U.S. Department of Justice
Office of Justice Programs
National Institute of Justice

National Institute of Justice

2021 Review and Revalidation of the First Step Act Risk Assessment Tool

December 2021

NCJ 303859

NIJ | National Institute of Justice
STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Table of Contents

Executive Summary

The First Step Act (FSA) of 2018 mandates the development and implementation of a risk and needs assessment system for use with each person in the custody of the Federal Bureau of Prisons (BOP). Section 3634 of Title I of the FSA requires the U.S. Department of Justice (USDOJ or the Department) to review, validate, and release publicly its risk and needs assessment system — the Prisoner Assessment Tool Targeting Estimated Risk and Needs (PATTERN) — on an annual basis. As previous reports detail, the National Institute of Justice (NIJ) contracted with Dr. Grant Duwe, Dr. Zachary Hamilton, and Dr. Alex Kigerl in April 2019 to develop PATTERN. NIJ contracted with Dr. Rhys Hester and Dr. Ryan Labrecque in August 2020 to serve as consultants to conduct the annual review and revalidation of PATTERN. This six-part report documents the activities and accomplishments completed to date regarding PATTERN.

Part 1 summarizes the development and initial review of PATTERN. As documented in NIJ's January 2021 report, discrepancies were identified with some of the measures used to create PATTERN version 1.2. NIJ's review and revalidation expert consultants collaborated with staff from BOP's Office of Research and Evaluation (ORE) to correct the discrepancies. Part 2 describes the study sample and statistical models that were estimated using the updated data to create PATTERN version 1.3. This section includes a summary of the new version 1.3 item weights (Table 2.2), item point assignments (Table 2.3), risk level categories (RLCs; Table 2.4), and descriptive statistics of risk scores and RLCs for the validation and revalidation samples (Table 2.7).

This report also includes the review and revalidation analyses related to the predictive validity (Part 3), dynamic validity (Part 4), and racial and ethnic neutrality (Part 5) of PATTERN 1.3 as required by the FSA. Predictive validity is assessed through Area Under the Curve (AUC) statistics (section 3.2.1), supplemented by risk level recidivism analyses (section 3.2.2) and predictive value analyses for the risk level groupings (section 3.2.3). Results suggest PATTERN 1.3 displays a high level of predictive accuracy, with AUCs ranging from 0.75 to 0.79 across validation and revalidation samples. The recidivism rates by RLC and predictive value analyses indicate that RLCs provide meaningful distinctions of levels of recidivism risk. Individuals in higher RLCs have increasingly higher average recidivism rates.

Dynamic validity is evaluated by assessing the extent to which individual risk scores and levels change from first to last assessment (section 4.2.1) and how these changes relate to recidivism (section 4.2.2). These analyses suggest that individuals are capable of changing risk score and level during confinement (Table 4.1, Figures 4.1 and 4.2), and that these changes relate to recidivism outcomes (Tables 4.2 and 4.3). Individuals who increased their risk scores and levels from first to last assessment were generally more likely to recidivate, whereas those who lowered their risk scores and RLCs were less likely to recidivate.

Racial and ethnic neutrality is examined through a comparison of AUCs (section 5.2.1) and predictive values (section 5.2.2) by race and ethnic group, as well as through differential prediction analyses which assesses a key question: *Do racial and ethnic subgroups have different probabilities of recidivism controlling for PATTERN score?* (section 5.2.3). PATTERN shows relatively high predictive accuracy across all five racial/ethnic groups (Table 5.1). The predictive

value (Table 5.2) and differential prediction results (Tables 5.3 and 5.4, Figure 5.1), however, are mixed and complex. The differential prediction analyses reveal statistically significant results in 28 of 48 tests (analyses of main effects). These include the overprediction of Black, Hispanic, and Asian males and females on some of the general recidivism tools and the underprediction of Black males and females and Native American males, relative to white individuals, on some of the violent recidivism tools. The magnitudes of differential prediction include:

- 6 to 7 percent relative overprediction for Black females on the general recidivism tool
- 12 to 15 percent relative underprediction of Native American males and females on the general recidivism tools
- 5 to 8 percent relative overprediction of Asian males on the general and violent recidivism tools

Finally, the level of differential prediction appeared more pronounced in more recent years, which may suggest a time trend. For example, there were no statistically significant results for Black males on the general recidivism tool in the fiscal year (FY) 2014 to FY 2015 validation sample, growing to a 2 percent relative overprediction in the FY 2016 revalidation sample and a 3 percent overprediction in the FY 2017 revalidation sample. Statistically significant results do not necessarily invalidate a tool, particularly with large sample sizes. However, due to the importance of the FSA mandate to examine the risk and needs assessment system for racial and ethnic neutrality, these results will be a central focus of subsequent review and revalidation efforts.

The NIJ consultants will also continue to investigate potential solutions for the differential prediction issues identified during this review, including testing emerging debiasing techniques and engaging with stakeholders to explore the most promising and supportable approaches. As required by the FSA mandate, a subsequent review and revalidation report including an additional cohort of individuals released from BOP custody in FY 2018 will be released in late 2022.

Part 1: Background

The First Step Act (FSA) of 2018 mandated the attorney general, in consultation with the Independent Review Committee (IRC),[1] to develop and implement a risk and needs assessment system for use with each person in the custody of the Federal Bureau of Prisons (BOP). Section 3634 of Title I of the FSA requires the U.S. Department of Justice (USDOJ or the Department) to review, validate, and release publicly its risk and needs assessment system — the Prisoner Assessment Tool Targeting Estimated Risk and Needs (PATTERN)[2] — on an annual basis. With this report, National Institute of Justice (NIJ) contractors Dr. Rhys Hester and Dr. Ryan Labrecque document the activities and accomplishments completed to date regarding PATTERN.

Part 1 summarizes the development and initial review of PATTERN. Part 2 describes revisions made to the risk instrument that were necessary due to data errors identified during a previous review of PATTERN. Parts 3, 4, and 5 report the results of the predictive validity, dynamic validity, and racial neutrality analyses, respectively, as mandated by sections 3631 and 3634 of Title I of the FSA. Finally, Part 6 reviews the main findings of the current assessment and discusses the next steps for the use and evaluation of PATTERN.

1.1 Development of PATTERN

PATTERN is a risk assessment system designed to periodically assess the risk for recidivism among individuals in BOP custody. PATTERN was initially developed and validated using a dataset of individuals released from BOP custody in fiscal year (FY) 2009 through FY 2015. The sample was split into a training set (individuals released between FY 2009 and FY 2013) and a validation sample (individuals released in FY 2014 and FY 2015). The developers of PATTERN — Dr. Grant Duwe, Dr. Zachary Hamilton, and Dr. Alex Kigerl — employed boosted regression to identify the best combination of variables for predicting general and violent recidivism with the training sample (see USDOJ, 2019, pp. 49-50). This process identified 17 measures across four tools: (1) general recidivism for males, (2) violent recidivism for males, (3) general recidivism for females, and (4) violent recidivism for females.[3] When evaluated on the validation sample, the four PATTERN tools were shown to "achieve a high level of predictive performance, surpassing what commonly is found for risk assessment tools for correctional populations in the United States" (USDOJ, 2019, p. 63).[4]

---

[1] The IRC is a statutorily mandated body of at least six experts on risk and needs systems. The IRC is charged with assisting the attorney general in carrying out the responsibilities of the FSA regarding the risk and needs assessment system.

[2] PATTERN is a risk assessment tool. For more information on the separate needs assessment system developed in response to the FSA, see BOP (2019).

[3] These included: age at first conviction; age at time of assessment; infraction convictions (any); infraction convictions (serious and violent); number of programs completed (any); number of technical or vocational courses; federal industry employment (UNICOR); drug treatment while incarcerated; drug education while incarcerated; noncompliance with financial responsibility; instant offense violent; sex offender (Walsh); and Bureau Risk and Verification Observation (BRAVO) initial criminal history score, history of violence, history of escapes, voluntary surrender, and initial education score.

[4] More specifically, the general male scale had an Area Under the Curve (AUC) value of 0.80, the violent male scale had an AUC value of 0.78, the general female scale had an AUC value of 0.79, and the violent female scale had an AUC value of 0.77.

Following the release of the July 2019 report, USDOJ solicited public comment and NIJ facilitated listening lessons with criminal justice stakeholders, advocates, and interested citizens to encourage a wide range of perspectives on PATTERN (see USDOJ, 2020a). Based on feedback received, several changes were made to PATTERN, including the addition of more dynamic — or changeable — factors (duration of time since last incident report and duration of time since last serious incident report) and the removal of two variables believed to exacerbate racial disparities (age at first arrest and voluntary surrender). Other changes included counting more types of treatment services in the programming measure (Adult Continuing Education [ACE]; Bureau Rehabilitation And Values Enhancement [BRAVE]; Challenge; Drug Education; Life Connections; Parenting; Skills; Sex Offender Residential Treatment; Sex Offender Non-Residential Treatment; Steps Toward Awareness, Growth, and Emotional Strength [STAGES]; and Step Down) and combining the technical and vocational courses variable with the federal industry employment variable (UNICOR) to create a single measure of work programming. Additionally, the drug treatment and drug education variables were altered.[5] These modifications to PATTERN also resulted in changes in the scoring criteria of the items[6] and cut points for the risk level categories (RLCs).[7]

The revised version of PATTERN (i.e., version 1.2) maintained a high level of predictive validity (see USDOJ, 2020a).[8] Version 1.2 was approved for use by the attorney general in consultation with the IRC. BOP developed a PATTERN field manual and risk score form, and in late January 2020, BOP began scoring the individuals in its custody with PATTERN.

1.2 Initial Review of PATTERN

To help fulfill the Title I requirements of the FSA, NIJ announced a competitive Consultant Statement of Work and selected two consultants to conduct the annual review and revalidation of PATTERN. NIJ contracted with Dr. Rhys Hester and Dr. Ryan Labrecque to review and revalidate PATTERN. In addition, Dr. Robert Morris was added to the project as a statistical consultant in spring 2021. Following their initial review of the PATTERN developmental datasets, syntax files, and other supporting documentation, the consultants identified several coding, specification, and scoring discrepancies.[9] After consulting with staff from BOP's Office of Research and Evaluation (ORE) and members of the development team about these discrepancies, the four PATTERN regression models were reestimated using the corrected versions of the variables to provisionally assess the magnitude of these issues on the performance of the tool. Results of these analyses

[5] Initially, PATTERN included two separate substance abuse measures: drug treatment while incarcerated (i.e., no need indicated, completed residential drug treatment during incarceration, completed drug treatment during incarceration, and need indicated but no treatment during incarceration), and drug education while incarcerated (i.e., yes or no). In PATTERN 1.2, these two variables were combined to create one measure of drug treatment while incarcerated (i.e., need indicated/no completion, completed nonresidential drug treatment, completed residential drug treatment, and no need indicated).

[6] For a full description of the differences between the initial and revised PATTERN, compare pages 53-56 in USDOJ (2019) to pages 37-39 in USDOJ (2020a).

[7] See also USDOJ (2020b).

[8] The general male scale had an AUC value of 0.79, the violent male scale had an AUC value of 0.78, the general female scale had an AUC value of 0.78, and the violent female scale had an AUC value of 0.77.

[9] For a detailed review, see NIJ (2021).

indicated that the coefficient weights differed across 37 of the 60 item possibilities when compared to version 1.2, which meant that the scoring weights assigned to nearly two-thirds of these items were different. Using the values of the coefficient weights from the corrected models, a provisional version 1.3 was created to assess the potential impact of the discrepancies. The total risk scores were then converted into the four RLCs — minimum, low, medium, and high — by replicating as closely as possible the BOP population distributions and recidivism rates of version 1.2.

Using data on the population of individuals who were incarcerated on November 28, 2020, the consultants and staff from BOP ORE independently calculated the risk scores and RLCs for PATTERN version 1.2 and provisional version 1.3. After both research teams obtained the same results, the RLCs derived from version 1.2 were compared to those from provisional version 1.3. Findings from this analysis revealed that approximately 10.9 percent of males and 9.8 percent of females were categorized into different overall PATTERN risk levels between the two versions. In determining one's overall category of risk, BOP takes the highest risk level from the general or violent instruments and then aggregates minimum and low risk into a lower risk level group and medium and high risk into a higher risk level group.[10] This distinction is important, as individuals in the lower risk group are eligible to receive earned time credits, whereas those in the higher risk group are not. For the November 2020 stock population analysis, about 4.2 percent of males and 4.5 percent of females were classified differently across the risk groups, which would influence the amount of earned time credits these individuals were eligible to receive. The impact of their assigned risk level group on their ability to receive earned time credit during this time, however, was mitigated because the COVID-19 pandemic impacted BOP's delivery of programming nationally.

Following this discovery, BOP updated its field guidance and scoring sheets with revisions made to the scoring typos published in the USDOJ 2020 report.[11] In addition, BOP reassessed 1,745 individuals whose RLCs were classified differently because of the scoring typos. BOP now uses a revised version of PATTERN 1.2 to assess risk for recidivism (i.e., version 1.2-R). These revisions addressed the typos but did not implement any new item weights.[12] A decision was made not to make any additional changes to PATTERN until a finalized 1.3 revision could be constructed with input from the IRC and other stakeholders and submitted to the attorney general for review and approval — a process that represents the purpose of the current report (see also NIJ, 2021).

---

[10] In practice this operationalization affects only a small percentage of individuals who are designated as minimum or low risk on the general tool but are medium or high risk on the violent tool, and thus are classified into the aggregated medium- or high-risk groupings. For example, among the 63,848 males in the FY 2014-FY 2015 validation sample, only 694 (or 1.1 percent) were classified in the higher risk group due to having a minimum/low general recidivism classification but a medium/high violent recidivism classification. For the 10,527 females in the sample, only 37 (or 0.3 percent) were similarly classified.

[11] It should be noted that the coding discrepancies detected were limited to the developmental analysis of the data, which was published in USDOJ (2020) and Hamilton et al. (2021). BOP, however, has been scoring these items correctly in practice all along.

[12] There were four corrections made to the BOP manual and scoring sheets. First, the point values for the time since last serious incident report item for the violent male risk instrument were changed from 0, 2, 4, and 6 to 0, 1, 2, and 3. Second, the point values for the time since last serious incident report item for the general female risk instrument were changed from 0, 1, 2, and 3 to 0, 2, 4, and 6. Third, the values for the criminal history points item for the violent female risk scale were changed from 0, 4, 8, 12, 16, and 20 to 0, 2, 4, 6, 8, and 10. Fourth, the cut points for the general male tool were changed from < 11 for minimum risk and 11-30 for low risk to < 9 for minimum risk and 9-30 for low risk.

To investigate reliability in scoring, BOP ORE conducted an internal analysis to assess the extent to which the RLCs generated from their PATTERN simulator program[13] matched those recorded by staff conducting the assessments by hand. Using the BOP in-custody population data from November 28, 2020, this analysis revealed substantial discrepancies: More than 20 percent of individuals in custody had a different risk level identified by the PATTERN simulator than the score recorded by BOP staff. These findings were replicated by the NIJ consultants. Staff completing the PATTERN assessments by hand digitally recorded the final RLC but not the individual item ratings or overall assessment score, which would have allowed for a more granular analysis of the scoring discrepancies. Nevertheless, ORE obtained a sample of PATTERN paper files from BOP regional offices to investigate the potential sources of the mismatches. Through this process, ORE discovered that the scoring discrepancies were more common in more complex items such as the incident report and programming variables but were also present less frequently across more straightforward items such as prior Walsh conviction.

BOP fully automated PATTERN in August 2021. The assessment scores are now automatically generated via a software application that leverages available information from the appropriate internal data sources, thereby eliminating reliance on staff to complete a form by hand. Automation has not only improved scoring reliability, but also reduced staff labor costs and increased the speed and efficiency of the assessments. As part of this development process, staff from ORE and the BOP Office of Information Technology have collaborated to determine from which specific data sources the item information should be retrieved to calculate the PATTERN scores.[14] All PATTERN scores will be updated using the automated tool at the review regularly scheduled for everyone in BOP custody.

Given these developments and the need to finalize a revised version of PATTERN with updated item weights, staff from ORE reconstructed the PATTERN developmental dataset in the spring of 2021, a process that included optimizing variable sourcing within BOP data systems. All changes were meant to improve the reliability and accuracy of the information used to score PATTERN. In the current report, the updated developmental training data — which include information on individuals released from FY 2009 through FY 2013 — are used to reestimate the PATTERN item weighting scheme using the same methodological approach employed in the initial development process. In addition, data from individuals released in FY 2014 and FY 2015 are used to validate PATTERN, and data from two new cohorts of individuals released from BOP custody during FY 2016 and FY 2017 are used to revalidate PATTERN as required by the FSA.

[13] The simulator program is a SAS syntax file developed by staff from ORE that uses historical administrative data to score PATTERN in an automated fashion.

[14] During the automation process, changes were made to the sources for some of the data elements previously used in the PATTERN simulator. In collaboration, the BOP Office of Information Technology's automation and ORE's simulator were both given minor updates to synchronize with each other so that the same results were obtained using two different programming languages.

Part 2: PATTERN 1.3 Revisions

This section discusses the revisions that were made to PATTERN in the current report. Section 2.1 describes the updated data and samples used in this investigation. Section 2.2 reports the analytic strategy for reconstructing PATTERN. Section 2.3 presents the new item weights (section 2.3.1), item point assignments (2.3.2), and cutoffs for RLCs for PATTERN 1.3 (section 2.3.3), compares the RLC population distributions and recidivism rates between versions 1.2 and 1.3 (section 2.3.4), and summarizes the risk scores and levels by assessment type for PATTERN 1.3 (section 2.3.5).

2.1 Data and Samples

The updated developmental dataset for the current analysis included 283,139 individuals who were released to the community between FY 2009 and FY 2015 (i.e., the developmental sample).[15] Following the development team's approach, individuals were only eligible for inclusion if their initial Bureau Risk and Verification Observation (BRAVO) assessment was available under the current version and policy (i.e., BOP Program Statement 5100.08; see BOP, 2006).[16] Additionally, individuals who were known to have died during the three-year follow-up period were excluded (n = 42).[17] A total of 224,967 individuals met the inclusion criteria: 150,592 who were released between FY 2009 and FY 2013 (i.e., the training sample) and 74,375 who were released between FY 2014 and FY 2015 (i.e., the validation sample). Given the statutory mandate to conduct annual revalidation analyses, BOP also provided two additional years of release data (FY 2016 and FY 2017). Of the 47,063 individuals released from BOP custody in FY 2016, 42,353 met the study inclusion criteria (i.e., the FY 2016 revalidation sample); 38,333 of 41,096 individuals released in FY 2017 also met the study inclusion criteria (i.e., the FY 2017 revalidation sample).

[15] Individuals who were scheduled for deportation upon release were not included in this dataset. The sample size is different from the 280,588 individuals in the original PATTERN developmental database for two reasons. First, a few individuals subsequently entered witness protection, so their records were dropped from all research data extracts. Second, whereas the previous data file included only the first release for individuals who were discharged from prison multiple times in the same fiscal year, the current dataset includes all releases regardless of timing within the fiscal years.

