Title
Conducting a Randomized Field Experiment for the California Department of Corrections: The Experience of the Inmate Classification Experiment

Author
Berk, Richard A

Publication Date
2004-01-12

Conducting a Randomized Field Experiment for the California Department of Corrections: The Experience of the Inmate Classification Experiment*

Richard Berk
Department of Statistics
UCLA

January 12, 2004

1 Introduction

The California Department of Corrections (CDC) has big problems. It houses more prisoners than any other state's corrections system: 160,000 inmates in 33 prisons and over 50 other facilities. The costs are enormous, including an average of about $30,000 per inmate per year and about $150,000 for each new cell built. The prisons are also very difficult to run. Each year about 25% of the inmates engage in some form of misconduct serious enough to document, and 2.5% commit an offense that would probably be a felony in the outside world.

One of the ways in which the CDC attempts to make the best use of its resources is to assign prisoners to facilities with varying levels of "security." Higher levels of security place more restrictions on inmates because greater human and physical resources are brought to bear. There are higher staff to inmate ratios and physical surroundings that reduce the chances of serious

*This paper was written for the NRC Panel on Evaluation Research for Criminal Justice Programs funded by the National Institute of Justice.

infractions. For example, in some high security facilities, inmates are housed one to a cell and are only allowed into the exercise yard in small groups. However, higher security facilities are more costly to build and run. It is important, therefore, to place each inmate in the least restrictive setting necessary to insure the well-being of that inmate, other inmates, and CDC personnel.

This paper discusses the implementation of a very large, randomized field experiment testing two different procedures through which inmates could be assigned to facilities with different security levels. By most any measure, the experiment was implemented in a textbook fashion and led to useful results. The question addressed is how this success was achieved.

2 Background to the Experiment

For several decades, the CDC has assigned inmates to "beds" through an objective "Inmate Classification System." Shortly after intake at the CDC Reception Center, background information is collected on each inmate: age, length of sentence, nature of the crime, prior incarceration in a CDC facility or under the California Youth Authority, and the like. This information is used to construct a classification score.

For about 75% of the inmates, the score determines placement in one of four security levels. For example, a score of 52 or greater means placement in a Level IV prison. A score between 28 and 51 leads to placement in a Level III prison. About 25% of the inmates are because of special concerns processed outside of classification system as "administrative placements." For example, all offenders sentenced to life without the possibility of parole (LWOP) are automatically assigned to at least level III settings. One rationale is that such inmates have little to lose by being difficult.

In the early 1990s, the UCLA Statistical Consulting Center was asked to evaluate the existing inmate classification system and to suggest possible improvements. Using a generalized regression discontinuity design and other approaches, we found that by and large the classification system was working as intended, but that a number of refinements could be made (Berk and de Leeuw, 1998). Those refinements included using several new background items to construct the classification score (e.g., gang activity), eliminating a few items that were not associated with conduct in prison (e.g., marital status), changing a bit the weights given to particular items (e.g., weighting

age more heavily), and making the rationale for administrative placements far more explicit by reformulating them as "mandatory minimums" (e.g., for LWOP prisoners).

These changes naturally led to the question of what impact the new system would have. In particular, would the new scoring system allocate inmates to beds so that there would be too many beds in some security levels and too few beds in others? While there is some flexibility in determining a given facility's security level, there are also very important constraints. For example, it would be very expensive and effectively impractical to construct a lethal perimeter around a work camp or to construct prisoner cells within a dormitory. A second issue was whether the revised classification system would sort inmates better by their predicted risk. For example, would greater distinctions be made between inmates who were likely to cause problems and inmates who would be unlikely to cause problems? A final question was what impact the new system would have on inmate misconduct. As important as the overall number of incidents and their nature, was where those incidents would occur. For example, might there be a reduction in infractions in Level I facilities?

Beginning in November of 1998, a randomized field experiment was launched to evaluate the revised classification system. Over 20,000 inmates admitted over the next 6 months were placed through the existing classification system or the revised system. Half were assigned at random to the old system and half were assigned at random to the revised system. The follow-up period for each inmate was 24 months from the date of admission. Details of the design can be found in Berk et al. (2003).

