Beyond Validity: Current Auditing Methods for Criminal Risk Assessments Do Not Consider Sequential Feedback Effects

Benjamin Laufer
ben.laufer@li.me
Lime
San Francisco, California

ABSTRACT

In the criminal legal context, risk assessment algorithms are touted as data-driven, well-tested tools. Studies known as validation tests are typically cited by practitioners to show that a particular risk assessment algorithm has predictive accuracy, establishes legitimate differences between risk groups, and maintains some measure of group fairness in treatment. To establish these important goals, most tests use a one-shot, single-point measurement. Using sentencing data from Philadelphia, we show empirically that decisions in the criminal legal domain are highly correlated with past and future decisions. Then, using a Polya Urn model, we explore the implication of feedback effects in sequential scoring-decision processes. We show through simulation that risk can propagate over sequential decisions in ways that are not captured by one-shot tests. For example, even a very small or undetectable level of bias in risk allocation can amplify over sequential risk-based decisions, leading to observable group differences after a number of decision iterations. Risk assessment tools operate in a highly complex and path-dependent process, fraught with historical inequity. We conclude from this study that these tools are not as data-driven as they seem, and call for improvements in auditing before these tools can be widely adopted.

CCS CONCEPTS

• **Computing Methodologies** → **Simulation Environments**; • **Computer systems organization** → **Embedded systems**.

KEYWORDS

Computational Social Science, Risk Assessment, Validation, Public Safety Assessment, Group Fairness, Predictive Accuracy, Sequential Decision, Scoring Decision System, Sentencing, High-Impact Algorithm

ACM Reference Format:
Benjamin Laufer. 2021. Beyond Validity: Current Auditing Methods for Criminal Risk Assessments Do Not Consider Sequential Feedback Effects. In *Proceedings of Conference on Fairness, Accountability, and Transparency (FAccT '21).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

FAccT '21, March 2021, Toronto, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

As machine learning techniques have developed to replicate human decision-making, their use has forced a reconciliation with existing decision policies: can statistics do better? Are the statistics unfair, and are they more unfair than the people?

A number of influential papers in 2015 [19, 20] suggested that accuracy in statistical forecasting methods can and should be used in 'important' contexts, where people's freedom or health or finances are on the line, since these algorithms come with demonstrable accuracy levels. These contexts include sentencing and pre-trial decisions, credit scoring, medical testing and selective education access. Since then, the release of a ProPublica investigation of a common bail algorithm [1] and retorts from the Criminology field [9, 13] have forced a reckoning among theorists and practitioners about what fairness goals can and cannot be achieved.

Researchers have emphasized shifting focus from predictions to treatment effects, acknowledging that many of these high-impact decisions are, indeed, highly impactful on individual life-courses [3]. This revelation introduces the relatively new and under-analyzed topic of fairness in relation to repeated decision processes. Individual studies have demonstrated that 'predictive feedback loops' can lead to disproportionate over-policing in certain neighborhoods [24], and that these loops can be modeled and simulated to demonstrate sub-optimal allocation in policing and compliance contexts [7, 11].

The sequential-decision context is truly the norm, rather than the outlier. In virtually all high-impact scoring or testing systems, these processes occur (or may occur) numerous times throughout individual life-courses and each are both highly dependent on the past and highly impactful on individuals' futures. In light of sequential dependence in high-impact algorithms, this paper analyzes current methods for validating scoring systems as accurate and fair.

In the criminal legal context, new risk assessment algorithms are touted as data-driven, well-tested tools and often cite one or multiple validation studies that demonstrate a tool's predictive accuracy and predictive parity between defendants of differing protected classes. Virtually all use a single-point-in-time, batch setting to analyze fairness and accountability concerns, with the exception of a few studies about how change in scores over time can better predict future scores [21, 22, 31, 35]. We show that these tests are not catered to the criminal legal domains, where decisions often occur sequentially at multiple times through a defendant's life. We take a close look at the statistical methods used by these studies, and show using simulation experiments that risk assessment tests

can fail at meeting a number of fairness definitions even while passing instantial validity tests.

1.1 Validation and One-Shot Testing

Risk assessment algorithms are developed and then tested for 'validity'. These experiments, formerly only concerned with predictive validity, now test various potential biases that algorithms may exhibit in new populations. Validation experiments have therefore become an important aspect of the risk-assessment development process, and validity is seen as a necessary requisite for any risk assessment algorithm in use. What does validity mean?

While there has been some controversy over the way in which risk assessment tools get developed,[1] remarkably little analysis has been conducted of the best practices for validation in risk assessment. As a result, many validation experiments resemble one another. Typically, the studies measure a tool's predictive capacity by analyzing post-conviction arrest rates over a short time-frame. They take a group of defendants released from the same jurisdiction in a given time-frame, and determine the average re-arrest rate of defendants with different risk scores over a typical period of one or two years. For example, Lowenkamp et al. conducted a validation experiment in which they tested the LSI-R and the LSI-Screening Version, which screens defendants to decide whether to administer the more in-depth LSI-R assessment [23]. Using a look-ahead period of 1.5 years, the study measured re-arrest rate and re-conviction rate, and found that a higher LSI-R score is positively correlated with future incarceration.

Interestingly, algorithmic risk assessments tend to find disparate validity levels when the same algorithm is used on racially distinct populations. Fass et al. in 2008 published validation data on the Level of Service Inventory - Revised (LSI-R) algorithm, as well as COMPAS [12]. Using a dataset of 975 offenders released into the community between 1999-2002 from New Jersey, the measurement period was 12 months. The purpose of the study was to see whether these algorithms, trained on mostly white populations, are invalid for a population like New Jersey, which has has "substantial minority" representation in incarceration. The study finds "inconsistent validity when tested on ethnic/racial populations" [12, 1095], meaning the predictive validity may suffer as the result of differences between the training cohort used to develop the algorithm and the actual demographic breakdown of a jurisdiction. Demichele et al. in "The Public Safety Assessment: A Re-Validation" use data from Kentucky provided by the Laurence and John Arnold Foundation, which developed the PSA. The study measured actual failure-to-appear, new criminal activity, and new violent criminal activity before a trial. They found that the PSA exhibited broad validity, but found a discrepancy based on race [8].

