Written Testimony of Melissa Hamilton to the

United States House of Representatives

Committee on the Judiciary

Subcommittee on Crime, Terrorism, and Homeland Security

Oversight Hearing on the Bureau of Prisons and Implementation of the First Step Act

**********

The Department of Justice, National Institute of Justice, and the Bureau of Prisons are to be commended for managing the challenges of a tight time frame given by the First Step Act on introducing a risk and needs system. Further, offering a gender-sensitive scoring system is supportable by scientific studies and is likely to withstand most legal challenges considering that criminal justice statistics consistently show that women recidivate at far lower rates than males.[1] The DOJ has released a risk and needs system, as required by the First Step Act. Yet, as with any newly developed risk assessment system, improvements can be made. I address herein three major points applicable to any risk assessment tool: transparency, validity, and fairness. I then make suggestions on how the risk and needs tool initially released can be modified to better serve the legislation's purposes, while also addressing broader concerns.

## *Transparency*

A critical foundation for successful implementation of a new risk assessment system is trust by stakeholders (e.g., judges, prosecutors, defense counsel, corrections officials). The black-box nature of most tools is commonly cited as undermining confidence that these tools have sufficient predictive ability, are reliably scored, and are equitable. Users and stakeholders who did not have faith in the tool can undermine the system by finding ways to score the tool differentially, blatantly or secretly override scores, ignore scores when making decisions, and otherwise challenge the system.[2] These responses would be unfortunate considering the many benefits that the First Step Act is designed to achieve for improving efficiencies, saving taxpayer dollars, incentivizing offenders to undertake beneficial programming, and protecting the public.

[1] *See generally* Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231 (2015), http://epubs.surrey.ac.uk/id/eprint/842342.

[2] Faye S. Taxman & Amy Dezember, *The Value and Importance of Risk and Need Assessment (RNA) in Corrections & Sentencing: An Overview of the Handbook, in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 22, 40 (Faye S. Taxman ed., 2017).

The BOP can bolster stakeholder trust by doing more to gain their buy-in. Transparency is "critically" important to stakeholder and user acceptance.[3] External trust and support can further leverage proper use by internal staff.[4] One reason to focus on the inclusion of stakeholders is that successful implementation will require a significant cultural shift in the criminal justice agencies involved and those who participate in it.

> Policymakers' need to know the subsequent strategies for public safety and recidivism reduction might begin with a simple question: Do risk assessment instruments reliably predict recidivism? The short answer, according to years and volumes of research, is resoundingly: yes. But we must be mindful of what saying yes may mean. Adoption of a risk assessment tool goes hand-in-hand with fundamentally altering approaches to reentry and correctional management, supervision, services, and more broadly criminal justice practice. Ultimately, the process of implementing risk assessments within an agency should consist of more than simply adding a tool to the agency portfolio; it should result in a shift of corrections culture, practices, and policies.[5]

Reference will be made herein to a report titled *The First Step Act of 2018: Risk and Needs Assessment System* issued by the DOJ on July 19, 2019 ("DOJ Report"). The DOJ Report implies that involved agencies have been transparent by reporting on a variety of information points about PATTERN. It is appreciated that the DOJ Report is clear about the point scoring system. It is also the case that the DOJ Report provides some statistics on its validation and recidivism results. However, the DOJ Report is lacking in so many other transparency areas and, as a result, the buy-in and confidence of key stakeholders may be lacking.

For example, the National Association of Criminal Defense Lawyers has submitted a FOIA request on October 8, 2019, to release the datasets on which PATTERN was developed and validated. Independent audits by third parties is consistently highlighted as a key mechanism for trust in risk assessment tools. Releasing criminal justice datasets for independent researchers to conduct audits is not novel. The U.S. Sentencing Commission, for example, releases its datasets quarterly and annually with a host of specific information on offending behavior, sociodemographic information, and sentencing outcomes. It is not evident that there have been any negative consequences to such practices.

The datasets requested are readily available. They have already been anonymized and delivered to the external PATTERN developers. Hence, there is no obvious burden to personnel

---

[3] Faye S. Taxman & Amy Dezember, *The Value and Importance of Risk and Need Assessment (RNA) in Corrections & Sentencing: An Overview of the Handbook, in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 22, 37 (Faye S. Taxman ed., 2017).

[4] Faye S. Taxman & Amy Dezember, *The Value and Importance of Risk and Need Assessment (RNA) in Corrections & Sentencing: An Overview of the Handbook, in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 22, 47 (Faye S. Taxman ed., 2017).

[5] Faye S. Taxman & Amy Dezember, *The Value and Importance of Risk and Need Assessment (RNA) in Corrections & Sentencing: An Overview of the Handbook, in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 22, 22 (Faye S. Taxman ed., 2017).

or resources for the BOP and NIJ to publicly release these datasets. Doing so would be consistent with the First Step Act's intention regarding public availability of detailed information about the BOP risk and needs system developed and implemented under the legislation. The datasets would allow researchers to replicate the data points in the DOJ report, which is particularly important considering the existence of multiple data errors evident therein. Independent researchers would also then be able to calculate a host of measures that are absent in the DOJ Report yet are relevant to a more holistic analysis of the validity, reliability, and equity of the PATTERN tool as it exists. Further discussion about what a third party audit may helpfully reveal is discussed later herein.

Issues with transparency are also evident with the DOJ Report containing citations to papers that are not publicly available (e.g., reports cited at note 8 on page 64 and note 25 on page 66). The Brennan Center for Justice in a submission to the NIJ concerning PATTERN dated September 3, 2019 complains that it requested release of information on the BRAVO/BRAVO-R tools that the DOJ Report indicates are foundations for PATTERN (e.g., note 8 on page 64 of the DOJ Report), yet were rebuffed because of proprietary claims. This initial assertion of secrecy is deeply concerning.

## *Validity*

The terms "validation" and a "validated tool" imply a strength of predictive ability. Yet, in reality, the terms only refer to a tool that can perform slightly better than chance. Thus, a validated tool simply means one that distinguishes recidivists from non-recidivists marginally better than the proverbial flip of a coin.[6]

The DOJ Report claims that PATTERN has been "validated." It uses as its primary basis for such claim a metric known as the area under the curve (AUC). The AUC is derived from a statistical plotting of true positives and false positives across a risk tool's rating system.[7] More specifically, the AUC is a discrimination index that represents the probability that a randomly selected recidivist received a higher risk classification than a randomly selected non-recidivist.[8] The size of the risk scale differential between them is irrelevant; as long as the risk classification of the

[6] KiDeuk Kim & Grant Duwe, *Improving the Performance of Risk Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 189, 217 (Faye S. Taxman ed., 2017).

