Pretrial Services Agency for the District of Columbia Risk Assessment Instrument Re-validation Project

Final Report

Contract No. PSA17C0043

Avinash Bhati, PhD
Maxarth LLC
12215 Fellowship Lane,
North Potomac, MD 20878

May 5, 2019

Acknowledgments

This work was conducted under contract PSA17C0043 from the Pretrial Services Agency for the District of Columbia. The effort would not have been possible without the help of PSA staff and executives. In particular, I would like to thank Michael Kainu, Kenneth Chen, Demond Tigs, Hala Maktabi, Sharon Banks, Michael Williams, Cathy Terry-Crusor, and Leslie Cooper for the numerous hours they spent helping craft this analysis as well as for providing valuable feedback. However, they are not responsible for any analysis conducted here. All errors and omissions are the author's alone.

Contents

List of Tables

List of Figures

Chapter 1

Project Overview

The Pretrial Services Agency for the District of Columbia (PSA)—an independent entity within the Court Services and Offender Supervisions Agency for the District of Columbia (CSOSA)—developed, validated, and deployed a risk assessment instrument (RAI) in 2012. That instrument has improved PSA's ability to identify defendants appropriate for release and those more suited for detention. On June 24, 2014, staff completed the 10,000th risk assessment using this instrument. In keeping with standard practice, PSA is in the process of re-validating the RAI and improving its functionality. In addition, PSA is interested in gauging the need to add additional features to the RAI, utilizing its data to formulate risk-based supervision strategies, and assess the extent of any predictive bias in its RAI. Maxarth LLC was awarded a contract (PSA17C0043) in September 2017 to assist PSA in addressing these needs.

1.1 Project Goals

The main GOAL of the *Risk Assessment Instrument Re-validation Project* was to provide PSA consulting services to address three interrelated tasks:

1. Re-validation of existing risk assessment instrument;
2. An evaluation of PSA's planned risk-based recommendation matrix; and
3. An investigation of predictive bias in PSA's risk assessment instrument.

1.2 Summary of Analysis

Detailed, micro-level, data were obtained from PSA over the course of the project. These included data on features currently used by the RAI as well as the underlying raw data in order to develop additional (refined) features for inclusion in the model (if needed). Data were analyzed for accuracy and consistency and the current risk assessment tool was assessed for continued validity. In addition, a series of analyses were conducted to study the extent of racial bias in the RAI predictions.

Detailed sequential data was also obtained from PSA on defendant conduct while under supervision and PSA's response to defendant conduct. These infraction and agency response data were used to develop a series of models that permit simulating the short-, medium-, and long-term implications of adopting different risk-based supervision strategies. The models were then used to assess the implications of PSA's proposed risk-based supervision plans. The models were also used to develop a series of detailed tables that can be used by PSA to further fine tune its plans.

The analysis conducted over the course of the last 12 months was shared with PSA research and operations staff as well as executives. Regular meetings resulted in shaping the analysis to be maximally consistent with PSA's needs. Detailed data tables and related materials were developed for PSA staff to consult while considering alternative risk-based supervision strategies.

1.3 Summary of Findings

Based on the analysis conducted, the main findings and recommendations from the project can be summarized as follows:

1. Re-validation Effort (Task 1):
   - (a) The current instrument performs well but can be improved upon.
   - (b) Many of the currently included items are based on charge level details about a defendant's criminal history. Data analysis suggests that these items may be problematic. Revised models were developed without these items and they are sufficiently predictive of pretrial misconduct.
   - (c) A revised 43-item RAI was found to predict pretrial misconduct with sufficient accuracy.

2. Risk-based Supervision Recommendation Matrix (Task 2):

   (a) Markov models of the behavior of defendants (infraction) and PSA's response to defendant conduct were developed. These models permit an analysis of the effects of different agency policies on the rate of successful case conclusion (without an FTA or Re-arrest).

   (b) The models suggest interesting patterns in the data that can be exploited by PSA when formulating their risk-based supervision strategy.

   (c) Specifically, the policy currently under consideration can be expected to reduce pretrial misconduct (FTA or Re-arrest) by a small amount, reduce the average number of infractions, and increase the likelihood of a case reaching disposition without misconduct. Data driven policy simulations also suggest that there is room for improvement.

   (d) Given that the current policy mostly deals with infractions, there is a limit to how much improvement the agency can see from revising responses to defendant conduct. Other aspect of pretrial supervision (e.g., incentives, preemptive responses, and other non-reactionary policy choices) may have equal or larger impacts.

   (e) The data suggest caution in formulating a rigid policy, especially when the client or case is complex (e.g., higher risk or clients with repeat infractions). Under these circumstances, providing supervision staff more discretion might be advisable.

3. Predictive Bias (Task 3):

   (a) Data on items included in the current risk assessment instrument were analyzed and several items were recommended for removal from the instrument. Very little predictive bias was uncovered in the current instrument.

   (b) The revised models developed were found to be racially unbiased with regard to errors (false positive and false negative). This is despite the instrument making recommendations at disparate rates.

1.4 Report outline

This report describes the analysis conducted in support of PSA's needs as well as the findings and recommendations resulting from the analysis. It is organized as follows. The next chapter describes the re-validation effort while the chapter following that describes the analysis conducted to detect predictive bias in the instrument. Chapter 4 presents findings to support a revised instrument that addresses shortcomings that were uncovered. Chapter 5 produces findings from the risk-based supervision task and presents recommendations. A series of appendices provide detailed background data on feature distributions and categorization assumptions made in this project.

Chapter 2

Current Instrument Performance

2.1 Introduction

This chapter describes the data used in the re-validation effort and presents findings from an assessment of the currently deployed instrument. This includes a comparison of the distribution of the features between the original study and the re-validation study, an assessment of the predictive efficacy of the currently deployed instrument, as well as an analysis of additional features.

2.2 Data and Methods

Data and methods used in this project largely mirror the original validation study conducted between 2010 and 2012.

2.2.1 Data Used

This effort used data exclusively from PRISM. PSA provided detailed data mirroring what was obtained for the original RAI validation project in 2010. This included detailed demographic data, data on current charge, data on pretrial release conditions, data on bench warrants and disposition (for those cases that were disposed), along with risk assessment data (on those cases for who the RAI was administered). The RAI data included detailed weights applied to the currently included attributes to compute overall risk scores and risk categories as well as the underlying raw data for each of the 70 attributes.

Finally, PSA provided misconduct data of interest–(i) any re-arrest for a non-traffic charge; (ii) a re-arrest for a dangerous, violent or domestic violent charge; (iii) a failure to appear; and (iv) an arrest for a domestic violence charge. The last outcome is of interest only for domestic violence cases.

Re-validation Sample

The re-validation (current) sample includes all cases filed between Oct 2014 and Oct 2017. The data were extracted on or about Nov 2017. As such, cases filed in the later months (e.g., fall 2017) may include many cases that have (i) not been disposed yet or (ii) have not had sufficient follow-up period to capture misconduct. As a result, for the purpose of this chapter, data were truncated to all cases filed between Oct 2014 and March 2017. This ensured that most of the cases in the sample were either disposed by Nov 2017 or had at least a six-month follow-up period to allow for misconduct to be observed.

A second limiting factor is that the RAI is not administered in every case. Most dockets classified as CTF (Traffic) or FUG (Fugitive) do not get as RAI as a routine matter. Of the nearly 61,000 records in the original file, about 10,000 did not have an RAI administered. The remaining 50,712 records had a completed RAI with scores computed and recommendations recorded. The analysis was, by design, limited to these cases.

A third limiting factor is that not all defendants are released pretrial. While the pretrial release rate in DC is very high (approximately 92%), for the remaining detained population there can be no observed misconduct (re-arrest or failure to appear) because they were not released at all. Hence, when analyzing pretrial misconduct this group is also removed from the sample.

The final sample–excluding defendants without an RAI administered, excluding pretrial detained defendants, and including only cases filed between Oct 2014 and March 2017–consisted of 38,466 records. Further, when analyzing the domestic violence outcome, only 5,144 domestic violence cases were included.

Original Validation Sample

For comparison purposes, a similar sample was constructed from the data used in the original RAI creation and validation study (conducted between 2010 and 2012). That original sample included cases filed between Oct 2007 and Oct 2010. The original sample has similarly been reduced to include cases with valid risk assessments, a pre-trial release, as well as sufficient follow-up period. With these limitations, the sample included 37,315 cases filed between Oct 2007 and March 2010.

2.2.2 Methodology

This section reviews the methodology used in the re-validation effort. As noted above, the methodology largely replicates the analysis conducted in the original study.

Scorecard Schema

The current deployment of the the RAI at PSA includes scores for Any Re-arrest (ARR), Dangerous/Violent Rearrest (DVD), Failure to appear (FTA) and Domestic Violence Re-arrest (DVO). The models for the first three outcomes are applied to all cases while the models for the last outcome are applied only to those charged with a domestic violence charge.[1]

The Scorecard Schema is a simple strategy of evaluating risk. Individual attributes are *scored* with a numeric value based on the attribute value and its relationship with risk. For example, for categorical variables, the score might be different for each category within the attribute. The computed scores are then combined (summed or averaged) to create a total score. Because higher individual item scores are assigned to higher risk categories, the total score represents a quantitative assessment of risk—the higher this score the higher the risk. This final score can then be converted into categories—using some method for computing cut-points—and used for decision making. This generic framework is the basis of the currently deployed RAI used by PRISM. The scores for each individual category and variable are computed based on a committee model framework.

Committee Models

While the Scorecard schema is a fairly straightforward strategy to deploy, the way the scores are computed as well as the way these scores are aggregated

[1] While the current deployment uses global models for the first three outcomes, at the time of initial estimation, several subset models were also developed and tested. Based partly on the findings and on their need, PSA has chosen to use the global models. The current analysis is limited to global models/samples.

to create the total score can vary from model to model. In PSA's current RAI deployment, a committee model approach was developed and used.

For each of the outcomes, the K predictor variables were first converted into mutually exclusive categories. Next, each of these categories was given a score equal to the average training data failure rate within that category. Because there are many attributes included in the analysis, there are many individual scores that need to be combined. A simple principle component analysis (PCA) was used to convert the K different scores into K orthogonal principle components. By design, the components are orthogonal and attributes scores that are highly correlated will load higher on the same components. However, not all components are equal in terms of their internal consistency/coherence. The strongest components have the highest eigenvalues while weaker ones have smaller eigenvalues. In the final step, the K components are combined into a single score using weights proportional to these eigenvalues.

The steps described above are labeled the committee model because we can consider each attribute as an expert. The PCA essentially combines the predictions of each expert into K committees (the components). Finally, predictions from each of the committees are combined based on how well experts in a committee agree with one another (i.e., strong committees). An individual expert contributes strongly to the final score when two conditions are met: (i) the expert contributes highly to a committee and (ii) the committee is a good one. In other words, if individual items contribute weakly to a good committee or strongly to a bad committee, their influence on the final score is negligible.

As a way to normalize the final score, the lowest weight from each attribute is subtracted from all categories in that item so that the final score has a lower bound of 0. Finally, the total score is also normalized so that the maximum value of the final score is 100.

While the committee model approach was used in the original validation effort, a regression-based approach to combining the individual scores was found to be as effective in the current effort. Therefore, for the revised model developed in this report, committee models were not used. However, the final models used still follow the *Scorecard Schema*—final scores are a weighed, linear sum of underlying scores—and so can be implemented with minimal modifications to the underlying IT infrastructure.

Predictive Accuracy

The standard method of assessing the predictive efficacy of risk assessment models is by conducting receiver operating characteristic (ROC) analysis. The ROC analysis yields an area under the curve (AUC) statistics that can range anywhere between 0.5 to 1. The AUC statistic combines the false positive and false negative rates over all possible cut points in a given scoring scheme. Hence, it is an overall measure of predictive performance. The higher the AUC score, the more accurate a model is. An AUC statistic of 0.5 implies the model classification is no better than random chance. Within the criminal justice setting, typical risk assessment tools have AUC statistic between 0.6 and 0.7. AUC statistics above 0.7 are desirable, but hard to obtain. AUC statistics above 0.8 or below 0.6 are very rare. This report uses the AUC statistic for re-validation purposes.

Item Relevance

Because data and underlying populations can change, any re-validation effort should assess whether the relevance or importance of items included in the RAI may have changed–as compared with when they were last estimated. Because PSA's RAI follows a *Scorecard Schema*–linear combination of weights applied to different categories of each item–it is possible to quantify the relative importance of any particular item by computing the variance for each of the individual scored items. For example, if all males were to get a weight of 0.23 and all females a weight of 0, then applying these weights to the full sample and computing the variance of this scored variable would quantify the amount of variation in the final score one can expect due to gender. Scored Items that have a higher variance contribute more to the final score than do scored items with a lower variance (all else being equal).

In a similar manner, we can use the categories for each item and compute the average observed misconduct rate within each category as a new scoring scheme. The variance of this scheme, although measured on a different scale, also measures the relative importance of the item. However, this scheme is based on the observed outcomes and not the weights currently used in the RAI. A comparison of the relative importance of the items on the current and new scoring schemes will shed light on whether, and to what extent, the ranking of variables may have changed since the last time the RAI was deployed. Because the two scheme are on different scales, a direct comparison is ill-advised. Instead, in this report, we compare the ranking of each of the items on these two schemes. Moreover, because an altered ranking of an *important* variable is more problematic than the altered ranking of a less important variables, the analysis also takes into account the importance of the variables (using currently observed misconduct rates). These comparisons are presented graphically and discussed in the next section.

2.3 Findings

This section discusses the findings from the re-validation effort. Where pertinent, comparisons are made with the original validation exercise from 2012. In response to PSA's desire to assess the contribution of new attributes like synthetic drugs and firearm charges, these additional analysis are also presented here. Finally, because a number of jurisdictions are testing a standardized pretrial RAI developed and made available by the John and Laura Arnold Foundation, this section also presents findings from applying that model to the PSA data.

2.3.1 Attribute Distributions

Appendix A shows the distribution of attributes for both the current (revalidation) sample as well as the original (validation) sample from 2010-2012. There appear to be several changes in the distributions of the attributes as compared to the data used in the original validation effort.

Among the current charge attributes, while general categories like felony or misdemeanor appear to remain stable, the distribution of several of the underlying detailed charge categories appears to have changed substantially. For example, nearly 29% of the re-validation sample have property related charges while only about 19% of the validation sample had property related charges. Several charge categories, on the other hand, appear to have reduced in prevalence. This includes sex crimes, sexual solicitation, drug possession or distribution, and domestic violence. This is not surprising given that the original sample covers a period nearly a decade ago. It is entirely possible that the underlying populations have changed or prosecutorial priorities have shifted.

Changes were also seen in the distributions of the prior arrest related attributes. With the exception of person and property crimes, where the prevalence of prior arrests appears to have increased, the proportion of cases having all other types of past arrests seems to have decreased. Notably, prior arrests related to weapons, dangerous crimes, violent charges, and sex related crimes, all seem to have decreased (compared to the validation sample). Derivative attributes (like Lambda - the number of past arrests divided by current age) are also altered because of these shifts.

The set of attributes based on past conviction histories shows a similar shift in distribution–with the exception of person and property related crimes, the current sample has lower rates of past convictions of almost all crime categories (as compared to the original validation sample). While it is feasible that such shifts are a result of shifting populations, what is more likely is that the data being used by the current deployment of the RAI is, in some sense, different from the data that was used to develop the original models.

Demographic and social indicators appear to be fairly stable in the data with two exceptions. The attribute flagging the presence of an emotional or physical problem with the defendant has changed very dramatically. While these problems only showed a prevalence of 6.4% (emotional) and 5.2% (physical) in the original sample, they have a prevalence of nearly 19% and 9% respectively in the re-validation sample. A look at the temporal plot for these attributes between Oct 2014 and Oct 2017 also shows that the trend has been gradually increasing since the original study.

2.3.2 Assessing Predictive Efficacy

While the attributes' distributions look different in the re-validation sample compared to the original sample that, in and of itself, does not imply that the models may be performing poorly. Indeed, it is very possible that the models are robust to distributional shifts. To that end, this section discusses the accuracy of the existing model using the currently deployed scores as well as by developing a fresh set of scores. In other words, the scores that are currently estimated in the deployed RAI are assessed for their ability to predict actual pretrial misconduct. Furthermore, this predictive efficacy is compared with the predictive efficacy of the same set of attributes if the models were re-estimated. Table 2.1 provides these AUC statistics.

Couple of things are worth noting here. First, the current weights (and resulting scores) perform fairly well, given that they were estimated from a sample from nearly a decade ago. Second, the re-estimated scores perform at least as well as the current scores. However, the performance of the revised

Table 2.1: Area Under the Curve statistics: current versus re-estimated scoring scheme.

	Current Score		Re-estimated Scores	
Outcome	AUC	95% Conf.	AUC	95% Conf.
ARR	0.66	(0.66,0.67)	0.67	(0.66,0.68)
DVD	0.64	(0.63,0.65)	0.65	(0.64,0.66)
FTA	0.64	(0.63,0.65)	0.64	(0.63,0.65)
DVO	0.59	(0.57,0.62)	0.63	(0.61,0.65)

weights is only marginally better. With the exception of the domestic violence models, the other misconduct model AUC statistics are improved only by a percentage point or less. The domestic violence model improves considerable (enhancing the AUC statistic from 0.59 to a bout 0.63).

These findings are not surprising given that the same set of attributes is used in the revised models and that the current and new scores are driven largely by a few main predictors. However, the relevance of some of the items changes when the revised models are estimated and this might be important.

2.3.3 Assessing Item Relevance

This section discusses the relevance of the 70 attributes included in the original models. While the predictive efficacy of the models (currently deployed and the re-estimated) might not differ much (in the aggregate), the weights of the variables are altered because of the shifting distributions in the underlying data. This section assesses the changes in relevance between the currently deployed RAI and a re-estimated version.

Figure 2.1 shows a scatter plot of the relevance of each of the 70 attributes measured using the current model (on the $x-$axis) against the relevance of the same 70 attributes measured using the observed misconduct rates (on the $y-$axis). The size of each bubble indicates the relative importance of the particular attribute based on observed misconduct rate.

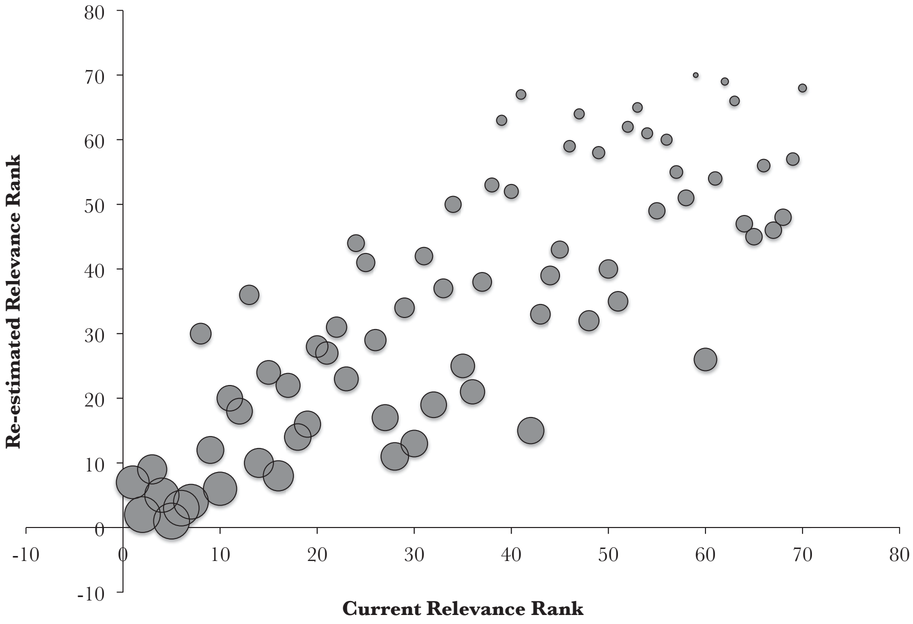


Figure 2.1: Feature importance rank: Current weights rank (*x*-axis) versus Re-estimated weights rank (*y*-axis), Outcome = Any Rearrest (Bubble size = re-estimated relevance).

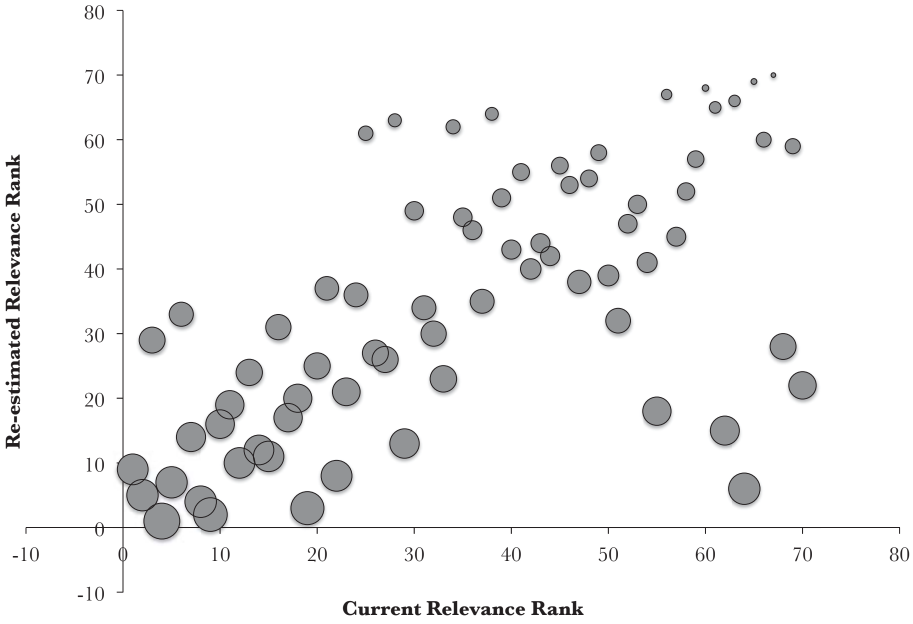


Figure 2.2: Feature importance rank: Current weights rank (*x*-axis) versus Re-estimated weights rank (*y*-axis), Outcome = Dangerous, Violent, or Domestic Violence Rearrest (Bubble size = re-estimated relevance).