[16] BRAVO is known within BOP as the classification system; it is a risk assessment system designed by BOP to predict serious misconduct in prison. This eligibility criterion was necessary because several of the PATTERN items were taken from BRAVO. This criterion mostly excluded individuals who were admitted to BOP custody prior to September 2006. This exclusion was necessary because four of the items in BRAVO (criminal history points, history of escapes, history of violence, and drug program status) are used to score PATTERN, and the old version of those items may not be comparable to the current version. This represents the most recent major change in BOP's data-collection process, so available data should have consistent meanings for all inmates after September 2006. See Program Statement 5100.8, Inmate Security Designation and Custody Classification, for more information. As a supplemental sensitivity analysis, results were examined using the closest proxies available for instrument scores on those individuals falling under earlier BRAVO policy versions. These results showed that for the FY 2016 release cohort, approximately 11 percent of the sample was excluded due to the policy selection criteria, and that the proxy results for these individuals were substantially similar to the results for the selection sample. For instance, the AUCs for the nonincluded subsample were all within 0.01 of the results from the selection sample, suggesting that PATTERN predicts similarly for these out-of-sample individuals.

[17] Although the USDOJ (2019, pp. 42-43) report indicated that individuals who died during the follow-up period were excluded, they were inadvertently retained in the analyzed sample.

The dataset included demographic information on gender (i.e., male, female), race/ethnicity (i.e., white, Black, Hispanic, Asian, Native American),[18] criminal history, education, programming, institutional behavior, and post-release recidivism. General recidivism is defined as a return to BOP custody or a rearrest within three years of release from BOP custody, excluding all traffic offenses except driving under the influence and driving while intoxicated. Violent recidivism is defined as a rearrest for a suspected act of violence within three years of release from BOP custody. The dataset also included information necessary to retrospectively score an initial (at intake) and final (prior to release) PATTERN assessment:[19]

- **Current age.** The number of years between the assessment date and the individual's date of birth, rounded down. This variable is then converted into six ordinal categories: *25 and younger*, *26 to 29*, *30 to 40*, *41 to 50*, *51 to 60*, or *61 and older*.

- **Walsh with conviction.** An identification as a sex offender based on the Adam Walsh Act criteria.[20]

- **Violent offense**. A current conviction for a violent offense, including but not limited to firearms violations, homicide, child abuse, robbery, sex trafficking, and sexual assault.[21]

- **Criminal history points**. The number of criminal history points taken from the most recent BRAVO available. This variable is then converted into six ordinal categories: *0 to 1 point*, *2 to 3 points*, *4 to 6 points*, *7 to 9 points*, *10 to 12 points*, or *13 or more points*.

- **History of escapes**. The number of years from last escape attempt by seriousness taken from the most recent BRAVO available. This variable is then converted into four ordinal categories: *None*, *greater than 10 years minor*, *5 to 10 years minor*, or *less than 5 years minor or any serious*.

- **History of violence**. The number of years from last act of violence by seriousness taken from the most recent BRAVO available. This variable is then converted into eight ordinal categories: *None*, *greater than 10 years minor*, *greater than 15 years serious*, *5 to 10 years minor*, *10 to 15 years serious*, *less than 5 years minor*, *5 to 10 years serious*, or *less than 5 years serious*.

- **Education status**. The highest grade level completed.[22] This variable is then converted into three ordinal categories: *High school degree or GED*, *enrolled and progressing in GED program*, or *no verified degree and not participating in GED program*.

[18] BOP data included race categories for white, Black, Asian, and Native American and an ethnicity indicator for Hispanic or non-Hispanic. Following the development team's operationalization, a race/ethnicity measure was created by classifying all ethnic Hispanic individuals as Hispanic regardless of race category; the four race categories are indicative of a non-Hispanic race.

[19] The presentation order of these variables has been modified from the prior reports to reflect the item names and order used by BOP to score PATTERN.

[20] See 34 U.S.C. § 20911, *et seq*.

[21] See BOP (2020).

[22] In the previous version of PATTERN, the education status variable was taken from the most recent BRAVO assessment available. The current version of this item comes from SENTRY because ORE and the BOP Office of

- **Drug program status.** This measure combines two sources of information: (1) Identification of a substance abuse problem from the most recent BRAVO available and (2) completion of residential or nonresidential drug programming during the current incarceration.[23] This variable is then converted into four ordinal categories: *No drug need indicated*, *completed residential drug treatment*, *completed nonresidential drug treatment*, or *need indicated but no drug treatment completed*.

- **All incident reports.** The number of guilty incident reports[24] within the past 120 months following one's incarceration date. This does not include incident reports occurring during pretrial, holdover, or from prior BOP incarcerations. The variable is then converted into four ordinal categories: *No incident*, *1 incident*, *2 incidents*, or *3 or more incidents*.

- **Serious incident reports.** The number of guilty serious and violent incident reports[25] within the past 120 months following one's incarceration date. This does not include incident reports occurring during pretrial, holdover, or from prior BOP incarcerations. The number of incidents is then converted into four ordinal categories: *No incident*, *1 incident*, *2 incidents*, or *3 or more incidents*.

- **Time since last incident report.** The number of months between the assessment date and the date of the most recent incident report, rounded down. Only incidents from the current incarceration are counted. This variable is then converted into four ordinal categories: *12+ months or no incident*, *7 to 12 months*, *3 to 6 months*, or *less than 3 months*.[26]

- **Time since last serious incident report.** The number of months between the assessment date and the date of most the recent serious or violent incident report, rounded down. Only incidents from the current incarceration are counted. This variable is then converted into four ordinal categories: *12+ months or no incident*, *7 to 12 months*, *3 to 6 months*, or *less than 3 months*.

- **Financial responsibility refuse.** Noncompliance with financial responsibility during incarceration for payment toward victim restitution and dependents.

Information Technology determined that this data source contained more up-to-date educational information. Whereas BRAVO information can be up to a year old between assessments, SENTRY provides information on GED attendance and completion on a daily basis.

[23] This measure does not include all of the evidence-based recidivism reduction drug programs and other drug-related productive activities currently available throughout BOP (see https://www.bop.gov/inmates/fsa/docs/ebrr_programs.pdf). These data were not available during the study observation period.

[24] This includes only incident reports, not acts. For example, if an incident report included multiple acts occurring at the same time (e.g., serious assault and possession of a weapon), it would only be counted as one incident, not two.

[25] This includes 100- and 200-level offenses that represent the most serious prohibited acts, such as killing, serious assault, arson, weapon possession, rioting, fighting, threatening, extortion, and drug-related infractions (for more information see BOP, 2011).

[26] More precisely, the four infraction-free categories are operationalized as 0 to 91 days, 92 to 212 days, 213 to 365 days, and 366 days or more.

- **Programs completed.** The number of ACE, BRAVE, Challenge, Drug Education, Life Connections, Parenting, Skills, Sex Offender Residential Treatment, Sex Offender Non-Residential Treatment, STAGES, and Step Down courses successfully completed during the current incarceration.[27] This variable is then converted into five ordinal categories: *No program*, *1 program*, *2 to 3 programs*, *4 to 10 programs*, or *11 or more programs*.

- **Work programs completed**. The number of technical and vocational courses completed during the current incarceration. In this measure, federal industry employment (UNICOR) is counted as a program completion if the individual worked at least one day. This variable is then converted into three ordinal categories: *No program*, *1 program*, or *2 or more programs*.

Table 2.1 summarizes the descriptive statistics for the total sample using the updated BOP data file. In comparison to the descriptive statistics reported in Hamilton et al. (2021), the two samples are generally similar across the variables examined, but there are several notable differences.[28] There are two sources for these discrepancies. First, the coding and specification errors documented in the NIJ (2021) report have been corrected in the current report.[29] Second, BOP has updated the data sources across several measures used in the current report to produce the most reliable and accurate data. Of note, there were fewer individuals identified as having a Walsh conviction in the current report than in the earlier report (7.5 vs. 10 percent). The distributions of the drug program status measures also varied. For instance, 7 percent of individuals were identified as completing either residential or nonresidential drug treatment in the current report compared to 21 percent in the earlier report.[30] There were also fewer incident reports (28.5 vs. 37 percent), serious incident reports (13.8 vs. 18 percent), incidents within the past 12 months (20.6 vs. 24 percent), and serious incidents within the last 12 months (8.9 vs. 11 percent) in the current report than in the earlier report. Finally, although the measures of programs and work programs

[27] This measure does not include all of the evidence-based recidivism reduction programs and productive activities currently available in BOP (see https://www.bop.gov/inmates/fsa/docs/ebrr_programs.pdf), as the data for these program were not available during the study observation period. Additionally, some of the programs currently included in this variable, such as ACE, are not considered evidence-based recidivism reduction programs or productive activities by the BOP policy, though they predictively correlate with recidivism.

[28] Hamilton et al. (2021) presents additional descriptive tables that were not published in the USDOJ (2020a) report. For more specific details about how the characteristics of the samples differ, compare Table 2.1 in this report to Table 1 in Hamilton et al. (2021, pp. 10-11).

[29] In the initial development of PATTERN, the programs completed variable counted program participation rather than program completion. The work programs completed variable also counted each day of participation in UNICOR as the completion of a separate work program. The drug program status variable counted individuals who completed both residential and nonresidential programming as completing nonresidential, whereas the developers intended to count the individual as completing the higher value residential programming. The all incident reports and all serious incident reports measures counted incidents even if they occurred beyond the 120-month window. The current age, criminal history points, history of escapes, history of violence, drug program status, all incident reports, serious incident reports, time since last incident report, time since last serious incident report, financial responsibility refuse, programs completed, and work programs completed variables were operationalized using the individual's release date, and the education status variable was operationalized using the initial assessment date rather than the current assessment date, which is done in practice. For more information see NIJ (2021, pp. 5-7).

[30] It is important to note that the drug treatment programs tend to be delivered just prior to release. As such, the current report reflects the percentage of individuals who completed programs at the time of their last assessment, whereas the prior report reflects the percentage of individuals who completed programs at the time of release.

completed were close, there were fewer individuals identified as having completed two or more work programs in the current report relative to the earlier report (4.6 vs. 8 percent).[31]

Table 2.1 also compares the summary statistics for the training, validation, and two revalidation samples. The four samples are similar across most metrics. They substantively differ, however, on several variables. Relative to the FY 2009 to FY 2013 training sample, individuals in the FY 2014 to FY 2015 validation sample and the FY 2016 and FY 2017 revalidation samples are more likely to have a high school degree or GED (by 6.1, 8.5, and 8.2 percent, respectively), more likely to have completed residential or nonresidential drug treatment (by 3.7, 5.4, and 4.7 percent), more likely to have one or more incident reports (by 6.3, 10.6, and 10.4 percent), more likely to have one or more serious incident reports (by 5.4, 9.2, and 9.3 percent), more likely to have completed one or more programs (by 7.5, 11.9, and 8.4 percent), and more likely to have completed one or more work programs (by 5.8, 9.5, and 8.6 percent).

[31] Although there also appears to be a difference in the history of escapes measure, this is due to a typo in the Hamilton et al. (2021) table, which has percentage totals that sum to 86 rather than 100.

Table 2.1. ***Descriptive statistics for the developmental, training, validation, and revalidation samples***

Measure		FY09-15 Developmental Sample % (*n* = 224,967)	FY09-13 Training Sample % (*n* = 150,592)	FY14-15 Validation Sample % (*n* = 74,375)	FY16 Revalidation Sample % (*n* = 42,353)	FY17 Revalidation Sample % (*n* = 38,333)
Gender	*Male*	85.1	84.8	85.8	86.8	86.5
	Female	14.9	15.2	14.2	13.2	13.5
Race/ethnicity	*White*	35.6	35.7	35.4	34.4	33.7
	Black	38.6	39.0	37.7	38.3	38.5
	Hispanic	20.2	19.6	21.4	22.0	22.4
	Asian	2.0	2.0	1.9	1.9	1.7
	Native American	3.6	3.6	3.5	3.4	3.7
Current age	*> 60*	3.7	3.5	4.2	4.4	4.6
	51-60	10.2	9.9	10.9	10.9	11.4
	41-50	20.8	20.6	21.2	22.1	22.3
	30-40	37.5	36.7	39.3	40.6	39.9
	26-29	14.4	15.1	13.0	12.6	12.6
	< 26	13.3	14.3	11.5	9.4	9.2
Walsh with conviction	*No*	92.5	93.1	91.3	91.4	89.8
	Yes	7.5	6.9	8.7	8.6	10.2
Violent offense	*No*	73.7	73.9	73.4	73.1	71.3
	Yes	26.3	26.1	26.6	26.9	28.7
Criminal history points	*0-1*	33.8	34.1	33.0	31.0	29.8
	2-3	13.3	13.2	13.4	13.8	13.5
	4-6	17.5	17.5	17.3	18.0	17.6
	7-9	12.6	12.5	12.7	13.2	13.3
	10-12	9.2	9.1	9.4	9.8	10.3
	13+	13.7	13.4	14.1	14.2	15.6
History of escapes	*None*	86.3	86.5	85.9	85.4	84.2
	> 10 years minor	4.5	4.2	5.0	5.6	6.2
	5-10 years minor	3.4	3.3	3.5	3.2	3.3
	< 5 years minor or any serious	5.8	5.9	5.6	5.7	6.3
History of violence	*None*	55.8	56.9	53.7	51.7	50.0
	> 10 years minor	4.9	4.7	5.3	5.8	5.5
	> 15 years serious	6.5	6.2	7.2	8.1	9.1
	5-10 years minor	5.4	5.4	5.3	5.4	5.3
	10-15 years serious	5.7	5.5	6.1	6.6	6.7
	< 5 years minor	8.5	8.4	8.8	9.3	10.0
	5-10 serious	8.0	7.8	8.4	8.5	8.8
	< 5 years serious	5.2	5.3	5.1	4.5	4.8

Education status	*Not enrolled*	23.4	24.9	20.2	18.0	19.0
	Enrolled in GED	12.1	12.6	11.1	11.0	10.3
	High school degree/GED	64.5	62.5	68.6	71.0	70.7
Drug program status	*Need indicated/No completion*	67.5	69.6	63.2	58.5	59.2
	Completed nonresidential drug treatment	4.6	3.9	6.0	6.7	5.9
	Completed residential drug treatment	2.4	1.8	3.4	4.4	4.5
	No need indicated	25.6	24.7	27.4	30.4	30.4
All incident reports	*0*	71.5	73.6	67.3	63.0	63.7
	1	15.1	14.6	16.1	17.1	16.2
	2	6.0	5.6	7.0	7.6	7.4
	3+	7.3	6.2	9.6	12.3	12.8
Serious incident reports	*0*	86.2	88.0	82.6	78.8	78.7
	1	9.4	8.5	11.3	13.2	12.8
	2	2.4	2.0	3.3	4.0	4.2
	3+	2.0	1.5	2.8	4.1	4.3
Time since last incident report	*12+ months*	79.4	79.6	78.9	78.0	77.2
	7-12 months	3.0	2.9	3.3	3.8	3.3
	3-6 months	4.7	4.7	4.7	4.5	4.4
	< 3 months	12.9	12.9	13.1	13.7	15.0
Time since last serious incident report	*12+ months*	91.1	91.6	90.2	89.3	88.6
	7-12 months	1.6	1.5	1.9	2.2	2.1
	3-6 months	2.4	2.4	2.6	2.5	2.6
	< 3 months	4.8	4.5	5.4	5.9	6.6
Financial responsibility refuse	*No*	95.9	95.7	96.3	96.6	96.3
	Yes	4.1	4.3	3.7	3.4	3.7
Programs completed	*0*	51.4	53.9	46.4	42.0	45.5
	1	19.5	20.0	18.5	17.4	16.5
	2-3	14.8	14.2	15.9	16.3	14.8
	4-10	11.8	10.1	15.2	18.5	17.5
	11+	2.4	1.7	3.9	5.7	5.7
Work programs completed	*0*	84.6	86.5	80.7	77.0	77.9
	1	10.9	9.9	12.8	14.4	14.2
	2+	4.6	3.6	6.6	8.5	7.9
Recidivism	*General*	47.0	47.7	45.4	44.5	44.6
	Violent	15.1	15.2	15.0	14.6	15.9

Note: Variable percentages do not all sum to 100.0 due to rounding.

2.2 Analytical Plan

This section details the reestimation and reconstruction of PATTERN with the updated data using the methodological approach previously employed. This procedure was necessary because the data previously used in the empirical models to develop version 1.2 contained discrepancies which have been corrected.[32] These data changes influence the coefficients returned from the regression analyses, which are used to calculate the item scoring weights. The results of the models presented in this report draw on the updated data to provide a revised scoring scheme (i.e., version 1.3).

The first task involved rerunning, with the updated data, the same machine learning boosted regression procedures used to develop PATTERN.[33] The coefficients derived from these analyses conducted on the training sample were used to construct the item weights across the four PATTERN tools (i.e., general male, violent male, general female, and violent female). The total risk scores were then converted into the four RLCs (i.e., minimum, low, medium, and high) by replicating as closely as possible the population distributions and recidivism rates as approved by the attorney general for version 1.2.[34]

2.3 Results

2.3.1 Boosted Regression Analyses

The revised version of PATTERN was developed using the updated data from the FY 2009 to FY 2013 training sample (N = 150,592).[35] Table 2.2 presents the items and coefficients from the boosted regression analyses, where spaces with hyphens indicate factors that were not retained in the models.[36] The retained coefficients ranged in value from 0.001 to 0.08.[37] Following the developmental team's methodological approach, the coefficients were then used as multipliers to

[32] For more information, see NIJ (2021).
[33] For more information on the boosted regression method, see USDOJ (2019, pp. 49-53).
[34] See Hamilton et al. (2021, p. 15).
[35] The data used to develop PATTERN include predictor information that was available at the final assessment, but no later.
[36] The boosted regression procedure uses a meta-algorithm that estimates multiple iterations of models, ultimately retaining items that contribute to prediction and dropping items that do not (see USDOJ, 2019). In comparison to version 1.2, there are some differences in the items that were retained in version 1.3: The general male model excluded time since last serious incident report; the violent male model excluded time since last serious incident report and financial responsibility refuse; the general female model included violent offense and excluded serious incident reports; and the violent female model included financial responsibility refuse and excluded all incident reports and serious incident reports.
[37] Although the coefficient values for the criminal history points item appear larger in magnitude in the current version relative to version 1.2, it should be noted that Hamilton et al. (2021) coded this variable as 0 = 0-1 points, 2 = 2-3 points, 4 = 4-6 points, 6 = 7-9 points, 8 = 10-12 points, and 10 = 13 or more points, which is consistent with how BRAVO codes the item, whereas to be more consistent with the other PATTERN items we coded the variable as 0 = 0-1 points, 1 = 2-3 points, 2 = 4-6 points, 3 = 7-9 points, 4 = 10-12 points, and 5 = 13 or more points. As such, while the coefficients in version 1.3 are double in size compared to version 1.2, the scoring of these items in practice is the same. To illustrate, a male scored on the general recidivism model with two criminal history points would be scored in version 1.2 as $2 \times 4 = 8$, whereas the same individual would be scored in version 1.3 as $1 \times 8 = 8$.

determine the item scores in PATTERN.[38] Correlations between 0.001 and 0.005 in value were rounded up to a multiplier of 0.01 in order to retain the variable in the assessment.

