INVESTIGATING ALGORITHMIC RISK AND RACE

Melissa Hamilton*

Abstract

Risk assessment algorithms lie at the heart of criminal justice reform to tackle mass incarceration. The newest application of risk tools centers on the pretrial stage as a means to reduce both reliance upon wealth-based bail systems and rates of pretrial detention. Yet the ability of risk assessment to achieve the reform movement's goals will be challenged if the risk tools do not perform equitably for minorities. To date, little is known about the racial fairness of these algorithms as they are used in the field. This Article offers an original empirical study of a popular risk assessment tool to evaluate its race-based performance. The case study is novel in employing a two-sample design with large datasets from diverse jurisdictions, one with a supermajority white population and the other a supermajority Black population.

Statistical analyses examine whether, in these jurisdictions, the algorithmic risk tool results in disparate impact, exhibits test bias, or displays differential validity in terms of unequal performance metrics for white versus Black defendants. Implications of the study results are informative to the broader knowledge base about risk assessment practices in the field. Results contribute to the debate about the topic of algorithmic fairness in an important setting where one's liberty interests may be infringed despite not being adjudicated guilty of any crime.

Table of Contents

* Melissa Hamilton is a Reader in Law & Criminal Justice at the University of Surrey School of Law and obtained a Ph.D in criminology and criminal justice from The University of Texas at Austin and a J.D. from The University of Texas School of Law. This research was supported by funding from the Koch Foundation.

Introduction

Policymakers often tout risk assessment as the catalyst for achieving modern criminal justice reform.[1] State-of-the-art prediction methods meld big data, statistical analyses, and technological advances.[2] Risk assessment tools are a primary output of the evidence-based practices movement, whereby developers use scientifically derived correlates of criminal offending and package them into computational algorithms.[3] In general, algorithmic risk tools are assumed to improve the transparency,

1. *See generally* Sarah Brayne & Angèle Christin, *Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts*, Soc. Probs. 1 (2020).
2. Angèle Christin, *Algorithms in Practice: Comparing Web Journalism and Criminal Justice*, 4(2) Big Data & Soc'y 1, 1–2 (2017).
3. Faye S. Taxman, *The Partially Clothed Emperor: Evidence-Based Practices*, 34 J. Contemp. Crim. Just. 97, 97–98 (2018).

objectivity, consistency, and fairness of decisions that triage offenders for management purposes.[4]

The United States simply incarcerates far more offenders than it needs to safeguard its citizens.[5] Criminal justice officials contribute to mass incarceration when they rely upon their own intuitions about offenders' risk, as the tendency is to overestimate dangerousness out of an abundance of caution.[6] Risk tools may temper such predispositions by offering a scientifically led method for identifying large segments of criminal justice populations who are at low risk of reoffending.[7] Algorithmic risk thereby offers a systematic method for reducing mass incarceration without endangering public safety.[8]

Notwithstanding many potential advantages, risk assessment tools may lead to negative consequences if they do not exhibit sufficiently accurate predictions or do not treat protected groups fairly.[9] Inaccurate algorithmic outcomes may yield too many false positives or false negatives, which can result in harm to offenders or to community members.[10] Thus, while reducing mass incarceration is a laudable goal overall, risk assessment tools may serve such an interest in ways that disparately burden minorities.[11] The potential for race-based discrimination is an emerging debate regarding the new risk assessment model.[12] The tolerance for discretion has allowed an environment in which criminal justice authorities often focus more scrutiny on minorities than is justified.[13] Concern is understandable to the extent that an algorithm simulates such mistreatment

4. Megan Stevenson & Sandra G. Mayson, *Pretrial Detention and Bail*, 3 Reforming Crim. Just. 21, 34 (2017).
5. Michael O'Hear, *Actuarial Risk Assessment at Sentencing, Potential Consequences for Mass Incarceration and Legitimacy*, 38 Behav. Sci. & L. 193, 193–94 (2020).
6. *See, e.g.*, Alexa Van Brunt & Locke E. Bowman, *Toward a Just Model of Pretrial Release: A History of Bail Reform and a Prescription for What's Next*, 108 J. Crim. L. & Criminology 701, 737 (2018); Malcolm M. Feeley, *How to Think about Criminal Court Reform*, 98 B.U. L. Rev. 673, 702 (2018); Patrick Liu et al., The Economics of Bail and Pretrial Detention 13 (2018).
7. Lauryn P. Gouldin, *Defining Flight Risk*, 85 U. Chi. L. Rev. 677, 681 (2018).
8. Christopher Bavitz et al., Assessing the Assessments: Lessons from Early State Experiences in the Procurement and Implementation of Risk Assessment Tools 1 (2018).
9. Nicholas Scurich & Daniel A. Krauss, *Public's Views of Risk Assessment Algorithms and Pretrial Decision Making*, 26 Psychol. Pub. Pol'y & L. 1, 1 (2020).
10. Paul Hayes et al., *Algorithms and Values in Justice and Security*, 35 AI & Soc'y. 533, 535 (2020).it represents one step in a value sensitive design based methodology (not incorporated here are empirical and technical investigations
11. Anupam Chander, *The Racist Algorithm?*, 115 Mich. L. Rev. 1023, 1034 (2017).
12. *See* Richard Berk, *Accuracy and Fairness for Juvenile Justice Risk Assessments*, 16 J. Empir. Legal Stud. 175, 175 (2019); Osonde Osaba & William Welser IV, An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence 19 (2017) *(positing a* "*reasonable algorithm*" may fail to result in fair and equitable treatment of diverse populations).
13. *See generally* Bernard E. Harcourt, Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age (2006).

of minorities, and then exacerbates the problem by replicating injustices on a more technological level and in a more systematically efficient way.[14]

Risk algorithms generate "automated suspicion" of individuals designated as high risk.[15] Indeed, algorithmic risk assessment is a forthright method of human profiling *meant* to discriminate in the sense of segregating individuals into disparate degrees of risk classifications based on their observable characteristics.[16] Critics thus charge that the risk assessment scheme has turned into one of the digital racialization of risk that overpredicts recidivism for minorities and reifies labeling minorities as dangerous and risky.[17] The Pretrial Justice Institute, once a prominent advocate for risk assessment, recently reversed course, explaining its epiphany: "There is no pretrial justice without racial justice."[18]

Due to these conflicting stances on the upsides and downsides of risk assessment, stakeholders are calling for independent researchers to audit risk tools as a form of due diligence.[19] To date, relatively few studies exist in the public realm that evaluate whether algorithmic risk tools perform in ways that are equitable to minority groups.[20] Results of independent research can productively inform policymakers and stakeholders who are interested in ensuring the utility and fairness of the algorithms they may employ.[21] This Article responds to these calls by reporting on an original empirical study evaluating the fair performance of a popular risk assessment tool that its owner markets as a national tool and in practice is used by jurisdictions across the country (i.e., the Public Safety Assessment). The main issue herein involves racial equality and

14. In the popular media, criminal justice risk algorithms are at times portrayed as "dystopian, science-fiction scenarios run awry." Arthur Rizer & Caleb Watney, *Artificial Intelligence Can Make our Jail System More Efficient, Equitable and Just*, 23 Tex. Rev. L. & Pol. 181, 183 (2018).
15. Roger Brownsword & Alon Harel, *Law, Liberty and Technology: Criminal Justice in the Context of Smart Machines*, 15 Int'l J.L. Context 107, 112 (2019).
16. Melissa Hamilton, Risk Assessment Tools in the Criminal Legal System—Theory and Practice: A Resource Guide 2–3 (2020).
17. Pamela Ugwudike, *Digital Prediction Technologies in the Criminal Justice System: The Implications of a "Race Neutral" Agenda*, 24 Theoretical Criminology 482, 483 (2020).
18. Pretrial Just. Inst., A Racial Equity Transformation: PJI's Rationale 1 (2019).
19. Bavitz et al., *supra* note 8, at 19.
20. Ugwudike, *supra* note 17, at 492; Naomi Murakawa, *Racial Innocence: Law, Social Science, and the Unknowing of Racism in the US Carceral State*, 15 Ann. Rev. L. & Soc. Sci. 473, 480 (2019); James T. McCafferty, *Unjust Disparities? The Impact of Race on Juvenile Risk Assessment Outcomes*, 29 Crim. Just. Pol'y Rev. 423, 427 (2018).
21. *See, e.g.*, Grant Duwe & Michael Rocque, *The Predictive Performance of Risk Assessment in Real Life: An External Validation of the MnSTARR*, Corrections Pol'y Prac. & Res. (2019); Faye S. Taxman, *Risk Assessment: Where do We Go From Here?*, in Handbook of Recidivism Risk/Needs Assessment Tools 271, 273, 277 (Jay P. Singh et al. eds., 2018).

for these purposes we are interested in potential differences in treatment of white versus Black individuals.[22]

This Article proceeds as follows. Part I reviews the advancement of algorithmic risk practices in criminal justice, with a focus on their importance in pretrial decisions. The pretrial context is critical as stakeholders are realizing how markedly pretrial detention contributes to mass incarceration, while reliance on money bail aggravates wealth-based and race-based disparities. The algorithmic risk wave shows promise in supporting pretrial justice reforms to increase pretrial release rates while acting as an alternative to bail systems. Yet such promise can only be achieved if the risk tools are shown to perform with adequate equity in field settings. The Part then reviews arguments for how, at least theoretically, algorithmic risk may specifically improve the chances of release for Black defendants, as statistics indicate that, historically, Black defendants are more likely to be detained pretrial and less able to afford the assigned bail amounts to secure their release.[23] Contrary perspectives concern reasons that, instead, Black defendants may be harmed when the algorithms learn on already biased data, imbed these structural inequalities, and further relaunder inequities through the guise of objective risk predictions.

Part II describes the Public Safety Assessment (PSA), its development, and the reasons for its popularity. We use here the PSA algorithm that predicts any new criminal arrest. The study is unique in evaluating two samples from diverse jurisdictions in Illinois and Kentucky that actively use the PSA to inform judges when making pretrial decisions on release. The combined dataset is over 200,000 defendants, which provides a large sample size to run robust analyses. An extraordinary advantage of the two-sample design is to be able to compare these contrasting sites on racial fairness grounds, as the Illinois dataset presents a supermajority Black sample while the Kentucky dataset offers a supermajority white sample. The main research questions are these: Does the tool result in disparate impact in disadvantaging Black individuals by placing a greater percentage of them into higher risk bins? Does the tool exhibit test bias in predictive accuracy by racial grouping? Does the tool exhibit differential validity by race? The latter question entails investigating the accuracy of the tool's predictions in general and then enquires whether the tool's accuracy and error rates materially vary between races.

22. While the legal concepts of disparate impact and racial fairness are central, this Article is not a constitutional review of the pretrial system or of the use of algorithms to assist in the criminal justice system. Those issues have been extensively debated elsewhere. *See generally* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 Duke L.J. 1043 (2019); Paul Heaton et al., *The Downstream Consequences of Misdemeanor Pretrial Detention*, 69 Stan. L. Rev. 711 (2017); Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 Am. Crim. L. Rev. 231 (2015).

23. *See infra* notes 107–111 and accompanying text.

Part III displays the results in responding to the research questions. Policy implications of the research findings are meant as informational points for academics, stakeholders, and practitioners for how a popular risk algorithm performs in real world settings. Suggestions provide potential avenues to improve the utility and fairness of the tool as a result. Key conclusions are relevant to other criminal justice settings, as well, where risk tools are considered.

The present study employs an analytical strategy that seeks to improve upon previous research by: (a) using datasets from racially contrasting sites; (b) investigating the potential for adverse impact; (c) employing the gold standard from the psychometric literature to check for test bias; (d) estimating multiple measures of performances for a more vigorous analysis; and (e) parsing a risk assessment tool's performance by subgroups defined by race.

I. Risk Assessment Tools as a Criminal Justice Reform

Risk assessment practices in criminal justice present an exemplary combination of forensic tests, laws pertaining to decisions such as pretrial release or sentencing, and policies contributing to mass incarceration. Forensic psychology is a branch that adapts psychological science to resolve legal questions.[24] An important novelty has involved the use of psychometry to create scientific rules and measures in the form of algorithms designed to predict offenders' risk of recidivism.[25]

The 'evidence-based practices movement' is the now-popular term to describe the turn to behavioral sciences data to improve risk-based classifications. Scientific studies targeting recidivism outcomes are benefiting from the compilation of large datasets (i.e., big data) of discharged offenders. Researchers track the offenders post-release, observe recidivism rates, and then statistically test which factors correlate with recidivism. Risk assessment tool developers use computer modeling to combine factors of sufficiently high correlation and weight them accordingly using increasingly complex algorithms.[26]

Algorithm-based risk assessment tools help inform officials in their decisions concerning offenders, such as whether to incarcerate or release.[27] A focus on incarceration is intentional. The United States is regrettably renowned for being the world's leader in mass incarceration, with a detention rate exceeding any other country.[28] One out of

24. Tess M. S. Neal et al., *Psychological Assessments in Legal Contexts: Are Courts Keeping "Junk Science" Out of the Courtroom?*, 20 Psychol. Sci. Pub. Int. 135, 139 (2019).
25. *Id.* at 135–36.
26. Melissa Hamilton, *Debating Algorithmic Fairness*, 52 UC Davis L. Rev. Online 261, 266 (2019).
27. Carolyn McKay, *Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, AI and Judicial Decision-Making*, 32 Current Issues Crim. Just. 22, 26 (2019).
28. Roy Walmsley, World Prison Population List 2 (2020).

every five prisoners globally reside in America's prisons and jails.[29] The situation has become untenable due to the sheer financial burdens of operating the prison industrial complex.[30] Consequently, bipartisan, public support has emerged to engage policy reforms to significantly reduce incarceration numbers.[31]

Algorithmic risk tools have become the weapon of choice to combat mass incarceration.[32] Indeed, risk assessment tools are now considered best practice to lead justice reform efforts.[33] Some research indicates that science-informed risk tools help alleviate high rates of imprisonment by convincing decisionmakers that substantially more subjects may be safely supervised in their communities.[34] The new algorithmic risk wave has been actively targeted to backend decisions, such as sentencing and parole.[35] More recently, the pretrial context has become a "flashpoint" in efforts to reduce mass incarceration.[36]

A. *Pretrial Reforms*

Pretrial detention is a significant driver of mass incarceration.[37] One-fifth of the nation's incarcerated population is comprised of individuals yet to be convicted.[38] At any time, about 500,000 inmates are defendants waiting in the country's jails for their trial dates.[39] One in five local jails are operating at or over capacity.[40] The churn is significant. In 2018 alone, local jails counted almost 11 million bookings.[41] America is a world leader here as well. While the country's total population is 4 percent of the global population, it houses 20 percent of the world's pretrial detainees.[42]

---

29. *Id.*
30. The costs to American taxpayers to operate America's prison systems are $80 billion a year. Nicole Lewis & Beatrix Lockwood, *The Hidden Cost of Incarceration*, The Marshall Project (Dec. 17, 2019, 5:00 AM), https://www.themarshallproject.org/2019/12/17/the-hidden-cost-of-incarceration [https://perma.cc/H2Z7-GGQD]; Urban Justice Center, The Prison Industrial Complex 1 (2018).
31. Brandon L. Garrett & John Monahan, *Judging Risk*, 108 Calif. L. Rev. 439, 447 (2020).
32. Feeley, *supra* note 6, at 690.
33. Kristin Bechtel et al., *A Meta-Analytic Review of Pretrial Research: Risk Assessment, Bond Type, and Interventions*, 42 Am. J. Crim. Just. 443, 446 (2017).
34. *Id.* at 447.
35. Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 Emory L.J. 59, 67 (2017).
36. Scurich & Krauss, *supra* note 9, at 5.
37. *See* Colin Doyle et al., Bail Reform: A Guide for State and Local Policymakers 7 (2019).
38. Will Dobbie & Crystal Yang, Proposals for Improving the U.S. Pretrial System 4 (2019).
39. *See* Zhen Zeng, Jail Inmates in 2018 1 (2020).
40. *Id.* at 8.
41. *Id.* at 1; Wendy Sawyer & Peter Wagner, Mass Incarceration: The Whole Pie 2020 5 (2020).
42. Doyle et al., *supra* note 37, at 7.

