An Evaluation of California's Inmate Classification System using a Generalized Regression Discontinuity Design

Richard A. Berk and Jan de Leeuw
Department of Statistics
UCLA

June, 1999[1]

[1]The research reported in this paper would have been impossible to undertake without the full cooperation the California Department of Corrections. Indeed, key staff were more like collaborators than clients, and we learned an enormous amount working with them on a regular basis. Thanks also go to the students and staff of the UCLA Statistical Consulting Center who at various times worked in the project.

Abstract

Published studies using the regression discontinuity design have been limited to cases in which linear regression is applied to a categorical treatment indicator and an equal interval outcome. This is unnecessarily narrow. We show here how a generalization the usual regression discontinuity design can be applied in a wider range of situations. We focus on the use of categorical treatment and response variables, but we consider the more general case of any regression relationship. We also show how a resampling sensitivity analysis may be usefully employed to address the credibility of the assumed assignment process. The broader formulation is applied to an evaluation of California's inmate classification system, which is used to allocate prisoners to different kinds of confinement.

1 Introduction

With the rapid growth of prison populations and firm budget constraints imposed by state legislatures, prison systems across the country have been looking for measures that might improve their efficiency. California is no different, and the California Department of Corrections (CDC) has been under pressure to seek to new ways to get more for less. Among the strategies being considered are methods to house prisoners so that more costly, higher security beds are allocated only to prisoners who truly need them. A first step would be an evaluation of how well the system currently allocates inmates to incarceration facilities.

In July of 1996, we were asked by the CDC to provide an analysis of how inmates are currently screened and placed. The key criterion for effective placement was defined by CDC as inmate misconduct in prison. Serious misconduct can substantially disrupt prison operations and put inmates and prison staff in harm's way. We were not asked to address later misconduct in the community outside of prison because the issues and research designs are quite different and would require rather different research designs.

Two specific questions naturally followed:

1. How well do current placement methods sort inmates by their potential for misconduct?

2. How effective currently are different placements in controlling prisoner misconduct?

While these questions might seem simple enough, we had access only to observational data with which to provide answers. We raised the possibility of randomized experiments, but for a variety of practical reasons, randomized experiments were at the time out of the question. However, because inmates were most often placed through a computed "classification score," there was the real prospect of employing a regression discontinuity design (Cook and Campbell 1979; Berk and Rauma 1983, Trochim 1984). That is, CDC's most common placement procedures assigned inmates to different kinds of housing on the basis of a known covariate. Under such circumstances, it is now well understood that conditioning on the assignment covariate alone can lead to unbiased estimates of treatment effects (Rubin 1977). Unfortunately, the outcome of interest was binary, and past applications of the regression discontinuity design had been limited to equal interval outcomes.

In response, we generalized in the usual regression discontinuity design to any regression function that is invariant across different interventions (in the categorical treatment case) or different doses (in the quantitative treatment case). We also developed a resampling sensitivity test to consider assumptions made about the assignment process. We report on these efforts below as part of our evaluations of CDC's inmate classification and placement system. A number of other issues and details can be found in our full report to the California Department of Corrections (Center for Statistics 1997a).

2 California's Inmate Classification System

2.1 A Brief Summary of CDC's Classification and Placement System

Each inmate is sent after sentencing to a CDC reception center. There, information is collected on a standardized form, including information thought to be related to the likelihood of later behavior problems in prison. The form is called either an 839 or 840 depending on whether the inmate is admitted following a new conviction or is being returned to custody to complete a sentence.

Among the items linked to potential behavior problems are an inmate's age, martial status, work history, prior CDC incarcerations, and the sentence length of the current commitment. Younger inmates, for instance, and inmates with longer sentences are believed to be more prone to misconduct.

The form is structured so that as the relevant information is entered, a simple formula may be applied to calculate an inmate's "classification score." This formula amounts to a linear combination of the items. Sentence length is the most influential component by far, accounting for almost 70% of the variance in classification score.

After the classification score is computed, placement may be undertaken by one of two procedures. If placement is to be determined by the classification score alone, the value of the score automatically leads to placement. The score range is divided into contiguous segments, and the segment into which a score falls defines the level of security required. For example, a score of 15 implies placement in a low security "Level I" facility. A score of 60 implies placement in a high security "Level IV" facility. Facilities are ranked from I to IV, with I the lowest security level and IV the highest security

level. CDC operates 32 incarceration facilities with several for each security level. Within each security level there are, in addition, often different kinds of placements having implications for misconduct. At the extreme, for instance, are "Secure Housing Units" located in Level IV prisons, which are used for holding unusually difficult or dangerous inmates. We will have a bit more to say about different placements within a given prison later.