Table 2.2. ***PATTERN version 1.3 coefficient weights by assessment type***

Item	General Male	Violent Male	General Female	Violent Female
1. Current age	0.07	0.04	0.06	0.01
2. Walsh with conviction	0.02	-	-	-
3. Violent offense	0.05	0.07	0.01	0.03
4. Criminal history points	0.08	0.03	0.08	0.01
5. History of escapes	0.03	0.02	0.03	0.01
6. History of violence	0.01	0.02	0.01	0.01
7. Education status	0.01	0.004	0.03	0.01
8. Drug program status	0.02	0.003	0.03	0.004
9. All incident reports	0.01	0.001	0.002	-
10. Serious incident reports	0.01	0.01	-	-
11. Time since last incident report	0.01	0.01	0.02	0.004
12. Time since last serious incident report	-	-	0.002	0.01
13. Financial responsibility refuse	0.02	-	0.03	0.01
14. Programs completed	0.03	0.01	0.02	0.002
15. Work programs completed	0.01	0.005	0.02	0.003

Note: Male training sample $N = 127{,}683$; female training sample $N = 22{,}909$.

2.3.2 PATTERN Item Point Assignments

Table 2.3 presents the item point assignments across all four PATTERN tools. As an example, the general male model assigns 0 points for individuals who are over 60 years old at the time of assessment, 7 for those 51-60 years old, 14 for those 41-50 years old, 21 for those 30-40 years old, 28 for those 26-29 years old, and 35 for those under 26 years old. Total PATTERN scores are then computed by summing the values across all the included items.

Table 2.3. ***Revised points assigned in the PATTERN 1.3 risk assessment models***

Item	Category	General Male	Violent Male	General Female	Violent Female
1. Current age	*> 60*	0	0	0	0
	51-60	7	4	6	1
	41-50	14	8	12	2
	30-40	21	12	18	3
	26-29	28	16	24	4
	< 26	35	20	30	5
2. Walsh with conviction	*No*	0			
	Yes	2			
3. Violent offense	*No*	0	0	0	0
	Yes	5	7	1	3
4. Criminal history points	*0-1*	0	0	0	0

[38] Two items were included in version 1.2 despite not being retained in the final analytical models. These were the time since last serious incident report item in the general male model and the financial responsibility refuse item in the violent male model. These variables are both listed as possessing a coefficient value of 0.01 in Table 2 of Hamilton et al. (2021, p. 15). There are another seven coefficients in this table that are presented as 0.01 despite being less than 0.005 in value. These include the all incident reports and time since last serious incident report items in the violent male model and the all incident reports, time since last incident report, programs completed, work programs completed, and drug program status items in the violent female model.

	2-3	8	3	8	1
	4-6	16	6	16	2
	7-9	24	9	24	3
	10-12	32	12	32	4
	13+	40	15	40	5
5. History of escapes	*None*	0	0	0	0
	> 10 years minor	3	2	3	1
	5-10 years minor	6	4	6	2
	< 5 years minor or any serious	9	6	9	3
6. History of violence	*None*	0	0	0	0
	> 10 years minor	1	2	1	1
	> 15 years serious	2	4	2	2
	5-10 years minor	3	6	3	3
	10-15 years serious	4	8	4	4
	< 5 years minor	5	10	5	5
	5-10 serious	6	12	6	6
	< 5 years serious	7	14	7	7
7. Education status	*Not enrolled*	0	0	0	0
	Enrolled in GED	-1	-1	-3	-1
	High school degree/GED	-2	-2	-6	-2
8. Drug program status	*Need indicated/No completion*	0	0	0	0
	Completed nonresidential drug treatment	-2	-1	-3	-1
	Completed residential drug treatment	-4	-2	-6	-2
	No need indicated	-6	-3	-9	-3
9. All incident reports	*0*	0	0	0	
	1	1	1	1	
	2	2	2	2	
	3+	3	3	3	
10. Serious incident reports	*0*	0	0		
	1	1	1		
	2	2	2		
	3+	3	3		
11. Time since last incident report	*12+ months*	0	0	0	0
	7-12 months	1	1	2	1
	3-6 months	2	2	4	2
	< 3 months	3	3	6	3
12. Time since last serious incident report	*12+ months*			0	0
	7-12 months			1	1
	3-6 months			2	2
	< 3 months			3	3
13. Financial responsibility refuse	*No*	0		0	0
	Yes	2		3	1
14. Programs completed	*0*	0	0	0	0
	1	-3	-1	-2	-1
	2-3	-6	-2	-4	-2
	4-10	-9	-3	-6	-3
	11+	-12	-4	-8	-4
15. Work programs completed	*0*	0	0	0	0
	1	-1	-1	-2	-1
	2+	-2	-2	-4	-2

2.3.3 Risk Level Category Cutoffs

The total number of possible points varies across the four PATTERN tools. The range of scores is 132 (-22 to 109) for the 14 items in the general male tool, 83 (-11 to 71) for the 12 items in the violent male tool, 130 (-27 to 102) for the 13 items in the general female tool, and 42 (-11 to 30) for the 12 items in the violent female tool. The risk level cut points were determined by replicating as closely as possible the population distributions and recidivism rates in version 1.2.[39]

Table 2.4 displays the range of scores that fall into the four RLCs of each assessment type. For example, males rated on the general recidivism tool who possess scores between -22 and 11 are considered minimum risk, 12 to 32 — low risk, 33 to 44 — medium risk, and 45 to 109 — high risk.

Table 2.4. ***PATTERN version 1.3 risk score and category information***

Risk Level	General Male	Violent Male	General Female	Violent Female
	(14 items)	**(12 items)**	**(13 items)**	**(12 items)**
Minimum	-22 to 11	-11 to 7	-27 to 10	-11 to 1
Low	12 to 32	8 to 24	11 to 34	2 to 11
Medium	33 to 44	25 to 31	35 to 51	12 to 17
High	45 to 109	32 to 71	52 to 102	18 to 30

2.3.4 Risk Level Population Distribution and Recidivism Rates

Tables 2.5 and 2.6 compare the population distributions and recidivism rates across the four RLCs for the general and violent risk scales between version 1.2 and 1.3 with the training sample. The cut points replicate the distributions and rates approved by the attorney general for use in version 1.2. Although most comparisons represent a substantively similar result (or a 1 to 3 percent difference), there were two exceptions. First, the violent recidivism rates for the high-risk group in the total, male, and female samples were all 31 percent in version 1.2, compared to 35, 35, and 42 percent respectively in version 1.3. Second, whereas medium-risk females had a 23 percent violent recidivism rate in version 1.2, the same category had an 18 percent violent recidivism rate in version 1.3.

Table 2.5. ***Comparison of general PATTERN RLC BOP population distributions and recidivism rates between versions 1.2 and 1.3, training sample (FY09-13)***

	Population Distributions (%)		Recidivism Rates (%)	

[39] Per the USDOJ (2019) and (2020a) reports, several variants of RLC cut points were presented, with consideration given to base rate multipliers and the population distributions of how many individuals would be classified into the RLCs under different iterations. Although much of the description of the original version of PATTERN (USDOJ, 2019) and PATTERN 1.2 (USDOJ, 2020a) revolves around the base rate multipliers (such as setting the threshold for high risk at a recidivism rate that is 1.5 times the overall base rate of recidivism), the base rate multipliers changed between versions while the population distribution appeared to remain the same. For example, the PATTERN multipliers for the minimum, low, and high RLCs were 0.25, 0.50, and 1.66 times the base rate. The PATTERN 1.2 revisions also required revising the RLCs; new multipliers were chosen (0.20, 0.50, and 1.50) which differed from the PATTERN 1.1 base rate multipliers, but which generated a similar population distribution. (Similar shifts were implemented for the violent tools, with original and revised multipliers of 0.33, 0.66, and 1.33 versus 0.20, 0.50, and 2.0, respectively, for minimum, low, and high thresholds.) Accordingly, it appears that the population distributions dictated the base rate multipliers, and thus the PATTERN 1.3 RLCs replicated those distributions.

	All		Male		Female		All		Male		Female	
	1.2	**1.3**	**1.2**	**1.3**	**1.2**	**1.3**	**1.2**	**1.3**	**1.2**	**1.3**	**1.2**	**1.3**
Minimum	17	17	14	14	34	34	10	11	10	11	10	11
Low	29	30	26	27	44	44	31	33	31	32	34	35
Medium	19	19	20	20	16	16	55	55	54	55	58	58
High	34	34	40	39	6	6	75	75	75	75	75	73

Note: Percentages do not all sum to 100.0 due to rounding.

Table 2.6. ***Comparison of violent PATTERN RLC BOP population distributions and recidivism rates between versions 1.2 and 1.3, training sample (FY09-13)***

	Population Distributions (%)						Recidivism Rates (%)					
	All		Male		Female		All		Male		Female	
	1.2	**1.3**	**1.2**	**1.3**	**1.2**	**1.3**	**1.2**	**1.3**	**1.2**	**1.3**	**1.2**	**1.3**
Minimum	20	20	14	15	51	48	1	2	1	2	1	2
Low	42	42	42	42	46	46	9	9	9	10	7	7
Medium	14	15	16	17	3	5	21	20	21	20	23	18
High	24	23	28	27	1	1	31	35	31	35	31	42

Note: Percentages do not all sum to 100.0 due to rounding.

2.3.5 Summary of Risk Score and Levels by Tool

Table 2.7 summarizes the total risk scores and RLCs for the first and last assessments of the four PATTERN tools in the validation and revalidation samples.[40] The mean total scores of the assessments in the validation sample are similar to those in the two revalidation samples. The largest mean difference in the total scores across the three samples was for the male general recidivism tool, which had mean total scores of 35.5 in the FY 2014 to FY 2015 validation sample and 34.7 in the FY 2016 revalidation sample (a difference of 1.2 percent). The proportions of individuals assigned to each RLC between the validation and revalidation samples were also highly comparable (i.e., all within 2 percentage points of each other). The mean final assessment scores were lower than the mean first assessment scores across all four tools in all three samples, with larger differences in the general versus violent scales. There were also higher percentages of minimum- and low-risk individuals at final assessment compared to first assessment across all four tools in all three samples. These findings emphasize that PATTERN risk scores and RLCs are consistent between samples — and further, that on average all three groups were rated to be at lower risk across all four tools during the last assessment compared to the first.

In the following sections, three types of revalidation analyses address the FSA mandates for the predictive validity (Part 3), dynamic validity (Part 4), and racial and ethnic neutrality (Part 5) of PATTERN 1.3. These sections each outline the analytical approaches employed and describe the specific FSA requirement that each sought to address.

[40] Because PATTERN relies on information collected as part of the BRAVO assessment, it has followed the same schedule of administration. The initial assessment occurs as part of the intake process. Reassessment occurs after seven months in custody and every 12 months thereafter. These results include the same first and last assessment information for 14.4 percent of the FY 2014 to FY 2015 validation sample, 12.4 percent of the FY 2016 revalidation sample, and 14.7 percent of the FY 2017 revalidation sample — individuals who were not incarcerated for long enough to receive a second set of PATTERN general and violent assessments.

Table 2.7. ***PATTERN total risk scores and risk level categories, by assessment type***

	FY14-15 Validation Sample		FY16 Revalidation Sample		FY17 Revalidation Sample	
Measure	**First Assessment**	**Last Assessment**	**First Assessment**	**Last Assessment**	**First Assessment**	**Last Assessment**
Male General Recidivism						
Mean risk score (SD)	40.9 (20.6)	35.5 (22.0)	41.1 (20.4)	34.7 (21.9)	41.8 (20.5)	35.8 (22.3)
Percent minimum	9.1	16.3	8.4	16.8	8.4	16.2
Percent low	24.8	27.0	25.0	28.3	24.1	26.6
Percent medium	21.2	19.1	21.4	19.3	20.8	19.0
Percent high	44.9	37.6	45.2	35.7	46.7	38.2
Male Violent Recidivism						
Mean risk score (SD)	24.1 (12.9)	22.0 (13.9)	24.3 (12.7)	21.6 (13.8)	24.7 (12.9)	22.3 (14.0)
Percent minimum	9.4	16.8	8.8	17.2	8.9	16.4
Percent low	42.9	40.5	43.0	41.7	41.5	39.8
Percent medium	18.0	16.3	18.1	15.7	18.1	16.2
Percent high	29.7	26.4	30.1	25.5	31.5	27.7
Female General Recidivism						
Mean risk score (SD)	24.4 (17.1)	18.3 (19.0)	25.1 (17.1)	18.5 (19.1)	25.7 (17.3)	19.6 (19.7)
Percent minimum	23.6	36.9	21.8	36.2	21.2	35.7
Percent low	51.0	43.0	50.5	42.9	50.2	41.6
Percent medium	17.9	14.8	19.9	15.5	20.1	16.1
Percent high	7.5	5.3	7.8	5.4	8.6	6.6
Female Violent Recidivism						
Mean risk score (SD)	4.0 (4.1)	1.9 (5.3)	4.2 (4.1)	1.8 (5.4)	4.3 (4.2)	2.2 (5.6)
Percent minimum	27.1	50.6	26.0	51.3	25.1	48.7
Percent low	66.8	44.0	67.9	43.4	68.0	44.4
Percent medium	5.8	4.8	6.0	4.8	6.5	6.2
Percent high	0.3	0.5	0.2	0.4	0.4	0.8

Note: SD = standard deviation. Percentages do not all sum to 100 due to rounding. Male FY14-15 validation sample $N = 63,848$; female FY14-15 validation sample $N = 10,527$; male FY16 revalidation sample $N = 36,758$; female FY16 revalidation sample $N = 5,595$; male FY17 revalidation sample $N = 33,168$; female FY16 revalidation sample $N = 5,165$.

Part 3: Predictive Validity

3.1 Analytical Plan

The second analytic task involved assessing the predictive validity of PATTERN as mandated by § 3631(b)(4)(D) of the FSA. All analyses were performed on the four separate PATTERN tools:

- Male general recidivism
- Male violent recidivism
- Female general recidivism
- Female violent recidivism

The validation and revalidation of PATTERN focused on the interpretation of the Area Under the Curve (AUC) statistics (section 3.2.1), supplemented by risk level recidivism analyses (section 3.2.2) and predictive value analyses for the risk level groupings (section 3.2.3). A focus on AUCs is consistent with the analyses presented in the PATTERN developmental reports (see USDOJ, 2019; USDOJ, 2020a). The value of the AUC is a widely recognized marker of predictive accuracy and represents the probability that a randomly selected individual who has recidivated has a higher PATTERN risk score than a randomly selected individual who has not recidivated.

The AUC is especially useful for comparing tools and examining whether changes to a tool translate into increased accuracy in the risk score. However, the AUC examines whether an overall risk score could distinguish between randomly drawn individuals who have and have not recidivated, which is less useful for gauging the accuracy of a risk tool as used in practice. For instance, the AUC does not provide information about the accuracy of the high-risk RLC designation. Thus, it is useful to examine the recidivism rates by RLC. Further, as a supplement to the AUC analysis, predictive values derived from the RLC categories were analyzed. More specifically, a summary statistic known as the positive predictive value (PPV) was used to address the question: *Among individuals designated medium and high risk, what percentage recidivated?* Similarly, the negative predictive value (NPV) was used to answer the question: *Among individuals designated minimum and low risk, what percentage remained crime free?* (see e.g. Berk, Heidari, Jabbari, Kearns, & Roth, 2021; Chouldechova, 2017). These statistics provide an intuitive way for the public and stakeholders to evaluate tool accuracy.[41]

3.2 Results

3.2.1 AUC Analyses

Table 3.1 presents the results from the AUC analyses that examined the relationship between the total PATTERN risk scores and the recidivism measures, which ranged from 0.71 to 0.79.[42] As evidenced in the table, both the first and last assessment scores were highly predictive of recidivism

[41] Predictive values address some accuracy limitations inherent in the AUC measure. See Singh (2013).
[42] Supplemental analysis indicated that AUCs and other analyses (including the racial and ethnic neutrality analyses discussed in section 5) were substantively similar when considering all observations in the dataset, and when selecting only those individuals statutorily eligible to receive early release time credits.

across all four PATTERN tools in the validation and revalidation samples. The AUCs were higher at final assessment (range = 0.75 to 0.79) than at first assessment (range = 0.71 to 0.76).

Table 3.1. ***PATTERN AUCs and 95% confidence intervals***

FY14-15 Validation Sample	**First Assessment**	**Last Assessment**
Male General Recidivism	0.763 [0.759, 0.767]	0.784 [0.781, 0.788]
Male Violent Recidivism	0.752 [0.748, 0.757]	0.767 [0.763, 0.772]
Female General Recidivism	0.756 [0.746, 0.765]	0.773 [0.764, 0.782]
Female Violent Recidivism	0.742 [0.722, 0.763]	0.750 [0.730, 0.771]
FY16 Revalidation Sample	**First Assessment**	**Last Assessment**
Male General Recidivism	0.761 [0.756, 0.766]	0.790 [0.785, 0.794]
Male Violent Recidivism	0.758 [0.752, 0.764]	0.777 [0.771, 0.783]
Female General Recidivism	0.741 [0.728, 0.754]	0.768 [0.755, 0.781]
Female Violent Recidivism	0.714 [0.685, 0.743]	0.749 [0.721, 0.776]
FY17 Revalidation Sample	**First Assessment**	**Last Assessment**
Male General Recidivism	0.757 [0.752, 0.762]	0.789 [0.784, 0.794]
Male Violent Recidivism	0.751 [0.745, 0.758]	0.770 [0.763, 0.776]
Female General Recidivism	0.754 [0.741, 0.768]	0.780 [0.767, 0.793]
Female Violent Recidivism	0.763 [0.737, 0.788]	0.782 [0.757, 0.807]

3.2.2 Risk Level Recidivism Analyses

Figures 3.1 and 3.2 display the rates of recidivism by PATTERN RLC among the validation and revalidation samples. These figures provide an important and assessable gauge of the accuracy of PATTERN's RLC designations. They reveal a steady increase in the rate of recidivism with each successively higher risk category. For example, approximately 10 to 11 percent of individuals in the validation and revalidation samples who were rated as minimum risk on the general male scale recidivated during the three-year follow-up period, compared to 30 to 31 percent in the low-risk group, 52 to 54 percent in the medium-risk group, and 74 to 75 percent in the high-risk group. Thus, the RLCs provide meaningful distinctions of the likelihood of recidivism.

