EVALUATING ALGORITHMIC RISK ASSESSMENT

Melissa Hamilton*

Algorithmic risk assessment is hailed as offering criminal justice officials a science-led system to triage offender populations to better manage low- versus high-risk individuals. Risk algorithms have reached the pretrial world as a best practices method to aid in reforms to reduce reliance upon money bail and to moderate pretrial detention's material contribution to mass incarceration. Still, these promises are elusive if algorithmic tools are unable to achieve sufficient accurate rates in predicting criminal justice failure. This article presents an empirical study of the most popular pretrial risk tool used in the United States. Developers promote the Public Safety Assessment (PSA) as a national tool. Little information is known about the PSA's developmental methodologies or performance statistics. The dearth of intelligence is alarming as the tool is being used in high-stakes decisions as to whether to detain individuals who have not yet been convicted of any crime. This study uncovers evidence of performance accuracy using a variety of validity metrics and, as a novel contribution, investigates the use of the tool in three diverse jurisdictions to evaluate how well the tool generalizes in real-world settings. Policy implications of the findings may be enlightening to officials, practitioners, and other stakeholders interested in pretrial justice as well as in the use of algorithmic risk across criminal justice decision points.

Keywords: *risk assessment, pretrial, algorithmic fairness, criminal justice reform*

*Reader of Law & Criminal Justice, University of Surrey School of Law; J.D., The University of Texas School of Law; Ph.D. (criminal justice), The University of Texas at Austin. This research was generously supported by funding from the Koch Foundation.

New Criminal Law Review, Vol. 24, Number 2, pps 156–211. ISSN 1933-4192, electronic ISSN 1933-4206. © 2021 by The Regents of the University of California. All rights reserved. Please direct all requests for permission to photocopy or reproduce article content through the University of California Press's Reprints and Permissions web page, <https://www.ucpress.edu/journals/reprints-permissions>. DOI: <https://doi.org/10.1525/nclr.2021.24.2.156>.

INTRODUCTION

Risk assessment is trending in the criminal justice system as a means to be smarter on crime.¹ The new wave of risk assessment is part of the evidence-based practices movement that relies upon science to identify strong predictors of criminal justice failure.² Basically, a tool developer uses a training sample to test for statistically significant factors to package into a final algorithmic prediction model.³ This form of algorithmic governance⁴ has become popular for driving how officials triage their populations to manage them accordingly.⁵

Proponents view risk assessment as a systematic, transparent, and objective system to predict future offending in order to reduce mass incarceration and without sacrificing public safety.⁶ The algorithmic turn has reached pretrial systems to respond to the significant contribution that pretrial detention numbers make to the excessive incarceration rates in the country,⁷ as well as serving as a recent reform movement to reduce reliance upon money bail.⁸

1. Richard Berk, *Accuracy and Fairness for Juvenile Justice Risk Assessments*, 16 J. EMPIR. LEG. STUD. 175, 175 (2019); Malcolm M. Feeley, *How to Think About Criminal Court Reform*, 98 B.U. L. REV. 673, 690 (2018) (“Algorithms are the new silver bullet of criminal justice reform.”).

2. Melissa Hamilton, *Debating Algorithmic Fairness*, 52 U.C. DAVIS L. REV. ONLINE 261, 266 (2019); Yoav Mehozay & Eran Fisher, *The Epistemology of Algorithmic Risk Assessment and the Path Towards a Non-Penology Penology*, 21 PUNISHMENT & SOC’Y 523, 531 (2019). “Evidence-based practice . . . is any correctional practice or intervention whose effectiveness at achieving its stated goal is supported by ‘empirically sound’ research of some kind.” Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 556 (2015).

3. Joanna Amirault & Patrick Lussier, *Population Heterogeneity, State Dependence and Sexual Offender Recidivism: The Aging Process and the Lost Predictive Impact of Prior Criminal Charges over Time*, 39 J. CRIM. JUST. 344, 344 (2011).

4. Kelly Hannah-Moffat, *Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates*, 23 THEORETICAL CRIMINOLOGY 453, 454 (2019).

5. John Monahan, *Preface: Recidivism Risk Assessment in the 21st Century*, in HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS xxiii, xxiii (Jay P. Singh et al. eds., 2018).

6. Megan Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 303, 305 (2018).

7. See *infra* notes 31–35 and accompanying text.

8. Shima Baradaran Baughman, *Dividing Bail Reform*, 105 IOWA L. REV. 947, 949 (2020).

Jurisdictions are scrambling to become part of the algorithmic risk generation to differentiate high-risk offenders, for whom incarceration may be merited, from low-risk offenders, who might be safely released.⁹ Many agencies are taking an off-the-shelf approach by adopting an available risk instrument and implementing it, but without empirically testing the tool's ability to function adequately within their own populations.¹⁰ This practice is unsound because a supposedly global risk assessment tool is not necessarily generalizable to other jurisdictions.¹¹ A tool with considerable error rates will affirmatively cause harms when its errors lead to release outcomes that then fail to protect the public, incarcerate unnecessarily those who do not pose a danger, and otherwise waste precious resources.

Critics are thereby urging independent researchers to audit risk tool capabilities to provide stakeholders with needed information on how well risk assessment tools actually perform in the field.¹² This article answers this call by offering a cross-validation study that analyzes how well the most popular algorithmic risk tool designed for pretrial settings performs in its predictive competencies. The Public Safety Assessment (PSA) is promoted by its owners as a national tool and is currently employed by court systems across the United States.¹³ Little is known about the predictive utility of the PSA as the owner has declined to release information on the instrument's developmental methodologies or its accuracy in the training samples.¹⁴ We seek to improve upon the knowledge base generally by analyzing this popular risk tool in three diverse settings and offering multiple performance metrics. The results complement the field by providing policymakers, criminal justice officials, and other stakeholders insight into

9. Sandra G. Mayson, *Dangerous Defendants*, 127 YALE L.J. 490, 494 (2018).

10. Robert Werth, *Risk and Punishment: The Recent History and Uncertain Future of Actuarial, Algorithmic, and "Evidence-Based" Penal Techniques*, 13 SOC. COMPASS 1, 9 (2019).

11. James T. McCafferty, *The Importance of Counties: Examining the Predictive Validity of a State Juvenile Risk Assessment Instrument*, 55 J. OFFENDER REHABILITATION 377, 378 (2016).

12. Alicia Solow-Niederman et al., *The Institutional Life of Algorithmic Risk Assessment*, 34 BERKELEY TECH. L.J. 705, 742 (2019); PARTNERSHIP ON AI, REPORT ON ALGORITHMIC RISK ASSESSMENT TOOLS IN THE U.S. CRIMINAL JUSTICE SYSTEM 31 (2019), <https://www.partnershiponai.org/wp-content/uploads/2019/04/Report-on-Algorithmic-Risk-Assessment-Tools.pdf>.

13. Kristian Lum et al., *The Impact of Overbooking on a Pre-Trial Risk Assessment Tool*, ARXIV (Jan. 23, 2020), <https://arxiv.org/pdf/2001.08793>.

14. *Infra* Section I.A.

the utility of a purportedly national algorithmic tool in practice and clarifying whether it can function consistently across settings.

This article proceeds as follows. Section I provides an overview of how algorithmic risk tools theoretically serve the pretrial justice reform movement and contextualizes why these goals may only be met if satisfactory accuracy rates are achieved. The text outlines various empirical and cultural reasons that a purportedly national tool may be weakly generalizable in practice. Then we outline the scientific, legal, and practical values in cross-validating any risk tool in the jurisdictions in which it is implemented to verify its level of accuracy. Nevertheless, the expectation for predictive ability across sites is nowhere near perfection for a tool to affirmatively offer some value. As a famous British statistician once remarked: “Essentially, all models are wrong, but some are useful.”¹⁵

Section II discusses the background of the cross-validation study. The PSA is a popular risk instrument used in pretrial settings across the country. Section II introduces the three datasets employed in the study, which derive from three jurisdictions: Illinois, Kentucky, and New Mexico. The demographic and geographic diversity of these sites keenly serves to challenge the operative ability of the tool across the country. The total sample size is 215,738 cases, which is large enough to allow for robust findings on the utility of the PSA. Three research questions structure the study: Does the PSA meaningfully classify individuals into risk levels (low, medium, high) according to their likelihood of failure? What are the PSA’s predictive accuracy rates? Does the PSA perform consistently across sites? Section II outlines the empirical methodologies designed to vigorously respond to these questions.

Section III presents the results of the analyses to respond to the research questions. Research findings explore multiple performance metrics to judge various abilities of the tool to classify individuals into risk levels and ascertain accuracy rates. The comparative analyses are reviewed to evaluate whether the tool is as generalizable as its developers claim. Policy implications of the results provide further context for how the information points elicited from the study may be generally useful to stakeholders who are interested in employing an algorithmic risk tool to maximize returns. Conclusions follow.

15. Ron Wasserstein, *George Box: A Model Statistician*, 7 SIGNIFICANCE 134, 134, (2010).

This study offers some novel contributions to the literature. Whereas most research in the risk assessment literature depends upon a single sample of convicted prisoners released from prison,¹⁶ this study employs three samples from varied jurisdictions of pretrial defendants discharged from local jails. At present, the few research reports about the performance of pretrial risk tools that provide any data about their predictive powers typically offer merely one or two statistical tests.¹⁷ This study augments the field with multiple classification and predictive performance measures in order to provide more dynamic analyses toward appraising the effectiveness of the PSA. This research operates as a case study, as well, with implications beyond pretrial practices by engaging policy issues to inform on strategies to improve practices when a tool offers some utility, yet is plagued by critical limitations in its capabilities. Fortunately, algorithmic risk tools and the practices surrounding their use are generally retrainable and adaptable.

I. PRETRIAL RISK ASSESSMENT

Criminal justice officials are keen to properly manage their correctional populations to strike a reasonable balance among interests in public safety, protecting individual rights, and the efficient use of limited resources.¹⁸ In recent decades, correctional management draws on an evolving set of risk assessment practices.¹⁹ Risk assessment here involves predicting an

16. Synøve Nygaard Andersen & Torbjørn Skardhamar, *Pick a Number: Mapping Recidivism Measures and Their Consequences*, 63 CRIME & DELINQ. 613, 617 (2017).

17. KiDeuk Kim & 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); T. Douglas et al., *Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data*, 42 EUR. PSYCHIATRY 134, 135 (2017); Jay P. Singh, *Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review*, 31 BEHAV. SCI. & L. 55, 61 (2013).

18. 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, 788–89 (2019); Michael L. Rich, *Limits on the Perfect Preventive State*, 46 CONN. L. REV. 883, 932–33 (2014); Singh, *supra* note 17, at 55.

19. Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 232 (2015), <http://epubs.surrey.ac.uk/id/eprint/842342>. See also Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. 439, 441 (2020) (referencing a “risk-assessment revolution in criminal justice” in America).

individual's potential for a negative criminal justice outcome, such as flight from justice or reoffending.²⁰

Predictions of future dangerousness have long been a part of criminal justice decision making to ameliorate harms.²¹ Historically, risk predictions have been based on intuition or the personal experience of the official responsible for making the relevant decision.²² With technological advances permitting the collection and processing of large datasets, risk tools offer criminal justice officials predicted outcomes informed by big data.²³ Contemporary risk tools use algorithms to sort individuals into categories based on their likelihood for failure.²⁴ An algorithm is simply an equation in which inputs are received and processed to produce a quantitatively derived output.²⁵ Tool developers package their selected risk factors and weight them according to their predictive strength, which then becomes the algorithm.²⁶

The new algorithmic risk wave has been heavily oriented to back-end decisions, such as sentencing and post-incarceration supervision.²⁷ More recent attention considers the potential benefits that algorithmic risk assessment practices provide in pretrial settings,²⁸ though its use in a front-end context is also the most controversial as a result.²⁹

20. Garrett & Monahan, *supra* note 19, at 450.

21. Jordan M. Hyatt et al., *Reform in Motion: The Promise and Perils of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing*, 49 DUQ. L. REV. 707, 724–25 (2011).

22. Klingele, *supra* note 2, at 556.

23. Ignacio Cofone, *Algorithmic Discrimination is an Information Problem*, 70 HASTINGS L.J. 1389, 1400 (2019); Hannah-Moffat, *supra* note 4, at 455; Feeley, *supra* note 1, at 689.

24. Alice Ristroph, *Criminal Law as Public Ordering*, 70 U. TORONTO L.J. 64, 81 (2020); Frank Pasquale & Glyn Cashwell, *Prediction, Persuasion, and the Jurisprudence of Behaviourism*, 68 U. TORONTO L.J. 63, 63 (2018); Kelly Hannah-Moffat et al., *Negotiated Risk: Actuarial Illusions and Discretion in Probation*, 24 CANADIAN J.L. & SOC'Y 391, 396 (2009).

25. Reuben Binns, *Algorithmic Accountability and Public Reason*, 31 PHIL. & TECH. 543, 545 (2018). See also OSONDE A. OSOBA ET AL., ALGORITHMIC EQUITY: A FRAMEWORK FOR SOCIAL APPLICATIONS, RAND ix (2019), https://www.rand.org/content/dam/rand/pubs/research_reports/RR2700/RR2708/RAND_RR2708.pdf (characterizing algorithms as “advanced decisionmaking or decision support artifacts that exhibit ‘intelligent behavior’”).

26. Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 CRIM. JUST. & BEHAV. 185, 186 (2019).

27. Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 67 (2017).