In the pretrial setting, judges are typically the final arbiters of decisions on whether to release jailed individuals.[43] From a constitutional perspective, infringing upon liberty interests by incarcerating significant numbers who have not been convicted of any offense invokes due process considerations.[44] The Supreme Court in *United States v. Salerno* accepted that an appropriate balance justified the practice: "In our society liberty is the norm, and detention prior to trial or without trial is the carefully limited exception."[45] While a reasonable rule of thumb might presume to release those who have not yet been convicted of any crime, judges tend to be risk averse and thus often order pretrial detention as a means to ensure the defendants' appearance for trial and/or to protect the community from potential offending if released.[46]

Even if discharge appears justified, the judge may order release contingent upon posting bail as security.[47] The bail amount is not necessarily linked to an estimate of the individual's ability to pay it or to the likelihood of criminal offending.[48] Many jurisdictions use definitive bail schedules with the specific monetary amount tied to the seriousness of the crime(s) of arrest.[49] These types of bail schedules use the pending criminal charge(s) as a proxy for the risk of failure (i.e., flight or rearrest), with the monetary sums intended to mitigate such risk.[50]

Bail systems produce negative consequences. Many defendants otherwise eligible for release remain behind bars simply because they cannot afford the required bail amount.[51] Detaining individuals who are unable to pay the bond raises questions about problematic wealth-based disparities.[52] Further, this scenario perverts the risk scheme as individu-

43. Brandon P. Martinez et al., *Time, Money, and Punishment: Institutional Racial-Ethnic Inequalities in Pretrial Detention and Case Outcomes*, 66 Crime & Delinq. 837, 839 (2020).
44. Hafsa S. Mansoor, *Guilty Until Proven Guilty: Effective Bail Reform as a Human Rights Imperative*, DePaul L. Rev. (2020); Glen J. Dalakian II, *Open the Jail Cell Doors, HAL: A Guarded Embrace of Pretrial Risk Assessment Instruments*, 87 Fordham L. Rev. 325, 350 (2018).
45. United States v. Salerno, 481 U.S. 739, 755 (1987).
46. Brandon Buskey, *Wrestling with Risk: The Questions Beyond Money Bail*, 98 N.C. L. Rev. 379, 384 (2020); Gouldin, *supra* note 7, at 680–81; Heaton et al., *supra* note 22, at 716–17.
47. Van Brunt & Bowman, *supra* note 6, at 728 n. 132.
48. Doyle et al., *supra* note 37, at 7.
49. Rizer & Watney, *supra* note 14, at 187, 193; John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 Wash. L. Rev. 1725, 1744 (2018).
50. Ram Submarian et al., Incarceration's Front Door: The Misuse of Jails in America 32 (2015).
51. Garrett & Monahan, *supra* note 31, at 442; Sarah Picard et al., Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness 3 (2019).
52. *See* Lauryn P. Gouldin, *Reforming Pretrial Decision-making*, 55 Wake Forest L. Rev. 857, 861 (2020).

als who continue to be detained for an inability to post bail can present lower risk profiles than those who are released.[53]

Strikingly, "pretrial decisions determine mostly everything,"[54] acting as "incarceration's front door."[55] Detained defendants are more likely to plead guilty, not necessarily because of factual guilt, but due to being disadvantaged in their ability to work with counsel or prepare for trial, while at the same time the coercive atmosphere of jail pressures them to plead in an attempt to resolve their unsettled statuses.[56] Facing such a pretrial model as this, it is assuredly "better to be guilty, dangerous, and rich than to be innocent, harmless, and poor."[57]

Bail systems that do not link the bond with the individual's ability to pay inevitably drive up their pretrial jail population numbers.[58] Pretrial detention exacerbates mass incarceration even further as it is correlated with increasing the likelihood of being convicted at trial and with receiving a longer prison sentence.[59] Indeed, imprisonment itself is criminogenic in nature in that those who are incarcerated—including in a preconviction context—are more likely to offend in the future.[60] In contrast, for many pretrial detainees, early release materially curtails their risk of failure, thereby allowing them to avoid the jail's revolving door, which otherwise further contributes to mass incarceration.[61]

Enter the reform movement adopting algorithmic risk tools. A risk-based decision format speaks to efforts to reduce the contribution of pretrial detention to mass incarceration. The algorithmic trend

53. Emily Tiry et al., Road Map to Pretrial Reforms 2 (2016); Rizer & Watney, *supra* note 14, at 187.
54. Bechtel et al., *supra* note 33, at 444 (quoting Candace McCoy, *Caleb was Right: Pretrial Decisions Determine Mostly Everything*, 12 Berkeley J. Crim. L. 135, 135 (2007)).
55. Sawyer & Wagner, *supra* note 41.
56. Doyle et al., *supra* note 37, at 8; Lydette S. Assefa, *Assessing Dangerousness Amidst Racial Stereotypes: An Analysis of the Role of Racial Bias in Bond Decisions and Ideas for Reform*, 108 J. Crim. L. & Criminology 653, 668 (2018); Ellen A. Donnelly & John M. MacDonald, *The Downstream Effects of Bail and Pretrial Detention on Racial Disparities in Incarceration*, 108 J. Crim. L. & Criminology 775, 789 (2018).
57. Brook Hopkins et al., *Principles of Pretrial Release: Reforming Bail Without Repeating its Harms*, 108 J. Crim. L. & Criminology 679, 680 (2018); Bryanna Fox et al., *Psychological Assessment of Risk in a County Jail: Implications for Reentry, Recidivism and Detention Practices in the USA*, 9 J. Crim. Psychol. 173, 174 (2019).
58. Shima Baradaran Baughman, *Dividing Bail Reform*, 105 Iowa L. Rev. 947, 1004 (2020).
59. Donnelly & MacDonald, *supra* note 56, at 791; Van Brunt & Bowman, *supra* note 6, at 745; Stevenson & Mayson, *supra* note 4, at 22 n. 5.
60. Léon Digard & Elizabeth Swavola, Justice Denied: The Harmful and Lasting Effects of Pretrial Detention 6 (2019); Koepke & Robinson, *supra* note 49, at 1746 n. 94.
61. Koepke & Robinson, *supra* note 49, at 1769–70.

also seeks to curtail reliance upon money bail in lieu of a risk-based release scheme.[62]

The algorithmic risk agenda is a form of e-government expected to offer multiple advantages.[63] Algorithmic risk tools may ameliorate human biases that, unconsciously or not, likely infect human decisions about future dangerousness.[64] Selections based on personal opinions of which individuals might succeed if released are prone to bias and reliance on heuristics.[65] Decisional shortcuts are even more likely for judges in a pretrial context as they typically gain little access to information, have few interactions with the individuals, and must make on-the-spot calls with limited time for reflection.[66] Pretrial detention hearings are hasty affairs, often occurring without the benefit of counsel who may otherwise provide the judge with more contextualizing information about the defendant's prospects for successful reentry.[67] Algorithmic outcomes instead offer the promise of objectivity to reduce various human biases.[68] The data-driven science underlying the tools provide an empirically-informed method to estimate a pretrial defendant's risk of criminal offending if released.[69]

In turn, algorithmic risk forecasts are more likely to be correct than human oracles.[70] Indeed, one of the strongest and consistent findings of psychological research is that judgments based on science are more accurate than those reliant upon individual intuition.[71] Judges'

62. Koepke & Robinson, *supra* note 49, at 1746; *see also* Stevenson & Mayson, *supra* note 4, at 23 (describing it as a "shift from the 'resource-based' system of money bail to a 'risk-based' system, in which pretrial interventions are tied to risk rather than wealth").
63. Cary Coglianese & Lavi M. Ben Dor, *AI in Adjudication and Administration*, 86 Brook. L. Rev. (forthcoming 2021).
64. Rizer & Watney, *supra* note 14, at 184; Seth J. Prins & Adam Reich, *Can we avoid reductionism in risk reduction?*, 22 Theoretical Criminology 258, 259 (2018); Stevenson & Mayson, *supra* note 4, at 34.
65. "Implicit bias often occurs when people are asked to resolve a complex issue with limited time and must resort to sterotypes, mental shortcuts, or other rules of thumb to quickly solve the issue. This cognitive process that relies on mental shortcuts or existing schemas is rapid, intuitive, automatic, and error-prone." Assefa, *supra* note 56, at 658.
66. David Arnold et al., *Racial Bias in Bail Decisions*, 133 Q. J. Econ. 1885, 1887 (2018).
67. Stevenson & Mayson, *supra* note 4, at 25 ("In practice, bail hearings are a messy affair.").
68. Prins & Reich, *supra* note 64, at 259.
69. Sarah L. Desmarais & Evan M. Lowder, Pretrial Risk Assessment Tools: A Primer for Judges, Prosecutors, and Defense Attorneys 1 (2019); Matthew DeMichele et al., *Public Safety Assessment: Predictive Utility and Differential Prediction by Race in Kentucky*, 19 Criminology & Pub. Pol'y 1, 1 (2020).
70. Coglianese & Ben Dor, *supra* note 63; Rizer & Watney, *supra* note 14, at 183; Prins & Reich, *supra* note 64 at 259; Stevenson & Mayson, *supra* note 4, at 34.
71. Karen Bogenschneider & Bret N. Bogenschneider, *Empirical Evidence from State Legislators: How, When, and Who Uses Research*, 26 Psychol. Pub. Pol'y & L. 413, 413 (2020).

uncorrected biases tend toward the direction of overestimating the potential to endanger public safety.[72] Algorithmic risk scores thereby act as a decision-support datapoint which may convince judges who otherwise are risk averse to discharge significantly more offenders who score at lower risk.[73] The evidence-based platform might increase release rates further by insulating judges from allegations that they wrongfully released an individual who endangered the public.[74] Thus, if a released offender commits some bad act, the scheme offers a strategy for humans to avoid blame by shifting responsibility to the algorithm.[75]

The standardization inherent in algorithmic outcomes offers several benefits: improving the efficiency of pretrial decision processes;[76] reducing inconsistencies across decisions;[77] making pretrial release decisions more equitable;[78] and rendering release decisions more legally justifiable.[79]

Notwithstanding the foregoing advantages, whether risk tools are sufficiently accurate is not a given, and questions are being raised on this score, as discussed further below. There is certainly a critical need to get risk assessment "right" to raise its legitimacy in facilitating release decisions to combat mass incarceration.[80] As a result, a better understanding of how algorithmic risk tools perform in pretrial settings is imperative to the mission of energizing release practices while keeping safety in mind.

Supplemental reasons to concentrate on the experience with risk algorithms in local jails are evident. Pretrial may be the test bed for the

72. Gouldin, *supra* note 7, at 680–81.
73. Scurich & Krauss, *supra* note 9, at 6; O'Hear, *supra* note 5, at 194; Desmarais & Lowder, *supra* note 69.
74. O'Hear, *supra* note 5, at 199; Megan T. Stevenson & Jennifer L. Doleac, Algorithmic Risk Assessment in the Hands of Humans 61 (2019); Stevenson & Mayson, *supra* note 4, at 34; Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 Fed. Sent'g Rep. 237, 237 (2015). The fear of political accountability is particularly salient in jurisdictions where judges are elected and worry about reelection prospects if they release individuals who commit new crimes and create bad publicity.
75. Harry Surden, *Ethics of AI in Law: Basic Questions*, *in* The Oxford Handbook of Ethics of AI 734 (Markus D. Dubber et al. eds., 2020); Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 Crim. Just. & Behav. 185, 203 (2019); Kelly Hannah-Moffat et al., *Negotiated Risk: Actuarial Illusions and Discretion in Probation*, 24 Can. J.L. & Soc'y 391, 398 (2009).
76. Coglianese & Ben Dor, *supra* note 63; Rizer & Watney, *supra* note 14, at 183; *see also* Stevenson & Mayson, *supra* note 4, at 34.
77. Cecelia Klingele, *Making Sense of Risk*, 38 Behav. Sci. & L. 218, 220 (2020); Guido Noto La Diega, *Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information*, 9 J. Intell. Prop. Info. Tech. & Elec. Com. L. 3, 33 (2018).
78. *See* Desmarais & Lowder, *supra* note 69, at 1.
79. Harcourt, *supra* note 74; *see* Stevenson & Doleac, *supra* note 74, at 6–7; *See* Stevenson & Mayson, *supra* note 4.
80. Garrett & Monahan, *supra* note 31, at 446.

broader criminal justice system looking to adopt algorithmic risk tools because pretrial decisions are less complicated than other decision points in the criminal justice process and pretrial outcomes are more quickly known and easier to measure.[81] These theoretical and practical circumstances facilitate research into the real-world operation of risk tools. Then, a pretrial risk tool is considered a "high stakes" instrument because the individual's liberty is at issue.[82] Any information from studies is therefore desirable concerning risk tool performance when used to inform outcomes that infringe constitutional interests. Empirical research investigating accuracy presents in the forensic sciences field with the terminology of validation studies.

B. *Validation Studies*

Agencies either develop their own risk tools or adopt one "off-the-shelf" by selecting a preexisting tool trained on other samples.[83] In either case, the tools' employment of sometimes sophisticated algorithms offer the guise of science and objectivity, which may unfortunately lead to a false sense of security about their utility.[84] To serve the mission of safely releasing more defendants, risk tools must be sufficiently accurate in how they perform in the field.[85]

Validation methods evaluate how well a tool performs with what it is designed to accomplish.[86] To date, relatively little is known about the accuracy of algorithmic risk tools as there has been insufficient audit or scrutiny of them.[87] Of the scant validation studies which have been publicly released on risk tools generally, the available evidence has shown low to moderate rates of accuracy.[88] In the criminal justice risk assessment

81. *See* Rizer & Watney, *supra* note 14, at 184.
82. Grant Duwe, *Better Practices in the Development and Validation of Recidivism Risk Assessments: The Minnesota Sex Offender Screening Tool–4*, 30 Crim. Just. Pol'y Rev. 538, 548 (2019).
83. Koepke & Robinson, *supra* note 49, at 1748; Zachary Hamilton et al., *Designed to Fit: The Development and Validation of the STRONG-R Recidivism Risk Assessment*, 43 Crim. Just. & Behav. 230, 231–32 (2016).
84. Coglianese & Ben Dor, *supra* note 69, at 26; *see* McCafferty, *supra* note 20, at 424.
85. *See* Neal et al., *supra* note 24, at 137.
86. Partnership on AI, Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System 14 (2019).
87. Sarah L. Desmarais et al., *Performance of Recidivism Assessment Instruments in U.S. Correctional Settings*, *in* Handbook of Risk/Needs Assessment Tools 3, 19 (Jay P. Singh et al. eds., 2018).
88. T. Douglas et al., *Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data*, 42 Eur. Psychiatry 134, 135 (2017); Seena Fazel et al., *Prediction of Violent Reoffending on Release from Prison: Derivation and External Validation of a Scalable Tool*, 3 Lancet Psychiatry 535, 535 (2016); *see also* James Hess & Susan Turner, *Accuracy of Risk Assessment in Corrections Population Management: Where's the Value Added?*, *in* Handbook on Risk and Need Assessment: Theory and Practice 93, 104 (Faye S. Taxman & Amy Dezember eds., 2016) (noting levels of accuracy considered acceptable for recidivist risk assessment are noticeably weaker than those expected and achieved in

field, the bar to validation is itself negligible. Too often, researchers conclude that a tool is validated even when it performs only slightly better than chance (i.e., the proverbial coin toss).[89]

Notwithstanding, the existence of a validation study regarding a particular tool's performance itself does not resolve the issue.[90] Unlike in other areas of scientific study, the development of risk tools in criminal justice have not relied upon independent, random samples. Instead, algorithms are often trained on convenience samples (i.e., the data was available).[91] This leads to selection bias.[92] When the algorithmic model is then tested on new samples in real-life settings, shrinkage in accuracy will result in the form of lower positive performance numbers and higher error rates.[93]

Moreover, validity statistics, whether from training samples or other external datasets, are not easily transferrable to new environs.[94] Accuracy rates may simply not replicate to other jurisdictions,[95] because of the existence of potentially risk-relevant differences in offenders, raters, or the availability of support services.[96] Variation in the particular legal environment, too, may influence how a risk assessment tool performs, such as peculiarities in its criminal laws, policing and prosecutorial policies and practices, judicial norms, and political sensitivities toward responses to crime.[97]

Stakeholders thereby should verify that a tool they intend to use is an appropriate fit to its environment. "Validation of risk prediction tools in different study populations than those used to develop the risk

---

other disciplines).