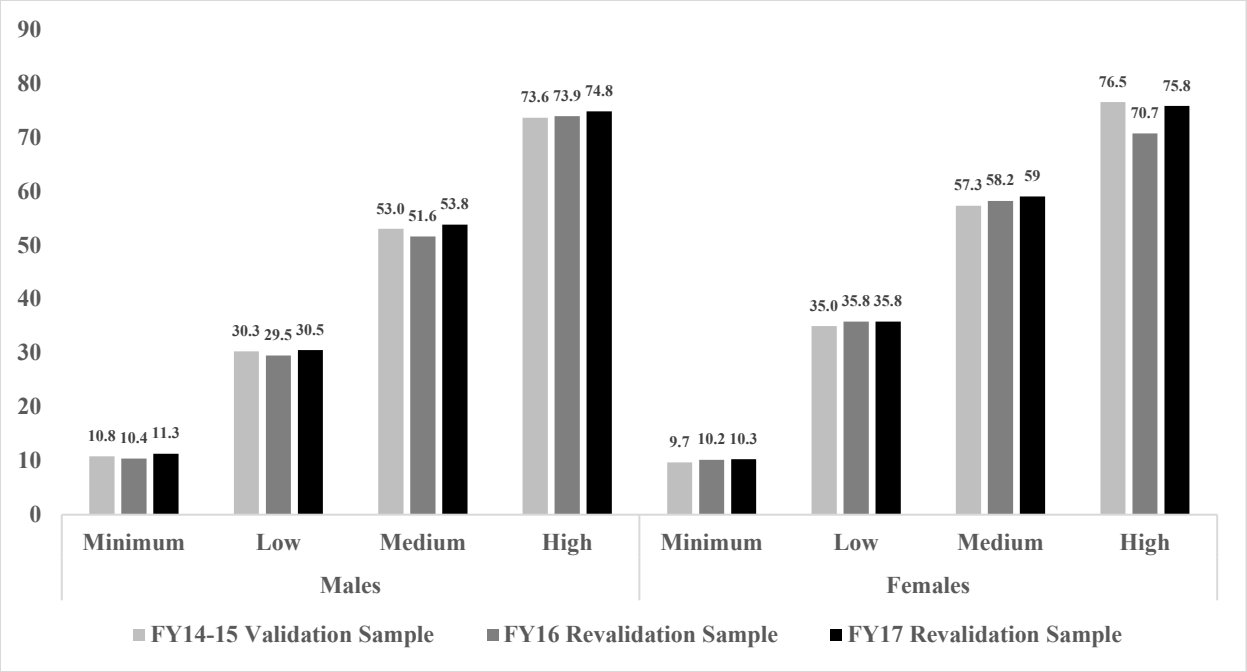


Figure 3.1. Percentage of males and females with general recidivism outcomes by PATTERN reassessment general risk level category in the validation and revalidation samples.

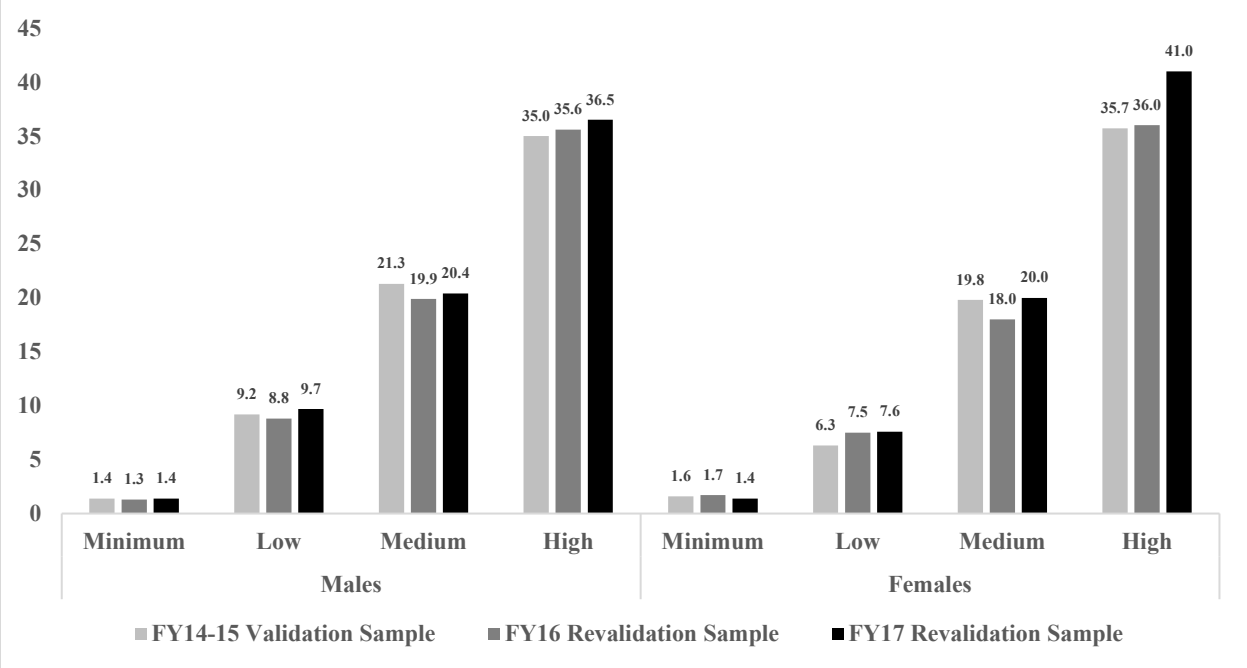


Figure 3.2. Percentage of males and females with violent recidivism outcomes by PATTERN reassessment violent risk level category in the validation and revalidation samples.

3.2.3 Predictive Value Analyses

PATTERN's institutional impact is based on the lower risk grouping (i.e., those in the minimum and low overall RLCs) and higher risk grouping (i.e., those in the medium and high overall RLCs), with time credit accrual and some programming priority based on these designations. As a final way to consider the accuracy of PATTERN, predictive values were calculated for the lower and higher risk groupings. The PPVs measure the recidivism rates of the higher risk groups, while the NPVs measure the success rates of the lower risk groups. Unlike AUCs, these predictive values do not lend themselves well to comparing accuracy between tools, because the predictive values are derived from the RLCs and depend on where the risk level cut points are drawn. However, predictive values provide important information about how accurately risk is predicted within RLCs and RLC groupings (i.e., minimum/low and medium/high) for the chosen cut points. The predictive values thus serve as measures of accuracy for the risk tool RLCs.

Table 3.2 displays the PPV and NPV metrics for PATTERN 1.3. These values were derived from 2 × 2 tables known as confusion tables or confusion matrices (see Berk et al., 2021). Since these statistics draw on the BOP risk grouping decision dichotomy, they provide a more meaningful practical comparison of the differences in prediction between the lower and higher risk groups.[43]

The PPV is the recidivism rate of the higher risk grouping. As such, it provides an intuitive measure of accuracy. The higher the PPV, the more accurate the higher risk grouping is in identifying individuals who recidivate. The PPV statistics are similar across the validation and revalidation samples. For the general tools, 66 to 68 percent of medium- and high-risk males recidivated during follow-up, as did 61 to 64 percent of medium- and high-risk females.

Table 3.2. ***Positive predictive values (PPV), negative predictive values (NPV), false positive rates (FPR), and false negative rates (FNR)***

FY14-15 Validation	***PPV***	***NPV***	***FPR***	***FNR***
Male General Recidivism	0.67	0.77	0.36	0.21
Male Violent Recidivism	0.30	0.93	0.36	0.24
Female General Recidivism	0.62	0.77	0.11	0.60
Female Violent Recidivism	0.22	0.96	0.04	0.76

FY16 Revalidation	***PPV***	***NPV***	***FPR***	***FNR***
Male General Recidivism	0.66	0.78	0.35	0.22
Male Violent Recidivism	0.30	0.93	0.35	0.24
Female General Recidivism	0.61	0.76	0.12	0.60
Female Violent Recidivism	0.20	0.96	0.04	0.80

[43] These analyses do not provide diagnostics on the combined decision of being minimum/low risk on both general and violent tools compared to being medium or high risk on either tool. First, it was most important to perform analysis on the underlying tools as developed. Second, there is substantial overlap between the general recidivism groupings and the combined general and violent groupings, meaning the general instrument is driving the process. For example, for FY 2016 and FY 2017, only 1.3 percent of males were classified as overall higher risk due to having a medium- or high-risk violent RLC despite a minimum- or low-risk general RLC. For females, less than half a percent were classified as overall higher risk due to cross-classification.

FY17 Revalidation	*PPV*	*NPV*	*FPR*	*FNR*
Male General Recidivism	0.68	0.77	0.36	0.20
Male Violent Recidivism	0.31	0.93	0.37	0.23
Female General Recidivism	0.64	0.76	0.12	0.56
Female Violent Recidivism	0.22	0.96	0.06	0.73

Note: PPV = positive predictive value (proportion of true positives out of all positive predictions); NPV = negative predictive value (proportion of true negatives out of all negative predictions); FPR = false positive rates (proportion of false positives out of all observed nonrecidivism); FNR = false negative rate (proportion of false negatives out of all observed recidivism).

The PPVs for the violent recidivism tools are significantly lower than for the general recidivism tools, with values between 20 and 30 percent depending on the tool and sample. This lower accuracy reflects the difficulty of predicting violent recidivism events, which have a relatively low base rate of occurrence.[44] For the validation sample, 47 percent of males and 31 percent of females recidivated for any type of offense, but only 17 percent of males and 5 percent of females recidivated for a violent offense. With low-prevalence events, risk tools tend to be less accurate in their positive predictions. However, these assessments tend to be accurate with negative predictions.

The NPV indicates the percentage of minimum- and low-risk individuals who did not recidivate (i.e., successes). Stated differently, the NPV is the combined success rate of the minimum and low RLCs. The higher the NPV, the more accurate the lower risk designations are in identifying individuals who do not recidivate. For the general recidivism tools, the NPVs are around 76 to 78 percent across the different tools and samples, meaning that 76 to 78 percent of individuals who are designated as minimum or low risk avoid recidivating within the three-year follow-up period. For the rarer violent outcomes, the NPVs improve to 93 percent for males and 96 percent for females, meaning that over 90 percent of minimum- and low-risk individuals are recidivism successes.[45]

Table 3.2 also provides false positive rates (FPRs) and false negative rates (FNRs). FPRs indicate the percentage of false positives among individuals who recidivate, while FNRs indicate the percentage of false negatives among individuals who do not recidivate. Since the 2 × 2 matrix for the predictive values and misclassification rates was constructed on the split between minimum/low and medium/high categories, the statistics in Table 3.2 do not reflect other possible

[44] Note that the AUCs for the violent recidivism tools are quite high (i.e., well over 0.70). The difference in conclusions drawn based on the AUCs and PPVs underscores the importance of having additional measures of accuracy. The AUCs confirm that for the violent tools, if individuals who did and did not recidivate were randomly selected, the individual who recidivated would have a higher PATTERN score more than 70 percent of the time. However, in more practical application, among those designated as medium or high risk of violent recidivism, the PPVs indicate an actual violent recidivism rate of around 20 to 30 percent. Conversely, 70 to 80 percent of individuals designated as medium or high risk for a violent recidivism event will not actually go on to have such an event. This higher error rate could have important implications when evaluating the accuracy of the violent recidivism tools.

[45] These high numbers are also largely a reflection of the rarity of the outcome measured. For instance, while PATTERN's female minimum- and low-risk designations have a 96 percent success rate, the violent recidivism rate for females is just 5 percent; one could simply predict that every female was at minimum or low risk of violent recidivism and be correct 95 percent of the time. This is a product of the low prevalence rate more than the discernment of the prediction.

groupings of the RLCs. For example, the PPVs for the high-risk group are higher than for all the other RLCs (indicating more accurate predictions), and the NPVs for the minimum-risk group are higher than for all the others.[46]

[46] Decisions related to PATTERN are primarily based on the low/medium threshold, so the full range of possibilities are not presented here. Interested readers may explore them using the distribution tables provided in Appendix A.

Part 4: Dynamic Validity

4.1 Analytical Plan

The third analytical task involved examining changes in PATTERN scores for individuals over time. This portion of the study addresses § 3631(b)(4)(C) of the FSA mandate. More specifically, this analysis compares changes in PATTERN scores and RLCs from initial assessment (at intake) to final assessment (prior to release).

PATTERN was developed with both static (i.e., unchangeable) and dynamic (i.e., changeable) risk factors. The inclusion of dynamic variables underscores a belief that people can alter their likelihood for future reoffending (Bonta, 2002). This means that the utility of PATTERN lies with its ability not only to predict recidivism, but also to provide meaningful information about changes in risk.

The USDOJ (2020) report demonstrated how an individual could change his or her PATTERN risk designations based on hypothetical examples of compliance, noncompliance, and a mixture of the two. The current analysis expands on this prior work by examining actual changes in PATTERN risk scores and levels from first to last assessment.[47]

Theoretically, changes in a person's risk score over time should correspond to differences in their probability for recidivism. For example, if one were to follow institutional rules and participate in programming and other work requirements, it is expected that one's risk scores would be lowered and one's likelihood for recidivism would be reduced. On the other hand, if one were to violate rules and not comply with programming and work expectations, it is anticipated that one's risk scores would be higher and one's likelihood for recidivism would increase. To test the dynamic validity of PATTERN, this study assesses the extent to which individual risk scores and levels changed from first to last assessment (section 4.2.1) and how these changes relate to recidivism (section 4.2.2).

4.2 Results

4.2.1 Changes in Risk Scores and Levels

Results in Table 4.1 show an overall reduction in the mean risk scores and assigned risk levels from first to last assessment across all four tools. The largest score reduction was found in the female general recidivism scale (7 to 8 points), followed by the male general recidivism scale (6 to 7 points) and the male and female violent recidivism scales (3 points). Although most individuals (between 62 and 71 percent) in the three samples remained in the same risk level across the four tools, approximately 23 to 35 percent had a lower risk level and 3 to 6 percent had a higher risk level at the last assessment compared to the first. These analyses confirm that changes in risk scores and levels are possible in both directions across the four PATTERN instruments and further

[47] Individuals with only one assessment available are excluded from the dynamic analyses. The sample sizes are 63,645 for the FY 2014 to FY 2015 validation sample, 37,086 for the FY 2016 revalidation sample, and 38,333 for the FY 2017 revalidation sample.

document that such changes are more likely to be associated with reductions rather than increases in risk.

Table 4.1. ***Change in PATTERN risk scores and levels from first to last assessment***

	FY14-15 Validation Sample	FY16 Revalidation Sample	FY17 Revalidation Sample
Male General Recidivism			
Mean change in risk score (SD)	-6.2 (8.3)	-7.3 (8.9)	-7.0 (9.0)
Percentage at lower risk level	28.0	32.4	30.1
Percentage at same risk level	69.3	65.0	67.2
Percentage at higher risk level	2.7	2.6	2.7
Male Violent Recidivism			
Mean change in risk score (SD)	-2.5 (5.5)	-3.1 (5.8)	-2.8 (5.9)
Percentage at lower risk level	23.2	26.1	24.5
Percentage at same risk level	70.8	68.4	69.3
Percentage at higher risk level	6.0	5.5	6.2
Female General Recidivism			
Mean change in risk score (SD)	-7.4 (8.3)	-7.9 (8.4)	-7.3 (8.6)
Percentage at lower risk level	28.8	30.6	30.8
Percentage at same risk level	67.3	66.3	65.0
Percentage at higher risk level	3.9	3.1	4.2
Female Violent Recidivism			
Mean change in risk score (SD)	-2.6 (3.4)	-2.9 (3.4)	-2.6 (3.5)
Percentage at lower risk level	32.7	34.6	32.8
Percentage at same risk level	63.6	61.8	62.3
Percentage at higher risk level	3.7	3.5	4.8

Note: SD = standard deviation. Percentages do not all sum to 100 due to rounding. Male FY14-15 validation sample $N = 54{,}974$; female FY14-15 validation sample $N = 8{,}671$; male FY16 revalidation sample $N = 32{,}425$; female FY16 revalidation sample $N = 4{,}661$; male FY17 revalidation sample $N = 28{,}433$; female FY16 revalidation sample $N = 4{,}260$.

Figures 4.1 and 4.2 illustrate the changes in the overall RLCs for males and females. These figures display the distribution of highest RLC assignment at first and last assessments by gender in the validation and revalidation samples. Figure 4.1 shows that males were more likely to receive a lower overall risk assignment during the last assessment compared to the first. These changes reflect a 10 to 12 percent increase in the proportion of minimum- and low-risk individuals, which indicates that males were more likely to become eligible for earned time credit over time. Figure 4.2 shows that females were also more likely to be rated as lower risk during the last assessment relative to the first. Although the magnitude of the increase in the proportion of minimum- and low-risk females (i.e., a 6 to 8 percent increase) was smaller than in the male sample, females were also much more likely to be identified as minimum or low risk during their first assessment (72 to 76 percent) compared to males (33 to 35 percent). Given the differences in the initial risk ratings between genders, females had less of an opportunity to be reassessed at a lower risk level compared to males.

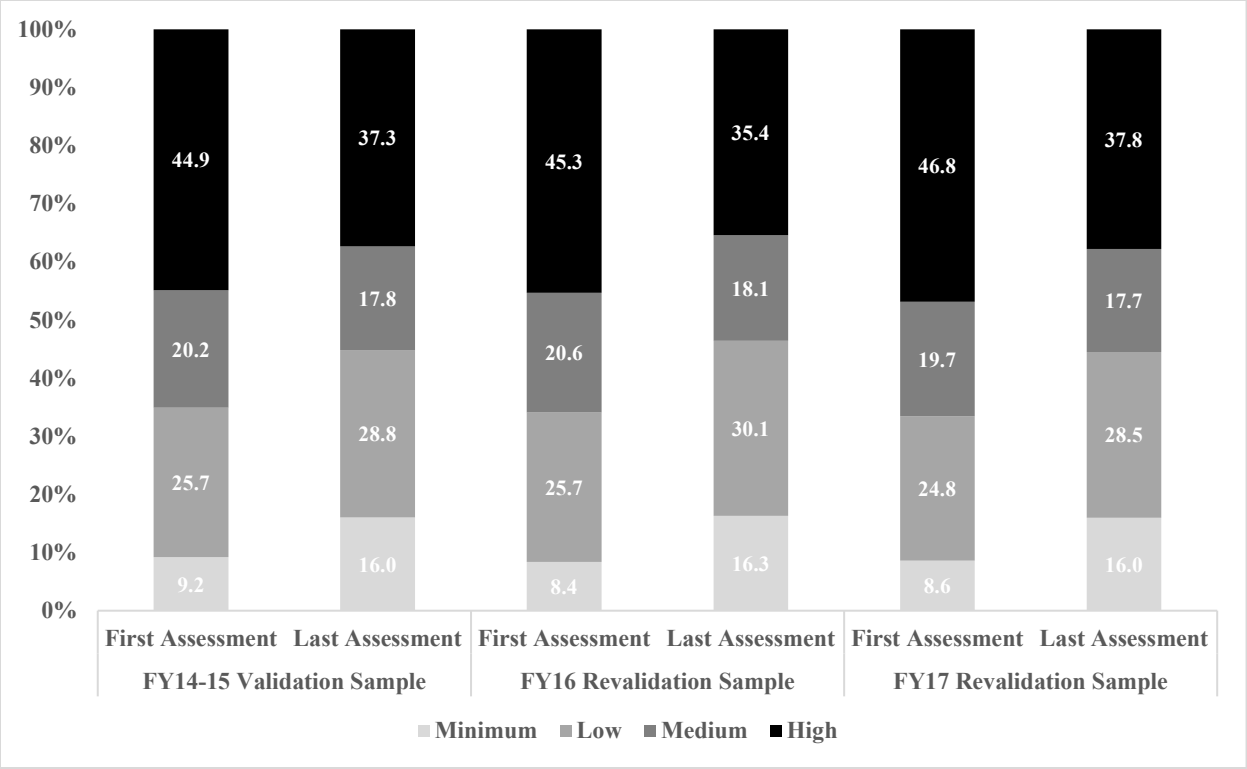


Figure 4.1. Percentage of males assigned to each of the PATTERN overall risk level categories in the validation and revalidation samples.