Beyond recidivism, a few studies have focused on the relationship between risk assessment-driven decisions and other life outcomes, including earnings and family life. Bruce Western and Sara McLanahan in 2000 published a study entitled "Fathers Behind Bars" that finds alarming impacts of incarceration on family life. A sentence to incarceration was found to lower the odds of parents living together by 50-70% [36]. Dobbie et al. published a study that demonstrated that pre-trial detention in Philadelphia on increased conviction rates, decreased future income projects and decreased the probability that defendants would receive government welfare benefits later in life [10]. The Prison Policy Initiative reports an unemployment rate above 27% for formerly incarcerated people, and find a particularly pronounced effects of incarceration on employment prospects for women of color [6].

Given the deeply impactful nature of risk-based decisions, validation experiments are surprisingly limited in scope. The outcome variable - typically rearrests in a one or two-year window - fail to capture the many ways that a risk-assessment can impact an individual's family, employment, income, and attitudes - all of which may be relevant in considering recidivism. Perhaps more importantly, the various aspects of life impacted by detention are precisely the risk factors that may get picked up by a subsequent judicial decision.

By treating risk assessment as instantial and analyzing longitudinal effects of a single assignment of risk, validation experiments are only observing part of the picture. When we consider the tangible impacts of judicial decisions and relate these impacts to future decisions, we see that there are possible feedback effects in the criminal system. The dependence of subsequent judicial decisions on prior judicial decisions is rampant. Sentencing guidelines suggest (and often require) judges to give longer sentences to repeat offenders, for example. The very notion of responsivity in criminal treatment requires periodic assessments that determine the 'progress' or treatment effect over time for a given defender, and shape punishment accordingly. However, treatment of sequential risk-assessments and the possible harms of feedback is missing from a literature that has so exhaustively debated whether incarceration has a criminogenic effect.

This paper explores how compounding in criminal justice impacts defendants. The treatment of risk assessment as innocuous, objective, statistical prediction has clouded rigorous theoretical exploration of lifetime compounding in criminal punishment. Using data from Philadelphia, we find that higher confinement sentences significantly increase cumulative future incarceration sentences for defendants. Synthesizing data from Philadelphia with a theoretical understanding of feedback in algorithmic risk assessment, we will discuss implications for judges and defendants.

1.2 Contributions

This paper is meant to critically evaluate the current vetting and auditing process for high-stakes, repeated-use risk assessment algorithms that are deployed in the U.S. criminal legal system.

First, we demonstrate one method of testing for sequential feedback in repeat-decision processes. Using the predictive scoring model as a control, we are able to see whether sentencing decisions may causally relate to future criminal punishment.

Second, we develop a generalized sequential scoring-decision model, which can be used in simulation experiments to test for possible compounding effects in group fairness, uncertainty, and punishment.

[1] In Philadelphia, for example, recidivism was being measured as re-arrest rate, and because of public opposition the sentencing commission began measuring it as subsequent conviction rate.

Finally, using simulation experiments, we demonstrate that a risk assessment can pass validity tests and still exhibit problems with predictive accuracy, group-fairness, and risk-group-difference.

The broader argument put forward by this paper is that current validation tests do not consider sequential feedback, and are therefore insufficient to approve criminal risk assessments for use. Algorithms used in the criminal legal system, credit system, and in other high-impact domains should test for unintended impacts when used repeatedly.

2 LITERATURE REVIEW

Significant work has been devoted to the impacts of in risk assessment and decision systems in criminal legal contexts. A smaller but still notably body of work exists specifically about feedback effects in repeated assessments and decisions.

2.1 The impacts of bail and sentencing decisions

Working within the social sciences, many economists, sociologists and criminologists have found deeply significant downstream effects of incarceration-related encounters and decisions. Starting with bail, there have been anumber of studies that show that bail decisions are profoudly impactful in a defendant's navigation through criminal legal procedures. Sacks and Ackerman [30] find that detention destabilizes family, increases expected incarceration length, and increases the likelihood of conviction. Dobbie et al. [10] find similar results: With compromised bargaining power, defendants who are detained before their trial are more likely to enter plea deals and incur guilty dispositions. Gupta et al. [14] find detention increases recidivism in Philadelphia, and another study found similar results in Texas [33, p672]. In Philadelphia, over half of people detained pretrial would be able to leave prison for a deposit of $1,000 or less, and many of these defendants are 'low-risk' - 60% of those held over three days were charged with non-violent crimes, and 28% just had a misdemeanor charge [32, p2]. Pretrial detention also increases expected court fees and sentence lengths [32]. A recently published study by Arnold et al. in 2018 used data from Miami and Philadelphia to find that judges exhibit significant racial bias in pre-trial release decisions, measured using offense rates of marginal white and black defendants [2]. Dobbie et al. [10] exploit randomness in Philadelphia court decisions to establish a causal impact of bail outcomes on criminal sentences and plea bargains.

On the question of whether incarceration lengths have a criminogenic effect, leading to higher incarceration prospects in the future, many studies have been conducted and have come to different conclusions. Camp and Gaes [5] find no criminogenic affect among 561 inmates in California with the 'same level of risk' who were distributed between Level I and Level III facilities - both were equally likely to be punished for misconduct in prison. Bhati and Piquero [4] attempt to estimate the impact of incarceration on subsequent offending trajectories, and find little criminogenic effect - the bulk of subsequent incapacitation came from some sort of violation of the terms of incarceration, such as parole. Nagin et al. [26] also observe a null or mildly criminogenic effect on future criminal behavior. Vieraitis et al. [34], using panel data over 30 years in 46 states, find a population deterrent effect of increased incarceration rates, but also find that increased prison release rates lead to higher rates of crime incidents, on average. Harding et al. [17] analyze the effects of imprisonment on felony convicts in Michigan and, using randomized judges to establish causal inference, find that a prison sentence increases the probability of subsequent imprisonment by 18-19%.

2.2 Group Fairness and Accumulated Disadvantage Studies

Disadvantage can accumulate over time. The notion of compounding effects in decision-making is intuitive – discrimination is instantiated when somebody consciously discriminates, but the effects of discrimination are often felt when the bias is more insidious and systemic. For example, even if gender-based discrimination is nearly undetectable at a single stage in a company's hiring or promotion process, executive teams tend to show remarkably little diversity [29]. Similar effects have been observed in education and wage rates, where a lifetime (or even inter-generational) time frame is needed to understand how bias becomes entrenched and can perpetuate over time.

