A Randomized Experiment Testing Inmate Classification Systems*

Richard A. Berk
Heather Ladd
Heidi Graziano

Department of Statistics
UCLA

April 26, 2002

1 Introduction

The California Department of Corrections (CDC) currently houses approximately 160,000 inmates in 33 institutions, 16 community corrections facilities (CCFs), 41 camps, and 8 prisoner mother facilities across the state. These facilities differ in many ways, including architectural design and construction, staffing, and program availability. However, they are each mandated to ensure public safety and institutional security.

A wide variety of housing options are provided within four levels of security. Inmates deemed to be most problematic are placed in the most restrictive settings requiring celled housing, a lethal perimeter, controlled movement and armed supervision within the housing units and dining halls. Inmates identified as less dangerous to staff, other inmates and the public are placed

*The research reported in this paper would have been impossible without the talents and efforts of our colleagues at the California Department of corrections: George Lehman, Maureen Tristan, Gloria Rea, Penny O'Daniel, Micki Mitchell, Mark Cook, and Terrence Newsome. They implemented the experiment and the data collection, as well as providing extensive comments on an earlier draft of this paper.

in less restrictive facilities, that can include dormitory housing, a non-lethal perimeter, and unarmed oversight.

Each housing option has a designated security level. The CDC uses an inmate classification score system to evaluate each inmate's need for supervision. The fundamental goal of the CDC classification system is to place an inmate in the least restrictive security level consistent with internal security and public safety. While the average cost for housing an inmate in CDC is over $25,000 a year, costs for more restrictive level of housing are significantly more expensive.

An earlier study conducted for CDC (Berk and de Leeuw, 1998) evaluated the existing inmate classification score system by which inmates were classified and then placed in different levels of security. While the data suggested that overall the procedures were placing inmates roughly consistent with Department's expectations, there was also evidence that improvements in the system could be made.

In this paper, we address how the inmate classification system was revised and then describe a randomized experiment to test the new system against the old one. Over 20,000 inmates took part in the experiment, which included two years of follow-up data. Key outcomes to be examined were the amount and type of misconduct in prison[1] and the implications of the new system for prison crowding; would the current distribution of beds by security levels suffice if inmates were distributed to levels differently? The study is one of the largest randomized trials ever undertaken, and certainly the very largest criminal justice randomized experiment. We will focus not just on the design, but its implementation in a prison setting where placement decisions can have very serious consequences. Finally, we report the central findings and policy implications.

2 Summary of the CDC's Current Classification System

The vast majority of inmates begin their sentences at a reception center where for each, a substantial amount of background information is collected on a

[1]These can range from minor violations such as failing to cooperate during a head count, to very serious violations such as assaulting a guard or another prisoner, selling drugs, or trying to escape.

standardized form, the CDC Form 839, "CDC Classification Score Sheet," commonly referred to as a CDC form 839 or simple "839." Much of the case factor information collected is thought to be related to the propensity for misconduct: sentence length, disciplinary history, work history, age, prior incarcerations and much more. The 839 assigns points to each of the background items. An inmates' total number of points constitutes a "classification score," which in turn, is used to help determine placements in one of four security levels. A higher score is supposed to reflects a greater proclivity to engage in misconduct or to attempt an escape, and therefore, the need for a higher level of security. For about 75% of inmates, placement in a security level is fully determined by the classification score. Inmates who score between 0-18 are placed in level I, 19 and 27 are placed in level II, 28-51 go to level III, and above 52 are placed in level IV.

Alternatively, about 25% of inmates are placed in a security level that is not necessarily consistent with classification scores. When the classification score is thought to not properly reflect the level of risk the inmate poses an "administrative placement" can follow. Approval to place an inmate in a security inconsistent with an inmate's classification score requires the "endorsement" of a department official (a Classification Staff Representative or CSR).

An admistrative placement considers both temporary and permanent case factors affecting inmate safety. An administrative placement is temporary when the administrative determinant is subject to time constraints, a potential change in case factors, or the receipt of additional information. For example, an inmate may be placed in a higher level of security pending the resolution of an active law enforcement felony hold likely to be exercised. Similarly, when an inmate's classification score falls within a security level that does not have available bed space, an inmate may receive a "population override" to an open bed in a security level above the level indicated by the classification score. This override is eliminated when beds at the original security level are available.

An administrative placement can also take special note of inmates who are convicted of predatory sex crimes, particularly violent crimes, or crimes for which the sentence is life without the possibility of parole (LWOPs). Such inmates are placed in at least level III facilities regardless of their classification score. Part of the rationale is the public relations difficulties that would follow should such an inmate escape. There is also the belief that the inmates serving LWOP are harder to control because they may feel they have little

to lose.

Finally, inmates are sometimes placed in one of two kinds of special facilities that do not formally correspond to a single security level. Inmates otherwise eligible for minimum security custody who have classification scores consistent with a Level I or Level II security level are eligible for placement at a CCF (Community Corrections Facility). Placement in a Security Housing Unit (SHU) is based on a departmental determination that the inmate's behavior endangers the safety of others or the security of the institution.[2] Placement in SHU is not based on the inmates classification score

2.1 Changes to the Existing System

Previous research (Berk and de Leeuw, 1998) coupled with less formal reviews internal to CDC led to a number of revisions of the existing inmate classification score instrument. The elimination of some items and the addition of others were suggested to better identify inmates with a proclivity for misconduct.

Several items were removed because they had no demonstrable association with misconduct in prison. The eliminated "stability factors" included an inmate's marital status, employment, education, and military service. Items indicating a successful escape were removed for the same reason coupled with the fact that successful escapes are very rare. Finally, whether or not an inmate had adjusted successfully to dormitory living in a past incarceration was removed. After years of severe crowding and the use of buildings not designed for housing inmates, it was no longer clear what inferences could be drawn.

Variables added because they were shown to be strongly related to misconduct included street gang or disruptive group activity, diagnosis of mental illness at a CDC reception center, age when first arrested, and prior incarceration. The earlier research and the day-to-day experiences of prison staff made clear that this meant young men with long arrest histories, gang activity, and/or a mental illness diagnosis.

