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

Predictive Bias Report

Contract No. PSA17C0043

Avinash Bhati, PhD
Maxarth LLC

12215 Fellowship Lane,
North Potomac, MD 20878

May 5, 2019

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 report presents findings of an analysis of racial differences in the Pretrial Services Agency's (PSA) recently re-validated risk assessment instrument. The instrument is designed to assess defendants on four dimensions of risk of re-arrest for any charge, re-arrest for a dangerous or violent charge, failure to appear, and re-arrest for a domestic violence charge (among defendants charged with domestic violence). While the agency requested a re-validation of all four risk scores, it intends to mostly rely on the any re-arrest (safety) and failure to appear (flight) risk scores.

Data and Methods

PSA's risk instrument is based on scores computed by aggregating weights applied to 43 risk items from five domains. These include items related to (i) criminal history, (ii) current charge, (iii) criminal justice system status, (iv) drug test results, and (v) defendant social and demographic attributes. Using a scorecard schema, each of the 43 items is weighted and summed to create an overall score for each dimension of risk. Each of the scores is further normed to lie between 0 and 100. The normed scores are finally converted into risk categories (low, medium high or very high). Details about the data used and the re-validation findings are documented in a final technical report (submitted to PSA).

Table 1 provides a breakdown of the data used for this study by race categories. Defendants with cases filed between Oct 2014 and Oct 2017 are 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 for a sufficient post-case filing follow-up period 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 of by Oct 2017. There were a total of 38,477 clients with cases filed between Oct 2014 and Mar 2017 with a pretrial release.

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

Eighty-six percent of the full sample were Black with the remainder mostly White (7.6%) or Hispanic (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.

Table 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

Table 2: Average risk scores and misconduct rates, by race.

	White	Black	Hispanic	Total
Risk Scores				
Any re-arrest	37.9	45.9	40.4	44.9
Dangerous/Violent re-arrest	46.9	53.2	50.2	52.5
Failure to appear	41.8	43.3	41.4	43.1
Domestic violence re-arrest	48.3	50.1	48.8	49.9
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%

The various sub-samples described above are similar to the overall sample—Blacks constituted a majority of the cases (between 85% and 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 2 shows the distribution of the re-validated risk scores, pretrial release rates, and misconduct rates by racial groups. In general, Blacks have much higher average risk scores than Whites or Hispanics. The average re-arrest score (for any crime) is 45.9 for Blacks, followed by 40.4 for Hispanics and 37.9 for Whites. The average dangerous/violent re-arrest score is 53.2 for Blacks, followed by 50.2 for Hispanics and 46.9 for Whites. Similar trends are seen for the FTA and domestic violence risk scores. Blacks and Hispanics have the lowest pretrial release rate (92.3% and 93.9% respectively) followed by Whites (95%).

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 2 lends some confidence regarding racial bias in the instrument. While the instrument does score Blacks 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.

Findings

A graphical analysis permits studying 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 1 through 4 present the distribution of the risk scores of the different race groups. Consistent with the aggregate risk scores, it is seen that Blacks 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 a skewed distribution among Whites and Hispanics but not among Blacks. Among Whites the peak score is around 30 with a relatively long tail to the right. Among Hispanics the peak is around 35 but with a slightly less pronounced tail to the right. Finally, among Blacks the distribution appears symmetrical around the peak of about 45. Therefore, the difference in the distribution of the any re-arrest score among Whites and Hispanics on the one hand and Blacks on the other are more pronounced than the means in Table 2 would suggest.

The dangerous/violent scores are symmetrical about their peak for all three racial groups—with the distributions for Hispanics shifted to the right of Whites and that for Blacks shifted further to the right. There is a small group of Whites and Hispanics who have a low score of about 25 that is missing among Blacks.

The next set of graphics (figure 5 through figure 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.