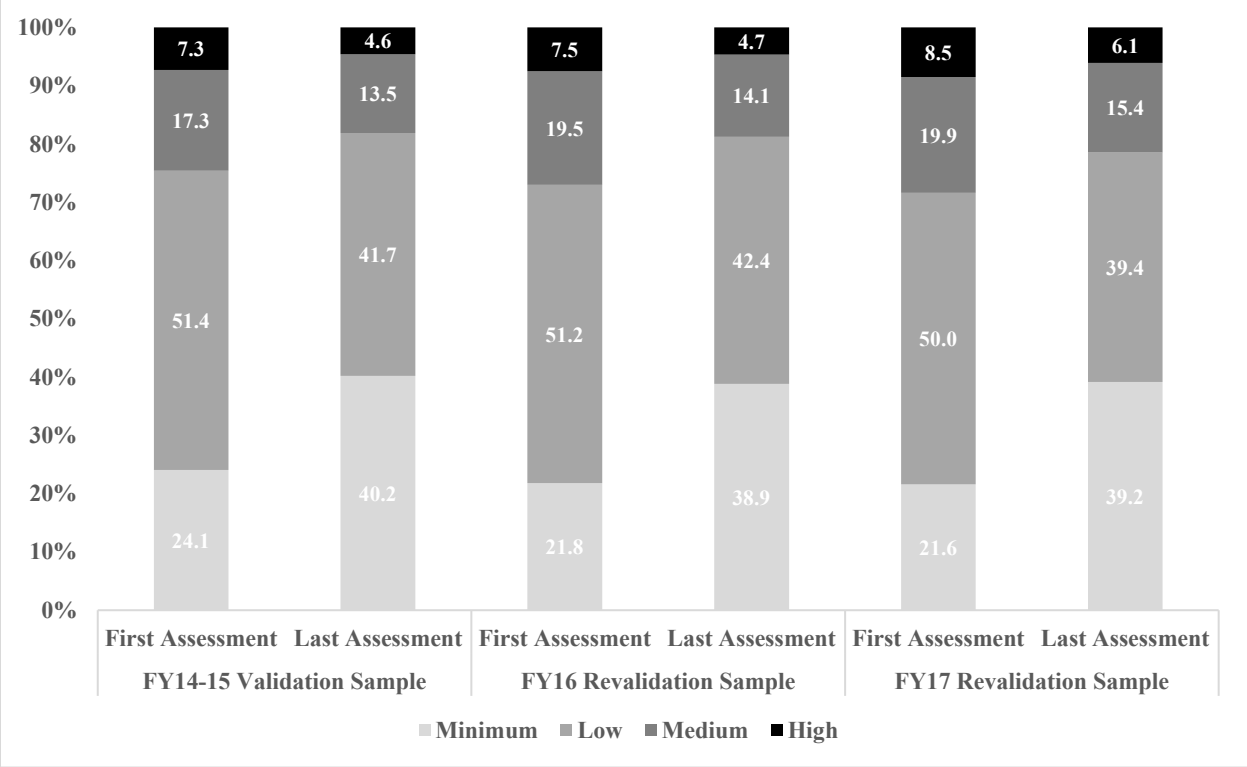


Figure 4.2. Percentage of females assigned to each of the PATTERN overall risk level categories in the validation and revalidation samples.

4.2.2 Dynamic Validity Analyses

In addition to documenting the nature and extent of changes in risk scores and levels between the first and last assessments, it is also important to examine how these changes relate to an individual's probability for recidivism. Table 4.2 reports the percentage of individuals who recidivated by their change in risk level status (i.e., lower, same, or higher risk level at final assessment compared to first) for all four PATTERN instruments in the validation and revalidation samples. Across all four instruments and all three samples, with one exception, individuals with a lower risk level at final assessment displayed the lowest probability for recidivism, whereas those with a higher risk level at final assessment had the highest probability for recidivism. In the FY 2017 revalidation sample, males whose risk level on the violent tool did not change from first to last assessment were slightly more likely to recidivate than those whose risk level increased from first to last assessment (19.3 vs. 19.1 percent, respectively). In totality, these findings emphasize that reducing one's risk level may lower one's likelihood for recidivism, and conversely, increasing one's risk level may heighten one's likelihood for recidivism.

Table 4.2. ***Percentage of individuals who recidivated by change in PATTERN risk level status from first to last assessment***

	FY14-15 Validation Sample (%)	FY16 Revalidation Sample (%)	FY17 Revalidation Sample (%)
Male General Recidivism			
Lower risk level	33.3	31.6	31.3
Same risk level	50.1	50.0	51.4
Higher risk level	54.1	57.0	56.8
Male Violent Recidivism			
Lower risk level	10.4	9.7	8.8
Same risk level	17.0	17.0	19.3
Higher risk level	21.6	19.8	19.1
Female General Recidivism			
Lower risk level	27.4	25.8	26.8
Same risk level	27.9	29.9	30.4
Higher risk level	47.2	44.8	52.1
Female Violent Recidivism			
Lower risk level	3.1	2.6	3.8
Same risk level	4.4	5.4	5.5
Higher risk level	11.3	10.4	12.0

Note: Lower risk level = lower risk level assigned at last assessment compared to first assessment; same risk level = same risk level assigned at last assessment as at first assessment; higher risk level = higher risk level assigned at last assessment compared to first assessment.

Next, analyses examined how changes in risk scores during one's period of confinement were associated with changes in the likelihood of recidivism. Tables 4.3 presents the results of the logistic regression analyses of the initial PATTERN assessment scores and the changes in scores from first to last assessment predicting recidivism in the male and female validation and revalidation samples. In all 12 of the logistic regression models, both the initial assessment score and the change in risk score predicted recidivism ($p < 0.001$). When holding the initial PATTERN scores constant, for every one-point increase in the total general and violent male scale scores from first to last assessment, there was a corresponding 6 and 7 percent increase in the odds of general and violent rearrest, respectively, within three years of release. For the female models, there was

similarly a 4 to 6 percent increase in the odds of general recidivism and an 11 to 14 percent increase in the odds of violent recidivism. These regression results confirm that increases in PATTERN risk scores during incarceration are associated with higher levels of recidivism and decreases in PATTERN risk scores are associated with lower levels of recidivism.

Table 4.3. ***Logistic regression of first PATTERN assessment score and change in score from first to last assessment predicting recidivism in the male and female validation and revalidation samples***

FY14-15 Validation Sample	**Male General**	**Male Violent**	**Female General**	**Female Violent**
First assessment score	1.06	1.08	1.06	1.21
Change in risk score	1.06	1.07	1.04	1.11
Constant	0.12	0.02	0.09	0.02
Model χ^2	14,534.19	7,134.03	1,707.08	320.25
Nagelkerke R^2	0.311	0.209	0.256	0.123

FY16 Revalidation Sample	**Male General**	**Male Violent**	**Female General**	**Female Violent**
First assessment score	1.06	1.09	1.06	1.20
Change in risk score	1.06	1.07	1.05	1.14
Constant	0.11	0.02	0.10	0.02
Model χ^2	8,832.52	4,359.22	894.70	182.35
Nagelkerke R^2	0.319	0.219	0.249	0.123

FY17 Revalidation Sample	**Male General**	**Male Violent**	**Female General**	**Female Violent**
First assessment score	1.06	1.09	1.06	1.23
Change in risk score	1.06	1.07	1.06	1.14
Constant	0.12	0.02	0.12	0.02
Model χ^2	7,871.24	3,996.41	916.09	248.70
Nagelkerke R^2	0.323	0.222	0.274	0.167

Note: Reported values are odds ratios. All findings are statistically significant at the 0.001 level.

Part 5: Racial and Ethnic Neutrality

5.1 Analytical Plan

The fourth and final analytic task involved assessing PATTERN for racial and ethnic neutrality. The FSA mandates that the Department's review of the risk and needs assessment system must include "an evaluation of the rates of recidivism among similarly classified prisoners to identify any unwarranted disparities, including disparities among similarly classified prisoners of different demographic groups, in such rates."[48]

The PATTERN instruments were evaluated using several approaches that reflect the current scientific standards for assessing instrument neutrality. This included a comparison of AUCs (section 5.2.1) and predictive values (section 5.2.2) by race and ethnic group.[49] In addition, the examinations focused on differential prediction analysis, which is assessed through a series of nested logistic regressions. This analysis measured whether PATTERN scores predict recidivism differently across racial and ethnic subgroups (section 5.2.3; see Chouldechova, 2017; Skeem & Lowenkamp, 2016). The approach was borrowed from the *Standards for Educational and Psychological Testing* developed by the American Education Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014).[50]

5.2 Results

5.2.1 AUC Analyses by Race

Table 5.1 reports the AUCs by race and ethnic group, which range from a low of 0.634 (Native American male, violent recidivism, FY 2017 revalidation sample) to a high of 0.862 (Asian female, general recidivism, FY 2017 revalidation sample). Most values fall in the 0.700 range (37 of 60, or 61.7%), followed by those in the 0.800 range (16 of 60, or 26.7%) and the 0.600 range (7 of 60, or 11.7%).[51] The AUCs are similar across the validation and revalidation samples.

[48] FSA § 3631(b)(4)(E).
[49] See also USDOJ (2019, p. 60) and USDOJ (2020, p. 9 and p. 11).
[50] Although other notions of fairness exist (see Berk et al., 2021), differential prediction analysis has emerged as the prevailing approach for examining group-based fairness in the criminal justice context. The methodology has been employed with the federal Post Conviction Risk Assessment (PCRA) (Skeem & Lowenkamp, 2016; 2020), the Arnold Public Safety Assessment (DeMichelle et al., 2021), the Level of Service Inventory-Revised (LSI-R) (Wilson & Gutierrez, 2014), and the Pennsylvania Commission on Sentencing's risk tool (PCS, 2018), among others.
[51] Recall that the AUC is a common accuracy statistic reflecting the likelihood that a randomly drawn individual who recidivated would have a higher risk score than a randomly drawn individual who did not recidivate. See section 3.1.

Table 5.1. ***PATTERN AUCs and 95% confidence intervals (CIs), by race/ethnicity***

	FY14-15 Validation Sample		FY16 Revalidation Sample		FY17 Revalidation Sample	
	AUC [95% CI]	***N***	**AUC [95% CI]**	***N***	**AUC [95% CI]**	***N***
Male General Recidivism						
White	0.797 [0.791, 0.803]	21,729	0.802 [0.794, 0.810]	12,096	0.808 [0.800, 0.816]	10,677
Black	0.756 [0.751, 0.762]	25,559	0.768 [0.761, 0.776]	14,910	0.766 [0.758, 0.774]	13,507
Hispanic	0.770 [0.762, 0.778]	13,153	0.778 [0.767, 0.788]	7,872	0.781 [0.771, 0.792]	7,237
Asian	0.826 [0.800, 0.853]	1,178	0.806 [0.768, 0.845]	650	0.805 [0.765, 0.844]	550
Native American	0.730 [0.706, 0.754]	2,229	0.754 [0.723, 0.785]	1,230	0.668 [0.630, 0.705]	1,197
Male Violent Recidivism						
White	0.788 [0.779, 0.796]	21,729	0.801 [0.790, 0.812]	12,096	0.798 [0.787, 0.810]	10,677
Black	0.729 [0.722, 0.736]	25,559	0.741 [0.732, 0.750]	14,910	0.739 [0.792, 0.748]	13,507
Hispanic	0.759 [0.747, 0.770]	13,153	0.774 [0.759, 0.789]	7,872	0.759 [0.743, 0.774]	7,237
Asian	0.823 [0.774, 0.872]	1,178	0.807 [0.743, 0.871]	650	0.803 [0.749, 0.856]	550
Native American	0.687 [0.664, 0.711]	2,229	0.676 [0.643, 0.708]	1,230	0.634 [0.599, 0.669]	1,197
Female General Recidivism						
White	0.781 [0.768, 0.795]	4,613	0.775 [0.756, 0.793]	2,483	0.784 [0.765, 0.804]	2,231
Black	0.788 [0.769, 0.807]	2,479	0.767 [0.740, 0.794]	1,302	0.809 [0.784, 0.835]	1,246
Hispanic	0.745 [0.725, 0.765]	2,764	0.742 [0.715, 0.769]	1,445	0.744 [0.716, 0.772]	1,359
Asian	0.803 [0.734, 0.872]	265	0.861 [0.797, 0.925]	147	0.862 [0.792, 0.932]	110
Native American	0.699 [0.648, 0.750]	406	0.746 [0.679, 0.812]	218	0.694 [0.620, 0.767]	219
Female Violent Recidivism						
White	0.731 [0.697, 0.764]	4,613	0.723 [0.678, 0.767]	2,483	0.779 [0.739, 0.819]	2,231
Black	0.780 [0.744, 0.815]	2,479	0.778 [0.732, 0.823]	1,302	0.834 [0.797, 0.871]	1,246
Hispanic	0.715 [0.674, 0.757]	2,764	0.735 [0.674, 0.797]	1,445	0.704 [0.641, 0.768]	1,359
Asian	0.748 [0.595, 0.902]	265	0.755 [0.554, 0.956]	147	0.824 [0.738, 0.910]	110
Native American	0.804 [0.721, 0.887]	406	0.739 [0.591, 0.887]	218	0.672 [0.585, 0.758]	219

5.2.2 Predictive Value Analyses by Race

Predictive values (PPVs and NPVs) were calculated separately for the five racial and ethnic groups (see Table 5.2). Recall that for these statistics, RLCs are aggregated to lower and higher risk groupings as is employed in practice. The PPV provides a measure of recidivism accuracy in the higher risk grouping, while the NPV provides a measure of nonrecidivism accuracy in the lower risk grouping.

For example, for the general female tool in the FY 2014 to FY 2015 validation sample, the PPVs indicate that 64 percent of white females who were classified as medium or high risk went on to recidivate, compared to 62 percent of Black females, 57 percent of Hispanic females, and 72 percent of Native American females. The NPVs indicate that 75 percent of white females in the minimum- or low-risk categories avoided recidivating, compared to 80 percent of Black females, 79 percent of Hispanic females, and 51 percent of Native American females. Thus, among medium- and high-risk females, Black and Hispanic individuals were less likely to recidivate than white individuals; among the minimum- and low-risk females, Black and Hispanic individuals were also more likely to be successful than white individuals.

As another example, the PPVs indicate that medium- and high-risk Black males recidivated at a slightly higher rate than medium- and high-risk white males in the FY 2014 to FY 2015 validation sample (68 vs. 66 percent). This disparity evened out in FY 2016 (both 66 percent) and reversed directions in FY 2017 (67 percent for Black males vs. 69 percent for white males). The NPV results show slightly higher success rates among minimum- and low-risk white individuals compared to Black individuals across the three samples (79 vs. 74 percent in FY 2014 to FY 2015, 79 percent vs. 75 percent in FY 2016, and 78 percent vs. 75 percent in FY 2017). An additional result for the general male tool is that the PPVs and NPVs among white, Black, and Hispanic males are roughly consistent with each other (in the 0.60s for PPVs and 0.70s for NPVs for each group). By comparison, the RLC groupings do a much better job at identifying higher risk Native American males (low 0.80s), but a much poorer job at identifying likely successes among minimum- and low-risk Native American males (0.44 to 0.56). For Asian men, the trend is the opposite: The groupings overclassify Asian males as high risk (PPVs from 0.55 to 0.59) compared to white, Black, and Hispanic males, while the NPVs indicate that Asian males designated minimum or low risk are on average more likely than white, Black, and Hispanic males to succeed (0.84 to 0.86).

Table 5.2. ***Positive and negative predictive values and false positive and negative rates, by race/ethnicity***

	FY14-15 Validation Sample				FY16 Revalidation Sample				FY17 Revalidation Sample			
	PPV	***NPV***	***FPR***	***FNR***	***PPV***	***NPV***	***FPR***	***FNR***	***PPV***	***NPV***	***FPR***	***FNR***
Male General Recidivism												
White	0.66	0.79	0.24	0.30	0.66	0.79	0.25	0.29	0.69	0.78	0.24	0.28
Black	0.68	0.74	0.52	0.14	0.66	0.75	0.48	0.15	0.67	0.75	0.50	0.14
Hispanic	0.62	0.77	0.34	0.26	0.61	0.79	0.30	0.28	0.64	0.78	0.32	0.25
Native American	0.80	0.56	0.55	0.14	0.83	0.52	0.52	0.15	0.81	0.44	0.63	0.15
Asian	0.59	0.86	0.18	0.33	0.55	0.86	0.18	0.37	0.57	0.84	0.18	0.39
Male Violent Recidivism												
White	0.25	0.95	0.26	0.30	0.26	0.95	0.26	0.30	0.27	0.95	0.27	0.29
Black	0.34	0.90	0.47	0.20	0.34	0.90	0.44	0.21	0.35	0.90	0.47	0.20
Hispanic	0.24	0.94	0.31	0.31	0.24	0.95	0.28	0.32	0.25	0.94	0.31	0.31
Native American	0.32	0.91	0.68	0.09	0.31[a]	0.92[a]	0.68[a]	0.09[a]	0.29	0.86	0.73	0.12
Asian	0.18[a]	0.98[a]	0.20[a]	0.29[a]	0.19[a]	0.96[a]	0.17[a]	0.45[a]	0.17[a]	0.96[a]	0.18[a]	0.50[a]
Female General Recidivism												
White	0.64	0.75	0.11	0.62	0.63	0.74	0.11	0.62	0.66	0.75	0.11	0.58
Black	0.62	0.80	0.13	0.50	0.57	0.80	0.16	0.51	0.63	0.81	0.14	0.45
Hispanic	0.57	0.79	0.09	0.66	0.59	0.78	0.10	0.65	0.58	0.76	0.12	0.62
Native American	0.72	0.51	0.21	0.59	0.76[a]	0.53[a]	0.18[a]	0.57[a]	0.73[a]	0.49[a]	0.26[a]	0.53[a]
Asian	0.52[a]	0.87[a]	0.06[a]	0.67[a]	0.71[a]	0.84[a]	0.03[a]	0.68[a]	0.67[a]	0.82[a]	0.03[a]	0.75[a]
Female Violent Recidivism												
White	0.16[a]	0.97[a]	0.03[a]	0.84[a]	0.13[a]	0.96[a]	0.03[a]	0.89[a]	0.24[a]	0.96[a]	0.04[a]	0.76[a]
Black	0.30	0.95	0.06	0.63	0.27[a]	0.94[a]	0.06[a]	0.72[a]	0.27	0.95	0.08	0.62
Hispanic	0.13[a]	0.96[a]	0.04[a]	0.88[a]	0.17[a]	0.96[a]	0.03[a]	0.86[a]	0.18[a]	0.96[a]	0.05[a]	0.79[a]
Native American	0.26[a]	0.96[a]	0.13[a]	0.42[a]	0.21[a]	0.97[a]	0.16[a]	0.36[a]	0.13[a]	0.89[a]	0.24[a]	0.72[a]
Asian	0.00[a]	0.98[a]	0.01[a]	1.00[a]	0.33[a]	0.97[a]	0.01[a]	0.83[a]	0.00[a]	0.98[a]	0.01[a]	1.00[a]

Note: PPV = positive predictive value. NPV = negative predictive value. FPR = false positive rate. FNR = false negative rate. The superscript [a] indicates that at least one of the 2 × 2 cells included fewer than 30 observations, so the generalizability to population estimates is less certain due to small sample size. Full distribution tables are provided in Appendix A.