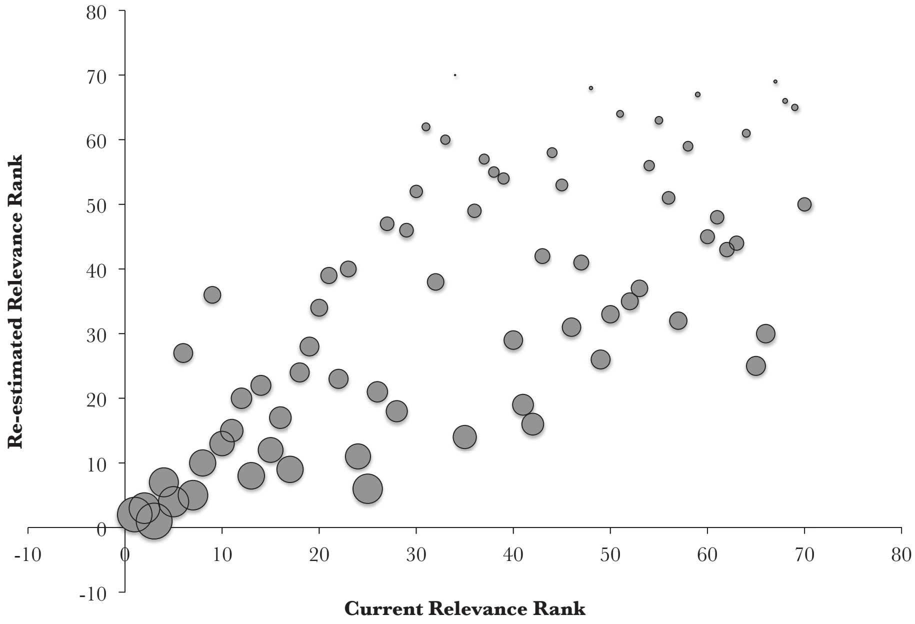


Figure 2.3: Feature importance rank: Current weights rank (*x*-axis) versus Re-estimated weights rank (*y*-axis), Outcome = Failure to Appear (Bubble size = re-estimated relevance).

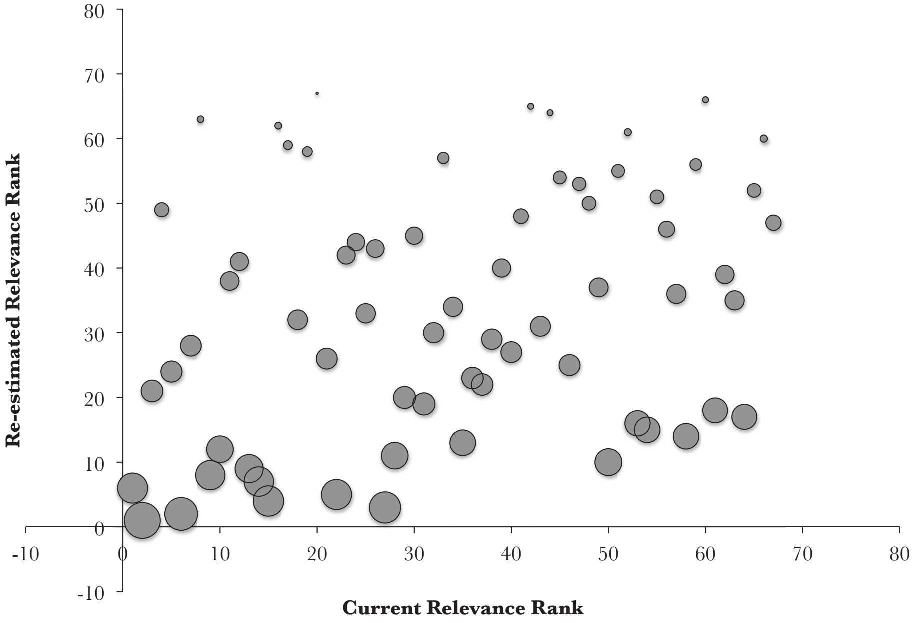


Figure 2.4: Feature importance rank: Current weights rank (x-axis) versus Re-estimated weights rank (y-axis), Outcome = Domestic Violence Rearrest (Bubble size = re-estimated relevance).

Two properties are worth noting in these plots. First, if the ranking of the relevance of the attributes does not change then all the bubbles would be clustered along the 45 degree line. When the scatter plot is dispersed away from the 45 degree line, this implies that the relevance of the attributes have changed. Second, the largest bubbles should be clustered around the origin (near the bottom left corner of the plot). In other words, even if the ranking of attributes change, the most important attributes continue to remain important.

Figure 2.1 shows that, by and large, these desirable properties are observed for the any rearrest outcome. While there is some deviation of the scatter plot away from the 45 degree line in the middle, the largest bubbles are clustered around the origin of the plot. This suggests that, while there are some changes in the relevance of the included attributes, these are mostly related to attributes that are less important. The most important attributes (the largest bubbles) continue to remain so. There appear to be one or two exceptions. For instance the medium sized bubble above 70 on the $x-$axis indicates that this variable is not very relevant in the current deployment of the RAI but should be. Its ranking (in terms of importance) changes from nearly 69 to about 15. This attribute happens to be the question about Physical problem. As noted in the previous section, the attribute has changed dramatically in the underlying data so a change in its relevance is not surprising.

A similar analysis of the dangerous, violent, and domestic violence models (Figure 2.2) shows that the relevance of items have changed somewhat more than the any rearrest model. Here there are two attributes that have gone from having very low relevance to having very high relevance. Upon closer examination, these two items are found to be the emotional and physical problem flags. Other than these, most of the important attributes retain their relevance while the less important attributes continue to be so.

The FTA model item relevance plot (Figure 2.3) shows a fairly robust/stable model. The plot is largely scattered about the 45 degree line and most of the large bubbles are clustered about the origin.

The Domestic Violence plot (Figure 2.4) shows that there are large changes in the relevance of items. The plot is not scattered about the 45 degree line and several of the medium to large bubbles are not located at the origin. This suggests large shifts in the weights of the attributes and therefore their contribution to the overall score. This is also not surprising given the AUC statistics presented in Table 2.1 where the biggest change in predictive efficacy was obtained in the DVO models.

The item relevance analysis, along with the analysis of the attribute distributions and the model AUC scores, suggest that there is sufficient shifts in the underlying data and underlying relationship between the predictors and the outcomes of interest to warrant revision of the weights. The revisions of weights alone may not, to be sure, improve model predictive performance much. However, the weights will be more consistent with current evidence regarding attributes and pretrial misconduct. Should the models, not just the weights, be revised, predictive performance may improve.

2.3.4 Implementation Issues

The comparisons of the data distributions discussed in the previous section were based on identically defined features between the original study and the re-validation effort. Despite that, the above analysis suggests that there might be some changes in the underlying population and that the relevance of several of the items in predicting misconduct may have changed. This section discusses some issues that were uncovered regarding the currently deployed instrument that shed some light on models, their current performance, and possible directions for model revisions.

Physical and Emotional Problem Measures

As part of the initial interview that is conducted by a defendant, PSA solicits self-report information on whether or not a suspect has any physical or emotional problems. These self-report items form a part of the risk assessment computations. While analyzing the underlying data, it became apparent that there has been sufficient shift in the proportion of defendants self-identifying themselves as suffering from physical or emotional issues. Upon further investigation, it was uncovered that this may stem from two distinct reasons.

First, the method of gathering information on physical and emotional health may have changed over time. For example, with a redesign of the data entry forms, PSA began to record “non applicable” or “null” responses that were initially left blank. As a result, the original data query scripts were producing misleading inputs into the risk assessment calculations. Fortunately, the weights for physical and emotional health are fairly low, and, as such, the risk assessments are still mostly correct (albeit with some noise). Moreover, this is a simple data implementation issues that can be resolved by modifying the query code.

Second, the language of the question that queries the physical and emotional health status of a defendant has also changed since 2012. While previously the query was about whether the defendant felt (s)he had a problem, the modified language asks if the defendant feels (s)he has a physical or emotional health problem "that might keep them from making their court appointments". The modification may result in a slight increase in the proportion of defendants claiming to have physical or emotional problems but it might also strengthen the links between this feature and failure to appear. Since defendants are directly identifying the feature when it will interfere with their ability to make court appearance. This second problem cannot be solved merely by fixing the query and will require re-estimation of the weights that capture the effect if these measures on pretrial misconduct and, as such, their relevance in the RAI will likely change.

Fortunately, the current implementation of the risk assessment instrument does not use very high weights for the Physical or Emotional Problem measures. As a result, despite the implementation issues, models' performances are not compromised.

Criminal History Measures

A second problem that was uncovered with the current implementation of the RAI is with the criminal history measures. During the original validation effort, in 2012, all criminal history measures used in the analysis/models were limited to a 10 year window. In other words, all models used prior arrests or convictions (be they for specific types of crimes or in the aggregate) only within the past 10 years. The main reason for this was because of unreliability of the older legacy criminal history data as well as to exclude very old criminal history records that might be irrelevant when assessing the defendants current risk. In addition, PRISM retains criminal history data in two separate tables—internal and external. The internal criminal history records pertain to arrests (and conviction) that are DC based and the external records pertain to arrests (and conviction) that might have occurred outside of DC (mostly in Maryland or Virginia). In conversations with PSA staff it was uncovered that in 2012 PSA was not confident with the validity of the external records and the older internal records. As a result only the internal records, and only those within a 10 year window, were used for the validation effort.

Over the years following the 2012 implementation, however, the reliability of the external records and the older internal records has improved. As a result, in the current implementation, these records contribute to the risk assessment. As can be seen from the AUC scores discussed earlier in this chapter, inclusion of these additional records in the criminal history measure does not necessarily harm the risk assessments. However, it does make the currently deployed model different from the originally estimated model. As with the mental and physical health measures, this can readily be remedied by revising the model weights to accurately reflect the underlying data, should they provide added predictive power over and above the 10 year history measures.

Charge Information

The final implementation issues that was uncovered pertains to the charge level information. The currently deployed model uses a total of 70 features. More than half of these features use charge level information. For example, the model includes 14 features that use the defendant's current charge and another 26 features are charge-specific criminal history measures (both arrest and convictions).

Charge information in maintained by PSA in a large look-up table that is updated manually whenever new charges are encountered by the system. The procedure works fine for current charge related features because these are more directly visible and new charges are picked up easily. However, the look-up table does not provide a robust method for classifying the historical charge records. This is because the look-up table was manually developed several years ago and, given the large number of unique charges in the historical data (recall that these will include deprecated charge codes as well as current ones), extremely rare charges may not have been classified well. However, these rare charges will continue to be incorrectly classified as "other" or some such generic category even if they are serious. Simple spelling variations etc. can also slip through the algorithm and contribute to error in the historical charge-related feature. Unfortunately, unless there is a labor intensive manual effort to correctly classify those charges, the data inaccuracies will continue. A simple coding fix will not solve the problem. While some NLP (Natural Language Processing) software or RegEx (Regular Expression) scripts may help diminish the concerns, a full manual validation would be difficult.

A related concern with charge information is that the external records continue to have very poor and unreliable charge descriptors. At best, these data can be reliably used to distinguish felony and misdemeanor charges. But anything more detailed promises to be error prone.

While every attempt was made with PSA staff to identify and tackle the source of the charge related data noise, it appears unlikely that the charge level criminal history details currently used by the RAI are a reliable, robust source of information about the defendant. As a result, in the revised model presented in Chapter 4, a number of different criminal history measures were developed that do not rely so heavily on charge level information. While making use of the felony and misdemeanor distinction, these measures rely more heavily on the distinction between internal and external records as well as on giving more/less weight to more recent records.

2.3.5 Additional Attributes

PSA was also interested in assessing the value of including two additional features—testing positive for synthetic drugs at lockup and whether a defendant was charged with a firearms charge—into the existing RAI model.

While PSA retains data on drug tests at lockup, detailed information on firearm related charges is lacking in the data. As noted in the previous section, charge information is available in a mixed format (mostly text). Hence, a murder charge will be recorded as “Murder” irrespective of whether it resulted from a firearm of some other weapon. As a result, additional attribute analysis could only be conducted for synthetic drug positive tests and not for the impact of firearm related charges.

The drug test at lockup data showed that very few cases tested positive for synthetic drugs. Upon closer analysis it was found that the positive drug tests all existed in a few months (late 2015 and fall of 2017). PSA staff confirmed that K2 (synthetic drug) testing only started in Oct 2015. Moreover, after initially starting the testing, it was conducted sporadically until mid-2017. As a result, there were very few cases that tested positive for K2. This trend is clearly visible in Figure 2.5. There were only a total of 200 cases that tested positive.

Despite the small samples, a simple unconditional analysis suggested that those testing positive for K2 did have higher misconduct rates than those that did not. Figure 2.6 shows that those testing positive for K2 had about 30% re-arrest and FTA rate and about 12% DVD re-arrest rate. These are the middle bars in the clusters of three bars in figure 2.6. The bar on the left, typically lower than the middle bar, shows the misconduct rate for those testing negative for K2. This suggests a large increase in misconduct among those testing

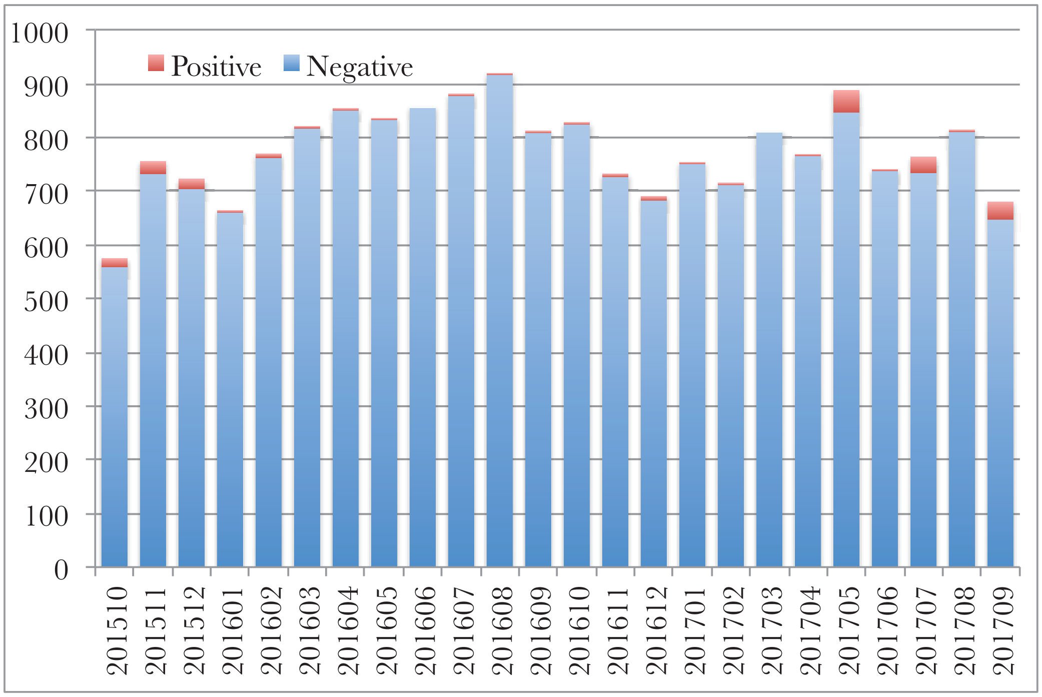


Figure 2.5: Trends in the number of clients testing positive for synthetic drugs.

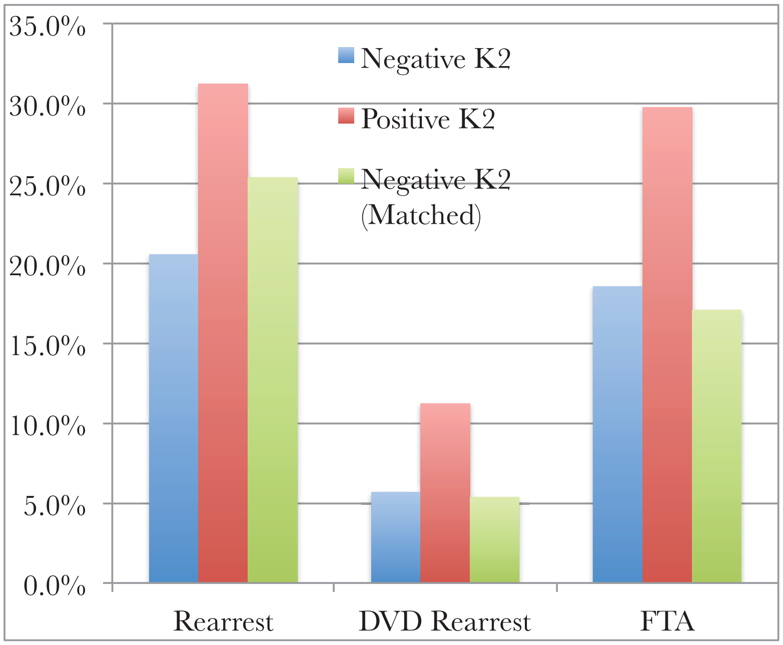


Figure 2.6: Risk associated with clients testing positive for synthetic drugs.

positive for K2. However, this relationship is completely unconditional and does not control for any other attributes. It is very likely that those testing positive for K2 are already very high risk and that K2 is merely a proxy for that risk level. To address this concern, a matched sample was developed based on risk level. From the sample of cases testing negative for K2, defendants were selected with the same risk profiles as the K2 positive ones. The outcomes for this "matched" sample is presented in the bar on the right. Despite some changes in the misconduct rates, there is still a large difference between the misconduct among those testing positive for K2 and defendants matched on risk level but who tested negative for K2. Hence, defendants testing positive for K2 did have higher misconduct rates even after controlling for (or adjusting for) risk level.

2.3.6 Arnold Foundation Model

The Laura and John Arnold Foundation has funded an effort to develop a pretrial risk assessment instrument that utilizes a minimal set of predictors (called the Public Safety Assessment–PSA). Information about the instru-

Table 2.2: The Laura and John Arnold Foundation PSA Model.

Attribute	FTA	ARR	DVD
Age at arrest		X	
Current violent offense			X
Current violent offense & 20 yrs or younger			X
Pending charge at time of offense	X	X	X
Prior misdemeanor conviction		X	
Prior felony conviction		X	
Prior (misdemeanor or felony) conviction	X		X
Prior violent conviction		X	X
Prior FTA in past two years	X	X	
Prior FTA older than two years	X		
Prior sentence to incarceration		X	

ment can be obtained from the Arnold Foundation (`www.arnoldfoundation.org`).[2]

Table 2.2 provides a list of the attributes included in their model and the outcomes they are used to predict. As part of the re-validation effort, the Arnold Foundation model was estimated using the PSA data and compared with PSA's currently deployed RAI. The variables used by the Arnold Foundation model were all available in the PSA data with one exception. Their model uses *Prior Sentence to Incarceration* to estimate the general rearrest risk. This measure was not directly available in the PSA data and was proxied by creating a flag combining prior conviction for any of the serious charges (dangerous, violent, person, weapon, and domestic violence). The Arnold PSA produces scores for three types of risk–(i) risk of FTA, (ii) risk of new criminal activity, and (iii) risk of new violent criminal activity. These three are, respectively, related to the PSA risk models for (i) FTA, (ii) ARR, and (iii) DVD.

The estimated AUC statistics for the Arnold model applied to the PSA data are presented in Table 2.3. Findings suggest that the Arnold PSA model performs almost as well as PSA's ARR and DVD models but substantially

[2] More specifically, details about the underlying variables and weights are available from `http://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf`.

Table 2.3: Predictive efficacy of the Laura and John Arnold Foundation PSA Model applied to DC PSA data

Outcome	AUC	95% Conf.
ARR	0.67	(0.66,0.67)
DVD	0.60	(0.59,0.61)
FTA	0.63	(0.62,0.64)

worse than PSA's DVD model. The ARR and FTA AUC statistics of 0.67 and 0.63 are at par with those estimated using the PSA model. However, the Arnold model produces an AUC statistic of only 0.60 for the DVD model while PSA's model generated an AUC statistics nearly 5 percent points higher (0.65). Therefore, unless the Arnold model is re-calibrated to reflect the District of Columbia population, it will not be very predictive of risk of violent re-arrests.

2.4 Conclusion

The analysis presented in this chapter was aimed as a baseline analysis that would shed light on how the existing RAI implementation at PSA is performing and whether, and to what extent, the underlying population and risk profiles may have changed.

2.4.1 Summary of Findings

Findings suggest that, while the underlying data have changed somewhat, this might reflect both a change in underlying population as well as a change in the source of the underlying data. This would warrant a re-estimation of the models to ensure the weight reflect the correct data. However, despite the changes in the underlying data, the predictive efficacy of the models is fairly robust. With one exception–domestic violence–PSA's RAI models perform fairly well and re-estimation of the models would improve predictive efficacy, but only marginally.