Often the classification score is not used to place inmates. Rather, some feature of the inmate or the crime leads to "administrative placements" determined by CDC policy. For example, sex offenders are typically kept in higher security facilities because a successful escape, even if very unlikely, would be a public relations disaster. By and large, administrative placements are made regardless of the classification score; inmates subject to administrative placements are processed through an alternative apparatus. However, some administrative placements are "overrides" of placements derived solely from classification scores. The majority of these are "population overrides," which occur when there is no bed available at the inmate's score-designated placement. From year to year, around 25% of all placements are by administrative determination and they, along with outright errors in placement, create significant problems for the evaluation. We will address those complications below.

2.2 Past Research on Inmate Classification and Placement

Ours is certainly not the first study of inmate classification and placements systems (Gearing 1979; Austin 1986; Austin and Alexander 1996; Buchanan et al. 1986; Levinson 1988; Jones 1992; Proctor 1994; Cowles and Gransky 1996). Nor is ours the first evaluation of California's system (Finchmp 1988; California Department of Corrections 1986).

By and large, past studies have found small but consistent associations between inmates' classification scores and various measures of misconduct in prison. However, the studies are very uneven in quality, so typical results are not necessarily credible. The two California studies also find those associations and evidence that placement in higher security levels may reduce the likelihood of misconduct. And the two California studies are among the stronger efforts reviewed.

Still, none of the studies we examined were able to capitalize on a regression-

discontinuity design, especially with the enhancements we introduce below. And none had the advantage of large and rich dataset we were able to exploit.

2.3 Data Available

For the analyses we report below, we rely most heavily on the "c-file" dataset constructed by CDC for the evaluation. A total of 3000 inmates admitted early in 1994 were selected for the study. Beginning in January 1994, inmates admitted to CDC were identified until a total of 3000 was reached. The necessary total was achieved in several months. The date to start the data collection was chosen because it was the most recent time after which one would have at least 18 months of follow-up data to consider inmate misconduct. Eighteen months was considered a minimum follow-up necessary to evaluate the impact of inmate placements.

For each inmate sampled, the file folder containing all relevant paper records was sought. Selected information contained therein, not already available in CDC's electronic databases, was transferred to coding forms by CDC staff experienced in working with those records. For example, CDC's electronic databases did not contain any fields for prior arrests, and these could be found in the paper files. The coded data were then key entered and merged with CDC's electronic data. The result was a single, enlarged electronic record for each inmate in the study.

Unfortunately, many folders had missing information on one or more key variables, and some folders could not be found at all. The total effective sample, therefore, was 2,746. Comparisons between distributions of variables found in the electronic data for the full c-file dataset and found in the c-file subset were very similar, suggesting that the missing data effectively were missing at random. This result was not surprising since folders are typically lost or misplaced through clerical error unrelated to the contents of the folders. For example, folders are sometimes "misfiled," which makes them very difficult to find.

There was also keen interest in comparable information on inmate sentenced under California's recent "3-strikes" legislation. In brief, defendants with a prior conviction for a serious felony were subject to automatic and large sentence length enhancements. Such inmates were called "2-strikers" by CDC. Defendants with two prior convictions for serious felonies were subject to a mandatory life sentence. Such inmates were called "3-strikers" by CDC. CDC was concerned that 3-strikers at least, having little to lose, would

perhaps become very difficult inmates.

In response, 2-strike and 3-strike inmates were added to the c-file dataset. Since there were relatively few 3-strikers, virtually all of them were included. A total of 1,000 folders for 3-strikers was sought, and the final sample was of size 734. A total of 1,000 folders of 2-strikers was sought, beginning with the admissions date of January, 1994. The final sample was 771. Again, comparisons between the distributions of variables on the electronic dataset for the full samples and the for smaller samples suggested that data effectively were missing at random.

The need to include all 3-strike inmates meant that some were added to the study late in the data collection process. As a result, such inmates would have less "time-at-risk" to get into trouble, given the fixed ending date of the study. Such 3-strike inmates would also tend to have higher classification scores than other inmates and be more likely placed in higher security institutions. Thus, time-at-risk is a potential confounder. Note that for this study, time-at-risk is determined by when an inmate arrives at a CDC reception center, and that date necessarily precedes assignment to a prison bed. Time-at-risk cannot, therefore, be a consequence of the placement or the assignment process. And as we consider at some length shortly, if the regression discontinuity design applies, confounders such as time-at-risk are not a problem.