When base rates of recidivism differ, it is impossible to achieve parity in PPVs/NPVs and FNRs/FPRs (Berk et al., 2021; Chouldechova, 2017). Goel et al. (2021, p. 16) note that "differences in false positive rates often tell us more about the underlying populations than about bias in the algorithm." Nevertheless, differences in FNRs and FPRs are important metrics for assessment to some observers, and thus they are also included in Table 5.2.[52]

5.2.3 Differential Prediction Analyses

The differential prediction analyses proceed with a series of four logistic regression models for each tool and in all three samples (FY 2014 to FY 2015, FY 2016, and FY 2017). Model 1 includes only the categorical race and ethnicity identifier as a predictor of recidivism, with white individuals serving as the reference group. This model tests for racial differences in the recidivism outcomes in the absence of other control variables. Model 2 includes only the PATTERN risk score and assesses if the score is independently able to predict recidivism.

Model 3 includes both the PATTERN risk score and the race and ethnicity identifier. In this model, a statistically significant result for Black or Hispanic individuals indicates that for a given risk score, a member of the given race or ethnic group has a statistically different probability of recidivism, on average, compared to a white individual with the same score. This could be in either a positive or negative direction. For example, a positive result for the Black indicator variable would indicate that compared to white individuals, and controlling for PATTERN score, Black individuals had a higher rate of observed recidivism (i.e., overprediction of risk for Black individuals relative to white individuals). Reciprocally, a negative directional result for the Black indicator variable would show that compared to white individuals, and controlling for PATTERN score, Black individuals had a lower rate of observed recidivism (i.e., underprediction of risk for Black individuals relative to white individuals). When a group has a higher rate of recidivism compared to the white group based on this test, the risk score underpredicts risk for that group relative to the white group (since the group actually recidivates at a higher rate for the same score). When a group has a lower rate of recidivism, the score overpredicts their risk relative to the white group.

In Model 4, an interaction term between race/ethnicity and the risk score is added to further test whether the relationship between race and recidivism varies significantly across changes in the risk score. As a hypothetical example, if Black individuals with low PATTERN scores were on average less likely to recidivate than white individuals with those same low scores (resulting in the overprediction of lower risk Black individuals compared to lower risk white individuals), but Black individuals with high PATTERN scores were more likely to recidivate than white individuals with high scores (underprediction of higher risk Black individuals compared to higher risk white individuals), this would constitute an interaction effect between race and the risk score. Not only would there be differential prediction, but the relative nature of the difference would vary with changes in the score value (see Berry, 2015).

Table 5.3 provides the results of Model 1 through Model 4 for the general male PATTERN instrument using the FY 2014 to FY 2015 validation sample. In Model 1, Black individuals have

[52] Appendix A presents the full distribution of recidivism outcomes by RLC separated by racial and ethnic group for the validation and revalidation samples.

an 85 percent higher odds of recidivism compared to white individuals when nothing else is controlled for, and Hispanic individuals have a 13 percent higher odds of recidivism compared to white individuals. Native American individuals have 286 percent higher odds of recidivism, while the odds of Asian individuals recidivating are 43 percent lower relative to white individuals. Model 2 confirms that higher PATTERN scores are associated with greater likelihoods of recidivism. In Model 3, when controlling for the PATTERN score, Black individuals have slightly (2 percent) lower odds of recidivism compared to white individuals, but the difference is not statistically significant ($p = 0.330$). Hispanic and Asian individuals have lower odds of recidivism compared to white individuals (10 and 33 percent respectively), and the result is statistically significant. This means the PATTERN score overpredicts the recidivism risk of Hispanic and Asian individuals compared to white individuals, and that when controlling for PATTERN score, Hispanic and Asian individuals have lower recidivism rates than white individuals. The interaction effects from Model 4 are not statistically significant for Hispanic individuals, meaning the relationship between race and recidivism did not change (to a statistically significant degree) as the risk score varied. The interaction effect is statistically significant for Asian individuals, indicating that the relationship between recidivism for Asian and white individuals changes across values of the PATTERN score. However, the substantive size of this interaction effect is small, at an odds ratio of 1.01.

Table 5.3. ***Logistic regression models of general male scores in the FY14-15 validation sample, by race/ethnicity***

	Model 1	Model 2	Model 3	Model 4
Assessment score	-	1.06***	1.06***	1.06***
Black	1.85***	-	0.98	0.93
Hispanic	1.13***	-	0.90***	0.83**
Native American	3.86***	-	2.08***	2.27**
Asian	0.57***	-	0.67***	0.48***
Black × assessment score	-	-	-	1.00
Hispanic × assessment score	-	-	-	1.00
Native American × assessment score	-	-	-	1.00
Asian × assessment score	-	-	-	1.01*
Constant	0.67***	0.12***	0.13***	0.13***
Model χ^2	1,963.73	16,981.38	17,256.26	17,265.95
Nagelkerke R^2	0.04	0.312	0.316	0.316

Note: Reported values are odds ratios. $N = 63{,}848$. White individuals are the referent group. *** $p \leq 0.001$. ** $p \leq 0.01$. * $p \leq 0.05$.

The four models presented in Table 5.3 were estimated for each tool and each sample. The findings for Models 3 and 4 were then consolidated into Table 5.4. This table also adds Average Marginal Effects (AMEs) estimates that convert the findings into percentage differences. For dummy variables, the AME is the mean of differences in predictions for each observation when moving from 0 to 1, leaving the rest of the data unchanged (see Williams, 2012). In this context, for example, the AME for Black individuals compared to the reference group of white individuals is interpreted as the mean percentage difference in predicted recidivism attributable to being Black rather than white, when holding the PATTERN score constant. Levels of statistical significance are given alongside the AME values. It is important to note that with large sample sizes such as this one, statistically significant findings are more likely even when substantive differences are small, and results should be interpreted in terms of both statistical and substantive significance.

Table 5.4. ***Differential prediction regression analyses — Summary of findings***

		FY14-15 Validation Sample						FY16 Revalidation Sample						FY17 Revalidation Sample					
		Model 3			Model 4			Model 3			Model 4			Model 3			Model 4		
		AME	OR	*p*	AME	OR	*p*	AME	OR	*p*	AME	OR	*p*	AME	OR	*p*	AME	OR	*p*
	Black	-0.46%	0.98		-0.39%	0.93		-2.10%	0.89	***	-2.02%	0.82	**	-3.04%	0.85	***	-3.01%	0.82	**
	Hispanic	-2.08%	0.90	***	-1.93%	0.83	***	-3.60%	0.82	***	-3.36%	0.73	***	-3.23%	0.84	***	-3.00%	0.73	***
	Native American	13.71%	2.08	***	14.27%	2.27	***	15.31%	2.30	***	15.61%	2.28	***	14.06%	2.17	***	17.76%	4.67	***
General	Asian	-7.40%	0.67	***	-5.75%	0.48	***	-8.27%	0.64	***	-7.75%	0.57	**	-7.74%	0.66	***	-6.97%	0.57	**
Male	Black × score		-			1.00			-			1.00			-			1.00	
	Hispanic × score		-			1.00			-			1.00			-			1.00	*
	Nat. Am. × score		-			1.00			-			1.00			-			0.98	***
	Asian × score		-			1.01	*		-			1.00			-			1.00	
	Black	4.97%	1.51	***	4.96%	1.69	***	4.31%	1.44	***	4.28%	1.77	***	4.74%	1.46	***	4.65%	1.70	***
	Hispanic	-0.22%	0.98		-0.23%	0.97		-0.61%	0.94		-0.60%	0.93		-0.04%	1.00		-0.06%	1.03	
	Native American	1.69%	1.16	**	2.60%	1.87	***	0.15%	1.01		1.82%	2.45	***	0.21%	1.02		3.34%	3.85	***
Violent	Asian	-5.19%	0.57	***	-4.67%	0.40	**	-2.90%	0.74		-2.67%	0.71		-3.26%	0.73		-3.26%	0.79	
Male	Black × score		-			1.00			-			0.99	*		-			0.99	
	Hispanic × score		-			1.00			-			1.00			-			1.00	
	Nat. Am. × score		-			0.99	**		-			0.98	***		-			0.96	***
	Asian × score		-			1.01			-			1.00			-			1.00	
	Black	-5.83%	0.71	***	-5.88%	0.70	***	-7.11%	0.66	***	-7.02%	0.74	*	-6.36%	0.69	***	-6.45%	0.61	**
	Hispanic	-5.76%	0.71	***	-5.83%	0.77	*	-5.27%	0.74	***	-5.29%	0.76		-4.36%	0.78	**	-4.42%	0.87	
	Native American	12.21%	1.91	***	13.63%	2.57	***	11.84%	1.86	***	11.33%	1.69		13.02%	2.00	***	15.82%	3.30	***
General	Asian	-9.20%	0.57	**	-8.82%	0.54	*	-3.96%	0.80		-1.08%	0.53		-3.82%	0.80		-1.24%	0.41	
Female	Black × score		-			1.00			-			1.00			-			1.00	
	Hispanic × score		-			1.00			-			1.00			-			1.00	
	Nat. Am. × score		-			0.99			-			1.00			-			0.98	*
	Asian × score		-			1.00			-			1.02			-			1.03	
	Black	1.24%	1.32	*	1.16%	1.11		2.11%	1.52	**	2.05%	1.23		0.49%	1.10		0.25%	0.95	
	Hispanic	0.19%	1.05		0.13%	1.13		-0.09%	0.98		-0.05%	0.93		-0.51%	0.90		-0.68%	1.26	
	Native American	-0.29%	0.93		-0.68%	0.64		-1.13%	0.74		-1.10%	0.67		0.35%	1.07		2.53%	2.83	**
Violent	Asian	-1.02%	0.75		-1.19%	0.77		1.39%	1.34		1.87%	1.11		-2.57%	0.51		-2.58%	0.55	
Female	Black × score		-			1.02			-			1.04			-			1.01	
	Hispanic × score		-			0.98			-			1.01			-			0.95	
	Nat. Am. × score		-			1.04			-			1.02			-			0.90	**
	Asian × score		-			0.99			-			1.05			-			0.99	

Note: AME = average marginal effects. OR = odds ratio. White individuals are the referent group. *** $p \leq 0.001$. ** $p \leq 0.01$. * $p \leq 0.05$.

Focusing first on the direct effects of race from the Model 3 analyses, there were 48 statistical tests conducted: 4 racial/ethnic groups (compared to white individuals) × 4 tools × 3 samples.[53] There were statistically significant results in 28 of these 48 tests (58 percent). Of these, 16 were overpredictions compared to white individuals, and 12 were underpredictions. Some of the larger conclusions to be drawn are:

- PATTERN overpredicted the risk of Black individuals relative to white individuals on the general recidivism tools. For males, the difference was not statistically significant in the FY 2014 to FY 2015 sample but was statistically significant in the FY 2016 sample, with a difference of 2 percent growing to 3 percent in the FY 2017 sample. PATTERN overpredicted recidivism for Black females compared to white females by 6 to 7 percent in all three samples.
- The violent recidivism tools underpredicted recidivism of Black individuals relative to white individuals by 4 to 5 percent for males and 1 to 2 percent for females (with no statistically significant difference in the FY 2017 sample).
- PATTERN overpredicted risk for Hispanic individuals relative to white individuals on the general recidivism tools by 2 to 6 percent. There were no statistically significant differences for Hispanic individuals with the violent recidivism tools.
- PATTERN underpredicted recidivism of Native American individuals relative to white individuals on the general recidivism tools in the range of 12 to 15 percent, with slight underprediction for the violent male tool (2 percent). There were no statistically significant Native American results on the violent female tool.
- General recidivism of Asian males was consistently overpredicted relative to white males in the range of 7 to 8 percent across the samples. For all but the FY 2016 violent female tool, Asian males and females had results in the direction of overprediction, but the results were not statistically significant in most other tests.

Finally, of the 48 interaction effect tests of Model 4, only nine were statistically significant. When statistically significant, the substantive effects were small, with odds ratios mostly around 2 percentage points. This indicates little evidence that the main effect relationships between race/ethnicity and recidivism varied substantially with changes in the score. Also of note, the level of differential prediction appeared more pronounced in the FY 2016 and FY 2017 samples than in the FY 2014 to FY 2015 sample, which may suggest evidence of a time trend.[54]

[53] By focusing on all five of the ethnic and racial groups in the data, these analyses go further than the race analyses previously reported in USDOJ (2019) and USDOJ (2020a). However, the differential prediction findings are not the result of the version 1.3 revisions or the changes in data sourcing from BOP's automation of PATTERN. The consultants performed the differential prediction analyses on the original 2019 PATTERN 1.2 dataset and obtained substantially similar results for the FY 2014 to FY 2015 validation sample. For example, for the general male tool, the main effect for Black individuals was statistically significant (OR = 0.95, AME = -1.0%); all other results from the version 1.3 analyses were the same as the original version 1.2 data in terms of direction and statistical significance, with similar substantive results.

[54] The unstandardized regression coefficients can also be used to calculate racial and ethnic imbalance points, i.e., the number of points it would take to equalize predictions for the race comparisons. These are calculated by dividing the unstandardized racial/ethnic logistic regression coefficient for a group by the unstandardized score coefficient. For example, for the general male tool and FY 2014 to FY 2015 validation sample, racial and ethnic groups could obtain scores equalized with white individuals by subtracting 1.4 points for Black individuals, subtracting 3.4 points for Hispanic individuals, adding 15.2 points for Native American individuals, and subtracting 13.5 points for Asian individuals. For the general female tool, balance could be achieved relative to white individuals by subtracting 6.1

These results are also displayed graphically below in Figure 5.1. The figure plots the differential prediction in the FY 2014 to FY 2015 validation sample for the general female model. If the tool is free from differential prediction, race-specific lines will overlap or at least be very closely matched; if there are differences, trend lines will separate, showing different recidivism rates by race for a given risk score.

In Figure 5.1, a vertical line is drawn at the PATTERN score of 35 to indicate the threshold between the lower and higher risk groupings (scores of 34 are low risk while scores of 35 are medium risk). Where the race and ethnicity trend lines intersect the vertical line, the figure shows the average marginal effect estimates of recidivism by the different racial and ethnic groups. Thus, at this low/medium RLC threshold, the difference represents an overprediction of about 10 percent for Black and Hispanic females compared to white females (54 percent vs. 44 percent). Taken together, the largest differences are in the middle score ranges; there is with less (or no) differential prediction in the tails of the distribution. The results presented in Table 5.4 underestimate the differential prediction at the risk level thresholds with the most impact. The differential prediction plots for all four tools within the validation and revalidation samples are provided in Appendix B.

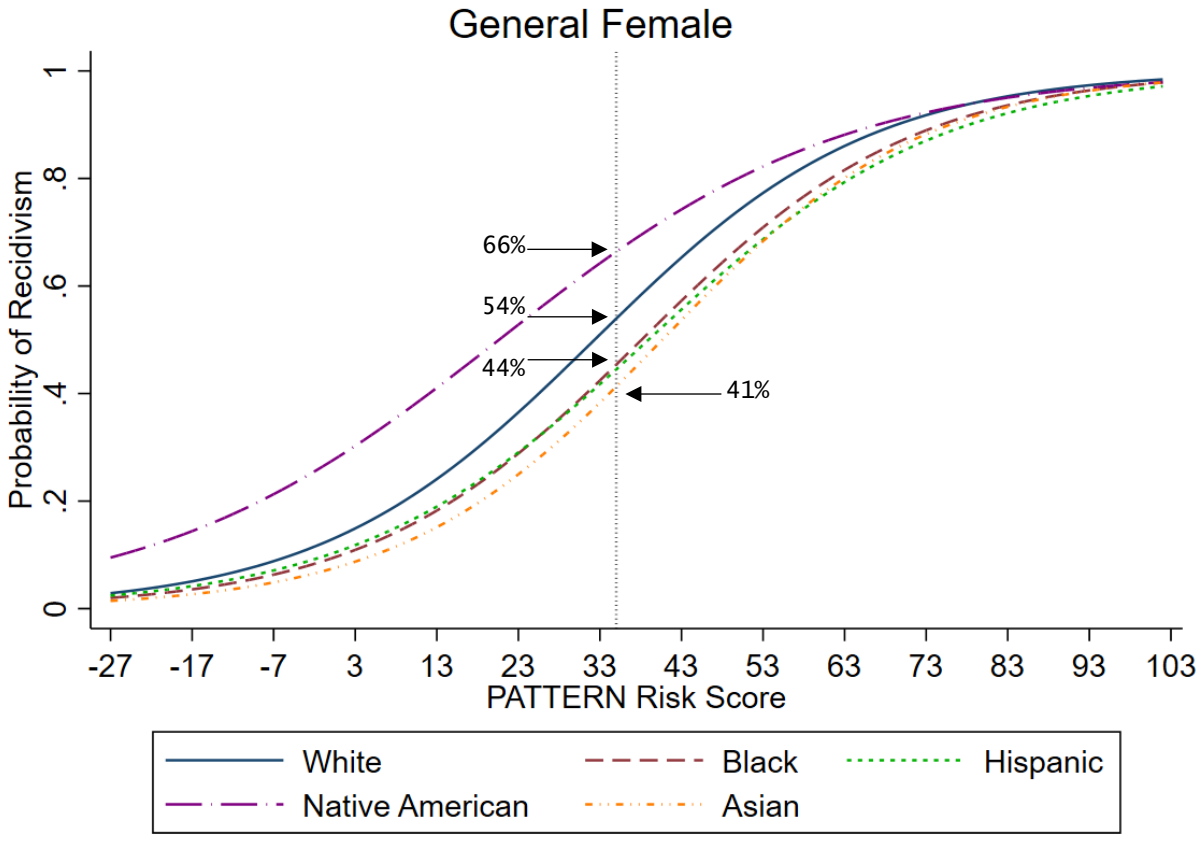


Figure 5.1. Differential prediction plot, general female PATTERN tool, FY14-15 validation sample.

points for Black individuals, subtracting 4.4 points for Hispanic individuals, adding 15.9 points for Native American individuals, and subtracting 10.4 points for Asian individuals.

Note: The gray vertical dotted line indicates a PATTERN score of 35, marking the risk level category threshold between low risk (score of 34) and medium risk (score of 35).

Part 6: Discussion and Conclusion

The results from this annual review and revalidation report indicate that PATTERN 1.3 displays a high level of predictive accuracy across all four of its tools. The PATTERN risk scores are accurate predictors of recidivism at both the first and last assessments. Scores from the last assessment are appreciably more accurate than those from the first. Risk level analyses further reveal an increased risk of recidivism for each successively higher RLC across all four tools. The PPVs indicate that for the male and female general tools, over 60 percent of individuals in the higher risk group (medium and high RLCs) recidivated. The NPVs indicate that over 70 percent of individuals in the lower risk group (minimum and low RLCs) did not recidivate. For the violent recidivism tools, the PPVs were lower (20 to 30 percent) and the NPVs higher (over 90 percent).