Finally, modifications were made to the scoring of existing items. First, the weight given to length-of-sentence was reduced, because the association between misconduct and length of sentence was very weak, after accounting for other background items such as age. Second, because there was a

[2]From the California Code of Regulations, Title 15, section 3341.5(c)

strong association between age and misconduct, more weight was given to the younger ages (measured at arrival to a reception center) shown to be most problematic.

Project staff also recommended the implementation of "mandatory minimum scores." The mandatory minimum score integrates administrative determinants representing certain permanent case factors into the inmate classification scoring system. As such, they are a threshold score overriding the classification score otherwise calculated for an inmate. The goal is to make such places more objective. For example, regardless of the calculated classification score, an LWOP inmate will be given at least 52 points leading to level IV housing. The mandatory minimums were as follows.

- 52 points: inmates sentenced to death
- 52 points: LWOP inmates
- 28 points: Inmates serving multiple life terms or life with specific circumstances
- 19 points: Inmates with a history of escape
- 19 points: Inmates committed for specific sex offenses or sex related behavior
- 19 points: Inmates found to be violent felons per statutory requirements
- 19 points: Inmates determined to meet criteria as a high notoriety inmate
- 19 points: Inmate serving a life sentence

Once changes in the items were determined, an effort was made to design the forms implementing the changes that would be easier to use and would, ideally, produce more accurate information. A number of different formats were proposed, each carefully reviewed by CDC staff experienced in how such forms are used in the field. Several of the most promising forms were field tested by institutional staff. In the end, there was a broad consensus that in addition to the technical improvements in the instrument, the new forms were far more user friendly than the existing forms.

In short, the primary goals of the new classification score system were to better predict inmate misconduct and place them accordingly, In addition, the new scoring system was designed to be easier to administer and less prone to recording and arithmetic errors.

3 Past Studies

Inmate classification systems, serving a variety of purposes, have long been part of the penal scene in the United States (Brennan, 1987a). "Objective" classification systems, roughly like the one used in California, are a more recent development, but are now common across the country. A nearly universal question is how well objective classification systems work.

Much has been written on objective classification systems, including their development and evaluation, (Austin, 1986; Brennan, 1987a; Kane, 1986; Alexander and Austin, 1992; Hardyman, Austin, and Tulloch, 2000; Hardyman and Adams-Fuller, 2001). There are, however, few reports of experimental evaluations of these systems. Of the evaluations that have been done, most have not been experimental in nature and several were flawed because of small and biased samples (Alexander and Austin, 1992).

For example, in 1987, the Washington Department of Corrections initiated one of the better randomized experiments using 488 medium custody inmates to test the effectiveness of a Prison Management Classification (PMC) system. The goals of the new system were to improve safety and operations (Austin, Baird, and Neuenfeldt, 1993). The research results suggested that the new system worked reasonably well. Unfortunately, all experimental inmates were assigned to a new facility so it was not clear how much of the treatment was the classification system and how much the new housing.

Quasi-experiments are more common. Thus, two quasi-experimental studies of classification were completed in Tennessee (Baird, 1993). The first, in 1984, compared the behavior of inmates classified to different levels but all treated as minimum custody. A key finding was that many more inmates than originally thought could be classified as minimum custody without affecting public and prison safety. In 1991, a follow-up study was completed reviewing the behavior of inmates classified as minimum custody but, for lack of beds, placed as medium custody. By and large, the original conclusions were still valid.

There seem to be four conclusions from past research: 1) there are a

number of reasons a priori for favoring objective classification systems, 2) existing objective systems are broadly similar, 3) rigorous evaluations of the systems are highly unusual and 4) the weight of the research evidence suggests objective systems, while superior to less formal procedures, could certainly be improved. All four conclusions are consistent with the rationale for the work reported here.

4 Study Design and Rationale

4.1 Some Legal and Political Issues

Clearly, a study testing a new way to assign inmates to different security levels entails substantial risks for prisoners and prison staff. These were carefully weighed against the potential benefits for a classification system that was safer and more cost-effective. On balance, CDC administrators felt the potential benefits were substantially greater than the potential costs. Because statutes governing CDC's implementation of regulatory changes allow discretion in conducting "pilot studies" involving no more than 10% of the total inmate population, the State Regulatory Office approved a two-year pilot project to test the revised inmate classification score forms. Plans for the study were thoroughly reviewed by stakeholders including CDC administrators, representatives of prison employee bargaining unions, several other California State agencies, California State legislative offices, and a wide variety of other interested parties. There was widespread agreement that the study was worth doing.

4.2 Selection of Subjects

Power analyses were undertaken that were unusually well informed because of the previous research cited above. A key concern was to have a sufficient number of level IV inmates because analyses by security level were anticipated, and level IV inmates typically constitute only about 5% of the inmate population. Overall attrition had to be addressed as well because time served by level I and II inmates was commonly less than the length of the 2-year follow-up. Finally, it was necessary to anticipate how the random assignment would be implemented. In particular, the implementation would have to fit as snugly as possible within the existing administrative structure to minimize

disruptions, errors and work load. Thus, for example, it was not practical to recruit a special group of reception center staff to implement the experiment, in part because a means would have had to be found to send a subset of incoming inmates to those staff members without raising undue concerns. Moreover, any alteration in existing intake procedures risked affecting other reception center activities (e.g., medical exams). In the end, it was decided that we would simply include all new felony commitments for six months as our subject pool, which ruled out such options as oversampling the relatively rare level IV prisoners.[3]

These and other considerations led to a target sample size of 20,000 new felony commitments overall, with half to be placed under the experimental classification score system and half under the existing classification score system. All new felon commitments arriving at the CDC Reception Centers between November 1, 1998 and April 30, 1999 were included in the study for an actual sample size of 21,734. One important asset of this approach was that the target sample size was reached as soon as possible, which meant that the follow-up data collection could be ended at the shortest possible time. Another advantage was that for a well-defined period, the entire reception process could be put on special footing. This simplified implementation enormously.