In terms of the additional predictors that PSA was interested in assessing, testing positive for synthetic drugs is related with misconduct, net of other risk factors.

Finally, a comparison of PSA's RAI with the simpler Arnold Foundation instrument suggests that PSA's RAI could be simplified to include a minimal set of attributes. The Arnold Foundation model would, however, need to be re-calibrated to reflect PSA's data. Otherwise, the models would result in a reduction in PSA's ability to predict dangerous and violent re-arrests.

2.4.2 Recommendations

Based on this analysis, it can be recommended that:

1. A revised set of models should be estimated as part of the re-validation effort;
2. The revised models should not rely on charge-specific criminal history measures;
3. The revised models will need to develop other measures from the criminal history data to better leverage it; and
4. The revised models should include synthetic drugs as one of the several drug-test related features that are include in the current model.
5. Finally, because the revised models will include different features as well as different weights, it is to be expected that the revised scores will have different range of values (than the current scores). As such, it is recommended that the cut-points used by the instrument for classifying individuals into low, moderate, high, very high risk categories should also be modified.

Chapter 3

Predictive Bias

3.1 Background

In recent years, the field of risk assessment within the criminal justice system has come under increasing scrutiny for the possibly biased algorithms underlying their risk assessment instruments. In particular, some have questioned whether the recommendations made by these instruments are racially neutral or favor non-minority groups.[1] This chapter reports on analysis conducted to address this issue.

3.2 Data and Methods

3.2.1 Data Used

Detailed descriptions of the data used for this analysis are provided in the previous chapter. Table 3.1 provides a more detailed break down of the data used for this study by race categories. Defendants with cases filed between Oct 2014 and Oct 2017 is the starting point of the sample. A total of 50,449 clients were assessed for risk within this period. However, not all of them were released pretrial—46,731 were released at some point prior to disposition. Data for misconduct was collected through the same period. In order to allow

[1] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). "Machine Bias. There is software that is used across the county to predict future criminals. And it is biased against blacks." `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Table 3.1: Sub-samples used in this report, by race.

	White	Black	Hispanic	Total
All defendants scored for risk*	3,857	43,268	2,636	50,449
	7.6%	85.8%	5.2%	
Pretrial release (PTRel)	3,657	39,943	2,474	46,731
	7.8%	85.5%	5.3%	
Domestic violence cases	225	5,623	303	6,222
	3.6%	90.4%	4.9%	
PTRel + Re-arrest sample**	3,031	32,793	2,094	38,477
	7.9%	85.2%	5.4%	

* 688 defendants with Race=Other are omitted from analysis
** Oct 2014 through March 2017

for enough follow-up period post case filing to observe pretrial misconduct, the analysis sample was further limited to only cases filed on or before March 31, 2017. This permits a minimum of 6 months follow-up for clients whose cases were not disposed off by Oct 2017. There were a total of 38,477 clients with cases filed between Oct 2014 and Mar 2017 with a pretrial release.

Eighty-six percent of the full sample were African Americans with the remainder mostly Whites (7.6%) or Hispanics (5.2%). There were an additional 688 cases where the defendants were coded as being Asian or other race (or were missing race information). These 688 cases have been removed from the analysis for convenience.

The various sub-samples described above are similar to the overall sample—Blacks constituted a majority of the cases (between 85% to 90%) while whites and Hispanics formed the remainder (about 8% and 5% respectively). The Domestic Violence sub-sample had a slightly different racial make-up—90.4% were Black while 3.6% were White and 4.9% were Hispanic.

Table 3.2 shows the distribution of the risk scores, pretrial release rates, and misconduct rates by racial subgroups. In general, Blacks have much higher average risk scores than Whites or Hispanics. The average re-arrest score (for any crime) is 37.3 for Blacks, followed by 24 for Hispanics and 20.5 for Whites. The average dangerous/violent rearrest score is 36 for Blacks, followed by 26.3 for Hispanics and 22.4 for Whites. Similar trends are seen for the FTA and domestic violence risk scores.

Table 3.2: Average risk scores and misconduct rates, by race.

	White	Black	Hispanic	Total
Average Risk Scores				
Any Rearrest	20.5	37.3	24.0	35.3
Dangerous/Violent Rearrest	22.4	36.0	26.3	34.5
Failure to Appear	24.6	36.2	26.5	34.8
Domestic Violence Rearrest	26.4	35.7	28.2	35.0
Pretrial Release Rate	94.8%	92.3%	93.9%	92.6%
Misconduct Rates				
Any re-arrest	13.3%	27.2%	21.4%	25.6%
Dangerous/Violent re-arrest	1.3%	4.5%	3.2%	4.1%
Failure to appear	19.1%	22.3%	20.3%	21.9%
Domestic violence re-arrest	7.2%	10.1%	9.4%	9.8%

Not surprisingly, the overall pretrial release rates follow these same trends—Blacks have the lowest pretrial release rate (92.3%) followed by Hispanics (93.9%) and Whites (94.8%).

Because the original RAI was developed and deployed using similar outcomes—any re-arrest, dangerous/violent re-arrest, failure to appear, and domestic violence re-arrest—the observed misconduct rates follow the same trends as the underlying average scores. In general, the observed misconduct among Black defendants is typically the highest, followed by Hispanics and then Whites. A cursory look at the estimates reported in Table 3.2 lends some confidence regarding racial bias in the instrument. While the instrument does score Black more severely than Hispanics and Whites, it appears that the scoring is consistent with observed misconduct rates. However, simple aggregate comparison might hide biases regarding differential predictive efficacy of the instrument or biases in terms of the errors that the instrument makes. The next section provides more detailed analysis of the data.

3.3 Findings

The simplest method to move beyond aggregate comparisons is to study the data graphically. That permits us to study the full distribution of the scores by race as well as the full distribution of the relationship between risk score and misconduct (also by race). Figures 3.1 through 3.4 present the distribution of the risk scores of the different race groups. Consistent with the aggregate risk scores, it is seen that Black score distributions for each of the scores are shifted to the right of Whites and Hispanics. However, there are some interesting nuances.

The any re-arrest score shows bi-modality among the Black defendants but not among White and Hispanic defendants. A majority of White and Hispanic defendants have scores in the 0 to 20 range and then the distribution tapers off towards the rights. In other words, there is one large cluster of individuals on the lower end of the score. The distribution for Black defendants shows two modes—there is a cluster, like White and Hispanic defendants—at the lower score(between 0 and 20). However, there is another cluster of Black defendants who have higher risk scores (about 40). It is this second cluster that appears to drive the overall higher scores for Blacks. The dangerous/violent scores show a similar trend but with a more pronounced second cluster among Blacks. The FTA score appears to show three clusters among Blacks (low about 18, medium about 30 and high about 60). Similarly, the domestic violence score shows a shifted distribution among Blacks relative to Whites and Hispanics. In each of these distribution graphs, we can see a cluster of Black defendants who are similar to the lower-risk White and Hispanic defendants (clustering around 20).

While this graphical analysis does not speak directly to the presence or absence of racial bias in the algorithms, it does highlight that while White and Hispanic defendants can be well described by a single mean score (the typical defendant), the same makes little sense for Black defendants.

The next set of graphics (figure 3.5 through figure 3.8) show a more detailed look at the relationship between risk scores and misconduct rates. To make the plots easier to understand, the underlying scores were first converted into quantiles (20 for each score). The average misconduct rate was then computed within each quantile and plotted for each race group. To ease exposition, fitted curves were also plotted to show the overall aggregate relationship between the misconduct rates and the quantiles.

As expected, all of these graphs show that the expected misconduct rate increases as the scores increase. Moreover, the rate of increase is very similar among Blacks, Whites and Hispanics. There is one exception, though. Figure 3.7 shows that the relationship between FTA rates and FTA scores is much steeper (better) among Whites than it is among Blacks and Hispanics except the very highest quantile. This suggests that the FTA risk score is a *better* model for Whites than it is for Blacks and Hispanics. On average, the FTA rates for higher risk White defendants are higher than the FTA rates among similarly scored high risk Black or Hispanic defendants.

The graphics analysis therefore suggests that the risk scores are distributed differently among Black defendants compared with White and Hispanics defendants. On the other hand, they also suggest that the relationship between risk and misconduct is fairly stable among defendants of all races with the exception of FTA.

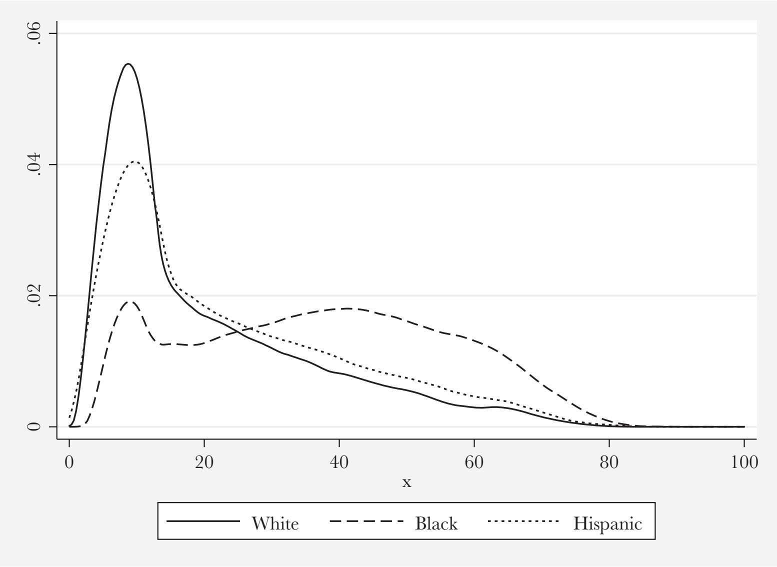


Figure 3.1: Distribution of Rearrest Risk Score, by Race.

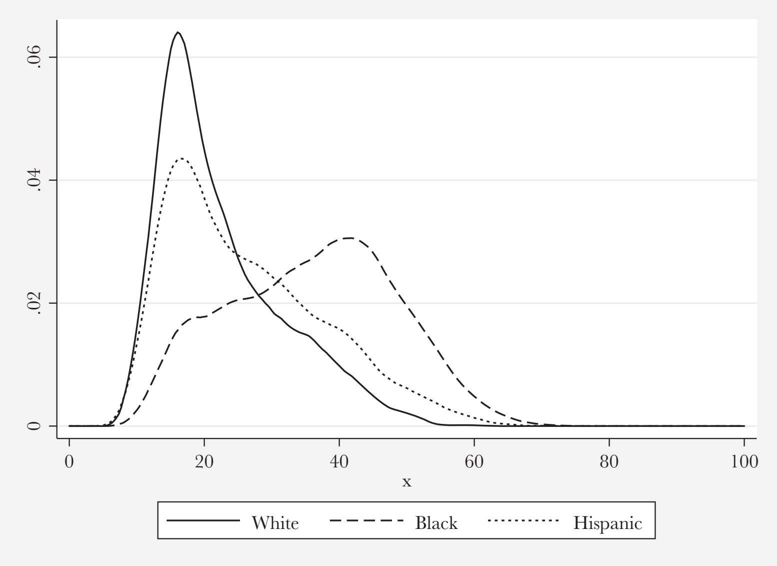


Figure 3.2: Distribution of Dangerous/Violent Rearrest Risk Score, by Race.

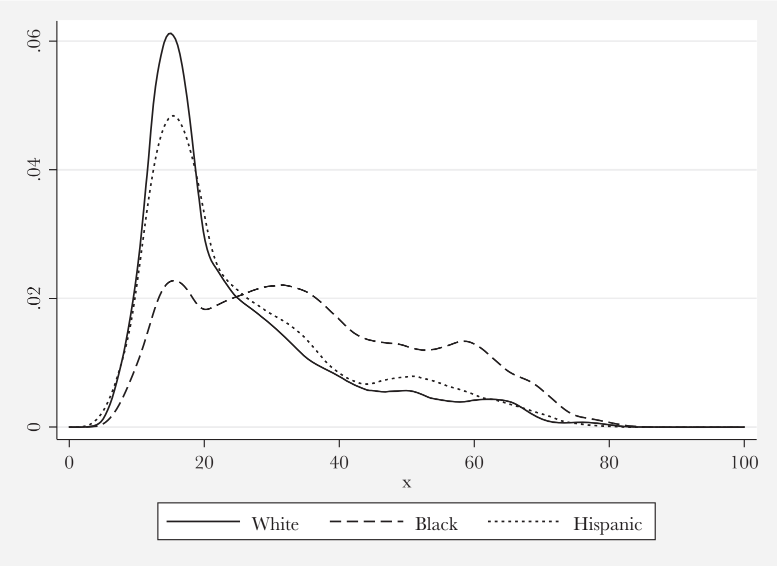


Figure 3.3: Distribution of Failure to Appear Risk Score, by Race.

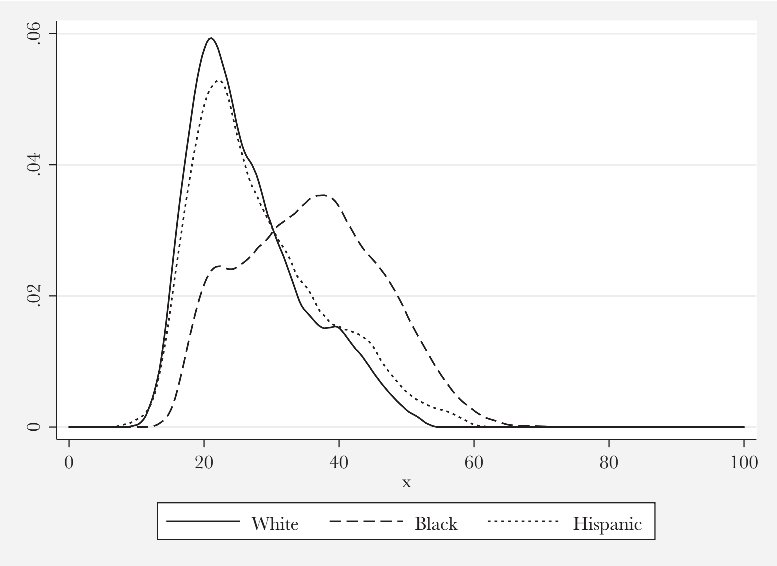


Figure 3.4: Distribution of Domestic Violence Rearrest Risk Score Among DVM Cases, by Race.

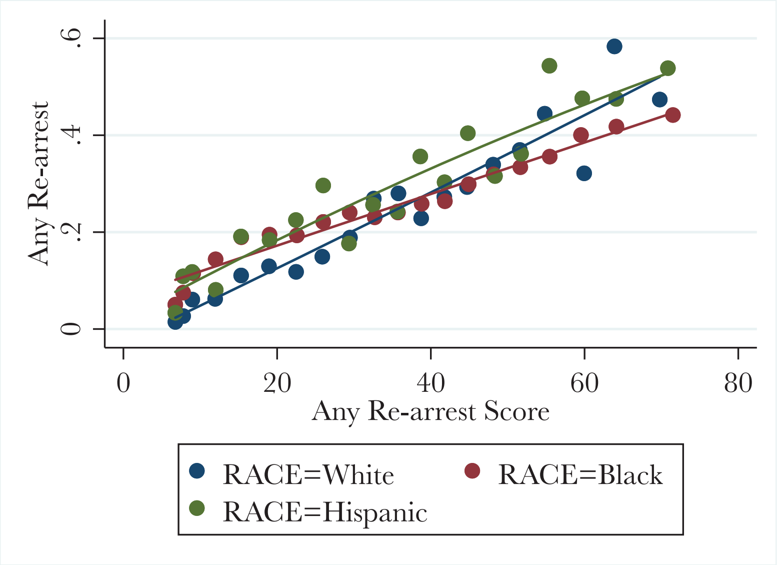


Figure 3.5: Rearrest Rate for 20 Quantiles of Rearrest Risk Scores, by Race.

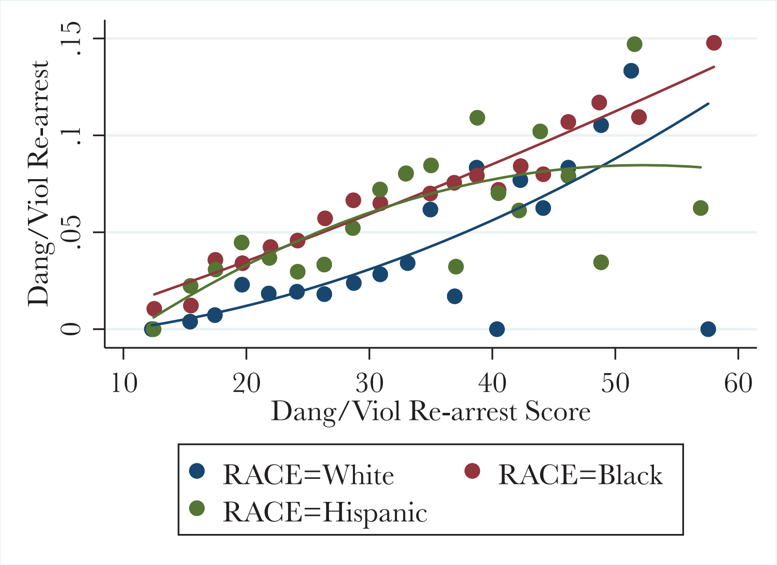


Figure 3.6: Dangerous/Violent Rearrest Rate for 20 Quantiles of Dangerous/Violent Rearrest Risk Scores, by Race.

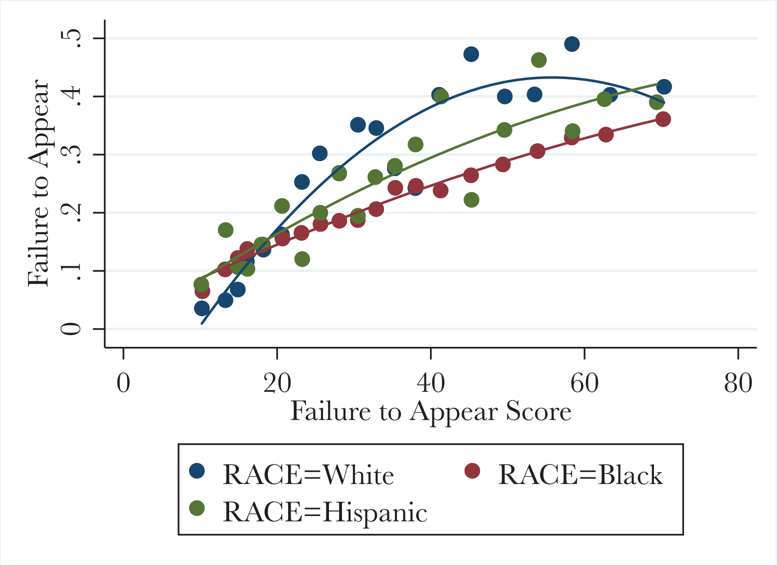


Figure 3.7: FTA Rate for 20 Quantiles of FTA Risk Scores, by Race.

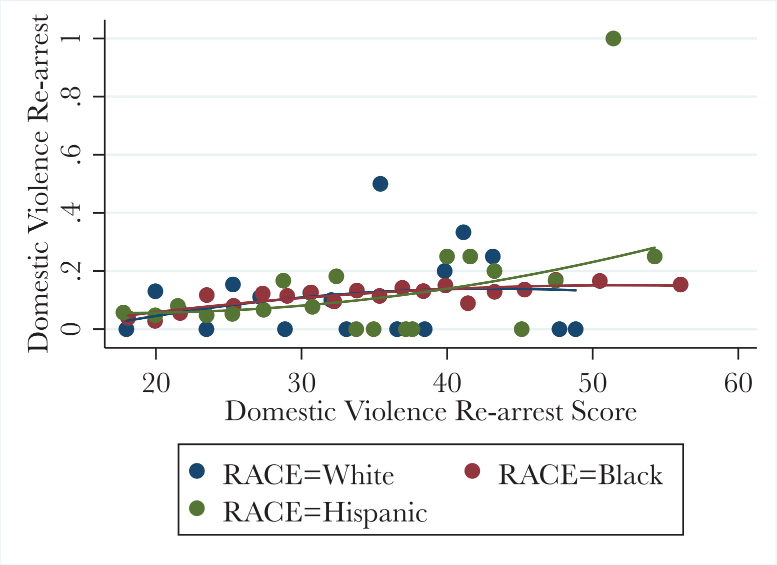


Figure 3.8: Domestic Violence Rearrest Rate for 20 Quantiles of Domestic Violence Rearrest Risk Scores (among DVM Cases), by Race.

While the graphical analysis described above sheds some light on the distribution of risk score and their relationship with misconduct for different racial groups, the analysis ignores the classification of risk (low, moderate, high, etc.) that is used by the agency in making decisions. In other words, while the distinction between the 19th and 20th quantile may be of interest, if both of these are collapsed into a Very High risk group then they matter little from an operational point of view. Table 3.3 shows the distribution of defendants of different races into different risk categories along with their observed misconduct rates.