Dynamic analyses further indicate that individuals are capable of changing their risk score and level during confinement, and that these changes relate to their recidivism outcome. Individuals who increased their risk score/level from the first to last assessment were generally more likely to recidivate, whereas those who lowered their risk score/level were less likely to recidivate.

PATTERN 1.3 also shows relatively high predictive accuracy across the five racial/ethnic groups examined, with AUCs above 0.70 for all but two of the 20 race/tool combinations for the FY 2014 to FY 2015 validation sample, all but one for the FY 2016 revalidation sample, and all but four for the FY 2017 revalidation sample. The predictive value and differential prediction results, however, are mixed and complex. It is important to emphasize that the differential prediction findings are not the result of changes made in version 1.3; they were also present in versions 1.2 and 1.2-R, and transitioning to PATTERN 1.3 will neither exacerbate nor solve the differential prediction issues (see footnote 53, *supra,* for a discussion of empirical results comparing differential prediction across versions). Due to the importance of the FSA mandate to examine the risk and needs assessment system for racial and ethnic neutrality, and to minimize racial disparities to the extent possible, these results will be a central focus of subsequent review and revalidation efforts.

6.1 Addressing Differential Prediction Concerns

The race and ethnicity findings from this review and revalidation are multifaceted. PATTERN overpredicts for nonwhite individuals relative to white individuals for some tools and underpredicts for others. Some differences relative to white individuals are modest, while others are more substantial — such as the up to 7 percent overprediction of Black female general recidivism, the 12 to 15 percent underprediction of Native American males and females for some tools,[55] and the 5 to 8 percent overprediction of Asian individuals for some tools.

[55] Until recently, tribal reservations were reportedly not required to provide arrest information to the National Law Enforcement Telecommunications System. With respect to the PATTERN tools, absence of this information could affect both the recidivism outcomes (unless very recent) and the criminal history predictor, with potential systematic underreporting of Native American criminal histories. With some of the criminal history missing, one would expect PATTERN to underestimate the risk and underpredict for such cases, though the problem may slowly correct itself as reporting improves. In addition, "for serious crimes the federal justice system can supersede tribal authority"; thus, the federal system processes cases involving Native American individuals who would otherwise be processed in state courts (Ulmer & Bradley, 2018, p. 752). Consequently, the federal Native American prison population may share similarities with state prison populations, and there may be systematic reporting and offender type differences at play with the Native American population.

Although risk instruments have been used in the criminal justice context for around 100 years, only more recently have issues of race and prediction been analyzed closely. In 2016, a ProPublica article by Angwin et al. (2016) sparked debate over whether the COMPAS risk tool was biased against Black individuals as used to inform pretrial release decisions in Broward County, Florida. Although Angwin and colleagues claimed the instrument was biased due to differences in FPRs, other scholars issued replies (Flores et al., 2016; see also Dieterich, Mendoza, & Brennan, 2016) demonstrating, through differential prediction analyses, that COMPAS predicted recidivism similarly for white and Black individuals.[56] Since then, several important works from scholars in the fields of statistics, machine learning, and predictive analytics have established that multiple definitions of racial fairness exist, and that in real-world applications, these notions of fairness conflict. When base rates of offending differ by group membership (such as race), a risk tool cannot simultaneously achieve all notions of fairness (Berk et al., 2021; Chouldechova, 2017; Klineberg, Mullainathan, & Raghavan, 2016).

The reasons why a tool cannot satisfy all definitions of fairness relate to the different risk-prediction outcomes one might focus on. To illustrate, assume a binary risk-prediction outcome such as lower risk versus higher risk. Given that individuals must fall into one of these two risk groups and that in a recidivism study everyone will have either recidivated or not recidivated, everyone in the study can be classified into one of four mutually exclusive outcomes: false positive, false negative, true positive, or true negative. Several research articles illustrate how these outcomes can be used to populate a 2 × 2 table (Berk et al., 2021; Chouldechova, 2017). The descriptive statistics reported in section 3 are derived from calculations in these 2 × 2 tables of false positives and negatives and true positive and negatives.

One notion of fairness might focus on the similarities or differences among groups in terms of the proportions of true positives among all of the positive predictions. In other words, are the PPVs the same for white, Black, Hispanic, Asian, and Native American individuals? In this case, the denominator of the equation is the overall number of positive predictions. Another notion of fairness might focus on the similarities or differences among groups with respect to the proportion of false positives among all those who did not recidivate. Here the denominator of the equation would be all false positives plus all true negatives constituting all individuals who did not recidivate.

Similar measures associated with distinct notions of fairness can be constructed for all negative predictions and for all individuals who recidivated. In essence, there are different dimensions of fairness related to an instrument's ability to identify those who did or did not recidivate out of different populations of all individuals who actually recidivated, who actually did not recidivate,

[56] Another often-cited milestone of this interest in race and prediction is a 2014 speech by U.S. Attorney General Eric Holder (Holder, 2014) warning against the misuse of risk assessments in criminal justice, though Attorney General Holder's views appear consistent with the employment of a risk instrument used to incentivize recidivism reduction programming and additional early release time for lower risk individuals. As Monahan and Skeem (2016, p. 681) observe, "Holder (2014) celebrated the momentum building behind data-driven justice reform and specifically supported the use of risk assessment in back-end applications designed to reduce risk: 'Data can help design paths for federal inmates to lower these risk assessments, and earn their way towards a reduced sentence, based on participation in programs that research shows can dramatically improve the odds of successful reentry.' In Holder's view, everyone — even high-risk inmates — should have the chance to reduce his or her prison time."

who were predicted to recidivate, and who were predicted not to recidivate. These measures are captured by the PPVs, NPVs, FPRs, and FNRs reported and discussed in section 3. (There are additional notions of fairness based on combinations of these four measures; see Berk et al., 2021 for a comprehensive review.) Since these measures use different denominators consisting of different combinations of true and false positives and true and false negatives, it is mathematically impossible to satisfy parity in all of them if base rates of recidivating differ across groups. Consequently, "the goal of complete race or gender neutrality is unachievable" (Berk et al., 2021, p. 20).

Importantly, the *Standards for Educational and Psychological Testing* do not suggest that a tool be invalidated or discarded based on statistically significant differential prediction results. According to the *Standards,* "subgroup mean differences do not in and of themselves indicate lack of fairness, but such differences should trigger follow-up studies, where feasible, to identify the potential causes of such differences," and "what constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws" (Standard 3.6, AERA et al., 2014, p. 65). The nature of limitations in criminal justice data and the complexities and concerns related to race, crime, and the criminal justice system provide a unique context for actuarial instruments. In education, a standardized test item might rightly be discarded over concerns of subgroup differences, particularly since test makers have a nearly limitless ability to generate new test questions and retain ones which are unbiased. In contrast, criminal justice data are imperfect and limited, which may have implications for how differential prediction results are ultimately received by the criminal justice community.

Although some risk tool validations have reported well-calibrated results comparing just two race groups (usually Black and white individuals; see e.g., Flores et al., 2016 [COMPAS]; Skeem & Lowenkamp, 2016 [most of the PCRA tools]), there are also several examples of tools that exhibit some degree of differential prediction, but which have not been invalidated on those grounds (see, e.g., DeMichele et al., 2020 [Arnold Public Safety Assessment]; Skeem & Lowenkamp, 2016 [PCRA];[57] PCS, 2018 [Pennsylvania Commission on Sentencing's risk tool for sentencing]).

PATTERN presents a novel application of differential prediction results in the criminal justice context for several reasons. Unlike most studies of criminal justice risk instruments, which analyze just two groups (usually white and Black individuals), BOP classifies the individuals in its custody according to the five racial and ethnic groups analyzed in this report. With some groups overpredicted compared to white individuals and others underpredicted relative to white individuals, addressing differential prediction through simple reweighting or altering the risk items is unlikely to succeed.[58]

---

[57] While there were no statistically significant results using a propensity-score-matched sample, in the unmatched supplemental analyses there were statistically significant results for the any arrest outcome, "which suggests overestimation of arrest for White offenders" (Skeem & Lowenkamp, 2016, p. 695).

[58] Even equalizing white, Black, and Hispanic calibration by simply dropping or reweighting items appears unlikely to succeed in addressing differential prediction. As a preliminary exploration, the NIJ contractors pursued a methodology for equalizing prediction by estimating regressions with item-by-race interaction terms predicting recidivism and then examining which specific item generated the most differential prediction. That item's score would then be adjusted, and the process repeated with a new regression model containing the remaining item × race interactions. While this approach was able to reduce bias for some comparisons, it is not a viable option with multiple groups, some of which are overpredicted and others underpredicted relative to white individuals. Mitigating

There are several suggestions for addressing racial fairness in predictive analytics. Many novel potential solutions have recently emerged in the literature. Since 2016, there has been "an unprecedented explosion" of research in fairness and prediction (Chouldechova & Roth, 2018, p. 1), much of it spawned by advances in machine learning, artificial intelligence, and big data. In the criminal justice context, scholars are exploring some of these solutions in academic articles, but they caution that adoption for real-world applications should be slow and deliberate:

> Corrections for unfairness combine technical challenges with policy challenges. We have currently no definitive responses to either. Progress will likely come in many small steps beginning with solutions from tractable, highly stylized formulations. One must avoid vague or unjustified claims or rushing these early results into the policy arena. Because there is a large market for solutions, the temptations will be substantial. (Berk et al., 2021, p. 31)

Similarly, Skeem and Lowenkamp (2020, p. 274) explore several debiasing options and achieve success in reducing bias but warn that "debiasing approaches like those we tested here should be used with caution."

Many of these solutions would require the explicit inclusion of race and ethnicity in the risk assessment and thus trigger equal protection concerns. One approach would be to create race-specific tools, just as PATTERN already includes gender-specific tools.[59] Since there are some factors that affect prediction differently by race and ethnicity, race-specific scales would likely generate even more accurate scores; recidivism-based thresholds could then be applied to satisfy this notion of racial fairness. However, the equal protection concerns raised by race-informed solutions deserve careful consideration and deliberation (see, e.g., Huq, 2018). As Goel and colleagues (2021, p. 15) note, "In practice, it can be legally challenging, though not impossible, to base risk assessments on race or gender." There are a number of other approaches discussed in the literature.[60]

Given the novelty and advancement of approaches for debiasing risk tools, and based on the results presented here, there are no simple solutions to this complex problem. It is likely that differential prediction can be reduced and possibly eliminated, but not without trade-offs. Some methods may result in lower overall predictive accuracy. Others may achieve both accuracy and fairness, but only by using race as a factor, which raises equal protection concerns. While NIJ's consultants have experimented with some solutions, the nature of their findings suggests that deliberate study and engagement with stakeholders and experts are warranted to identify an optimal path forward.

6.2 Next Steps

differences with one group may in turn exacerbate differences with another (e.g., with the violent male tool, in which Black recidivism is underpredicted but Hispanic and white predictions are currently about equal).

[59] For discussions of the importance of gender-specific risk instruments, see Funk (1999), Hamilton et al. (2016), and Salisbury et al. (2009).

[60] For examples, see Berk et al. (2021), Goel et al. (2021), Huq (2019), and Skeem and Lowenkamp (2021).

Since PATTERN version 1.3 addresses the coding and specification errors affecting versions 1.2 and 1.2-R, and since version 1.3 does not exacerbate the existing differential prediction issues, it is recommended that the Department adopt PATTERN 1.3. The NIJ consultants will continue to investigate potential solutions for the differential prediction issues identified during this review. These efforts will include testing the emerging debiasing techniques discussed in section 6.1 and engaging with stakeholders to explore the most promising and supportable approaches to reducing differential prediction through potential revisions to PATTERN.

A subsequent review and revalidation report including an additional cohort of individuals released from BOP custody in FY 2018 will be released in late 2022. As required by Section 3634 of the FSA, this report will assess and summarize the predictive validity, dynamic validity, and racial and ethnic neutrality of PATTERN.

More generally, NIJ's consultants will continue to collaborate with USDOJ subject matter experts, BOP staff, and the IRC to explore whether further refinements to items and the scoring scheme of PATTERN may help improve the equitability, efficiency, and predictive validity of the risk assessment system. This will include an exploration of the inclusion of additional information (e.g., more recent programming data) and how it may be used to improve prediction and fairness. All additional recommended refinements to the risk assessment tool will be submitted to the attorney general for review and consideration.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Angwin. J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica. Retrieved from www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods and Research, 50*(1), 3-44.

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior, 2*(1), 435-463.

Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice and Behavior, 29*(4), 355-379.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153-163.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint, arXiv:1810.08810*.

DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., & Comfort, M. (2020). Public safety assessment: Predictive utility and differential prediction by race in Kentucky. *Criminology & Public Policy, 19*(2), 409-431.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Federal Bureau of Prisons (BOP). (2006). *Inmate security designation and custody classification* (policy number 5100.08). Washington, DC: U.S. Department of Justice, Federal Bureau of Prisons.

Federal Bureau of Prisons (BOP). (2011). *Inmate discipline program* (policy number 5270.09). Washington, DC: U.S. Department of Justice, Federal Bureau of Prisons.

Federal Bureau of Prisons (BOP). (2019). *Key components of the Federal Bureau of Prisons' current needs assessment system.* Washington, DC: U.S. Department of Justice, Federal Bureau of Prisons.

Federal Bureau of Prisons (BOP). (2020). *Violent offense codes for PATTERN risk assessment*. Washington, DC: U.S. Department of Justice, Federal Bureau of Prisons.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *Federal Probation, 80*(2), 38-46.

Funk, S. J. (1999). Risk assessment for juveniles on probation: A focus on gender. *Criminal Justice and Behavior, 26*(1), 44-68.

Goel, S., Shroff, R., Skeem, J., & Slobogin, C. (2021). The accuracy, equity, and jurisprudence of criminal risk assessment. In R. Vogl, ed. *Research Handbook on Big Data Law*. Northampton, MA: Edward Elgar.

Hamilton, Z., Duwe, G., Kigerl, A., Gwinn, J., Langan, N., & Dollar, C. (2021). Tailoring to a mandate: The development and validation of the Prisoner Assessment Tool Targeting Estimated Risk and Needs (PATTERN). *Justice Quarterly*. In press.

Hamilton, Z., Kigerl, A., Campagna, M., Barnoski, R., Lee, S., Van Wormer, J., & Block, L. (2016). Designed to fit: The development and validation of the STRONG-R recidivism risk assessment. *Criminal Justice and Behavior, 43*(2), 230-263.

Holder, E. (2014). *Attorney General Eric Holder speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting.* Washington, DC: U.S. Department of Justice. Retrieved from http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th.

Huq, A. Z. (2018). Racial equity in algorithmic criminal justice. *Duke Law Journal, 68*(6), 1043-1134.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint, arXiv:1609.05807*.

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, 12*, 489-513.

National Institute of Justice (NIJ). (2021). *2020 review and revalidation of the First Step Act risk assessment tool*. Washington, DC: U.S. Department of Justice, National Institute of Justice.

Pennsylvania Commission on Sentencing. (2018). *Racial impact analysis of the proposed risk assessment scales*. Retrieved from http://pcs.la.psu.edu/publications-and-research/risk-assessment/phase-iii-reports/racial-impact-analysis-of-proposed-risk-assessment-scales-may-2018/view.

Salisbury, E. J., Van Voorhis, P., & Spiropoulos, G. V. (2009). The predictive validity of a gender-responsive needs assessment: An exploratory study. *Crime and Delinquency, 55*(4), 550-585.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences and the Law, 31*(1), 8-22.

Skeem, J. L., & Lowenkamp, C. T. (2020). Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences and the Law, 38*(3), 259-278.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology, 54*(4), 680-712.

Ulmer, J. T., & Bradley, M. S. (2018). Punishment in Indian country: Ironies of federal punishment of Native Americans. *Justice Quarterly, 35*(5), 751-781.

U.S. Department of Justice (USDOJ). (2019). *The First Step Act of 2018: Risk and needs assessment system*. Washington, DC: U.S. Department of Justice, Office of the Attorney General.

U.S. Department of Justice (USDOJ). (2020a). *The First Step Act of 2018: Risk and needs assessment system – Update*. Washington, DC: U.S. Department of Justice, Office of the Attorney General.

U.S. Department of Justice (USDOJ). (2020b). *First Step Act implementation fiscal year 2020 90-day report.* Washington, DC: U.S. Department of Justice, Office of the Attorney General.

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal, 12*(2), 308-331.