2.4 Inmate Misconduct

When an inmate is reported for some kind of serious misconduct, the reporting staff member fills out a form called a 115. The key information on those forms becomes part of CDC's electronic database. The violations are rather heterogeneous. At the low end are acts such as not standing to count, failure to obey an order, and failure to report for an assignment. While these are not by themselves acts of violence, they can be highly disruptive and if left unchecked, can lead to ungovernable institutions. Violence is then likely to follow. At the high end are acts such as trafficking in narcotics, fighting, assault on staff or inmate with a weapon, and inciting a riot. All of the acts at the high end are very rare and cannot be properly used alone as outcome variables. There was also no interest in distinguishing among the low end violations. So, with the agreement of CDC, we defined a "failure" as any violation recorded on a 115. As an empirical matter, inmates who engaged in the most serious forms of misconduct were also more likely to

engage in the least serious forms. The most difficult and dangerous inmates were substantially over-represented among all forms of misconduct. Thus, if the inmates were found who were at higher risk to any 115 violation, among them would likely be the really problematic cases.

Violations recorded on the 115 forms were fairly common. 22% of the 3-strikers, 46% of the 2-strikers, and 26% of the ordinary inmates had recorded 115's. The overall figure was 29%. On its face, it would seem that the 2-strikers are the most difficult inmates, but one must keep in mind that misconduct is in principle a function of the inclinations of an inmate and the security level to which he is assigned. Thus, 2-strikers may be no more inclined toward misconduct than 3-strikers, but 3-strikers, as "lifers," are typically housed in higher security settings. Recall that sentence length is a key component of the classification score.

The median classification score for the c-file data, including strikers, is 25. The interquartile range is 30. When the strikers are removed, the median is 21, and the interquartile range is 13. Clearly, including the strikers introduces a large number of inmates with higher than average classification scores. Approximately 1% of the classification scores had values larger than 79, some ranging well above 100. These outliers were deleted from the analysis, although the substantive conclusions are unchanged with them included. CDC felt that such high scores were likely to be errors or very atypical inmates.

Previous research on the California system (California Department of Corrections 1986) and our own analyses (UCLA Center for Statistics 1997a) suggested that of the four security levels, a comparison between level IV (the highest security level) and the other three would be the most instructive. For example, working with the entire inmate population from 1988 to 1996 and using only the electronic data, we focused on inmates with classification scores two points above or below the three thresholds between the four security levels. For all practical purposes, a four point spread is unrelated to the risks of misconduct so that within that range, inmates are effectively assigned at random to security level.

Nearly 40% of the inmates just below the level IV threshold engaged in misconduct, while only about 30% of the inmates just above the level IV threshold engaged in misconduct (for an odds ratio of about .65 of level IV compared to level III). The differences between the other levels were far smaller, which when coupled with past research and other evidence, suggested concentrating on the comparison between level IV and the other three levels.

A classification score of greater than 51 is required for placement in level IV. For the c-file dataset without strikers, 5.9% of the inmates were assigned to level IV facilities. When the strikers are added, 18.7% were assigned to level IV facilities. This differences has implications for statistical power that we will examine shortly.

3 Regression Discontinuity Design and Analysis

Given the two questions posed at the beginning of this paper and the ways in which CDC places inmates, it initially seemed reasonable to proceed within a regression discontinuity framework. The misconduct response variable could be regressed on inmate classification score and one or more binary variables representing the security level to which inmates are assigned. The relationship between the response variable and the classification score would address whether the classification score usefully sorted inmates by "risk." The relationship between the response variable and the binary variable(s) for treatment(s) would address whether placement might affect misconduct.

It is well known that if the relationship between an equal interval response variable and the assignment covariate is the same for experimental and control subjects, "assignment by a covariate" can lead to unbiased estimates of the average impact of the treatment (Rubin 1977). No other covariates need be considered.

The basic logic is as follows. Let y be a response variable (e.g., misconduct in prison). Let x be the covariate by which subjects are assigned to treatments (e.g., the classification score). Let z be a treatment indicator variable (e.g., one of two prison security levels). And let u be any other variable that may be related to both the treatment variable z and the response variable y (e.g., gang membership). That is, u is a potential confounder. Now, it is always true that $p(uz|x) = p(u|zx)p(z|x)$, where $p(.)$ is a probability density or probability mass function. But when assignment to treatments is by covariate x, $p(u|zx) = p(u|x)$, because z is a function of x. Thus, $p(uz|x) = p(u|x)p(z|x)$. Conditional on the assignment variable, the treatment indicator variable z and the confounder u are independent. As a result, one does not have to condition on u to obtain an unbiased estimate of the treatment effect, even though u may still be related to the response

after conditioning on x.