The data support the general findings from the graphical analysis. In general, Blacks are more concentrated in the higher risk categories than Whites and Hispanics. However, there is some variation in the misconduct rates among Blacks, Whites and Hispanics for different categories. Misconduct rates within risk categories are more similar between Blacks and Hispanics than Whites. With few exceptions, the misconduct rates among Blacks and Hispanics are within a few percentage points of each other within each risk category. However, there are differences between the misconduct rates between Blacks and Whites within risk categories. With the exception of Domestic Violence scores, the misconduct rate for Whites is typically lower than Blacks among the low risk categories while it is higher than Blacks among the high risk groups. The difference, while present, are not large with the exception of the FTA score and outcome. Consistent with the graphical analysis discussed earlier, the FTA rates among high risk White defendants is much higher than Black or Hispanic defendants in those high risk categories. This suggests that, even though Black defendants are classified as higher risk of FTA at higher rates than White defendants (33% of Blacks are classified as high or very high risk of FTA compared to only 10% of Whites), observed FTA rates among Whites is nearly 10-12 percentage points higher than Blacks and Hispanics.

The graphical analysis as well as the more detailed analysis of risk categories presented above point towards a generally bias free instrument with the exception of the FTA instrument. To take a closer look at that, the next set of tables look at statistics computed related to errors. These measures are all based on of a 2 X 2 contingency table. On one axis is the predicted risk levels (Positive or Negative prediction) while on the other axis is the observed misconduct (whether the predictions were True or False). The table below shows the possible combinations:

Table 3.3: Percent distribution in risk score based categories and misconduct rates, by race.

	% in Risk Category			Misconduct Rate*		
	White	Black	Hispanic	White	Black	Hispanic
Any Rearrest						
Very Low	4.1%	1.0%	3.1%	3.2%	4.6%	5.4%
Low	55.8%	20.3%	47.1%	4.9%	12.4%	10.2%
Moderate	29.1%	38.1%	32.8%	18.1%	23.4%	25.0%
High	7.5%	25.2%	11.8%	36.4%	32.7%	36.6%
Very High	3.4%	15.4%	5.2%	43.9%	39.4%	52.4%
Dangerous/Violent Rearrest						
Very Low	0.1%	0.0%	0.1%	–	–	–
Low	63.6%	21.9%	48.2%	0.6%	2.9%	1.5%
Moderate	33.3%	50.6%	42.8%	3.0%	6.9%	6.3%
High	3.0%	23.9%	8.1%	9.8%	10.8%	10.1%
Very High	0.1%	3.6%	0.8%	–	14.8%	5.6%
Failure to Appear						
Very Low	4.1%	2.4%	5.3%	5.3%	6.3%	5.9%
Low	55.7%	23.3%	47.7%	9.1%	13.2%	14.0%
Moderate	28.4%	42.2%	31.5%	30.5%	20.9%	24.0%
High	8.7%	23.1%	12.5%	42.0%	30.8%	34.7%
Very High	3.0%	9.0%	3.0%	43.8%	32.6%	37.8%
Domestic Violence Rearrest						
Very Low	–	–	–	–	–	–
Low	34.2%	10.8%	29.6%	5.7%	3.4%	6.1%
Moderate	37.8%	34.7%	37.8%	7.6%	11.8%	7.8%
High	9.3%	27.7%	11.5%	17.6%	13.7%	13.3%
Very High	0.0%	7.0%	2.0%	–	15.9%	–

* Only calculated when cell size >= 20

	No-Misconduct	Misconduct
Low Risk	True negative	False negative
High Risk	False positive	True positive

A combination of the two axes produces four categories—true negative (TN), false negative (FN), true positive (TP) and false positive (FP)—and a number of statistics have been developed that use these categories.

The main criteria used to assess the efficacy of risk assessment instruments within the criminal justice systems is the Area Under the Curve (AUC) statistic. This number is based on the concepts of Sensitivity and Specificity. These are defined as:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.1}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.2}$$

Specificity is computed as the proportion of those who had an observed misconduct that were assessed to be at high risk of misconduct while sensitivity is the proportion of those who did not have a misconduct that were assessed to be at low risk of misconduct. In other words, these are ways of gauging how good the instrument is at isolating high risk among those who failed or isolating low risk among those who did not fail. The AUC statistic is a combination of the two quantities for all possible cut-points (or categories) into one aggregate measure. The higher the AUC score the better the RAI is at separating out the high and low risk defendants. Table 3.4 shows the calculated AUC scores for each of the scores (in the first three column) and the categories (last three column) for each racial groups. To make comparison possible, the table also shows the 95% confidence (lower and upper) bounds for each of these numbers.

The table shows a distinct pattern. The risk assessment instruments are all better at separating the White defendants into high and low risk groups than Black defendants. In other words, the instruments are more *specific* and/or more *sensitive* when assessing Whites than while assessing Blacks. In general all of the AUC scores are higher among Whites than among Blacks and Hispanics. This is true even after considering the confidence bounds around the estimates.

One of the draw back of the AUC statistic is that it is based on all possible cut-points or categories. This is unrealistic as one would never envision using

Table 3.4: Area Under the Curve (AUC) statistics, by race.

	Raw Scores			Risk Categories		
	White	Black	Hispanic	White	Black	Hispanic
Any Rearrest						
AUC statistic	0.79	0.65	0.72	0.75	0.63	0.70
95% low bound	0.77	0.64	0.69	0.72	0.63	0.67
95% high bound	0.81	0.66	0.74	0.77	0.64	0.72
Dangerous/Violent Rearrest						
AUC statistic	0.76	0.63	0.68	0.72	0.62	0.68
95% low bound	0.70	0.62	0.63	0.66	0.61	0.63
95% high bound	0.82	0.64	0.73	0.79	0.63	0.72
Failure to Appear						
AUC statistic	0.75	0.63	0.66	0.71	0.62	0.64
95% low bound	0.72	0.63	0.62	0.69	0.61	0.61
95% high bound	0.77	0.64	0.69	0.74	0.62	0.67
Domestic Violence Rearrest						
AUC statistic	0.65	0.58	0.63	0.60	0.58	0.61
95% low bound	0.51	0.55	0.49	0.43	0.55	0.47
95% high bound	0.80	0.61	0.77	0.76	0.60	0.75

a low cut-point (e.g., 20 in our scores) to identify high risk defendants. But the AUC score is computed for all possible categories. Moreover, it is a measure of the ability of the instrument to score observed failures with a high value and score observed non-failures with a low value. A more prospective measure of predictive efficacy is to use the current cut-points or risk classes and to base calculations on those assessed of being at high risk or those assessed of being at low risk. These more direct measures are the False Discovery Rate and the False Omission Rate. These are defined as:

$$\text{False discovery rate (FDR)} = \frac{FP}{TP+FP} \tag{3.3}$$

$$\text{False omission rate (FOR)} = \frac{FN}{TN+FN} \tag{3.4}$$

The *FDR* is the proportion of those scored at a high risk of misconduct who did not have a misconduct and the *FOR* is the proportion of those scored at a low risk of misconduct who did have a misconduct. Table 3.5 shows these calculations using the risk categories computed from the underlying risk scores. These numbers paint a slightly different picture with regards to the errors committed by the RAI. In general, the FDR is only slightly higher among Blacks compared with Whites and Hispanics. There are two exceptions. First, the FDR for the Dangerous/violent instrument is marginally higher among Whites than Blacks. Second, and more problematic, the FTA FDR is 57.4% among Whites but 68.6% among Blacks. This means, the error made by the instrument is about 11.2 percentage point higher among Blacks than Whites.

The false negative rates (FOR) paint a similar picture—in general, the FOR is higher among Black defendants than Whites (with the any rearrest FOR difference being 7.5 percentage points)—suggesting that when it identifies low risk defendants, the error rate among Blacks is higher than Whites.

3.4 Conclusion

3.4.1 Summary of Findings

The analysis presented in this chapter was designed to assess algorithmic bias in the RAI as currently deployed by PSA. While the RAI is currently undergoing a re-validation, and will be revised, the analysis here suggests that the

Table 3.5: False Positive and False Negative rates using individual risk score based categories, by race.

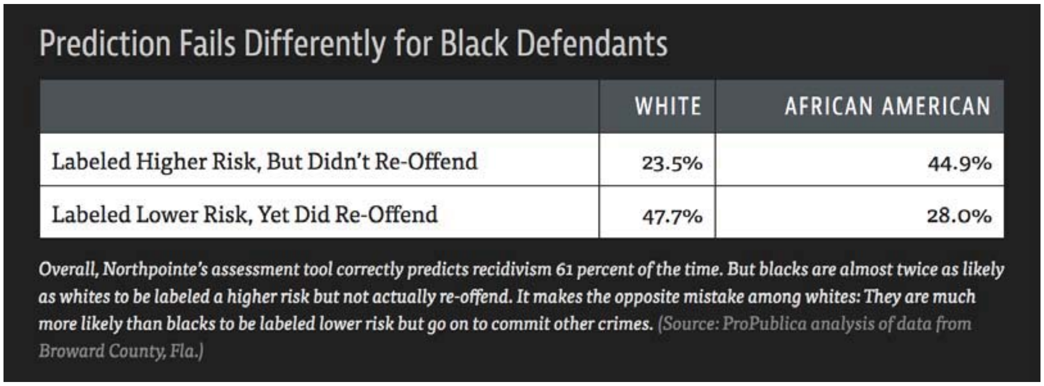
	False Positive Rate*			False Negative Rate**		
	White	Black	Hispanic	White	Black	Hispanic
Any Rearrest	61.1%	64.5%	58.0%	4.7%	11.9%	9.8%
Dang/Viol Rear.	90.3%	88.7%	90.4%	0.6%	2.9%	1.5%
Failure to Appear	57.4%	68.6%	64.6%	8.8%	12.4%	13.1%
Dom Viol Rear.	82.4%	85.9%	82.9%	5.7%	3.4%	6.1%

* Very high + high together are predicted to fail
** Very low + low are predicted to not fail

RAI, as it is currently implemented, is mostly unbiased. While risk scores and misconduct rates vary by race, the relationship between the risk scores and observed misconduct remains fairly stable across race. With the exception of the FTA instrument, where some differences are observed, the predictive efficacy as well as the errors made by the instrument is fairly consistent across different races.

It should be noted that the error differences found in the FTA tool are small compared to the egregious biases that have been reported elsewhere and that are at the heart of the concern in the field. Figure 3.9, for example, shows what ProPublic's analysis concluded about the implementation of COMPASS in Broward County Florida.[2] In comparing the FPR, they found a 20 percentage point gap in favor of Whites and, similarly when comparing FNR, they found a 20 percentage point gap in favor of Whites. That is, Blacks were two times more likely than Whites to be *falsely* identified as high risk while Whites were two times more likely than Blacks to be *falsely* identified as low risk. The analysis of PSA's data suggests that the FTA tool might be falsely identifying Blacks as high risk at slightly higher rates than Whites (70% compared to 60%). However, the same tool also falsely identifies Blacks as low risk at slightly higher rates than white (12% compared to 8%). Moreover, these small differences are not found in the safety related scores and measures.

[2] Their analysis is not without controversy, though. For example, Flores, Lowenkamp, and Bechtel (2017) have criticized the ProPublica analysis as being flawed. See `http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf`.



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Figure 3.9: Screen shot from ProPublica's Article Describing Predictive Bias Findings.

3.4.2 Recommendations

While the purpose of this analysis was to assess the predictive bias in the current deployment of the RAI used by PSA, the analysis presented here should be replicated for the revised instruments. Moreover, it should be a part of any future re-validation efforts that PSA might undertake.

Chapter 4

Revised Instrument

4.1 Introduction

Building on the analysis presented in the previous chapter, this chapter presents a revised risk assessment instrument that addresses some of the issues identified with the currently deployed instrument. More specifically, the revised instrument is a slightly simplified model (several problematic features are dropped while some are added) and the methods used to weight the features are different.

4.2 Data and Methods

The data used for developing the revised models are the same as describe in previous chapters. However, instead of using the overall number of prior arrests and convictions, the criminal history measures are revised.

4.2.1 Revised Features

Table 4.1 summarizes the changes that are proposed for the revised RAI. The same domains are covered by the revised RAI as in the current deployment—criminal history, current charge, criminal justice status, lockup drug test, and demographic/social indicators. Of these, there is no revision to the current charge and criminal justice status domains. Within the lockup drug test domain, synthetic drug test results have been added and the overall drug test

Table 4.1: Features (43) proposed for the revised instrument.

Criminal History (11)
of Internal Convictions (Felony Charges) within last 10 years
of Internal Convictions (Misd. Charges) within last 10 years
of Internal Convictions (Felony Charges) more than 10 years ago
of Internal Convictions (Misd. Charges) more than 10 years ago
of External Conviction within last 10 years
of External Convictions more than 10 years ago
Lambda Internal (# of Internal Arrests / Current Age)
Lambda External (# of External Arrests / Current Age)
Prior Bench Warrants
Juvenile Arrests
Age at First Arrest
Current Charge (14 - same as before)
Criminal Justice Status (3 - same as before)
Lockup Drug Tests (5 - Synthetic Drugs added)
Demographic/Social (10 - past SSU assignment added)

compliance feature has been dropped. Within the demographic/social indicators domain, prior assignment to specialized supervision unit has been added as an available, objective measure of mental health problems.

The most drastic changes are in the criminal history domain. All features that included charge specific arrest and conviction histories have been dropped. Instead, the criminal history domain includes four features based on internal conviction histories—(i) Felony within the last 10 years, (ii) Misdemeanor within the last 10 years, (iii) Felony more than 10 years old and (iv) misdemeanor convictions more than 10 years old. In a similar manner, the convictions from the external criminal history files are also broken in to the more recent (last 10 years) and the distant past (more than 10 years ago). However, because the robustness of the felony and misdemeanor distinction is questionable in the external data, the measures do not make that distinction. Because the two sources of criminal history (internal and external) provide two qualitatively different sources of prior involvement in crime, the lambda measures are also defined differently for the internal and external sources. However, the lambda measures are defined using all arrests (not just convictions). That way, any additional information that arrest data may contain (over and above the convictions) will be included in the models. Finally, prior bench warrants, juvenile arrests and age at first arrest are included like in the current deployment.

A detailed list of each of the underlying features, their definitions, as well as a comparison of the old and new features are provided in Appendix B

4.2.2 Weighting Method

Similar to the committee models, the first step in the revised model is to convert each attribute into a categorical variable with a set of mutually exclusive categories. Feature-specific scores are then computed based on the misconduct data–again just like in the committee models. Finally, these feature specific scores are combined into a single misconduct specific score. The weighting method used to combine these scores is different from the committee models. Rather than use principal component analysis to combine these scores, the revised models use a tradition logistic regression model to estimate coefficients that will combine the individual scores into a single measure. This measure is then re-normed to get a final score ranging from 0 and 100 (similar to the currently deployed RAI).

4.3 Findings

Similar to the comparisons presented in the previous two chapters, this section discusses the predictive efficacy and predictive bias of the revised models.

4.3.1 Predictive Efficacy

Table 4.2 shows the AUC statistics for the four outcomes—any re-arrest, dangerous/violent re-arrest, FTA, and domestic violence re-arrest—using the revised model.[1] The AUC statistic are much improved now, compared to the current deployment. In fact, the revised models for the ARR and FTA outcomes have AUC statistics above the 0.7 threshold that is considered good within the criminal justice field. While the DVD and DVO model do not

[1] The AUC statistics reported here were computed for a random 20% test sample while the remaining 80% was used to train the model. To compute the final weights, all 100% of the available data was used.

Table 4.2: Predictive efficacy of the revised instrument.

	AUC Stat.*	95% Low	95% Hi
Any Re-arrest	0.74	0.73	0.75
Dang/Viol Re-arrest	0.69	0.66	0.72
FTA	0.71	0.69	0.72
DomViol Re-arrest	0.65	0.58	0.71

* Computed on a 20% test sample.

make that threshold, there is a substantial increase over the currently deployed model. Note that the AUC statistic for these same outcomes using the re-weighted version of the old model (Table 2.1) had AUC scores in the 0.63 to 0.67 range. Hence, simply re-weighting the older models would not yield gains in predictive efficacy. In fact, the revised models with the new criminal history features are probably rendering the models more predictive.

4.3.2 Predictive Bias

As noted in the last chapter, while the current deployment seems fairly robust to predictive bias, it is important to check the revised models to ensure that desirable property still exists. Figures 4.1 to 4.4 show the distribution of the estimated risk scores in the sample at hand. These distributions can be compared with those presented in figures 3.1 through 3.4. Even though the data, models, and methods are different, the scores are on a common scale (0 to 100).

The distributions of the estimated revised scores show a couple of things. First, the distributions are all unimodal. While the distribution for the revised ARR score for Blacks is shifted to the right of the Whites and Hispanics, it only has one mode. This is different from the current score–which has a bimodal distribution. The remaining three scores—DVD, FTA and DVO—show similar unimodal distributions and show some differences between the races.

Figures 4.5 through 4.8 show the relationship between the revised scores and misconduct. As with figures 3.5 through 3.8. the scores are fist grouped into 20 quantiles and then misconduct rates are computed within these groups. The plots show the estimated misconduct rate versus the average

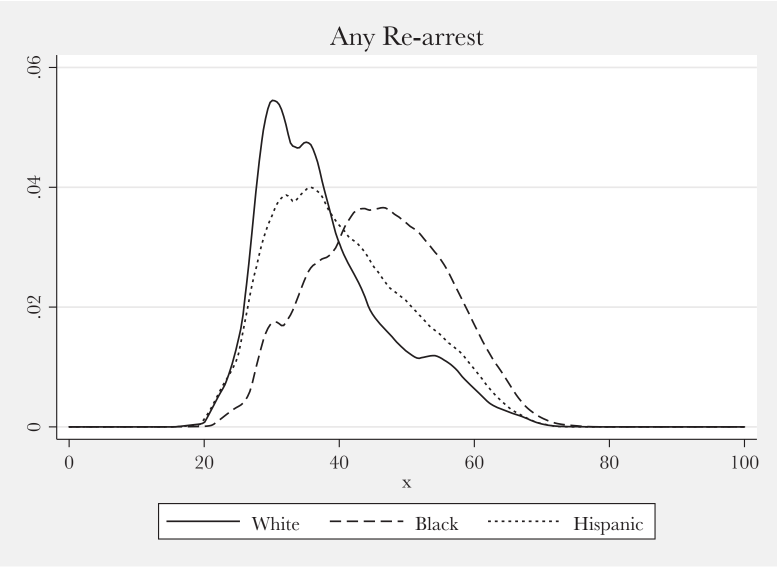


Figure 4.1: Distribution of Rearrest Risk Score, by Race.

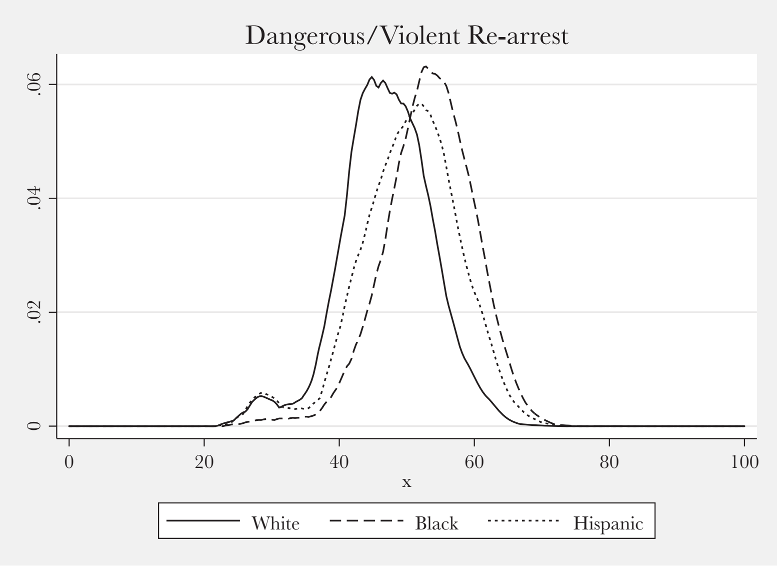


Figure 4.2: Distribution of Dangerous/Violent Rearrest Risk Score, by Race.

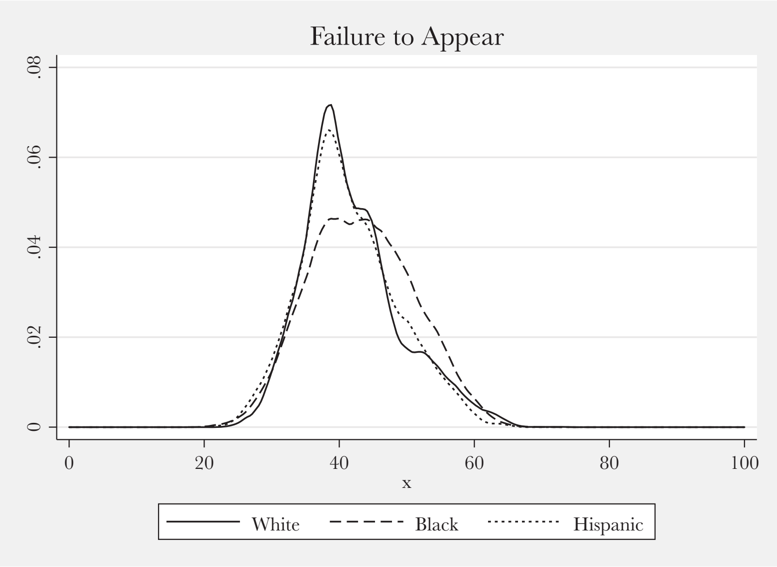


Figure 4.3: Distribution of Failure to Appear Risk Score, by Race.