[7] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 15 (2013).

[8] Jay P. Singh et al., *Measurement of Predictive Validity in Violence Risk Assessment Studies*, 31 BEHAV. SCI. & L. 55, 64 (2013).

recidivism is even minimally higher, it will count positively toward the AUC.[9] The PATTERN AUC shows it performs better than chance and thus the DOJ Report declares it is thereby validated.

However, the AUC has serious limitations and thus cannot present a holistic portrait of a tool's abilities.[10] Validity has two main features: discrimination and calibration. *Discrimination* indicates the tool's ability to distinguish recidivists from non-recidivists. Discrimination, thus, represents the tool's relative accuracy (in terms of the ability to differentiate recidivisms from non-recidivists). Discrimination is retrospective in nature as it is calculated after the recidivists and non-recidivists are identified. In other words, discrimination determines how well the tool would have classified the recidivists versus the non-recidivists.

In contrast, *calibration* concerns how accurate the tool statistically estimates recidivism, and it measures the tool's absolute predictive accuracy. Calibration is prospective (i.e., forward looking) by indicating how well a tool predicts future recidivism. Hence, discrimination and calibration offer distinct contributions to judging a tool's validity. As a result, a tool may vary in how well it meets either of these metrics.

> A scale that ranks well, but systematically overestimates or underestimates risk might have good discriminative properties but be poorly calibrated to the population under examination; in contrast, a very simple scale (e.g., one that merely divided offenders into ever violent/never violent, or male/female groups) might be very well-calibrated but have only modest discriminative validity.[11]

An analogy may be useful here. Take a bathroom scale. Suggest you have two people, one of whom is clearly heavier than the other. Each gets on the scale and the weight given the heavier person is, indeed, higher than that of the lighter person. But the scale mistakenly begins at 50 pounds rather than 0 pounds. Thus the scale gives the heavier person a weight of 250 pounds when in reality the person weights 200 pounds. The scale discriminates between the two but is not calibrated well. It overestimates actual weight.

The AUC is a simple metric for discrimination. It provides little information on calibration. The AUC has further limits. The AUC cannot calculate how well an instrument selects those at medium or high risk.[12] The AUC can indicate performance better than chance even if no recidivists were ranked as medium and/or high risk. The AUC regrettable fails to distinguish between types of errors. Whether the errors are predominantly false positives or false negatives is simply not picked up in this single statistic. False positives are wrongful predictions of recidivism while false negatives are wrongful predictions of non-recidivism. But these differences likely matter to officials who generally have an interest in whether they prefer a higher rate of false positives

[9] Philip D. Howard, *The Effect of Sample Heterogeneity and Risk Categorization on Area Under the Curve Predictive Validity Metrics*, 44 CRIM. JUST. & BEHAV. 103, 107-08 (2017).

[10] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 16-18 (2013).

[11] Philip D. Howard, *The Effect of Sample Heterogeneity and Risk Categorization on Area Under the Curve Predictive Validity Metrics*, 44 CRIM. JUST. & BEHAV. 103, 105 (2017).

[12] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 17 (2013).

versus false negatives.[13] Another flaw is that AUC accuracy rates between groups may be comparable, but the type of error may differ between groups. As an example, equivalent AUCs will mask whether one group has a higher rate of false positives, yet another a higher rate of false negatives.[14]

The disadvantages of the AUC in judging a tool's validity is well known. Here, we can draw on prior reports by the two external consultants hired to create PATTERN's: Dr. Zachary Hamilton and Dr. Grant Duwe. Each has previously written about such limitations. Hamilton has previously warned about "the mark of 'validated' [by an AUC as] a criterion that is rather easily achieved" as being slightly better than chance even though this may be a "*trivial* and *unexceptional accomplishment*."[15] Duwe agrees, writing that the AUC mark of validation may be "not practically meaningful" and that reliance upon it is "simple-minded."[16] Further, Duwe notes that the AUC is "casually misinterpreted as the probability that the [tool] accurately predicts who will reoffend and who will not. In other words, it is a measure of how well a risk assessment instrument can rank order recidivists and non-recidivists regardless of the predicted (absolute) risk."[17]

> [T]he AUC has the disadvantage of not being able to tell how likely individuals are to reoffend, which is a feature of well-calibrated risk assessment instruments. Calibration refers to the extent to which the predicted probabilities of risk agree with the occurrence of an actual outcome.[18]

Hence, Duwe argues that researchers "recognize the importance of not relying solely on the summary AUC" to evaluate a tool's performance.[19] Indeed, in the same article and others, Duwe actually computes preferred validity measures. In agreement, as a result of these limitations Hamilton has encouraged tool "creators to look to additional metrics of discrimination, calibration and accuracy to assure users of their instrument's strengths and weaknesses."[20] In sum, a claim about achieving a certain AUC level is far from a conclusive or holistic endorsement to support a claim that a tool is well-validated. For some unknown reason, the DOJ Report fails to follow through on providing evidence of these better visions of how well PATTERN performs from discrimination and calibration purposes.

---

[13] Jorge M. Lobo et al., *AUC: A Misleading Measure of the Performance of Predictive Distribution Models*, 17 GLOBAL ECOLOGY & BIOGEOGRAPHY 145, 146 (2008).

[14] Solon Barocas et al., *Big Data, Data Science, and Civil Rights* (2017), https://arxiv.org/pdf/1706.03102.

[15] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 569 (Faye S. Taxman ed., 2017) (emphasis in original).

[16] KiDeuk Kim & Grant Duwe, *Improving the Performance of Risk Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 189, 216 (Faye S. Taxman ed., 2017).

[17] KiDeuk Kim & Grant Duwe, *Improving the Performance of Risk Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 189, 206 (Faye S. Taxman ed., 2017).

[18] KiDeuk Kim & Grant Duwe, *Improving the Performance of Risk Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 189, 206-07 (Faye S. Taxman ed., 2017).

[19] KiDeuk Kim & Grant Duwe, *Improving the Performance of Risk Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 189, 217 (Faye S. Taxman ed., 2017).

[20] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 596 (Faye S. Taxman ed., 2017).

*Error Rates*

The DOJ Report contains some numerical evidence that permits a third party to calculate relevant performance metrics beyond the AUC. Several measures of discrimination and calibration require the use of what is referred to in the risk assessment literature as a contingency table. It is also known as a 2 × 2 table because it has two rows and two columns. Table 1 provides the elements to a contingency table.