We shall now be far more specific and examine how this result can be applied in practice. We begin by considering for any experimental subject the joint distribution of the outcome and one or more assignment variables, given the treatment. There are an arbitrary number of discrete treatments or quantitative treatments such as doses. Each treatment $z \in \mathcal{Z}$ corresponds to a joint density (or probability mass function) $p(xy|z)$ of the outcome y and one or more assignment variables x. If all subjects received treatment z, then we would observe a simple random sample from $p(xy|z)$. But, the treatment does not influence the pre-treatment measurements, so must have $p(x|z) = p(x)$ for all z. Hence, we can work with the distribution of the outcome y given x and z because $p(xy|z) = p(y|xz)p(x)$.

In the regression discontinuity design, treatment $\underline{z}$ is a function of $\underline{x}$, say $\underline{z} = \phi(\underline{x})$, where the underlining indicates random variables. That is, x is a realization of $\underline{x}$, z is a realization of $\underline{z}$, and so on. This leads to truncated $p^\star(xy|z)$ versions of the densities, where

$$p^\star(xy|z) = \begin{cases} 0 & \text{if } x \notin \phi^{-1}(z) \\ \frac{p(xy|z)}{p(z)} & \text{if } x \in \phi^{-1}(z), \end{cases}$$

and where

$$p(z) = \int_{x \in \phi^{-1}(z)} p(x)dx.$$

The joint density after assignment is thus

$$p^\star(xy) = \int_z p(z)p^\star(xy|z)dz = p(xy|\phi(x)).$$

The likelihood for n independent trials is now simply

$$\mathcal{L}(\{x_i, y_i\}) = \prod_{i=1}^{n} p(x_i y_i|\phi(x_i)).$$

This can also be written as

$$\mathcal{L}(\{x_i, y_i\}) = \prod_{k=1}^{\ell} \prod_{i=1}^{n_k} p(x_i y_i|z_k),$$

where we observe $\ell \leq n$ different values $z_k \in \mathcal{Z}$, with frequencies $n_1, \ldots, n_k$. Since $\underline{z} = \phi(\underline{x})$, the n_k are realizations of random variables. But the likelihood is precisely the same if we design the experiment by fixing the n_k, and observe n_k realizations in each treatment group. Thus, maximum likelihood methods for multiple group independent trials can be used. This result is due to Visser and Deleeuw (1984).

We now continue our specification, by assuming that $p(y_i|x_i z_k)$ in the conditional likelihood

$$\mathcal{L}(\{y_i|x_i\}) = \prod_{k=1}^{\ell} \prod_{i=1}^{n_k} p(y_i|x_i z_k),$$

satisfies the usual canonical generalized linear model form (see McCullagh and Nelder 1989, section 2.2). We use an indicator variable (a "dummy") Z to code treatment, and we have $\mu = E(\underline{y}) = X\beta + Z\alpha$. Because of the canonical link, the statistics $b = X'y$ and $a = Z'y$ are sufficient for β and α. Note that there are multiple treatment realizations and multiple assignment variables. Under suitable conditions, the sufficient statistics will be normally distributed (even if the numbers in the treatment groups are random), and the maximum likelihood estimates are consistent and asymptotically normal.

This last result is the analogy to Rubin's, but in this case the treatment effects α are consistently estimated by applying the usual maximum likelihood (i.e. iterative generalized least squares) methods within the regression discontinuity design. For a binary outcome, this means applying logistic regression.

In application to follow, we assume the usual linear logistic relationship. We will regress an overall measure of misconduct in prison on two explanatory variables: inmate classification score and a binary variable indicating whether an inmate was assigned to the highest security level. We also considered a) what happens when inmates who are placed by administrative decisions are added to the mix and b) what happens if one allows for errors in placement.

4 Results

The upper section of Table 1 shows the results of a logistic regression in which the presence or absence of a failure is regressed on classification score and a binary variable for whether an inmate is housed in a level IV facility. The

outcome is coded so that "1" represents a violation and "0" represents no violation. The binary variable for placements is coded so that "1" represents placement in level IV and "0" represents placement in any of the lower levels.

In the upper section of the table are the regression coefficients, standard errors and odds multipliers for the full c-file dataset with administrative placements excluded. In principle, therefore, the regression-discontinuity design applies. The odds multiplier of 1.02 for the classification score implies that for every additional 10 points, the odds of misconduct are increased by a factor of 1.22. Thus, if placement did not affect misconduct, one would expect the odds of misconduct to be about twice as large for an inmates with scores placing them in level IV facilities than for inmates with scores placing them in level I facilities. This is an important difference from CDC's point of view. But the odds multiplier of .47 for level IV placements indicates that the inmates in level IV facilities have their odds of misconduct cut by a factor of about half compared to inmates placed in any of the lower security levels. Placement in level IV seems to matter. Indeed, the "suppressor" effect of level IV is estimated to approximately cancel out the increase in risk associated with level IV inmates.