28. Kristin Bechtel et al., *A Meta-Analytic Review of Pretrial Research: Risk Assessment, Bond Type, and Interventions*, 42 AM. J. CRIM. JUST. 443, 444 (2017).

29. Jennifer L. Skeem & Christopher T. Lowenkamp, *Race, Risk, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 682 (2016).

A. Pretrial Detention Decisions

Critical attention toward reforming pretrial justice is thriving.³⁰ An important reason is that pretrial detention is a primary driver of mass incarceration. “Pretrial detention is responsible for all of the net jail growth in the last twenty years.”³¹ Three-quarters of jail inmates have not been convicted of any crime.³² This translates to over 500,000 individuals at any time residing in jails around the country awaiting court action over pending charges.³³ America is an unfortunate leader in the use of pretrial detention. The United States represents 4 percent of the world’s population yet houses 20 percent of the world’s pretrial jail population.³⁴ The American pretrial incarceration rate far exceeds that of any European country.³⁵

Detaining the unconvicted appears unfair,³⁶ but it is lawful in some circumstances. Criminal procedure rules generally permit judges to order pretrial detention based on an estimation that, if released, the defendant

30. SARAH L. DESMARAI & EVAN M. LOWDER, PRETRIAL RISK ASSESSMENT TOOLS: A PRIMER FOR JUDGES, PROSECUTORS, AND DEFENSE ATTORNEYS 3 (2019), <http://www.safetyandjusticechallenge.org/wp-content/uploads/2019/02/Pretrial-Risk-Assessment-Primer-February-2019.pdf>; Paul Heaton et al., *The Downstream Consequences of Misdemeanor Pretrial Detention*, 69 STAN. L. REV. 711, 718 (2017); Bechtel et al., *supra* note 28, at 444.

31. WENDY SAWYER & PETER WAGNER, PRISON POLICY INITIATIVE, MASS INCARCERATION: THE WHOLE PIE 2020 (2020), <https://www.prisonpolicy.org/reports/pie2020.html>.

32. *Id.*

33. Joe Russo et al., *Data-Informed Jails: Challenges and Opportunities*, RAND 3 (2020), https://www.rand.org/content/dam/rand/pubs/research_reports/RRAI00/RAIA08-1/RAND_RRAI08-1.pdf; SAWYER & WAGNER, *supra* note 31.

34. COLIN DOYLE ET AL., BAIL REFORM: A GUIDE FOR STATE AND LOCAL POLICYMAKERS 7 (2019), <https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=9a804d1d-f9be-e0f0-b7cd-cf487ec70339&forceDialog=0>.

35. ROY WALMSLEY, WORLD PRISON BRIEF, WORLD PRE-TRIAL/REMAND IMPRISONMENT LIST 6 tbl. 1, II-14 tbl. 4 (4th ed. 2020), https://www.prisonstudies.org/sites/default/files/resources/downloads/world_pre-trial_list_4th_edn_final.pdf.

36. “Detaining a person before he is found guilty of a crime is a particularly draconian decision for the state to make. Pretrial detention is one of the most severe things that can happen to a person: the detainee immediately loses his freedom, and can also lose his family, health, home, job, and community ties.” OPEN SOC’Y JUST. INITIATIVE, PRESUMPTION OF GUILT: THE GLOBAL OVERUSE OF PRETRIAL DETENTION 7 (2014), <https://www.justiceinitiative.org/uploads/de4c18f8-ccc1-4eba-9374-e5c850a07efd/presumption-guilt-09032014.pdf>.

will abscond or threaten the public by committing a new offense.³⁷ Judges tend to err on the side of caution and order detention, thereby boosting the numbers of individuals incarcerated.³⁸

Even if a judge determines that release is justified, it may be contingent upon posting a financial bond as a means of securing the individual's return for court dates without new arrest charges.³⁹ The amount of bond is not always tied to an estimate of the individual's ability to pay it or to the likelihood of compliance with the law.⁴⁰ Instead, many jurisdictions use strict bail schedules with the exact monetary amount tied to the seriousness of the crime(s) alleged.⁴¹ These use the pending criminal charge as a proxy for the risk of flight or rearrest, with the monies meant to mitigate such risk.⁴² Consequently, many defendants otherwise eligible for release remain behind bars simply because they cannot afford the bond.⁴³ This scenario subverts the foundations of risk management in that individuals who remain incarcerated because they cannot post bail may present a lower risk profile than those who are released.⁴⁴ In turn, bail systems contribute significantly to jail population counts.⁴⁵

37. Garrett & Monahan, *supra* note 19, at 449; Doaa Abu Elyounes, *Bail or Jail? Judicial Versus Algorithmic Decision-Making in the Pretrial System*, 21 COLUM. SCI. & TECH. L. REV. 376, 379 (2020), <https://journals.library.columbia.edu/index.php/stlr/article/view/6838/3595>; 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, 738 (2018). The Supreme Court has officially approved this form of preventive detention based on judicial guestimates. *United States v. Salerno*, 481 U.S. 739, 746–51 (1987).

38. Eckhouse et al., *supra* note 26, at 186; Van Brunt & Bowman, *supra* note 37, at 737.

39. Van Brunt & Bowman, *supra* note 37, at 728 n. 132.

40. DOYLE ET AL., *supra* note 34, at 7.

41. John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 WASH. L. REV. 1725, 1744 (2018).

42. RAM SUBRAMANIAN ET AL., VERA INST. JUST., INCARCERATION'S FRONT DOOR: THE MISUSE OF JAILS IN AMERICA 32 (2015), https://www.vera.org/downloads/publications/incarcerations-front-door-report_02.pdf.

43. Garrett & Monahan, *supra* note 19, at 442; SARAH PICARD ET AL., CENTER FOR COURT INNOVATION, BEYOND THE ALGORITHM: PRETRIAL REFORM, RISK ASSESSMENT, AND RACIAL FAIRNESS 3 (2019), <https://www.courtinnovation.org/sites/default/files/media/document/2019/beyondthealgorithm.pdf>.

44. Emily Tiry et al., *Road Map to Pretrial Reforms: Profile Analysis of New Hampshire's Pretrial Defendant Population*, URBAN 2 (2016), https://www.urban.org/research/publication/road-map-pretrial-reforms/view/full_report.

45. Baughman, *supra* note 8, at 949.

The foregoing strains in the age of enduring mass incarceration have led to progressive reforms of pretrial practices across the fifty states.⁴⁶ The adoption of an algorithmic risk tool is a main foundation of this new wave of pretrial reform.⁴⁷ Indeed, the pretrial context is one of the two most popular new settings for algorithmic risk to operate today (along with sentencing).⁴⁸ The *Science Bench Book for Judges* from the National Judicial College conceptualizes risk assessment algorithms as the “most ubiquitous form of predictive technology used in the criminal pretrial context,” serving as a critical weapon in the judiciary’s “arsenal of justice.”⁴⁹

Reformers advocate that the adoption of algorithmic risk tools offers several advantages, with some overlap between these theoretical prospects. First, proponents expect that reliance upon scientifically informed algorithms will reduce the use of detention and monetary bail to ensure appearance at trial and to protect communities.⁵⁰ Drawing on quantitative algorithmic tool outcomes, judges will be better able to isolate those who pose a high risk of failure, and thus may be more confident to increase release rates of the rest.⁵¹ A sort of formula is suggested here: “(1) Maximize public safety and (2) Maximize court appearance while (3) Maximizing release from custody.”⁵² A related value is that scientific research indicates

46. Koepke & Robinson, *supra* note 41, at 1747.

47. Mitali Nagrecha et al., *Court Culture and Criminal Law Reform*, 69 DUKE L.J. ONLINE (2020), <https://dlj.law.duke.edu/2020/04/courtculture/>; Koepke & Robinson, *supra* note 41, at 1729.

48. CHRISTOPHER BAVITZ ET AL., ASSESSING THE ASSESSMENTS: LESSONS FROM EARLY STATE EXPERIENCES IN THE PROCUREMENT AND IMPLEMENTATION OF RISK ASSESSMENT TOOLS I (2018), <http://nrs.harvard.edu/urn-3:HUL.InstRepos:37883502>.

49. Veronica Alicea-Galvan, *Pre-Trial Criminal*, in SCIENCE BENCH BOOK FOR JUDGES 201, 205 (Peggy Hora et al. eds., 2020), <https://www.judges.org/wp-content/uploads/Chapter6-SBB.pdf>.

50. Glen J. Dalakian, *Open the Jail Cell Doors, HAL: A Guarded Embrace of Pretrial Risk Assessment Instruments*, 87 FORDHAM L. REV. 325, 341 (2018); Koepke & Robinson, *supra* note 41, at 1729.

51. PICARD ET AL., *supra* note 43, at 3; Eckhouse et al., *supra* note 26, at 202; Mayson, *supra* note 9, at 515; Megan Stevenson & Sandra G. Mayson, *Pretrial Detention and Bail*, in REFORMING CRIMINAL JUSTICE: A REPORT BY THE ACADEMY FOR JUSTICE 21, 23 (Erik Luna ed., 2017), https://law.asu.edu/sites/default/files/pdf/academy_for_justice/2_Reforming-Criminal-Justice_Vol_3_Pretrial-Detention-and-Bail.pdf.

52. MICHAEL R. JONES, PRETRIAL JUST. INST., UNSECURED BONDS: THE MOST EFFECTIVE AND EFFICIENT PRETRIAL RELEASE OPTION 4 (2013), <https://www.ncjrs.gov/App/AbstractDB/AbstractDBDetails.aspx?id=269164>.

that the pretrial release of low-risk individuals measurably improves their chances of staying crime-free.⁵³

Second, algorithms may ameliorate errors produced by subjective human intuitions.⁵⁴ Historically, judgements on future dangerousness have been based on the gut instinct or the personal experience of the official responsible for making the relevant management decision.⁵⁵ Critics challenge those sorts of predictions as suffering from implicit and explicit human biases.⁵⁶ Plus, detention decision making may rely upon oversimplified heuristics whereas judges in a pretrial context typically have few interactions with the individuals, gain limited access to relevant information, and must make on-the-spot calls with scant time for reflection.⁵⁷ The fact that defendants are often unrepresented at the stage of the proceedings in which pretrial detention decisions are made means that counsel is

53. Marie VanNostrand & Gena Keebler, *Pretrial Risk Assessment in Federal Court*, 73(2) FED. PROB. 3, 6 (2009).

54. Elyounes, *supra* note 37; Jeffrey C. Singer et al., *A Convergent Approach to Sex Offender Risk Assessment*, in THE WILEY-BLACKWELL HANDBOOK OF LEGAL AND ETHICAL ASPECTS OF SEX OFFENDER TREATMENT AND MANAGEMENT 341, 343 (Karen Harrison & Bernadette Rainey eds., 2013).

55. Klingele, *supra* note 2, at 556.

56. PRETRIAL JUST. INST., PRETRIAL RISK ASSESSMENT CAN PRODUCE RACE-NEUTRAL RESULTS 2 (2017), https://www.ncsc.org/__data/assets/pdf_file/0012/1650/pretrial-risk-assessment-can-produce-race-neutral-results-pji-2017.ashx.pdf; Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 242 STAN. TECH. L. REV. 242, 267 (2019), <https://law.stanford.edu/publications/developing-artificially-intelligent-justice-stanford-technology-law-review/>; Eckhouse et al., *supra* note 26, at 202; Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 U.C. DAVIS L. REV. 1067, 1070 (2018); Stephanie Bornstein, *Antidiscriminatory Algorithms*, 70 ALA. L. REV. 519, 523 (2018); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 116 (2017). See also Chelsea Barabas, *Beyond Bias: Re-Imagining the Terms of "Ethical AI" in Criminal Law*, 12 GEO. J.L. MOD. CRITICAL RACE PERSP. 84, 95 (2020) ("implicit bias is understood as a phenomenon which 'is somehow apart from us yet can infect our decision-making . . . as opposed to something that is variously, but systematically cultivated and maintained'") (citing Anna Lauren Hoffman, *Where Fairness Fails: On Data, Algorithms, and the Limits of Antidiscrimination Discourse*, 22 INFO. COMMUN. & SOC'Y 900 (2019)).

57. Hafsa S. Mansoor, *Guilty Until Proven Guilty: Effective Bail Reform as a Human Rights Imperative*, DEPAUL L. REV. (forthcoming 2021); David Arnold et al., *Racial Bias in Bail Decisions*, 133 Q. J. ECON. 1885, 1922 (2018). The speedy nature of pretrial decision-making is not solely a negative attribute. Quick decisions may lead to earlier releases, thereby reducing the negative consequences of pretrial detention. Paul Heaton, *The Expansive Reach of Pretrial Detention*, 98 N.C. L. REV. 369, 375 (2020).

typically unable to offer contextualizing data to counter these biases and heuristics.⁵⁸ Thus, algorithmic risk instruments are seen as far more objective information-processing devices.⁵⁹

Third, algorithmic risk tools are often lauded for reducing the arbitrariness of pretrial release decisions.⁶⁰ Human judgements may vary for reasons that algorithms cannot, such as by the person being tired, hungry, or distracted.⁶¹ Algorithms are not intended to entirely usurp human involvement in criminal justice outcomes.⁶² Still, algorithms can aid by providing standardization and helping to structure the human aspect of decision making.⁶³ A related advantage is that the use of systematic risk tools makes release decisions more defensible in legal and ethical terms.⁶⁴ For instance, a risk tool that specifically predicts failure to appear would seem fit for purpose when applied in a jurisdiction that strives to use it to aid in

58. Baughman, *supra* note 8, at 962; Mansoor, *supra* note 57.

59. Jessica Gabel Cino, *Deploying the Secret Police: The Use of Algorithms in the Criminal Justice System*, 34 GA. ST. U. L. REV. 1073, 1078 (2018); Faye S. Taxman, *The Partially Clothed Emperor: Evidence-Based Practices*, 34 J. CONTEMP. CRIM. JUST. 97, 99 (2018).