Thus, statistical methods that try to find instances of discrimination may not capture biases that compound over repeated decisions. Another challenge for research is the difficulty of developing rigorous models of systemic effects. These processes can be highly complex because they involve information about history – something that traditional regression techniques lack. In a text entitled "Measuring Racial Discrimination" by the National Research Council in 2004, a chapter devoted to compounding effects concedes that the field is under-analyzed. The text observes, "Measures of discrimination that focus on episodic discrimination at a particular place and point in time may provide very limited information on the effect of dynamic, cumulative discrimination" [27, p226]. As a result, more research is needed, despite modeling difficulties. The authors write:

> Relatively little research has attempted to model or estimate cumulative effects. In part, this is because modeling and estimating dynamic processes that occur over time can be extremely difficult. The difficulty is particularly great if one is trying to estimate causal effects over time. [27, p224]

Indeed, theorists have found that survey and panel experimentation usually have not been able to capture the accumulating disadvantage that can cyclically affect a group of people, or cause divergent levels of wealth or status in society [25]. Instantiated experiments are unable to capture the dynamic nature of cumulative effects, and therefore often underestimate coefficients that determine measure of inequity.

2.3 Feedback Loops and Fairness in Sequential Machine Learning

A few studies have specifically concerned themselves with the idea that sequential decision-processes can exhibit feedback effects. In the criminology space, much of this inquiry originated with a finding by Lum and Isaac [24] that PredPol, a commonly used

software for police monitoring, exhibited feedback effects that could lead to certain neighborhoods being constantly patrolled and others never getting visited. Ensign et al. [11] extended this work using theoretical (and some simulated) Pólya Urn models to explain the policing disparities observed by Lum and Isaac [24].

More recently, D'Amour et al. [7] argue for more focus formal notions of fairness in dynamic decision environments, given that these domains can be difficult to analyze empirically. In this paper, we demonstrate empirically that the measure of 'risk' at points throughout criminal legal encounters do exhibit path-dependent dynamics and then use simulation, as suggested by D'Amour et al. [7], to demonstrate why the auditing procedures for current risk assessment algorithms are not adequate.

3 EVIDENCE OF FEEDBACK IN SENTENCING

We use criminal sentencing data in Philadelphia to provide empirical evidence that decisions in the criminal legal system are not only *informed* by risk but can *impact* formal measures of criminal risk. $n = 12,066$ court docket summaries from 2011 were pulled from the Philadelphia court system's website. The dockets contain demographic information, historical arrest and court outcomes in adult court, crime severity, disposition, sentence, and updated future court encounters and outcomes.

To test the impact of incarceration decisions on life-courses in Philadelphia, we leverage the fact that Philadelphia has *not* used algorithms to dictate sentencing decisions. Instead, we use risk factors as controls (covariates) and allow random variation in judicial sentencing decisions to understand the impact of disparate sentences on defendants *who otherwise have the same risk scores and severity of crime.* In doing so, we attempt to answer the following question empirically: Given two defendants with identical risk factors, how are differences in prison sentencing associated with cumulative future incarceration rates, measured up to 2 years after release?

For each individual defendant with a at least one charge, we compute the three Public Safety Assessment scores, which predict the probability of new criminal activity, new violent criminal activity. The scoring methodology is provided in the appendix. To control for the severity of a given crime, we include covariates to representing the typology of offense committed - felony and misdemeanor dummies, the 'degree' of the felony/misdemeanor, cross-terms, and the total number of guilty charges incurred.

We use a linear regression model with covariates X reported above. We aim to find the average incremental treatment effect of an extra day of sentenced prison time on the expected cumulative duration of prison sentences accrued until 2 years after the minimum sentenced time in prison. The treatment variable x_1 is measured using maximum sentences.

The potential for unobserved variable bias is important to note here, because judges may be seeing factors that are not reported in court docket summaries but may be relevant for sentencing. In particular, it is likely that judges cater sentences to different crimes that have the same grade, and may also cater sentences to particular combinations of multiple crimes that hold relevance for future incarceration prospects. To make sure our results are not representing our own shortcomings in modelling crime severity, we perform a second regression where we limit the sample to only defendants who commit the same crime, and who only are found guilty of that particular crime. We choose the most common crime in Philadelphia, "Manufacture, Delivery, or Possession with Intent to Manufacture or Deliver" - a non-violent felony with degree = 0. For our second regression, we take out factors that have to do with current criminal severity, since everybody is convicted with the same crime.

A regression was performed for all cases in the Court of Common Pleas, and an additional regression was performed on only those cases which have an identical, single guilty disposition for drug dealing. With models described above, we test for the average incremental treatment effect of a day in prison on the expected cumulative length of prison sentences, measured until two years after the minimum prison sentence. Results are reported below.

Regression results indicate that an additional day of sentencing is associated with 0.129 more days in prison sentences accrued two years after release, on average. For non-violent drug felony offenders, the estimated effect of incarceration is 0.094 extra days of prison time, on average. The regression that included all types of crime was statistically significant with $p < 0.01$, whereas the drug-only regression was statistically significant at $p < 0.05$.

While the results do provide evidence of a criminogenic impact of incarceration, it's important to note the possible alternative explanations for the observed treatment effect. First, unobserved variables might be influencing judge decisions. If judges use factors that were not controlled for and statistically correlate with future crime rates, we might observe the correlation in sentencing, which would suggest a causal relationship that is not only explained by differences in sentencing rates. We included the second regression because we were concerned that judges may use more granular information on the type of charge to decide a sentence. Another unobserved variable that may currently influence judicial decisions, since it is being adopted as part of Philadelphia's new sentencing tools, is juvenile delinquency history. Unless juveniles were tried in adult court, their record in inaccessible. While such a practice on face value seems to confirm our claim that sequential decisions in criminal justice compound (and are highly sensitive to inital conditions), being able to include juvenile information as another risk control would improve our confidence in the regression conclusions.

The results suggest that path-dependence plays an important role in carceral decision processes. Sentencing and criminal treatment decisions have profound impacts on the life-course, which is not necessarily captured by a one-shot measure of criminal risk. However, batch methods are being used to train risk scorers, and little attention is given to compounding effects.

4 MODEL PROBLEM SETTING

We offer a model of repeated high-impact decisions that will help us simulate the purpose and pitfalls of validation tests. We use a binary observation-decision system that allows each decision to impact the underlying propensity for a failed observation.