Table 1

A Contingency Table to Compute Error Rates

| | | **Outcome** | | |
|---|---|---|---|---|
| | | Recidivist | Non-Recidivist | |
| **Assessment** | High/Medium Risk | True Positives (TP) | False Positives (FP) | *False Discovery Rate (FDR)* |
| | Minimum/Low Risk | False Negatives (FN) | True Negatives (TN) | *False Omission Rate (FOR)* |
| | | *False Negative Rate (FNR)* | *False Positive Rate (FPR)* | |

The internal boxes in the contingency table require four statistics:

- TP are true positives, those correctly predicted to recidivate.
- FP are false positives, those wrongly predicted to recidivate.
- TN are true negatives, those correctly predicted not to recidivate.
- FN are false negatives, those wrongly predicted not to recidivate.

These four measures (TP, FP, TN, FN) require that one create two dichotomous (dividing a whole into two parts) factors. One is whether one is a recidivist or a non-recidivist, as in recidivist = yes or no. The other is the risk level. PATTERN has four risk levels (minimum, low, medium, high). The most appropriate way to divide risk levels is to combine medium and high into a group as the predicted recidivists and then combine the minimum with low into another grouping as the predicted non-recidivists. These recidivist and non-recidivists groupings are justified as the First Step Act provides substantive consequences to those whose risk scores land them in minimum and low, on the one hand, versus in medium and high categories, on the other.

The external figures in the contingency table are error rates which are computed using the numbers from the internal boxes. Table 2 conceptualizes them.

Table 2

| | | |
|---|---|---|
| False Positive Rate (FPR) | Retrospective; a measure of discrimination | Of those known to be non-recidivists, what percentage was classified as higher risk? |
| False Negative Rate (FNR) | Retrospective; a measure of discrimination | Of those known to be recidivists, what percentage was classified as lower risk? |
| False Discovery Rate (FDR) | Prospective; a measure of calibration | Of those classified as higher risk, what percentage did not recidivate? |
| False Omission Rate (FOR) | Prospective; a measure of calibration | Of those classified as lower risk, what percentage become recidivists? |

Notice that the first two of these measures (FPR and FNR) are retrospective in nature in that they are computed after knowing which offenders actually did (or did not) recidivate. The FPR and FNR are calculated in the columns in the contingency table. The second two (FDR and FOR) are prospective in that they look at the groups as predicted to reoffend (or not) and whether they actually did go onto recidivate. The FDR and FOR are calculated in the rows of the contingency tables. The more exact computations of these error rates are compiled in Table 3.

Table 3

| *Definition(s)* | *Measure* | *Calculation* |
|---|---|---|
| False negative error rate balance | False Negative Rate (FNR) | $\frac{FN}{FN+TP}$ |
| False positive error rate balance | False Positive Rate (FPR) | $\frac{FP}{FP+TN}$ |
| Type I Error | False Discovery Rate (FDR) | $\frac{FP}{FP+TP}$ |
| Type II Error | False Omission Rate (FOR) | $\frac{FN}{FN+TN}$ |

Before revealing the various error rates in PATTERN, a bit more context is required. The DOJ Report outlines four risk scales: (1) general recidivism for males, (2) violent recidivism for males, (3) general recidivism for females, (4) violent recidivism for females. These scales have much overlap in risk factors, but not entirely, and their numeric scoring systems (i.e., points assigned) vary across the four.[21] However, these four scales do not determine a person's final risk category that drives whether they are eligible for significant rewards under the First Step Act. Instead, final risk categories were created by combining the risk outcomes from the general and violent

[21] DOJ Report Chapter 3, Table 2.

recidivism scales (by gender). According to the DOJ Report, it appears that the final risk categories are assigned as follows:

Final (Combined) Risk Outcome

- If minimum risk on both scales, then minimum risk outcome overall
- If less than medium risk on both scales, then low risk outcome overall
- If high risk on either scale, then high risk outcome overall
- All other cases, medium risk outcome overall

The error rates provided below, therefore, use the final risk categories as the relevant measure. In sort of reverse engineering the data provided in the DOJ Report[22], these error rates are shown here in Tables 4(a) and 4(b) for general and violent recidivism, respectively. (For these tables, the genders are combined as they are in relevant part of the DOJ Report.)

Table 4(a)

Error Rates in PATTERN

General Recidivism

| | | **Outcome** | | |
|---|---|---|---|---|
| | | Recidivist | Non-Recidivist | |
| **Assessment** | High/Medium Risk | 25179 (TP) | 13083 (FP) | 34% *(FDR)* |
| | Minimum/Low Risk | 7711 (FN) | 27607 (TN) | 22% *(FOR)* |
| | | 23% *(FNR)* | 32% *(FPR)* | |

Notice the FPR and FDR, both of which are about erroneously classifying individuals as higher risk (combining medium and high risk categories), are at about an error rate of one-third (i.e., 32% and 34%). The FNR and FOR, which are about erroneously classifying individuals as lower risk (combining minimum and low risk) occurs over one-fifth of the time (i.e., 23% and 22%). One can observe that for general recidivism, false positives are more "acceptable" than false negatives.

Two other relevant metrics can supplement our understanding. One is the cost ratio, which shows whether a tool produces a greater number of false positives over false negatives. Indeed, the cost ratio of false positives over false negatives for PATTERN general recidivism is 1.7, indicating almost twice as many false positives as false negatives. Another way to judge calibration is to compare a tool's predicted recidivism rate to the actual observed recidivism rate.

[22] DOJ Report Chapter 3, Table 5.

We can observe a calibration issue as the tool predicts a 52% general recidivism rate while 45% did reoffend.[23] In other words, PATTERN over overpredicts general recidivism risk.

Next, Table 4(b) provides error rates for PATTERN with violent recidivism.

Table 4(b)

Violent Recidivism

| Assessment | Outcome: Recidivist | Outcome: Non-Recidivist | |
|---|---|---|---|
| High/Medium Risk | 9352 (TP) | 28910 (FP) | 76% *(FDR)* |
| Minimum/Low Risk | 1383 (FN) | 33935 (TN) | 4% *(FOR)* |
| | 13% *(FNR)* | 46% *(FPR)* | |

PATTERN has significantly large error rates for its higher risk attributions for violent recidivism. The retrospective False Positive Rate is 46% while the prospective False Discovery Rate is 76%. The error rates for lower risk attributions are small. Clearly, the final risk categories are not reasonably predicting violent reoffending.

The cost ratio of false positives over false negatives is 20.9, indicating a significant preference for false positives. Then, PATTERN predicts 52% will violently recidivate, while only 15% did. This is a significant overprediction of violent reoffending overall.

It would likely be informative to stakeholders to understand what such error rates may be across racial groups. In other words, is the false positive rate higher or lower for Whites as compared to African-Americans or Hispanics? This type of analysis would help confirm the DOJ Report's claim that PATTERN is racially fair. Unfortunately, the DOJ Report does not provide the numbers to be able to make those calculations. This is a significant example of a gap in transparency.