89. Hamilton et al., *supra* note 83, at 231.
90. Neal et al., *supra* note 24, at 137 ("Validation is an ongoing effort consisting of collecting, analyzing, and synthesizing various sources of evidence about how a particular tool performs in different sets of circumstances.").
91. Pari McGarraugh, *Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment is Appropriate at Sentencing and Inappropriate at Parole*, 97 Minn. L. Rev. 1079, 1096–97 (2013).
92. Michael R. Elliott & Richard Valliant, *Inference for Nonprobability Samples*, 32 Stat. Sci. 249, 252 (2017).
93. Melissa Hamilton, *Judicial Gatekeeping on Scientific Validity with Risk Assessment Tools*, 38 Behav. Sci. & L. 226, 239 (2020); *see* George Szmukler et al., *Risk Assessment and Receiver Operating Characteristic Curves*, 42 Psychol. Med. 895, 897 (2012); Stephen D. Gottfredson & Laura J. Moriarity, *Statistical Risk Assessment: Old Problems and New Applications*, 52 Crime & Delinq. 178, 185–86 (2006).
94. Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings*, 13 Psychol. Serv. 206, 207 (2016).
95. *See* Ehsan Bokhari & Lawrence Hubert, *The Lack of Cross-Validation Can Lead to Inflated Results and Spurious Conclusions: A Re-Analysis of the MacArthur Violence Risk Assessment Study*, 35 J. Classification 147, 168 (2018).
96. Desmarais et al., *supra* note 94.
97. *See* Michael F. Campagna et al., *The Contours of Assessment: Considering Aspects that Influence Prediction Performance*, Corrections Pol'y Prac. & Res., Nov. 2019, at 1, 4 https://www.tandfonline.com/doi/full/10.1080/23774657.2019.1690401 [https://perma.cc/Z3VK-UM4S].

prediction tool (e.g., different countries, different races, or subpopulations of patients with specific comorbidities) is necessary to determine the generalizability of the risk prediction tools."[98] This means that crossvalidating a tool in new population(s) in which it will be deployed is important because of the potential unrepresentativeness of prior validation samples.[99] For these reasons, the study offered herein uses samples from two jurisdictions considering the possibility/probability that results will not be identical.

In sum, although it is understandable that an agency may not have the resources, skills, or time to construct its own tool from scratch, it is important that an off-the-shelf tool be tested on its own population.[100] A completed validation study should also lead to discussion and debate among stakeholders about whether the observed error rates of the tool in its jurisdiction are tolerable.[101] The instrument may require retooling for the specific jurisdiction to improve performance metrics.[102] A technological advancement offered by the algorithmic method is that it can be retrained on new data and thus produce a better fit to the local jurisdiction.[103]

An additional important lens to apply is to review the tool's performance in the local jurisdiction by parsing *sub*group accuracy. Even if a tool demonstrates satisfactory overall accuracy in the local population, attention should be paid to whether accuracy rates vary across the sample's subpopulations.[104] "Differential validity" exists when test accuracy significantly differs between groups.[105] Experts thereby call for more independent research on differential validity as varying accuracy statistics by group undermines a tool's value and equity.[106] Most attention on potential group differences concentrates on race.

98. Cynthia S. Crowson et al., *Assessing Calibration of Prognostic Risk Scores*, 25 Stat. Methods Med. Res. 1692, 1692 (2016).
99. Gottfredson & Moriarty, *supra* note 93.
100. Desmarais & Lowder, *supra* note 69, at 7; Campagna et al., *supra* note 97; Hamilton et al., *supra* note 83, at 258; Matthew Fennessy & Matthew T. Huss, *Predicting Success in a Large Sample of Federal Pretrial Offenders: The Influence of Ethnicity*, 40 Crim. Just. & Behav. 40, 41 (2013).
101. *See* Charlotte Hopkinson, *Using* Daubert *to Evaluate Evidence-Based Sentencing*, 103 Cornell L. Rev. 723, 744 (2018).
102. Hamilton et al., *supra* note 83, at 258.
103. Reuben Binns, *Algorithmic Decision-making: A Guide For Lawyers*, 25 Jud. Rev. 2, 5 (2020).
104. Klingele, *supra* note 77, at 223; Binns, *supra* note 103, at 6 ("If any sub-set of the population differs from the majority in terms of the characteristics used to make a prediction, the model will likely make more errors in the predictions it makes about them.").
105. Christopher M. Berry et al., *Can Racial/Ethnic Subgroup Criterion-to-Test Standard Deviant Ratios Account for Conflicting Differential Validity and Differential Prediction Evidence for Cognitive Ability Tests?*, 87 J. Occupational & Organizational. Psychol. 208, 209 (2014).
106. Duwe & Rocque, *supra* note 21.

C. *Fairness to Racial Minorities*

Pretrial detention disproportionately ensnares Black citizens.[107] Black arrestees are more likely to be detained pretrial compared to similarly situated whites.[108] One reason is that money bail systems disproportionately harm Black individuals because of greater barriers of access to the financial resources necessary to successfully secure their release.[109] Another explanation is the application of race-based biases when humans drive pretrial outcomes.[110]

America has a long history of promoting stereotypes connecting criminality and violence with darkness of skin.[111] Despite their professional training, even judges use heuristics that can draw on racialized stereotypes of Black criminality.[112] Empirical works specifically find evidence of racial bias in pretrial decisions, with detriments typically imposed on Black defendants.[113] One particular study found that judges making pretrial bail decisions operated on unfounded stereotypes that exaggerated the danger of releasing Black defendants, and that this was the case with both white and Black judges.[114] In the pretrial context specifically, these types of race-based biases and stereotypes often go unchecked because judges there enjoy considerable discretionary authority with little accountability.[115]

In theory, entirely or partially replacing bail systems with risk assessment schemes will ameliorate the wealth-based imbalances in releases for Black defendants. Proponents also hope that risk assessment tools will reduce racial bias in the criminal justice system by swapping human intuition with algorithmic predictions.[116] Risk tool developers make such promises. "Vendors promote [algorithmic] models to the public and to the agencies that use them as the answer to human bias, arguing

107. *See* Wendy Sawyer, *How Race Impacts Who is Detained Pretrial*, Prison Policy Initiative (Oct. 9, 2019), https://www.prisonpolicy.org/blog/2019/10/09/pretrial_race [https://perma.cc/3BFZ-QAWM].
108. Stephen Demuth & Darrell Steffensmeier, *The Impact of Gender and Race-Ethnicity in the Pretrial Release Process*, 51 Soc. Probs. 222, 222 (2004).
109. Pretrial Just. Inst., Pretrial Risk Assessment can Produce Race-Neutral Results 2 (2017); Stevenson & Mayson, *supra* note 4, at 30; *see* Muhammad B. Sardar, *Give Me Liberty or Give Me . . . Alternatives?: Ending Cash Bail and its Impact on Pretrial Incarceration*, 84 Brook. L. Rev. 1421, 1423 (2019).
110. Doyle et al., *supra* note 37; *see* Emily Berman, *A Government of Laws and Not of Machines*, 98 B.U. L. Rev. 1277, 1327 (2018).
111. *See* Assefa, *supra* note 56, at 658.
112. Martinez et al., *supra* note 43, at 841.
113. Stevenson & Mayson, *supra* note 4, at 30; *see generally* Meghan Sacks et al., *Sentenced to Pretrial Detention: A Study of Bail Decisions and Outcomes*, 40 Am. J. Crim. Just. 661 (2015); Tina L. Freiburger et al., *The Impact of Race on the Pretrial Decision*, 35 Am. J. Crim. Just. 76 (2010).
114. Arnold et al., *supra* note 66, at 1889.
115. *See* Assefa, *supra* note 56, at 658.
116. Bavitz et al., *supra* note 8, at 20.

that computers cannot harbor personal animus or individual prejudice based on race."[117]

These premises—that mathematical algorithms can in reality work to eliminate bias and negate discrimination—are naive.[118] Stakeholders should be wary that the algorithmic risk turn will achieve their expected reduction in bias; indeed, without proper oversight, risk tools may instead boost detention rates for Black defendants. These warnings require some background to risk tool development practices to explain them.

The methods developers use in training their algorithms can produce race-based differentials in which accuracy rates may be substantially weaker for Black individuals. Risk assessment tools typically are trained on largely white male samples.[119] Consequently, these tools will incorporate risk factors that are more salient with whites than with other races or ethnicities,[120] thereby likely producing more accurate predictions for whites than other groups.[121] These circumstances thereby skew validation metrics as some studies have found a positive association between the percentage of whites in the sample tested and better overall results for the tool's predictive accuracy.[122] Yet overall accuracy rates may obscure differential performance. This problem remains hidden when external validation studies also are disproportionately composed of white

117. Eckhouse et al., *supra* note 75, at 186.

118. Noto La Diega, *supra* note 77, at 8; *see also* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671, 680 (2016) ("biased training data leads to discriminatory models").

119. Bobbie Ticknor & Jessica J. Warner, *Evaluating the Accuracy of SORNA: Testing for Classification Errors and Racial Bias*, 31 Crim. Just. Pol'y Rev. 3, 8 (2020); Thomas Cohen & Christopher Lowenkamp, *Revalidation of the Federal PTRA: Testing the PTRA for Predictive Biases*, 46 Crim. Just. & Behav. 234, 238 (2019); Seena Fazel & Stål Bjørkly, *Methodological Considerations in Risk Assessment Research*, *in* International Perspectives On Violence Risk Assessment 16, 18 (Jay P. Singh et al. eds., 2016).

120. Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 Psychol. Public Pol'y & L. 427, 429 (2016); *see also* Timnit Gebru, *Race and Gender*, *in* The Oxford Handbook of Ethics of AI 3 (Marcus D. Dubber et al. eds., 2020) (noting that the "myth of scientific objectivity" behind intelligence tests indicating differential intelligence between races and sexes obscures "that the IQ test in and of itself was designed by [W]hite men whose concept of 'smartness' or 'genius' was shaped, centered and evaluated on specific types of [W]hite men").

121. Jay P. Singh et al., *A Comparative Study of Violence Risk Assessment Tools: A Systematic Review and Metaregression Analysis of 68 Studies Involving 25,980 Participants*, 31 Clinical Psychol. Rev. 499, 501 (2011) (citing studies); Jay P. Singh & Seena Fazel, *Forensic Risk Assessment: A Metareview*, 37 Crim. Just. & Behav. 965, 978 (2010) (citing studies).

122. Whitney Bianca Threadcraft-Walker et al., *Gender, Race/Ethnicity and Prediction: Risk in Behavioral Assessment*, 54 J. Crim. Just. 12, 13 (2018) (citing studies); Howard Henderson et al., *Psychometric Racial and Ethnic Predictive Inequities*, 46 J. Black Stud. 462, 463 (2015) (citing studies).

samples.[123] If researchers fail to disclose accuracy rates by racial groupings, any evidence of differential validity is concealed. Thus, despite the aim of algorithmic risk tools to improve the accuracy of risk predictions overall, observers question whether risk tool performance is generalizable across racial groups such that accuracy rates (and error rates) are equivalent between them.[124]

Rigorous research on the presence of algorithmic fairness for racial minorities is new.[125] Interest is growing because, where officials are using algorithmic-based risk tools to inform decisions bearing significant consequences, it is advisable to conduct due diligence to study the specific impacts on minorities.[126] Of the few validation studies on pretrial tools focusing on racial disparities that exist, some evidence has shown that Black individuals are more likely to be ranked at higher risk by at least some of the tools, though this result is not consistently observed.[127] Any tool producing higher risk scores for minorities is problematic in potentially representing disparate impact.[128] One reason for the potential for increased risk scores is that algorithms often incorporate predictive factors that are not racially neutral.

1. Racialized Risk Factors

Contrary to popular expectations, risk tools that were developed using quite sophisticated methods, such as machine learning technologies, may still implicitly incorporate racialized data.[129] Algorithms that do not expressly include race as an explicit variable may output the same predictions as if they did by including one or more factors that are strongly correlated with race.[130] Algorithms that predict human behavior learn to exploit patterns in big data to tease out social categories—such as race—and track evidence associated with them.[131] When highly correlat-

123. *See* Shepherd & Lewis-Fernandez, *supra* note 120, at 428.
124. McCafferty, *supra* note 20, at 424.
125. Jennifer Skeem & Christopher Lowenkamp, *Using Algorithms to Address Trade-Offs Inherent in Predicting Recidivism*, 38 Behav. Sci. & L. 259, 260 (2020); Ticknor & Warner, *supra* note 119, at 7.
126. Bavitz et al., *supra* note 8, at 21; Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 Cumb. L. Rev. 67, 67 (2017); *see also* Chelsea Barabas, *Beyond Bias: Re-Imagining the Terms of "Ethical AI" in Criminal Law*, 12 Geo. J.L. Modern Critical Race Persp. 83, 96 (2020) ("accuracy has become a fetishized measure of a tool's worth" that is promoted as valuable despite potentially harming marginalized groups).
127. Threadcraft-Walker et al., *supra* note 122, at 13 (citing studies).
128. Skeem & Lowenkamp, *supra* note 125, at 262; Cohen & Lowenkamp, *supra* note 119, at 235.
129. Stevenson & Mayson, *supra* note 4, at 36–37.
130. Samuel Yeom & Michael Carl Tschantz, Avoiding Disparity Amplification under Different Worldviews 1 (2018); Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-making Processes*, 31 Phil. & Tech. 611, 616 (2018) ("excluded attributes can often be implicit in non-excluded ones").
131. Betsy Anne Williams et al., *How Algorithms Discriminate Based on Data They Lack*, 8 J. Info. Pol'y 78, 87 (2018).

ed, protected class membership is embedded into other data points in the form of "redundant codings."[132] In other words, racial categories become part of the final algorithms through proxy variables.[133]

An explanation for why correlated factors could result in Black individuals being assigned *higher* risk classifications is when an algorithm "bakes in" external disparities in treatment which negatively impact minorities.[134] Prediction acts like a mirror in which past behavior is expected to be repeated and thus foretells future behavior.[135] The circularity continues whereby algorithms then "generate futures that resemble the history that informs them."[136] When the algorithm incorporates factors representing preexisting biases produced by structural inequalities, without intervention, the algorithm will inevitably learn and replicate those biases.[137] The algorithm thereby recycles racial disparities as "bias in, bias out."[138]

A more focused explanation of the potential for disparate impact on Black individuals specifically is that risk assessment tools (like the PSA investigated herein) tend to heavily rely upon criminal history measures.[139] The fact that Black defendants are far more likely to have a lengthier criminal record is partly a manifestation of structural inequalities in American society. Black individuals bear significantly higher arrest rates and conviction rates than whites.[140] To illustrate, compared to 8 percent of American adults with felony convictions, almost one-third of adult Black men have felony convictions.[141] To confirm the connection, at least one recent study of a risk tool officially used to rate prisoners in the federal criminal justice system found a strong positive association between minority status and higher risk scores which was largely explained by the criminal history predictors used.[142]

It is possible for there to be differential involvement whereby minorities are more likely to engage in criminal activity.[143] But this does not necessarily render the use of criminal history in algorithms to be entirely

132. Barocas & Selbst, *supra* note 118, at 691.
133. *See generally* Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 Mich. L. Rev. 291 (2020).
134. Eckhouse et al., *supra* note 75, at 196.
135. Ticknor & Warner, *supra* note 119, at 9; Sandra G. Mayson, *Bias in, Bias out*, 128 Yale L.J. 2218, 2224 (2019).
136. Kaya Naomi Williams, *Public, Safety, Risk*, 44 Soc. Just. 36, 45 (2017).
137. Mayson, *supra* note 135, at 2284; Surden, *supra* note 75, at 3; Binns, *supra* note 103, at 6 ("systems trained on data that reflect an unequal and unjust society are likely to replicate and perpetuate it").
138. Berk, *supra* note 12, at 175.
139. *See* Stevenson & Doleac, *supra* note 74, at 1.
140. *See* The Sentencing Project, Report to United Nations on Racial Disparities in the U.S. Criminal Justice System 1 (2018) (citing studies).
141. Murakawa, *supra* note 20, at 476 (citing study).
142. Skeem & Lowenkamp, *supra* note 125, at 268.
143. Riccardo Fogliato et al., *Fairness Evaluation in Presence of Biased Noisy Labels*, 23 Proceedings of Machine Learning Research 1, 1 (2020).

fair.[144] Criminal activity may be more reflective of limited opportunities rather than entirely individual choices: structural inequalities (economic, social, political) may push minorities toward antisocial behavior.[145]

Nonetheless, it is perhaps more pertinent that the algorithms tend to use biased measures of criminal offending that disproportionately identify Black individuals who may not actually have committed crimes and/or who may not commit crimes at rates higher than whites.[146] When the outcome of interest involves *crime*, the measurement of this failure event is innately biased.[147] There is simply no theoretical or practical way to measure crime *per se*. Formal statistics, such as number of arrests and convictions, do not equate to actual crime rates. Formal records undercount crime and overcount crime at the same time. Not all crimes are reported in the first place. Even convictions may be factually erroneous. The alternatives of informal crime calculations, such as through self-reports and victimization surveys, also are profoundly prone to error.