We can imagine this context as being a repeated parole decision, where an officer uses a risk score at each meeting to decide whether to impose a more restrictive policy on a parolee (e.g. curfew), thus

Table 1: Criminogenic Effect of Confinement

	Dependent Variable: 2-year min. cum. sentence	
	All charges	Only M/P/D
Treatment Variable:		
confinement_max	0.1286***	0.0938**
	(0.0249)	(0.0451)
Risk Factors:		
fta_score	-17.4921	-103.3698***
	(21.7752)	(38.4002)
nca_score	17.8048	66.3430***
	(13.5443)	(24.6393)
nvca_score	-12.3632	47.1921
	(23.4130)	(39.2216)
number_prior_crimes	0.7477	0.1927
	(5.8815)	(11.2610)
number_prior_violent	-7.3534	-19.4765
	(6.8279)	(13.1349)
prior_incarceration_flag	3.5053	-21.0556
	(23.1688)	(40.9863)
num_prior_arrests	4.9809**	4.6289
	(2.4096)	(5.1231)
prior_m	9.4684	33.6151
	(18.8997)	(31.9502)
prior_f	54.2776***	4.5463
	(18.3891)	(32.4370)
Demographics:		
age	-0.0128***	-0.0163***
	(0.0021)	(0.0041)
male_flag	45.5795**	65.2217
	(21.9017)	(52.8662)
black_flag	-7.8428	-12.8142
	(14.6471)	(25.5853)
plea_flag	-70.8348***	-155.8573***
	(19.8014)	(56.1067)
Current Crime Severity:		
felony_flag	-64.1872	
	(40.2028)	
misdemeanor_flag	-17.5152	
	(35.4556)	
degree	-33.2990	
	(24.0823)	
(felony_flag)(degree)	40.5926*	
	(23.1846)	
(misdemeanor_flag)(degree)	19.8258	
	(18.0401)	
count_guilty_charges	-19.5394**	
	(7.6839)	
current_violent_charge	29.0483	
	(17.7124)	
N	6215	1473
R^2	0.008	0.033
Adjusted R^2	0.007	0.023
F-statistic	7.323***	3.521***
Standard errors in parentheses.	$^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$	

limiting employment opportunities and increasing the probability of unlawful behavior. Each periodic parole meeting there is some observation of whether the rules were broken, a re-assessment of risk, and a new binary treatment decision. The context also has parallels in credit decisions, regulatory compliance checks, ad clicks, and more.

4.1 General Modelling Assumptions

We begin with a simple model of risk-needs driven decisions. Given that existing risk assessment services emphasize their wide applicability, some algorithms are adopted at numerous stages in criminal proceedings. Other jurisdictions may use different assessments for policing, bail, sentencing and parole. Starting simple, we model risk assessments as instantaneous binary decisions that are separated in time. Each decision occurs sequentially, and the outcome is either "high risk" or "low risk", as visualized in Figure 1.

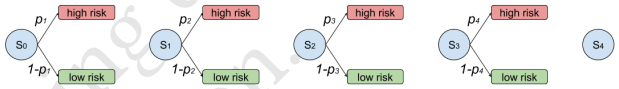


Figure 1: Sequential decision context diagram

We assume here that risk assessments are conducted T times throughout a person's life, and that the assessment r_t measures some underlying probability of future criminality $p_t \in [0, 1]$. The risk assessment r fully dictates a decision d_t, which denotes some choice of high-risk or low-risk treatment (e.g. increased surveillance, or prison security level):

$$d_t \in \begin{cases} 1, & \textit{if defendant is classified high-risk} \\ 0, & \textit{if defendant is classified low-risk} \end{cases}$$

We model each assessment using the current state of the world before decision t, denoted S_{t-1}.

The assessment is a random variable and not deterministic because risk assessment algorithms do not solely determine defendant outcomes - the ultimate decision is still up to a judge, who references the risk assessment score as part of the broader pre-trial policy decision.

We wish to explore the possibility that outcomes of assessments may impact and alter future assessments. As such, our model must enable us to analyze cases where the outcome variable X_i may impact the probability of high-risk classification for $X_{i+1}, X_{i+2}, ..., X_N$. The probability of a high-risk classification at decision i can thus be thought of as a function of some defendant information D_i (gender, race, age) and the history prior decisions, H_i. We write the current state of beliefs at i as $S_i = \{D_i, H_i\}$. We more accurately portray this dependence on the history of decisions as a branching process, rather than a sequence of decisions, in Figure 2.

Every major risk assessment algorithm uses information about criminal history to assess risk. PSA, for example, measures a defendant's number of prior misdemeanors, felonies, convictions, and violent convictions. These numbers add various point values to a

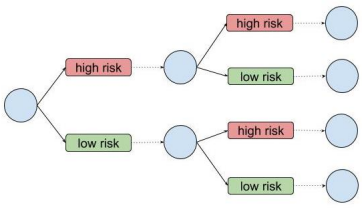


Figure 2: Branching and Path Dependence in a Binary Risk Classification Scorer

risk assessment score, and a threshold value may determine pre-trial detention or cash bail amounts. Therefore, the PSA and most (if not all) other algorithms have a reinforcement effect. After an individual is convicted with a felony charge, every subsequent risk assessment for the rest of his life will use his criminal history to increase his risk score. Thus, initial assessments of risk can hold more 'weight' in determining lifetime treatment than later assessments. If a person is identified as high-risk in their first encounter with the criminal system, known effects on future crime rates, employment, family life, taxes, and other features will increase the likelihood of subsequent encounters.

This property of *reinforcement* is key to modeling our system. The process is not Markovian: history matters, and our state of beliefs changes over time. Instead, we understand the changing effects of sequential risk-assessments as an Urn process, derived from the classic Pólya Urn model in mathematics [28].

4.1.1 *Dependence and Reinforcement.*

Let's say each risk assessment decision affects subsequent decisions as follows: If X_{i-1} is the risk-assessment outcome for decision $i-1$, the subsequent probability of a high-risk decision p_i is a weighted average between p_{i-1}, the prior probability, and X_{i-1}, the most recent classification:

$$p_i = p_{i-1}\left[\gamma_i\right] + X_{i-1}\left[1-\gamma_i\right], \quad i \in \{2, ..., N\}, \quad \gamma_i \in [0, 1]$$

This means that we model updates in risk score by averaging the prior assumed risk and the outcome of a new assessment. The X_{i-1} term can be thought of as the marginal effect of a new classification on defendant risk. To model reinforcement, we allow γ_i to increase as i increases, letting prior risk score p_{i-1} hold more importance as a defendant is older and has more history. This should make intuitive sense - if a defendant has lived out most of his life with a certain propensity for criminal activity ('risk'), the effect of a new assessment should carry less weight.