*Dimensions of Risk*

Importantly, one should be aware of the significantly limited nature of what PATTERN predicts. Even advocates of evidence-based practices acknowledge that a key question is—measuring the risk of what?[24] PATTERN predicts the probability of a single event: any recidivism, which it defines as any arrest (serious or nonserious) or return to prison (e.g., technical violation).

[23] The 45% general recidivism figure here is slightly off of the 47% recidivism statistic reported in the DOJ Report because the exact numbers in this document were based on percentages given in the DOJ Report that did not include decimals to allow more precise figures here.

[24] Jordan M. Hyatt et al., *Reform in Motion: The Promise and Perils of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing*, 49 DUQ. L. REV. 707, 743 (2011).

Presumably, though, the concept of risk that is the foundation of the First Step Act is not some singular feature focused solely on an abstract likelihood of being charged with some infraction at some point in the future. Instead, at least six different dimensions of risk are conceivably pertinent. Probability is only one of them. The other important dimensions include offense type (e.g., terrorism, violent, property, white collar, drugs), severity of harm, imminence, frequency, and duration of offending.[25] Unfortunately, PATTERN ignores most of those additional, yet important, dimensions. While PATTERN has a violence scale, it counts any serious or nonserious arrest for a violence-related offense, which presumably then includes minor threats or assaults.

*Reliability*

Another critical component of accuracy is entirely missing from the DOJ Report. Reliability refers here to consistency in scoring a tool across evaluators. A desirable trait exists whereby the same individual will receive the same score by different evaluators. A PATTERN developer has confirmed the salience of this aspect, writing in another article: "One of the important first steps in implementing a risk assessment instrument is to ensure that the instrument is administered consistently by those who collect and score risk factors."[26] Indeed, the developer (correctly) concedes that reliability and validity are the two principal properties to evaluate the ability of the instrument.[27] An unreliable tool will undermine its validity. The most common metric is the inter-rater reliability score to check for the degree of consistency in scoring between raters.[28] As the DOJ Report makes no mention of inter-rater reliability, the (limited) validity measures it does provide remain suspect.

## *Fairness*

Algorithmic fairness is of growing interest as algorithms pervade society and public institutions. The literature now offers many different definitions about how to conceptualize and test for algorithmic fairness. The usual interest here is fairness across sociodemographic groups.

*Statistical/Demographic Parity*

One of the most popular group fairness definitions is statistical parity. Statistical parity exists when the percentages of offenders predicted to recidivate and those predicted not to recidivate are the same across groups.[29] Hypothetically, if 40% of those assessed in one group are predicted

[25] Michael H. Fogel, *Violence Risk Assessment Evaluation: Practices and Procedures, in* HANDBOOK OF VIOLENCE RISK ASSESSMENT AND TREATMENT: NEW APPROACHES FOR FORENSIC MENTAL HEALTH PROFESSIONALS 41, 43 (Joel T. Andrade ed., 2009).

[26] Grant Duwe, *Why Inter-Rater Reliability Matters for Recidivism Risk Assessment* 2 (2017), https://psrac.bja.ojp.gov/ojpasset/Documents/PB-Interrater-Reliability.pdf.

[27] *Id*.

[28] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 568 (Faye S. Taxman ed., 2017).

[29] Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments*, 16 J. EMPIRICAL LEG. STUD. 175, 184 (2019).

to recidivate, statistical parity would require that the tool predict 40% of the other group to recidivate. The literature also refers to this measure of equity as demographic parity if the groups at issue are distinguished by some demographic characteristic (e.g., race, class, gender).[30] A lack of demographic parity suggests disparate treatment.

How does PATTERN fare on demographic parity? The same dichotomous groupings mentioned earlier are used here. PATTERN's combined medium and high risk categories comprise the predicted recidivists, while the minimum and low categories combined represent the predicted non-recidivists. Table 5(a)[31] shows the rates of predictions across race/ethnicity for males.

Table 5(a)

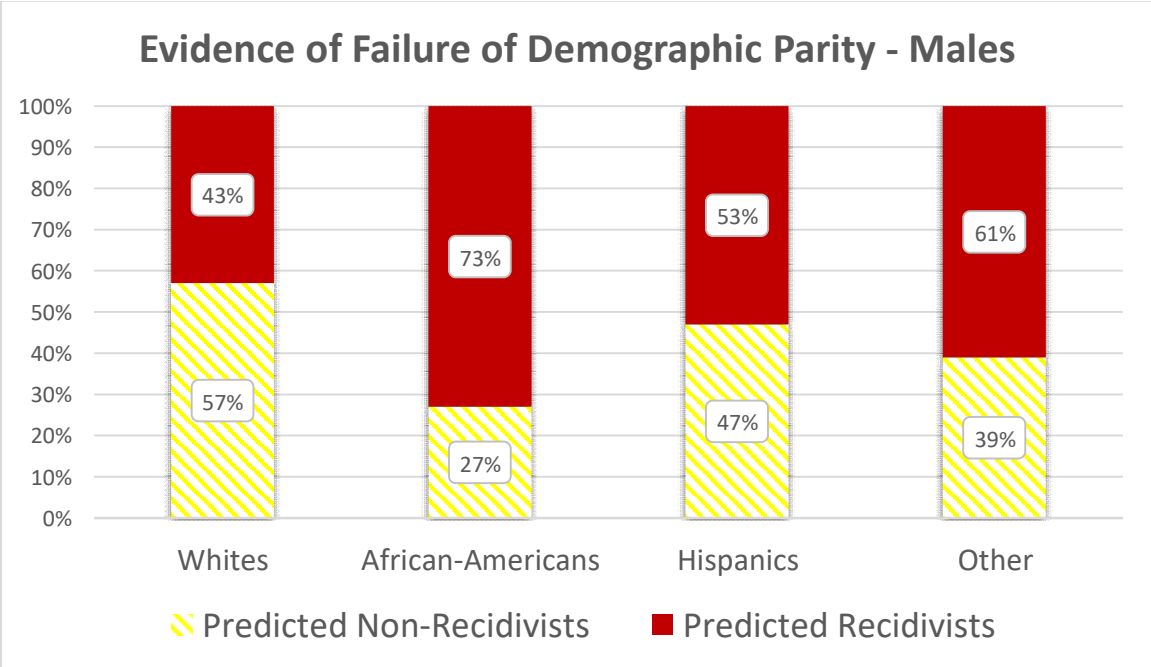



Table 5(a) evidences the lack of statistical/demographic parity of PATTERN for males. The tool predicts unequal proportions of recidivists across racial/ethnic groups, from 43% for White males, 53% for Hispanic males, and a significantly larger 73% for African-American males. Table 5(a) reflects another inequality. The First Step Act provides the greatest benefits to those scoring in the minimum and low risk categories. Notice that only 27% of African-Americans are able to gain the greatest benefits from the First Step Act, while more than twice that percentage, 57%, of Whites are eligible for those benefits. Indeed, PATTERN is significantly less likely (not shown in Table 5(a)) to classify White males into the extreme "High" risk category at 29%, compared to

[30] *See, e.g.*, James E. Johndrow & Kristian Lum, *An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction*, 13 ANNALS APPLIED STAT. 189 (2019), https://www.e-publications.org/ims/submission/AOAS/user/submissionFile/30728?confirm=1d6331c2.