4.3 Randomization and Placement of the Inmates

CDC ID numbers were used to divide the subjects into experimental (new classification score system) and control (existing classification score system) groups. Unique ID numbers are assigned sequentially at each reception center. Inmates receiving odd prison numbers were assigned to the experimental group and inmates receiving even numbers were assigned to the control group. All subjects were informed verbally and in writing that they were part of a study on the CDC classification system. Each was also received a copy of the classification score forms used to determine his or her classification score and were advised in writing that the assigned correctional counselor would

[3]There were many potential complications associated with oversampling. For example, one would have to compute a classification score first to determine who the level IV inmates were. And since computing that score was the major administrative burden in study implementation, oversampling would not actually save significant time or resources. We settled on collecting a sufficiently large sample overall to have the requisite number of level IV inmates.

be able to provide answers to most of their questions. Further, the inmates were advised that they could review the complete manual on the use of experimental classification score form. Project staff provided a copy of the manual for each facility law library to which inmates had access. During the project period, the CDC received no reported cases of inmates challenging participation in the study.

An intake classification score form, called CDC Form 839 (or "an 839"), was filled out for each inmate in the study. The control group version was used to record and tabulate intake information for the control group (color-coded yellow), from which placements were determined. The experimental group version played the same role for the experimental group (color-coded orange). However, both forms were filled out for all subjects, even through only one form would guide placement. The rationale was to permit answers to counterfactual "what if" questions, such as how a particular type of experimental inmate would have been placed under the existing classification score system. We will exploit such counterfactual information below.

4.4 Data Collection

Inmate intake forms were key-entered so that a machine readable file was produced. Also key-entered were CDC "reclassification" forms (CDC Form 840, called an "840"). About a year after reception, the performance of each inmate is reviewed, and an 840 filled out. Because the CDC requires that inmates be evaluated at least annually, the classification score, housing assignment, and performance of each inmate are reviewed by a classification committee to update the 840. Recorded are both favorable and unfavorable credits and points assessed as a result of disciplinary violations during the preceding period.

The annual review is designed to evaluate the inmate's behavior, update the classification score, and consider any need for a change in placement. An inmate who is free of any disciplinary actions and demonstrates positive participation in an inmate program during the period reviewed, earns points that are deducted from the classification score. Conversely, an inmate who has been found guilty of one or more disciplinary violations during the period of review has points added.

Point reductions often result in a transfer to a lower security level when the inmate's score falls within a range associated with such a level. A score increase can have the opposite impact. If an inmate completes his or her

sentence and is released to parole before the annual reclassification review, the inmate may have no 840.[4] Thus, an inmate may have no 840, one 840, or several.

The 840s used for the experimental subjects were largely the same as the 840 used for the control subjects. The main difference was to increase a bit the weight given to inmate behavior since the last review so that good conduct could be better rewarded. For both the experimental and control inmates, the 840s permit one to determine how an inmates classification score changes over time.

The 840s document the endorsement of an inmate to a different facility or to a different security level within an institutional complex. With the use of the CDC Movement History File, which essentially records an inmate's placements over time, and the endorsed location documented on the 840, one can determine how and when a change in the classification score translates into a change in housing.

To further supplement the data on the 839 and 840, data were extracted and key-entered from the CDC Form 115, "Rules Violation Report" (called a "115"). A 115 is completed when staff observe an inmate engaged in some form of prison misconduct. Inmate disciplinary violations range from minor violation such as failing to report to an assignment to serious violations such as battery on a correctional officer or on another inmate, trafficking in drugs, or attempting an escape.

For the experiment, project staff audited every 115 received by any of the 21,734 inmates during the 24 months they were part of the study. Project staff compiled and entered data to record the date of the misconduct, a description of the specific act, the inmate's housing location and security level at the time of the incident, the determination of whether the violation was serious or only "administrative," the "division level" if the 115 was serious, the mental health status of the inmate at the time of the violation, and whether the violation was drug-related and/or alcohol-related.[5]

[4]Inmates returned to complete their sentences continued to be included in the study. Their classification scores were updated on an 840. Inmates returned with a new conviction and sentence were not included in the study because they were, in effect, a new admission after the study intake period.

[5]Serious offenses recorded on the 115 are placed in one of several broad categories, such as "narcotics trafficking." These are called "division levels."

5 Study Implementation

Pages could be written about how the study was implemented. For this paper, three points can be made. First, project personnel worked closely with CDC staff to provide thorough training in the new instrument. In addition, a special manual was provide to all and a hotline established where pressing questions could be quickly answered.

Second, the randomization process was regularly monitored by on-site observation and statistical analyses of preliminary data. No important problems were uncovered during the six months of intake.

Third, all intake forms were thoroughly reviewed as part of routine CDC procedure with supplemental reviews by project staff. Project staff were stationed at the reception facilities during the six months of intake for a minimum of three days a week. As time permitted, cases completed by reception center personnel were given a complete review by project staff. After all forms were later key-entered, the project staff audited nearly 100% of all 839s and 840s. Finally, all intake forms were subject to a number of logical checks with specially written computer algorithms. If errors were found through any of these processes, the case file was reviewed and corrections made. In addition, records were kept of the corrections made to determine if there were systematic errors in the data collection process. No such errors were found.

6 Findings

6.1 Data on Randomization

If the experiment was implemented as designed, there should be equal numbers of inmates in the two groups, experimental and control, and the composition of those groups will be effectively the same. Our data indicate that this was the case. There are 10,877 inmates in the experimental group and 10,857 inmates in the control group. The expected 50-50 split is approximated extremely well. Composition of the two groups was also very similar. For example, there were 1,861 gang members among the experimental group and 1,839 gang members among the control group. That is a split of 50.3% versus 49.7%. Another way to show the split is that 17.1% of the experimental group and 16.9% of the control group are gang members. Consider