60. Monika Zalnieriute et al., *From Rule of Law to Statute Drafting: Legal Issues for Algorithms in Government Decision-Making*, in THE CAMBRIDGE HANDBOOK ON THE LAW OF ALGORITHMS 251 (Woodrow Barfield ed., 2020), https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3380072_code722134.pdf?abstractid=3380072&mirid=1; Re & Solow-Niederman, *supra* note 56, at 267; Eckhouse et al., *supra* note 26, at 202; Simmons, *supra* note 56, at 1070. *See also* PARTNERSHIP ON AI, *supra* note 12, at 7 (indicating criminal justice risk tools “provide rigor and reproducibility for life-critical decisions”).

61. R. Barry Ruback et al., *Communicating Risk Information at Criminal Sentencing in Pennsylvania: An Experimental Analysis*, 80 FED. PROB. 47, 47 (2016). *See also* Ozkan Eren & Maci Mocan, *Emotional Judges and Unlucky Juveniles*, 10 ECON. J. APPLIED ECON. 171, 194 (2018) (empirical study finding judges issued longer sentences to juvenile defendants the week after a popular college football team experienced an unexpected loss, with increased severity from judges who were alumni of the college).

62. Reuben Binns, *Algorithmic Decision-making: A Guide for Lawyers*, 25 JUD. REV. 2, 6 (2020); Simmons, *supra* note 56, at 1101.

63. Re & Solow-Niederman, *supra* note 56, at 246; Eckhouse et al., *supra* note 26, at 202; Taxman, *supra* note 59, at 99. *See also* Faye S. Taxman & Amy DeZemmer, *The Value and Importance of Risk and Need Assessment (RNA) in Corrections & Sentencing: An Overview of the Handbook*, in HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 22, 23 (Faye S. Taxman ed., 2017) (“Standardized [risk] tools, in theory, offer a science-based approach to regulate decision-making to avoid or minimize biases, decrease unnecessary discretion, improve proper use of resources, and/or increase fairness.”).

64. Brian P Schaefer & Tom Hughes, *Examining Judicial Pretrial Release Decisions: The Influence of Risk Assessments and Race*, 20 CRIMINOLOGY CRIM. JUST. L. & SOC’Y 47, 49 (2019); Tiry et al., *supra* note 44, at 2; Hannah-Moffat et al., *supra* note 24, at 400.

determining whether imposing a cash bail condition is necessary to ensure court date compliance.⁶⁵ The tool thereby makes the requirement of a financial bond more legally justifiable from a wealth disparity perspective and thus seemingly more rational than an impersonal bail schedule.⁶⁶

Fourth, the adoption of a risk tool in pretrial decision making may save significant monies. Pretrial incarceration is expensive.⁶⁷ Detaining those pending trial costs over \$13 billion annually in the United States.⁶⁸ If risk tools lead to increasing the rate of pretrial release, states can curtail unnecessary costs of incarceration⁶⁹ without endangering the public.⁷⁰ The time may be ripe for algorithmic risk reforms as more limited use of pretrial detention is enjoying broad public support in the United States.⁷¹ The political climate is ideal, too, with bipartisan support to pare jail populations because of the resulting burdensome institutional costs and the weighty financial harms to individuals, families, and communities.⁷² Notably, the foundational cross-party consensus is not a moral judgement decrying the legitimacy of incarcerating large numbers of people presumed innocent, but is tied to the “virtue of frugality” in that cutbacks to the use of imprisonment lead to more conservative resource allocations.⁷³ Other forms of savings from risk tools are possible. Algorithmic risk tools make pretrial detention decision making more efficient in process and time.⁷⁴ Plus, risk tool adoption could shield jurisdictions from the stresses and

65. Cecelia Klingele, *Making Sense of Risk*, 38 BEHAV. SCI. & L. 218, 223–24 (2020).

66. *Id.* at 223–24.

67. Brandon L. Garrett, *Evidence-Informed Criminal Justice*, 86 GEO. WASH. L. REV. 1490, 1506 (2018).

68. Schaefer & Hughes, *supra* note 64, at 48. *See also* DOYLE ET AL., *supra* note 34, at 8.

69. Michael Baglivio et al., *The Usefulness of a General Risk Assessment, the Static Risk Assessment (SRA), in Predicting Pretrial Failure: Examining Predictive Ability Across Gender and Race*, 3 JUST. EVAL. J. 1, <https://doi.org/10.1080/24751979.2019.1668242>.

70. Stevenson, *supra* note 6, at 305.

71. WILL DOBBIE & CRYSTAL YANG, THE HAMILTON PROJECT, PROPOSALS FOR IMPROVING THE U.S. PRETRIAL SYSTEM 7 (2019), https://www.hamiltonproject.org/assets/files/DobbieYang_PP_20190319.pdf.

72. Dalakian, *supra* note 50, at 336.

73. Klingele, *supra* note 2, at 565.

74. PARTNERSHIP ON AI, *supra* note 12, at 7; Simmons, *supra* note 56, at 1074; Stephen D. Gottfredson & Laura J. Moriarity, *Statistical Risk Assessment: Old Problems and New Applications*, 52 CRIME & DELINQ. 178, 179 (2006).

expenses from recent civil rights litigation challenging practices relying heavily on money bail as the key to freedom.⁷⁵

There is also the potential for considerable accuracy rates for algorithmic prediction,⁷⁶ which is discussed in more detail below. With the foregoing advantages risk tools potentially offer, their use now is proclaimed to represent industry best practice.⁷⁷ As a consequence, the American Bar Association and the National Association of Pretrial Services Agencies support the implementation of risk assessment tools in pretrial to facilitate releases and reduce reliance upon money bail as a means of assuring defendants' appearances for scheduled court dates.⁷⁸ The National Association of Counties, the Conference of State Court Administrators, and the Conference of Chief Judges likewise promote pretrial risk assessment tools.⁷⁹ The rise in the use of algorithmic tools is evident. The Pretrial Justice Institute estimates that one quarter of Americans reside in jurisdictions that use an evidence-based risk tool in a pretrial context.⁸⁰

Despite these potential improvements in pretrial justice and the buy-in of so many stakeholder groups, a critical question remains open. Does scientific research confirm that algorithmic tools are sufficiently accurate to serve their intended purposes? Without confirmation of some minimally acceptable level of accuracy, the algorithmic enterprise constitutes a policy hype without concrete substance.

Regrettably, the deployment of algorithmic risk tools is outpacing the empirical evidence to substantiate their abilities.⁸¹ The Vera Institute recently warned that the lack of rigorous research on the use of risk tools

75. See DOYLE ET AL., *supra* note 34, at 8.

76. DOBBIE & YANG, *supra* note 71, at 7; DESMARAIS & LOWDER, *supra* note 30, at 4; Dalakian, *supra* note 50, at 332.

77. Mehozay & Fisher, *supra* note 2, at 531; Bechtel et al., *supra* note 28, at 446; Hannah-Moffat et al., *supra* note 24, at 397.

78. TODD SHAUGHNESSY ET AL., REPORT TO THE UTAH JUDICIAL COUNCIL ON PRETRIAL RELEASE AND SUPERVISION PRACTICES 16 (2015).

79. Garrett & Monahan, *supra* note 19, at 452.

80. PRETRIAL JUST. INST., THE STATE OF PRETRIAL JUSTICE IN AMERICA 13 (2017), <https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=484affbc-d944-5abb-535f-b171d091a3c8&forceDialog=1>.

81. Seth J. Prins & Adam Reich, *Can We Avoid Reductionism in Risk Reduction?*, 22 THEORETICAL CRIMINOLOGY 258, 259 (2018).

in pretrial settings signals that their practical value has yet to be proven.⁸² The fact that risk tool developers may be satisfied with their tools does not oblige criminal justice officials to find the science is sufficiently ready for real-world use. “While researchers may accept a moderate or less performance for a tool, for the practice of justice this should not be accepted or tolerated since it diminishes the value of the . . . tool.”⁸³ A weak or highly inaccurate tool offers little utility, and the resulting waste of resources may further overwhelm any benefits the tool may otherwise offer.⁸⁴ Inaccuracies undermine effective population management by leading to the release of those who breach the law, while unnecessarily incarcerating individuals likely to have been compliant.⁸⁵ These types of errors undermine the basic credibility of the criminal justice system,⁸⁶ particularly if they are avoidable.

The urgency that risk assessment practices be sound is particularly salient in this time of mass incarceration.⁸⁷ Politicians, regulators, legal practitioners, private foundations, and academics are thereby raising concerns, enquiring about how accurate these tools actually are in real-world settings.⁸⁸

82. Léon Digard & Elizabeth Swavola, *Justice Denied: The Harmful and Lasting Effects of Pretrial Detention*, VERA 9 (2019), <https://www.vera.org/downloads/publications/Justice-Denied-Evidence-Brief.pdf>.

83. Taxman & Dezember, *supra* note 63, at 37. 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 157, 173 (Faye S. Taxman ed., 2017) (noting the levels of accuracy that have been considered acceptable for recidivist risk assessment are noticeably weaker than expected and achieved in other disciplines).

84. Berk, *supra* note 1, at 175.

85. Paul Hayes et al., *Algorithms and Values in Justice and Security*, 35 AI & SOC. 533–55 (2020), <https://doi.org/10.1007/s00146-019-00932-9>.

86. Eaglin, *supra* note 27, at 92.

87. Garrett & Monahan, *supra* note 19, at 446.

88. Brandon L. Garrett & Megan Stevenson, *Open Risk Assessment*, 38 BEHAV. SCI. & L. 279, 280–81 (2020); Harry Surden, *Ethics of AI in Law: Basic Questions*, in THE OXFORD HANDBOOK OF ETHICS OF AI 719 (Markus D. Dubber et al. eds., 2020), https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3441303_code709715.pdf?abstractid=3441303&mirid=1; BAVITZ ET AL., *supra* note 48, at 3; Feeley, *supra* note 1, at 689; Cino, *supra* note 59, at 1079; Eaglin, *supra* note 27, at 106. See also Binns, *supra* note 25, at 546–47 (enquiring whether algorithmic models are “useful guesswork”).

B. Risk Tool Validation

In the forensic sciences field, the concept of tool accuracy is often discussed in conjunction with the term “validation.” Validation involves an empirical review of whether the forensic tool functions as intended.⁸⁹ Validation connotes acceptance and confidence in the tool.⁹⁰

Because of the high-stakes nature of criminal justice outcomes, an assumption that an algorithmic risk tool will have a sufficient degree of validity is unwarranted and naive.⁹¹ One reason to question the ability to achieve an acceptable level of accuracy is related to the risk assessment enterprise’s essential assumption that criminal behavior is innately predictable.⁹² This supposition contrasts with the philosophy of the criminal law system in common law countries, which has always endorsed the concept of human free will.⁹³ People are fundamentally unreliable beings. Human behavior is hard to predict as humans are active, interactive, reactive, and adaptive creatures.⁹⁴ There is absolutely no certainty in whether a person will or will not commit a future act.⁹⁵ No matter how sophisticated an algorithm, it can never be anywhere near perfect when forecasting human compliance with the law. Hence, it is reasonable in the first instance to expect that a risk tool will be burdened with error rates.

Another basis for challenge is that relatively few validation studies exist (at least in the public realm) to allow for a sufficient knowledge base to confirm that algorithmic tools offer a considerable degree of accuracy.⁹⁶ Of

89. Natalie Ram, *Innovating Criminal Justice*, 112 NW. U. L. REV. 659, 688 (2018).

90. Zachary Hamilton et al., *Customizing Criminal Justice Assessments*, in HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 536, 567 (Faye S. Taxman ed., 2017).

91. Hayes et al., *supra* note 85.

92. Hannah-Moffat et al., *supra* note 24, at 395.

93. Stephen Schafer, *The Problem of Free Will in Criminology*, 67 J. CRIM. L. & CRIMINOLOGY 481, 481–82 (1976) (citing IMMANUEL KANT, KRITIK DER PRAKTISCHEN JERUNET (1788)).

94. See Mehozay & Fisher, *supra* note 2, at 537. See also Hess & Turner, *supra* note 83, at 174 (“Humans are agents, not molecules; they use strategy to pursue goals and express their social identities.”).