Using the above intuition, we'll start by assuming the following relationship between γ_i and i (the number of encounters with the criminal justice system):

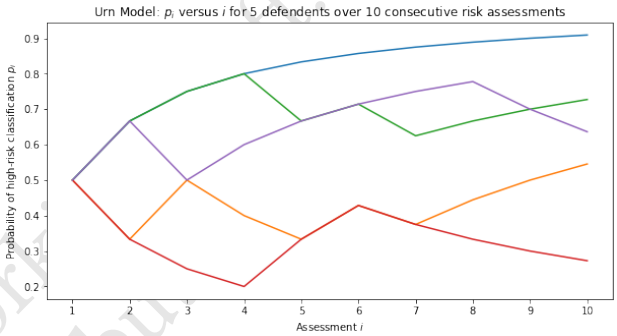
$$\gamma_i = \frac{i}{i+1}$$

To understand the equation above, let's consider the value of γ_i for varying i. In a first encounter with criminal courts where $i = 1$, we'd have $\gamma_1 = \frac{1}{2}$. Risk assessment outcome X_1 would thus have a very strong impact on future risk assessments. When i is high, however, γ_i approaches 1 and new assessments would diminish in weight. This is the reinforcement property we're seeking - the more decisions that go by, the less weighty they are in determining a person's lifetime experience with the state's criminal system.

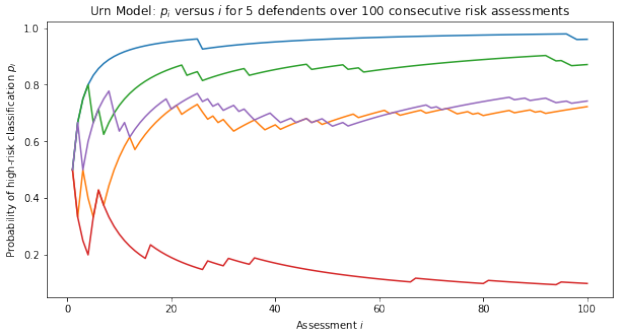
Thus, our formula for $P(X_i|D, H_i)$ is:

$$P(X_i|p_{i-1}, X_{i-1}) = p_{i-1}\left[\frac{i}{i+1}\right] + X_{i-1}\left[\frac{1}{i+1}\right], \quad i \in \{2, ..., N\} \quad (1)$$

Let's assume temporarily that every defendant starts off with a probability of high-risk classification $p_1 = \frac{1}{2}$. We model the effect of sequential risk-assessments for different defendants by implementing our iterative equation. Below are sample paths for 5 defendants who are subject to ten periodic, evenly spaced assessments over time:



In the plot above, each color represents an individual who encounters criminal risk assessments throughout their life. Notice that this plot behaves in accordance with the reinforcement effect - initial assessments have large effects on p_i, and later assessments only marginally change the course of the risk level. Indeed, the for very large i the risk level approaches a straight-line, meaning that the system reaches a stable propensity for criminal activity. Below are the paths of the same five defendants, this time over a total of 100 assessments (so 90 additional assessments):



While it is unrealistic that a single person would have one hundred exactly evenly spaced and identical assessments throughout their lives, the behavior of our model seems to cohere with our knowledge of risk-assessments - their output impacts future assessments in a way that reinforces their classification. In other words, people detained after being identified as high-risk are more likely to re-offend, spend time in jail, have financial trouble, lose employment, or receive a guilty charge - all of which will affect their level of 'risk'.

4.1.2 *Pòlya's Urn Generalization.*

The model derived above is an Urn process. Borrowing a few theorems from probability theory, we can begin to understand the large-scale, long-term effects that might come about when algorithms are used consecutively throughout a person's life.

Pòlya's Urn can be used to model path-dependent branching processes that are 'exchangeable', meaning the order of prior events does not matter.[2] The model asks what the long-term distribution of blue balls will be in the following random process:

- An urn contains R_t red balls and B_t blue balls. Start at $t = 0$, with an initial mix of R_0 and B_0 balls.
- for iteration $t \in \{1, ..., T\}$:
  - Pick a ball randomly from the urn.
  - For the ball picked, return it and k additional balls of the same color to the urn.

4.1.3 *Urn Equivalence to a Risk Assessment Model.*

We can model reinforcement in algorithmic decision-making as an urn process. Our basic defendant model replicates exactly the basic Pòlya process with $R_0 = 1$, $B_0 = 1$, and $k = 1$. We derive the equivalence in the two processes below.

Denote the color of the ball selected by pick $i \in \{1, 2, ..., N\}$ as:

$$\tilde{X}_i \in \begin{cases} 1, & \textit{if blue ball is picked} \\ 0, & \textit{if red ball is picked} \end{cases}$$

Assuming each ball is picked with equal probability, the probability of picking blue in is given by:

$$P(\tilde{X}_i = 1) = \frac{B_{i-1}}{B_{i-1} + R_{i-1}}$$

The total number of ball in the urn is $n_i = R_i + B_i$. The probability of picking blue given all prior picks is denoted as $\tilde{p}_i$. We can always find $\tilde{p}_i$ by dividing the number of blue balls in the urn by the total number of balls. We've shown that $p_i = \frac{B_{i-1}}{n_{i-1}}$. After the i^{th} pick, what will be the probability of picking blue? We inevitably add k balls into the urn, so $n_i = n_{i-1} + k$. In the event that our pick is red, we still have B_{i-1} blue balls, so the probability of picking blue decreases to $\frac{B_{i-1}}{n_{i-1}+k}$. If we do pick blue, however, the probability increases to $\frac{B_{i-1}+k}{n_{i-1}+k}$. Thus, the probability of picking blue on the $(i+1)^{th}$ pick, given B_0, n_0 and $\tilde{X}_1$, is:

$$\tilde{p}_{i+1} = \frac{B_{i-1} + \tilde{X}_i k}{n_{i-1} + k}$$

With a bit of algebra, we can define this probability in terms of the probability for the prior pick:

$$\tilde{p}_{i+1} = \frac{B_{i-1}}{n_{i-1} + k} + \tilde{X}_i \frac{k}{n_{i-1} + k} = \left[\frac{B_{i-1}}{n_{i-1}}\right] \frac{n_{i-1}}{n_{i-1} + k} + \tilde{X}_i \frac{k}{n_{i-1} + k}$$

$$\therefore \; \tilde{p}_{i+1} = \tilde{p}_i \frac{n_{i-1}}{n_{i-1} + k} + \tilde{X}_i \frac{k}{n_{i-1} + k}$$

[2] This is an assumption that may not hold true for our case, because many algorithms care about how *recent* a historical event took place. PSA, for example, cares about prior failures to appear in court in the past two years. However, for the most part, algorithms consider the aggregate number of historical events - number of prior felonies, misdemeanors, convictions, etc. These indicators are all *exchangeable* in the sense that it doesn't matter when in the defendant's life they occurred.