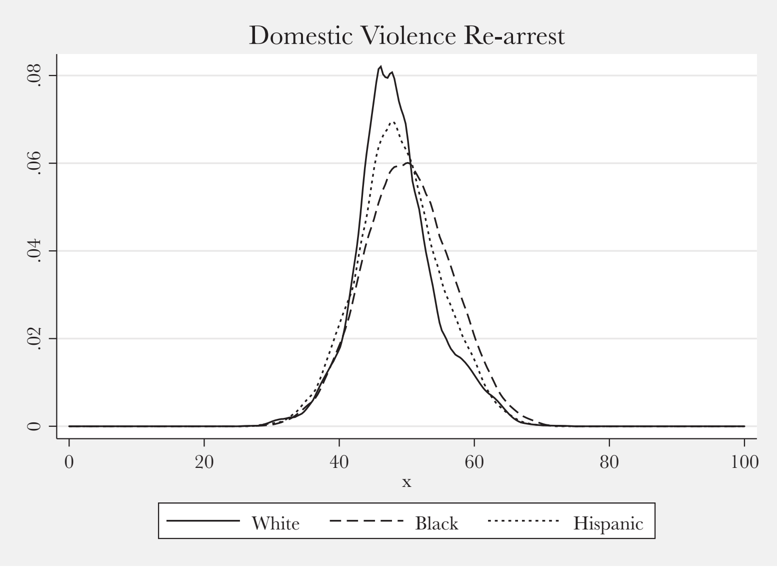


Figure 4.4: Distribution of Domestic Violence Rearrest Risk Score Among DVM Cases, by Race.

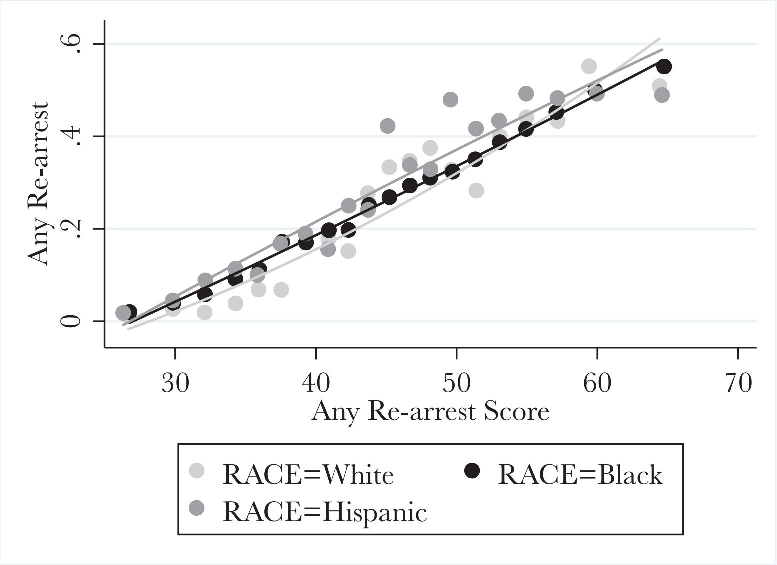


Figure 4.5: Rearrest Rate for 20 Quantiles of Rearrest Risk Scores, by Race.

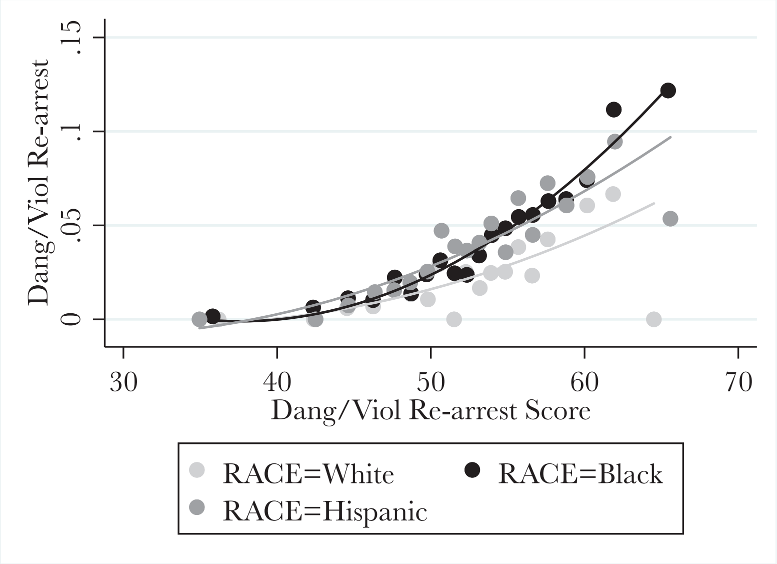


Figure 4.6: Dangerous/Violent Rearrest Rate for 20 Quantiles of Dangerous/Violent Rearrest Risk Scores, by Race.

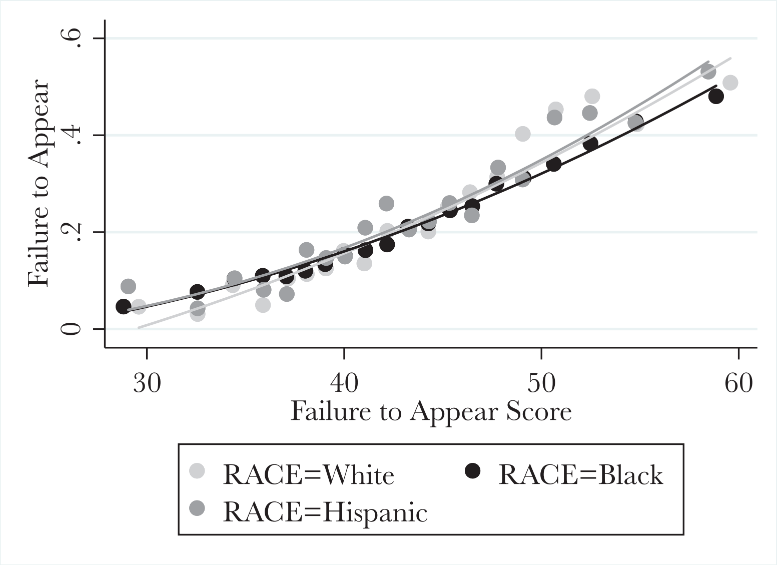


Figure 4.7: FTA Rate for 20 Quantiles of FTA Risk Scores, by Race.

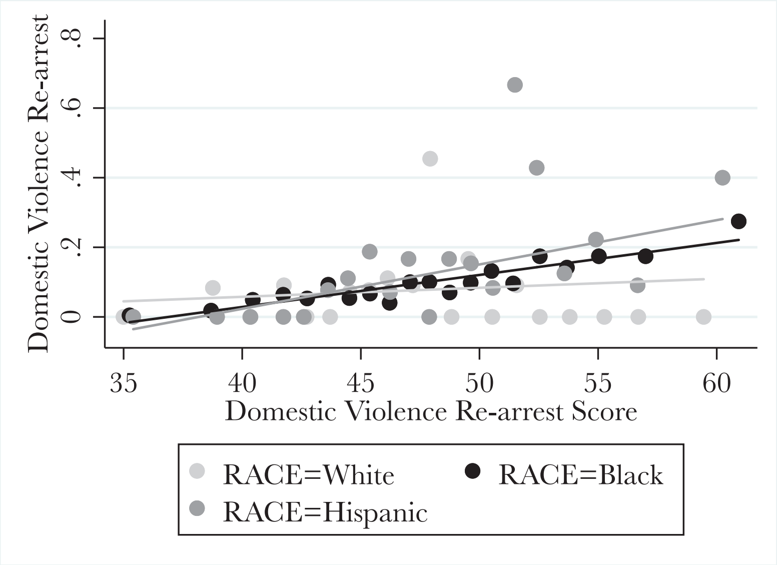


Figure 4.8: Domestic Violence Rearrest Rate for 20 Quantiles of Domestic Violence Rearrest Risk Scores (among DVM Cases), by Race.

Table 4.3: Percent distribution in revised risk score based categories and misconduct rates, by race.

	% in Risk Category			Misconduct Rate		
	White	Black	Hispanic	White	Black	Hispanic
Any Rearrest						
Low	66.7%	29.1%	53.2%	4.2%	9.9%	9.1%
Med	22.3%	42.4%	31.6%	26.7%	27.4%	30.9%
High	9.9%	25.5%	14.0%	45.1%	44.4%	47.6%
Vhigh	1.1%	3.0%	1.2%	–	58.8%	–
Dangerous/Violent Rearrest						
Low	73.1%	34.5%	51.3%	0.7%	1.5%	1.2%
Med	24.5%	50.7%	40.0%	2.5%	4.8%	5.2%
High	2.3%	13.1%	8.0%	6.1%	10.1%	6.3%
Vhigh	0.1%	1.7%	0.7%	–	13.8%	–
Failure to Appear						
Low	39.8%	31.3%	40.4%	8.2%	9.7%	9.7%
Med	42.4%	40.6%	41.2%	19.6%	20.8%	21.9%
High	17.0%	27.6%	18.1%	45.1%	38.2%	40.9%
Vhigh	0.9%	0.5%	0.3%	–	50.3%	–
Domestic Violence Rearrest						
Low	32.6%	24.4%	29.4%	4.6%	5.1%	4.2%
Med	47.3%	41.7%	45.5%	12.0%	10.5%	15.3%
High	11.2%	20.3%	15.8%	–	20.5%	–
Vhigh	2.7%	4.8%	2.5%	–	22.2%	–

– ($N < 30$)

Table 4.4: False Discovery and False Omission rates using revised risk score based categories, by race.

	False Discovery Rate*			False Omission Rate**		
	White	Black	Hispanic	White	Black	Hispanic
Any Rearrest	53.9%	54.1%	52.0%	4.2%	9.9%	9.1%
Dang/Viol Rear.	94.2%	89.4%	93.2%	–	1.5%	–
Failure to Appear	54.7%	61.6%	59.1%	8.2%	9.7%	9.7%
Dom Viol Rear.	–	79.3%	–	–	5.1%	–

* Very high + high together are predicted to fail
** low are predicted to not fail
– (N<30)

scores. The plots, with one exception, show remarkable similarity in the relationships for each of the races. The only exception is the domestic violence model (figure 4.8). Because of the sparseness of the data, it is possible that the model is unstable.

Consistent with the graphical analysis, Table 4.3 shows that the distribution of misconduct rates are fairly uniform by race. While the risk scores may be distributed differently by race—e.g., only about 9.9% of Whites are at high risk of any re-arrest but nearly 25.5% of Blacks are at high risk of any re-arrest—the observed misconduct rates among the different groups are very similar (about 44.4% for Blacks versus 45.1% for Whites). Indeed, with very few exceptions, the misconduct rates for each of the racial groups are within 6%-7% of one another. This is in contrast to Table 3.3 in Chapter 3 where the differences were more pronounced.

Finally, Table 4.4 shows a comparison of the errors (false discovery and false omission rates), by race, for each of the outcomes. This tables shows very clearly that the revised models produce very similar errors, by race. If we use the High and Very High risk categories to identify failures, then the models are equally erroneous for each of the races for all of the outcomes. The difference in the FDR between Blacks and Whites is highest for the FTA model (6.9%) and the difference in the FOR between Blacks and Whites is highest for the any rearrest model (5.7%). Unlike the current RAI, where we had found double digit differences in the error rates between Blacks and White FTA models (see Table 3.5 in Chapter 3), here the large biases are not observed. While no risk assessment model is perfect, the errors made by this revised model is largely unbiased with respect to race. Moreover, it retains this property while producing AUC scores above 0.70.

4.4 Conclusion

A more detailed discussion of predictive bias analysis for the revised and re-validated instrument is presented in a separate report.[2]

4.4.1 Summary of Findings

The estimated scores from the revised models developed have several desirable properties:

1. The revised model is simpler (only 43 features as opposed to 70);
2. The revised model is slightly more accurate; and
3. The revised model improves the racial parity of the instrument. While there were very little observed differences in the classification error rates by the current RAI, the revised model eliminates them where they were more pronounced (e.g., FTA instrument false positive rates and the Any Re-arrest instrument false negative rates).

4.4.2 Recommendations

Based on the findings presented in this chapter, the revised models appear to be robust and should be adopted in the next deployment.

The only exception is the domestic violence model. Because PSA does not utilize the domestic violence scores to make decisions with regards to the domestic violence population (that it is designed for), PSA might consider dropping this model altogether. Its predictive efficacy is much lower than the other models and the graphical analysis suggests that the predicted scores may not be completely devoid of racial bias.

[2] Bhati, Avinash (2019) *Pretrial Supervision Agency for the District of Columbia, Risk Assessment Instrument Re-validation Project: Predictive Bias Report.* Report submitted to the Pretrial Services Agency for the District of Columbia.

Chapter 5

Risk-based Supervision

5.1 Introduction

While PSA has been using a validated risk assessment instrument for several years, it has used it primarily to make release recommendations to the courts. PSA's supervision activities are organized primarily based on court ordered conditions of release. In recent year, PSA has sought to leverage its risk assessment efforts and to utilize risk information in supervising defendants. In 2015, the Office of Strategic Development at PSA conducted a "Red Zone Project" identifying pretrial release phases where defendants might be at higher or lower risk (safety or flight). Building on that effort the Office of Operations at PSA is developing a series of risk-based case management strategies that are designed to align case management with defendant risk. This risk-based case management model will utilize a series of graduated sanctions and incentives to respond to defendant conduct under pretrial supervision.

As part of the RAI re-validation project, PSA was interested in analyzing its administrative data to provide guidance to the Office of Operation in developing, analyzing, and assessing the efficacy of such risk-based case management strategies. This chapter describes the analysis conducted to support PSA in that effort.

5.2 Data and Methods

5.2.1 Data Used

While the main data source described in previous chapters was also used in this analysis, additional data from PRISM was accessed and used. Defendants released on pretrial release between Oct 2014 and Oct 2017 continues to be the core cohort of interest.

In addition to static information about these clients, a series of dynamic data about these clients were also accessed and used. Two main data sources were accessed—(i) Response to defendant conduct and (ii) Supervision Logs.

The Supervision Logs store ad-hoc, semi-structured data entered into PRISM by supervision officers. While the data is rich, it is not easily accessible nor does it contain the needed information to analyze supervision strategies.

The Response to Defendant Conduct data, on the other hand, contains detailed data on defendant infractions (conduct) and the agency's recorded response to that conduct. These data were detailed enough to permit a dynamic analysis of the different pathways that defendants may follow while under supervision. Specifically, the data could be structured into a sequence of distinct conduct-events, for each defendant, that permitted the development of detailed Markov models. Markov models are appropriate to use when one is interested in studying the dynamics—short-, medium-, or long-term implications of choices on outcome of interest. In the current project, this meant the ability to study how various choices made by PSA in responding to defendant infractions might steer defendants along different pathways—leading either to a successful termination of pretrial supervision or an unsuccessful pretrial supervision (with either a failure to appear or a re-arrest event occurring while the defendant was still under pretrial supervision).

5.2.2 Markov Models

Markov Chains

The basic analytical framework relies on the concept of Markov chains. Markov chains are a very powerful way to study the process by which a system evolves over time. Markov chains are used to study the transitions of units (persons, cases, etc.) through a sequence of stages (or 'states' or 'events') over time. The time interval over which we study this process may be discrete—e.g., states recorded at the end of every month—or it may be continuous—e.g., pretrial infraction events that can occur anytime while under pretrial release.

The basic model is build by defining a (categorical) variable Y that identifies the specific state a unit is in at a particular time. The 'order' of the Markov chain is determined by the number of prior events that need to be recorded for developing a Markov chain transition matrix. For example, a first order Markov model would require the current event (Y_t) as well as the previous event (Y_{t-1}). Taken together, these two (the current and previous event) are used to define the transitions as:

$$\begin{array}{ccccc} & Y_{1,t} & Y_{2,t} & \dots & Y_{J,t} \\ Y_{1,t-1} & p_{11} & p_{12} & \cdots & p_{1J} \\ Y_{2,t-1} & p_{21} & p_{22} & \cdots & p_{2J} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_{J,t-1} & p_{J1} & p_{J2} & \cdots & p_{JJ} \end{array} \tag{5.1}$$

where $Y_{1,t}, \dots, Y_{J,t}$ reflect the current events, $Y_{1,t-1}, \dots, Y_{J,t-1}$ reflect the last state, and $p_{jj'}$ reflects the transition probability from state j to state j'. The transition probabilities are in fact conditional probabilities—i.e., $p_{jj'} = \Pr(Y_{j',t}|Y_{j,t-1})$ is the probability of transitioning to state j' from state j. In the case of pretrial supervision and defendant conduct, this means a transition probability of having infraction type j' after infraction type j.

Two critical assumptions about first order Markov chains include:

- Mutually exclusive and exhaustive states: A unit can only occupy one of a finite number of mutually exclusive states at any time. This assumption implies that the probabilities must sum to 1 within each row of the transition matrix. In other words, the unit must move from the last period state to one or another new state or remain in the same state.

- Time-independence (stationarity): The second assumption states that the probability of occupying any state depends only on the states occupied in the last time period (and not any period before that). Of course, one can define higher order Markov processes that depend on more than just the past period. Higher order Markov processes may be more realistic in many setting but they become more analytically intractable with increasing order.

One of the implications of limiting the order of the Markov process to just 1 (stationarity) is the ability to project probabilities into the future by taking powers of the transition matrix. As an example, if we consider a 2×2 transition matrix

$$P = \begin{array}{cc} & \begin{array}{cc} \mathrm{A} & \mathrm{B} \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{bmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{bmatrix} \end{array}$$

which satisfies the two properties above. Each of the rows sum to 1 and the states are mutually exclusive and exhaustive. The probabilities reflect the transition from state A to B and back over time. If a unit was in state A, then the probability of this unit being in state A next period is 0.3 and the probability that it will move to state B is 0.7. Hence, it will either stay at A or move to B (exhaustive states). Similarly, if the unit was previously in state B, the probability that it will move to state A next period is 0.2, and the probability that it will stay in state B is 0.8.

Now, given that this is a first order stationary process, we can also estimate the probability of the unit occupying state A or B two periods out by simple matrix multiplication.

$$P^2 = P \times P = \begin{bmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{bmatrix} \times \begin{bmatrix} 0.3 & 0.7 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.23 & 0.77 \\ 0.22 & 0.78 \end{bmatrix}$$

so that the probability that a unit in state A at time t will still be in state A at time $t + 2$ is 0.23 and the probability that it will have moved to state B at time $t + 2$ is 0.77. Similarly for three periods out, we get

$$P^3 = P \times P \times P = \begin{bmatrix} 0.223 & 0.777 \\ 0.222 & 0.778 \end{bmatrix}$$

As can be seen, as time progresses, the probabilities of occupying state A or B converge to a unique vector *irrespective* of the lagged state (approximately 0.22 for occupying state A and 0.77 of occupying state B). This is the stationarity property. Hence, given a stationary process, one can study the evolution of the system in the short, medium or the long run.

In the example shown above, all transitions had non-zero entries. This is not required. In larger transition matrices it is entirely possible to have some of the entries be zero and some be 1. Since the entries measure probabilities, they cannot be less than 0 or more than 1. Having some zeros in the Markov process allows one to distinguish between transient and recurring states. In the above example, all states were recurrent because a unit could go from A to B and return to A or B. However, had one of the entries been 0, this would imply that the other entry in that row would have to be 1. For example

$$\begin{bmatrix} 0.3 & 0.7 \\ 0 & 1 \end{bmatrix}$$

implies that, if a unit was previously in state A, then the probability of it staying in state A next period is 0.3 and the probability of it moving to state B is 0.7. However, if the unit was in state B, it will never go to state A *and* will forever remain in state B. State B is referred to as an *absorbing* state while state A is referred to as a *transient* state. Elaboration on these two features (transient and absorbing states) are needed to accurately model any criminal justice system.

Absorbing Markov Chains

In the current analysis, the key absorbing states of interest are whether a defendant has an FTA or is re-arrested or has his/her case disposed. Once either of these events occurs, we consider the process to be completed. While, in reality, PSA will continue to supervise clients even after they have an FTA, the Markov models treat these misconduct events as *absorbing* states because that is the behavior PSA would like to discourage. Treating them as absorbing states in the Markov models is useful because it allows us to model and analyze different strategies for guiding defendants towards pathways that increase the probability of reaching a successful absorption state (case disposition) instead of an unsuccessful one (FTA or re-arrest).

Absorbing Markov chains have some additional properties that are extremely useful in studying the dynamics of the system. Consider a more complicated transition matrix with a few absorbing states and several transient states. Here transient states are like defendant infractions and absorbing states are like FTA, re-arrest, or case disposition.

$$\begin{array}{cccccc}
 & A & B & C & D & E \\
A & 0.6 & 0.1 & 0.1 & 0.1 & 0.1 \\
B & 0 & 0.7 & 0.2 & 0 & 0.1 \\
C & 0 & 0 & 0.9 & 0.1 & 0 \\
D & 0 & 0 & 0 & 1 & 0 \\
E & 0 & 0 & 0 & 0 & 1
\end{array}$$

This Markov process has 5 states (A through E) of which A, B and C are recurring states (clients can have these conducts multiple times) and D and E are absorbing states (units stay in these states once they arrive there). This matrix can be written in, what is referred to as, its canonical form

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$$

where each of the components are

$$Q = \begin{bmatrix} 0.6 & 0.1 & 0.1 \\ 0 & 0.7 & 0.2 \\ 0 & 0 & 0.9 \end{bmatrix} \qquad R = \begin{bmatrix} 0.1 & 0.1 \\ 0 & 0.1 \\ 0.1 & 0 \end{bmatrix}$$

Q contains transition probabilities between the transient states in the top left of the matrix (states A/B/C to A/B/C), R consists of all transition probabilities from the transient states to the absorbing states in the top right of the matrix (states A/B/C to states D/E), 0 is just a matrix of zeros in the lower left of the matrix (impossible transitions from D/E to A/B/C) and I is the identity matrix of the absorbing states at the lower right of the matrix (mandatory transitions from D to D and E to E).