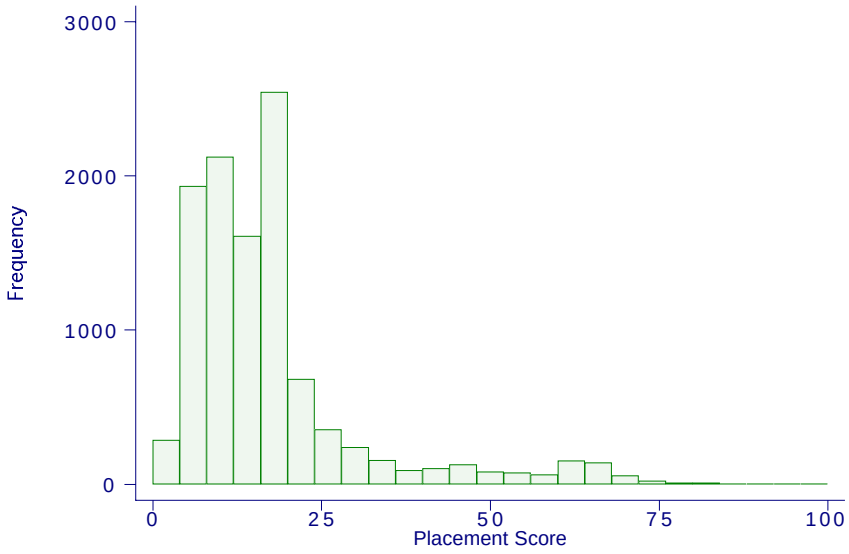

Figure 1: Placement Score Distributions for Controls

the number of inmates under 21 years of age. There are 1225 such inmates in the experimental group versus 1202 in the control group. That is a 50.5% to 49.5% split. 11.2% of the experimentals and 11.1% of the controls were under the age of 21. Regardless of the background variable chosen, the two groups were balanced. The experimental and control groups were effectively identical.

6.2 Treatment Effects on the Size and Mix of Inmate Populations

Figures 1 and 2 show the score distributions for the control and experimental groups respectively. Both distributions are skewed to the right with the mass of data below a score of 25, as would be expected. Most inmates are incarcerated for a relatively short period of time and are not high risk. Low classification scores follow. The two score distributions are much the same except that the experimental scores are shifted a bit to the right.

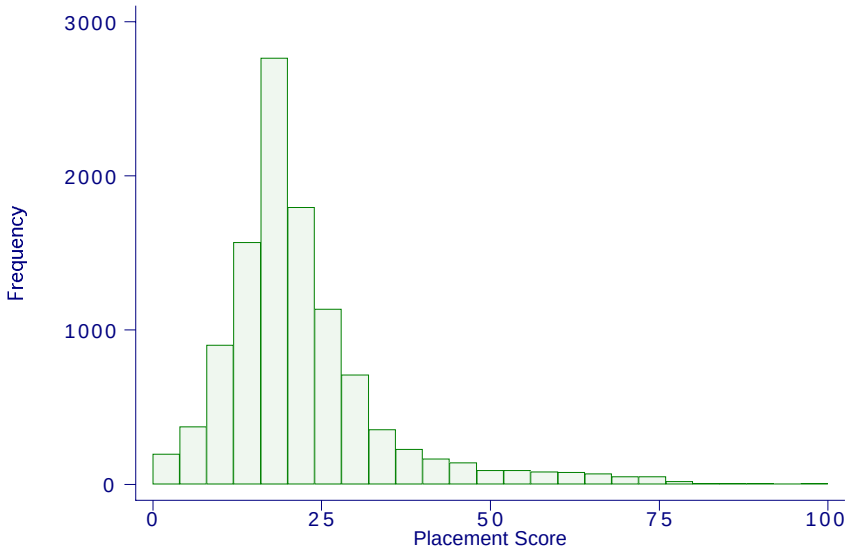

Figure 2: Placement Score Distribution for Experimentals

If one compares scores by initial endorsed placement level, it is easily seen that the shift in scores for the experimental inmates is primarily in the lower levels. For example, within level I, the median score increases from 11 to 15 and the mean from 11 to 15. Level II shows a slight difference favoring the experimental group, but levels III and IV have similar central tendencies for the experimental and control groups. Note that these are aggregate results and do not imply that the new forms simply increased lower scores a few points on the average. We will see shortly that a lot more is going on.

6.2.1 Items Most Affecting the Classification Score Distributions

Which items in the classification score were driving these score totals? Note that the importance of an item in practice is a function of the weight given to that item in the scoring system and the amount of empirical variation in that item among incoming prisoners.

We computed the average percentage that each item contributed to the total score. The analysis was done separately in each security level in part because average overall scores that served as the base vary substantially by level.

For the control inmates, the story is simple. Points awarded for longer

sentences dominate the classification score. It accounts for about 30% of the score in level I and about 80% in level IV. Even for level I inmates, term length is more than twice as important as any other other item. By design, the new system was intended to make term length far less important, and increase the impact of other items; term length was in past studies not found to be nearly as useful an indicator of misconduct as its weight suggested. What actually happened?

For the experimental inmates, in level I, more than half the total score on the average was determined by the inmate's age at first arrest and at reception. Another 25% was explained by points awarded for longer sentences. Having no serious misconduct charges during prior incarcerations was responsible for 10% of the total. None of the other score variables individually contributed more than 5%. The story in level II was much the same.

The overall pattern in level III was similar to that of levels II and I except that on the average 34% of the total score was explained by sentence and another 31% of the variation was explained by the age of the inmate at first arrest. Having no serious misconduct charges during prior incarcerations counted for 5% of the total. Points given for being a gang member starts to count in level III, explaining 8% of the total score.

In level IV, 63% of the total score on the average is determined by sentence length. This is reasonable since inmates assigned to level IV generally commit the worst crimes thereby receiving the longest sentences.

In short, the score distributions are determined by relatively few classification items. Term length dominated the old classification system. Under the new system, the impact of points for longer sentences are far less important. New items included to better predict misconduct take up the slack.

6.2.2 Impact on the Placement of Inmates

One of the key issues raised by the experiment was the potential impact of the new classification form on initial placement. Table 1 shows the actual placements for the experimental and control groups.[6] Note that we have included for now Reception Center (RC) placements, which represent the few inmates who were not placed in a regular CDC prison.[7] We have also

[6]Technically, the "initial placement" is actually an inmate's "endorsed location."

[7]The majority of inmates paroling from RC, who are not placed in a regular housing unit, are inmates who complete their sentences while in RC and are therefore, released from RC to parole. A few have pending court obligations and are ordered back to a county

at this point included Community Corrections Facilities (CCF), which are level I placements, and Secure Housing Unit (SHU) placements, which, as we mentioned earlier, are formally outside of the classification system. Later analyses will focus on the four security levels because placement in those levels is what the experiment was meant to address, and it is those levels that affect the vast majority of inmates.

For the experimental group, there is a significant decline in the relative size of the level I population from about 37% to 29% and a significant increase in the relative size of the level III population from about 12% to 20%. Given the large sample sizes, such disparities are easily large enough to reject the null hypothesis of no difference. How might such a shift come about?