95. Klingele, *supra* note 65, at 223; Klingele, *supra* note 2, at 561.

96. Garrett & Stevenson, *supra* note 88, at 280; Koepke & Robinson, *supra* note 41, at 1795; William D. Burrell, *Risk and Needs Assessment in Probation and Parole: The Persistent Gap Between Promise and Practice*, in HANDBOOK ON RISK AND NEED ASSESSMENT: THEORY AND PRACTICE 54, 80 (Faye S. Taxman ed., 2017).

the studies that are available, results tend to represent the particular tool's performance on the sample for which it was originally trained.⁹⁷ But a tool tends to function at its best with its training sample because developers often actively reconfigure their algorithms to improve accuracy rates.⁹⁸ The tool can thus become too specialized on its training data, resulting in overfitting.⁹⁹ "Overfitting is a statistical problem that is analogous to learning the answer to all the questions on an exam by heart, without having actually understood the true principles that made them correct."¹⁰⁰ Consequently, shrinkage in being less precise is quite likely when the tool is used on new samples, thereby exhibiting higher error rates.¹⁰¹

It is concerning when a jurisdiction adopts a tool trained on an external sample as this "off-the-shelf" tool may perform quite poorly on its own population and thus be ineffective.¹⁰² One study of a few selected risk tools found evidence that each operated across sample sets with wide variations in predictive ability.¹⁰³ Otherwise, potential disparities in algorithmic risk accuracy remains an understudied topic in the forensic sciences.¹⁰⁴

97. Taxman, *supra* note 59, at 100; Douglas et al., *supra* note 17, at 135.

98. Sarah L. Desmarais, *Risk Assessment in the Age of Evidence-Based Practice and Policy*, 16 INT'L J. FORENSIC MENTAL HEALTH 18, 19 (2017). "Instruments are likely to produce better results under closely supervised research conditions than in regular clinical practice, and that in studies in which an instrument is first developed, its predictive accuracy is optimized and is inevitably less when repeated in an independent sample." George Szmukler & Nikolas Rose, *Risk Assessment in Mental Health Care: Values and Costs*, 33 BEHAV. SCI. & L. 125, 130 (2013).

99. Binns, *supra* note 25, at 546; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PENN. L. REV. 633, 684 (2012).

100. PARTNERSHIP ON AI, *supra* note 12, at 13.

101. McCafferty, *supra* note 11, at 379. *See also* Gottfredson & Moriarity, *supra* note 74, at 185 ("Prediction methods are intended to estimate, based on some group of people available for study, how members of other similar groups will behave. There is a danger, however, of overestimating the extent to which relations found in one sample can be used to explain relations in a similar sample.").

102. Brian K. Lovins et al., *Validating the Ohio Risk Assessment System Community Supervision Tool with a Diverse Sample from Texas*, 3 CORRECTIONS 186, 188 (2018). *See also* BAVITZ ET AL., *supra* note 48, at 19 (noting government officials "should carefully consider whether validation studies performed in other jurisdictions will indicate predictive accuracy in their own: the fact that a tool is accurate elsewhere does not necessarily mean it will be predictive here").

103. Faye S. Taxman, *Risk Assessment: Where Do We Go From Here?*, in HANDBOOK OF RECIDIVISM RISK/NEEDS ASSESSMENT TOOLS 271, 274 (Jay P. Singh et al. eds., 2018).

104. Sarah L. Desmarais et al., *Performance of Recidivism Assessment Instruments in U.S. Correctional Settings*, in HANDBOOK OF RISK/NEEDS ASSESSMENT TOOLS 3, 15 (Jay P. Singh et al. eds., 2018); McCafferty, *supra* note 11, at 377.

Contextual considerations further explain why an algorithmic risk tool may not generalize well.¹⁰⁵ Datasets on criminal offending are often unique to their geographic and cultural locations.¹⁰⁶ Notably, “[a]nalytic methods can only work with the data given. Unmeasured heterogeneity across offenders, neighborhoods, and jurisdictions could possibly affect predictive accuracy.”¹⁰⁷ Absent the existence of homogeneous sample sets, there is a natural limit on a tool’s attainable level of mastery.¹⁰⁸ Predictive risk factors, even the more objective ones, are not necessarily universal.¹⁰⁹ Criminal acts and their correlates can vary depending on personal characteristics and experiences, times, geographies, sites, environments, and circumstances.¹¹⁰ To date, risk instruments do not incorporate any population drivers of crime as predictors (e.g., poverty rates, drug problems, social unrest) that might distinguish certain risk factors that are more relevant to localized areas with different crime problems.¹¹¹

Unfortunately, many jurisdictions do not take heed of these potential obstacles.¹¹² A recent national survey of agencies using an algorithm-led tool to inform pretrial decisions revealed that most of the jurisdictions had not ensured that the tool they were using was validated on their own

105. Cofone, *supra* note 23, at 1402 (“An algorithm can only be as good as the data that it is fed.”).

106. Desmarais et al., *supra* note 104, at 3–4. *See also* Eaglin, *supra* note 27, at 72 (“[D]ata only tells the story of the places from where it derives.”).

107. Hess & Turner, *supra* note 83, at 162. *See also* PARTNERSHIP ON AI, *supra* note 12, at 14 (“[J]ust because a tool uses data collected from the real world does not automatically make its findings valid.”)

108. Hess & Turner, *supra* note 83, at 178.

109. Arthur Rizer & Caleb Watney, *Artificial Intelligence Can Make Our Jail System More Efficient, Equitable, and Just*, 23 TEX. REV. L. & POL. 181, 192 (2019).

110. Keith Soothill, *Sex Offender Recidivism*, 39 CRIME & JUST. 145, 176 (2010). *See also* Prins & Reich, *supra* note 81, at 265 (“While the criminogenic risk patterns of individuals in a given area may predict who *among them* is more likely to engage in crime or be arrested (i.e. inter-individual variation), they almost certainly do not explain differences between different groups or areas (between-population variation) or over time.”) (emphasis in original).

111. *See* Michael F. Campagna et al., *The Contours of Assessment: Considering Aspects that Influence Prediction Performance*, CORRECTIONS POL’Y PRAC. & RES. (forthcoming 2021).

112. One reason for few validation studies in local jurisdictions is that pretrial practices often are performed at the county level, and their “[d]ata collection practices on pretrial outcomes . . . are notoriously varied, often haphazard, and sometimes totally absent.” Koepke & Robinson, *supra* note 41, at 1795.

populations, either preceding or following implementation.¹¹³ In practice, “the application of risk knowledge is often haphazard: jurisdictions frequently deploy pre-existing screening tools in settings for which they were neither designed nor calibrated.”¹¹⁴ The Wisconsin Supreme Court in *State v. Loomis*, the leading case on the topic to date, weighed in on this issue.¹¹⁵ The Wisconsin opinion approved the use of algorithmic risk assessment as informing sentencing decisions, but only when given to the judge with a written warning that the particular tool was a generic one whose predictive abilities had yet to be evaluated on a population of Wisconsin prisoners.¹¹⁶ The court further surmised that if a cross-validation study on Wisconsin were performed and were to show weak results, then the warning would require appropriate modification,¹¹⁷ implying that a judge should thus place less reliance upon the tool.

C. The Need for Cross-Validation

Officials should be more wary than they are when they look at implementing a tool not developed on their own populations, even if the tool is purportedly a generic one meant to be applicable across jurisdictions.¹¹⁸ Cross-validation of the tool in different samples and locations is crucial to determining whether an off-the-shelf risk tool is generalizable in practice and whether it offers an acceptable degree of proficiency.¹¹⁹ Experts call this a “global-local tension”: officials should not be swayed by a high global accuracy rate if the tool functions poorly in a specific local context.¹²⁰ Best practices require that any supposedly global tool be evaluated “with reference to the particular [local] conditions of application.”¹²¹ The issue is akin

113. Media Mobilizing Project, *Survey Summary: Pretrial Risk Assessments*, PRETRIAL RISK (2020), <https://pretrialrisk.com/our-research/interview-results/>.

114. Prins & Reich, *supra* note 81, at 260. See also Koepke & Robinson, *supra* note 41, at 1756 (likewise describing a “hazardous” exercise to use a tool developed on another jurisdiction’s data).

115. *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016), *cert. denied*, 137 S.Ct. 2290 (2017).

116. *Id.* at 769.

117. *Id.* at 770.

118. Bechtel et al., *supra* note 28, at 447.

119. Rizer & Watney, *supra* note 109, at 192; Cynthia S. Crowson et al., *Assessing Calibration of Prognostic Risk Scores*, 25 STAT. METHODS MED. RES. 1692, 1692 (2016).

120. Solow-Niederman et al., *supra* note 12, at 709.

121. *Id.*

to the legal requirement of “fitness” in terms of relevance in evidentiary considerations of the science to the issue, here being: Is there a correspondence between the population on which the tool was developed and the one to which the tool will be applied?¹²²

Cross-validation of an off-the-shelf risk instrument offers various advantages. At a high level, predictive accuracy studies serve the public interest in the accountability of criminal justice practices.¹²³ Only risk tools whose abilities are transparent and are confirmed with a sufficient degree of competency can foster the public’s perception in the fairness of detention decisions.¹²⁴

It is important, too, for more information about a tool’s capabilities to be made available to stakeholders to improve the competence of the decisions that the predictions are designed to advise.¹²⁵ Risk tools are often misunderstood, and thus officials may be either naively opposed or, instead, overly confident in the algorithm’s abilities.¹²⁶ Where cross-validation studies of a particular tool begin to demonstrate a sufficient level of accuracy across populations, that tool will encourage naysayers’ acceptance and concurrence with predictive outcomes.¹²⁷ Not only will a tool showing strong utility inspire greater confidence, such evidence may help deflect legal challenges.¹²⁸

122. Nicholas Scurich, *An Introduction to the Assessment of Violence Risk*, in INTERNATIONAL PERSPECTIVES ON VIOLENCE RISK ASSESSMENT 3, 10 (Jay P. Singh et al. eds., 2016). See also Hamilton et al., *supra* note 90, at 536 (“Because outcomes and populations of interest differ from one jurisdiction to another, taking an instrument ‘off-the-shelf’ and applying it to a new population is not recommended. It is like jamming a square peg down a round hole.”).

123. Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges*, 31 PHIL. & TECH. 611, 619 (2018).

124. Rizer & Watney, *supra* note 109, at 184. See generally PRETRIAL JUST. INST., *supra* note 80, at 13.

125. Klingele, *supra* note 65, at 219 (“[I]n order to use these tools in ways that promote equity and justice, end users of these tools need to possess a general understanding of how risk prediction tools work and what they do—and do not—reveal that has relevance to the legal decisions they make.”); BAVITZ ET AL., *supra* note 48, at 19.

126. Prins & Reich, *supra* note 81, at 259. See also Rhys Hester, *Risk Assessment Savvy: The Imperative of Appreciating Accuracy and Outcome*, 38 BEHAV. SCI. & L. 246, 248 (2020) (suggesting some judges will reject algorithmic risk tools as pseudo-science).

127. Koepke & Robinson, *supra* note 41, at 1802.

128. Lovins et al., *supra* note 102, at 187.

In contrast, cross-validation studies may also correct for problems when predictive technology results in “automation bias,” which occurs when stakeholders assume algorithms are not error-prone and thus place *too much* reliance on tool outcomes.¹²⁹ Without context for how a tool can actually perform on their own populations, officials may be influenced by the scientific mystique of the tool, accept its predictions on faith, and thus be unduly optimistic about its effectiveness.¹³⁰ “By ensuring that everyone in the system knows how accurate the tool is, they are better [able] to make decisions that give the tool appropriate, rather than exclusive, weighting.”¹³¹ To the extent that available cross-validation studies reveal significant error rates, their existence serves due process interests by allowing defendants to challenge the accuracy of the algorithm.¹³² In sum, judges and legal practitioners, even if they do not become experts in the science of risk assessment, should be mindful of “what it is the oracle is predicting” and “of how often this oracle is right and wrong.”¹³³

A word of caution is justified here about how the terms “validation” and “validated” are often bandied about in the criminal justice risk assessment literature. Tool developers and other advocates often portray a selected risk assessment tool as “validated,” implying that to be the end of the matter on its scientific and legal acceptability.¹³⁴ Certainly, many stakeholders and users appear to take such pronouncements at face value,¹³⁵ largely because the ideology of *validation* is assumed to have greater significance than

129. OSOBA ET AL., *supra* note 25, at 3; Simmons, *supra* note 56, at 1101; BAVITZ ET AL., *supra* note 48, at 3.

130. Hester, *supra* note 125, at 248; 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, 148 (2018); Douglas et al., *supra* note 17, at 135.

131. BAVITZ ET AL., *supra* note 48, at 20. *See also* Klingele, *supra* note 65, at 223 (informing judges on what a tool’s risk bins actually mean allows them better to determine how much weight to give any particular classification).

132. BAVITZ ET AL., *supra* note 48, at 22; Íñigo De Miguel Beriain, *Does the Use of Risk Assessments in Sentences Respect the Right to Due Process? A Critical Analysis of the Wisconsin v. Loomis Ruling*, 17 LAW PROBABILITY & RISK 45, 51 (2018). *See generally* John Philipsborn, *A Basic Risk Assessment Toolbox: Aiming for Adequate Lawyering During the Spread of Risk Assessments*, CHAMPION (Jan./Feb. 2020).

133. Hester, *supra* note 125, at 248.

134. Hamilton et al., *supra* note 90, at 536.

135. Kim & Duwe, *supra* note 17, at 217–18.

deserved.¹³⁶ Instead, any declaration by tool developers or interested parties regarding a tool being “validated” is insufficient on its own merit to justify a tool’s admissibility as scientific evidence or to operate as a signal that it deserves much weight.¹³⁷ Risk tool proponents unfortunately use the term “validated” even when they are relying upon one or two metrics that provide an extremely limited, and potentially misleading, perspective on the proficiency and reliability of these tools.¹³⁸ Researchers have tended to use the claim of “validation” when the tool performs slightly better than chance.¹³⁹ Yet statistical significance does not necessarily equate to practical significance. Some experts are just recently acknowledging that at least a few of the available tools do not perform to a reasonable standard to justify their utility for informing important criminal justice decisions that bear great legal consequences and infringe personal rights.¹⁴⁰

Validation is never truly complete anyway.¹⁴¹ Validation research is an ongoing process as times, sites, populations, evaluators, and criminogenic risk factors change.¹⁴² Hence, any proclamation about a tool generically being “validated” is “inappropriate and misleading. More appropriate is the review or advertisement that indicates substantial data exist demonstrating

136. Hamilton et al., *supra* note 90, at 567.

137. Melissa Hamilton, *Judicial Gatekeeping on Scientific Validity of Risk Assessment Tools*, 38 BEHAV. SCI. & L. 226, 227 (2020).