The short- and medium-term evolution of this process can be studied as explained above, by taking higher powers of the matrix. These higher powers will move the system slowly (or rapidly) towards the absorbing states. Instead of a long-term equilibrium vector, as was seen in the 2×2 non-absorbing example explained previously, taking higher powers of the absorbing Markov chain moves the system towards the absorbing states. In other words, rather than settle on an equilibrium transition vector, the system converges to the absorbing states (i.e., the unit is eventually absorbed). There are two extremely powerful properties of this canonical presentation that can help us derive very useful properties of the process under study.

1. Expected number of 'visits': The *fundamental* matrix, as it is called, gives the expected number of times a unit is expected to visit each of the transient states, given that a unit just entered one of these states. It is defined as
$$F = (I - Q)^{-1}$$
The identity matrix in this case has the dimensionality of the Q matrix. This matrix is termed the fundamental matrix and is a long-run approximation of the number of times a unit can be expected to 'hit' a particular transient state before eventually getting absorbed (in one or another absorbing state). Because Q is a 3×3 matrix, F has the same dimensionality. This means F gives an estimate of number of times a unit can expect to visit each of the 3 transient states (before being absorbed), given that it started from each of the transient states. If the transient states are infractions, then this gives us an estimate of the number of infractions prior to eventual absorption.

2. Eventual Absorption Probabilities: The product of the fundamental matrix and the R matrix yields an estimate of the probability of eventually being absorbed in each of the absorbing states, given that the unit started from one of the three transient states.
$$\Pr(\text{Eventual Absorption} \mid \text{Starting State}) = F \times R = (I - Q)^{-1} \times R$$

These two properties form the basis on which we can analyze long term implications of defendant conduct.

Markov Decision Process

The final component of the model needed to study implication of PSA's choice on defendant outcome is a decomposition of the transition matrix. In transition from one infraction to the next, there are two distinct components—(a) the agency's response to defendant conduct and (b) the defendants response to the agency's action. In other words, when we observe a defendant who has an infraction, the agency records some response to that and the defendant's next conduct reflects his/her response to the reaction. Put differently, the transition from state j to j' really has two sub-components to it–the probability that the agency will respond to the conduct j by action a and the probability that the defendant will respond to this action by response d. Hence,

$$P = A \times D$$

where A reflect choices made by the agency and D reflects choices made by the defendant.

The working assumption for Markov Decision Processes is that D is in the control of the defendant and cannot be changed. However, the agency can modify A and, in doing so, alter P. The ability to modify P means that the agency is able to modify the fundamental matrix F and therefore the Pr(Eventual Absorption | Starting State). These quantities, described in the previous section, are the values of interest. In formulating different policies, the agency is modifying A in such a way as to maximize the chance that defendants will exit supervision successfully (and minimize the chance that they will exit unsuccessfully).

Once the data are structured correctly, the computations of each of these matrices is straight-forward and they may be computed for various subpopulations of interest. Each of the computations produces two critical estimates of interest—(i) the estimated probability of eventual absorption into the various states (e.g., FTA, re-arrest, or case disposition) and (ii) the estimated number of infractions before that eventual absorption. The findings discussed in the remainder of this chapter make exclusive use of these estimates.

5.3 Findings

This section discusses the findings from estimating and assessing the implications of the Markov model on DC's pretrial population.

Defendant conduct categories include infractions relating to Contact, Drug Testing, Electronic Supervision, Group Sessions, and Other. Appendix D provides a detail list of the underlying conducts that constitute each of these categories. In addition to specific infraction types, where pertinent or relevant, these categories are further refined. For example, as part of their planned graduated supervision strategy, PSA has designed different responses for upto the fifth contact or drug testing infractions. The analysis is conducted by defining the first, second, third, fourth or fifth (or higher) contact infractions as distinct states. This allows the models to provide estimates of probability of absorption starting from each distinct contact infraction.

Agency responses to these conducts include three types:

1. Active with Client Contact—these include responses like "Verbal Warning", "Written Warning", or "Referral to MH treatment" where the agency's response directly involves the client;
2. Active without Client Contact—these include responses like "Request for removal from Program" or "Recommend Judicial Review" where the agency responds indirectly to the conduct (without client contact); and
3. Passive—these include responses like "Invalid EM alert" or "No Response Required" where the agency's response does not involve the client at all.

Appendix E provides a detailed list of all responses included in the analysis and how they are categorized.

Finally, PSA was interested in assessing whether the dynamics were different for different program types (teams or queue names). These include Substance Use Disorder (SUD), General Supervision, Specialized Supervision Unit (SSU), and HISP. A detailed list of specific programs that are grouped within each of these categories is provided in Appendix C

5.3.1 Risk Dynamics

Table 5.1 shows the basic findings from the Markov models, applied to the full sample as well as to various risk categories. The last row in this table shows the average estimates from the model applied to the full sample. There are a total of 45,846 defendants included in the sample and they have a combined total of 229,319 events in the data. Events include entering pretrial supervision, a series of infraction, and exit (absorption) events.

There is wide variation in the type and number of infractions that individuals commit but, on average, sample members have 3 infractions before being absorbed (i.e., either have a re-arrest and FTA or a case disposition event). Nearly 64% of these clients will exit supervision without an FTA or re-arrest event. About 17.2% of them will have a re-arrest event prior to case disposition and another 18.8% will have an FTA.[1]

[1] Note, that the 17.2% and 18.8% reflect the first of these two events. Some of the 18.8% who first have an FTA could eventually have a re-arrest event as well.

Table 5.1: Probability of exit from Pretrial status, by risk level and exit type.

				Probability of exit by ...		
	# Events	# Cases	# Infr.	Rearrest	FTA	CaseDisp
Low Risk	20,602	6,142	1.35	6.9%	9.7%	83.3%
Moderate Risk	87,700	19,093	2.59	14.2%	16.9%	68.9%
High Risk	73,442	12,626	3.82	21.1%	22.3%	56.6%
Very High Risk	47,575	7,985	3.96	24.5%	23.6%	51.8%
Full Sample	229,319	45,846	3.00	17.2%	18.8%	64.0%

The upper panel of Table 5.1 shows the same computations for different risk groups within the sample. For example, there were 19,093 clients classified as Moderate Risk and they had a total of 87,700 events among them. On average they had 2.6 infraction per client and they had a slightly higher than average chance of completing their supervision successfully (69%). They had about 14% chance of re-arrest and about 17% chance of an FTA.

As one would expect, the models show that lower risk clients typically have lower number of infractions, have lower re-arrest or FTA (misconduct) probabilities and higher chances of a successful completion of pretrial supervision. The lowest risk clients had a 83% chance of successfully completing supervision while the Very High risk clients only had a 51.8% chance of successful supervision completion.

Table 5.2 shows the same calculations computed by program type. The largest number of clients were supervised under General Supervision (31,887) and they accounted for a total of 108,956 events. On average, they had only about 1.3 infractions prior to either a misconduct or successful completion of pretrial supervision. These clients had a 65% chance of completing pretrial supervision successfully but nearly 16% chance of being re-arrested or 18.6% chance of an FTA.

Substance Abuse Disorder (SUD) clients had a similar probability of successfully completing pretrial supervision (65%). However, they were expected to accumulate many more infraction (5.45 on average). They were also more likely to have an FTA (21%) but slightly less like to be re-arrested (14.4%) compared to the General Supervision clients.

Specialized Supervision Unit (SSU) clients (those assessed in need of mental health problems) were expected to successfully complete pretrial supervi-

Table 5.2: Probability of exit from Pretrial status, by program type and exit type.

				Probability of exit by ...		
	# Events	# Cases	# Infr.	Rearrest	FTA	CaseDisp
SUD	11,744	1,300	5.45	14.4%	21.1%	64.5%
General Supervision	108,956	31,887	2.28	16.0%	18.6%	65.4%
SSU	40,929	8,227	3.83	26.4%	28.8%	44.9%
HISP	43,917	4,820	10.22	22.3%	18.1%	59.5%
Traffic Safety	10,959	3,807	1.30	7.3%	10.4%	82.3%
Clients with MH Prob*	53,759	8,998	3.97	24.2%	25.3%	50.5%
... w/ MH Prob*not in SSU	28,956	6,375	3.97	22.3%	23.9%	53.8%
... in SSU w/o MH Prob*	16,126	3,436	3.48	26.0%	31.3%	42.7%

* Self identified with emotional problems.

sion with the lowest rates (45%) and had the highest misconduct rates (26.4% rearrest and 28.8% FTA). Notably, though, this was despite the fact that they only accumulated 3.8 infractions per client—much lower than the average infraction accumulated by the SUD clients (5.5). HISP clients, though successfully completing pretrial supervision at higher rates than the SSU clients (60% compared to 45%), however, accumulated over 10 infraction per client.[2]

PSA executives were keen to assess if the SSU clients were different from those that self identify themselves as having MH (emotional) problems. The lower panel in Table 5.2 shows these estimates. When compared with the overall sample, this group does have elevated risk of re-arrest (24%) and FTA (25%) and lowered chance of successfully completing pretrial supervision (50%). However, the SSU clients are an even higher risk group than those self identifying themselves as having MH problems. In all likelihood this is because of the formal diagnosis that is conducted before a client enters SSU while the self-reported MH needs may be erroneous. Indeed, a surprising finding here is that the group of clients who are in SSU but who self-report themselves as *not* having MH issues (last row in Table 5.2) are at the highest risk of misconduct and lowest risk of successful completion of pretrial supervision. Only about 42% of them are expected to complete pretrial supervision successfully.

The analysis presented in tables 5.1 and 5.2 shows that risk and program type matters and that the SSU clients appear to be at a particularly high risk of unsuccessful pretrial supervision. In order to take a closer look at these two dimensions, Table 5.3 combines these two and shows estimates by program type and risk level.

The directional relationship between risk level and infractions, misconduct, and successful completion of pretrial supervision holds within most of the program types. That is, typically, Case Disposition rates are higher for lower risk groups and misconduct (re-arrest and FTA) rates are higher for the higher risk groups. Also, typically, the higher risk groups have more infractions than the lower risk groups. The exception to this rule appears to be the SSU clients. Within this group, the FTA rates follow and inverted "U" shape. For example, the FTA rates among the very lowest and the very highest risk groups are comparable (27% each). Indeed, the highest FTA rates appear to be the moderate and High risk categories within the SSU group. Similarly, the moderate and high risk groups among the SSU clients seem to accumulate

[2] This may be attributable to a large number of false EM alerts that the system triggers that are not necessarily infractions.

Table 5.3: Probability of exit from Pretrial status, by program type, risk level, and exit type.

				Probability of exit by ...		
	# Events	# Cases	# Infr.	Rearrest	FTA	CaseDisp
SUD						
Low	592	57	...	...	...	...
Mod	4,183	449	5.95	10.8%	15.7%	73.5%
High	4,320	484	5.95	14.6%	19.2%	66.1%
VHigh	2,649	310	4.32	17.1%	27.1%	55.8%
General Supervision						
Low	11,491	4,412	1.24	7.2%	10.0%	82.8%
Mod	46,290	13,921	2.02	13.8%	16.8%	69.4%
High	33,679	8,812	2.88	20.8%	23.2%	56.0%
VHigh	17,496	4,742	3.07	22.3%	24.1%	53.6%
SSU						
Low	1,523	329	3.18	13.0%	27.1%	59.8%
Mod	11,272	2,276	3.81	20.5%	30.6%	48.9%
High	14,343	2,741	4.12	27.2%	29.2%	43.6%
VHigh	13,791	2,881	3.63	31.4%	27.3%	41.3%
HISP						
Low	1,673	226	8.74	14.8%	13.4%	71.8%
Mod	14,945	1,870	8.74	20.3%	15.4%	64.3%
High	16,192	1,703	10.67	23.1%	18.9%	58.0%
VHigh	11,107	1,021	12.63	25.6%	22.4%	52.0%
Traffic Safety						
Low	2,690	1,137	0.60	3.84%	5.39%	90.77%
Mod	5,465	1,861	1.33	7.71%	10.01%	82.27%
High	2,151	587	2.71	12.21%	17.21%	70.58%
VHigh	653	222	2.30	12.47%	27.49%	60.04%

more (on average) infractions than the very highest risk group.

5.3.2 Policy Simulations

The Markov models were next used to develop a series of simulations to assess whether, and to what extent, modifying agency response to defendant conduct could increase the chances of defendants completing their pretrial supervision successfully. In other words, could they be guided into different pathways that would minimize their chances of misconduct. Table 5.4 shows these simulated estimates for all non SSU clients while Table 5.5 shows the simulated estimates for the SSU clients.

Two policy simulations were developed in this analysis. First, PSA's Office of Operation has developed a tentative case management strategy that involves a series of graduated responses to defendant conduct (infractions). A copy of the policy was obtained from Operation staff and its recommendations were converted into a simulated variant of the D matrix described above. If we assume that the A matrix remains stable (i.e., can be estimated from the data), then we can estimate a simulated transition matrix P. Recall that the overall transition matrix is a combination of how the agency responds to defendant conduct and then how the defendant acts based on the response. The simulations, therefore, provide a way to assess how overall misconduct rates, successful completion rates, and the estimated number of infractions might change should the agency pursue a graduated response policy different from what it is currently doing.

More specifically, with regards to contact violations, the proposed policy recommends that minimum risk clients get a verbal warning for their first/second contact infraction; a written warning for their third/fourth contact violation; get modified reporting requirements for their fifth contact violation and the court be notified for any subsequent contact infraction. Medium, High or Very High risk clients get a verbal warning for their first/second contact infraction; written warning for the third; modified reporting requirement for the fourth; and court notification for the fifth or high contact infraction. With respect to drug testing infractions, the proposed policy recommends that clients of all risk level get a verbal or written warning for the first/second drug testing infraction; referral to substance use assessment or drug treatment referral in response to their third or fourth drug test infraction; and that the court be notified in response to their fifth or subsequent drug test infraction.[3]

Second, a completely data driven (empirical) policy was developed by cherry picking only those responses that produced better than average risk-category specific estimated outcomes. This was done as a simulation to compute, what might be considered, a best case scenario. In other words, if the agency had retrospectively selected the best alternative and defendants had responded like they always do, then how different could the successful completion rates and misconduct rates have been.

These two policies are referred to in Tables 5.4 and 5.5 as the *Proposed Policy* and the *Data Driven Policy* respectively. The first row in each panel in these tables shows the current average (as is) while the next two show the simulated estimates. The last two rows show the difference between current practice and the simulated policies. These are the simulated effects of these policies. Negative effects are desirable under the # of infractions and misconduct (rearrest or FTA) columns while positive effects are desirable under the CaseDisp (successful completion of supervision) column.

Table 5.4 shows that the proposed policy could have an overall small impact on Non SSU clients (with an increase in successful completion rates of 1.86%). The drop in misconduct rate is mostly from FTA (-1.2%) and some from re-arrest (-0.65%). The proposed policy would have very little, if any, effect on the low risk clients. It would have some effect on the moderate, high or very high risk clients. On the other hand, the data driven policy suggests that using a completely empirical approach of pursing those alternative (be they active or passive) that show the most promise in the data, the overall success pretrial supervision completion rate could be increased by about 5%. Here too, the bigger effect would be on reducing FTA rates and some on reducing rearrest rate.

The same policy simulations conducted for the SSU clients show slightly different patterns. The estimates are presented in Table 5.5. For one thing, the overall gains using the empirical, data-driven policy are higher—there could be an increase in successful completion rates of about 8% with this population. More interestingly, the low risk clients stand to gain as much as the higher risk clients. In a manner paralleling the non SSU clients, the proposed policy has a more modest effect on FTA rates than on re-arrest rates. And, these

[3] While the proposed policy includes other elements—e.g., different types of contact proposed for different risk level and phase of supervision; incentives in response to compliance; or different types of drug testing infractions—data needed to simulate these more nuanced features of the policy were not available.

Table 5.4: Probability of exit from Pretrial status, by simulated policy, risk level, and exit type: Non SSU clients.

		Probability of exit by ...		
	# Infr.	Rearrest	FTA	CaseDisp
All Non SSU Clients				
Current Practice	3.051	16.31%	17.67%	66.02%
Proposed Policy	2.488	15.66%	16.47%	67.87%
Data Driven Policy	3.182	14.77%	13.94%	71.29%
Proposed Policy Effect	-0.563	-0.65%	-1.20%	1.86%
Data Driven Policy Effect	0.131	-1.54%	-3.73%	5.27%
Low Risk				
Current Practice	1.28	6.67%	8.97%	84.36%
Proposed Policy	1.07	6.18%	8.64%	85.18%
Data Driven Policy	1.19	5.69%	5.99%	88.32%
Proposed Policy Effect	-0.210	-0.50%	-0.33%	0.82%
Data Driven Policy Effect	-0.084	-0.98%	-2.98%	3.96%
Moderate Risk				
Current Practice	2.47	13.55%	15.37%	71.08%
Proposed Policy	2.14	12.79%	14.34%	72.86%
Data Driven Policy	2.44	11.96%	12.78%	75.26%
Proposed Policy Effect	-0.336	-0.75%	-1.03%	1.78%
Data Driven Policy Effect	-0.033	-1.59%	-2.59%	4.18%
High Risk				
Current Practice	3.75	19.83%	20.88%	59.29%
Proposed Policy	3.15	18.99%	19.46%	61.55%
Data Driven Policy	4.26	18.53%	15.75%	65.72%
Proposed Policy Effect	-0.599	-0.85%	-1.42%	2.26%
Data Driven Policy Effect	0.506	-1.30%	-5.13%	6.42%
Very High Risk				
Current Practice	4.14	21.83%	22.20%	55.97%
Proposed Policy	2.93	21.66%	20.47%	57.87%
Data Driven Policy	4.10	19.65%	17.91%	62.44%
Proposed Policy Effect	-1.212	-0.17%	-1.73%	1.90%
Data Driven Policy Effect	-0.033	-2.18%	-4.30%	6.48%

Table 5.5: Probability of exit from Pretrial status, by simulated policy, risk level, and exit type: SSU clients.

		Probability of exit by ...		
	# Infr.	Rearrest	FTA	CaseDisp
All SSU Clients				
Current Practice	3.836	26.24%	28.85%	44.91%
Proposed Policy	3.653	24.59%	27.31%	48.10%
Data Driven Policy	0.715	25.32%	21.83%	52.85%
Proposed Policy Effect	-0.184	-1.64%	-1.54%	3.19%
Data Driven Policy Effect	-3.122	-0.91%	-7.03%	7.94%
Low Risk				
Current Practice	3.18	13.04%	27.13%	59.83%
Proposed Policy	3.19	14.47%	25.68%	59.85%
Data Driven Policy	3.47	13.22%	17.49%	69.29%
Proposed Policy Effect	0.019	1.43%	-1.45%	0.02%
Data Driven Policy Effect	0.299	0.17%	-9.64%	9.47%
Moderate Risk				
Current Practice	3.81	20.45%	30.60%	48.95%
Proposed Policy	3.95	19.60%	27.28%	53.12%
Data Driven Policy	0.59	19.75%	25.75%	54.50%
Proposed Policy Effect	0.146	-0.85%	-3.32%	4.17%
Data Driven Policy Effect	-3.217	-0.70%	-4.85%	5.55%
High Risk				
Current Practice	4.12	27.18%	29.19%	43.63%
Proposed Policy	3.66	26.15%	27.68%	46.16%
Data Driven Policy	0.61	26.40%	21.12%	52.48%
Proposed Policy Effect	-0.468	-1.03%	-1.51%	2.54%
Data Driven Policy Effect	-3.513	-0.79%	-8.07%	8.86%
Very High Risk				
Current Practice	3.63	31.44%	27.27%	41.30%
Proposed Policy	3.45	28.16%	27.13%	44.71%
Data Driven Policy	0.62	30.10%	19.83%	50.07%
Proposed Policy Effect	-0.180	-3.27%	-0.14%	3.41%
Data Driven Policy Effect	-3.014	-1.34%	-7.43%	8.77%

effects are more heavily skewed towards the higher risk groups.

5.4 Response Types and Diversity

While the previous section provided simulated evidence that the proposed strategy could increase successful pretrial supervision completion rates, it also suggested that there are other strategies that might be pursued. This section presents some evidence to inform those other strategies.

The graphics presented here plot the estimated success probabilities for various strategies, by risk category and infraction type. Figure 5.1 shows the estimated success rates for non SSU clients for various responses types to contact infractions.[4]

Several observations are worth highlighting here. First, as is to be expected, the success probabilities are typically higher among the lower risk groups than among higher risk groups. Second, for the contact related infractions, the active responses with client contact (the blue bars) are usually the best option (i.e., have the highest chance of resulting in success). Third, for the first contact infraction, irrespective of risk level, the active responses without client contact and passive responses are equally effective (the red and green bars are about the same height). By about the third contact infraction, the passive responses become less and less appealing. By the fourth contact infraction, the passive responses are the lowest category in terms of success probabilities.

Figure 5.2 shows the same estimates for the drug testing infractions (among non SSU clients). Here the findings are very different. It is no longer clear if active responses are superior to passive responses. While it is still the case that success probabilities are higher among the lower risk clients, the height of the blue, red, and green bars are about the same for most categories. In other words, there is no evidence that active responses (with or without client contact) are any better than passive responses for drug testing related infractions among non Mental Health treatment clients.