It is important to stress that while the experiment was designed by our group in the Department of Statistics at UCLA, and the data were analyzed by that same group, the study was conducted by CDC personnel. During the course of the study, there was at UCLA ongoing monitoring of the experiment's implementation, but the day-to-day work of of running the study was in CDC's hands.

3 Findings of the Experiment

The experiment provided a rich set of conclusions.

1. The revised inmate classification forms were well received by CDC staff. They made sense, reflected changes consistent with common under-

standings (e.g., gang activity really mattered), and were easier than the old forms to use.

2. The process of converting administrative placements under the old system to mandatory minimums under the revised system was a success. The rationale for placements outside of the an inmates classification score were now easily understood by inmates and CDC personnel.

3. Classifications scores were about 4 point higher on the average under the revised system. The increase was caused by small increases in score values at the lower ranges (i.e., less than 20 points). The lower tail was moved a bit to the right.

4. Table 1 shows that the majority of inmates would be placed the same under both systems. For those who would have been placed differently, the shifts were typically only one level up or down. (Table not shown).

5. Table 2 shows that overall, there was under the revised system a net decline in the number of inmates initially assigned to Level I facilities and a net increase in the number of inmates initially assigned to Level III facilities. One implication was there might be some space problems under the revised system: too many Level I beds and too few Level III beds. RC is the Reception Center, CCF is Community Corrections Facilities, and SHU is the Secure Housing Unit. CCF is considered a Level I placement, and the others are not relevant for this discussion.

6. Tables 3 and 4 show that the revised classification score sorted inmates substantially better by level of risk. The analyses are based on a generalized regression discontinunity design. The odds multiplier for classification score is 1.09 for the experimentals and 1.06 for the controls. Consider two inmates who differ in score by 20 points. For the experimentals, the inmate with the higher score has an odds of misconduct that is 5.60 times greater (i.e., $1.09^{20} = 5.60$) than the inmate with the lower score. For the controls, the 20 additional points translates into risk that is only 3.20 times greater (i.e., $1.06^{20} = 3.20$).

7. The revised classification score sorted inmates substantially better by level of risk when serous misconduct was the sole concern, not all misconduct. (Tables not shown.)

8. Tables 3 and 4 show that under both the existing and revised classifications systems, there was strong evidence that placement in a Level IV facility, compared all of the other facilities reduced the amount of misconduct substantially.

9. As shown in Table 5, under the revised classification system, the incidence of misconduct declined a bit in Level I settings and increased a bit in Level III setting. More difficult inmates were moved upward in the system taking their proclivity for misconduct with them. There were no changes in the incidence of misconduct for Levels II and IV. (See Berk et al., 2003, for a detailed discussion.)

10. One can use ensemble methods in statistics to predict rather well the very few inmates who are likely to commit the most serious offenses (e.g., assault). (See Berk and Baek, 2003, for the details.)

With these key findings in hand, the CDC almost immediately began the process of shifting to the revised system. However, many levels of approval were necessary, a large number of administrative regulations had to be changed, CDC staff needed to be retrained, new forms were required, and new data systems had to be constructed. The transition was expected to take about a year and as of now, is about half completed.

4 Why Did the Experiment Succeed?

There were a larger number of related reasons why the study worked. Each will be briefly considered.

4.1 "Preconditions"

1. "Over-incarceration" had been a very salient topic in the state legislature, the legislative analysts's office, and among major stakeholders. There were two issues: the high costs of placing inmates in more restrictive settings than necessary and the ethical issues this raised. The notion was that if a careful study was done, "over-incarceration" could be reduced substantially.

2. In the 1990s, there were many public relations disasters for the CDC, including several very embarrassing escapes and grisly crimes committed by former inmates on parole. CDC was also gaining a reputation for stalling when asked for potentially damaging information and even for fabricating data. I participated in several meetings in which legislative staffs openly accused CDC officials of stonewalling on key information, and in more informal conversations staff from various fiscal oversight agencies was told that often the "numbers" from the CDC did not "add up." As a result, the CDC needed some credibility. Their future budgets would depend in part on regaining the confidence of key legislators, government watchdog agencies, and stakeholders.