138. Kim & Duwe, *supra* note 17, at 215. *See also* Stefan Bogaerts et al., *Predictive Validity of the HKT-R Risk Assessment Tool: Two and 5-Year Violent Recidivism in a Nationwide Sample of Dutch Forensic Psychiatric Patients*, 62 INT’L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 2259, 2261 (2018) (naming the area under the curve [AUC] as the most commonly used single metric to judge algorithmic risk tools).

139. Hamilton et al., *supra* note 90, at 569. *See also* Kim & Duwe, *supra* note 17, at 217 (“[T]he field of risk assessment tends to regard ‘validation’ as a matter of degree, not as a matter of kind, which creates a tendency to consider an instrument ‘valid’ as long as it outperforms a fair coin.”).

140. Taxman & Dezember, *supra* note 63, at 37; Kim & Duwe, *supra* note 17, at 189.

141. Joel S. Milner et al., *Prediction Issues for Practitioners*, in ASSESSING DANGEROUSNESS: DOMESTIC VIOLENCE OFFENDERS AND CHILD ABUSERS 33, 41 (Jacquelyn C. Campbell & Jill Theresa Messing eds., 3d ed. 2017).

142. Tess M.S. Neal et al., *Psychological Assessments in Legal Contexts: Are Courts Keeping “Junk Science” Out of the Courtroom?*, 20 PSYCHOL. SCI. PUB. INTEREST 135, 137 (2019) (“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.”).

the test has some degree of validity for a specific application with a specific population or populations.”¹⁴³

Hence, a blanket estimate about any particular tool’s global abilities is of limited value. Practitioners should not much care how well a particular tool performs in other jurisdictions if it does not stand up well in their own locales.¹⁴⁴ Cross-validation studies thus must focus on localized data, not just observe its aggregate performance across sites.¹⁴⁵ This type of research has the potential for added benefits. Even if the particular tool does not initially appear to be sufficiently generalizable, repeatedly subjecting the tool to external validations can improve consistency if the information from such studies leads to adaptations to better fit different cultures, populations, and subjects.¹⁴⁶

Instrument tailoring is the process of taking an “off-the-shelf” tool and adapting the standard set of items to fit the agency’s local diction, crime categories or statutes, and adding additional items needed for the population or system variations. This process can be viewed as an assessment of the *jurisdiction’s responsivity to the tool*. Variations in population demographics may require additional items or alterations to response definitions to help “fit” the assessment to an agency’s needs.¹⁴⁷

In sum, the greater the fit of the tool to the specific population, the better the predictive accuracy statistics will likely result.¹⁴⁸

Not only is there a need to evaluate predictive ability in the specific population, it is important to test it on up-to-date data.¹⁴⁹ Repeat validations at the same site from time-to-time is valuable.¹⁵⁰ As the Wisconsin Supreme Court in *Loomis* warned, a risk tool should be “constantly monitored and re-normed for accuracy due to changing populations and sub-

143. Milner et al., *supra* note 141, at 41–42.

144. Kim & Duwe, *supra* note 17, at 217.

145. Solow-Niederman et al., *supra* note 12, at 742.

146. Juan José López-Ossorio et al., *Validation and Calibration of the Spanish Police Intimate Partner Violence Risk Assessment System (VioGén)*, 34 J. POLICE CRIM. PSYCHOL. 439, 440 (2019).

147. Hamilton et al., *supra* note 90, at 557–58 (emphasis in original).

148. *Id.* at 572.

149. DOBBIE & YANG, *supra* note 71, at 7.

150. Beriain, *supra* note 132, at 52.

populations.”¹⁵¹ Revalidation and recalibration are thereby necessary to ensure the efficacy of the tool and increase opportunities to achieve the gains intended.¹⁵²

Based on the foregoing concerns, further investigations are imperative to educate officials and stakeholders on the validity of risk tools with the samples for which they are intended for use (or in practice being used).¹⁵³ Experts further encourage policymakers not to rely on developers’ claims about their own tools, but instead to invite independent validation studies.¹⁵⁴ Audits conducted by objective third parties bring transparency,¹⁵⁵ enhance accountability, and provide courts with information they can apply in exercising needed oversight.¹⁵⁶ Cross-validation results should be publicly released and open to public debate about whether such tools rise to a sufficiently high degree of accuracy for the seriousness of the decisions they inform.¹⁵⁷ The public may then view tools performing at

151. *Loomis v. Wisconsin*, 881 N.W.2d 749, 763–64 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

152. DOBBIE & YANG, *supra* note 71, at 7; DESMARAIS & LOWDER, *supra* note 30, at 9; Garrett & Monahan, *supra* note 19, at 490.

153. BAVITZ ET AL., *supra* note 48, at 19; Alisha Caldwell Jimenez et al., *Validation and Application of the LS/CMI in Nebraska Probation*, 45 CRIM. JUST. & BEHAV. 863–84 (2018), <https://doi.org/10.1177/0093854818763231>.

154. BAVITZ ET AL., *supra* note 48, at 19; Garrett & Monahan, *supra* note 19, at 490; Ram, *supra* note 89, at 688; Seena Fazel & Stål Bjørkly, *Methodological Considerations in Risk Assessment Research*, in INTERNATIONAL PERSPECTIVES ON VIOLENCE RISK ASSESSMENT 16, 22 (Jay P. Singh et al. eds., 2016).

155. Emily Berman, *A Government of Laws and Not of Machines*, 98 B.U. L. REV. 1277, 1307 (2018); Stevenson & Mayson, *supra* note 51, at 47. *See also* PARTNERSHIP ON AI, *supra* note 12, at 29 (encouraging jurisdictions to release their datasets to permit “independent research and public discourse required to evaluate their effectiveness”).

156. Jeff Ward, *10 Things Judges Should Know About AI*, 103 JUDICATURE 12, 17 (2019); Katherine Freeman, *Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis*, 18 N.C. J.L. & TECH. 75, 101 (2016); Kroll et al., *supra* note 99, at 702. *See also* David G. Robinson et al., *Pretrial Risk Assessments: A Practical Guide for Judges*, 57(3) JUDGES J. 8, 11 (2018) (“Judicial leadership is an essential part of a growing national conversation about the real-world impact of algorithmic decision making in criminal justice.”).

157. PARTNERSHIP ON AI, *supra* note 12, at 31. “We should not make significant decisions based on tools that are not designed, calibrated, audited for results, or reviewed by independent parties and the public writ large.” The Leadership Conference, *Six Things to Know About Algorithm-Based Decision-Making Tools* 1, CIVILRIGHTSORG (2018), <http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-6Things.pdf>.

adequate levels as legitimate and thus more acceptable.¹⁵⁸ The cross-validation study herein responds to these calls for more field research on algorithmic risk tool performances, and this article acts as the public dissemination to inform stakeholder debate on relevant algorithmic risk topics.

II. RESEARCH QUESTIONS AND METHODOLOGICAL APPROACHES

We conducted a cross-validation study of the Public Safety Assessment (PSA), an algorithmic tool designed explicitly for pretrial settings. PSA developers intend for the tool to serve jurisdictions across the United States and thus have made it freely available.¹⁵⁹ In this section, we briefly describe the development and content of the PSA and introduce the three datasets. The main research aims are presented and the methodologies to address them are delineated.

A. The Public Safety Assessment

Arnold Ventures (now generally referred to as the Arnold Foundation) funded the development of the PSA.¹⁶⁰ The intent was to create an instrument that was “accurate, inexpensive to administer, easy to use, and scalable nationally . . . that could be adopted by judges and jurisdictions anywhere in America.”¹⁶¹ To that end, at least forty jurisdictions around the country to date have incorporated the PSA into their pretrial routines, making this tool the most widely used in pretrial settings.¹⁶²

158. Simmons, *supra* note 56, at 1071.

159. Stevenson, *supra* note 6, at 307.

160. Matthew DeMichele et al., *Public Safety Assessment: Predictive Utility and Differential Prediction by Race in Kentucky*, 19 CRIMINOLOGY & PUB. POL’Y 1, 6 (2020).

161. Arnold Foundation, *Developing a National Model for Pretrial Risk Assessment*, ARNOLD FOUNDATION 3 (2013), http://craftmediabucket.s3.amazonaws.com/uploads/PDFs/LJAF-research-summary_PSA-Court_4_1.pdf.

162. PRETRIAL JUST. INST., SCAN OF PRETRIAL PRACTICES 27 (2019), <https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=24bb2bc4-84ed-7324-929c-d0637db43c9a&forceDialog=0>.

The PSA was trained on a dataset of almost 750,000 pretrial defendants released in the United States, drawn from an original database involving over three hundred jurisdictions.¹⁶³ Developers statistically tested hundreds of predictive factors before selecting among those with the strongest correlations to be included and weighted in algorithms.¹⁶⁴ The PSA offers three scales, with the one of interest here being the scale that predicts failure to appear (FTA).¹⁶⁵ More specifically, this scale defines the outcome of interest as whether a released individual missed any court date prior to the disposition of their criminal case.¹⁶⁶

The FTA scale is a simple one.¹⁶⁷ The tool scores and weights four risk factors: (a) the existence of a pending charge at the time of the offense of arrest, (b) prior conviction, (c) prior failure to appear in the last two years, and (d) prior failure to appear older than two years. The tool's simplicity and/or its entire dependency on criminal history/compliance are not inherently causes for concern. There is value in a parsimonious tool with a few, but strongly predictive factors.¹⁶⁸ The PSA is an example of a "shorthand" screen that has recently emerged.¹⁶⁹ Then criminal history information with their objective and easily scored items can be preferable over more subjectively scored factors.¹⁷⁰

The PSA is designed to be efficiently administered based on available records and without an interview with the defendant.¹⁷¹ By not relying upon defendants' input, the tool eliminates any skew that might otherwise occur by "seasoned" offenders who may try to manipulate the assessment process.¹⁷² The PSA represents an actuarial instrument in which risk predictions are derived from an algorithm with no discretion in selecting,

163. Arnold Foundation, *supra* note 161, at 3.

164. *Id.* at 3–4.

165. The PSA also offers a scale to predict any new arrest and a flag to predict any new violent arrest.

166. DeMichele et al., *supra* note 160, at 7.

167. Rizer & Watney, *supra* note 109, at 191.

168. Georgios Georgiou, *Weights Matter: Improving the Predictive Validity of Risk Assessments for Criminal Offenders*, 58 J. OFFENDER REHABILITATION 92, 94 (2019).

169. Hannah-Moffat et al., *supra* note 24, at 392.

170. Georgiou, *supra* note 168, at 94.

171. DeMichele et al., *supra* note 160, at 6.

172. Hannah-Moffat et al., *supra* note 24, at 402.

measuring, or scoring items.¹⁷³ The PSA is thereby offered as a systematic, standardized, and efficient method for evaluating risk.¹⁷⁴

The tool appears to be useful to judges in practice. A recent survey of thirty jurisdictions using the PSA found that nearly 80 percent agreed that the PSA “often” or “always” informs their release decisions, and over 60 percent of judges reported that they “often” concurred with PSA outcomes.¹⁷⁵

Despite the PSA instrument itself being publicly available, little information on its development is known, such as the statistical methodologies employed, rationale for the choice of datasets, or validation information.¹⁷⁶ The Arnold Foundation simply states that it conducted validations, without additional detail.¹⁷⁷ When enterprising researchers sent open records requests to jurisdictions enquiring about their use of the PSA, officials generally rebuffed these appeals, contending they were restrained by agreements with the Arnold Foundation from releasing such information.¹⁷⁸ In turn, the Arnold Foundation justifies not disclosing the developmental dataset or validation results by arguing that it is bound by confidentiality agreements with the agencies who shared their data for these purposes.¹⁷⁹

One post-development validation study of the PSA was recently released by the Research Triangle Institute (RTI).¹⁸⁰ Caution is advised, however, as the RTI conducted the research with funding and data provided exclusively by the Arnold Foundation itself.¹⁸¹ Further, as will be addressed below, the RTI provided only two measures of performance ability, which

173. Gina M. Vincent et al., *The Validity of Risk Assessment Instruments for Transition-Age Youth*, 87 J. CONSULTING & CLINICAL PSYCHOL. 171 (2019).

174. The historical trajectory of risk tool development involves four generations. The first employed clinical judgements on risk. The PSA represents a second generation risk tool as it is actuarial in nature and reliant upon static factors. The PSA does not include dynamic factors or address needs or responsivity as do third and fourth generation instruments.

175. Matthew DeMichele et al., *What do Criminal Justice Professionals Think about Risk Assessment at Pretrial?*, 83 FED. PROB. 32, 37 (2019).

176. Koepke & Robinson, *supra* note 41, at 1803.

177. Arnold Ventures, Public Safety Assessment FAQs (“PSA 101”) at *2 (2019), https://craftmediabucket.s3.amazonaws.com/uploads/Public-Safety-Assessment-101_190319_140124.pdf.