Figure 5.3 shows the estimated success probabilities for responding to contact infractions by SSU clients. Here we find that the passive responses appear to be the best choice when responding to low-risk clients, in particular

[4] The responses are grouped into three categories—active with client contact, active without client contact, and passive. See Appendix E for a detailed list of responses classified within each category.

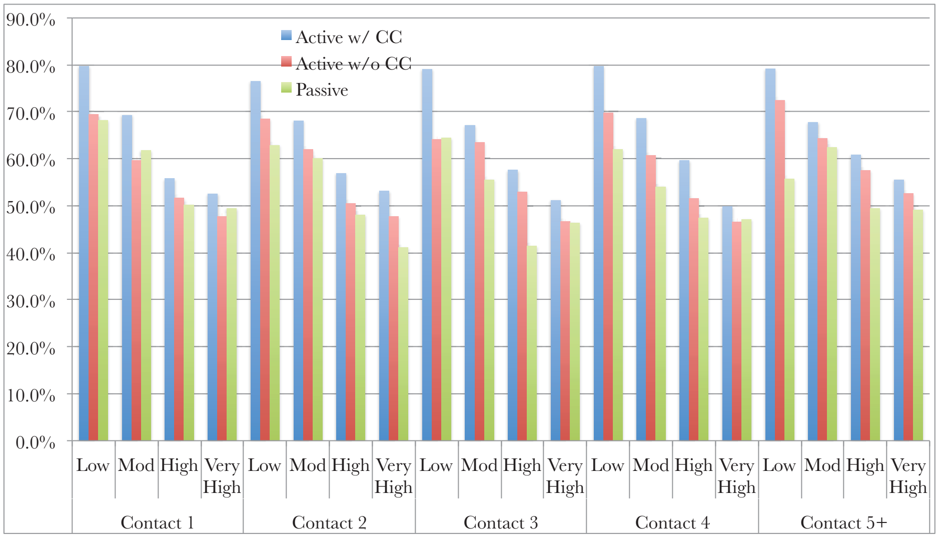


Figure 5.1: Probability of successful exit from pretrial supervision among non SSU clients with contact infractions, by agency response and risk level.

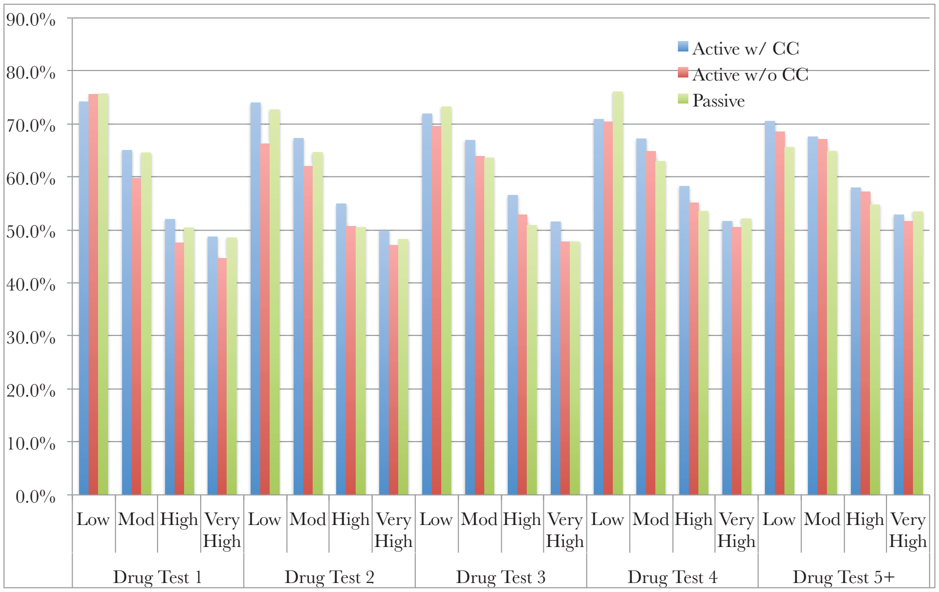


Figure 5.2: Probability of successful exit from pretrial supervision among non SSU clients with Drug Test infractions, by agency response and risk level.

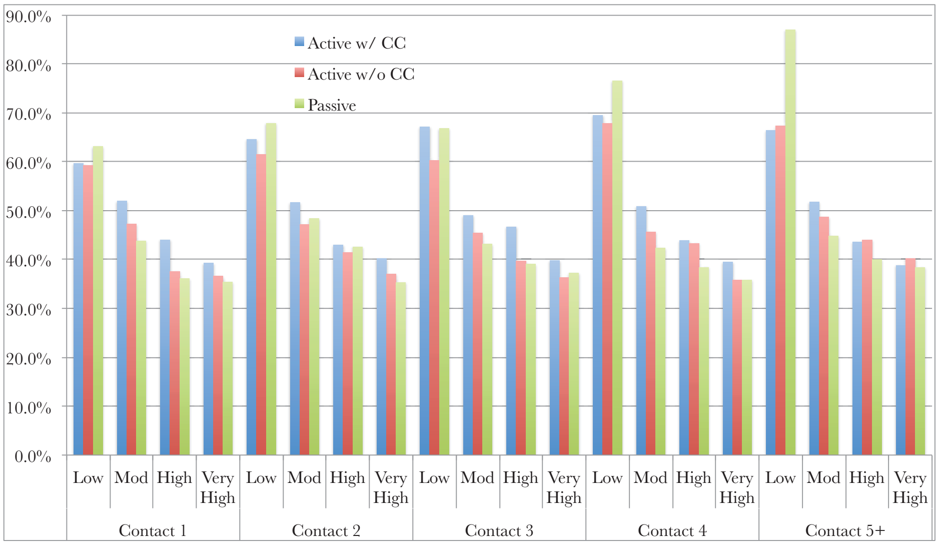


Figure 5.3: Probability of successful exit from pretrial supervision among SSU clients with contact infractions, by agency response and risk level.

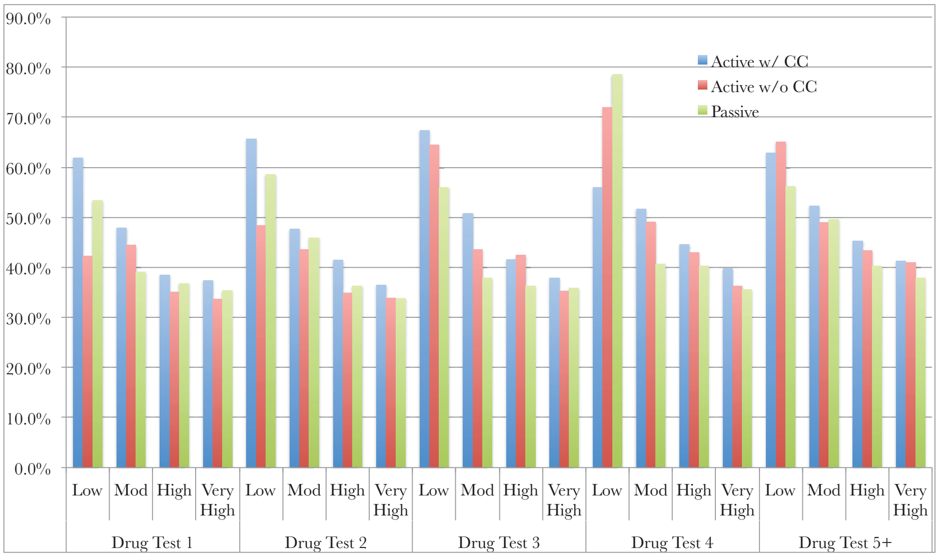


Figure 5.4: Probability of successful exit from pretrial supervision among SSU clients with Drug Test infractions, by agency response and risk level.

among clients with repeat infractions. However, for all other risk levels, typically, active responses to contact infractions appears to offer the best chance of guiding the client towards successful completion of pretrial supervision.

The findings for SSU clients with drug testing infractions (Figure 5.4) are more mixed. While success probabilities continue to be higher among the lower risk clients, there appear to be no discernible patterns in the active or passive response types (with the anomalous exception of the fourth infraction, where the success probabilities of passive responses to low risk clients has the highest success probability). Among the high and very high risk categories, the success probabilities of the various responses types are almost identical.

The next set of graphics provide some insights into the issue of discretion. While using empirical data to inform policy choices, it becomes apparent that the pretrial supervision population is fairly diverse. Even though PSA is attempting to account for this diversity by creating risk-based supervision strategies, the underlying assumption is that a particular policy will be formulated and applied to a specific risk level. The Markov models estimated in this effort provide a way to assess whether there is a need to expand the discretion of pretrial supervision staff by formulating a policy with several options. Towards that end, the next set of graphics were developed to assess the *number* of responses that might help increase the chance of successful pretrial supervision. Unlike figures 5.1 through 5.4, where specific response types were analyzed, figures 5.1 through 5.6 plot the number of responses that have a better than average success probability.

Figure 5.5 shows that, irrespective of the infraction number, the number of agency responses that could increase the chance of clients successfully completing pretrial supervision after contact infractions increase with risk level. This is because the bars are typically higher in Figure 5.5 for the higher risk groups than the lower. Indeed, the figure shows an interaction effect between the risk level and infraction number. Most options are available for the high/very high risk clients on their first contact infraction. These numbers decline somewhat as the contact infraction number increases. In other words, as the client continues to have contact infractions, the number of available responses that could help the client reduces. By the fourth contact infraction, few promising options are available.

When responding to drug test infraction for the same population (non SSU clients), a slightly different trend is observed (Figure 5.6). The number of promising options available for high/very high risk clients remains fairly stable through repeat drug test infractions. However, the number of such

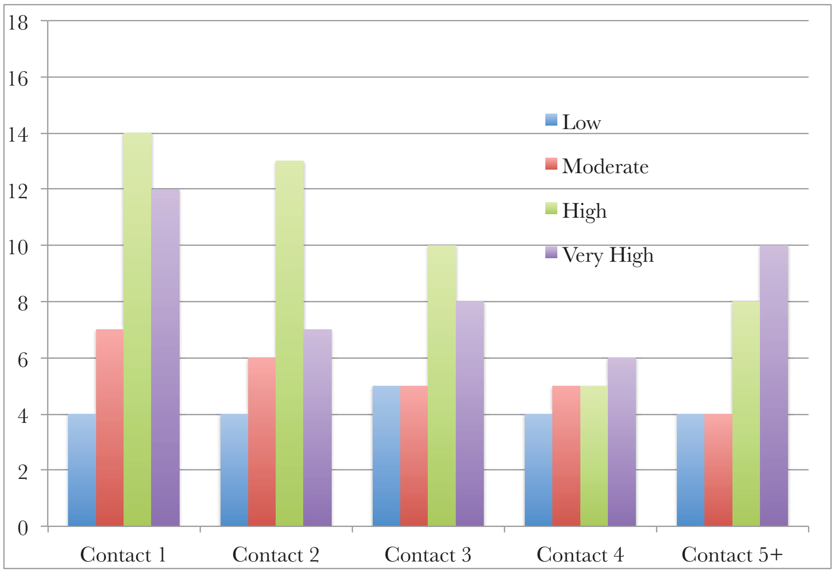


Figure 5.5: Number of different agency responses to contact infractions that produce better than average success probabilities among non SSU clients, by risk level.

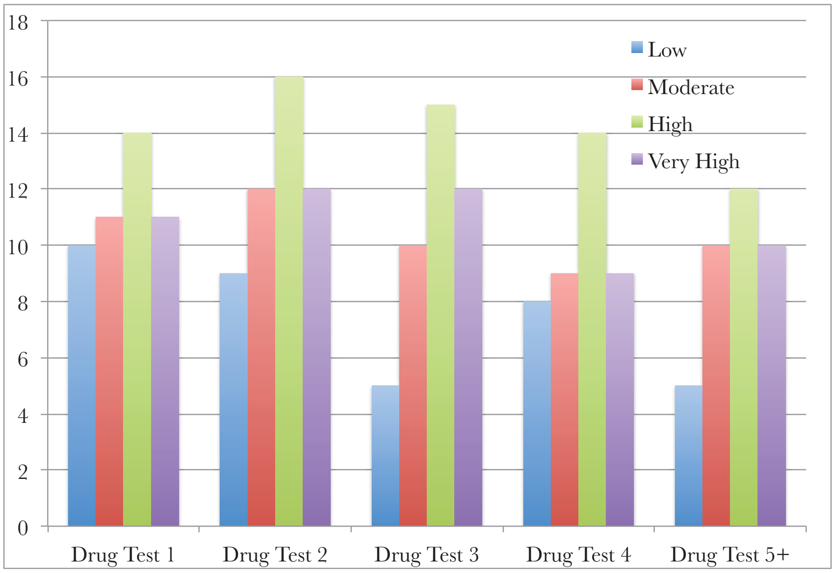


Figure 5.6: Number of different agency responses to drug test infractions that produce better than average success probabilities among non SSU clients, by risk level.

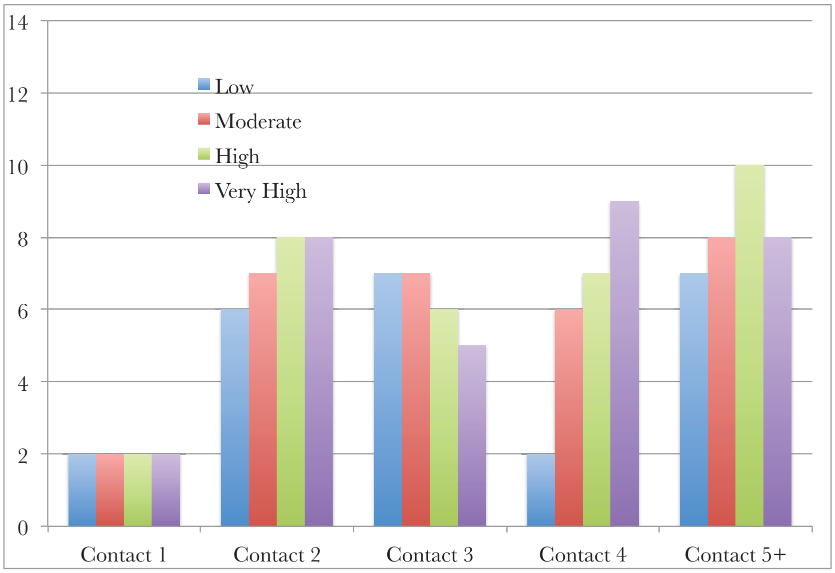


Figure 5.7: Number of different agency responses to contact infractions that produce better than average success probabilities among SSU clients, by risk level.

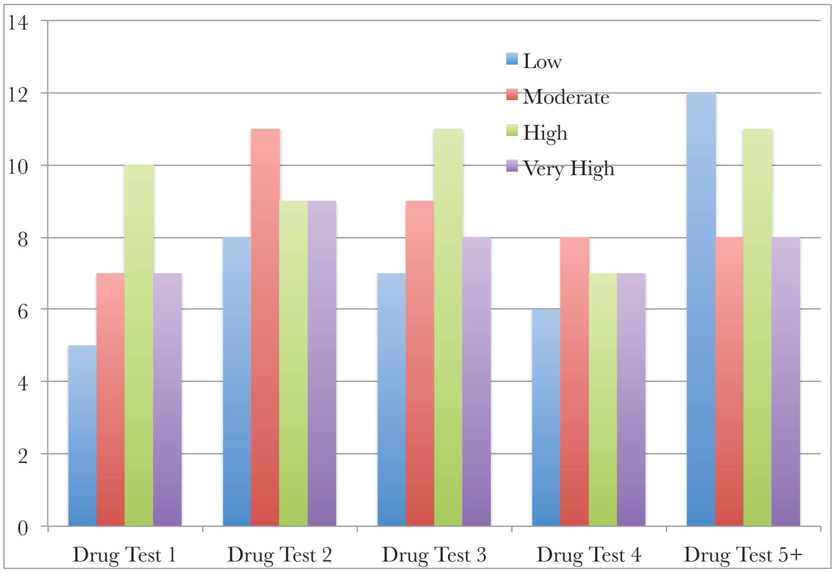


Figure 5.8: Number of different agency responses to drug test infractions that produce better than average success probabilities among SSU clients, by risk level.

choices among the low risk defendants declines as the client has repeated infractions.

Figure 5.7 show yet another kind of pattern. Among SSU clients, there are very few options that produce beneficial results for the first contact infraction. However, as the SSU client has the second, third, and subsequent infractions, the number responses that might work increases. Interestingly, there is no consistent difference by risk level for this population and infraction type.

Finally, Figure 5.8 shows that, while the number of viable options for responding to drug test related infractions is typically higher among higher risk clients, these numbers do not change with the infraction number.

5.5 Conclusion

This chapter described an effort to develop a set of Markov models to help PSA formulate its risk-based supervision strategy. Detailed sequential data on defendant conduct as well as the agency's response to that conduct was obtained from PRISM and was re-structured to estimate the Markov models. The models were used to study the dynamics or risk and how different pathways lead towards successful completion of pretrial supervision or towards misconduct. In particular, the models were used to develop simulations that helped quantify the effects of various policies. Relevant parts of a policy currently under consideration by Office of Operations were simulated. Findings and recommendations are summarized below.

5.5.1 Summary of Findings

1. Overall, the data show expected dynamic patterns. Risk is associated with pathways leading to higher misconduct rates and lower rates of successful pretrial supervision completion.

2. The relation of risk with dynamic patterns is observed within several of the program categories with one exception—Specialized Supervision Unit. Among the SSU clients, the relationship between designated risk levels and misconduct was not as clear cut.

3. Policy simulations show some differences in how SSU and Non-SSU clients respond to various choices made by the agency.

4. Based on PSA feedback, agency response to defendant conduct was classified into three categories—(i) active with client contact, (ii) active without client contact, and (iii) passive. Analysis suggests that the way clients respond to these distinct types of agency responses depends on the type of infraction (contact versus drug test) the infraction number (first, second, etc.), risk level, and the type of client (SSU versus Non SSU). Active responses were most likely to help Non-SSU clients when they engaged in contact violations. The differences between active and passive responses were the least obvious when dealing with drug test violation, in particular among Non SSU clients.

5. The diversity of promising responses showed some interesting links with client complexity. The number of responses that could produce higher than average success probabilities increased, in general, with the complexity of the situation—higher risk level of the client, subsequent infractions, or a combination of the two.

5.5.2 Limitations

While the analysis presented here provides some insights, there are several limitations to the analysis that should be noted.

1. The analysis did not include any incentives. The data collected by PSA only include client infractions and the sanctions that are imposed in response to those infractions. However, in recent years PSA has started exploring incentives as well (by rewarding clients for good conduct). There was insufficient data, at this point, to explore this issue and, as such, this chapter is unable to provide any insights about those policies.

2. The analysis is also unable to provide any guidance on preemptive measures. The Markov models and the Markov Decision Process employed here deals with events (defendant conduct) and agency responses to those event. However, like most supervision strategies, preemptive measures can be extremely useful. Examples include different orientation strategies that are not in response to any infraction or practices like sending defendant reminders (text or phone call) in advance of court appointments. While PSA is developing and pursuing such policies as its strategy, there is very little data in PRISM that can be analyzed to extract information and inform policy.

5.5.3 Recommendations

Based on the analysis described in this chapter as well as the numerous meeting with PSA staff and executives, the following set of recommendations can be made:

1. Use detailed tables provided to inform policy formulation. As a result of the data assembled for this analysis, detailed sets of data tables were estimated and compiled for PSA. These include tables providing expected success probabilities resulting from every response (detailed as well as category) to every infraction for every risk group. While some of the cells in these detailed tables are sparse (small sample sizes), the tables can provide valuable insights to guide PSA staff when formulating their policies. The tables have been delivered to PSA as Excel sheets and, because of the volume of data, are not reproduced in this report.

2. Design pilot projects to study additional aspects of risk-based supervision strategies that PSA currently lacks adequate data on. As noted in the limitations section above, PSA is interested in pursuing preemptive strategies as well as incentives. PSA should focus efforts on developing and piloting some programs so that empirical evidence can be collected and brought to bear on the topic as soon as possible.

3. Balance rules versus discretion in formulating policy. While PSA's focus is on developing a policy for pursuing risk-based supervision, the pretrial supervision population is extremely diverse. Hence, what works for one client may not for another. While general rules are valuable, PSA should carefully balance rules versus discretion when formulating its policy. There are two ways to do that.

   (a) PSA may opt to only recommend response types (active/passive) instead of developing very strict guidelines (e.g., verbal response for second contact infraction). This will permit the supervision officer to gauge the client and make a customized recommendation.

   (b) PSA can develop a policy that offers supervision officers freedom in selecting from a menu of possible choices (e.g., referral to substance use assessment OR referral for treatment in response to third drug testing infraction). A wider list of response choices can be developed for each infraction type based on the detailed data sheets that have been provided to PSA.