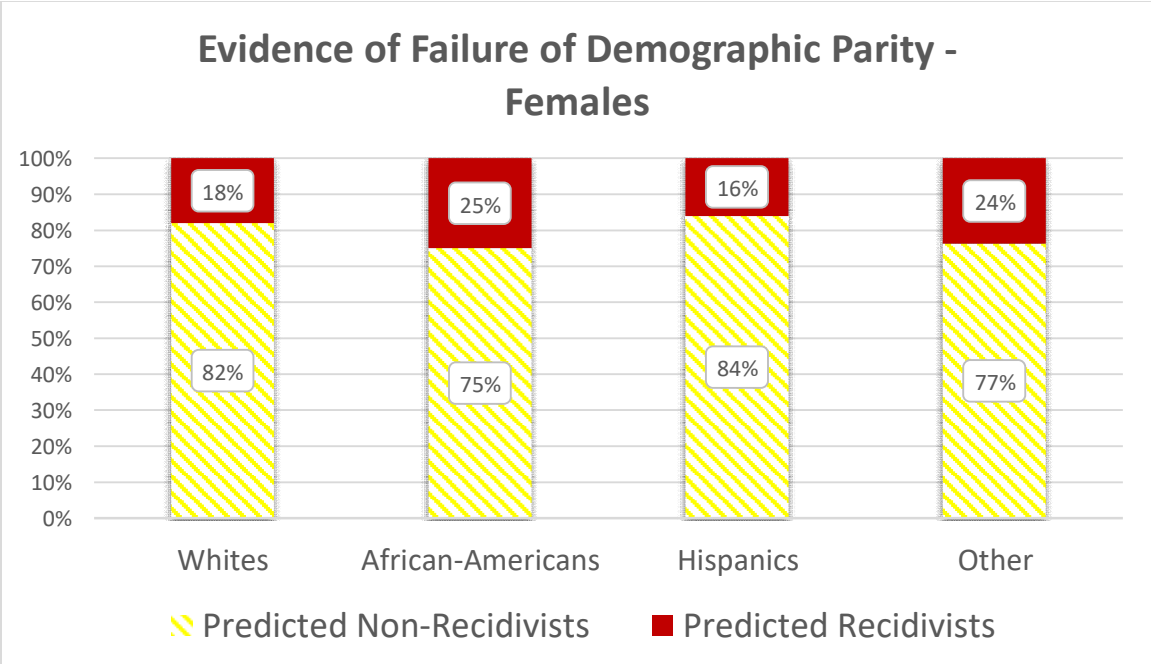
[31] This table is calculated from the DOJ Report Table 8. The percentages do not exactly match as the DOJ Report table does not add to 100% likely because of rounding.

53% of African-American males. This means that a African-American males are far less likely to ever earn the best incentives and rewards from the First Step Act.

The DOJ Report acknowledges a racial/ethnic disparity for males, though implicitly. The document uses the Relative Rate Index but does not discuss the adverse result in the text. Here, instead of again using the four racial/ethnic categories, the DOJ Report creates two groups: Whites and non-Whites. An RRI over 1.0 indicates disparity between groups.[32] The DOJ Report indicates an RRI of 1.54 for males,[33] which reflects racial disparity. Non-Whites are one-and-a-half times more likely to be assessed as medium/high risk than Whites. Thus, non-Whites face substantially reduced opportunities to gain early release credits through the First Step Act. It is odd that RRIs were computed in this way of Whites versus non-Whites. A DOJ technical manual indicates that best practices dictate that the RRI be calculated separately for each minority group that comprises at least one percent of the total population scored.[34] Thus, the document would more appropriately have calculated RRI's to compare Whites with African-Americans and Hispanics separately. The "Other" category used is also of concern. Native-Americans and Asian-Americans are often ignored racial groupings, yet each comprises greater than one percent of federal defendants and thus their numbers should separately have been included.

A vision of demographic parity for females is provided in Table 5(b).

Table 5(b)




[32] Zachary Hamilton et al., *Recrafting Youth Risk Assessment: Developing the Modified Positive Achievement Change Tool for Iowa*, DEVIANT BEHAV. (forthcoming 2019).

[33] DOJ Report Chapter 3, Table 8.

[34] William Feyerherm et al., *Identification and Monitoring, in* Dept. of Just. Office of Juvenile Justice and Delinquency Prevention, DISPROPORTIONATE MINORITY CONTACT TECHNICAL ASSISTANCE MANUAL 1-1, 1-2, 3 (4th ed. 2009).

Demographic parity remains a problem for females, though to a lesser degree than males. PATTERN predicted that 16% of Hispanic females would be recidivists, compared to 18% for White females and 25% for African-American females. Then, considering that predicted non-recidivists combine the minimum and low risk categories, African-American females benefit less often from the First Step Act's incentives and rewards. The reported Relative Rate Index with Whites versus non-Whites does not indicate a significant disparity for females. But with Hispanic females having a higher rate than White females of being eligible for early release credit, this washes out the potential disparity that an RRI computation separately for Whites versus African-Americans might show. It would also be useful to understand how Native-American and Asian-American females fare.

Overall, these sorts of racial/ethnic disparities are concerning. The risk assessment literature has been progressive as of late in crafting and testing various ways to ameliorate disparities. But to do so one would need to employ more data points than are provided in the DOJ Report. An interested party with access to the underlying datasets could look for which risk factors in PATTERN might be driving the disparities and then work on corrections that try to save predictive ability while reducing group inequities. For example, if criminal history measures account for a substantial portion of the discrepancies for African-Americans, then modifying criminal history in risk-sensitive ways may improve the tool and its equitable outcomes. Likely options could be to discount prior history with crack cocaine or marijuana arrests, both of which tend to be associated with differential policing practices in poorer areas. Excising misdemeanors from criminal history and recidivism definitions may also result in more equitably fair scores. As well, if the educational score suggests that it is a proxy for minority neighborhood disadvantage, then a fix to it might drive down disparate impact.