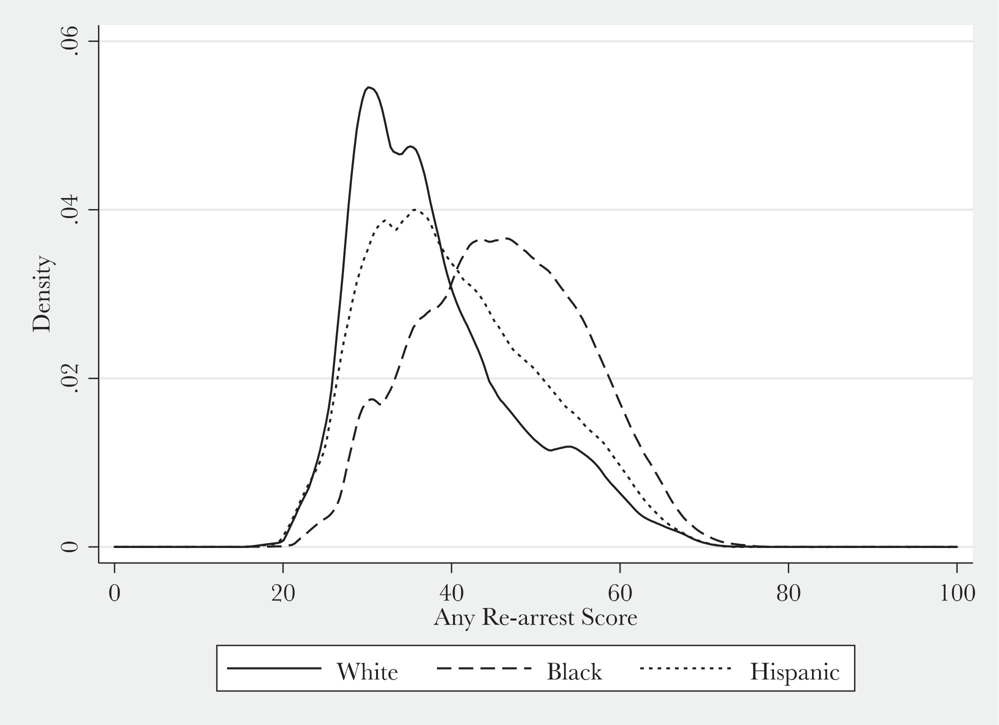


Figure 1: Distribution of Any Re-arrest score, by race.