Appendices

Appendix A

Data Distribution

	Revalidation Study 201410/201703		Original Study 200710/201003	
	N	%	N	%
Pretrial release samples	38,466	100.0%	37,315	100.0%
Misconduct rate				
Any re-arrest		24.5%		20.5%
Dangerous/Violent Re-arrest		6.7%		6.9%
FTA		21.5%		17.5%
Domestic Violence Re-arrest		11.1%		11.0%
Attribute (Category)				
Current Charges Include				
Felony				
None	28,852	75.0%	28,061	75.2%
1+	9,614	25.0%	9,254	24.8%
Misdemeanor				
None	7,191	18.7%	5,290	14.2%
1+	31,275	81.3%	32,025	85.8%
Person				
None	25,676	66.7%	25,305	67.8%
1+	12,790	33.3%	12,010	32.2%
Property				
None	27,375	71.2%	30,212	81.0%
1+	11,091	28.8%	7,103	19.0%
Weapon				
None	34,311	89.2%	32,479	87.0%
1+	4,155	10.8%	4,836	13.0%
Dangerous				
None	31,996	83.2%	30,262	81.1%
1+	6,470	16.8%	7,053	18.9%
Violent				
None	35,434	92.1%	34,611	92.8%
1+	3,032	7.9%	2,704	7.2%
Sex Crime				
None	37,129	96.5%	33,996	91.1%
1+	1,337	3.5%	3319	8.9%

	Revalidation Study 201410/201703		Original Study 200710/201003	
	N	%	N	%
Sexual Solicitation				
None	37,521	97.5%	34,268	91.8%
1+	945	2.5%	3047	8.2%
Drug Distribution				
None	36,088	93.8%	33,278	89.2%
1+	2,378	6.2%	4,037	10.8%
Drug Possession				
None	35,429	92.1%	26,228	70.3%
1+	3,037	7.9%	11,087	29.7%
DV (Non-person)				
None	33,292	86.5%	30,742	82.4%
1+	5,174	13.5%	6573	17.6%
DV (Person)				
None	33,992	88.4%	31,454	84.3%
1+	4,474	11.6%	5861	15.7%
Criminal Contempt				
None	37,558	97.6%	35,989	96.4%
1+	908	2.4%	1326	3.6%
Prior Arrests				
Felony				
None	26,829	69.7%	23,062	61.8%
1/2	3,844	10.0%	5,655	15.2%
3+	7,793	20.3%	8,598	23.0%
Misdemeanor				
None	16,850	43.8%	15,969	42.8%
1/2	6,477	16.8%	8,097	21.7%
3+	15,139	39.4%	13,249	35.5%
Person				
None	23,048	59.9%	23,073	61.8%
1+	15,418	40.1%	14,242	38.2%

	Revalidation Study 201410/201703		Original Study 200710/201003	
	N	%	N	%
Property				
None	24,214	62.9%	25,622	68.7%
1+	14,252	37.1%	11,693	31.3%
Weapon				
None	31,755	82.6%	28,670	76.8%
1+	6,711	17.4%	8,645	23.2%
Dangerous				
None	27,190	70.7%	21,845	58.5%
1+	11,276	29.3%	15,470	41.5%
Violent				
None	31,508	81.9%	27,114	72.7%
1+	6,958	18.1%	10,201	27.3%
Sex Crime				
None	36,192	94.1%	33,879	90.8%
1+	2,274	5.9%	3,436	9.2%
Sexual Solicitation				
None	36,863	95.8%	34,331	92.0%
1+	1,603	4.2%	2,984	8.0%
Drug Distribution				
None	33,182	86.3%	29,089	78.0%
1+	5,284	13.7%	8,226	22.0%
Drug Possession				
None	28,023	72.9%	24,272	65.0%
1+	10,443	27.1%	13,043	35.0%
DV (Non-person)				
None	33,336	86.7%	34,851	93.4%
1+	5,130	13.3%	2,464	6.6%
DV (Person)				
None	31,239	81.2%	29,048	77.8%
1+	7,227	18.8%	8,267	22.2%

	Revalidation Study		Original Study	
	201410/201703		200710/201003	
	N	%	N	%
Criminal Contempt				
None	34,335	89.3%	33,986	91.1%
1+	4,131	10.7%	3,329	8.9%
Bail reform act (BRA)				
None	31,692	82.4%	30,848	82.7%
1+	6,774	17.6%	6,467	17.3%
Escape				
None	37,334	97.1%	35,002	93.8%
1+	1,132	2.9%	2,313	6.2%
Traffic				
None	37,676	97.9%	32,790	87.9%
1+	790	2.1%	4,525	12.1%
Juvenile				
None	30,664	79.7%	34,140	91.5%
1+	7,802	20.3%	3,175	8.5%
Age at first arrest				
min/17	97	0.3%	2867	7.7%
18/24	9,661	25.1%	8,688	23.3%
25/34	12,245	31.8%	8,986	24.1%
35/44	6,987	18.2%	8,044	21.6%
45/max	9,476	24.6%	8,730	23.4%
Lambda (Prior arrests per year age)				
min/0.1	21,883	56.9%	18,072	48.4%
>0.1/.25	7,162	18.6%	9,392	25.2%
>.25/.5	5,652	14.7%	6,422	17.2%
>.5/max	3,757	9.8%	3,429	9.2%
Prior Bench Warrants				
None	18,550	48.2%	29,777	79.8%
1/2	7,828	20.4%	5,854	15.7%
3+	12,088	31.4%	1,684	4.5%

	Revalidation Study 201410/201703		Original Study 200710/201003	
	N	%	N	%
Prior Convictions				
Felony				
None	29,263	76.1%	27,383	73.4%
1/2	2,783	7.2%	3,784	10.1%
3+	6,420	16.7%	6,148	16.5%
Misdemeanor				
None	23,051	59.9%	23,083	61.9%
1/2	4,972	12.9%	7,105	19.0%
3+	10,443	27.1%	7,127	19.1%
Person				
None	28,000	72.8%	29,583	79.3%
1+	10,466	27.2%	7,732	20.7%
Property				
None	29,031	75.5%	31,007	83.1%
1+	9,435	24.5%	6,308	16.9%
Weapon				
None	34,200	88.9%	32,816	87.9%
1+	4,266	11.1%	4,499	12.1%
Dangerous				
None	30,081	78.2%	27,432	73.5%
1+	8,385	21.8%	9,883	26.5%
Violent				
None	33,811	87.9%	32,683	87.6%
1+	4,655	12.1%	4,632	12.4%
Sex Crime				
None	37,152	96.6%	35,332	94.7%
1+	1,314	3.4%	1,983	5.3%

	Revalidation Study 201410/201703		Original Study 200710/201003	
	N	%	N	%
Sexual Solicitation				
None	37,595	97.7%	35,613	95.4%
1+	871	2.3%	1,702	4.6%
Drug Distribution				
None	34,539	89.8%	31,444	84.3%
1+	3,927	10.2%	5,871	15.7%
Drug Possession				
None	31,423	81.7%	28,688	76.9%
1+	7,043	18.3%	8,627	23.1%
DV (Non-person)				
None	35,520	92.3%	35,934	96.3%
1+	2,946	7.7%	1,381	3.7%
DV (Person)				
None	34,467	89.6%	33,825	90.6%
1+	3,999	10.4%	3,490	9.4%
Criminal Contempt				
None	35,483	92.2%	35,020	93.8%
1+	2,983	7.8%	2,295	6.2%
Bail reform act (BRA)				
None	33,967	88.3%	32,299	86.6%
1+	4,499	11.7%	5,016	13.4%
Escape				
None	37,632	97.8%	35,443	95.0%
1+	834	2.2%	1,872	5.0%
Traffic				
None	37,874	98.5%	34,606	92.7%
1+	592	1.5%	2,709	7.3%
Juvenile				
None	32,334	84.1%	35,098	94.1%
1+	6,132	15.9%	2,217	5.9%

	Revalidation Study 201410/201703		Original Study 200710/201003	
	N	%	N	%
Current Criminal Justice Status				
Pending Criminal Charge				
None	26,515	68.9%	26,049	69.8%
1+	11,951	31.1%	11,266	30.2%
Pending Dangerous/Violent Charge				
None	35,252	91.6%	35,151	94.2%
1+	3,214	8.4%	2,164	5.8%
Currently on Probation/Parole				
No	31,848	82.8%	30,780	82.5%
Yes	6,618	17.2%	6,535	17.5%
Demographic/Social Indicators				
Gender				
Male/Unknown	30,648	79.7%	30,228	81.0%
Female	7,818	20.3%	7,087	19.0%
Current Age				
min/24	9,751	25.3%	8,671	23.2%
25/34	12,248	31.8%	10,215	27.4%
35/44	6,989	18.2%	7,955	21.3%
45/max	9,478	24.6%	10,474	28.1%
US Citizen				
No	279	0.7%	947	2.5%
Yes	19,181	49.9%	24,325	65.2%
Null	19,006	49.4%	12,043	32.3%
DC Resident				
No	15,908	41.4%	17,383	46.6%
Yes	22,558	58.6%	19,932	53.4%
Employment Status				
Unemployed	14,065	36.6%	12,426	33.3%
Employed	7,991	20.8%	9,273	24.9%
Other	16,410	42.7%	15,616	41.8%

	Revalidation Study 201410/201703		Original Study 200710/201003	
	N	%	N	%
Total # of Children				
None	20,662	53.7%	21,531	57.7%
1/5	16,707	43.4%	14,832	39.7%
6+	1,097	2.9%	952	2.6%
Live with Children				
No	30,798	80.1%	30,625	82.1%
Yes	7,668	19.9%	6,690	17.9%
Emotional Problems				
No	31,267	81.3%	34,922	93.6%
Yes	7,199	18.7%	2,393	6.4%
Physical Problems				
No	34,907	90.7%	35,374	94.8%
Yes	3,559	9.3%	1,941	5.2%
Lockup Drug Test				
Amp				
Neg	20,967	54.5%	26,509	71.0%
Pos	310	0.8%	486	1.3%
Null	17,189	44.7%	10,320	27.7%
Coc				
Neg	18,161	47.2%	17,953	48.1%
Pos	3,116	8.1%	9,043	24.2%
Null	17,189	44.7%	10,319	27.7%
Opi				
Neg	19,698	51.2%	24,322	65.2%
Pos	1,579	4.1%	2,671	7.2%
Null	17,189	44.7%	10,322	27.7%
Pcp				
Neg	19,116	49.7%	24,216	64.9%
Pos	2,161	5.6%	2,779	7.4%
Null	17,189	44.7%	10,320	27.7%
Drug test compliant				
No	14,813	38.5%	16,412	44.0%
Yes	15,389	40.0%	14,739	39.5%
Null	8,264	21.5%	6,164	16.5%

Appendix B

Old vs Revised Risk Instrument Features

Feature	Label	New	Old	Change Explanation (between original old and new43)
Current Charge				
CCFelony	Current case includes a Felony charge	x	x	No change
CCMisd	Current case includes a Misd charge	x	x	No change
CCPerson	Current case includes a Person charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCProperty	Current case includes a Property charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCWeapon	Current case includes a Weapons charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCDang	Current case includes a Dangerous charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCViolent	Current case includes a Violent charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCSexCrime	Current case includes a Sex Crime charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCSexSol	Current case includes a Sexual Solicitation charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCDrgDist	Current case includes a Drug Distribution charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCDrgPos	Current case includes a Drug Possession charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCDomNPer	Current case includes a Nonperson Domestic Violence charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCDomPer	Current case includes a Person Domestic Violence charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
CCCrmCtmp	Current case includes a Criminal Contempt charge	x	x	No change (dropped for interim36 because of concerns, but added back in)
Current CJ Status				
CJPendCrm	Defendant has pending Criminal charge	x	x	No change
CJPendDV	Defendant has pending Dangerous or Violent charge	x	x	No change
CJProPar	Defendant is currently on probation/parole	x	x	No change
Demographic/Social Indicators				
DSFemale	Defendant is female	x	x	No change
DSAgeCat	Defendant age	x	x	No change
DSUSCit	Defendant is US citizen	x	x	No change
DSDCRes	Defendant is DC Resident	x	x	No change
DSEmpType	Defendant's employment status	x	x	No change
DSTotChild	Total number of children defendant has	x	x	No change
DSLivWChld	Does defendant live with children	x	x	No change
DSEmoProb	Defendant has emotional problems	x	x	No change
DSPhysProb	Defendant has physical problems	x	x	No change
PSSU	Defendant has past assignment to Specialised Supervision Unit	x		Added to include objective measure of MH status
Luckup Drug Test Results				
LUAmp	Tested positive for Amphetamines @ lockup	x	x	No change
LUCoc	Tested positive for Cocain @ lockup	x	x	No change
LUOpi	Tested positive for Opioids @ lockup	x	x	No change
LUPcp	Tested positive for Pcp @ lockup	x	x	No change
LUK2	Tested positive for K2 @ lockup	x		Added per agency request
LUDrgComp	Overall compliance @ lockup drug test		x	Dropped because this measure is redundant with the individual drug-specific items
Criminal History Measures				
CHIConF10	# of Misd charges for which convicted (Internal) within last 10 years	x		Added in lieu of detailed charge specific criminal history measures that are dropped
CHIConM10	# of Felony charges for which convicted (Internal) within last 10 years	x		Added in lieu of detailed charge specific criminal history measures that are dropped
CHIConF11	# of Misd charges for which convicted (Internal) more than 10 years ago	x		Added in lieu of detailed charge specific criminal history measures that are dropped
CHIConM11	# of Felony charges for which convicted (Internal) more than 10 years age	x		Added in lieu of detailed charge specific criminal history measures that are dropped
CHECon10	# of charges for which convicted (External) within last 10 years	x		Added in lieu of detailed charge specific criminal history measures that are dropped
CHECon11	# of charges for which convicted (External) more than 10 years ago	x		Added in lieu of detailed charge specific criminal history measures that are dropped
CHIArF10	# of Misd charges for which arrested (Internal) within last 10 years			Added for interim36 but eventually dropped because of correlation with convictions
CHIArM10	# of Felony charges for which arrested (Internal) within last 10 years			Added for interim36 but eventually dropped because of correlation with convictions

Feature	Label	New	Old	Change Explanation (between original old and new43)
CHIArF11	# of Misd charges for which arrested (Internal) more than 10 years ago			Added for interim36 but eventually dropped because of correlation with convictions
CHIArM11	# of Felony charges for which arrested (Internal) more than 10 years age			Added for interim36 but eventually dropped because of correlation with convictions
CHEAr10	# of charges for which arrested (External) within last 10 years			Added for interim36 but eventually dropped because of correlation with convictions
CHEAr11	# of charges for which arrested (External) more than 10 years ago			Added for interim36 but eventually dropped because of correlation with convictions
CHIArPerYr	Lambda Internal (# of Internal arrest charges / Current Age)	x		Added in lieu of detailed charge specific crimina history measures that are dropped
CHEArPerYr	Lambda Extrtnal (# of External arrest chanrges / Current Age)	x		Added in lieu of detailed charge specific crimina history measures that are dropped
CHArPerYr	Lambda (# of total arrest charges / Current Age)		x	Dropped because different (internal and external) measures are now included
CHBW	# Prior Bench Warrants	x	x	No change
CHArJuv	# Juvenile Arrests	x	x	No change
CHAgeFirst	Age at First Arrest	x	x	No change
CHArFel	# of Felony charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArMisd	# of Misd charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArPerson	# of Person charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArProp	# of Property charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArWeap	# of Weapons charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArDang	# of Dangerous charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArVio	# of Violent charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArSexCrm	# of Sex Crimes charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArSexSol	# of Sexual Solicitation charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArDrgDis	# of Drug Distribution charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArDrgPos	# of Drug Possession charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArDVNPer	# of Non-person Domestic Violence charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArDVPer	# of Person Domestic Violence charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArCC	# of Criminal Contemp charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArBra	# of Bail Reform Act charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArEsc	# of Escape charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHArTraf	# of Serious Traffic charges for which ever arrested		x	Dropped because high correlation and issues with charge information
CHConFel	# of Felony charges for which ever convicted		x	Dropped because different internal/external and <10/10+ versions are now included
CHConMisd	# of Misd charges for which ever convicted		x	Dropped because different internal/external and <10/10+ versions are now included
CHConPer	# of Person charges for which ever convicted		x	Dropped because of issues with charge information
CHConProp	# of Property charges for which ever convicted		x	Dropped because of issues with charge information
CHConWeap	# of Weapons charges for which ever convicted		x	Dropped because of issues with charge information
CHConDang	# of Dangerous charges for which ever convicted		x	Dropped because of issues with charge information
CHConVio	# of Violent charges for which ever convicted		x	Dropped because of issues with charge information
CHConSexCr	# of Sex Crimes charges for which ever convicted		x	Dropped because of issues with charge information
CHConSexSo	# of Sexual Solicitation charges for which ever convicted		x	Dropped because of issues with charge information
CHConDDis	# of Drug Distribution charges for which ever convicted		x	Dropped because of issues with charge information
CHConDPos	# of Drug Possession charges for which ever convicted		x	Dropped because of issues with charge information
CHConDVNP	# of Non-person Domestic Violence charges for which ever convicted		x	Dropped because of issues with charge information
CHConDVP	# of Person Domestic Violence charges for which ever convicted		x	Dropped because of issues with charge information
CHConCC	# of Criminal Contemp charges for which ever convicted		x	Dropped because of issues with charge information
CHConBRA	# of Bail Reform Act charges for which ever convicted		x	Dropped because of issues with charge information
CHConEsc	# of Felony charges for which ever convicted		x	Dropped because of issues with charge information
CHConTraf	# of Felony charges for which ever convicted		x	Dropped because of issues with charge information
CHConJuv	# of Juvenile convictions		x	Dropped beccause conviction information is inconcistent with arrest information

Appendix C

Program Categories

Program Categories/Queue Names	138,346	100%
Substance Use Disorder (SUD)	10,606	7.7%
1015 PSA Drug Court		
1021 SC Sanctions Based Testing and Treatment		
1033 PSA SC New Directions		
General Supervision	56,957	41.2%
1018 SC General Supervision - Extensive		
1073 SC General Supervision - Monitor		
1097 Work Release - Extended		
1098 Work Release - Fairview		
1099 Work Release - Hope Village		
1112 Work Release - Extended w/ GPS		
1113 Work Release - Hope Village w/ GPS		
1114 Work Release - Fairview w/ GPS		
Specialized Supervision Unit (SSU)	29,134	21.0%
1042 SC Specialized Supervision Options		
1045 SC Specialized Supervision Unit		
District Court	120	0.1%
1048 USDC General Supervision		
1051 USDC Sanctions Based Testing and Treatment		
1126 USDC High Intensity Supervision Program		
High Intensity Supervision Program (HISP)	36,973	26.7%
1092 High Intensity Supervision Program		
1127 SC GPS Only		
Traffic Safety	4,556	3.3%
1124 PSA Traffic Safety Supervision		

Appendix D

Defendant Conduct Types

Infraction Categories/Types	138,408	100%
Contact	30,614	22.1%
Contact infraction - Face to Face		
Contact infraction - In Person		
Contact infraction - Telephone		
Drug Testing	72,020	52.0%
Drug testing infraction		
Electronic Supervision	28,822	20.8%
EM Alert		
EM infraction		
Group Sessions	1,670	1.2%
Fail to appear - group		
Other	5,282	3.8%
Contact infraction		
Contact infraction - Address Verificati		
Contact infraction - Mental Health Asse		
Contact infraction - Social Services As		
Contact infraction - Substance Abuse As		
Curfew infraction		
Escape or Abscond HWH		
Fail to abide by stay away condition		
Fail to appear - Court		
Fail to complete orientation process		
Fail to comply with substance abuse tre		
Fail to report for drug evaluation or d		
Fail to verify address		
Inpatient treatment failure		
Loss of contact		
Other infraction		
Re-Arrest		

Appendix E

Active and Passive Agency Responses

Responses

- 🟩 Behavior contract
- 🟩 Clinical session
- 🟩 Clinical session & group processing
- 🟩 Clinical staffing with participant
- 🟨 Community Service (4-hour)
- 🟩 Daily reporting
- 🟨 Email notification to Court
- 🟩 Enhanced treatment
- 🟧 Evaluate for HISP
- 🟩 Group Assist
- 🟩 Group processing
- 🟩 Home Confinement
- 🟩 Increase reporting condition
- 🟩 Increased drug testing
- 🟧 Invalid EM Alert
- 🟨 Loss of Contact Investigation
- 🟨 Modification of treatment
- 🟧 No Response Required
- 🟧 Other: staffing driven
- 🟨 Outside Meetings (3 groups)
- 🟧 Pending Documentation/Confirmation
- 🟩 Placement in drug treatment
- 🟩 Placement in mental health treatment
- 🟨 Re-assessment
- 🟨 Recommend Jail/CellBlock/JuryBox
- 🟨 Recommend additional court hearings
- 🟨 Recommend judicial rev/warn/admonishment
- 🟨 Reconsider treatment plan or strategy
- 🟩 Redirection group
- 🟩 Referral mental health assessment
- 🟩 Referral substance abuse assessment
- 🟩 Removal from HISP
- 🟩 Reorientation
- 🟩 Report by telephone
- 🟩 Report for weekly drug testing
- 🟨 Request for Removal from PSA Supervision
- 🟨 Request for Show cause hearing
- 🟨 Request for removal from program
- 🟨 Restrict social passes (HWH)
- 🟩 Restricted curfew
- 🟨 Sanction Based Treatment
- 🟩 Self-Help
- 🟩 Spot drug testing
- 🟨 Submitted report to court
- 🟩 Suspension of Treatment groups
- 🟨 Unable to respond – attempts made
- 🟧 Unable to respond – no contact information
- 🟩 Verbal warning
- 🟩 Weekly reporting
- 🟨 Workbook-Treatment
- 🟨 Written assignment
- 🟩 Written warning
- 🟧 _NULL

Response Types

- 🟩 Actve w/ client contact
- 🟨 Active w/o client contact
- 🟧 Passive