These results depend on the assumed linear functional form in the log-odds. So, we tested whether quadratic and cubic terms in classification score were necessary. They were not. The model seemed to be linear in the log of the odds, which simplified the interpretation of the results.

How well does the regression discontinuity design live up to its advance billing? Consider again the potential confounder time-at-risk. If the regression-discontinuity design holds, controlling for time-at-risk should not alter the overall conclusions reported immediately above.

All inmates are reviewed about every 9 months, and the information about misconduct (including no misconduct) is recorded. That information is only obtained when the review is done, and an inmate must be behind bars at that time. Thus, for each inmate there is a first review, a second review and so on, until release. Within each review period, inmates have comparable time-at-risk.

We had access to a much larger dataset with which it was possible undertake separate analyses for each of several review periods and thus for each, condition on time-at-risk. These data included CDC inmates admitted between 1988 and 1995. The mix of inmates, population pressures, and administrative policies were quite different even 5 years ago, so the larger dataset is not fully comparable to the dataset we constructed for this study.

For example, there were no inmates sentenced under the 3-strikes statute. (For more information see Center for Statistics; 1997a.) Nevertheless, we analyzed the data by review period using the same approach described above. For the first review period, the odds multiplier for the level IV treatment was .71. For the second review period, the odds multiplier for the level IV treatment is .70. For the third review period, (which is beyond the length of the follow-up period we used above), the odds multiplier for the level IV treatment was .71. For the fourth review period, the odds multiplier for the level IV treatment was .77. All of the odds multiplier were several times larger than their standard errors and showed the same kind of treatment effect reported in Table 1. As anticipated, conditioning on time-at-risk does not alter the conclusion that placement in level IV housing seems reduce inmate misconduct.

The lower section of Table 1 shows the results when administrative determinants are included. The figures are almost the same indicating that including inmates not assigned by a known covariate does not change the conclusions. An examination of the reasons for administrative placements suggests why this is be so. A large fraction of the administrative placements resulted from practical exigencies such as too few beds in some facilities and too many beds in others. Moreover, these kinds of placements were most common among inmates who would not ordinarily be placed in level IV facilities. Thus, the shuffling occurred primarily *within the control group*, not between the control group and the experimental group.

However, the close inspection of "out-of-level placements" revealed a number of inconsistencies in the data. Some inmates were placed with no recorded rationale and some were placed in inappropriate facilities, given their classification score. Depending on the details of how one counts such problems, as many as 10% of the inmates may have been placed improperly out-of-level. Alternatively, much of the problem could result from errors in the data, not in what actually transpired. In data systems as large and complicated as CDC's, staffed by people of widely varying skill and motivation, a certain amount of inaccuracy is to be expected.

4.1 Sensitivity Analysis

In order to explore the possible implications of such errors for the credibility of our findings, we undertook a series of sensitivity analyses. Building on the spirit of Rosenbaum's work (Rosenbaum 1996), we conducted simulations of

Variable	Coefficient	Stand. Error	Odds Multiplier
Admins Out (N=3918)			
Level IV	-.761	.138	0.47
Score	.025	.003	1.02
Admins In (N=4251)			
Level IV	-.717	.149	0.48
Score	.024	.003	1.02

Table 1: Logistic Regression Results for Data with Administrative Determinants Included and Not Included

the impact of inmate misclassifications. Under the regression discontinuity design, assignment is fully determined given a classification score. If an inmate scores above the level IV threshold, the probability of assignment to the experimental group is 1.0. If an inmate scores at or below the level IV threshold, the probability of assignment to the control group is 1.0. But if assignment can sometimes occur by error, or in response to unobserved random variables, the assignment is no longer certain. And if the errors in assignment tend place inmates who belong in the experimental group in the control group and inmates who belong in the control group in the experimental group, bias can result. The assignment covariate no longer properly adjusts for "pre-existing" differences between the experimental subjects and the control subjects.

To simulate the implications of different levels of biasing misassignment, we examined an assignment process in which the probability (π) that the experimental group would be reassigned to the experimental condition was less than 1.0. Likewise, we examined an assignment process in which the probability that the control group would be reassigned to the control condition was less than 1.0. In effect, we were simultaneously "diluting" the experimental and control groups.

We could have also simulated conditional misassignment probabilities. For example, we might have allowed the misassignmeant probabilities to depend upon the computed classification score, so that inmates nearer the threshold between a level IV assignment and a level III assignment would have a higher probability of misassignment. However, we had no information that this was a plausible assumption. And the unconditional approach we

chose to implement was actually a more telling test. Moving inmates near to the threshold to one or the other side would not make much of a difference in the results. Moving a sufficient number of inmates far away from the threshold might.