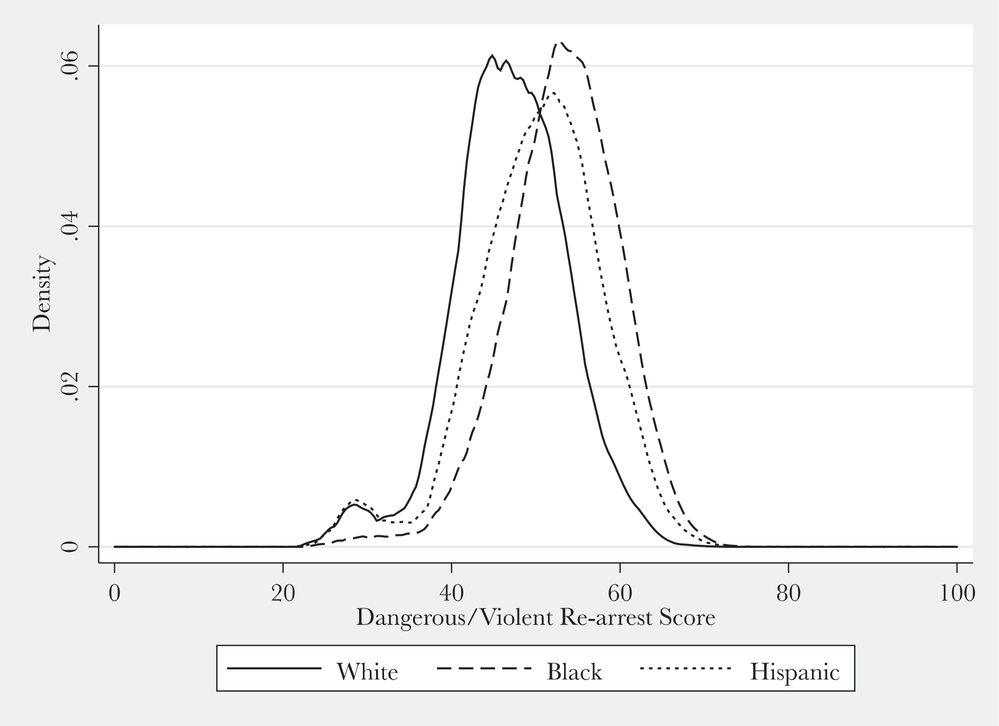


Figure 2: Distribution of Dangerous/Violent Re-arrest score, by race.

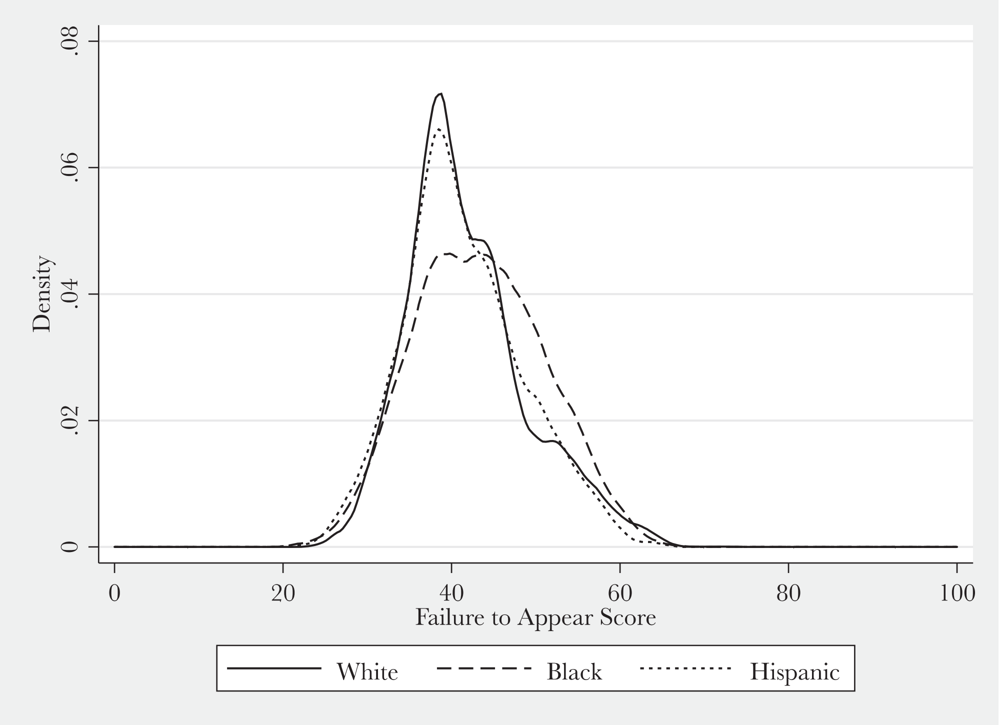


Figure 3: Distribution of Failure to Appear risk score, by race.