As crime itself cannot be accurately measured, tool developers must resort to using proxies. By definition, any proxy is a fundamentally inaccurate measure of the phenomenon it is trying to represent.[148] Moreover, this does not mean a proxy will produce similar errors across groups in that it will under count or over count at the same rates. For example, the fact that Black individuals may on average have a more serious criminal history does not necessarily mean a higher true crime

144. Donnelly & MacDonald, *supra* note 56, at 786 ("The cumulative disadvantage framework offers an alternative, systems-level view of the criminal case processing of defendants. Cumulative disadvantage refers to a process of intensifying inequality among individuals that grows over time through negative interactions with the criminal justice system.").

145. Vincent Chiao, *Fairness, Accountability and Transparency: Notes on Algorithmic Decision-Making in Criminal Justice*, 15 Int'l J.L. Context 126, 129 (2019); Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 Criminology 680, 685 (2016) ("Race reflects longstanding patterns of social and economic inequality in the United States (e.g., differences in social networks/resources, neighborhoods, education, and employment). Although poverty and inequality do not inevitably lead to crime, they 'involve circumstances that do contribute to criminal behavior.'") (citing Samuel Walker et al., The Color of Justice: Race, Ethnicity, and Crime in America 99 (2011)).

146. Regarding practices which criminalize Black life:

> In the field of criminology, the interpretation of crime data has always served as a critical point of departure between positivist subfields of the discipline—seeking to measure and manage criminal behavior in 'risky' populations—and critical scholars, who conceive of crime as primarily the by-product of criminalizing discourses and practices carried out by the carceral state.

Barabas, *supra* note 126, at 86 (2020).

147. Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, Stanford Computational Policy Lab 1, 3 (2018).

148. The Law Society Commission On The Use Of Algorithms In The Justice System, Algorithms in the Criminal Justice System 18 (2019).

rate than whites; it may simply reflect a heightened scrutiny of them by police and prosecutors.[149] For example, police in large urban areas have often focused attention on open-air drug markets, which are skewed towards the method of drug dealing by Black individuals, rather than residence-based dealing preferred by whites.[150]

How the tool counts crime in either its criminal history predictors or its forecasted failure event (e.g., counting recidivism) is of consequence. Developers tend to prefer arrest records as a chosen proxy because they are more readily available and require shorter followup periods to track than convictions.[151] Hence, the choice of proxy may be reflective of easier and quicker access to arrest records. An arrest is assuredly a biased estimate of crime as "one of the least procedurally protected instances of contact with the criminal justice system."[152] The low evidentiary bar of probable cause renders arrest outcomes highly unreliable and problematic in itself.

Notably, the use of arrests as the proxy to crime increases the likelihood of a tool producing disparate impact for minorities. Arrest data undeniably evidence the discretionary behaviors of the police. "Officers use discretion in enforcement decisions (e.g., deciding whom to stop, search, question, and arrest) just as police officers and prosecutors use discretion in charging (e.g., simple assault vs. felonious assault). The underlying data reflect[] these judgement calls."[153]

More specifically, the significant weight risk assessment algorithms tend to place on prior arrests means that the tools fundamentally are more about replicating *police actions* or forecasting *police responses* than they are predicting true criminal behavior, and this focus rests disproportionately negative consequences on minorities.[154] As an illustration, a person living in an overpoliced, minority neighborhood is likely to attract a greater number of arrests and convictions and thus receive a higher risk score than a person who resides in a less policed area with a more white population, regardless of their actual criminal behavior.[155] The training data will thereby learn on what amounts to overpolicing practices in minority neighborhoods and underpolicing in mainly white, upper class areas.[156] As a result, the algorithm may overestimate the offending risk

149. Eaglin, *supra* note 35, at 95.
150. Chiao, *supra* note 145, at 127.
151. Eaglin, *supra* note 35, at 78.
152. *Id.* at 94.
153. Exec. Off. of the President, Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights 22 (2016).
154. Binns, *supra* note 103, at 4–5 (noting risk tools predict policing rather than crime); Ugwudike, *supra* note 17, at 22–23 (conceptualizing reliance on arrests is not necessarily indicative of criminal offending but of racially tainted interactions with criminal justice officials).
155. Megan Stevenson, *Assessing Risk Assessment in Action*, 103 Minn. L. Rev. 303, 328 (2018).
156. Robert Werth, *Risk and Punishment: The Recent History and Uncertain Future of Actuarial, Algorithmic, and "Evidence-Based" Penal Techniques*, 13 Sociol.

of minorities while at the same time underestimate the risk of whites.[157] Hence, instead of representing differential involvement in crime, the arrest proxy is picking up differential selection by policing agencies, thereby introducing target variable bias.[158]

In sum, algorithms that train on racially skewed policing data will reflect those biases and the outputs will as a result be racially skewed, likely leading to further discriminatory actions.[159] In the foregoing ways the "objective" guise of the algorithms thereby creates a pernicious feedback loop[160] and effectively "launders" racial inequalities.[161]

2. The Trope of the Race-Free Tool

The foregoing is not to suggest that race-based impacts are intended by tool developers.[162] But adverse effects are tied to the outcomes of a practice, without requiring discriminatory motive or intent.[163] The lack of attention to potential disproportionality may be because data scientists are not often trained in civil rights law.[164] At the same time, constitutional law experts may be dissuaded from investigating algorithms for racial undertones because of a lack of statistical skills.[165]

Developers may claim (and often do) that their tools are taint-free by not incorporating race as an express predictive factor.[166] Yet such a

---

Compass. 1, 9 (2019). For example, local law enforcement may focus on street crimes committed in low-income communities while ignoring more sheltered criminal activity occurring in affluent white neighborhoods. Barabas, *supra* note 126.

157. Corbett-Davies & Goel, *supra* note 147, at 18.
158. Fogliato et al., *supra* note 143 (noting other disciplines using the terms differential outcome measurement bias or differential outcome misclassification bias).
159. Hayes et al., *supra* note 10; Martinez et al., *supra* note 43, at 841–42.
160. Marius Miron et al., *Evaluating Causes of Algorithmic Bias in Juvenile Criminal Recidivism*, Artificial Intelligence & L. 1, 5 (2020) ("feedback loops which reinforce previous biases . . . occur in environments characterized by complex interactions such as policing systems"); *see also* Hayes et al., *supra* note 10 at 536; Prins & Reich, *supra* note 64, at 259 (expressing concern that algorithms "will likely only reproduce, or may even exacerbate, the injustices of contemporary criminal justice policy under a more 'objective' guise"); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. Penn. L. Rev. 633, 680 (2012) (noting algorithms conceal discrimination while entrenching bias into the future).
161. Murakawa, *supra* note 20, at 480.
162. Surden, *supra* note 75.
163. MacCarthy, *supra* note 126, at 80.
164. *See* Ugwudike, *supra* note 17 at 6 ("digital prediction technologie . . . , which make no direct reference to race, are blindly endorsed as unbiased and compliant with race equality laws. Potential conducts of racial bias are thus ignored" but scholars writing from the field of law point out racial equality laws are not able to remedy all structural productions of racial inequalities); Gebru, *supra* note 120, at 2 ("the predominant thought that scientists are 'objective' clouds them from being self-critical and analyzing what predominant discriminatory view of the day they could be encoding").
165. Barocas & Selbst, *supra* note 118, at 6.
166. Kroll et al., *supra* note 160, at 685 (conceptualizing algorithms that do not expressly include race is simply race-blind and naive as other factors can include

simplistic assertion presents as "fairness through blindness."[167] A critic has surmised that such a denial of race effects is a conjured form of racial innocence as these tools are developed on datasets in which systemic racism has already been baked.[168] The trope of race neutrality is thereby "mythical" as structural inequalities and societal histories of racial discrimination that embed into the risk tools remain unaddressed.[169]

For the foregoing reasons, the assumption that algorithmic tools will virtually ameliorate harms from biased human predictions is myopic. In fact, it is possible that algorithmic risk itself produces greater harms. While human bias is acted out on a case-by-case basis, an algorithm's efficiency means it can discriminate on a more systematic basis and on a larger scale.[170]

These concerns about racial disparities have spawned a new literature on how to identify group fairness via algorithmic devices.[171] A host of definitions have been offered by academics deriving from various disciplines, such as law, machine learning engineers, forensic scientists, and criminologists. Many of the algorithmic group definitions will be introduced in the analytical strategy below.

The promises and perils of the algorithmic risk reform movement in ameliorating the harms of pretrial detention and money bail have divided groups otherwise similarly aligned in their interests in protecting criminal defendants' civil rights. The Pretrial Justice Institute, originally a strong advocate for risk assessment, surprised reformers when it recently announced its shift to an opposition stance upon realizing the tools deepened structural inequalities.[172] Human Rights Watch has urged abandoning the algorith-

implicit bias). Indeed, the owner of the PSA, the tool studied in this Article, makes such an assertion, proclaiming that the PSA does not promote racial bias because it does not include race as a predictor. Arnold Ventures, Public Safety Assessment Faqs (PSA 101) 2 (2019).

167. MacCarthy, *supra* note 126, at 89–90; Ugwudike, *supra* note 17, at 8 (alternatively referring to "bias elimination fallacy").

168. Murakawa, *supra* note 20, at 480; Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 U.C. Davis L. Rev. 1067, 1075 (2018) (warning algorithms appear neutral and thus mask that they may be based on racially biased policing data); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. Rev. 54, 58 (2018) ("The idea that algorithmic decisionmaking, like laws, are objective and neutral obscures a complex situation. It refuses to grapple with the causes and effects of systematic and structural inequality, and thus risks missing how AI can have disparate impacts on particular groups.").

169. Ugwudike, *supra* note 17.

170. Noto La Diega, supra note 77, at 8; Indrė Žliobaitė, *Measuring Discrimination in Algorithmic Decision Making*, 31 Data Mining & Knowledge Discovery 1060, 1063 (2017).

171. *See, e.g.*, Richard Berk et al., *Fairness in Criminal Justice Settings: The State of the Art*, 50 Soc. Methods & Res. 3 (2021).

172. Pretrial Justice Inst., Updated Position on Pretrial Risk Assessment Tools 1 (2020). News outlets noticed the change in position. Dawn R. Wolfe, *Criminal Justice Group Drops Support for Pretrial Risk Assessment Tools as Ohio Justices Seek to Block Their Use*, The Appeal (Feb.12, 2020), https://theappeal.

mic risk platform because of the potential for racial bias.[173] The influential Leadership Conference on Civil and Human Rights formally announced it (with over 100 civil rights, social justice, and digital rights groups signing on) is against pretrial risk assessment tools due to reliance upon biased data that will serve to perpetuate disadvantage upon minorities.[174]

On the other hand, a different coalition of prominent criminal defense groups (e.g., National Association of Criminal Defense Lawyers, American Council of Chief Defenders) advocate the use of algorithmic tools in pretrial, though in an updated statement is more nuanced: "[We] support the use of a validated pretrial risk assessment as a component of a fair pretrial release system, in any jurisdiction where it is evident . . . that it will serve to reduce unnecessary detention and help to eliminate racial and ethnic bias in the outcome of the pretrial decisions."[175]

Empirical research into the potential for race-based variances with risk tool predictions is in its relative infancy.[176] Tracking outcomes to determine how well a risk tool labels individuals is important to understanding whether decisions based on tool outcomes are equitable across races in real world settings.[177] The promises of algorithmic risk tools are undermined if there is evidence of disparate impact, if the test appears biased, and/or if the predictive validity meaningfully varies across racial groups.[178] Consequently, the issue of algorithmic fairness remains unsettled and stakeholders are thus calling for independent audits of risk tools to evaluate their performances with respect to racial issues and

org/criminal-justice-group-drops-support-for-pretrial-risk-assessment-tools-as-ohio-justices-seek-to-block-their-use [https://perma.cc/3X7A-4KZ4]; Tom Simonite, *Algorithms Were Supposed to Fix the Bail System. They Haven't*, Wired (Feb. 19, 2020, 8:00 AM), https://www.wired.com/story/algorithms-supposed-fix-bail-system-they-havent [https://perma.cc/ZMG8-CTMU].

173. Marsha Slough et al., *Human Rights Watch Statement to the California Judicial Council Pretrial Reform and Operations Workgroup*, Hum. Rts. Watch (Feb. 11, 2020), https://www.hrw.org/news/2020/02/11/human-rights-watch-statement-california-judicial-council-pretrial-reform-and [https://perma.cc/MB7X-DKCH] ("Replacing money bail with risk assessment will not address the problems of unnecessary pretrial incarceration, will not improve the racial and class discrimination of the system, and will not result in fairer outcomes. A court system committed to justice should not use them.").

174. *See generally*, The Leadership Conference Educ. Fund, The Use of Pretrial "Risk Assessment" Instruments: A Shared Statement of Civil Rights Concerns (2018).

175. Gideon's Promise et al., Joint Statement: Pretrial Risk Assessment Instruments 2 (2019).

176. Kevin R. Reitz, *The Compelling Case for Low-Violence-Risk Preclusion in American Prison Policy*, 38 Behav. Sci. & L. 207, 214 (2020) (citing studies); Doaa Abu Elyounes, *Bail or Jail? Judicial Versus Algorithmic Decision-Making in the Pretrial System*, Colum. Sci. Tech. L. Rev. (2020); Fennessy & Huss, *supra* note 100, at 41 (citing studies).

177. Koepke & Robinson, *supra* note 49, at 1796.

178. Taxman, *supra* note 21, at 277 ("Given recent concerns raised in the U.S. regarding the potential racial bias of justice policies and practices, the ability to demonstrate that the tools are racially neutral is critical.").

thereby work toward developing a stronger knowledge base.[179] Indeed, in the most prominent case to date addressing the role of risk assessment in criminal justice,[180] the Wisconsin Supreme Court in *State v. Loomis* expressly warned of a potential legal downside if a tool were to disproportionately classify minorities.[181]

II. Background to a Study of Algorithmic Fairness

The present study is of the Public Safety Assessment (PSA), a popular algorithmic risk instrument designed specifically for use by judges in making pretrial release decisions. We use datasets from two jurisdictions to examine the utility and fairness of the PSA across sites and races. Three research questions orient the study, with the empirical methodologies designed to address them. Blended into the methodology Part are certain terms from the blossoming academic literature developing algorithmic fairness nomenclature (e.g., demographic parity, equalized odds, error rate balance).