This means the joint distribution of $\underline{x}$ and $\underline{z}$, which used to be

$$\begin{pmatrix} & \underline{z}=0 & \underline{z}=1 \\ \underline{x}<x_0 & p(x) & 0 \\ \underline{x}\geq x_0 & 0 & p(x) \end{pmatrix},$$

now becomes

$$\begin{pmatrix} & \underline{z}=0 & \underline{z}=1 \\ \underline{x}<x_0 & \pi p(x) & (1-\pi)p(x) \\ \underline{x}\geq x_0 & (1-\pi)p(x) & \pi p(x) \end{pmatrix}.$$

The log likelihood is now

$$\mathcal{L}=\sum_{i=1}^{n}\log\text{prob}(\underline{y}_i=y_i\ \wedge\ \underline{x}_i=x_i\ \wedge\ \underline{z}_i=z_i)=$$
$$\sum_{i=1}^{n}\log\text{prob}(\underline{x}_i=x_i)+\sum_{i=1}^{n}y_i(\alpha+\gamma z_i+\beta x_i)-\log\{1+\exp(\alpha+\gamma z_i+\beta x_i)\}+$$
$$n_+\log\pi+n_-\log(1-\pi),$$

where n_+ and n_- are the number of cases for which $\underline{x}$ and $\underline{z}$ are and are not "in agreement." Clearly, ordinary logistic regression is misspecified, and will lead to biased estimates.

We used three different reassignment probabilities π: .95, .85, and .80. These applied to both the experimental and control groups and represented the probability of reassignment to the observed assigned group: experimentals to the experimental group and controls to the control group, respectively. For each probability, 100 trials were simulated, and the logistic regression coefficients and their standard errors stored. The results of these simulations can be found in Tables 2 through 5.

In each table, the first set of results (for π=1.0) serve as a baseline, since they represent the outcome when the data are simply taken at face value; no simulation is performed. Following are the results for probabilities of .95, .85 and .80 for both the experimental and control groups. For each of these

Variable	Coefficient	Stand. Error	Odds Multiplier
For π=1.0			
Level IV	-.732	.149	0.48
Score	.024	.003	1.02
For π=.95			
Level IV	-.431	.119	0.65
Score	.019	.002	1.02
For π=.85			
Level IV	-.189	.092	0.83
Score	.015	.002	1.02
For π=.80			
Level IV	-.146	.085	0.86
Score	.014	.002	1.01

Table 2: Sensitivity Analysis for Full Sample of Inmates Placed by Classification Score: Means for 100 Simulation Trials and Different Misclassification Probabilities (N=3918)

simulations, the average regression coefficient, average standard error and average odds multiplier are shown. Each is based on the 100 replicates.

The tables differ in the database used. We begin with the full c-file dataset and then drop, respectively, the 2-strikers, the 3-strikers, and both sets of strikers. The goal of looking at subsets of the data is to explore the robustness of our findings in the face of possible interaction effects with striker status. For each analysis, inmates placed by administrative decisions are not included. Despite the results in Table 1, we felt that including administrative determinants would have unnecessarily complicated matters.

Beginning with Table 2, we see that as the experimental and control groups are increasingly diluted, the estimated size of the level IV effect declines. The standard errors do not not substantially change, but we "lose" the effect by about the time the probability of reassignment reaches .80. In contrast, while the relationship between classification score and misconduct is also reduced, it remains large relative to its standard error and substantively important. In short, the treatment effect is lost when about 20% of the controls are misplaced as experimentals and about 20% of the experimentals are misplaced as controls. Whether this is a likely level of error in practice

Variable	Coefficient	Stand. Error	Odds Multiplier
For π=1.0			
Level IV	-.718	.172	0.48
Score	.024	.003	1.02
For π=.95			
Level IV	-.429	.139	0.65
Score	.019	.002	1.02
For π=.85			
Level IV	-.196	.101	0.82
Score	.015	.002	1.02
For π=.80			
Level IV	-.161	.099	0.85
Score	.014	.002	1.01

Table 3: Sensitivity Analysis for the Sample of Inmates Placed by Classification Score with 2-Strikers Dropped: Means for 100 Simulation Trials and Different Misclassification Probabilities (N=3147)

is a question to which we will return.

In Table 3, the 2-strikers are dropped from the dataset. While 771 cases are deleted, very few are lost from level IV placements. Only 8% of 2-strikers were assigned to security level IV. And one can see that the results look much the same as those in Table 2. Eliminating the 2-strikers does not change the overall conclusions.