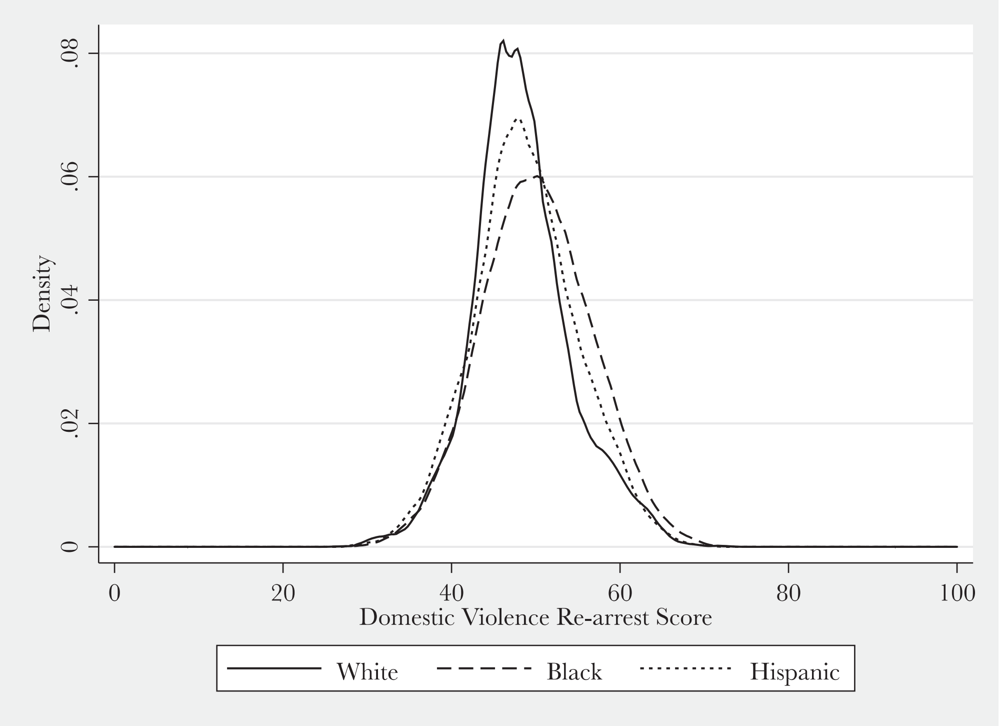


Figure 4: Distribution of Domestic Violence Re-arrest risk score, by race.

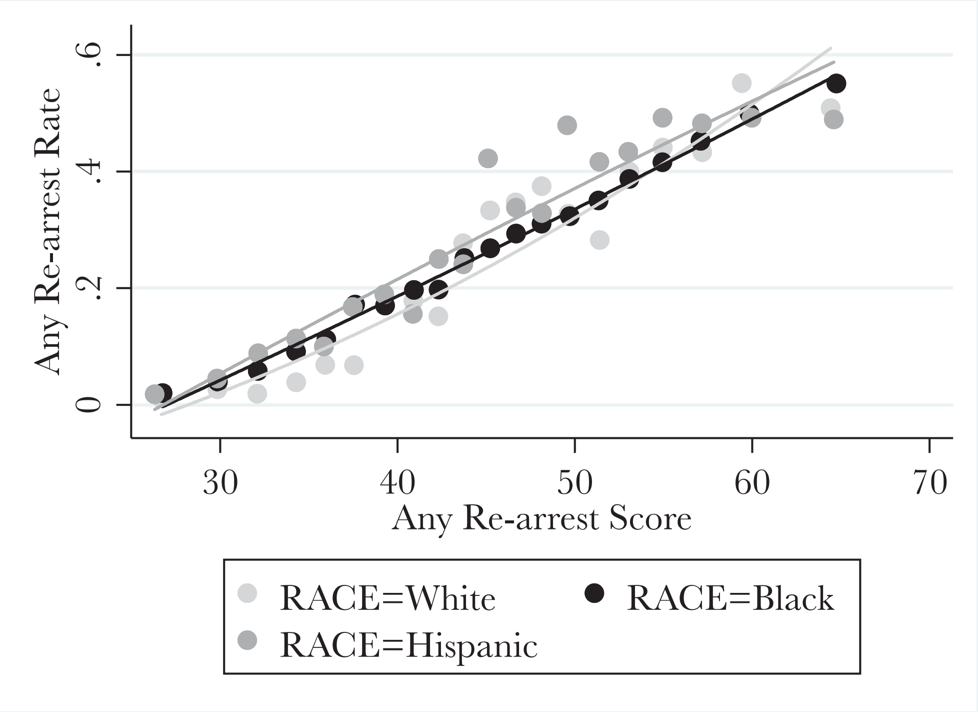


Figure 5: Any Re-arrest rate for 20 quantiles of Any Re-arrest scores, by race.