Recall that all inmates were scored under both systems. Thus, for every placement there is considerable information on the hypothetical placement that could have been made but was not. That is, for all experimental inmates, we have information on how they would have been placed under the existing classification system, and for all control inmates we have information on how they would have been placed under the revised classification system.

Table 2 is constructed by comparing the actual placement of each inmate to the hypothetical placement (i.e., without population overrides) under the classification system that was not applied to them. The table shows that although there is some displacement of inmates, overall the majority of inmates would receive the same initial placement under either system. The major exception for the experimental group is in level III where only 52% of the inmates would have been placed the same under the original system.

If one goes a step farther and tabulates for the experimental group the actual placement against what the placement would have been under the existing classification system (table not shown), it is readily apparent that while some inmates are placed very differently under the new system (e.g., in level IV instead of level I), a majority who are placed differently shift up or down one level. Thus, for example, of the level III inmates who would have been placed differently, 55% would have been placed in level II, and 27% would have been placed in level I.

From Table 2, it is level I for the control group where the major changes occur; only 71% would have been placed in the same level under both systems. Tabulating for the controls the actual placement against the placement that would have occured under the new system (table not shown) again shows that

jail for another offense.

the majority of those inmates who would have been placed differently, would have changed only one level. Thus, for example, of those level I inmates who would have been placed differently, 74% would have been placed in level II, and 26% would have been placed in level III. Clearly, there is an important shift upward overall under the new system from level I to levels II and III. As before, given the very large sample sizes, all such percentage comparisons easily lead to a rejection of the null hypothesis of no difference.

Initial Placement	Controls	Experimentals	Total
RC	2.26%	2.57%	2.42%(525)
CCF	14.89%	12.93%	13.91%(3023)
Level I	37.35%	29.05%	33.20%(7215)
Level II	28.81%	31.33%	30.07%(6536)
Level III	11.67%	19.79%	15.74%(3420)
Level IV	4.77%	4.06%	4.42%(960)
SHU	0.25%	0.26%	0.25%(55)
Total	100%(10857)	100%(10877)	100%(21734)

Table 1: Initial Placements for the Experimentals and Controls Separately (Full Sample)

Initial Placement	Experimentals	Controls
RC	100.00%	100.00%
CCF	99.93%	94.56%
Level I	96.93%	70.60%
Level II	79.84%	83.09%
Level III	52.11%	84.93%
Level IV	88.91%	80.69%
SHU	100.00%	100.00%

Table 2: Percentage of Experimental and Control Inmates for whom the Actual Initial Placement was the same as the Hypothetical Initial Placement

Shifts of the sort just described can have important implications for crowding insofar as the new classification system allocates inmates initially in a manner inconsistent with available beds. Equally important is how the new

distribution of inmates affects which kind of inmates are sent to which kinds of facilities. For example, is the new system really placing "gang-bangers" in more secure settings, as intended?

We focus here on levels I and III because that is where differences in placements between the new and existing system are most pronounced. For level III placements, 25% of the control group were linked to gang activity compared to 44% of the experimental group; 38% of the control group were under 27 compared to 58% of the experimental group; 16% of the control group were under the age of 21 compared to 29% of the experimental group. Finally, 18% of the control group had a history of mental illness compared to 14% of the experimental group. All but the last of these patterns are consistent with the intent of the new classification system. For inmates with a history of mental illness, it is likely that there are other common features of such individuals that were both unanticipated and mitigated the impact of the mental illness designation.

For level I placements, 14% of the control group were determined to be involved in street gang activity compared with 5% of the experimental group. In addition, 29% of the control group were under the age of 27 compared to 16% of the experimental group; 10% of the control group were under the age of 21 compared with 4% of the experimental group. Finally, 3.6% of the control group had mental illness status compared with 2.4% of the experimental group. The pattern for level I inmate is less dramatic than for level III inmates, primarily because it was relatively rare to find in level I institutions many inmates with gang activity or a history of mental illness. However, the impact on age is apparent: younger inmates are generally shifted upward under the experimental classification score system.

6.2.3 Impact on the Mix of Inmates

Given the impacts of the new classification system on initial placement, what might the longer term implications be? Recall that misconduct recorded in Form 115 can lead to movement to higher security levels while good conduct recognized on Form 840 can lead to movement into lower security levels. Using 24 months of data from the Movement Files, one can see how the populations in the different levels will change over time. Figure 3 contains six graphs showing the number of experimental and control inmates in each placement location for the length of the follow-up period. There is one graph each for the Reception Center (RC), Community Corrections Facilities (CCF), and

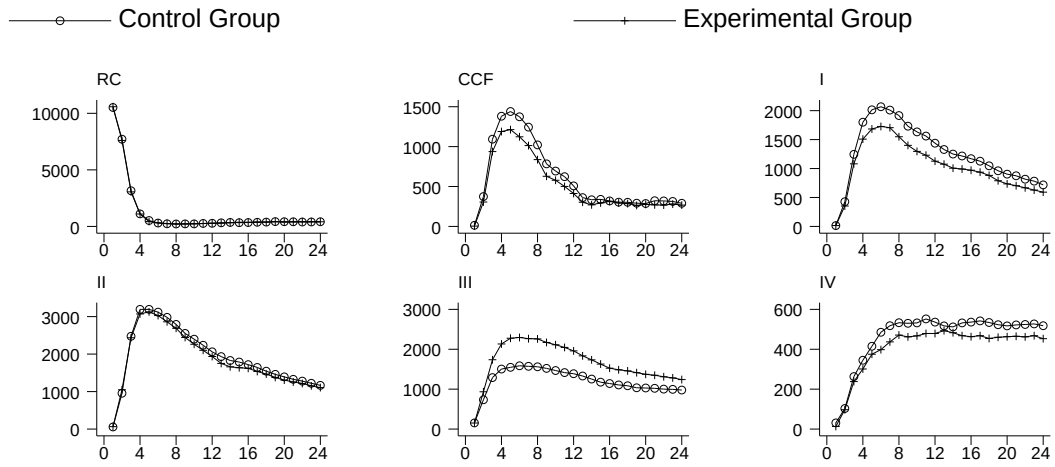

Figure 3: Number of Inmates in each Location by Months after Admission

each of the four security levels I through IV.

1. The Reception Center population drops to zero for both the experimental and control groups by month six as would be expected because the Reception Center is the holding place for the inmates until assignment to a bed in one of the institutions. Both groups show the same distribution over the 24 months.