Notably, there are other ways to assess for test bias. The DOJ Report suggests one of them when it notes that "[t]o be racially unbiased or neutral, the tool should ensure race and ethnicity have no effect on the tool's outcomes, specifically the prediction of whether an individual will recidivate, once the tool is controlled."[35] The description and its accompanying footnote is referring to what has been nicknamed the Cleary method involving hierarchical regression models.[36] This type of test is respected, yet for some reason the DOJ Report does not actually use it. One with access to the underlying datasets could usefully employ this method to assess racial/ethnic bias further.

[35] DOJ Report, at 28.

[36] Melissa Hamilton, *The Biased Algorithm*, 56 AM. CRIM. L. REV. 1553 (2019), http://epubs.surrey.ac.uk/id/eprint/852008.

*Differential Calibration*

The algorithmic fairness definition of equal calibration requires that tool outcomes mean the same thing across groups. Table 6(a)[37] shows general recidivism rates for males by PATTERN risk category.

Table 6(a)

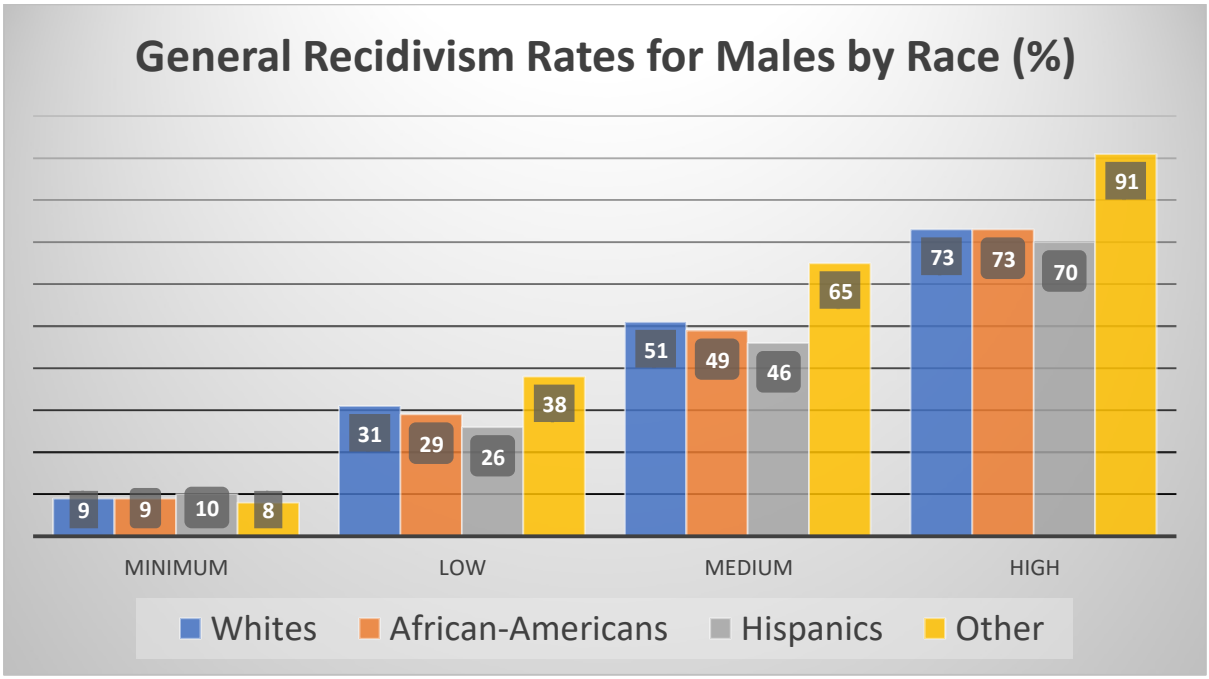
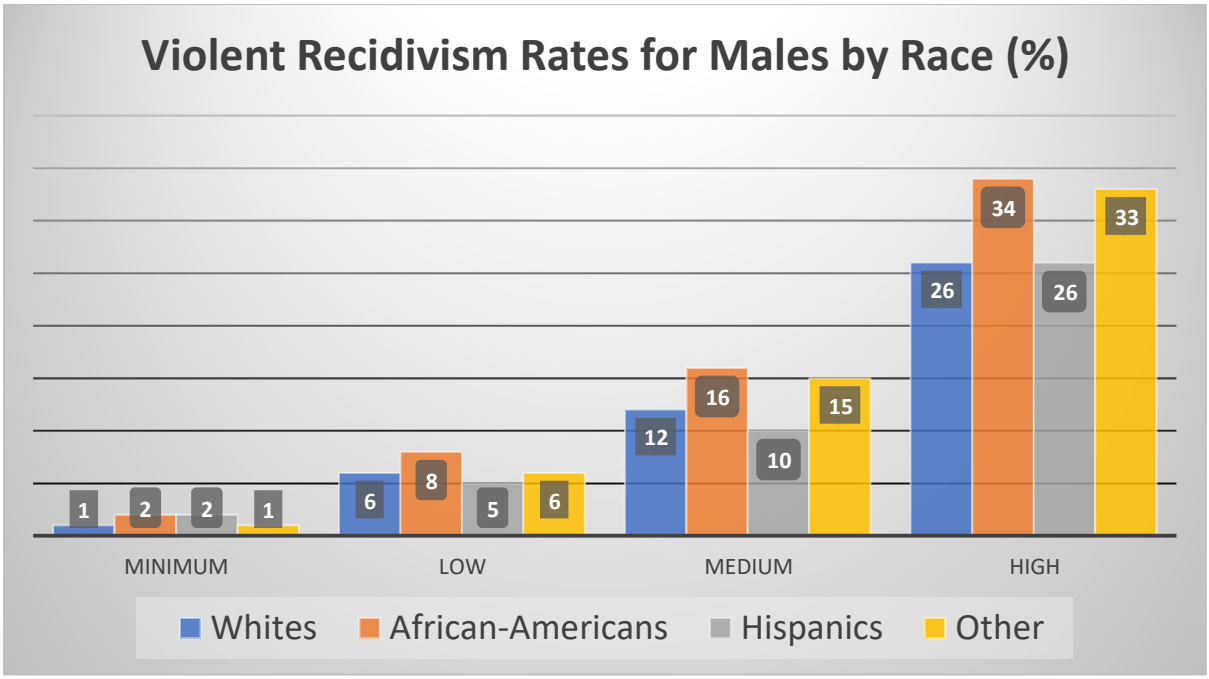



Table 6(a) indicates that risk categories do not mean the same thing across racial/ethnic lines for males. For example, a low risk outcome sees a 26% general recidivism rate for Hispanics but a 38% for Other. A medium risk outcome means a 46% chance of reoffending for Hispanics but a 65% general recidivism rate for Other. Table 6(a) also indicates that PATTERN tends to overpredict risk for African-American and Hispanic males while underpredicting for Other.