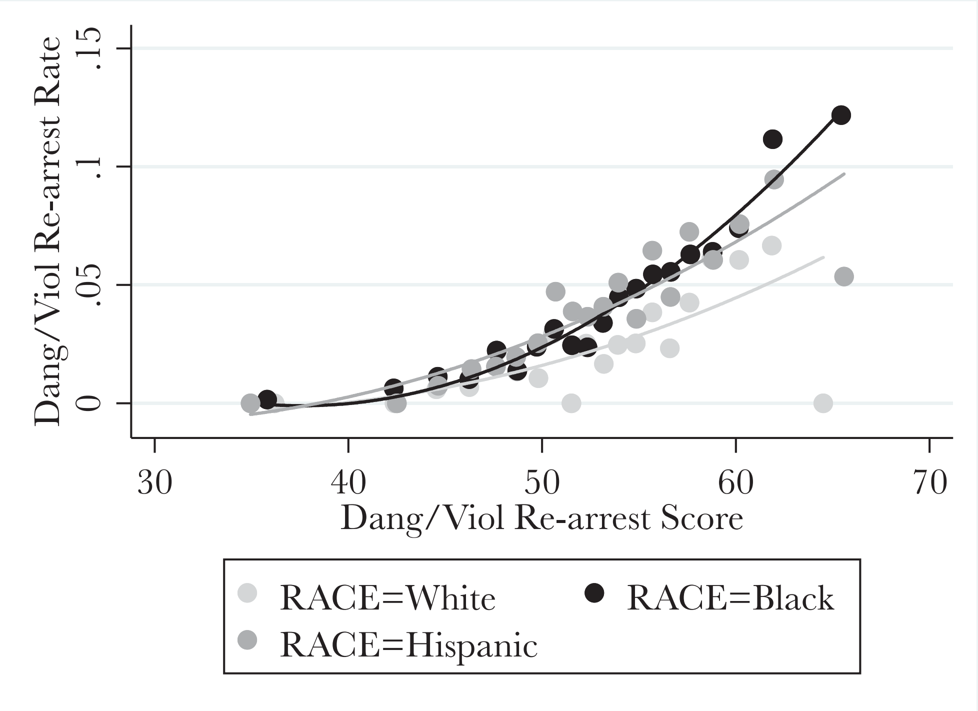


Figure 6: Dangerous/Violent Re-arrest rate for 20 quantiles of Dangerous/Violent Re-arrest risk scores, by race.