2. The population in the CCF increases till about month 6 at which time the population starts to decrease. This is, of course, a necessary consequence of the 6 month intake at the beginning of the study. There are a maximum of about 200 more inmates in the control group than the experimental group, but the control group exits the CCF more

quickly. Thus, by the end of the study, the numbers of inmates are approximately equal for the two groups.

3. For level I, we again see an increase in the number of inmates till month 6 for both groups of inmates. There are approximately 200 more control inmates in Level I than experimental at that point. The gap between the two groups closes over time so, by month 24, the numbers are practically the same.

4. Level II shows the increase to month 6 followed by a decrease to month 24. However, the distributions for the two groups are basically the same.

5. Level III shows an increase in population till month 6 with the experimental group having about 500 more inmates than the control. The gap between the two groups closes as the distributions decrease. It appears that if the trend were to continue, the numbers of control and experimental inmates should be approximately equal by month 30.

6. Level IV shows an increase in population till month 6 and then tapers off but does not decrease as might be expected. There are about 100 more controls than experimentals, a gap that is roughly constant over the length of the study.

One message from Figure 3 is that differences in placement patterns for the experimental and control inmates generally decrease over time. This would make sense if the net percentage of inmates leaving a given level were about the same for both groups. The group with the greater initial number would shrink faster toward zero.

Another message is that during the follow-up period, a substantial fraction of the inmates in the study were released from prison because their terms expired. Indeed, about a quarter of the experimental inmates and about a quarter of the control inmates were released on parole within a year after they arrived at a CDC reception center. By the end of the second year, those figures were approximately 80% each. This "attrition" was fully expected based on CDC's own studies, and was one of the reasons why such a large sample of inmates was required. But movement of inmates out of prison needs to be consided part of the explanation for the declining curves in addition to

movement to other security levels.[8] The major exception is level IV inmates, most of whom remain in level IV for the entire 24 months of the study. This too is really no surprise because level IV placements depend substantially on term length, and inmates with long terms to serve are assigned a large number of points. It would take the typical level IV inmates several years of good behavior to significantly reduce their point totals[9]

6.3 Misconduct

6.3.1 Comparing Misconduct for the Experimental and Control Inmates

One of the central issues of the study in-custody misconduct: would the new forms better sort inmates into different categories of risk and then after placement reduce, or at least not increase, the amount of misconduct. We looked at misconduct in a number of ways. To begin, using whether or not an inmate had a CDC Form 115 during the 2-year study, we compared the experiences of inmates placed by the new and existing procedures. For both the experimental and control groups, 34% of the inmates engaged in misconduct. Clearly, there were no overall differences.

There are two broad types of 115s: "administrative" and "serious." Administrative 115s range from not reporting on time for a class or a job to making threats. Serious 115s range from possession of a deadly weapon to manslaughter or murder. Administrative 115s were recorded for 18.7% of the control inmates and 18.4% of the experimental inmates. Serious 115s were recorded for 25.8% of both the experimental and control inmates. Clearly, there are once again no important differences.

Although the majority of inmates did not engage in misconduct during the 18 months we have observed, some received more than one 115. When using the total number of 115's committed as the outcome, the conclusions are the same: the control and experimental inmates each have about half of the total number of 115s and of the administrative and serious 115s as well.

Given the substantial changes in how the inmates were placed under the new forms, the lack of any differences in misconduct between the experi-

[8]Note that the attrition did not differ for the experimental and control groups. Thus, the attrition does not affect the study of treatment effects.

[9]The very large point total also are usually well above the threshold between a level III placement and a level IV placement.

mental and control inmates is somewhat perplexing. One possible inference is that placements do not affect misconduct very much. If true, random assignment virtually guarantees that the amount of misconduct for the experimental inmates will be the same as the amount of misconduct for the control inmates. Another possible inference is effectively the opposite; the placement environments in the different security levels work so well, and adjusted so quickly to the changes in the mix of inmates, that the overall level of social control was maintained. Stated so baldly, both inferences seem implausible. In an effort to better understand, we examined misconduct for the experimentals and controls as a function of initial placement. Emphasis now shifts to the role of security level since that is the primary concern of the CDC inmate classification system. Table 3 show the results.

Initial Placement	Experimentals	Controls
Level I	26%	31%
Level II	30%	33%
Level III	54%	50%
Level IV	54%	54%

Table 3: Percentage of Experimental and Control Inmates Engaging in Misconduct by Initial Placement

Tables 3 reveals in level IV similar percentages of misconduct for the control and experimental groups. The levels where the control and experimental groups differ a bit are I, II and III. In levels I and II, the control group engages in more misconduct. In level III, the experimental group engages in more misconduct. These findings are not surprising given that one of the goals of the new system was to place inmates at high risk for in-custody misconduct into more secure settings. Thus, while the overall levels of misconduct are the same for the inmates placed under the new and existing classification system, the new system shifted the misconduct into the higher security level.[10]

[10]Note that since the vast majority of inmates are in level I or II housing, the misconduct percentages in those levels largely determine of the overall misconduct rate.

6.3.2 Including the Role of Classification Score and Placement

But, there is much more to the story. Misconduct is a function of an inmate's proclivity to get into trouble and the nature of the setting in which he is placed.[11] One needs to try to separate these two distinct effects. Indeed, a failure to do so has been a major flaw in much past research.

Using a generalized regression discontinuity design (Berk and de Leeuw, 1998), one can consider misconduct as a function of initial placement and classification score. In this instance, the design leads naturally to a logistic regression with placement and classification score as predictors. Note that since classification score is the vehicle by which placement is undertaken, estimates of the effect of placement are in principle unbiased without including any other covariates. This may seem counter-intuitive, but the underlying rationale as been accepted for well over a generation (Campbell and Stanley, 1963). A formal proof and further details can be found in Berk and de Leeuw (1998).

Tables 4 and 5 show the results of logistic regressions using the existance of a CDC Form 115 as the response variable and initial placement and classification score as the explanatory variables. Table 4 reports the results for the experimental group and Table 5 reports the results for the control group. For both regressions, level I serves as the reference category for security level. All administrative placements were excluded from the logistic regressions so that classification score in fact determined fully an inmate's initial placement, as the generalized regression discontinuity design requires. However, the results are much the same if administrative placements are included.