3. The UCLA Statistical Consulting Center had immediate credibility. It had legitimacy a "straight shooter" with no vested interest in the outcome. It had the reputation of doing sophisticated statistical analyses and being able to make those analyses accessible to policy makers. The Center was seen as part of the state "team" because it was housed in the state University System. Other possible collaborators were dismissed as carpet baggers, technically suspect, or ideologically tainted. The Center also had extensive hands-on experience working with city, county, and state agencies and understood a lot of the politics. Finally, the Center was known to be uninterested in grandstanding but would speak up if asked.

4. The CDC assigned excellent people to the project, who were smart, knowledgeable and thoroughly professional, and them let them do the work. There was no interference from higher up along the way.

5. Funding from the legislature was available with most of it going the CDC. The UCLA Statistical Consulting Center got a very small piece of the pie (but enough to cover costs). This pleased the CDC and further enhanced the Center's credibility.

6. CDC was prepared to operate with a long time horizon of over 5 years. This allowed for a several studies as precursors to the experiment. It also permitted good working relationships to develop as needed and meant that findings from earlier work could usefully inform the design of later work. However, the long time horizon did not materialize all at once. Rather, the work progressed in three steps, each taking well over

a year: 1) an initial evaluation of the classification system using observational data, 2) the design of a revised classification system "tested" with simulations, and 3) a randomized experiment comparing the existing classification system to the revised classification system. Step 2 evolved from the experience with step 1, and step 3 evolved from the experience with step 2. Contracts were written one step at a time; none of the parties knew there would be a step 2 or a step 3 when the work began.

4.2 Project Continuity and Maintenance

While over the course of the project there was some personnel turnover, continuity was well maintained. Several key people were involved in the project from start to finish, and before any experienced people left, they trained their replacements. The training included not just the details of the job to be done, but extensive background information including the politics and history of the study. As a result, the project's institutional memory was maintained.

There was also, in effect, a fire wall between the money allocated to the project and other CDC uses. When the legislature provided the necessary funding, the funding was earmarked for the research. It could not be reallocated to other CDC needs.

4.3 RCTs as a Method

The randomized experimental design was easily explained and credible as an evenhanded way to learn what impact the revised classification system might have. It just made good sense that if you wanted to find out how well something worked, you went out and tried it. And the random assigned was easily seen as a fair lottery in which each inmate had the same chance of being placed under the new system or the old, and in which the mix of inmates under the two systems would be approximately the same.

4.4 Feasibility and the CDC

Under California statutes, the CDC has the right to conduct "pilot studies" involving up to 10% of the inmate population. Once the experiment was labeled a pilot study, there were effectively no legal obstacles. However, before

proceeding, the CDC did its political homework. Several meetings were held with stakeholders to explain the project and address any concerns. All of these gatherings went well, once the study and its rationale were explained.

The study was also explained to each new inmate who was to be included. Each was told about the experiment and that he or she would be at random assigned to a CDC facility under the existing system or a revised one. But, the CDC is a total institution. Once a decision was made to proceed with the experiment, inmates had little choice but to cooperate. While inmates could certainly have protested and even tried to bring legal action, the study was completed without any overt complaints. Given all of the other concerns new inmates have, it is likely that the study was an easily overlooked detail.

The CDC is in an important sense a paramilitary institution as well. Orders are given and at least formally followed. But just as with police departments, there are ways to subvert orders if they are seen to be unreasonable. For this study, a great effort was made to get CDC staff not just on board, but to buy into the research. The pitch was that the new forms would be easier to use, would better reflect risk factors that "everyone knows" are important, and ultimately reduce the dangers faced by inmates and corrections officers. This message was delivered by experienced and well-respected CDC staff, not members of the UCLA Statistical Consulting Center. A sincere effort was also made to elicit suggestions from CDC staff about how the study might be most effectively fielded and about ways to improve the classification instrument. In the end, support from rank-and-file CDC personnel was exemplary.

4.5 The Role of the UCLA Statistical Consulting Center

From the start, all important decisions were to be made by the CDC. It was their study, and they would have to live with its consequences. The job of the Statistical Consulting Center was consulting. When called upon for its expertise, the intent was to explain the range of options, specify the possible consequences of each choice, and then let the Department decide what to do. However, the CDC's credibility with stakeholders, legislators, and the governor's office depended heavily on the Consulting Center's continued participation and its willingness to defend publicly all technical decisions. As a result, there was not a single instance in which after a healthy give and take,

decisions affecting the study were made without a solid consensus. In the process, CDC staff involved in the study became remarkably sophisticated about a wide variety of statistical issues and Consulting Center staff became no less sophisticated about corrections policy and practice.