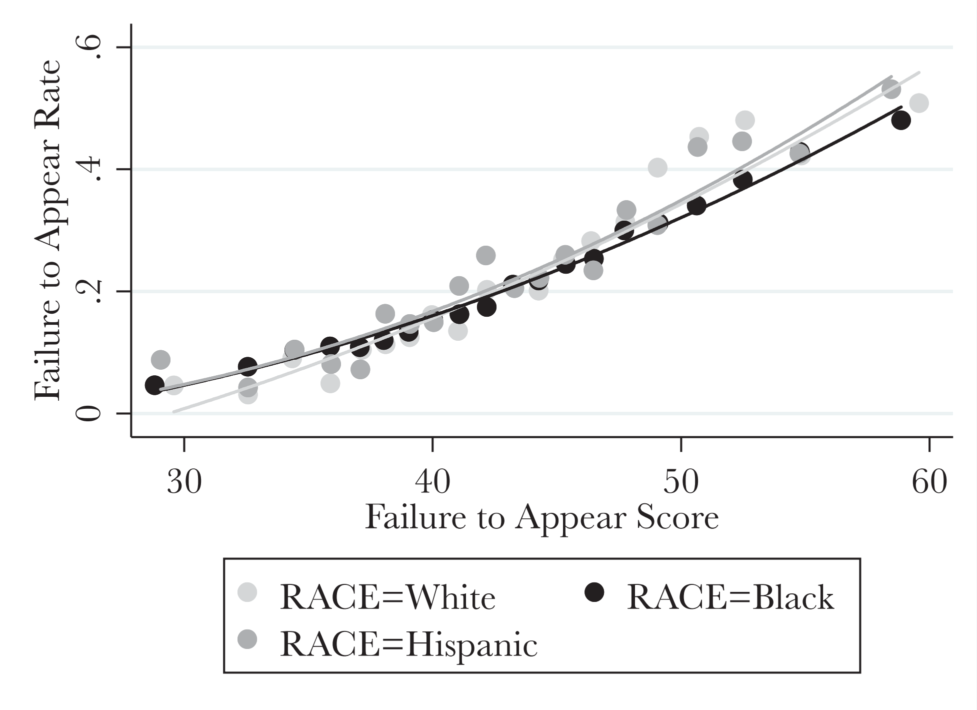


Figure 7: FTA rate for 20 quantiles of FTA risk scores, by race

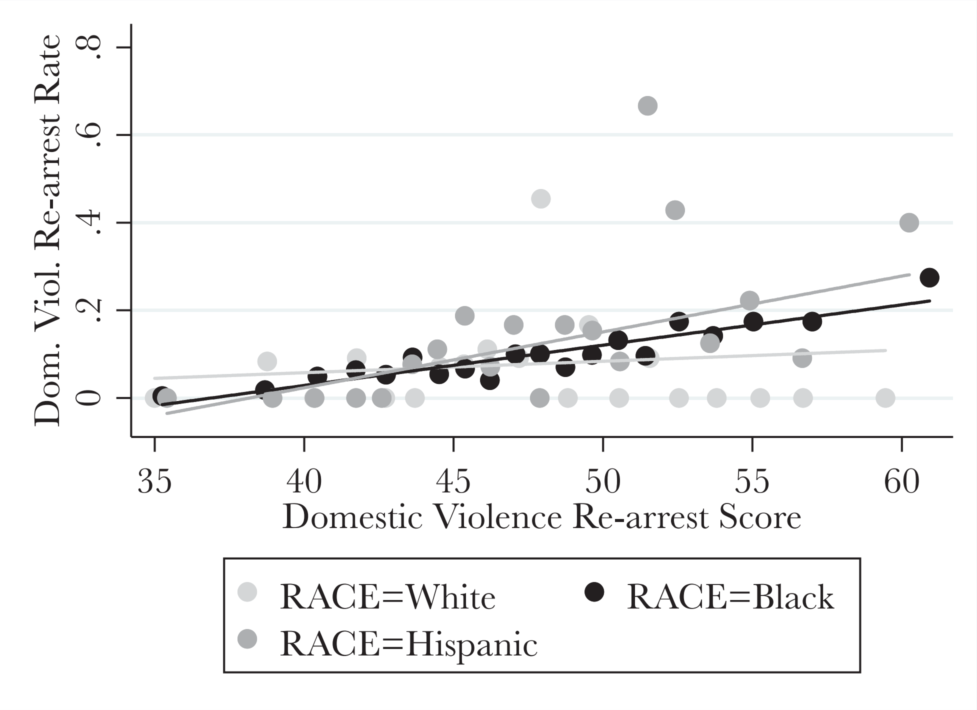


Figure 8: Domestic Violence Re-arrest rate for 20 quantiles of Domestic Violent Re-arrest risk scores (DVM cases), by race.

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. With the exception of the domestic violence risk scores, the binned scatter plots in figures 5 through 8 show that the relationship between risk scores and misconduct is fairly tight (well clustered around the fitted trend line). The somewhat large dispersion of the domestic violence scatter plot can be a result of the small number of cases included in the analysis.