Predictor	Coefficient	Std. Error	Multiplier
Score	.086	.005	1.09
Level II	-0.17	.071	0.84
Level III	-0.21	.114	0.81
Level IV	-2.97	.264	0.05
Constant	-2.40	.093	–

Table 4: Misconduct Logistic Regression for Experimental Inmates – Administrative Placements Excluded (N=13453)

[11] Recall that while female inmates are given classification scores just like the men, there are no security levels in facilities for women.

Predictor	Coefficient	Std. Error	Multiplier
Score	.059	.005	1.06
Level II	.064	.093	1.07
Level III	-0.66	.159	0.52
Level IV	-2.12	.285	0.12
Constant	-1.49	.070	–

Table 5: Misconduct Logistic Regression for Control Inmates – Administrative Placements Excluded (N=13453)

From Table 4, we see for the experimental inmates an odds multiplier for classification score of 1.09. This means that for each additional point received, an inmate's odds of misconduct are multiplied by a factor of 1.09. This may seem like a small effect, but from one level to the next, inmates' scores can vary by 20 points or more. Consider two inmates who differ in score by 20 points. For the inmate with the higher score, the odds of misconduct are 5.6 times greater (i.e., $1.09^{20} = 5.604$) than for the inmate with the lower score.

Table 5 contains the parallel analysis for the control inmates. One can see that classification score is less effective in sorting inmates by the risk of misconduct; the odds multiplier of 1.06. This difference (1.09 versus 1.06) may seem small, but it is significant when comparing inmates with substantially differences in score. For the experimental group, 20 additional points translated into a risk of misconduct that was 5.6 times larger. For the control inmates, the 20 additional points translates into risk that is only 3.2 times greater. Clearly, the new classification system makes greater distinctions between inmates with respect to the risk of misconduct.

Analyses of misconduct versus classification score were also undertaken for each level individually. A "matching" analysis of this sort (rather than relying on covariance adjustments) is a more robust analysis, made possible here by the large samples. In each case, the experimental score performed better at sorting inmates by their level of risk.

The experiment was not designed to study the impact of placement on misconduct and therefore, any such analysis must be interpreted with caution; there was no random assignment to security level. Still the apparent impact of security level broadly makes sense.

One can see from Table 4, that level IV compared to level I has a sub-

stantial impact on the odds of misconduct. An initial placement in level IV rather than level I, reduces the odds of misconduct by a factor of .05. This is a large reduction, roughly equivalent to the increase in the risk of misconduct associated 35 addition classification points (i.e., $.05^{-1} = 20 \approx 1.09^{35}$. However, consistent with the findings in Table 3, such reductions are not large enough to compensate for the increases in risk compared to level I. The classification scores of level IV inmates are generally more than 35 points greater than the classification scores of level I inmates. Hence, it is not surprising to find the highest rate of misconduct in level IV facilities.

In Table 4 there is also a hint of "suppressor effects" in levels II and III when compared to level I. The coefficients are roughly twice the standard errors and the multipliers large enough to be of some practical interest. Still, CDC officials are quick to point out that while there is somewhat greater control over inmates in levels II and III compared to level I, level IV housing is substantially more restrictive than the lower levels. In short, under the revised classification system, the suppressor effects are consistent with the way CDC currently allocates its social control resources.

For the control inmates, Table 5 shows that under the existing classification system there is also a strong suppressor effect in level IV compared to level I. However, the results for the other two levels are somewhat different. The suppressor effect for Level III is substantial while the suppressor effect for level II is no longer apparent.

These differences between the experimental and control inmates could result rather easily for the random error introduced by random assignment. Still, there may be a plausible explanation. Inmates placed under either the revised or existing classification system were housed within the same prison system. Both groups experienced the same prison environment. However, the revised system altered somewhat which kinds of inmates were sent to which security levels. In particular, the revised system shifted a large number of inmates more likely to be difficult from level I and level II upward. Thus, the fit was different and one might expect to find that the suppressor effects in levels II and III declined relative to level I . And that is what one sees when Table 5 is compared to Table 4.

It is important to stress that we have examined the impact of initial placement only. About half the inmates in the study remained in their initial placements until the 12 month evaluation, and about a third remained in their initial placements during a 2-year follow-up. Ideally, one might like to explore the impact of each placement. Unfortunately, this is extremely

difficult to do because one would have determine the time spent in each and then allow for different placement sequences. For example, 6 months in a level III faculity followed by 6 months in a level IV facility implies something very different from 6 months in a level IV facility followed by 6 months in a level III facility. Clearly, there would be a large number of possible sequences of placements and even with our large sample size, statistical power would be very low. But once again, the goal of the experiment was not to examine in great detail the effect of placements on misconduct. Whatever we are able to learn about such processes is a bonus.

6.3.3 Unpacking the Items in the Classification Score

We were also interested in knowing which items used to construct the classification score are most strongly associated with future misconduct. The logistic regressions in Tables 4 and 5 were rerun with the classification score items substituted for overall classification score. The six items that seem to be most strongly associated with any misconduct were age of first arrest, age at reception, mental illness, prior jail sentences, a prior sentence with the California Youth Authority, and a prior CDC sentence. Gang activity almost made the cut, but was highly correlated with the variables corresponding to age and added little new information. When serious misconduct is used as the response variable, gang activity becomes an important predictor even with the age variables included. In contrast to the implications of old system, points computed from the nomial sentence length were found to be unimportant in predicting misconduct.

In short, inmates who are engaged in gang activity, young, and who have long histories of contact with the criminal justice system tended to get in the most trouble. Mental illness also counts. Once these variables are factored in, other items such as offense type and sentence length do not contribute much. We also find no evidence that earlier good behavior predicts less misconduct. Points for successful completion of a prior minimum custody incarceration and having no serious disciplinaries in the last twelve months were not associated with less misconduct. One implications is that in the future it might be possible to further simplify how the classification score is calculated.

6.4 Reclassification

Finally, we turn to what might be learned from CDC form 840. Recall that such forms are filled out as a routine matter approximately every 12 months while an inmate is incarcerated. They are also filled out if there is any reason, such as inmate misconduct, to compute a new classification score.