178. Koepke & Robinson, *supra* note 41, at 1803.

179. Arnold Ventures, *supra* note 177, at *3.

180. *See generally* DeMichele et al., *supra* note 160.

181. *Id.* at 11.

are not sufficient to appreciate the PSA's overall utility. Instead, the cross-validation study we conducted and report herein offers multiple metrics and thereby more robust analyses of the PSA's performance from varied perspectives.

B. The Datasets

This secondary data analysis used archival data on the PSA's FTA risk outcomes and court appearance compliance from three diverse jurisdictions that use the tool to inform judges when making decisions on pretrial release for individual defendants. The samples are drawn from the states of Illinois, Kentucky, and New Mexico.

1. Illinois Dataset

The state of Illinois has been at the forefront of bail reform. In 1963, the state was one of the first to eliminate the commercial bail bond industry by allowing individuals to post 10 percent of a secured money bond to the court.¹⁸² Legislation in 1987 required each circuit court in the state to implement a pretrial services agency to operate in its jurisdiction.¹⁸³ Thirty years thereafter, the state's Bail Reform Act of 2017 reduced 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."¹⁸⁴ This bail reform bill encouraged 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.¹⁸⁵

This first dataset derives from Cook County, home to the city of Chicago. Cook County transformed its release decision practices by formally adopting algorithmic risk in its pretrial processes. Pressured by pending civil rights litigation and grassroots community activism to adopt progressive measures to ameliorate the harms of pretrial detention,¹⁸⁶ the chief

182. DOYLE ET AL., *supra* note 34, at 53.

183. Pretrial Services Act, 725 ILL. COMP. STAT. 185/1.

184. Pub. Act 100-0001, S.B. 2034, 100th Gen. Assembly (2017).

185. 725 ILL. COMP. STAT. 5/110-5(f) (referencing a risk assessment evaluation of a pretrial defendant using a recognized evidence-based instrument).

186. Van Brunt & Bowman, *supra* note 37, at 762.

judge of the Cook County Circuit Court in 2017 issued an order to require that judges use risk assessment tool results in decisions on release.¹⁸⁷ Cook County had at the time of the chief judge's order already begun preparing to use predictive technology. In anticipation, Cook County authorities in March 2016 adopted the PSA as its foundational risk assessment tool for use on its pretrial population.¹⁸⁸

The Illinois sample set includes 44,939 arrestees who were booked, scored with the PSA FTA risk scale, and released between July 1, 2016, and December 31, 2018. This dataset was made public as multiple electronic spreadsheets by the Chief Judge of the Cook County judicial system in conjunction with a governmental report concluding that rates of failure to appear and rearrest did not increase after implementation of the PSA.¹⁸⁹ Such governmental report did not address the tool's empirical validity in terms of its predictive abilities.

The racial makeup of the Cook County dataset consists of whites (32%), blacks (66%), and other (2%). Most are male (86%). The dataset includes an ordinal age variable, with 25 years and younger (31%), age 26–35 (30%), age 36–45 (17%), and age 46 and above (21%).¹⁹⁰

2. Kentucky Dataset

Kentucky is known as a pretrial reform pioneer.¹⁹¹ The state has used some form of a risk tool in its pretrial processes since 1976, though judicial consideration of the tool was optional.¹⁹² In 2011, the state legislature enacted a law designed to reduce incarceration rates by requiring that judges employ an evidence-based risk assessment tool.¹⁹³

187. DOYLE ET AL., *supra* note 34, at 52 (referencing OFFICE OF THE CHIEF JUDGE, CIRCUIT COURT OF COOK COUNTY, GENERAL ORDER NO. 18.8A—PROCEDURES FOR BAIL HEARINGS AND PRETRIAL RELEASE (July 17, 2017)).

188. DOYLE ET AL., *supra* note 34, at 56.

189. CIRCUIT COURT OF COOK COUNTY, BAIL REFORM IN COOK COUNTY: AN EXAMINATION OF GENERAL ORDER 18.8A AND BAIL IN FELONY CASES 2 (2019), [http://www.cookcountycourt.org/Portals/o/Statistics/Bail Reform/Bail Reform Report FINAL - Published 05.9.19.pdf](http://www.cookcountycourt.org/Portals/o/Statistics/Bail%20Reform/Bail%20Reform%20Report%20FINAL%20Published%2005.9.19.pdf).

190. The percentages do not sum to 100% due to rounding.

191. Stevenson, *supra* note 6, at 307.

192. *Id.* at 308.

193. Public Safety and Offender Accountability Act, H.B. 463, 2011 Gen. Assemb. Reg. Sess. (Ky. 2011) (codified at KY. REV. STAT. ANN. § 532.007 (2019)).

Kentucky was the first jurisdiction to utilize the PSA in actual cases.¹⁹⁴ During the state's implementation, PSA researchers worked alongside state officials to investigate the tool's performance and adjusted the tool as a result of what they learned.¹⁹⁵

The data for this sample were extracted from a manuscript concerning the use of the PSA in Kentucky pretrial cases.¹⁹⁶ The underlying manuscript presents the only publicly available validation study of the PSA's accuracy. However, the validation information provided by RTI was limited to the tool's discriminative ability in the form of observing whether failure rates increased incrementally by PSA score and calculating the Area Under the Curve (AUC, discussed further below).¹⁹⁷ Perhaps as a result of their restricted inquiries, the authors explicitly called for more research on validating the PSA with local datasets.¹⁹⁸ The cross-validation study offered herein responds by calculating various additional statistics to more fully examine the PSA's performance in Kentucky, and adds further value by comparing results of the other two sites.

This second dataset tracked 164,407 adults who were booked into jails across Kentucky between July 1, 2013, and December 30, 2014, received a PSA score, and were released pending trial.¹⁹⁹ In the Kentucky dataset, the racial makeup is whites (81%), blacks (17%), and another race or race unknown (2%). For gender, the sample comprises males (69%), females (31%), and a small number whose gender was not recorded. The average age is 34.

194. DeMichele et al., *supra* note 160, at 12.

195. *Id.* at 6.

196. *Id.* at 13 tbl. 4.

197. *Id.*

198. *Id.* at 18 (“[O]ur results provide information about the instrument’s performance in Kentucky. Research is needed throughout the jurisdictions using the PSA to assess accuracy and evaluate predictive bias using local data.”).

199. The preexisting validation study in Kentucky included sample numbers and the percentages of FTAs at each PSA FTA score. DeMichele et al., *supra* note 160, at 12 tbl. 3. From these datapoints, we were able to calculate the numbers of defendants who were assessed at each PSA FTA risk level and the numbers within each level who failed versus did not fail. Because of this rounding process, we lost a small number of defendants, such that the original dataset of 164,597 was reduced to a final 164,407 cases as reported herein.

3. New Mexico Dataset

For decades, New Mexico's pretrial system has permitted commercial bail bonds companies to operate and has used bail schedules.²⁰⁰ In a groundbreaking 2014 ruling, though, the New Mexico Supreme Court limited the application of the bail schedules by precluding judges from ordering a bail amount solely on the severity of the offense charged.²⁰¹ Instead, the court ruled that pretrial judges were required to conduct an individualized analysis of whether the defendant is likely to flee or reoffend if released pending trial.²⁰² In conjunction with issuing the critical opinion, the court created an advisory committee to review the state's practices with bail and criminal procedure.²⁰³ The committee's recommendations resulted in a state constitutional amendment that passed in 2017 to reform New Mexico's pretrial systems by precluding judges from detaining individuals solely because they could not afford to post bail, while also permitting the detention of those who pose a risk of failing to appear or committing a new offense if released.²⁰⁴

This third dataset is from Bernalillo County, the most populous county in the state of New Mexico, and whose county seat is the city of Albuquerque. In late 2016, when the state constitutional amendment had passed the legislature but was pending approval from voters, the chief judge of the municipal court in Bernalillo County contracted with the Arnold Foundation to adopt the PSA for use in pretrial evaluations for release.²⁰⁵

Bernalillo County implemented the PSA in its pretrial decision routines in June 2017.²⁰⁶ We created this sample set by extracting relevant data from

200. DOYLE ET AL., *supra* note 34, at 64, 67.

201. *State v. Brown*, 338 P.3d 1276, 1292 (N.M. 2014).

202. *Id.*

203. DOYLE ET AL., *supra* note 34, at 65.

204. NEW MEXICO STATISTICAL ANALYSIS CENTER, BAIL REFORM: BASELINE MEASURES I (2019) (referring to N.M. Const. art. II, § 13), <https://isr.unm.edu/reports/2019/bail-reform-baseline-measures.pdf>.

205. *Collins v. Daniels*, D.C. No. 1:17-CV-00776-RJ-KK, at *7 (10th Cir. Feb. 25, 2020). After the judgement, the New Mexico Supreme Court revised the Rules of Criminal Procedure to formally require a pretrial risk assessment instrument be used in pretrial release decisions. N.M. R. Crim. Proc. 5-401(C).

206. ELISE FERGUSON ET AL., BERNALILLO COUNTY PUBLIC SAFETY ASSESSMENT REVIEW – JULY 2017 TO MARCH 2017, at *2 (2019), [https://www.bernco.gov/uploads/files/PublicSafety/Bernalillo County Public Safety Assessment Review - July 2017 to March 2019%20C2%20Ao.pdf](https://www.bernco.gov/uploads/files/PublicSafety/Bernalillo%20County%20Public%20Safety%20Assessment%20Review%20July%202017%20to%20March%202019%20C2%20Ao.pdf).

a manuscript produced by county authorities.²⁰⁷ The county's document itself did not examine any validation measures, but instead was devoted to evaluating the correspondence between risk scores with recommendations for release and assigned supervision levels.²⁰⁸ The sample set includes 6,392 felony cases filed between July 1, 2017, and March 31, 2019, where the defendants were released pending their trial dates.²⁰⁹ Information on race, gender, or age was not available with this New Mexico dataset.

* * *

Thus, the datasets with a combined 215,738 cases derive from three unique sites. The fact that two datasets represent counties and the third an entire state does not necessarily introduce bias. In Kentucky, pretrial services operate at the state level, rather than a series of more localized versions.²¹⁰ Hence, while it is possible for local variations to exist in courtroom practices, the implementation of the PSA at the state level suggests reasonable comparability with the other datasets. Besides, one of the goals of this study is to evaluate a risk tool that its developer purports to be appropriate for use across the country, without stipulations as to size or type of pretrial agency. Indeed, the Arnold Foundation expressly approves of the use of the PSA in both county and statewide pretrial offices.²¹¹

C. Research Questions

This study is meant to elicit information on the PSA's performance abilities to assess the risk of failure to appear by released pretrial defendants. The research questions are as follows:

1. Do the PSA risk levels (low, medium, high) meaningfully classify individuals into unique groups according to their likelihood of failure to appear?
2. How accurate is the PSA in correctly predicting failure to appear? Correspondingly, how accurate is the PSA in correctly predicting success in terms of compliance with court dates?

207. See generally *id.*

208. *Id.* at *2.

209. The data were extracted from the report. *Id.* at *8 tbl. 4.

210. Stevenson, *supra* note 6, at 342.

211. *Supra* Section I.B.

3. Does the PSA perform similarly across the three samples? This question regards whether differential validity is present. Differential validity refers to the existence of disparities in how the PSA performs in separate samples.²¹² We expect that there will be meaningful variation between samples on the predictive ability of PSA and thus that differential validity will exist. Anticipated variances may be due to heterogeneity across the sites regarding the following attributes: population characteristics (e.g., sociodemographic makeup), degrees of associations between underlying risk factors and failure to appear, and base rates of failure to appear.²¹³

D. Methods

We imported the three datasets into Excel and SPSS version 25.0 for analyses. As risk tools commonly stratify individuals into ordinal risk bins,²¹⁴ and decision makers often prefer a three-risk-bin strategy (e.g., low, medium, high),²¹⁵ it is important to examine the predictive abilities of risk bins. Adopting Cook County's preferred strategy in this regard, we used a three-risk-bin allocation for PSA's FTA scores of low (1–2 points), medium (3–4 points), and high (5–6 points). With data on the numbers of defendants in each risk bin and corresponding failure to appear rates, we computed measures to inform on the performance abilities of the PSA's FTA scale in the jurisdictions to respond to the three research questions.

We analyzed multiple dimensions of the tool's discrimination and calibration abilities. 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.²¹⁶ Instead, in the

212. See Raymond A. Katzell & Frank J. Dyer, *Differential Validity Revived*, 62 J. APPLIED PSYCHOL. 137, 137 (1977) (indicating a test of the validity coefficients between two samples that rejects the hypothesis of similarity means differential validity).

213. Other potential reasons for variation include differences in the quality of criminal history records, quality of sources used to identify failure to appear, standards for court appearances, environmental factors that may moderate outcomes across sites, and availability and strength of supervisory service. Validation studies of pretrial risk tools rarely control for these types of population characteristics.

214. Jeremy Lualen et al., *The Predictive Validity of the Post-Conviction Risk Assessment Among Federal Offenders*, 43 CRIM. JUST. & BEHAV. 1173, 1174 (2016).

215. 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).