When $k = 1$ and $R_0 = B_0 = 1$, how does n_i behave? It starts at $n_0 = 2$, and after each pick it increments by $k = 1$. Thus, $n_i = 2 + i$. Equivalently, $n_{i-1} = 1 + i$, and $n_{i-2} = i$. Using the relationship derived above, a shift in index yields the probability of picking blue $\tilde{p}_i$ for $i \in \{2, ..., N\}$:

$$\tilde{p}_i = \tilde{p}_{i-1} \frac{n_{i-2}}{n_{i-2} + k} + \tilde{X}_{i-1} \frac{k}{n_{i-2} + k} = \tilde{p}_{i-1} \left[\frac{i}{i+1}\right] + \tilde{X}_{i-1} \left[\frac{1}{i+1}\right] \tag{2}$$

Notice the equivalence to equation 1. We've shown the probability for picking blue at each iteration of the classic Pólya Urn process exactly equals the probability of a high-risk classification in our simple model of sequential risk assessments, where $\tilde{p}_i = p_i$ and $\tilde{X}_i = X_i$.

4.2 Long Run Behavior

When we say that a sequence of random decisions might exhibit *reinforcement*, we now know that this means something deeper mathematically. Random processes with reinforcement behave in certain ways that might be problematic in the context of criminal policy. We have a general sense that algorithmic decisions in criminal justice impact defendants profoundly, and likely impact future encounters with law enforcement. Leveraging insights from probability theory, we can begin to understand the danger of policies that have compounding effects.

To start, we analyze the long-term treatment of individuals that are subject to sequential risk-based decisions. In Robin Pemantle's "A Survey of Random Processes with Reinforcement" (2006), the following theorem is reported about Pòlya's Urn process:

> Theorem 2.1: The random variable $p_i = \frac{B_i}{B_i + R_i}$ converges almost surely for large i to a limit P. The distribution of P is: $P \sim \beta(a, b)$ where $a = \frac{B_0}{k}$ and $b = \frac{R_0}{k}$. In the case where $a = b = 1$, the limit variable P is uniform on $[0, 1]$. [28]

Theorem 2.1 lays out how we can expect our modeled risk assessments to behave over many iterations. If one person undergoes risk assessments numerous times throughout their life, they may end up in radically different places depending on the risk-assessment outcome. They may be able to steer clear of subsequent confinement and re-arrest, or they may be continuously surveiled and repeatedly penalized by the state.

For a preliminary understanding of how inter-dependence in repeated risk assessments can impact a population, we use our initial modeling assumption that $p_1 = 0.5$ (so $B_0 = R_0$ and $a = b$), and imagine varying the parameter that determines the bearing of prior assessments on updated assessments, k (which defines γ). If we decrease k to 0.1 so that $a = b = \frac{B_0}{k} = 10$, we have the following long-term distribution for defendant risk. See Figures 3 and 4.

When decisions have little impact on people's lives (and potential subsequent risk assessments), we see consistency in long-term outcomes. Everyone starts with a risk score of 0.5, and all end up somewhere near there even after many assessments.

However, if algorithmic-driven decisions are more sensitive to the effect of prior decisions with $a = b = \frac{B_0}{k} = 0.1$, then we can see very problematic behavior in the long term. See Figures 5 and 6.

Figure 3: PDF of long term risk level when $k = 0.1$

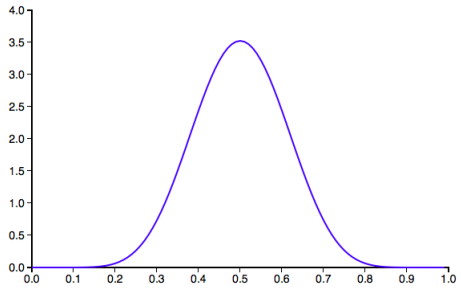


Figure 4: Urn Model Plot, p_i versus i for 30 defendants over 15 consecutive risk assessments, $k = 0.1$

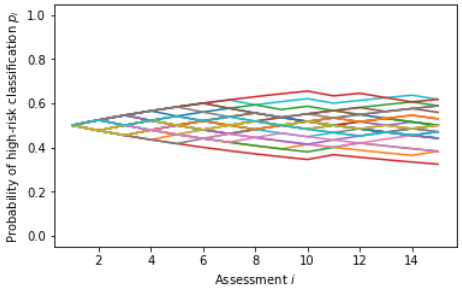


Figure 5: PDF of long term risk level when $k = 10$

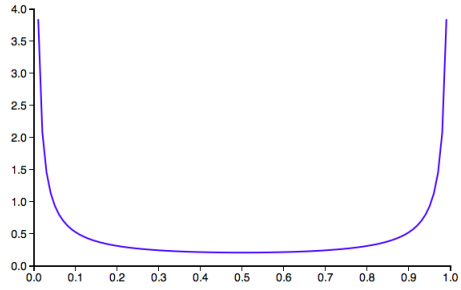
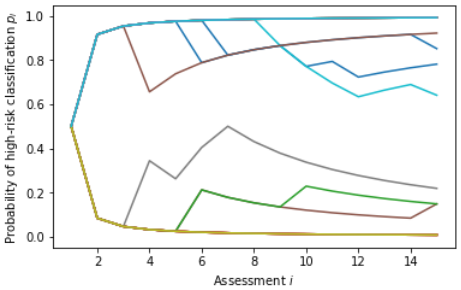


Figure 6: Urn Model Plot, p_i versus i for 30 defendants over 15 consecutive risk assessments, $k = 10$



In this second case, we begin with defendants that are identical in attributes, with an initial probability of high-risk classification $p_1 = 0.5$. However, simply because of the effect of risk-based decision making, defendants end up with radically different risk levels, and are highly likely to be pushed to an extreme (no criminal risk, 0, and extreme criminal risk, 1).