The differential calibration is worse for violent recidivism in males, as seen in Table 6(b).[38]

[37] DOJ Report Chapter 3, Table 9.
[38] DOJ Report Chapter 3, Table 10.

Table 6(b)




Again, the PATTERN risk categories do not mean the same thing across racial/ethnic groups for males for violent recidivism. The most significant difference is at the high risk category whereby 26% of Whites and Hispanics violently reoffended, compared to 34% for African-American males. Table 6(b) indicates that PATTERN underpredicts violent recidivism for African-American and Other males in the medium and high categories. We next can observe how well PATTERN is calibrated for females by race/ethnicity for general recidivism in Table 6(c).

Table 6(c)

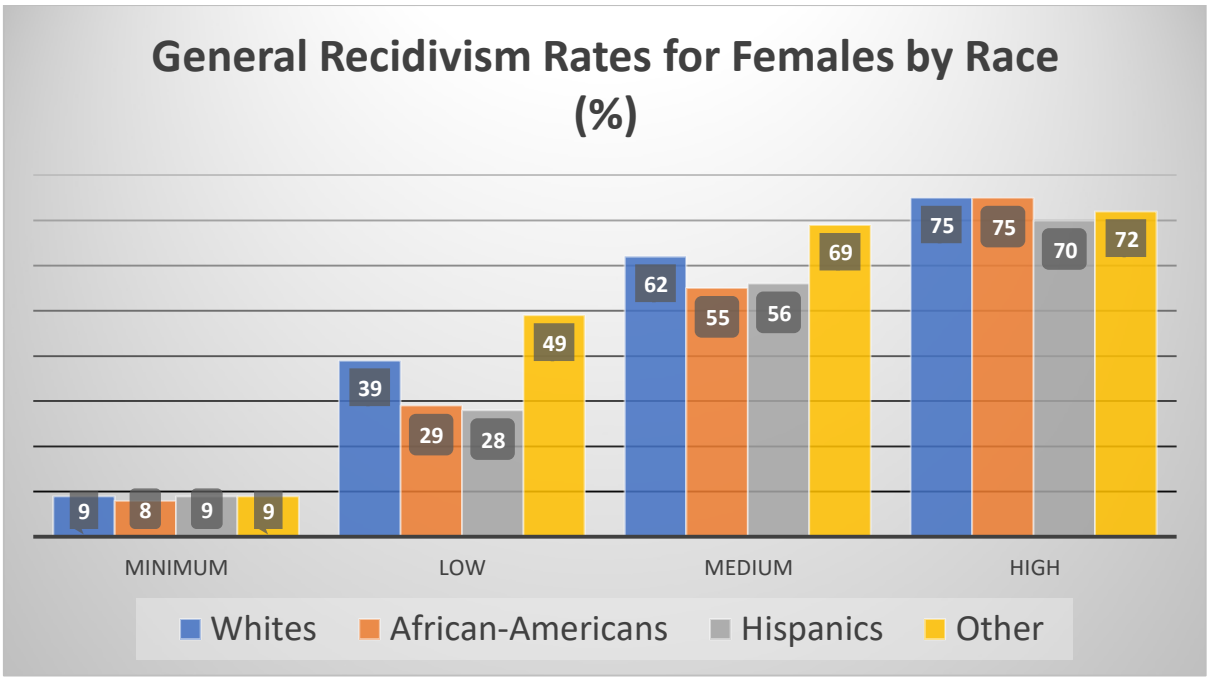
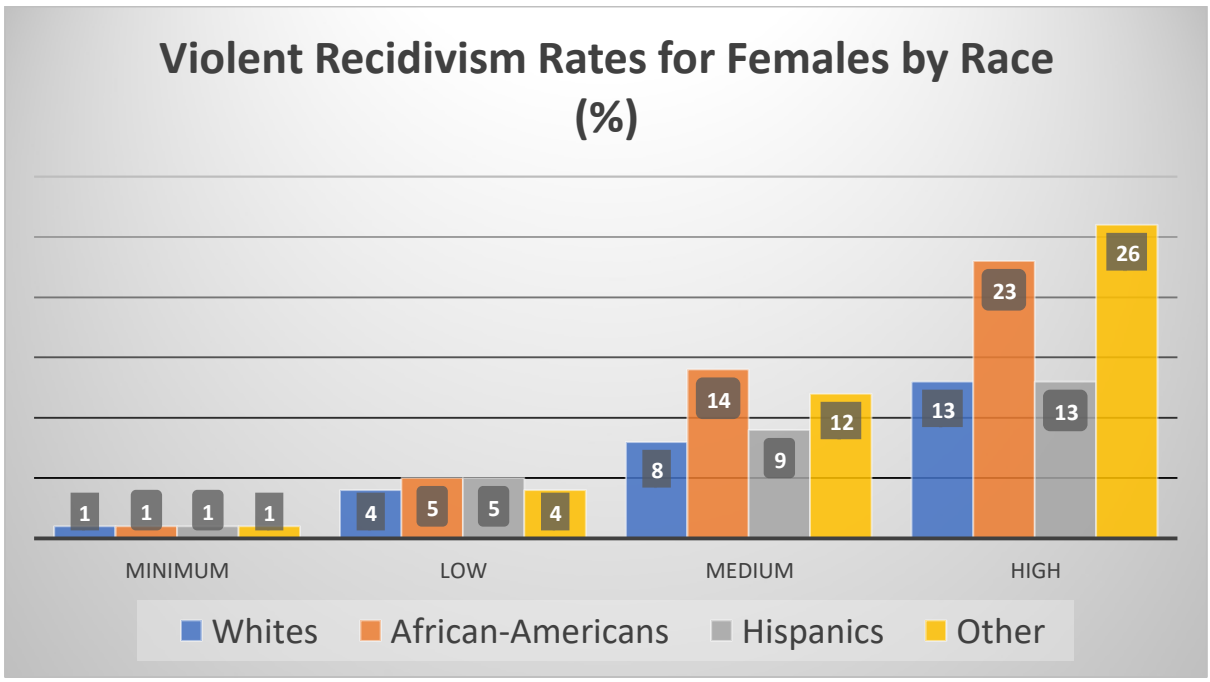



Table 6(c) indicates the most significant problem is in the low risk category, with significant variations across racial/ethnic groups for women. The situation, though, is more dire for violent recidivism in females, as indicated in Table 6(d).