To summarize, the graphic analysis suggests that the risk scores are distributed slightly differently among Black defendants compared with White and Hispanic defendants. On the other hand, the analysis also suggests that the relationship between risk scores and misconduct rates is fairly stable among defendants of all races.

Table 3: Percent distribution of categories based on revised risk scores and observed misconduct rates, by race.

	% in risk category			Misconduct Rate		
	White	Black	Hispanic	White	Black	Hispanic
Any re-arrest						
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 re-arrest						
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 violent re-arrest						
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)

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 shows the distribution of defendants of different races into different risk categories along with their observed misconduct rates. The risk categories are defined using a resource constraint model—i.e., in such a way as to mimic the distribution of low, medium, high, and very high risk categories from the July 2016 through March 2017 period.[2]

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. 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. There are slightly larger 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 or about the same as Blacks among the high risk groups. The difference, while present, are not large.

The graphical analysis as well as the more detailed analysis of risk categories presented above point towards a generally bias free 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 a 2 X 2 contingency table. On one axis is the predicted risk levels (Low Risk or High Risk) while on the other axis is the observed misconduct (No Misconduct or Observed Misconduct). The table below shows the possible combinations:

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

[2] The cut-points for risk categories were last revised in July 2016. The agency was interested in the new categories mimicking the distributions of the existing four categories. Prior to July 2016, the agency used five risk categories.

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:

$$\begin{aligned} Specificity &= \frac{TN}{TN+FP} \\ Sensitivity &= \frac{TP}{TP+FN} \end{aligned}$$

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 4 shows the calculated AUC statistics for each of the instruments by race (in the first three columns) and for all races combined (in the last column). To make comparison possible, the table also shows the 95% confidence (lower and upper) bounds for each of these numbers.

Table 4: Area Under the Curve (AUC) statistics, by race.

	White	Black	Hispanic	All Races
Any re-arrest				
AUC statistic	0.83	0.71	0.77	0.73
95% low bound	0.81	0.71	0.74	0.73
95% high bound	0.85	0.72	0.79	0.74
Dangerous/violent re-arrest				
AUC statistic	0.76	0.71	0.72	0.72
95% low bound	0.70	0.69	0.67	0.71
95% high bound	0.82	0.72	0.77	0.73
Failure to appear				
AUC statistic	0.73	0.70	0.70	0.70
95% low bound	0.71	0.69	0.67	0.69
95% high bound	0.75	0.70	0.73	0.71

Domestic violence re-arrest				
AUC statistic	0.60	0.68	0.76	0.68
95% low bound	0.48	0.66	0.68	0.66
95% high bound	0.72	0.71	0.84	0.70

With the exception of the domestic violence score, the risk scores typically have an AUC score above 0.70 for all risk dimensions and all race groups. Within criminal justice settings, AUC scores at or above 0.70 are deemed desirable. The table shows a distinct racial pattern, however. With the exception of the domestic violence score, the risk assessment instruments are all better at separating the White defendants into high and low risk groups than Black or Hispanic defendants. In other words, the instruments are more *specific* and/or more *sensitive* when assessing Whites than while assessing Blacks or Hispanics. In general, all of the AUC scores are higher among Whites than among Blacks and Hispanics.

The pattern is reversed for the domestic violence score. As was noted earlier, however, the domestic violence results should be viewed with caution. In particular, AUC statistics are sensitive to small sample sizes and given the small number of cases available for the domestic violence analysis, AUC statistics computed for the domestic violence instrument by race may not be very reliable.