4.6 Mistakes Made

While it is certainly possible to find in retrospect some things that might have been done better, none that have surfaced to date would likely have had any important impact on the study's results. For example, our power analyses undertaken before the study began were probably a bit too conservative. We could have managed with somewhat smaller sample, perhaps as few 15,000 inmates rather than a bit more than 20,000. But because most of the required data were already being collected as part of the CDC's routine management procedures, the marginal cost of each addition inmate was small. The main expense came from a variety of data quality control procedures for each case, some of which were labor intensive.

5 General Lessons about RTCs in Criminal Justice Research

Perhaps the major lesson is that a necessary condition for effective randomized field experiments is that a long time horizon is essential. For the inmate classification experiment, the research took over seven years. About half that time was spent doing what later turned out to be preliminary studies drawing on observational data. It was this work that provided a rationale for the experiment and ideas for how the existing inmate classification system could be improved. Equally important, working with the CDC staff during this period established the mutual expertise and trust on which the experiment depended. The importance of this kind of human and social capital is too often overlooked when project timetables are constructed.

But in the end, you also have to be lucky. For the inmate classification experiment, all of the essential pieces were in place as the study began and during the course of the project, Murphy's Law seemed not to apply. In particular, had the current budget crisis in California occurred 5 years earlier, it is very unlikely that the experiment would have been undertaken.

Initial Placement	Experimentals	Controls
RC	100.00%	100.00%
CCF	99.93%	94.24%
Level I	97.33%	68.20%
Level II	81.06%	83.19%
Level III	52.51%	85.40%
Level IV	88.63%	81.23%
SHU	100.00%	100.00%

Table 1: Percentage of Experimental and Control Inmates for whom the Actual Initial Placement was the same as the Hypothetical Initial Placement

Initial Placement	Controls	Experimentals	Total
RC	2.25%	2.55%	2.40%(463)
CCF	15.82%	13.82%	14.82%(2863)
Level I	33.51%	25.54%	29.53%(5705)
Level II	30.36%	31.92%	31.14%(6016)
Level III	12.55%	21.42%	16.99%(3282)
Level IV	5.24%	4.46%	4.85%(937)
SHU	0.26%	0.29%	0.27%(55)
Total	100%(9656)	100%(9662)	100%(19318)

Table 2: Initial Placements for the Experimentals and Controls Separately

Predictor	Coefficient	Std. Error	Multiplier
Score	.080	.005	1.09
Level II	-0.08	.075	0.92
Level III	-0.27	.117	0.76
Level IV	-2.99	0.27	0.05
Constant	-2.25	.010	–

Table 3: Misconduct Logistic Regression for Experimental Inmates – Level I is the References Category and Mandatory Placements are Excluded (N=6121)

Predictor	Coefficient	Std. Error	Multiplier
Score	.059	.005	1.06
Level II	0.05	.095	1.05
Level III	-0.78	.163	0.46
Level IV	-2.27	.295	0.10
Constant	-1.38	.075	–

Table 4: Misconduct Logistic Regression for Control Inmates – Level I is the Reference Category and Administrative Placements are Excluded (N=5177)

Initial Placement	Experimentals	Controls
Level I	29%	34%
Level II	30%	33%
Level III	53%	48%
Level IV	50%	52%

Table 5: Percentage of Experimental and Control Inmates Engaging in Misconduct by Initial Placement

6 References

Berk, R.A. and J. de Leeuw (1998) "An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinuity Design." *Journal of the American Statistical Association*, Volume 94, Number 448: 1045-1052.

Berk, R.A., Ladd, H., Graziano, H. and J. Baek (2003) "A Randomized Experiment Testing Inmate Classification System." *Journal of Criminology and Public Policy*, 2, No. 2: 215-242.

Berk, R.A., and J Baek (2003) "Ensemble Procedures for Finding HighRisk Prison Inmates." UCLA Department of Statistics preprint #367.