Table 4 shows that when we drop the 3-strikers rather than the 2-strikers, the story changes. Deleting the 3-strikers means that 735 cases are lost, but more important, 72% of them are level IV placements. Less than 10% of the inmates are now in the experimental group, which at least implies a substantial loss of statistical power for estimates of the treatment effect. While the association between the classification score and misconduct holds firm, estimates of the impact of level IV placements are small in absolute size and small relative to the standard errors (although in each case the sign remains negative).

In Table 5 both sets of strikers are dropped. Once again, the importance of classification score holds, but the impact of a level IV placement is estimated to be small. A total of approximately 1500 cases has been dropped

Variable	Coefficient	Stand. Error	Odds Multiplier
For π=1.0			
Level IV	-.115	.204	0.89
Score	.039	.004	1.04
For π=.95			
Level IV	-.062	.143	0.94
Score	.039	.003	1.04
For π=.85			
Level IV	-.020	.105	0.98
Score	.038	.003	1.04
For π=.80			
Level IV	-.015	.095	0.99
Score	.038	.003	1.04

Table 4: Sensitivity Analysis for Sample of Inmates Placed by Classification Score with 3 Strikers Dropped: Means for 100 Simulation Trials and Different Misclassification Probabilities (N=3187)

Variable	Coefficient	Stand. Error	Odds Multiplier
For π=1.0			
Level IV	-.287	.254	0.75
Score	.042	.002	1.04
For π=.95			
Level IV	-.118	.177	0.89
Score	.040	.004	1.04
For π=.85			
Level IV	-.022	.127	0.98
Score	.039	.003	1.04
For π=.80			
Level IV	-.065	.117	0.94
Score	.039	.003	1.04

Table 5: Sensitivity Analysis for the Sample of Inmates Placed by Classification Score with 2 Strikers and 3 Strikers Dropped: Means for 100 Simulation Trials and Different Misclassification Probabilities (N=2416)

but perhaps more important, less than 6% of the sample now falls in the experimental group.

5 Discussion

It is clear that the CDC classification score is associated with inmate misconduct for the full dataset and when the strikers are dropped from the analysis. The association also remains in our simulations based on different probabilities of reassignment. Clearly, the relationship is quite robust to errors in inmate placement. In short, the classification scores seems to perform roughly as its designers intended.

The story for estimates of the impact of level IV placements is more complicated. Our sensitivity analysis shows that even large estimates of level IV effects can disappear if the regression discontinuity design is degraded through misassignment. If the reassignment probability drops much below .80, null findings can dominate. Consequently, a key question is whether reassignment probabilities below .80 are likely. The evidence we have suggests they are not. First, when placements we inferred from inmate classification scores were compared to placements recorded in the data (with administrative placements removed), disparities were found in far less than 10% of the placements. That is, our derived placements agreed with the recorded placements more than 90% of the time. Unfortunately, it is possible that some unknown number of the matches were false positives if both classification score and placement were inaccurately recorded. Second, spot checking done when the inmate records were coded for analysis revealed relatively few data errors for classification score and placement. Finally, while CDC's data are certainly not free of error, those errors tend to be found in fields that are not administratively essential. Classification score and placement are among the very most important fields because so much is at stake for both CDC and the inmate. In short, the weight of evidence suggests that misclassification rates of more than 10% seem to be unlikely.

The apparent interaction effects are a bit more curious. After conditioning on classification score, inmate placement under the regression discontinuity design is uncorrelated with all "pre-existing" variables, including striker status. Confounding with the treatment variable is eliminated. But it appears that treatment may be especially effective for 3-strikers. A simple explanation was suggested earlier; when the 3-strike inmates are eliminated from the

dataset, there is very little variance left in the treatment variable and consequently, very little statistical power. Thus, a null finding when 3-strikers are eliminated from the dataset may be just what one should expect.

Alternatively, there could be something about *other* interventions imposed on 3-strikers that makes them better risks. However, we have been unable to discover what those other interventions might be. All 3-strikers are treated the same way as other prisoners. There are no special constraints imposed nor special programs of any kind. There is also no evidence that misconduct is defined less broadly for 3-strikers or that they face stiffer punishments when caught. For example, confinement in a "special housing unit" is not more common among 3-strikers, and confinement in a special housing unit does not seem to be related to the likelihood of misconduct anyway, once security level is taken into account.

Moreover, there does not seem to be anything special about the backgrounds of 3-strikers that might enhance the possible impact of level IV placements. Indeed, all 3-strikers were at one time 2-strikers, and 2-strikers are not especially good risks. In fact, they may be worse risks than either 3-strikers or the general population of inmates (Center for Statistics 1997b). In addition, as we noted earlier, sentence length is actually associated with *increases* in risk, other things equal (Center for Statistics 1997a), and 3-strikers are clearly facing very long terms. Finally, CDC research cited earlier undertaken well before the 3-strikes statute was passed also found level IV effects. That is, there seem to have been level IV suppressor effects before the advent of the 3-strikes legislation.