216. BLACK'S LAW DICTIONARY 543 (D. Garner ed., 9th ed. 1990).

genre of predictive technologies the term reflects the ability to distinguish between different outcomes.²¹⁷

Many validation studies regrettably rely solely upon one measure of discrimination (the AUC, discussed below).²¹⁸ Calibration accuracy is rarely assessed or reported.²¹⁹ Yet, each offers a distinct contribution to surveying a tool's accuracy and performance. A tool may vary in how well it meets either of these dimensions.²²⁰

Discrimination reflects how well a tool classifies by distinguishing between failures and non-failures.²²¹ Discrimination is retrospective in nature as it is calculated after the failures and non-failures have been identified.²²² Discrimination thereby represents the tool's relative accuracy.²²³ In contrast, calibration measures how well the tool's predicted outcomes match the actual outcomes in terms of failure, and thus presents the tool's achievement on absolute accuracy.²²⁴ Calibration is forward-looking; it measures how well the tool predicts future failure. Suppose a particular tool predicted that 40 percent of the assessed population would fail. The tool would be well-calibrated if, indeed, about 40 out of every 100 actually did fail. In contrast, if 20 or 50 percent of the population failed, the tool instead would exhibit poor calibration, and thus be less credible because the tool would be over- or under-calibrated, respectively.

Albeit unfortunately rarely accomplished, best practices in reporting on the utility of a criterion-referenced prediction tool (such as the PSA)

217. OXFORD ENGLISH DICTIONARY (3rd ed. 2013).

218. 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, 546–57 (2019); L. Maaik Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 11–12 (2017).

219. 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 154, at 22.

220. See Melissa Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, 56 AM. CRIM. L. REV. 1533, 1574–75 (2019) (showing results of a risk tool with varying performances for discrimination and calibration).

221. Helmus & Babchishin, *supra* note 218, at 11.

222. *Id.*

223. *Id.*

224. Crowson et al., *supra* note 119, at 1692. The term “failure” here is not intended as a value statement about an individual's character. Instead, failure reflects that criminal justice risk assessment is fundamentally about predicting some negative event defined as such by the criminal justice system (e.g., arrest, conviction, parole violation). Failure also more acutely is relevant as the outcome of interest in this study is failure to appear.

endorse engaging measures to assess both discrimination and calibration.²²⁵ “A scale that ranks well, but systematically overestimates or underestimates risk might have good discriminative properties but be poorly calibrated to the population under examination.”²²⁶ Thus, we have chosen to calculate multiple measures for discrimination and calibration because each is meant to evaluate a somewhat distinct aspect of the tool’s performance ability. Employing more than one perspective provides a more holistic analysis of the PSA’s utility.

1. Discriminative Ability

Research question I concerns the tool’s ability to effectively discriminate in its classifications. The query seeks to understand how well the PSA was able to differentiate groups of defendants according to their risk of failing to appear. Stakeholders likely have an interest in the risk bin strategy (low, medium, high) offering some meaningful distinctions.

Several statistical computations were run in response. For each sample, we calculated the percentages of failure to appear at each ordinal risk bin (low, medium, high) to observe whether failure rates appreciably increased in a monotonic fashion in the medium- and high-risk bins. This offers a simple method to visually confirm that there appears some meaning to the risk bin distinctions.

For a more statistically robust view, we then employed a logistic regression model for each sample with the dependent variable as failure to appear (failure = 1; non-failure = 0). The predictors were the risk bin dummy variables with the reference category of low risk. The resulting coefficients were translated into more interpretable odds ratios to indicate the magnitude of deviations in failure rates for the medium- and the high-risk bins as compared to low-risk.²²⁷ Larger odds ratios reflect better discrimination in terms of relative accuracy.²²⁸ Next, an overall chi-square statistic for each

225. Hanson, *supra* note 219, at 27; Fazel & Bjørkly, *supra* note 154, at 22.

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

227. Helmus & Babchishin, *supra* note 218, at 14 (indicating logistic regression is a discrimination method for quantifying the change in the rate of failure as the risk level increases).

228. Seung C. Lee & R. Karl Hanson, *Similar Predictive Accuracy of the Static-99R Risk Tool for White, Black, and Hispanic Sex Offenders in California*, 44 CRIM. JUST. & BEHAV. 1125, 1131 (2017).

regression model (i.e., for each sample) tested whether, for that site, the PSA created distinct groups with the overall three-risk-bin scheme according to their likelihood of failure (with statistical significance). Thus, each chi-square statistic signals if the tool is a good classification model fit of the ordinal risk bin strategy to each dataset.

To supplement the analyses, we calculated three additional overall discrimination statistics. Somers' d measures the strength of association between a single ordinal predictor involving the three risk bins and the binary outcome of failure to appear.²²⁹ In other words, Somers' d indicates how well the ranking system (i.e., low, medium, high) performs. Somers' d ranges from 0 (no association) to 1 (perfect association). For each sample set, Somers' d was calculated with a chi-square test to assess if there was a statistically significant relationship between the hierarchical ordinal bin structure and failure rates.

Next, we employed the Dispersion Index for Risk (DIFR). The DIFR is a base rate dispersion metric that evaluates the effectiveness of a tool's classification system as partitioned into various groups, and the extent to which group outcome rates vary from the base rate for the entire sample.²³⁰ Thus, the DIFR indicates potency in terms of how well a tool distinguishes multiple groupings with significantly disparate failure rates.²³¹ The DIFR has a low value of 0 but has no upper end, theoretically reaching into infinity.²³² Accordingly, no statistical significance test exists to judge the DIFR.

229. Taxman, *supra* note 103, at 273.

230. CHRIS BAIRD ET AL., A COMPARISON OF RISK ASSESSMENT INSTRUMENTS IN JUVENILE JUSTICE 19 (2013), https://www.nccdglobal.org/sites/default/files/publication_pdf/nccd_fire_report.pdf. See also Baglivio et al., *supra* note 69 ("The DIFR essentially adjusts for sample size and assesses how an entire sample is divided by risk, and the extent to which the pretrial failure rate of subgroups (low-risk, moderate-risk, etc.) varies from the total sample base rates.").

231. KRISTEN JOHNSON ET AL., DEVELOPING AN ACTUARIAL RISK ASSESSMENT TO INFORM THE DECISIONS MADE BY ADULT PROTECTIVE SERVICE WORKERS B4 (2010), <https://www.issuelab.org/resources/5494/5494.pdf>.

232. Eric Silver et al., *Constructing Actuarial Devices for Predicting Recidivism: A Comparison of Methods*, 27 CRIM. JUST. & BEHAV. 733, 751 (2000) ("DIFR is a weighted, composite log-odds that assesses the degree of difference between the base rate of failure in the total sample and the base rates of failure in each of the risk classes produced by the screening device.").

The next global discrimination index is the Area Under the Curve (AUC).²³³ The AUC is the most frequently utilized statistic in risk assessment validation studies.²³⁴ The AUC here represents the probability that a randomly selected defendant who failed to appear receives a higher risk classification than a randomly selected non-failure. An AUC of 0.5 indicates no better accuracy than chance, and 1.0 signifies perfect discrimination (i.e., all failures are classified in a higher risk bin than all non-failures).²³⁵ Again, we used the ordinal risk bin strategy, as is common in the relevant literature.²³⁶

No statistical standard exists for how close to 1.0 an AUC must be to signify a sufficiently accurate test in terms of discriminative utility. Determining the degree of acceptable performance per the AUC is, instead, a value judgement.²³⁷ Some risk assessment scholars refer to AUCs of .56, .64, and .71 as the thresholds for small, medium, and large effect sizes, respectively.²³⁸ (An effect size simply refers to the magnitude of the relationship between two variables.) Using this guidance, an AUC of .56 means that failures are ranked at greater likelihood than non-failures in 56 percent of cases. This .56 is better than chance, but only slightly so, which is why it is labeled a small effect size.

An alternative framework for labeling effect sizes has recently been suggested as AUCs presenting as poor (up to .54), fair (.55–.63), good (.64–.70), and excellent (.71 and higher).²³⁹ On the other hand, a more conservative conceptualization regards AUCs as indicating a tool's discriminative ability is a failure (.50–.59), poor (.60–.69), fair (.70–.79), good (.80–.89), and excellent (.90 and above).²⁴⁰

233. Critical limitations exist with the AUC, as noted elsewhere. Hamilton, *supra* note 137, at 235–37. Nevertheless, this statistic is included herein as it is the most popular metric in the scientific literature on risk assessment tool validation.

234. Kim & Duwe, *supra* note 17, at 206.

235. Helmus & Babchishin, *supra* note 218, at 12.

236. Thomas Cohen & Christopher Lowenkamp, *Revalidation of the Federal PTR: Testing the PTR for Predictive Biases*, 46 CRIM. JUST. & BEHAV. 234, 245 tbl. 2 (2019); PICARD ET AL., *supra* note 43, at 6–7.

237. 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).

238. Helmus & Babchishin, *supra* note 218, at 12.

239. DeMichele et al., *supra* note 160, at 8.

240. Helmus & Babchishin, *supra* note 218, at 12.

2. Calibration Ability

Research question 2 enquires about the tool's predictive (i.e., forecasting) ability. In response, multiple measures of calibration were tested to convey the PSA's levels of absolute accuracy in prospective predictions. One signifier of model calibration is the Hosmer-Lemeshow (H-L) goodness-of-fit test. The H-L test partitions the sample into subgroups based on the tool's predicted probabilities of failure, and then compares within subgroups the predicted probability with the (actual) observed rates of failure.²⁴¹ A chi-square statistic indicates the statistical significance of the equivalence between predicted and observed rates.²⁴² Equivalence represents a solidly positive attribute of a tool as it indicates strong calibration ability.

For the H-L test, the null hypothesis is that the model fits the data well.²⁴³ In other words, the null hypothesis is that the actual and predicted failure rates are similar within the subgroups. The direction of the statistical significance test operates rather uniquely with the H-L test. An H-L test p value $< .05$ rejects the null hypothesis, reflecting statistically significant differences between predicted and observed outcomes, and thus signifies poor model fit. In contrast, a non-significant H-L result ($p > .05$) accepts the null hypothesis of a good overall fit in that the observed and expected recidivism rates in risk bins are equivalent.²⁴⁴

We previously discussed deploying logistic regression models to evaluate model fit. The difference is that the logistic models concern classification ability in judging the tool's relative accuracy, whereas the H-L test is about predictive ability in evaluating the tool's absolute accuracy.

Although the H-L test is informative, it is best not to definitively conclude whether a model fits the data based on a statistically significant result from a single goodness-of-fit test. A preferred practice is to also look at other empirical techniques as each may be testing an alternative way of

241. DAVID W. HOSMER ET AL., *APPLIED LOGISTIC REGRESSION* 157 (3d ed. 2013).

242. Crowson et al., *supra* note 119, at 1693.

243. Wei Yu et al., *A Modified Hosmer-Lemeshow Test for Large Data Sets*, 46 COMM. STAT. THEORY & METHODS 11813, 11813 (2017).

244. 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*, 33 SEXUAL ABUSE 63 (2020), <https://doi.org/10.1177/1079063219841901>.

observing the data. We therefore also offer an overall accuracy measure that combined true positives (failures correctly classified as to fail) plus true negatives (non-failures correctly classified as not to fail) and divides the sum by the population size. In other words, this is the single rate at which the tool accurately predicted outcomes. Yet this global accuracy measure is limited, too, in that it obscures whether the predictions are better at true positives or true negatives.

Thus, we calculated two additional calibration statistics that focus separately on true positives and on true negatives.²⁴⁵ The Positive Predictive Value (PPV) is the forecasting true positive rate, reflecting the proportion of higher-risk predictions who go on to fail.²⁴⁶ The Negative Predictive Value (NPV) is the forecasting true negative rate, presenting the proportion of lower-risk predictions who do not fail.²⁴⁷ The PPV is a high-risk calibration measure, and the NPV is a low-risk calibration measure.²⁴⁸ In Bayes rule terms, the PPV is the posterior probability of failure given a prediction of failure, and the NPV is the posterior probability of success given a prediction of non-failure.²⁴⁹

Unlike the H-L test, overall accuracy, PPVs, and NPVs require a choice of cut-point from the scale because they rely upon a dichotomous perspective of whether the tool predicted the individual to fail or not to fail. We use here the PSA's risk bin strategy to create two distinct cut-points. This form of comparing the predictive ability of tools using

245. We decided to report on the predictive accuracy statistics using calibration measures rather than the discriminative classification metrics. The latter would be sensitivity (given known failures, the rate they were classified as higher risk) and specificity (given known successes, the rate they were classified as lower risk). Forecasted predictors are more relevant than retrospective diagnostics. Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 12 (2013). Plus, forecasted predictors account for differences in base rates, whereas classification statistics generally do not. If base rates differ, forecasted predictors should actually not be identical between the groups. A group with a higher base rate *should be* predicted to fail more often. Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, SOC. METHODS & RES. 3 (2021).

246. Daniel J. Neller & Richard I. Frederick, *Classification Accuracy of Actuarial Risk Assessment Instruments*, 31 BEHAV. SCI. & L. 141, 143 (2013).

247. Singh, *supra* note 245, at 10 tbl. 1.

248. *Id.* at 11 fig. 1.

249. Kristian Linnert et al., *Quantifying the Accuracy of a Diagnostic Test or Marker*, 58 CLINICAL CHEM. 1292, 1296 (2012).

contrasting ordinal binning strategies has merit.²⁵⁰ Whereas one method of parsing the data may indicate weak calibration, another may exhibit strong calibration. The contrast is informative as to where to place more confidence. Here, one cut-point compared low risk with a merged combination of medium- and high-risk groupings (low vs. medium/high) (herein, Cut-point A). Cut-point A isolates the low-risk group from the medium- and high-risk bins. The other combined low and medium bins to compare to high risk (low/medium vs. high) (herein, Cut-point B). Cut-point B thereby isolates the high risk.