Appendix A: Distribution Tables

Table A1. ***Risk level categories and recidivism by race and ethnicity, FY 2014-2015 validation sample***

	Male General		Male Violent		Female General		Female Violent	
	Did Not Recidivate	*Recidivated*	*Did Not Recidivate*	*Recidivated*	*Did Not Recidivate*	*Recidivated*	*Did Not Recidivate*	*Recidivated*
White								
Minimum	5,395	673	6,249	80	1,622	174	2,436	38
Low	4,455	1,956	8,041	662	1,148	761	1,839	117
Medium	1,532	1,755	2,292	551	273	415	150	27
High	1,617	4,346	2,708	1,146	53	167	4	2
Black								
Minimum	1,635	202	1,820	37	789	77	1,133	24
Low	3,876	1,723	8,564	1,170	704	298	1,038	79
Medium	2,534	2,860	3,875	1,239	189	210	126	47
High	3,350	9,379	5,363	3,491	43	169	18	14
Hispanic								
Minimum	1,740	207	2,003	29	897	99	1,381	20
Low	3,188	1,252	5,971	484	943	389	1,173	83
Medium	1,438	1,441	1,612	340	161	188	87	13
High	1,120	2,767	1,911	803	27	60	6	1
Native American								
Minimum	95	18	85	1	50	17	102	2
Low	186	200	449	54	92	117	223	11
Medium	143	359	306	81	34	69	42	15
High	196	1,032	810	443	3	24	8	3
Asian								
Minimum	387	25	427	2	153	10	187	2
Low	309	84	465	18	52	21	69	4
Medium	87	63	99	11	10	12	3	0
High	67	156	119	37	4	3	0	0

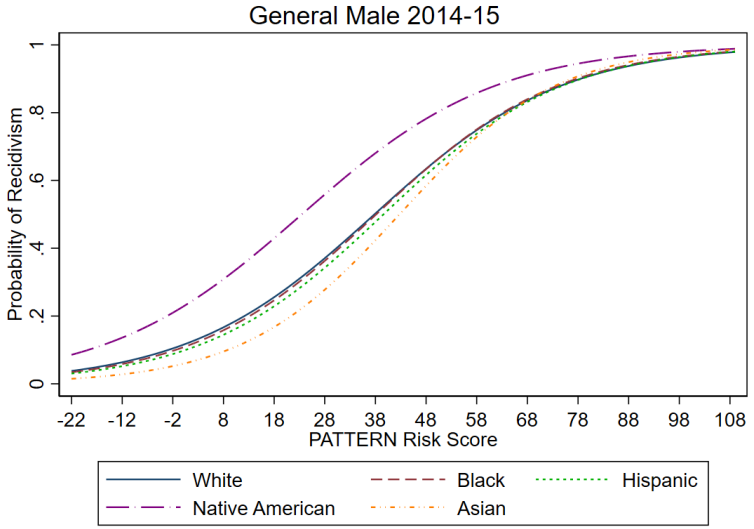
Table A2. ***Risk level categories and recidivism by race and ethnicity, FY 2016 revalidation sample***

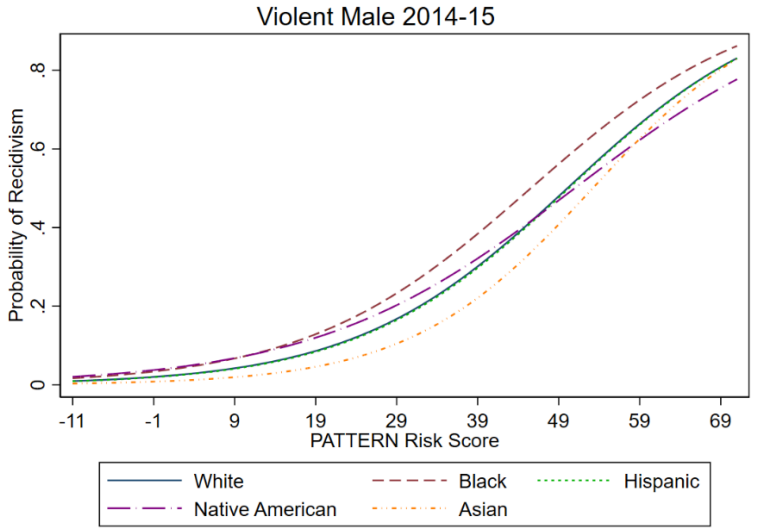
	Male General		Male Violent		Female General		Female Violent	
	Did Not Recidivate	*Recidivated*	*Did Not Recidivate*	*Recidivated*	*Did Not Recidivate*	*Recidivated*	*Did Not Recidivate*	*Recidivated*
White								
Minimum	3,015	359	3,440	42	831	99	1,320	23
Low	2,342	1,079	4,478	364	623	422	979	71
Medium	912	1,048	1,263	263	153	229	73	11
High	884	2,457	1,539	707	34	92	5	1
Black								
Minimum	1,078	117	1,175	21	411	39	615	11
Low	2,586	1,088	5,280	675	365	155	511	62
Medium	1,538	1,548	2,219	652	106	127	71	23
High	1,839	5,116	2,939	1,949	37	62	4	5
Hispanic								
Minimum	1,165	140	1,315	14	460	56	721	11
Low	2,124	727	3,686	276	483	205	616	43
Medium	849	813	923	182	86	103	41	9
High	582	1,472	1,032	444	13	39	4	0
Native American								
Minimum	47	11	43	2	34	7	62	1
Low	103	125	258	25	45	62	109	4
Medium	79	200	167	53	14	37	30	6
High	84	581	459	223	3	16	3	3
Asian								
Minimum	230	15	250	2	85	6	106	2
Low	164	48	251	18	27	15	33	3
Medium	52	42	55	4	3	7	2	1
High	35	64	50	20	1	3	0	0

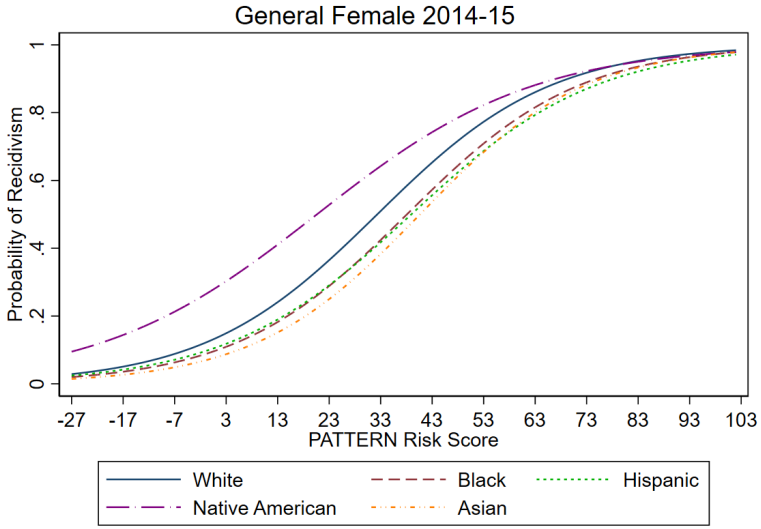
Table A3. ***Risk level categories and recidivism by race and ethnicity, FY 2017 revalidation sample***

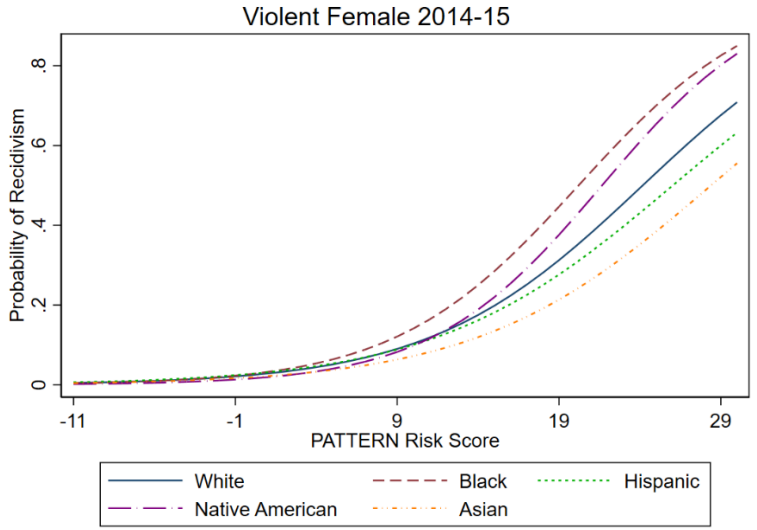
	Male General		Male Violent		Female General		Female Violent	
	Did Not Recidivate	*Recidivated*	*Did Not Recidivate*	*Recidivated*	*Did Not Recidivate*	*Recidivated*	*Did Not Recidivate*	*Recidivated*
White								
Minimum	2,596	344	3,000	30	773	95	1,142	17
Low	2,001	955	3,804	345	534	349	898	67
Medium	710	909	1,130	243	133	221	79	19
High	768	2,394	1,441	684	30	96	2	7
Black								
Minimum	943	117	1,008	27	410	35	587	5
Low	2,155	911	4,471	602	332	136	483	49
Medium	1,329	1,422	1,987	614	97	119	80	25
High	1,723	4,907	2,845	1,953	27	90	9	8
Hispanic								
Minimum	1,015	124	1,121	19	382	45	621	14
Low	1,759	654	3,218	271	453	212	609	39
Medium	750	821	931	180	91	106	56	13
High	569	1,545	1,020	477	20	50	6	1
Native American								
Minimum	34	14	38	1	30	12	58	0
Low	71	121	203	37	36	57	90	18
Medium	71	198	168	44	18	40	40	7
High	111	577	481	225	5	21	6	0
Asian								
Minimum	168	9	189	0	59	2	71	0
Low	155	51	230	19	24	16	36	2
Medium	42	36	54	11	3	6	1	0
High	30	59	39	8	0	0	0	0

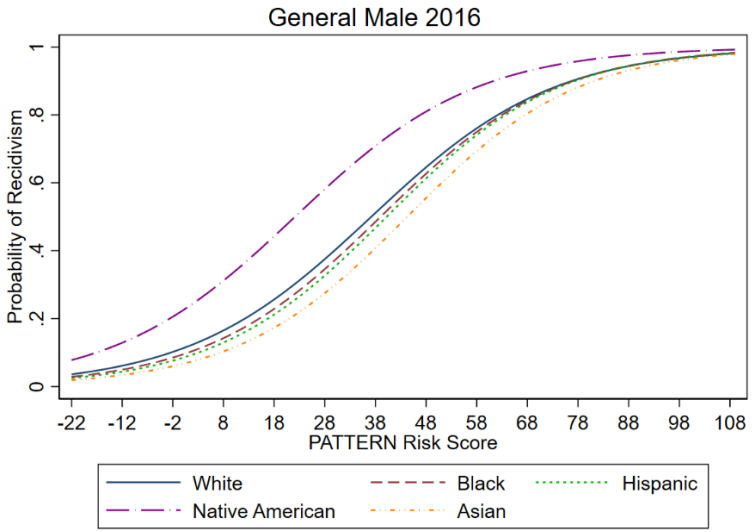
Appendix B: Differential Prediction Plots

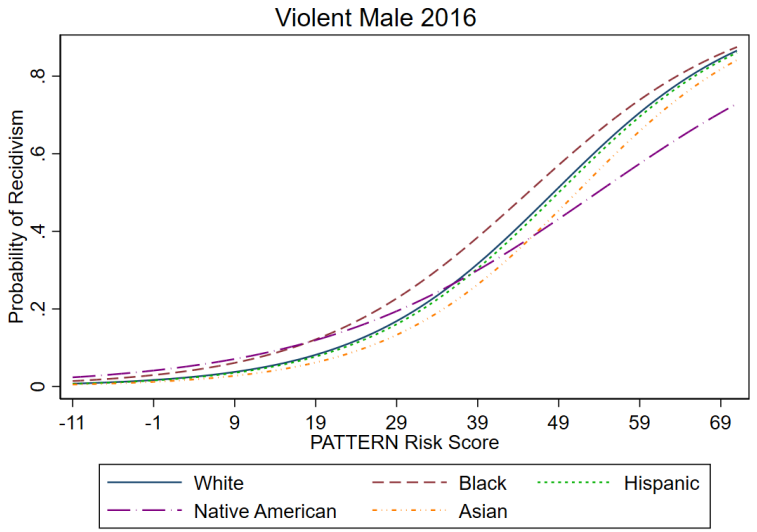


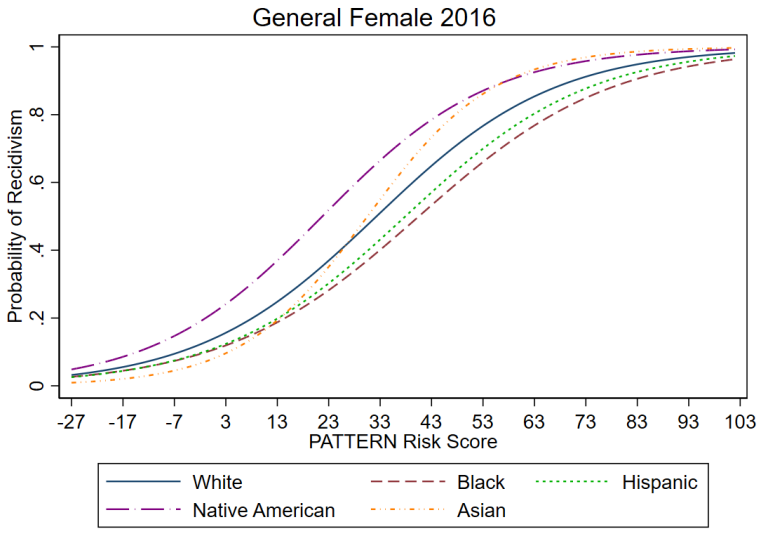


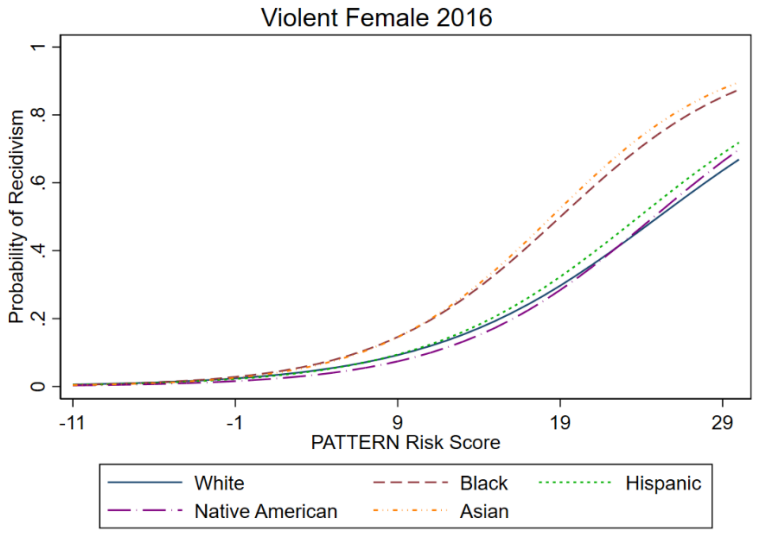


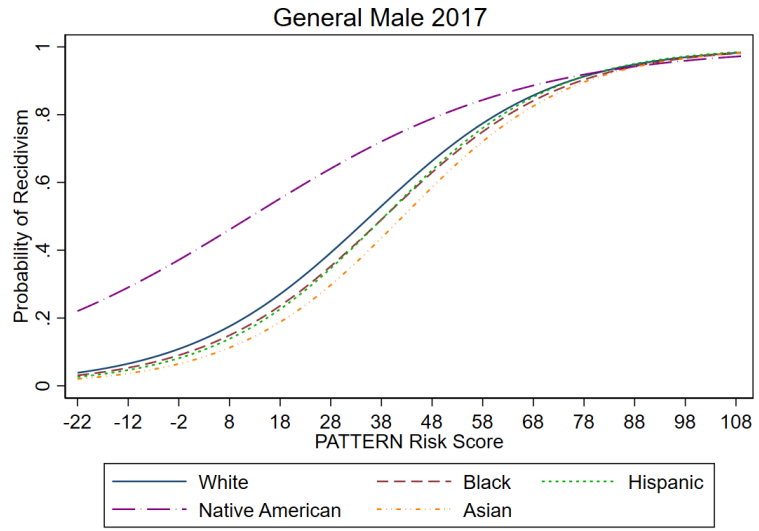


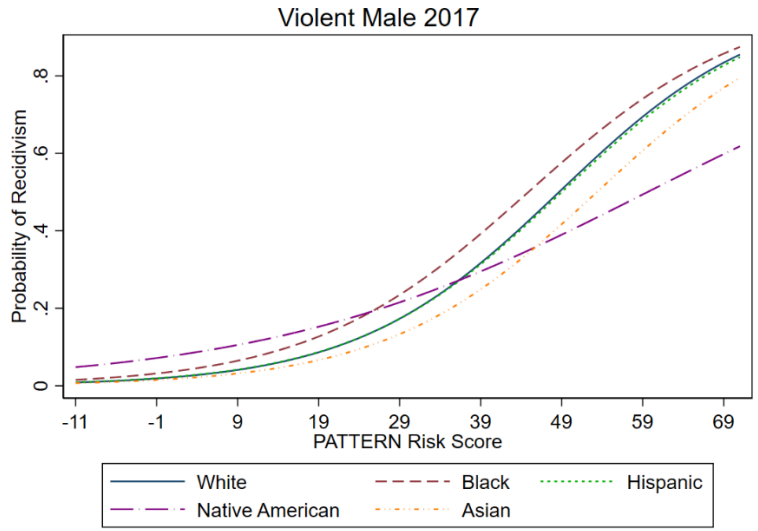

General Male 2016
Probability of Recidivism
0
.2
.4
.6
.8
1
-22 -12 -2 8 18 28 38 48 58 68 78 88 98 108
PATTERN Risk Score
White
Black
Hispanic
Native American
Asian

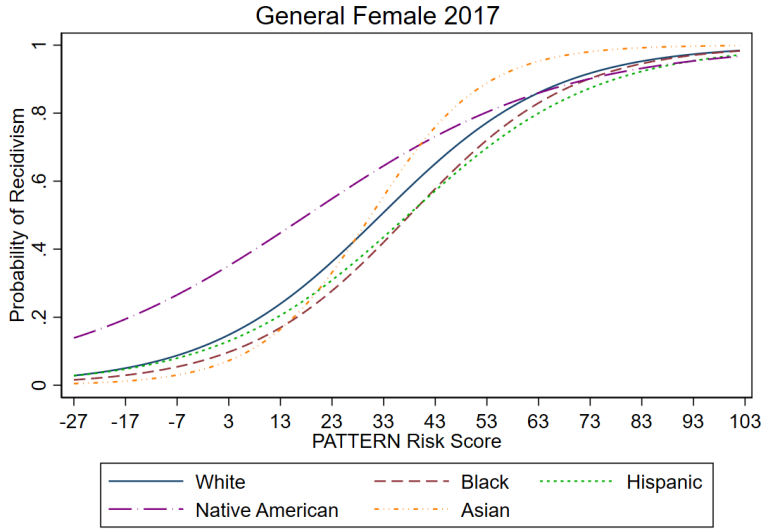

Violent Male 2016
Probability of Recidivism
0
.2
.4
.6
.8
-11 -1 9 19 29 39 49 59 69
PATTERN Risk Score
White
Black
Hispanic
Native American
Asian

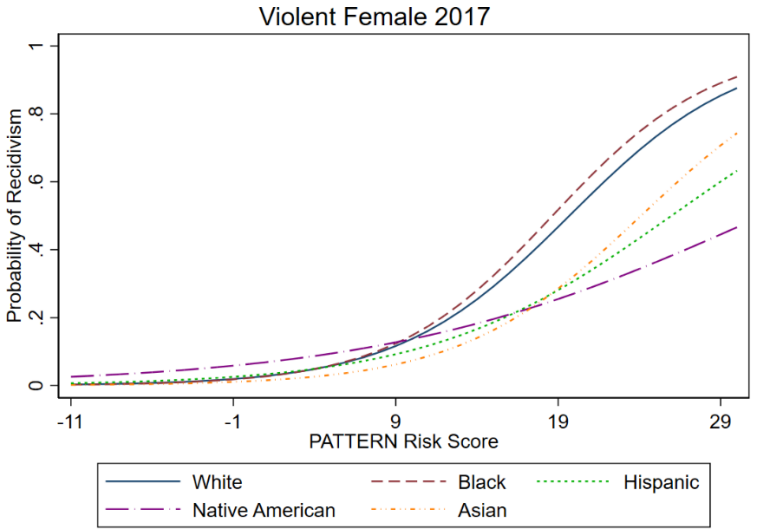

General Female 2016
Probability of Recidivism
0
.2
.4
.6
.8
1
-27 -17 -7 3 13 23 33 43 53 63 73 83 93 103
PATTERN Risk Score
White
Black
Hispanic
Native American
Asian


Violent Female 2016
Probability of Recidivism
0
.2
.4
.6
.8
1
-11 -1 9 19 29
PATTERN Risk Score
White
Black
Hispanic
Native American
Asian


General Male 2017
Probability of Recidivism
0
.2
.4
.6
.8
1
-22 -12 -2 8 18 28 38 48 58 68 78 88 98 108
PATTERN Risk Score
White
Black
Hispanic
Native American
Asian


Violent Male 2017
Probability of Recidivism
0
.2
.4
.6
.8
-11 -1 9 19 29 39 49 59 69
PATTERN Risk Score
White
Black
Hispanic
Native American
Asian


General Female 2017
Probability of Recidivism
0
.2
.4
.6
.8
1
-27 -17 -7 3 13 23 33 43 53 63 73 83 93 103
PATTERN Risk Score
White
Black
Hispanic
Native American
Asian


Violent Female 2017
Probability of Recidivism
0
.2
.4
.6
.8
1
-11 -1 9 19 29
PATTERN Risk Score
White
Black
Hispanic
Native American
Asian