In summary, when the regression discontinuity design is intact, and when there is sufficient variance in the treatment variable, the balance of evidence supports an interpretation in which assignment to level IV reduces the odds of misconduct.

6 Conclusions

We have shown in this paper how a simple extension of regression-discontinuity designs can be applied to research on prison classification systems. Our extension involves the use of binary response variables, although we also considered the more general regression case. When we applied the generalized regression-discontinuity design in an evaluation of the inmate classification system used by the State of California, we concluded that the existing clas-

sification scores usefully sort inmates by levels of risk and that the highest security level may well reduce the odds of misconduct compared to placement in the lower three security levels.

However, a far more powerful study addressing these and other related issues is now underway. The California Department of Corrections has a large randomized experiment in progress testing the old classification system against a new one we helped design. A total of 20,000 inmates have been assigned at random, 10,000 placed into CDC institutions using the old classification system and 10,000 placed into CDC institutions using the new classification system. Inmates are being followed for 18 months with a combination of CDC's usual data collection instruments and new instruments that will collect information not previously available in a systematic form. Preliminary results from a 6-month follow-up will be provided to the State legislature early in the year 2000 with final results available early in 2001.

7 References

Austin, J. (1986), "How Well is your Classification System Operating: A Practical Approach," *Crime and Delinquency*, 32, 326-338.

Austin, J. and Alexander, J. (1996), *Evaluation of the Colorado Department of Corrections Inmate Classification System*, Colorado Department of Corrections (unpublished manuscript).

Berk, R.A. and Rauma, D. (1983), "Capitalizing on Non-Random Assignment to Treatment: A Regression Discontinuity Analysis of a Crime Control Program," *Journal of the American Statistical Association*,78, 21-28.

Buchanan, R.A., Whitlow, K.L. and Austin, J. (1986), National evaluation of objective prison classification systems: the current state of the art. *Crime and Delinquency* **32**, 3, 341-356.

California Department of Corrections. (1986), *Inmate Classification System Study: Final Report*, California Department of Corrections (unpublished manuscript).

Center for Statistics. (1997a), *An Evaluation of CDC's Prisoner Classification System: Part 1 — The General Population of Inmates*, Los

Angeles, CA: UCLA Center for Statistics (unpublished manuscript).

Center for Statistics. (1997b), *An evaluation of CDC's prisoner classification system: Part III — Inmates Sentenced under California's "Three-Strikes" Legislation*, Los Angeles, CA: UCLA Center for Statistics (unpublished manuscript).

Cook, T.D, and Campbell, D.T. (1979), *Quasi-Experimentation Design and Analysis Issues for Field Settings*, Skokie, Il: Rand McNally.

Cowles, E. and Gransky, L. (1996), *A Review of Incarcerated Illinois Class 4 Felony Offenders: Are Alternative Sanctions Appropriate?*, Center Legal Studies, University of Illinois at Springfield (unpublished manuscript).

Finchamo, D.A. (1988), *An examination of the Validity of the California Department of Corrections' Custodial Classification System*, Dissertation, Claremont Graduate School.

Gearing, M.L. (1979), "The MMPI as a Primary Differentiator and Predictor of Behavior in Prison: A Methodological Critique and Review of the Recent Literature," *Psychological Bulletin*, 86, 451-475.

Jones, N. (1992), *Inmate Classification Procedures: An Evaluative Analysis (North Carolina Division of Prisons' Procedures Compared with Established Goals and Objectives and with Standards Presented by the National Institute of Corrections Model)*, North Carolina Department of Corrections (unpublished manuscript).

Levinson, C.L. (1988), "Developments in the Classification Process: Quay's AIMS Approach," *Criminal Justice and Behavior*, 15, 34-44.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, Second Edition. New York: Chapman & Hall.

Proctor, J.L. (1994), "Evaluating a Modified Version of the Federal Prison System's Inmate Classification Model: An Assessment of Predictive Validity," *Criminal Justice and Behavior*, 21, 198-207.

Rosenbaum, P. (1996), *Observational Studies*, New York: Springer-Verlag.

Rubin, D.B. (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 34-58.

Trochim, W.M.K. (1984), *Research Design for Program Evaluation: the Regression Discontinuity Approach*, Beverly Hills, CA: Sage Publications.

Visser R.A., and de Leeuw, J. (1984), "Maximum Likelihood Analysis for Generalized Regression-Discontinuity Design," *Journal of Educational Statistics*, 9, 45-60.