One of the main drawbacks 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 a low cut-point (e.g., 20 in our scores) to identify high risk defendants. But the AUC score is computed for all possible cut-points in the data. Moreover, it is a retrospective 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 realistic cut-points or risk categories 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:

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

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

The *FDR* is the proportion of those categorized at a high risk of misconduct who did not have a misconduct and the *FOR* is the proportion of those categorized at a low risk of misconduct who did have a misconduct. Table 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 very similar across all racial groups with one exception. The FDR for the FTA instrument among Blacks is 61.6% while among Whites it is 54.7%—a difference of about 6.9 percent points in favor of Whites. This means that the FTA instrument is incorrectly identifying Blacks as high risk at a slightly higher rate than Whites. In this instance, the discrepancy appears to favor White defendants. A similar pattern is seen in the FOR for all scores and all races, with one exception. The FOR for the any re-arrest score among Black defendants (9.9%) is about 5.7 percent points higher than that for White defendants (4.2%). This means that the any re-arrest instrument has a tendency to incorrectly score Black defendants as *low* risk when in-fact they get re-arrested. In this instance, the discrepancy appears to favor Black defendants.

Table 5: False Discovery and False Omission rates using categories based on each risk score, by race.

	False Discovery Rate*			False Omission Rate**		
	White	Black	Hisp.	White	Black	Hisp.
Any re-arrest	53.9%	54.1%	52.0%	4.2%	9.9%	9.1%
Dangerous/violent re-arrest	94.2%	89.4%	93.2%	…	1.5%	…
Failure to appear	54.7%	61.6%	59.1%	8.2%	9.7%	9.7%
Domestic violence re-arrest	…	79.3%	…	…	5.1%	…

* Very high + high categories predicted to fail

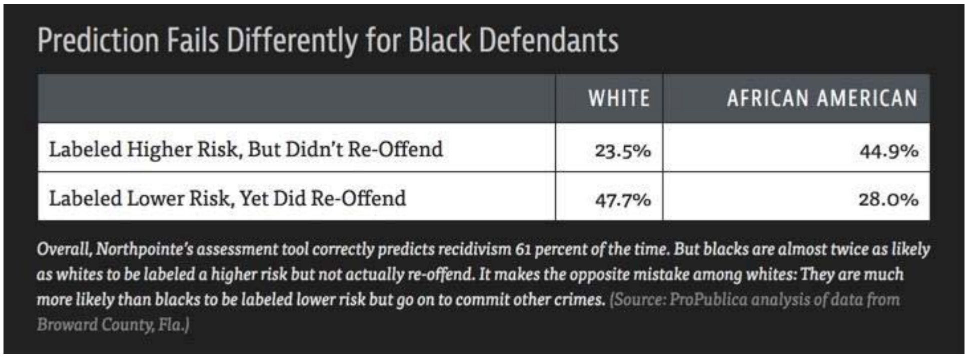
** Low risk category predicted to not fail

… (N<30)

While some differences are natural in all risk assessment instruments, it should be noted that the false discovery and false omission rates for all racial groups are typically within less than six to seven percentage points of one another.

Conclusion

The analysis reported in this report was designed to assess algorithmic bias in PSA's recently re-validated risk assessment instrument. While risk scores and misconduct rates vary by race, the relationship between risk scores and observed misconduct remains fairly stable across race. Moreover, while the predictive efficacy of the instruments are generally better among White defendants, the errors made by the instruments are fairly consistent across different races.



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 9: Screen shot from ProPublica's article describing predictive bias.

It should be noted that the error differences found in PSA's risk assessment instrument are small compared to the biases that have been reported elsewhere and that are at the heart of the concern in the field. Figure 9, for example, shows what ProPublica's analysis concluded about the implementation of COMPASS in Broward County Florida.[3] They found a twenty percentage-point FDR gap in favor of Whites and a similar twenty percentage-point FOR gap also in favor of Whites. That is, Blacks were much more likely than Whites to be *falsely* identified as high risk while Whites were much more likely than Blacks to be *falsely* identified as low risk. The analysis of PSA's data suggests that the re-validated risk assessment instrument is largely unbiased in assessing risk of misconduct for defendants of different racial groups.

[3] See, however, Flores, Lowenkamp, and Bechtel (2017) for a critique of the ProPublica analysis (`http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf`).