A. *Public Safety Assessment*

The nonprofit Laura and John Arnold Foundation funded the development of the PSA with an intent to design an efficient, evidence-based instrument to assist in a pretrial context:

> From the beginning, we believed that an easy-to-use, data-driven risk assessment could greatly assist judges in determining whether to release or detain defendants who appear before them. And that this could be transformative. In particular, we believed that switching from a system based solely on instinct and experience to one in which judges have access to scientific, objective risk assessment tools could further our central goals of increasing public safety, reducing crime, and making the most effective, fair, and efficient use of public resources.[182]

The PSA's developmental sample drew on data from approximately 750,000 pretrial defendants released from 300 jurisdictions.[183] The PSA is purported to be a "universal tool,"[184] meaning it is intended to be used in

179. Gideon's Promise et al., *supra* note 175; Duwe, *supra* note 82, at 548; Bavitz, *supra* note 8; McCafferty, *supra* note 20, at 425.

180. Brandon L. Garrett & Megan Stevenson, *Open Risk Assessment*, 38 Behav. Sci. & L. 279, 280 (2020).

181. State v. Loomis, 881 N.W.2d 749, 763–64 (Wis. 2016), *cert. denied* 137 S.Ct. 2290 (2017); *see also* Klingele, *supra* note 77, at 223 ("Despite calls from researchers for jurisdictions to periodically re-norm their tools for changing populations and sub-populations, few jurisdictions have done so, which means that in those places where the tool has not been tested on local populations, it may lack validity.").

182. Arnold Found., Developing a National Model for Pretrial Risk Assessment 2 (2013).

183. Advancing Pretrial Policy and Research, About the Public Safety Assessment, https://advancingpretrial.org/psa/about [https://perma.cc/4AEA-SNV].

184. Jessica Reichart & Alysson Gatens, *An Examination of Illinois and National Pretrial Practices, Detention, and Reform Efforts*, Ill. Crim. Just. Info. Auth. Res. Hub 6 (2018), https://icjia.illinois.gov/researchhub/files/

pretrial jurisdictions across the country.[185] The PSA is the most commonly used pretrial risk assessment tool in the United States,[186] having been adopted in four states and many major metropolitan areas.[187] According to a recent report, of the jail systems using a pretrial risk tool, almost one-third employed the PSA.[188] The PSA is actuarial in nature, reliant solely upon static factors, and designed to be scored based on available criminal history records without requiring an interview with the individual scored.[189]

The PSA offers a scale to predict any new criminal arrest. This scale represents a "broad-band" instrument in that it predicts general recidivism rather than any specific type of recidivist activity.[190] The new criminal arrest instrument contains seven factors: age at current arrest, pending charge at time of offense, prior misdemeanor conviction, prior felony conviction, prior violent conviction, prior failure to appear in the prior two years, and prior sentence to incarceration.[191] The PSA produces raw scores of one to thirteen, which are converted into final scores ranging from one to six points, with more points indicating a higher likelihood of arrest.

Despite calls for risk tool developers to make training data available for independent researchers to audit,[192] the PSA developers have not released their developmental datasets. The PSA's owner has reported that in its initial validation studies, defendants recidivated at similar rates across races and that the PSA did not overclassify minority risk levels.[193] However, this is a summary assertion, and to date, no other verifying information has been offered. Experts advise that skepticism is appropriate when proponents make claims about a tool's performance without providing proof to support such assertions.[194] Only one validation study of the PSA has been publicly released. It comes from the state of Kentucky, but it provides only a constricted perspective on tool performance measures. The study is still valuable in that it provides the

Pretrial_Article_060718-191011T20093352.pdf [https://perma.cc/28ZU-CF92].

185. DeMichele et al., *supra* note 69, at 6.
186. Pretrial Just. Inst., *supra* note 109.
187. Coglianese & Ben Dor, *supra* note 63.
188. Scan of Pretrial Practices Survey, Pretrial Just. Inst. (2019), https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=24bb2bc4-84ed-7324-929c-d0637db43c9a&forceDialog=1 [https://perma.cc/GZT8-2XDF].
189. Matthew DeMichele et al., *What do Criminal Justice Professionals Think About Risk Assessment At Pretrial?*, 83 Fed. Probat. 32, 35 (2019).
190. Duwe, *supra* note 82, at 539.
191. DeMichele et al., *supra* note 69, at 415.
192. Stephanie Wykstra, *What is "Fair"?: Algorithms in Criminal Justice*, 34 Issues Sci. & Tech. 21, 23 (2018); The Leadership Conference Educ. Fund, *supra* note, 174 at 3; Partnership on AI, *supra* note 86, at 31 (2019).
193. Arnold Found., *supra* note 182, at 5.
194. Kroll et al., *supra* note 160, at 683 (denoting assertions about good performance, without offering supporting facts, "just words on paper").

source of one of the datasets used herein, from which more enlightening measures of performance were derived and are presented.

B. *Research Aims*

This study was designed to investigate the PSA new criminal arrest scale's performance ability with a concentration on investigating racial disparity. The research questions are these:

(1) Does the PSA result in adverse impact on Black defendants through differences in the rates of higher risk classifications predicting new criminal arrests?

(2) Does the PSA new criminal arrest scale exhibit test bias by race?

(3) Does the PSA new criminal arrest scale exhibit differential validity by race? Inherent in this question is to elicit the tool's predictive abilities and then to compare those statistics by race. Accordingly, the study will be able to provide important information on the tool's accuracy overall, thus offering benefits to the knowledge base on risk tools even outside the context of racial issues.

These research questions are addressed through an analytical strategy that incorporates relevant empirical methods, legal standards, and algorithmic fairness definitions. The strategy is revealed after introducing the underlying datasets.

C. *Datasets*

Our study uses archival datasets from two jurisdictions that systematically use the PSA to inform pretrial release decisions. From the archival datasets, we excluded defendants who were not identified as white or Black as those two racial groupings are of interest in this study.

1. Illinois Dataset

The first sample is from a large jurisdiction in Illinois. The state recently embraced a pretrial reform platform. The Illinois Bail Reform Act of 2017 reduces reliance upon money bail because "decision-making behind pre-trial release shall not focus on a person's wealth and ability to afford monetary bail but shall instead focus on a person's threat to public safety or risk of failure to appear before a court of appropriate jurisdiction."[195] The Act encourages counties to adopt an evidence-based risk assessment tool to aid in the effort to estimate an individual's likelihood of dangerousness and failure to appear.[196]

This dataset is from Cook County. Cook County is home to the city of Chicago and stands as the second largest county by population size in the United States, with an estimate in 2019 of more than 5 million residents.[197] The Cook County Jail is one of the largest, single-site jails in the

195. 2017 Ill. Laws, *available at*: https://www.ilga.gov/legislation/publicacts/100/100-0001.htm [https://perma.cc/5P6G-K2UW].

196. 725 Ill. Comp. Stat. 5/110–5(f) (2018) (referencing a risk assessment evaluation of a pretrial defendant using a recognized evidence-based instrument).

197. Annual Estimates of the Resident Population for Counties in Illinois: April 1,

country,[198] with annual expenditures of over $330 million.[199] Its central bond court processes over 33,500 pretrial cases annually.[200]

Cook County is now regarded as "leading the way in pretrial justice reforms."[201] Pressured by pending civil rights litigation and grassroots community activism to adopt progressive measures to ameliorate the harms of pretrial detention,[202] the chief judge of the Cook County Circuit Court in 2017 issued an order mandating judges use risk assessment tool results to inform decisions on release.[203] The Coalition to End Money Bond, a community advocacy group, heralded the order as signaling the "turning point" in the battle to reform unfair pretrial detention outcomes.[204] Cook County had at the time of the chief judge's order already been preparing to use predictive technology. In March 2016, Cook County authorities adopted the PSA as its foundational risk assessment tool for use on its pretrial population.[205]

Pretrial services staff score the PSA and provide results to the judge at the defendant's bond hearing.[206] In Cook County, pretrial detention decisions are speedy, lasting on average 100 seconds.[207]

This archive was made publicly accessible on the government's website as a big data trove of electronic spreadsheets. The purpose behind the disclosure was to support a document by county officials concluding that a boost in release rates in conjunction with the adoption of the PSA did not result in increased arrest rates.[208] This governmental document did not reference or discuss any performance metrics or racial fairness issues.

2010 to July 1, 2019, U.S. Census Bureau, Population Div., (Mar. 2020), https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html [https://perma.cc/BZ5P-A6BR] (follow "Illinois" hyperlink under "Annual Estimates of Resident Population for Counties: April 1, 2010 to July 1, 2019" then open downloaded excel spreadsheet).

198. Assefa, *supra* note 56, at 655.

199. Coal. to End Money Bond, Monitoring Cook County's Central Bond Court: A Community Courtwatching Initiative, August–October 2017 10–11 (2018).

200. *Id.* at 15.

201. Coal. to End Money Bond, Money Bail is Ransom: End Pretrial Incarceration 3 (2019).

202. Van Brunt & Bowman, *supra* note 6, at 762; Coal. to End Money Bond, *supra* note 199, at 4–5.

203. Doyle et al., *supra* note 37, at 52 (referencing Office of the Chief Judge, Circuit Court of Cook Cty, General Order No. 18.8a—Procedures for Bail Hearings and Pretrial Release (2017)).

204. Coal. to End Money Bond, *supra* note 199, at 6.

205. Doyle et al., *supra* note 37, at 56.

206. Assefa, *supra* note 56, at 662.

207. *Id.* A court watch program of civilians reported that many cases lasted less than thirty seconds and were demeaning to the accused. Coal. to End Money Bond, *supra* note 199, at 50.

208. State of Ill. Circuit Court of Cook Cty., Bail Reform in Cook County: An Examination of General Order 18.8A and Bail in Felony Cases 1 (May 2019) [https://perma.cc/2QAN-64QU].

The full sample size is of 48,318 offenders booked into the Cook County jail and scored on the PSA between July 1, 2016 and December 31, 2018, whether released or not. The racial makeup of the full sample is asymmetrical with 8,041 (17 percent) white individuals and 40,277 (83 percent) Black individuals. The full sample is used only in response to the first research question about disparate impact.

The selected use of the full sample is because the main analyses are focused on only released offenders. Thus, we created a subset of 36,298 offenders with PSA new criminal arrest predictions who were discharged during the time of study. The racial makeup of this subset of released defendants is also unbalanced with 6,569 (18 percent) white individuals and 29,727 (82 percent) Black individuals.

2. Kentucky Dataset

The second site includes pretrial defendants across the entire state of Kentucky. The United States census estimates in 2019 over 4 million residents lived in the state.[209] Observers hail Kentucky as exemplifying a "successful implementation of evidence-based pretrial assessments."[210] Kentucky was an early adopter of algorithmic risk assessment generally and its risk-informed decisionmaking scheme has become a model for other jurisdictions across the country.[211] In 2011, a new state law mandated the use of an evidence-based risk tool to inform decisions on requiring pretrial bonds.[212] Pretrial services in Kentucky replaced its local risk tool with the PSA in mid-2013.[213]

Kentucky was the first jurisdiction to formally adopt the PSA in its pretrial decision practices.[214] PSA researchers worked alongside Kentucky officials in the early stages and made what they found to be appropriates adjustments to the PSA.[215] Kentucky's single pretrial services system operates on a statewide basis and is responsible for scoring the PSA on jail entrants within 12 hours of entry.[216] Pretrial services process over 100,000 bookings annually.[217]

209. Annual Estimates of the Resident Population for Counties in Illinois: April 1, 2010 to July 1, 2019, U.S. Census Bureau, Population Div., (Mar. 2020), https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html [https://perma.cc/BZ5P-A6BR] (follow "Kentucky" hyperlink under "Annual Estimates of Resident Population for Counties: April 1, 2010 to July 1, 2019" then open downloaded excel spreadsheet).
210. Arthur W. Pepin, Evidence-Based Pretrial Release 8–9 (2012).
211. Stevenson, *supra* note 155, at 342–43.
212. Ky. Rev. Stat. Ann. § 431.066(2) (West 2012).
213. Stevenson, *supra* note 155, at 344.
214. DeMichele et al., *supra* note 69, at 414.
215. *Id.*
216. Pepin, *supra* note 210, at 8.
217. Arnold Found., Results from the First Six Months of the Public Safety Assessment—Court in Kentucky 5 (2014), https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=42196c1e-c574-0af2-dcd5-0420aa7c5e8f&forceDialog=0 [https://perma.cc/H959-MC58].

The statistics to populate this sample were extracted from data provided in the first released validation study of the PSA by DeMichele et al.[218] The DeMichele publication provided some information on the discrimination accuracy of the PSA arrest scale overall and then comparatively for white and Black arrestees. Specifically, these researchers noted the Area Under the Curve (AUC, discussed further below) across the PSA scores, rates of arrest by PSA score, and regressions for performance by race. The study reported herein differs in that instead of using the PSA score range, it uses PSA risk bins to compare to the Illinois dataset, and offers additional statistical measures to more holistically evaluate the tool's performance in Kentucky concerning its accuracy and comparative application based on race.

This sample includes 161,330 adult offenders who were booked into jail in Kentucky between July 1, 2013 and December 30, 2014, scored on the PSA, and released pending trial. In contrast to the urbanized Illinois sample, Kentucky is a mostly rural jurisdiction and with a smaller Black population ratio compared to the national average.[219] The racial makeup of this dataset is 133,647 (83 percent) white individuals and 27,683 (17 percent) Black individuals.

For purposes of the question on disparate impact, while we had the full dataset for Illinois of released and nonreleased inmates, we did not have access to the PSA predictions in Kentucky for nonreleased defendants. Thus, we could not create a similar dataset for Kentucky, but this gap is not problematic in context, as discussed herein.

D. *Analytical Strategy*

The data were imported into Excel and SPSS version 25.0 for analyses. Each dataset was then split into two, one representing whites and the other focused on Black individuals, yielding these subsets: (1) Illinois (a) white and (b) Black; then (2) Kentucky (a) white and (b) Black.

Risk tools often stratify individuals into ordinal risk bins,[220] and decisionmakers typically prefer a three-risk bin allocation (e.g., low, medium, high).[221] Consequently, it is constructive to examine the utility of the preferred risk binning strategy. Adopting Illinois officials' practice in this regard, we employed a three-risk bin scheme for PSA's new criminal arrest scores of low (1–2 points), medium (3–4 points), and high (5–6 points). Based on calculations of the numbers of defendants in each risk bin and their corresponding new arrest rates, we computed measures

218. As a result of missing data for the variables used in the analyses herein, the current sample size is 161,173 versus 161,330 in the original set, a loss of a small 157 subjects.

219. DeMichele et al., *supra* note 69, at 17–18.

220. Jeremy Luallen et al., *The Predictive Validity of the Post-Conviction Risk Assessment Among Federal Offenders*, 43 Crim. Just. & Behav. 1173, 1174 (2016).

221. Ashley B. Batastini et al., *Communicating Violence Risk During Testimony: Do Different Formats Lead to Different Perceptions Among Jurors?*, 25 Psychol. Pub. Pol'y & L. 92, 93 (2019).

to inform on the performance abilities of the PSA's new criminal arrest scale to respond to the research questions.

A burgeoning literature in the legal and data sciences fields is offering statistical measures to judge various aspects of algorithmic fairness in terms of treatment and performance across demographic groups.[222] We engage with several of them in this study as incorporated into this methodology section.

1. Disparate Impact

Research question one inquires about the presence of disparate impact. The issue of disparate impact is discussed in the algorithmic fairness literature with the terminology of "statistical parity" and, alternatively, "demographic parity" or "equal acceptance rates."[223] Statistical parity requires that a tool assigns individuals to any risk bin at equal rates for all demographic groups.[224] A lack of statistical parity is generally indicative of disparate impact.[225] Disparate impact does not require discriminatory motivation or intent[226] and thus applies to a facially neutral policy or practice,[227] such as the PSA.