Of course, these results are purely theoretical and do not come from real observed processes. But they motivate the importance of scrutinizing how algorithms are used in practice. Algorithms may be validated to ensure that biases are mitigated to a certain confidence threshold. But even tiny disparities in the system described by the second plot above can profoundly impact outcomes.

4.3 Modelling Unequal Treatment

Many critics of risk assessment tools have expressed concern that these tools may encode biases that have historically characterized United States law enforcement. So far, our analysis of compounding effects has shown that these tools can lead to radically disparate treatment between people who began with the same risk factors. However, the analysis has not yet touched on existing and historical inequity. If a *biased* risk assessment tool were used, *and* it exhibited compounding effects, how might we expect bias to propagate over time? We can use our urn model to answer this question theoretically.[3]

4.3.1 *Disparate Initial Conditions.*

Risk assessment tools claim to add a level of consistency and 'objectivity' that judges lack without algorithmic assistance. Since judges have historically been biased in certain ways, many algorithmic tools boast that their improved accuracy can allow more people (of all groups) to leave detention pre-trial without increasing crime rates.

Even if we assume that our algorithm perfectly predicts risk and is able to eschew any kind of racially encoded bias, we know factually that risk is unevenly distributed across race.[4] A randomly selected black individual who finds himself arrested for a crime, therefore, is more likely to be labeled as high risk than an average white person in the same circumstances[5].

What are the long-term impacts of adopting algorithmic risk-assessments when risk is unevenly distributed across racial groups? How can our simple model of sequential risk assessments help us understand compounding effects and biased treatment?

Our first line of inquiry will look at the initial risk score that a defendant receives in a first encounter with the criminal justice system. Recalling our sequential decision-making model, we were able to describe the entire system with two quantities: the initial 'risk level' p_1 and the system's sensitivity to new decisions, $\frac{n_0}{k}$. What happens when we change the initial risk level, p_0, among defendants, and allow the rest of the process to remain the same?

Let's start by looking at what the expected value of our risk level, p_i, will be for time-step i, assuming only the prior risk p_{i-1}. We

[3] [18] discussees lowering the number of black people incarcerated as a potential goal for algorithmic criminal decisions.
[4] See [16].
[5] [15, The Virtues of Randomization] demonstrates that, as long as there is profiling, the arrested population will not accurately represent the true offending population demographically (absent perfect crime detection).

have from equation 2 that:

$$\tilde{p}_{i+1} = \tilde{p}_i \frac{n_{i-1}}{n_{i-1}+k} + \tilde{X}_i \frac{k}{n_{i-1}+k}$$

Taking the expectation over the linear equation:

$$E(p_{i+1}) = \frac{n_{i-1}}{n_{i-1}+k} E(p_i) + \frac{k}{n_{i-1}+k} E(X_i)$$

Using our knowledge that an indicator variable has expectation equal to its probability of being 1, we know:

$$E(p_{i+1}) = \frac{n_{i-1}}{n_{i-1}+k} p_i + \frac{k}{n_{i-1}+k} p_i = \frac{n_{i-1}+k}{n_{i-1}+k} p_i = p_i$$

Therefore, for any $p_i \in [0,1]$, the urn process maintains the same expected risk level, no matter how convergent or divergent the risk becomes over sequential decisions. This means that if black individuals are, on average, more likely to be labeled as high-risk individuals, our model of algorithmic risk assessments will not rectify these inequalities over time.

Some, including Kleinberg, believe that algorithmic risk assessment can lower the number of black people incarcerated [18]. Note that this is different from rectifying *inequalities* that exist in assessments: as long as the rate of white defendants decreases by the same rate proportion, the system is still treating more black people as high-risk than whites.

However, it is important to note that varying the initial probability of conviction does not lead to divergent effects for white and black people. The static expected risk for both groups implies that an initial bias will not perpetuate or magnify biases over time, according to our model. Purportedly unbiased algorithms can perpetuate and codify existing biases, therefore, but are unlikely to lead to divergent treatment as the result of initial conditions, according to our model.

4.3.2 *Entrenched Algorithmic Bias.*

Say, instead of assuming different initial probabilities of high-risk classifications for white and black folks, we instead assume that the algorithm itself produces biased judgments each time it makes a decision. Since no algorithm in use takes in race as an explicit variable, we may assume that race is reconstructed using correlated variables. Before, our urn model looked at risk assessments as a weighted average of prior risk belief and a random variable representing the most recent risk-assessment result. Now, let's add a race indicator to our weighting system. Now, each decision is a function of prior risk, the outcome of the most recent assessment, and the race of the defendant. If we denote the race of the defendant as a variable R, and write simply:

$$R \in \begin{cases} 1, & \textit{if defendant is black} \\ 0, & \textit{if defendant is white} \end{cases}$$

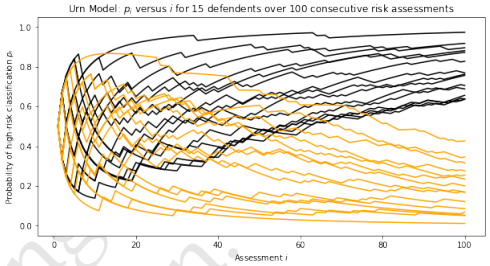
Then we can write the biased risk level at decision i as p_i^b, defined below:

$$p_i^b = p_{i-1}^b [\gamma_i] + R [\rho] + X_{i-1}^b [1 - \gamma_i - \rho],$$
$$i \in \{2, ..., N\}, \quad \gamma_i \in [0,1], \quad \rho \in [0, 1-\gamma_i]$$

We don't assume ρ to depend on i, as we might assume ρ to be a function of static features that do not change over time - education level, age at first arrest, family criminal history, etc.

When this is the case, we see that the bias affects every step in the algorithm and our system converges almost surely to 1 for black people and 0 for whites, so long as $\rho > 0$. Below are simulated risk assessments for adding a weight of 0.01 to each assessment - a level of bias that could go undetected in statistical validation experiments.

Figure 7: Urn Model p_i versus i for 15 defendants over 100 cumulative risk assessments, where two groups are plotted with differential treatment at each step



5 DISCUSSION

Understanding that sequential feedback-effects exist in criminal legal decisions forces us to re-evaluate the ways that validations are currently used.