Table 6(d)



For violent recidivism, PATTERN significantly varies in performance in the medium and high risk groups for women. This means that PATTERN risk categories just do not perform similarly across racial groups for women.

## *Improvements*

The PATTERN risk tool can better achieve the goals of the First Step Act through a variety of steps.

1. It is of questionable value and equity to include all arrests. One of the developers has written that risk assessment used by a department of corrections for sentenced prisoners ought rightly to focus on reincarcerations and reconvictions on violent crimes and felonies, rather than on less serious offenses or arrests.[39] Faye Taxman, one of the Independent Review Committee members, has similarly critiqued tools when "measures of criminal history tend to treat all crimes the same without prioritizing more serious criminal behavior."[40] Misdemeanors, technical violations, and arrests are poor proxies to actual, serious offending. By definition, any proxy measures for crime will be fundamentally inaccurate.[41] Arrests, in particular, are troubling because of their relatively low evidentiary bar and the fact they may instead reflect differential and discriminatory policing practices that disproportionately target minorities. Thus, to echo suggestions by other contributors, the tool might better focus on serious and violent convictions.
2. The significant error rates in overpredicting recidivism undermines the goal of the First Step Act to incentivize all offenders to undertake rehabilitative programming and otherwise reduce their risk levels. There is an easy fix. Changing the cut-points higher between low and medium would be the single most important resolution. This would automatically increase the number of prisoners eligible to earn and apply early release credits. In reality, there is no single method for determining cut-points. One of PATTERN's developers has written about this. The method chosen for PATTERN of using fractions of the base rates is, as he has admitted, "somewhat arbitrary."[42] Using another option or simply changing these fractions may yield fewer false positives and permit more prisoners to gain the advantages of engaging with needs-based programming.

[39] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 562-64 (Faye S. Taxman ed., 2017).

[40] Faye S. Taxman & Amy Dezember, *The Value and Importance of Risk and Need Assessment (RNA) in Corrections & Sentencing: An Overview of the Handbook, in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 22, 32 (Faye S. Taxman ed., 2017).

[41] MICHAEL VEALE, THE LAW SOCIETY, ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM 18 (2019).

[42] Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, *in* HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 561 (Faye S. Taxman ed., 2017).

3. Cut-points should be increased for another reason. Increasing cut-points can reduce the significant false positive rates and false discovery rates. The size of these error rates conflict with the goals of the First Step Act.
4. PATTERN could better represent the evidence-informed literature by incorporating additional protective factors that moderate (lessen) the salience of a risk factor or promotive factors that predict desistence (the flip side of a risk factor).[43]
5. Empirical research confirms the age-crime curve. As a result, the factor of age at assessment should have gradated negative points as age increases. Currently, the age at assessment only have positive points, all indicating increased risk with the sole exception of those over 60 years, which have zero points. Other tools more appropriate deduct points in significant increments as age increases.
6. Research indicates that the salience of criminal history as a risk factor fades over time. As a result, some mechanism should be seriously considered to reduce the points for criminal history as such events become stale over time.
7. Weights for dynamic factors should be substantially increased to more realistically allow prisoners to lower their risk scores with risk-reducing activities. The DOJ Report asserts that 99% of prisoners can reduce their risk category to low. As the current point totals significantly weight static factors, this statistic is implausible.
8. The final Risk Level Categories could be modified to allow more offenders to gain greater advantages of the First Step Act by assuming a lower risk outcome. For example, a person who scored low risk on the violent recidivism tool and medium risk on the general recidivism tool should be assigned a low risk final category rather than medium as currently applied.
9. It appears that many of the factors overlap without justification provided. This raised some questions about double and triple counting of the same events.
    a. Can the same violent offense be counted multiple times, such as in the Criminal History Score, Infraction Convictions, Instant Offense Violent, History of Violence, and/or Sex Offender?
    b. Can the same sex offense be counted multiple times, such as in the Criminal History, Infraction Convictions, Instant Offense Violent, History of Violence, and/or Sex Offender?
    c. Can a single infraction count as both any infraction and then again as a violent/serious infraction?
    d. Can a single infraction also count in the BRAVO criminal history factor and/or the sex offense (such as on a reassessment)?
    e. Young age can be counted twice for young prisoners whose index arrest/conviction (the DOJ Report is inconsistent on whether it is age at first

[43] John Monahan & Jennifer Skeem, *Risk Assessment in Criminal Sentencing*, 12 ANN. REV. CRIM. PSYCHOL. 489 (2016).

arrest or age at first conviction) was their first. Should this double-counting be ameliorated?

f. Can a person with multiple violent priors receive multiple point scores in this single variable? For example, in the male general recidivism tool, if a defendant had a minor violent offense <5 years plus a serious violent offense >15 years, would the defendant be scored as 5 or 7?

10. PATTERN increases points for a history of a sex offense (this appears not to require a formal charge or conviction). But separate DOJ reports on national samples have found that sex offenders are at lower risk of general recidivism.[44] Thus, this factor may not be supported by empirical research as a valid risk factor and should be excised as a result.
11. What is the empirical basis for including the violent offense factor? This appears to be contrary to DOJ studies of national samples of offenders released whereby violent offenders at a lower risk of general recidivism compared to other types of offenders? Is this factor really operating as a policy override for other purposes?[45]
12. A stand-alone mechanism for disputing risk scores must be established. The current plan appears to be to simply apply the current prisoner grievance system. This is insufficient and inapplicable. Algorithmic risk assessment practices require their own processes to challenge.
13. Information on overrides of PATTERN outcomes is required.
14. The datasets should be public released. Statisticians could effectively mine them to determine how the tool fares across algorithmic fairness definitions. Some of those definitions were outlined earlier. Additional ones are available in the literature that can be informative. These include the ability to test group fairness (e.g., race, ethnicity, gender, age) on such measures as balance for the positive class, balance for the negative class, diagnostic odds ratios, and correlations. Further, understanding how missing data was treated would be helpful.
15. Independent reviews of the datasets could allow data scientists to suggest ways to ameliorate the racial biases that are evident. For example, an audit could reveal which risk factors correlate with race/ethnicity and thus should be removed and/or modified to improve algorithmic fairness across effected groups.

Thank-you for the opportunity to submit this written testimony to the Subcommittee.

[44] Matthew R. Durose et al., Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010 (2014) (Special Report, U.S. Dept. of Just.).

[45] Mariel Alper and Matthew R. Durose, 2018 Update on Prisoner Recidivism: A 9-Year Follow-up Period (2006-2014) (2019) (Special Report, U.S. Dept. of Just.).