The statistical parity definition from the algorithmic fairness field holds that any difference in rates of assignment suggests disparate impact, though without specifying any particular threshold of difference. Among potentially useful criteria, three methods to evaluate the existence of disparate impact appear relevant to the task. Statistical/demographic parity can be observed by calculating and comparing the rates of assignment to any risk levels (low, medium, high) by racial groups to determine whether the discrepancies appear important from a practical perspective. This is still, though, a judgment call as to what gap may be practically meaningful. The second method is to investigate whether the discrepancies rise to the level of statistical significance. Here, the null hypothesis is that the proportions are equal between races and is judged by a z-test.[228] For this study, we apply a conservative alpha level of .001

222. Hamilton, *supra* note 26, at 266; Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 Sociol. Methods Res. 3, 16 (2021).
223. Dana Pessach & Erez Schmueli, *Algorithmic Fairness*, 4 (2020).
224. Shira Mitchell et al., *Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions*, 8 Ann. Rev. Statistics & Application 1, 8 (2019); James E. Johndrow & Kristian Lum, *An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction*, 13 Annals Applied Stat. 189, 214 (2019).
225. Yeom & Tschantz, *supra* note 130, at 1; Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 Criminology 680, 685 (2016).
226. Mark MacCarthy, *supra* note 126, at 80.
227. Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 Ala. L. Rev. 519, 554 (2018).
228. The square of the z-test statistic for the equivalence between independent proportions is identical to the chi-square statistic χ^2. A Martín Andrés & I. Herranz Tejedor, *The Equivalence of Two Proportions Revisited*, 31 J. Applied Stat. 61, 63 (2004).

to reject the null hypothesis. This statistical test is complementary of the first method because, with large sample sizes, statistical significance may not equate with practical significance (or with small sample sizes, there may be meaningful differences in proportions that are not statistically significant).[229]

The third method for determining if there are meaningful differences in classification by race borrows a statistical baseline from employment law's adverse impact test. Adverse impact is suggested when a specific policy, practice, or procedure results in the selection for some advantage of members of one group at a significantly greater rate than members of a disadvantaged group.[230] Employment law's numerical rule of thumb refers to a four-fifths standard: A procedure produces a disproportionate adverse impact on a protected group if its members are selected at a rate less than eighty percent of the rate for a nonprotected group's members.[231]

As applied here, the relevant procedure is based on each jurisdiction's official policy to adopt the PSA to actively inform decisions on pretrial detention. The selection at a low-risk level offers the potential benefit of increasing an individual's chance of release. Thus, evidence of adverse impact is indicated if Black individuals are assigned a low-risk classification level at a rate less than 80 percent of the rate for whites. Conversely, a high-risk label means the individual is likely subject to more severe penal consequences, and thereby avoiding the high-risk classification constitutes a benefit.[232] Thus, we flip the 80 percent rule accordingly. This alternative perspective suggests that adverse impact exists if whites are classified into the high-risk bin at a rate less than 80 percent of the rate of Black individuals.

This question is the single time that we use the full dataset in Illinois. We first calculated the 80 percent threshold in the samples of *released* defendants. However, as judges have discretionary authority to release (or not) for reasons that are not entirely dependent upon PSA predictions, the results for released defendants may not truly represent whether the PSA's algorithm itself results in adverse impact. Hence, we took the opportunity to also compute disparate impact statistics for the full Illinois sample of assessed defendants, released or not. We did not have access to the Kentucky sample of nonreleased defendants and, thus, were not able to do the same there.

2. Test Bias

The second research question enquires about the existence of test bias. Test bias in psychometric instruments exists if there are systematic errors in test outcomes based on group membership.[233] Bias in this sense

229. Kevin Tobia, Note, *Disparate Statistics*, 126 Yale L.J. 2382, 2406 (2017).
230. Wayne F. Cascio & Herman Aguinis, Applied Psychology in Human Resources Management 169 (1987).
231. MacCarthy *supra* note 126, at 67; Barocas & Selbst, *supra* note 118, at 701.
232. Ugwudike, *supra* note 17, at 491.
233. Cecil R. Reynolds & Lisa Suzuki, *Bias in Psychological Assessment: An Empirical*

is distinct from human biases that may result from cognitive failings or personal animus.[234] Instead, test bias here reflects structural discrimination based on population inequalities that are embedded into the algorithms.[235]

Scientists examining test bias in education (e.g., aptitude tests) and psychology (e.g., psychometric tests) have standardized a robust empirical methodology.[236] Selected researchers in criminal justice have recently begun to apply this methodological practice to evaluate group bias in recidivism risk tools.[237]

This gold standard for evaluating test bias, endorsed by the American Psychological Association, involves a series of nested models of regression equations involving the test, the group(s) of interest, and an interaction term (test × group) as predictors of test outcomes.[238] A regression is a statistical method to evaluate the relationship between one or more predictors with a response (outcome) variable.[239] An interaction term refers to the product of two predictor variables to determine whether the effect on the outcome of either predictor is moderated by the presence of the other (i.e., changes its strength or direction of influence).[240]

The nested model structure here utilized variables labeled as Black (coded as Black=1, white=0), dummy variable coding for the PSA risk bins (with low-risk as the reference category), and interactions between them with two variables (Black × medium-risk bin and then Black × high-risk bin). Per the gold standard, a four-model structure was employed with the outcome variable being new criminal arrest (arrest=1, no arrest=0). Model 1 tested only the Black variable; Model 2 tested just PSA risk bin dummy variables; Model 3 included both the Black and dummy PSA risk bins; and Model 4 retained Black and the dummy PSA bins, while adding the interaction terms. The interaction terms indicate

Review and Recommendations, *in* Handbook of Psychology: Assessment Psychology 82, 83 (J.R. Graham et al., eds. 2013).

234. *See*, *supra* notes 64–66 and accompanying text.

235. Ben Green, *The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness*, *in* FAT* 2020—Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 594, 598 (2020).

236. Nathan R. Kuncel & David M. Kleiger, *Predictive Bias in Work and Educational Settings*, *in* The Oxford Handbook of Personnel Assessment and Selection 462, 463 (Neal Schmitt ed., 2012) (confirming endorsements also from the National Council on Measurement in Education and the American Educational Research Association).

237. Jennifer Skeem et al., *Gender, Risk, Assessment, and Sanctioning: The Cost of Treating Women Like Men*, 40 L. & Hum. Behav. 580, 585 (2016).

238. Jeanne A. Teresi & Richard N. Jones, *Bias in Psychological Assessment and Other Measures*, *in* 1 APA Handbook of Testing and Assessment in Psychology 139, 144 (2013).

239. Ronet D. Bachman & Raymond Paternoster, Statistics for Criminology and Criminal Justice 675 (Jessica Miller et al. eds., 1997).

240. James Jaccard, Interaction Effects in Logistic Regression 12 (C. Deborah Laughton et al. eds., 2001).

whether any differences in prediction in the ordinal risk bins are a function of race.[241]

This moderated multiple regression modeling is useful to determine whether the form of the relationship between PSA predictions and new criminal arrests exhibited by regression lines vary or depend on the individual's race.[242] In other words, the method determines whether the tool and racial group interact in predicting the likelihood of arrest. More specifically, this method is interested in the existence of variances in the regression lines' intercepts and/or slopes.[243] A significant group difference in either the intercept or the slope represents a deviation from a common regression line.[244] When regression lines do deviate, it means the tool exhibits differential predictive ability and thus, according to the dictates of the modeling structure, test bias exists.[245] In the four model structure, unequal intercepts are indicated by a statistically significant effect of the Black variable in Model 3, while unequal slopes are observed in the existence of a statistically significant effect of any interaction term in Model 4.[246]

A difference in the intercept but not the slope means that one group's regression line lies above the other throughout the test score range, indicating that, comparatively, the test consistently underpredicts risk for the group with the higher regression line.[247] This situation indicates a discrepancy in the tool's predicted outcome for each of the two groups, the size of which does not change as tool scores rise. Unequal slopes likewise portend that the tool's predictions differ for individuals depending on the group to which they belong, but the degree of difference varies by which risk classification is applied.[248]

The importance of this modeling is that unequal intercepts or slopes means that the tool is not race-free and thereby represents test bias. Test bias is alternatively referred to in other disciplines as "group bias,"[249] "predictive bias,"[250] and "differential prediction."[251] Notably, none

241. Henderson et al., *supra* note 122, at 471.
242. Christopher M. Berry, *Differential Validity and Differential Prediction of Cognitive Ability Tests: Understanding Test Bias in the Employment Context*, 2 Ann. Rev. Org. Psychol. & Org. Behav. 435, 444 (2015).
243. Skeem & Lowenkamp, *supra* note 145, at 692.
244. Reynolds & Suzuki, *supra* note 233, at 101.
245. Russell T. Warne et al., *Exploring the Various Interpretations of "Test Bias,"* 20 Cultural Diversity & Ethnic Minority Psychol. 570, 572 (2014).
246. Berry, *supra* note 242, at 439.
247. *Id.* at 443.
248. Cascio & Aguinis, *supra* note 230, at 175.
249. Adam W. Meade & Michael Fetzer, *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*, 12 Org. Res. Methods 738, 738 (2009).
250. Berry, *supra* note 242, at 439 tbl. 1; Henderson et al., *supra* note 122, at 463; Teresi & Jones, *supra* note 238, at 143.
251. Skeem & Lowenkamp, *supra* note 145, at 685; Meade & Fetzer, *supra* note 249, at 738, 740.

of these terms inherently mean that the test exhibits unacceptably poor accuracy. Even if a test meaningfully predicts true risk for two or more groups, differential performance occurs if the test does so in unequal ways.[252] Then variances in application in the direction of negative effects on minorities is another reflection of the potential for disparate impact, here regarding the tool's predictive outcomes.[253]

3. Differential Validity

Research question three inquires about the presence of differential validity to determine if the PSA's accuracy rates are dissimilar based on race. We analyzed multiple dimensions of the PSA tool's discriminative and calibration abilities. Many validation studies regrettably rely solely upon one measure of discrimination.[254] Calibration accuracy is rarely assessed or reported.[255] However, each offers a unique contribution to surveying a tool's accuracy and performance. A tool may vary in how well it meets either of these dimensions.[256]

Discrimination reflects a tool's relative accuracy in how well it distinguishes between known failures and nonfailures.[257] In contrast, calibration measures absolute accuracy in how well the tool's predicted outcomes match the actual outcomes in terms of failure.[258] Best practices in reporting on the utility of a criterion-referenced prediction tool (such as the PSA) endorse engaging measures that address both discrimination and calibration.[259]

a. Discriminative Ability

We calculated for every subsample its post-release arrest rate at each ordinal risk bin to observe whether arrest rates incrementally increased in higher risk bins. For this tool to discriminate well, rates of arrest should noticeably and materially increase from low to medium to high risk. The term "discrimination" as used herein is not meant in the usual legal sense of assigning or denying privileges or benefits to a group based on some protected attribute, such as race or gender.[260] Instead, in

252. Berry *supra* note 242, at 446.
253. Meade & Fetzer, *supra* note 249 at 741.
254. Jay P. Singh, *Five Opportunities for Innovation in Violence Risk Assessment Research*, 1 J. Threat Assessment & Mgmt. 179, 180 (2014).
255. R. Karl Hanson, *Assessing the Calibration of Actuarial Risk Scales: A Primer on the E/O Index*, 44 Crim. Just. & Behav. 26, 27 (2017); Fazel & Bjørkly, *supra* note 119, at 22.
256. L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 Crim. Just. & Behav. 8, 11 (2017).
257. Singh, *supra* note 254, at 180.
258. Duwe, *supra* note 82, at 547.
259. *Id.*; Seena Fazel & Achim Wolf, *Selecting a Risk Assessment Tool to Use in Practice: A 10-Point Guide*, 21 Evid. Based Mental Health 41, 42 (2018); Hanson, *supra* note 255, at 33.
260. Black's Law Dictionary 543 (Bryan A Garner ed., 9th ed., 1990).

the genre of predictive technologies, the term reflects the ability to distinguish between different outcomes.[261]

The Area Under the Curve (AUC) is a global measure of a tool's discriminative ability.[262] More specifically, the AUC here represents the probability that a randomly selected arrestee receives a higher risk classification than a randomly selected non-arrestee.[263] An AUC of 0.5 indicates no better accuracy than chance and 1.0 signifies perfect discrimination (i.e., all those rearrested are classified as higher risk than all those not rearrested).[264]

No statistical standard exists for how close to 1.0 an AUC must be to signify a sufficiently accurate test in terms of discriminatory utility. Determining the degree of acceptable performance per the AUC is, instead, a value judgment.[265] Some risk assessment scholars refer to AUCs of .56, .64, and .71 as the thresholds for small, medium, and large effect sizes, respectively.[266] On the other hand, a more conservative conceptualization regards AUCs as failing (below .60), poor (.60 to .69), fair/moderate (.70 to .79), good (.81 to .89), and excellent (.90 and above).[267] We analyzed the AUC using the ordinal risk bin strategy, as is common in the literature.[268]

The global AUC statistic does not distinguish relative ability between true positives (failures correctly classified to fail) and true negatives (nonfailures correctly classified to not fail).[269] Thus, we computed the True Positive Rate (TPR) which indicates the proportion of individuals with new criminal charges correctly predicted to be arrested.[270] The True Negative Rate (TNR) is the proportion of those without new criminal charges who are correctly predicted to succeed.

These three discrimination statistics (i.e., AUC, TPR, TNR) require a choice of cut-point from the scale because they rely upon a dichotomous

261. Oxford English Dictionary (3rd ed. 2013).
262. Hamilton et al., *supra* note 83, at 246.
263. Helmus & Babchishin, *supra* note 256, at 12.
264. *Id.*
265. Sheldon X. Zhang et al., *An Analysis of Prisoner Reentry and Parole Risk Using COMPAS and Traditional Criminal History Measures*, 60 Crime & Delinq. 167, 171–72 (2014).
266. Helmus & Babchishin, *supra* note 256, at 12.
267. Milena Abbiati et al., *Validity and Predictive Accuracy of the Structured Assessment of Protective Factors for Violence Risk in Criminal Forensic Evaluations: A Swiss Cross-Validation Retrospective Study*, 44 Crim. Just. & Behav. 493, 501 (2017).
268. Gregório Hertz et al., *Cross-Validation of the Revised Version of the Violence Risk Appraisal Guide (VRAG-R) in a Sample of Individuals Convicted of Sexual Offenses*, Sexual Abuse (2021); *See* Mona J.E. Danner et al., Risk-Based Pretrial Release Recommendation and Supervision Guidelines 13 (2015).
269. Daryl G. Kroner, *The Roles of the Risk Estimate and Clinical Information in Risk Assessments*, *in* The Wiley International Handbook of Correctional Psychology 446, 451 (Devon L.L. Polaschek et al. eds., 2019).
270. Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 Behav. Sci. & L. 8, 9 (2013).

perspective of whether the tool predicted the individual to be arrested or not to be arrested. We used here the PSA's risk bin strategy by combining the low- and medium-risk bins with the prediction being non-arrest and then using just the high-risk bin classification as representing those predicted to be arrested.

To supplement the foregoing analyses, we calculated an additional overall discrimination statistic. Somers' *d* is a correlation statistic that measures the strength between a single ordinal predictor (here involving risk bins) and the binary outcome of arrest.[271] For any subsample, Somers' *d* was calculated along with a statistical significance test (using an alpha level of .001) to assess the relationship between ordinal bins and arrest.

To evaluate the existence of differential discriminative validity, we compared the foregoing statistics between racial groups within each site by corresponding *z*-tests using a conservative alpha level of .001 to reject the null hypothesis that the statistics were equal between races. If the AUCs are equivalent by race, the tool achieves what is known in the algorithmic fairness literature as "AUC parity."[272] In the algorithmic fairness literature, when TPRs and TNRs are equivalent across groups, the tool achieves the fairness definition of "equal opportunity" or its alternative name of "equalized odds."[273] The reciprocals of TPRs and TNRs represent error rates and if they are not equal, the tool does not in the algorithmic risk literature achieve "error rate balance" and the situation presents as disparate mistreatment.[274] Then we tested whether there were significant differences in the Somers' *d* statistics between races (*z*-tests), the presence of which would signify differential validity.[275]

b. Calibration

Measures of calibration were evaluated to discern the PSA's levels of absolute accuracy. We offer an overall accuracy measure of the proportion of cases the tool correctly predicted in terms of true positives plus true negatives. The algorithmic fairness notion of "overall accuracy equality" requires equivalent accuracy rates.[276] Yet this global accuracy measure is limited in that it obscures whether the predictions are better at forecasting true positives or true negatives.