This paper's empirical results suggest that when a defendant is sentenced to an extra day in prison in Philadelphia, they can expect to spend more than one extra day in prison over the course of their lifetime. There are numerous explanations for why this may be the case, and there are numerous implications for policy-makers.

The effect of prison time on future encounters with criminal punishment implies that algorithmic risk-assessment tools cannot be assessed using instantial experiments at one time in a defendant's life. We find that defendants tried in Philadelphia's Court of Common Pleas can expect to be arrested more than two more times in the future, regardless of the number of times they've been arrested in the past. If larger sentences are associated with greater prison time, it is likely that longer sentences hold bearing on future risk assessment. A more severe sentence may lead parole officers to have more discretion over parolees. It may increase a defendant's association with other criminals. This kind of dependence between decisions is clear from sentencing tables and three-strikes rules, which recommend that judges give exaggerated sentences to repeat-offenders.

Since judicial decisions appear to feed into one another sequentially over a defendant's life time, it is important to consider models that encompass compounding effects. Risk assessment algorithms and validation experiments fail to adequately address the potential of feedback effects over time. Rigorously considering the impacts if dependent, sequential decisions will be necessary for any high-stakes algorithm that makes decisions temporally. In the forthcoming section, we explore the possibility of compounding disadvantage and model problematic effects that may arise, undetected by instantial validation techniques.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* 23 (2016).

[2] David Arnold, Will Dobbie, and Crystal S Yang. 2018. Racial bias in bail decisions. *The Quarterly Journal of Economics* 133, 4 (2018), 1885–1932.

[3] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*. 62–76.

[4] Avinash Singh Bhati and Alex R Piquero. 2007. Estimating the impact of incarceration on subsequent offending trajectories: Deterrent, criminogenic, or null effect. *J. Crim. L. & Criminology* 98 (2007), 207.

[5] Scott D Camp and Gerald G Gaes. 2005. Criminogenic effects of the prison environment on inmate behavior: Some experimental evidence. *Crime & Delinquency* 51, 3 (2005), 425–442.

[6] Lucius Couloute and Daniel Kopf. 2018. Out of Prison Out of Work: Unemployment among formerly incarcerated people. *Prison Policy Initiative* (2018). https://www.prisonpolicy.org/reports/outofwork.html

[7] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[8] Matthew DeMichele, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. 2018. The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky. (2018).

[9] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* (2016).

[10] Will Dobbie, Jacob Goldin, and Crystal S Yang. 2018. The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review* 108, 2 (2018), 201–40.

[11] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*. 160–171.

[12] Tracy L Fass, Kirk Heilbrun, David DeMatteo, and Ralph Fretz. 2008. The LSI-R and the COMPAS: Validation data on two risk-needs tools. *Criminal Justice and Behavior* 35, 9 (2008), 1095–1108.

[13] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 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. *Fed. Probation* 80 (2016), 38.

[14] Arpit Gupta, Christopher Hansman, and Ethan Frenchman. 2016. The heavy costs of high bail: Evidence from judge randomization. *The Journal of Legal Studies* 45, 2 (2016), 471–505.

[15] Bernard E Harcourt. 2008. *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press.

[16] Bernard E Harcourt. 2014. Risk as a proxy for race: The dangers of risk assessment. *Fed. Sent'g Rep.* 27 (2014), 237.

[17] David J Harding, Jeffrey D Morenoff, Anh P Nguyen, and Shawn D Bushway. 2017. Short-and long-term effects of imprisonment on future felony convictions and prison admissions. *Proceedings of the National Academy of Sciences* 114, 42 (2017), 11103–11108.

[18] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2017), 237–293.

[19] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.

[20] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[21] Ryan M Labrecque, Paula Smith, Brian K Lovins, and Edward J Latessa. 2014. The importance of reassessment: How changes in the LSI-R risk score can improve the prediction of recidivism. *Journal of Offender Rehabilitation* 53, 2 (2014), 116–128.

[22] Edward J Latessa. 2016. Does Change in Risk Matter: Yes, It Does, and We Can Measure It. *Criminology & Pub. Pol'y* 15 (2016), 297.

[23] Christopher T Lowenkamp, Brian Lovins, and Edward J Latessa. 2009. Validating the level of service inventory—Revised and the level of service inventory: Screening version with a sample of probationers. *The Prison Journal* 89, 2 (2009), 192–204.

[24] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[25] Christopher J Lyons and Becky Pettit. 2011. Compounded disadvantage: Race, incarceration, and wage growth. *Social Problems* 58, 2 (2011), 257–280.

[26] Daniel S Nagin, Francis T Cullen, and Cheryl Lero Jonson. 2009. Imprisonment and reoffending. *Crime and justice* 38, 1 (2009), 115–200.

[27] D. The National Research Council Paer. 2005. Measuring Racial Discrimination. *SOCIAL FORCES* 83, 4 (2005), 1780.

[28] Robin Pemantle et al. 2007. A survey of random processes with reinforcement. *Probability surveys* 4 (2007), 1–79.

[29] Belinda Probert. 2005. 'I just couldn't fit it in': Gender and unequal outcomes in academic careers. *Gender, Work & Organization* 12, 1 (2005), 50–72.

[30] Meghan Sacks and Alissa R Ackerman. 2012. Pretrial detention and guilty pleas: if they cannot afford bail they must be guilty. *Criminal Justice Studies* 25, 3 (2012), 265–278.

[31] Jennifer L Skeem and Christopher T Lowenkamp. 2016. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology* 54, 4 (2016), 680–712.

[32] Megan T Stevenson. 2018. Distortion of justice: How the inability to pay bail affects case outcomes. *The Journal of Law, Economics, and Organization* 34, 4 (2018), 511–542.

[33] Megan T Stevenson and Sandra G Mayson. 2017. Bail reform: New directions for pretrial detention and release. (2017).

[34] Lynne M Vieraitis, Tomislav V Kovandzic, and Thomas B Marvell. 2007. The criminogenic effects of imprisonment: Evidence from state panel data, 1974–2002. *Criminology & Public Policy* 6, 3 (2007), 589–622.

[35] Brenda Vose. 2016. Risk assessment and reassessment: An evidence-based approach to offender management. *Criminology & Pub. Pol'y* 15 (2016), 301.

[36] Bruce Western and Sara McClanahan. 2000. Fathers behind bars: The impact of incarceration on family formation. (2000).