The revised classification system did not make important changes in the 840 form. The main alteration was to allow classification points to be deducted a bit more rapidly when there was no reported misconduct. In fact, this is what the data show. A regression of each inmate's 840 score on his 839 score showed that for both the experimentals and controls, the regression coefficient was effectively 1.0. This is not surprising because few classification scores change dramatically between the two assessments. For the control inmates, the intercept was about -2.0. The classification score had declined on the average by about 2 points. For the experimental inmates, the intercept was a little smaller than -4.0. The classification score had declined on the average by about 4 points. With the average classification score in our data of less than 20, during the first year or so the classification score for the control inmates dropped by about 10% while the classification score for the experimental inmates dropped by about 20%. This is just about what the new forms were designed to accomplish.

7 Conclusions and Policy Implications.

It would seem that there are a number of conclusions whose policy implications are relatively straightforward.

1. The experiment was well executed. Indeed, it was in many ways a textbook example of a very large field experiment. Working in a total institution like a prison surely helps, but the CDC also invested considerable resources in the project. There was also the key advantage of several preliminary studies. The moral is clear: large scale randomized experiments can be conducted well in prison settings when there is the commitment to do so. And because randomized experiments are widely understood to be the "gold standard" in program evaluation when causal inference is central, randomized experiments should always be seriously considered when the effectiveness of prison programs is of interest.

2. The revised inmate classification forms were well received by prison staff. Anecdotal evidence indicates that the new forms were more user friendly than the old forms. Quality control oversight indicated that the new forms were also less prone to recording and computational errors. Finally, the new forms were also preferred by the CDC staff bargaining organizations. In short, staff "buy-in" did not seem to be a problem.

3. Given the prison setting, it is difficult to know what conclusions to draw from the absence of any challenge to the revised forms or the experiment from inmates or inmate advocacy groups. Perhaps the earlier meetings with stakeholders helped. Or perhaps, the changes in the classification forms were too small to seem important.

4. Converting placement "overrides" under the existing system to "mandatory minimums" under the revised system proved to be a simple and effective means to make explicit decisions that previously had been difficult to track. Mandatory minimums made the new system more "transparent."

5. Under the revised system, inmates who engaged in gang activities, who were young, and who had had long histories of contact with the criminal justice system were anticipated to be among the most likely to get into trouble. Mental health problems could also be an important factor. In fact, the data from the experimental group supported the use of these indicators. Gang activity and mental illness were not considered in the old classification system, and age was not weighted as heavily.

6. A number of classification indicators popular with corrections officials and prison researchers a generation ago were discarded: marital status, education, service in the military and employment history. In fact, they were virtually unrelated to prison misconduct among the control group. Perhaps when rehabilitation was a more significant part of the prison agenda, these indicators were more relevant.

7. The majority of inmates had the same endorsed placements under the existing and revised systems. For those with different placements, the placement shifts were typically one level. Thus, the new classification system was by design, and in fact, a refinement of the existing procedures, not a wholesale reformulation. Given that earlier research had

shown the existing classification procedures to be functioning reasonably well, both the plan and the outcome made sense. This argues more generally for the usefulness of preliminary studies.

8. Overall, there was under the revised system a net decline in the population initially assigned to level I facilities and a net increase in the population initially assigned to level III facilities. The net decline in level I and increase in level III dissipated over time so the the net changes in population distributions did not accumulate. Neither result was surprising once the revised instrument was designed. Still, there may well be a need to reconfigure some facilities in response to these modest population shifts. And therein lies an important lesson: with housing pure and simple such a critical component of any prison system, it is very difficult to tinker with any feature of prison life and not affect the fit between the number of beds and inmate needs.

9. The revised classification score sorted inmates substantially better by level of risk. Thus, there was clear evidence that one can improve on an inmate classification system that was already well respected by prison administrators across the country.

10. Under both the existing and revised classification systems there was strong evidence for "suppressor effects" in level IV compared to all other endorsed placements. The architectural design of prisons coupled with staffing and administrative procedures really matter for the safety of inmates and staff and indeed more generally, if order is to be maintained.

11. Under the revised classification system, inmate scores declined a bit more quickly over time. Consequently, downward movement to lower levels of security will occur more rapidly.

The California Department of Corrections is currently making plans to implement the revised classification system. Drafts of new administrative procedures have been written, materials for retraining intake staff have been designed, and "sign-off" has been achieved throughout the Department. There have also been meetings with stakeholders explaining the changes underway and the research supporting these changes. It is likely that the revised system will become the operational system by early 2003.

8 References

Alexander, Jack and James Austin. *Handbook for evaluating prison classifications systems.* San Francisco: National Council on Crime and Delinquency.

Austin, James. 1986. Evaluating how well your classification system is operating: A practical approach. In *Crime & Delinquency* 32, No. 3, ed. Lawrence A. Bennett. Newbury Park, Calif.: Sage Publications.

Austin, James, Christopher Baird, and Deborah Neuenfeldt. 1993. Classification for internal management purposes: The Washington experience. In *Classification: A tool for managing today's offenders.* American Correctional Association.

Baird, Christopher. 1993. Objective classification in Tennessee: Management, effectiveness, and planning issues. *Classification: A Tool for Managing Today's Offenders.* American Correctional Association.

Brennan, Timothy. 1987a Classification: An overview of selected methodological issues. In *Prediction and clasification: Criminal justice decision making.* Chicago: University of Chicago Press.

Brennan, Timothy. 1993. Risk Assessment: An evaluation of statistical classification methods. In *Classification: A tool for Managing Today's Offenders.* American Correctional Association.

Campbell, D.T., and J.C. Stanley 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Hardyman, Patricia L., James Austin, and Owan C. Tulloch. 2000. *Revalidating External Classification Systems: The Experience of Seven States and Model for Classification Reform.* Report submitted to the National Institute of Corrections. Washington, D.C.: The Institute on Crime, Justice and Corrections at The George Washington University.

Hardyman, Patricia L., and Terri Adams-Fuller. 2001. National Institute of Corrections Prison Classification Peer Training and Strategy Session: What's happening with prison classification systems? September 6-7, 2000 Proceedings.

Kane, Thomas R. 1986. The validity of prison classification: An introduction to practical considerations and research issues. In *Crime & delinquency* 32, No. 3, ed. Lawrence A. Bennette. Newbury Park, Calif.: Sage Publications.