Hence, we calculated two additional calibration statistics. The Positive Predictive Value (PPV) is the proportion of higher risk predictions who were arrested.[277] The Negative Predictive Value (NPV) then presents the proportion of lower risk predictions who were not

271. Taxman, *supra* note 21, at 273.
272. Mitchell, *supra* note 224.
273. *Id.*
274. Pessach, *supra* note 223.
275. Berry, *supra* note 105, at 209.
276. Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, FairWare'18, 4 (2018).
277. Daniel J. Neller & Richard I. Frederick, *Classification Accuracy of Actuarial Risk Assessment Instruments*, 31 Behav. Sci. & L. 141, 143 (2013).

arrested.[278] The overall accuracy, PPV, and NPV require a dichotomous cut-off to create two categories of those predicted to have a new criminal arrest and those predicted not to be arrested. We used the same cut-point as earlier applied with only those with a high-risk assignment predicted to be arrested. Differential calibration performance in these metrics is evaluated through the differences between proportions by race via z-scores at a conservative $p < .001$ to reject the null hypothesis of equal percentages.

TPRs/TNRs from the discriminative ability discussion may appear similar to PPVs/NPVs. They use the same underlying numbers (e.g., true positives and true negatives) but are calculated through contrasting equations. TPR asks of those with a new criminal arrest, what percentage were classified as high risk? PPV instead asks of those classified as high risk, what percentage were arrested? Then TNR asks of those without a new arrest, what percentage was classified as lower risk? NPV instead asks of those predicted as lower risk, what percentage were not arrested?

In the algorithmic fairness literature, equivalence in PPVs and NPVs is evidence of "predictive parity"[279] and "well-calibration."[280] Predictive parity exists, for example, when those designated high risk in each group have similar arrest rates, as is indicated by the PPVs.[281] Predictive parity also requires that those predicted not to be rearrested in each group have similar arrest rates, indicated by the NPVs.[282]

III. Evaluating the Algorithm for Race Effects

The following presents the statistical analyses to answer the three main research questions. The results raise important policy concerns that are highlighted and discussed in terms of their implications for pretrial risk practices. The findings are instructive to the use of algorithmic risk tools at other criminal justice decision points, as well. The information gleaned from the following uncovers how a widely used risk tool operates and the potential for race-based effects in real world terms.

278. G. Szmukler et al., *supra* note 93, at 896. In Bayes rule terms, the PPV is the posterior probability of failure given a prediction of failure; the NPV is the posterior probability of success given a prediction of nonfailure. Kristian Linnet et al., *Quantifying the Accuracy of a Diagnostic Test or Marker*, 58 Clin. Chem. 1292, 1296 (2012).
279. Pratyush Garget et al., *Fairness Metrics: A Comparative Analysis*, 4 (2020), https://arxiv.org/pdf/2001.07864.
280. Till Speicher et al., *A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices*, 3 (2018), https://arxiv.org/pdf/1807.00787.
281. Ugwudike, *supra* note 17, at 14; Alexandria Chouldechova, *Fair Predictions with Disparate Impact: A Methodological Primer*, 5 Big Data 153, 155 (2017).
282. The term of conditional use accuracy equality applies if the tool achieves equal PPVs and NPVs across groups. Verma, *supra* note 276.

1. Disparate Impact

Research question one inquired about the existence of disparate impact. Table 1 provides summary statistics on the proportions of released individuals placed into the PSA risk bins and indicates whether the differences by race for each sample (i.e., Illinois and Kentucky) by risk bin are statistically significant.

Table 1: Risk Bin Assignments of Released Defendants

	Illinois			Kentucky		
	White	*Black*	*p*	*White*	*Black*	*p*
Risk Bin						
Low	48.4%	29.9%	*	46.6%	33.4%	*
Medium	42.5%	57.6%	*	*	50.4%	*
High	9.1%	12.5%	*	9.9%	16.2%	*

* $p < .001$

The PSA at each site places a smaller proportion of whites into the high-risk bin, with statistically significant differences. For Illinois, the high-risk bin is unequal: (9 percent white, 13 percent Black) ($p < .001$). For Kentucky, the high-risk bin is also unequal: (10 percent white, 16 percent Black) ($p < .001$). When one combines the medium- and high-risk bins, a significantly lower percentage of whites are assigned to them: Illinois: (52 percent white, 70 percent Black) (not shown in table, $p < .001$); and then for Kentucky: (53 percent white, 67 percent Black) (not shown in table, $p < .001$).

In contrast, the PSA places a greater proportion of whites into the low-risk bin. At each site, the PSA assigns low risk to approximately one-half of white defendants and approximately one-third of Black defendants ($p < .001$).

Thus, these results reflect disparate impact in that Black individuals are classified at substantially greater rates into the higher risk bins in both samples. Still, these are the *released* defendants, and the differences may not entirely reflect the PSA as there are other factors that enter into judicial decisions to release. Thus, we employed the full dataset of all defendants (released or not) scored on the PSA at the Illinois site during the time of study. While we did not have access to similar information in Kentucky, it is still informative to at least check with one site that the differences were not necessarily an artifact of release decisions. Table 2 provides the risk bin classifications by race for all defendants in the Illinois sample.

Table 2: Risk Bin Assignments of All Scored Defendants in Illinois

	Illinois		
	White	*Black*	*p*
Risk Bin			
Low	42.1%	24.6%	*
Medium	44.7%	58.2%	*
High	13.2%	17.3%	*

Table 2 confirms the existence of disparate impact in Illinois as Black defendants, both released and nonreleased, are disproportionately assigned to the medium- ($p < .001$) and high-risk bins ($p < .001$), while being significantly less likely to be placed in the low-risk bin: (white: 42 percent, Black: 25 percent) ($p < .001$). That is, two in five white defendants are assigned low-risk scores while two in eight Black defendants are assigned to the low-risk bin.

Returning to the main datasets of released defendants, we review the 80 percent rule of thumb demonstrating adverse impact from employment law. In the low-risk bin, the selection ratios for this favorable outcome for Black to white individuals is 62 percent in Illinois and 72 percent in Kentucky, which are less than the 80 percent baseline. At least in the employment discrimination world, these results exhibit prima facie evidence of adverse impact in the beneficial attribute of selection into the low-risk classification.[283] We invert the analysis considering the 80 percent rule for avoiding the negative outcome of a high-risk classification. In the high-risk bin, the selection ratios of white to Black defendants is 73 percent in Illinois and 61 percent in Kentucky. These rates are below the 80 percent criteria as well, thus signifying adverse impact by this legal measure with respect to the high-risk binning strategy.

A check on the rule of thumb for adverse impact in the full Illinois sample is of interest. In the full sample (combining released and nonreleased defendants), the selection ratio into the low-risk bin is 58 percent Black to white defendants, while the selection into the high-risk grouping ratio is 76 percent of white to Black defendants. These statistics from the full sample in Illinois confirm disparate impact on Black individuals as being disadvantaged in high-risk placements, and that such is a manifestation of the risk tool and not a fiction created by judicial release decisions.

In sum, these findings demonstrate disparate impact of the PSA classification binning outcomes in disserving Black individuals compared to whites in three ways: (1) in a practical sense, with observable discrepancies in the rates of selection into high-risk and into low-risk bins, (2) in an empirical sense, with disparities in proportions in high- and low-risk bins having statistically significant z-test statistics at a conservative $p < .001$ level, and (3) in a legal benchmark sense, with selection ratios of less than 80 percent in high- and low-risk bins using the popular adverse impact statistical standard.

2. Test Bias

The second research question inquired about the existence of test bias. This required the nested regression model structure promoted by the fields of psychology and education. Table 3 contains the models for the Illinois dataset. The p-values are here aligned toward the statistical significance of individual coefficients in each model.

283. Mayson, *supra* note 135, at 2242.

Table 3: Test Bias Models for Racial Fairness—Illinois

	Model 1	Model 2	Model 3	Model 4
Predictor				
Black	1.636*		1.481*	1.747*
Low risk				
Medium risk		1.793*	1.718*	2.021*
High risk		2.345*	2.247*	3.252*
Black × Low				
Black × Medium				.820
Black × High				.647*
Constant	.151*	.146*	.108*	.095*
Model Statistics				
-2LL	34745	34431	34327	34315
Nagelkerke R^2	.008	.021	.026	.027
n=36,298				

*$p < .001$

In the Illinois sample, in Model 1 the Black variable indicates that an individual being Black, on its own, is predictive of a greater likelihood of a new criminal arrest. (This finding is consistent with the fact that in Illinois, Black individuals have a higher rate of new criminal arrests than whites, as will be shown further down in Table 5.) Model 2 shows that the medium- and high-risk bins significantly and positively predict greater odds of being arrested. The Model 2 results support the utility of the PSA risk bins, irrespective of race.

The Black variable in Model 3, which controls for risk bin predictions, is significant ($p < .001$), and because the odds ratio is greater than one, the result means that the PSA underpredicts for Black defendants in the low-risk bin. This finding of bias in the intercept importantly signifies test bias.

In Model 4, the interaction term for Black × medium (not shown in the table, $p = .025$) is not statistically significant at the conservative .001 level but would have been at a more lenient, yet typical, .05 level. The statistically significant interaction for Black × high risk is statistically significant at the conservative level ($p < .001$), and it signals that the influence of a high-risk bin prediction is moderated by race. The tool is simply not calibrated equally for these racial groups. In empirical terms, this result means that, in addition to bias in the intercept as previously mentioned, there is bias in the slope of the relationship between the PSA and arrest between the races. Both interaction terms are less than one, indicating that the medium- and high-risk bins overestimated risk for Black compared to white defendants. These terms also mean that the degree of underprediction for Black defendants, as indicated by the intercept, weakens as the risk bins increase from low to medium and to high bins.

Hence, Table 3 provides evidence for the racialized performance of the PSA in Illinois, which in the psychometric literature qualifies as test bias. The form of the relationship between the PSA and race is simply unequal between the races in Illinois. Table 4 contains the test bias models for the Kentucky sample.

Table 4: Test Bias Models for Racial Fairness—Kentucky

	Model 1	Model 2	Model 3	Model 4
Predictor				
Black	1.048		.904*	.848*
Low risk				
Medium risk		2.272*	2.286*	2.271*
High risk		4.262*	4.314*	4.221*
Black × Low				
Black × Medium				1.062
Black × High				1.132
Constant	.120*	.064*	.065*	.066*
Model Statistics				
-2LL	110176	106244	106221	106218
Nagelkerke R^2	.000	.049	.049	.049
n=161,330				

* < *.001*

For Kentucky, Model 1 indicates there is no relationship between Black and arrest when just the race variable is used. (As will be shown in Table 5, this result is consistent with the base rates of arrest being equivalent by race at this site.) Model 2 confirms the ability of the PSA to predict arrest across races in the medium- and high-risk bins, with statistically significant odds ratios ($p < .001$).

For Model 3, the statistically significant race factor ($p < .001$) shows racial bias in the intercept whereby, as reflected in an odds ratio less than one, the PSA overpredicts risk for Black defendants. The interaction terms in Model 4 are not statistically significant, meaning that there is no difference in slopes. Thus, test bias for Kentucky is limited to the intercept.

Overall, the test bias results are not fully compatible. While both sites show bias in the intercepts, confirming the existence of test bias, the directions of the bias are conflicting. In the low-risk bin, the PSA in Illinois underpredicts for Black defendants while in Kentucky it overpredicts for Black defendants. There is bias in the slope in Illinois, indicating that the form of the relationship between the PSA and arrest is not the same for each race, plus the differences vary by risk level. No slope bias, though, exists in the Kentucky sample, such that the overprediction for Black individuals was at a consistent level across risk bins. These findings

of test bias technically qualify as differential validity as well, though we deployed supplemental metrics that evaluate differential validity through varied lenses.

3. Differential Validity

The third research question considered whether there is evidence of differential performance in terms of the PSA's abilities on discrimination and calibration. Throughout this Subpart, while the main interest is on racial comparators, the findings also provide some valuable insight into the performance abilities of the PSA as a risk assessment tool.

a. Discriminative Validity

Table 5 contains the results for discriminative validity of the PSA in the two sites and by racial groups within each site. For each of the four subgroups, the table provides base rates of new criminal arrest, rates of arrest at each risk bin, AUCs, TPRs, TNRs, and Somers' *d* correlations. The *p*-values indicate statistically significant differences between races in each sample.

Table 5: Discrimination Measures for New Criminal Arrest

	Illinois			Kentucky		
	White	***Black***	***p***	***White***	***Black***	***p***
Base rate	13.2%	19.9%	*	10.7%	11.1%	ns
Low	8.6%	14.2%	*	6.2%	5.3%	*
Medium	16.1%	21.5%	*	13.0%	11.9%	*
High	23.5%	25.8%	ns	21.7%	21.0%	ns
AUC	.608	.563	*	.630	.637	ns
True positive rate	16.3%	16.2%	ns	20.2%	30.5%	*
True negative rate	92.0%	88.5%	*	91.3%	85.6%	*
Somers' d	.216	.126	*	.260	.274	ns

** $p < .001$; ns = not significant*

In the Illinois sample, the overall base rates indicate that whites have a lesser rate of new criminal arrests: (13 percent white, 20 percent Black) ($p < .001$). The Kentucky sample presents no difference in the base rate of new arrests with an 11 percent rate for both races (n.s.). Per Table 5, at both sites and for both races, the PSA shows discriminatory ability whereby arrest rates increase in a linear fashion from low to medium to high risk. Hence, there is some discriminative utility of the PSA risk bins overall.

Reviewing performances by race, significant differences exist for the Illinois sample. At each risk bin, whites in Illinois are arrested at reduced rates: low risk (9 percent white, 14 percent Black) ($p < .001$), medium risk (16 percent white, 22 percent Black) ($p < .001$), high risk (24 percent white, 26 percent Black) (n.s.). The high-risk difference was not, however, significant.

For Kentucky, whites were more likely to be arrested in the low-risk bin (6 percent white, 5 percent Black) ($p < .001$) and medium-risk bin (13 percent white, 12 percent Black) ($p < .001$). These results showing statistical significance are influenced by large sample sizes, whereby practically the distinctions are relatively small. In the Kentucky sample, arrest rates are similar for the high-risk bin (22 percent white, 21 percent Black) (n.s.).

The AUCs in the Illinois dataset indicate better discriminatory ability for whites with an AUC of .608, compared to an AUC for Black individuals of .563 ($p < .001$). In contrast, the AUCs shows equivalent discriminatory ability for the races in the Kentucky sample with an AUC for whites of .630 and for Black individuals of .637 (n.s.). Hence, while there is not AUC parity in the Illinois sample, AUC parity exists for Kentucky.

A reflection upon classification accuracy overall appears justified. Across subsamples, the AUCs range from 56 to 64 percent in the ability to classify arrestees higher than nonarrestees. At the low end, with a 56 percent rate of correct classifications, the PSA does not, from a practical perspective, perform materially better than chance. These AUCs mean that the PSA has appreciable classification error rates from 36 to 44 percent. As judged by the relevant risk assessment literature, the AUCs exhibit, in one conceptualization, small effect sizes in three of the subsamples and borders on the threshold of medium effect size for the fourth. In the more conservative labeling, the AUC for Black individuals in Illinois is judged as failing, while the AUCs for the rest exhibit poor performance.

The next statistics break apart classification accuracy for true positives and true negatives. The true positive rates were equivalent in Illinois at 16 percent (n.s.), but not in Kentucky where whites subject to new criminal arrests were less likely to be correctly classified as high-risk (white: 20 percent, Black: 31 percent) ($p < .001$). True negative rates indicated differential discriminatory validity at both sites with nonarrested whites being correctly classified as lower risk at larger rates than Black individuals ($p < .001$). Hence, except for the true positive rate in Illinois, the PSA does not comply with the algorithmic fairness definitions of equal opportunity or equalized odds, and taking the reciprocals of the TNRs and TNRs, it also means there is not error rate balance. These results suggest disparate mistreatment of Black individuals in the TNRs in both sites, but disparate treatment of whites in the TPR in Kentucky.

The Somers' *d* measure of association between the ordinal risk bins and arrest rates ranges from .126 to .274. The Somers' *d* indicates differential validity by race in Illinois, whereby the strength of the relationship between PSA and arrests is significantly stronger for whites, with a statistic of .216 compared to.126 ($p < .001$) for Black individuals. In contrast, the Somers' *d* statistics do not vary by race in Kentucky, with a statistic of .260 for whites and .274 for Black individuals (n.s.), and thus the strength

of the relationship is equivalent.[284] Differential validity on this measure is thereby indicated in Illinois but not in Kentucky.

In sum, there is strong evidence that the PSA's ability to classify varies by race, though not with every statistic in Table 5. With respect to certain of those statistics (i.e., AUC, true positive rate, true negative rate, and Somers' d), classification ability is the persistently weakest with Black individuals in Illinois. This finding is notable because the Illinois site represents a supermajority Black population.

b. Calibration

Table 6 contains the calibration metrics for overall accuracy, PPVs, and NPVs. The *p*-values indicate if the differences between proportions by race are statistically significant.

Table 6: Calibration Measures for New Criminal Arrest

	Illinois			Kentucky		
	White	***Black***	***p***	***White***	***Black***	***p***
Overall accuracy	82.0%	74.1%	*	83.7%	79.5%	*
Positive predictive value	23.5%	25.8%	*	21.7%	21.0%	ns
Negative predictive value	879%	81.0%	*	90.5%	90.8%	ns

** $p < .001$; ns = not significant*

The results for overall accuracy demonstrate that the PSA's predictive accuracy (combining true positives and true negatives) was stronger for whites in the two samples ($p < .001$). The p-values indicate that neither sample achieves the algorithmic fairness criteria of overall accuracy equality.

Results were mixed when parsing accuracy rates for predictions of arrest versus predictions of no arrest. In Illinois, the differential PPVs indicate the PSA was worse at positively predicting arrest for whites (white: 24 percent, Black: 26 percent) ($p < .001$). On the measure of PPVs, the tool does not offer predictive parity in Illinois, though the variance is, in practical terms, small in size. For Kentucky, the PPVs were not statistically different by race (white: 22 percent, Black: 21 percent) (n.s.) and thus on this measure achieve predictive parity.

Noticeably, the accuracy of the PSA for predicting new criminal arrests is quite low. The reciprocal statistics (i.e., 1 – PPVs) indicate forecasting false positive rates of at least 74 percent.

Results from the NPVs for predicting non-arrests are far more accurate. In Illinois, the NPVs did a better job at predicting nonarrest for whites (white: 88 percent, Black: 81 percent) ($p < .001$). In Kentucky, the NPVs were equivalent at 91 percent (n.s.). The NPVs by race reflect differential calibration in Illinois, but not in Kentucky. The forecasting false negative error rates (i.e., 1 – NPVs) are 9 to 19 percent.

284. The *z*-test value was $p = .007$, and thus would have been considered statistically significant at an alpha of .05.

In terms of the PSA's ability in general, with contrasting PPV and NPV results, the PSA exhibits predictive parity and is well calibrated in Kentucky, but not so in Illinois. Further, the results mean that the PSA performs at an appreciable level at predicting success in terms of nonarrest with an accuracy rate of at least 80 percent, leaving a forecasting false negative rate across samples of less than 20 percent. Compared to an error rate on positively predicting arrest at 74 percent, the PSA as a tool is far better at predicting non-arrests than it is at predicting arrests.

A. *Policy Implications*

A clear conclusion from this study is that this tool is not race-free. Despite assertions by the tool's owner that the PSA is independent of racial inequalities,[285] study results reveal that (1) risk bin outcomes create disparate impact, (2) the gold standard for statistically evaluating test bias made evident that group bias does exist, and (3) the tool does not perform consistently in classification and prediction for white and Black individuals.

Disparate impact is evident in that at both sites, Black individuals are less likely to receive low-risk classifications and to avoid the higher risk classifications. This observation does not, on its own, dictate a constitutional violation. As an illustration, the law on adverse impact in employment law allows an employer to demonstrate a legitimate, nondiscriminatory justification for a practice that disproportionately impacts a protected group.[286] Here, the justification would likely be couched in terms of the numerous advantages of algorithmic risk to criminal justice reforms of money bail and reduced pretrial detention, as outlined earlier herein.[287] Nonetheless, the result is informative to the political, normative, and ethical discussions that surround any criminal justice policy which, no matter how well intended, further disenfranchises minorities.

The gold standard of modeling test bias revealed group bias at both sites, but interestingly in opposing directions. Test bias is inequitable to Black individuals in Kentucky by overpredicting risk yet test bias favored Black individuals in Illinois by underpredicting risk. These results raise ethical and legal concerns in terms of race-based advantages and disadvantages.

The ideology of the algorithmic fairness literature is flourishing, and many definitional offerings to delineate the existence of algorithmic fairness (or unfairness) are emerging. This study revealed violations of several of them. The statistical measures used to detect disparate impact also show, for example, that the tool does not comply with the algorithmic fairness terms of statistical parity, demographic parity, and equal

285. Arnold Found., *supra* note 182.
286. Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 Ga. L. Rev. 109, 161 (2017).
287. *Supra* Subpart I.A.

acceptance rates. The tool fails to assign equal rates of whites and Black individuals to any of the risk bins.

The algorithmic fairness results in evaluating differential validity measures are not always consistent. Evidence of site effect is substantial in that there were fewer instances of differential performance in Kentucky than in Illinois. The tool performs more racially equitably in Kentucky. This conclusion is not entirely surprising considering PSA developers at least in part trained the final PSA scale in Kentucky, and thus may have overfitted the tool to Kentucky data for all demographic groups. The shrinkage in performance for both races is marked, as evidenced herein when transferring the tool to Illinois. Weakness in performance is also applicable to racial differences. The PSA failed on many of the algorithmic fairness definitions in Illinois (e.g., AUC parity, predictive parity). While the PSA was more racially fair in Kentucky, it still violated certain algorithmic fairness definitions (e.g., equal opportunity, overall accuracy validity). It is notable that better overall performance exists in the sample that is a supermajority white population (i.e., Kentucky), which confirms observations that risk tools tend to perform better on largely white samples. As the PSA is used by judges to make important liberty infringing decisions, the foregoing results strongly suggest the need for judicial oversight[288] or other external governance.[289]

It is beyond the scope of this Article to substantiate ways to ameliorate racial disparities in the tool. Yet it is evident that any retooling must orient to the fact that the PSA's predictors and its outcome are crime centric. Decidedly, the placement of a greater proportion of Black individuals into the medium- and high-risk bins are a conspicuous consequence of criminal history. The new criminal arrest scale includes an age variable, but it represents only two out of a total possible thirteen points in raw scoring. Eighty-five percent of the risk classification, then, is based on criminal history measures. As a result, it is clear that it is criminal history that is driving the disparate impact. For the many reasons discussed earlier, this may not reflect actual disparities in rates of offending by race.[290] An evident policy issue here is for stakeholders to reconsider relying on a tool that is so heavily dependent on criminal history, considering the clear results shown herein of the detrimental labels connoting dangerousness being disproportionately assigned to Black individuals. PSA outcomes are used in decisions in which personal liberty is at issue. Considering officials adopted the tool with express encouragement for judges to respect the tool, it is foreseeable that the tool will reduce the likelihood of release of Black compared to white individuals.

288. Binns, *supra* note 103, at 7 (anticipating increase in judicial scrutiny, but noting "[i]t remains to be seen how administrative law standards of reasonableness, rationality, relevancy, intelligibility and adequacy can apply to decisions influenced by systems based on machine learning, and whether new standards may be needed to hold those who deploy them accountable").

289. Taxman, *supra* note 21, at 282.

290. *Supra* Subpart I.C.

At a very high level, possibilities for retooling include modifications to how these variables are chosen, operationalized, and evidenced. It seems justifiable, too, to engage in further research to add supplemental, noncriminal history predictive factors, particularly those that might be stronger predictors of arrest for Black individuals. Investigating whether a substitute for the outcome variable that defines dangerousness as an arrest might moderate racial differentials is warranted (e.g., arrest for only serious offenses, convictions). Stakeholders might also consider some form of algorithmic affirmative action to adjust factors, their scores, or weights by race, though this form of reengineering would be novel and controversial for a host of reasons, as is considered elsewhere.[291]

Racial differences aside, the study reveals some affirming demonstrations the PSA offers some utility. In terms of discriminatory ability, the risk binning strategy is positively correlated with new criminal arrests and the tool is more likely than not to rank arrestees in a higher risk grouping than non-arrestees. The PSA performs quite well in predicting success in terms of not being arrested on a new charge as shown by large accuracy rates of at least 80 percent in the low- and medium-risk classifications (i.e., the NPVs). At least in the jurisdictions sampled, a PSA prediction of low or medium risk thereby deserves a healthy degree of confidence. For judges wishing to be somewhat confident that individuals rated as low or medium risk are not likely to have a new arrest, the PSA may thus be appreciated as a legitimate tool to initially screen out risk, and thus release, those defendants who are not scored as high risk.

A corresponding policy issue regards the findings that the PSA scale performs with extremely large error rates in predicting arrest (as opposed to non-arrest). Weak accuracy performance for high-risk attributions signifies that the PSA tool should not be the only data point specifically in decisions to *deny* release. Thus, judges faced with individuals with a "high-risk" outcome should look for more information to supplement that PSA outcome before determining the individual should be detained to protect the public.

Other results from a general view of the PSA (outside of racial issues) may be informative outside the context of this tool and in other decisionmaking settings. This study confirms the benefit and thereby need for independent crossvalidation research and the downside for relying upon a developer's assertions about equal performance across groups. In general, the results highlight the need for localized validation studies. The evidence here indicates the presence of site effect. When comparing two quite different jurisdictions from sociodemographic and institutional resource considerations, the same tool did not perform equally as well based on site or on race. The tool could well be considered for revision based on each site's characteristics that correlate to criminal offending.

291. *See generally* Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 Geo. L.J. 803, 803 (2020); Skeem & Lowenkamp, *supra* note 145 (illustrating algorithmic affirmative action on a federal dataset).

The evidence may be instructive to stakeholders who might not be cognizant of what level of probability is associated with each risk category.[292] In this study, arrest rates for the PSA are these: low risk (5–14 percent) to medium risk (12–22 percent) to high risk (21–26 percent). Thus, this information is broadly useful to allow stakeholders a rough idea of the percentile likelihood of failure by category, at least for this popular risk tool. While this data on the failure rates by risk bin are important, the tale is also cautionary. Simply put, the risk bins do not mean the same thing across samples. Stakeholders should be aware of applying the appropriate offense rates rather than erroneously substituting proportional failures from a different sample.

A related caution is that there exists a concerning overlap in base rate ranges across the risk bins. A 14 percent likelihood of arrest is low risk in one subsample yet qualifies as medium risk in another. A 22 percent likelihood of arrest is medium risk for one group yet high risk in another. Such a conflation should be informative to stakeholders. These findings highlight the idea that risk bins might reasonably be better promoted as achieving a stronger efficacy in providing relative predictions and that it is not as salient for absolutist likelihood estimates.[293]

The probabilities associated with the high-risk bin raise a related policy issue. The PSA's high-risk bin seemingly fails to isolate to the degree of likelihood of failure that stakeholders might anticipate. Across all subsamples, fewer than 30 percent of defendants classified at high risk were arrested anew after release. Hence, over 70 percent of individuals designated as "high risk" were instead successful in remaining free of a new arrest. While no normative consensus exists in the criminal justice risk field on the degree of probability that ought to qualify as high risk,[294] this finding at least raises the question as to whether stakeholders would regard a 21 to 26 percent probability as justifying such a disparaging status label.[295] If the PSA's designation of "high risk" triggers even an informal presumption of pretrial detention, the consequence is that a significant percentage of individuals will be detained unnecessarily. This point is even more important when considering that the PSA's new criminal arrest scale is predicting any arrest, not just *violent* arrests or even

292. Stevenson, *supra* note 155, at 306 ("Judges may not understand exactly what the risk score is measuring, or what level of statistical risk is associated with each risk category.").

293. Klingele, *supra* note 77, at 219.

294. Nicholas Scurich, *The Case Against Categorical Risk Estimates*, 36 Behav. Sci. & L. 554, 558 (2018); Stephanie A. Evans & Karen L. Salekin, *Violence Risk Communication: What Do Judges and Forensic Clinicians Prefer and Understand?*, 3 J. Threat Assessment & Mgmt. 143, 156 (2016); J.C. Oleson et al., *Training to See Risk: Measuring the Accuracy of Clinical and Actuarial Risk Assessments Among Federal Probation Officers*, 75 Fed. Prob. 52, 55 (2011).

295. David G. Robinson et al., *Pretrial Risk Assessments: A Practical Guide for Judges*, 57 Judges J. 8, 9 (2018) (warning judges who use risk assessment: "What statistical probability does the "high-risk" label correspond to in your jurisdiction? It might be lower than you think.").

serious arrests. These points should inform relevant public debate as to the degree of probability and of the severity of the predicted offenses which should ethically and normatively qualify as sufficiently risky to justify incarceration for those whose guilt has not been determined.[296] Such a large error rate for predicting any type of arrest means that the lofty goals of introducing algorithmic risk predictions to ameliorate mass incarceration will suffer.

B. *Advantages and Limitations*

This study offers some novel contributions to the literature. Findings critically add to the small knowledge base that evaluates racial bias in risk instruments.[297] Whereas most research in the risk assessment literature depends upon a single sample of convicted prisoners released from prison,[298] this study employs a novel two-sample design involving diverse jurisdictions of pretrial defendants discharged from local jails. The contrasting samples permitted a unique perspective on tool performance with a supermajority white sample compared to a supermajority Black sample.

At present, the few research reports about pretrial risk tool performance that provide any data about their predictive powers typically offer merely one or two statistical measures.[299] We adopted the best practices model of reporting on a variety of measures to judge the PSA's utility for new arrests in its relative ability to discriminate the risk of those who failed versus those who did not and then its absolute ability in terms of how well calibrated the PSA is in its predictions of new arrests. Moreover, the analyses were bifurcated to address differential validity by sample and by race. This permitted further exploration into the potential for disparate impact or discriminatory outcomes by racial grouping.

Several limitations should be mentioned, though it is important to note that these tend to plague other risk assessment validation studies conducted by independent third parties who cannot control data collection. The datasets did not allow for a consistent or fixed followup period. No interrater reliability scores were available. The units of analysis are cases rather than individuals. Thus, it is possible that some individuals may be counted more than once if they were released, rearrested, and released again in the time periods studied. Then we could not control for

296. Stevenson & Mayson, *supra* note 4, at 39.
297. Taxman, *supra* note 21, at 277.
298. Synøve Nygaard Andersen & Torbjørn Skardhamar, *Pick a Number: Mapping Recidivism Measures and Their Consequences*, 63 Crime & Delinq. 613, 617 (2017).
299. Kim KiDeuk & Grant Duwe, *Improving the Performance of Risk Assessments*, *in* Handbook on Risk and Need Assessment: Theory And Practice 189, 215 (Faye S. Taxman ed., 2017); Douglas, *supra* note 88, at 135; Jay P. Singh, *Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review*, 31 Behav. Sci. & L. 55, 61 (2013).

individual conditions of release (e.g., restrictions, service offerings) which may correlate with new arrests.

Conclusions

Hopes for progressive reform by adopting risk algorithms to ameliorate race-based disparities are aspirational. The reality may somewhat differ. The independent study offered herein suggests caution. A popular risk assessment tool widely used across jurisdictions provides some utility in predicting success in terms of released defendants not being rearrested. Plus, the tool has some ability to classify whereby its low-, medium-, and high-risk bins are related to higher arrest rates in the direction expected.

Because the tool exhibits substantial error rates in predicting new arrests, one should proceed with caution. With respect to the focus on racial disparities, ample evidence is provided herein of disparate performance for white versus Black individuals in ways that, in most cases, disadvantage Black individuals. The information should spark more discussion and debate. These results do not necessarily require that the algorithmic risk turn be abruptly abandoned. Instead, more care can be taken to retool algorithms to reduce racial inequities while serving the reform movement's goal of increasing the release rates of white and Black pretrial defendants.