Finally, an alternative measure is calibration-in-the-large, which defines a tool as well-calibrated if the expected rate of failure is equivalent to the observed rate of failure.²⁵¹ We will call this the E/O test, with the “E” standing for expected rate and the “O” for the observed rate. We compute it for each sample as a whole. The E/O tests used Cut-Point A and Cut-Point B as well, in which the expected rate is the proportion of those predicted to fail, and the observed rate is equal to the base rate of failure to appear for the entire sample. Stakeholders might be interested in which cut-point strategy provides a more optimal choice. A jurisdiction would likely prefer to employ a cutoff that more closely aligns the expected rate with the observed rate of its own population. The E/O test may appear similar to the PPV, yet they are focused on slightly different sets of data. The PPV asks of those predicted to fail, what percentage did fail? Thus, the PPV is only calculated on the subsample of those predicted to fail (i.e., it excludes those predicted to succeed). The E/O test version of calibration views the entire dataset and compares in the entire sample the percentage predicted to fail to the base rate of failure. In other words, a tool is well-calibrated in the E/O test if the expected rate of failure is similar to the failure rate of the entire sample.

3. Differential Validity

The third research question calls for investigating differential validity by comparing performances across the three samples to determine how

250. 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, 503 (2011), <http://dx.doi.org/10.1016/j.cpr.2010.11.009>.

251. Crowson et al., *supra* note 119, at 1692. *See also* Berk, *supra* note 1, at 180–81 (contrasting predicted proportions with observed proportions of recidivism in entire samples).

generalizable the PSA is in practice. To assess the existence of statistically significant variations across the three samples, chi-square statistics were calculated for the following measures, all of which were introduced earlier: (a) proportions of failures in the low-, medium-, and high-risk groups, (b) Somers' *d* scores, (c) AUCs, (d) overall accuracy statistics, and (e) forecasting true positive and true negative rates (i.e., PPVs, NPVs); and (f) expected and observed failure rates in the E/O test. No significance tests exist to statistically evaluate differential validity for the overall chi-square statistics for the regression models with ordinal risk bin predictors, the DIFRs, or the H-L test. However, we can observe size distinctions in each of them to more qualitatively adjudge the presence of differential validity.

III. RISK ASSESSMENT TOOL PERFORMANCE

This section presents the results of the cross-validation study and discusses their relevance and importance in terms of the PSA's utility in the field. Policy implications are offered to situate the results in contributing to the broader literature and informing stakeholders on important issues in the developing algorithmic risk reform movement.

A. Public Safety Assessment Performance

This study computed multiple measures to review the performance of the PSA scale across three samples to respond to the research questions posed in terms of discriminative classification ability, calibration accuracy, and testing for differential validity. This section begins by addressing the first two research questions in order, and the comparative analysis about differential validity will be discussed in context within each of those two responses.

1. Discriminative Classification Ability

Summary statistics demonstrate the relative risk binning outcomes across the samples. Table 1 contains data on the numbers and percentages of the sample sets that were classified into each of the ordinal risk bins.

Within each sample, there are roughly equivalent rates within each of the low and medium bins. In all samples, the PSA classified a substantial majority of defendants in either the low- or medium-risk bins (77–95% calculated from table statistics).

Table 1. PSA classification risk bins for failure to appear.

Risk Bin	Illinois (n = 44,939)		Kentucky (n = 164,407)		New Mexico (n = 6,392)		χ^2
	n	%	n	%	n	%**	
Low	21,529	48%	72,147	44%	2,280	36%	*
Medium	21,151	47%	67,839	41%	2,597	41%	*
High	2,259	5%	24,421	15%	1,515	24%	*

* $p < .001$.

** The proportions do not sum to 100% because of rounding.

Although Table 1 statistics do not clearly provide evidence of differential validity, it does give an initial hint that the samples vary in risk-relevant ways. The tool classified a minority number of offenders as high-risk, though the rates varied greatly, from 5 to 24 percent. While one in twenty defendants in the Illinois sample fell into high-risk, one in four did so in the New Mexico sample. To be clear, though, these results do not mean the tool is assessing the populations of offenders in these jurisdictions in unequal groups. The samples here are only of *released* individuals, not all persons booked into jail who received PSA outcomes whether released or not. Judges retain discretion in release decisions, and thus, it is not expected that the risk classifications entirely dictate the rate at which individuals are released in any of these jurisdictions.

Table 2 provides results for various statistical measures to judge how risk bins performed in stratifying known failures to appear across samples. Notice that the base rates of failures varied from 15 to 19 percent.

a. Overall Discriminative Ability. For each of the three samples, the failure rates increased in a linear fashion as risk levels increased. Across samples, the odds of failure in the medium-risk groups were approximately two times that of the odds in the low-risk group. The odds of failure for the high-risk groups were three to four times that of the odds for low-risk. For each sample, the odds ratios in the medium-risk and in the high-risk bins were statistically significant (not shown in table, $p < .001$ in each sample). Overall chi-square statistics for each of the three logistic regression models indicated good classification model fit, thereby showing statistical significance in the ordinal risk bin strategy overall ($p < .001$). These statistics

Table 2. PSA risk bin performance.

	Illinois			Kentucky			New Mexico			χ^2
	n	%	OR**	n	%	OR	n	%	OR	
Base rate of FTA	7,691	17.1%	—	24,293	14.8%	—	1,181	18.5%	—	*
Low	2,965	13.8%	—	6,357	8.8%	—	247	10.8%	—	*
Medium	4,075	19.3%	1.49	11,188	16.5%	2.04	509	19.6%	2.01	*
High	651	28.8%	2.54	6,748	27.6%	3.95	425	28.1%	3.21	n.s.
Overall χ^2	456.713*			5,402.201*			182.840*			n/a
Somers' <i>d</i>	.129			.262			.109			*
DIFR	.252			.501			.462			n/a
AUC	.564			.631			.618			*

* $p < .001$; n.s. = not significant; n/a = not applicable.
** OR = odds ratio.

showed the PSA’s risk binning strategy is able in each sample to establish risk levels in unique groups according to their likelihood for failure to appear.

Somers’ *d* statistics for each sample found positive associations between FTA ordinal risk bins and failure rates (.13 to .26) at a statistically significant level (not shown in the table, $p < .001$ for each sample). Still, statistical significance in such large sample sizes does not equate with an appreciable degree of practical utility. On a scale from 0.0 to 1.0, the largest magnitude of the Somers’ *d* with the Kentucky sample (.26) is, in practical terms, not outstanding. A prior meta-analysis of risk assessment tools used across criminal justice decision points found median Somers’ *d* scores ranging from .23 to .55.²⁵² That the Kentucky sample is at the low end and with the Somers’ *d* statistics for the two other samples far below the smallest value in that meta-analysis suggest a relatively weak association from a practical perspective between the PSA risk bins and failures when compared to available tools.

252. Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings*, 13 PSYCHOL. SERV. 206, 214 tbl. 5 (2016).

The DIFRs (which have no associated statistical significance test) also reflected weak discriminatory abilities. DIFRs as obtained here (.25 to .50) are viewed in the risk assessment literature, as small in degree.²⁵³

The AUCs ranged from .56 to .63. The tool thereby ranked failures as higher risk than it ranks successes about 60 percent of the time. If we compute the reciprocals, the PSA exhibited a classification error rate in which about 40 percent of the failures to appear were not ranked in a higher risk bin than the successes. As with the Somers' *d* and the DIFR results, the AUCs indicated limitations in the tool's risk bin classification ability. Using the more generous judgements on effect sizes, these AUCs are regarded as small in size and reflecting a fair performance. But applying the more conservative judgements on the strength of the AUC, as suggested is more appropriate when the decisions have significant consequences to defendants such as earning release,²⁵⁴ the AUC for Illinois indicates failing, while the other two sample AUC rates are poor.

Overall, there exists mixed support for the meaningful ability of the PSA to classify defendants into low, medium, and high risk for failure to appear. The PSA ordinal risk bin strategy showed statistical significance across relevant tests (i.e., odds ratios, overall chi-square statistics for regression models, and Somers' *d*). But several of the associations (i.e., Somers' *d*, DIFRs, and AUCs) were relatively weak from a practical perspective, and the tool underperforms on those measures compared to other risk tool validation studies.

b. Differential Discriminative Validity. The results are clearer and more consistent for the existence of differential validity in unequal discrimination performance of the PSA across samples. Chi-square results indicated

253. Jennifer Skeem et al., *CA-YASI Predictive Utility: How Well do Scores and Classifications Predict Youths' Infractions and Re-Arrest?* 20–21 (2013), http://risk-resilience.berkeley.edu/sites/default/files/journal-articles/files/ca-yasi_predictive_utility_-_how_well_do_scores_and_classifications_predict_youths_infractions_and_arrests_2013.pdf. See also Baglivio et al., *supra* note 69 (citing studies for DIFRs from 1.33 to 1.77 for adult sex offenders) and reporting statistics of .37 and .38 from its own study of a pretrial tool used with adults).

254. Milena Abbiati et al., *Predicting Physically Violent Misconduct in Prison: A Comparison of Four Risk Assessment Instruments*, 37 BEHAV. SCI. & L. 61, 68 (2019).

substantial differences in the failure rates across samples in the low- and medium-risk groups ($p < .001$). The high-risk group showed equivalent failure rates (n.s.).

Using an alternative analysis with the logistic regression models, the risk bin strategy was more or less effective across samples. For example, although the high-risk bin failure rates were equivalent, there were material differences in how well the high-risk bin distinguished from the low-risk bin. Here, the odds of failure for the high-risk bin was between 2.54 and 3.95 times the odds of failure for the low-risk bin, indicating differential validity.

Chi-square analyses to compare sample results revealed differential validity in the strength of the association of the ordinal risk bin scheme overall, as indicated by the Somers' d ranging from .13 to .26 ($p < .001$) and by the lack of AUC parity with results from 56 to 63 percent accuracy in distinguishing failures as higher risk levels than non-failures ($p < .001$).

As the DIFR is not a hypothesis test and these results are not proportions, no exact statistic exists to empirically evaluate if these results significantly vary between samples. Nonetheless, differences in DIFR scores indicated more or less potency in the risk bin classification system's ability to partition into groups with wide distinctions from the overall base rates.

Noticeably, the Kentucky dataset performed stronger in its discriminative ability across all measures of discrimination (i.e., odds ratios, logistic regression model chi-square statistics, Somers' d , DIFRs, and AUCs) with one exception: there was no difference in the odds ratios at the medium-risk bin for Kentucky and New Mexico. The relevance of the tool's achievement is discussed further below. Differential validity by site was also shown in that the consistently weakest discrimination performances across statistical measures existed in the Illinois sample.

2. Predictive Accuracy

The second research question called for an investigation into the predictive accuracy of the PSA. Table 3 provides results for the calibration performance across samples.

We first address the PSA's general abilities regarding its calibration in terms of absolute performance. Then we turn to whether calibration metrics vary across the samples.

Table 3. PSA calibration measures.

	Illinois	Kentucky	New Mexico	χ^2
Hosmer-Lemeshow χ^2 (p)	4.287 (.038)	3.834 (.050)	2.924 (.087)	n/a
Cut-point A: low v. medium/high risk				
Overall accuracy	51.8%	50.9%	46.4%	*
Positive Predictive Value	20.2%	19.4%	22.7%	*
Negative Predictive Value	86.2%	91.2%	89.2%	*
Predicted failure rate	52.1%	56.1%	64.3%	*
Observed failure rate	17.1%	14.8%	18.5%	*
Cut-point B: low/medium v. high risk				
Overall accuracy	80.8%	78.6%	71.1%	*
Positive Predictive Value	28.8%	27.6%	28.1%	*
Negative Predictive Value	83.5%	87.5%	84.5%	*
Predicted failure rate	5.0%	14.8%	23.7%	*
Observed failure rate	17.1%	14.8%	18.5%	*

* $p < .001$; n/a = not applicable.

a. Calibration Ability. The Hosmer-Lemeshow results on the degree of equivalence between predicted and observed rates of failure were mixed. The H-L null hypothesis is that predicted and observed rates are equivalent and thus provide a good model fit. The PSA scale provided a good calibration fit for the data for the New Mexico sample ($p > .05$) and a possibly good fit for Kentucky as the p -value was on the threshold of statistical significance ($p = .05$). In contrast, the H-L test showed a poor calibration fit for the Illinois sample ($p < .05$).²⁵⁵

We calculated the next calibration statistics at two cut-points. Overall accuracy (true positives and true negatives) at Cut-point A ranged from 46 to 52 percent. This result is unimpressive, with the PSA at this threshold

255. Still, the H-L test may not be the final answer. The creators of the H-L test have acknowledged that it might work best on a tool whose outcome is frequent but not too frequent, with the suggestion of base rates between 25–75%. HOSMER, *supra* note 241, at 167. As the failure rates in these three samples were below 20%, one might question the salience of the H-L results here.

operating at or below the level of chance. Raising the threshold to Cut-point B, overall accuracy rose to 71–81 percent. Thus, treating only the high-risk group as predicted failures appreciably enhanced overall fidelity. Still, these statistics are vague in that they cannot reflect how well the tool predicts failures versus non-failures. For this, we move onto the PPVs and NPVs, which represent the forecasting true positive rates and the forecasting true negative rates, respectively.

Cut-point A predicted those in the medium- and high-risk groups to fail to appear. At Cut-point A, the PPVs demonstrate that 19–23 percent of those predicted to fail did fail. In other words, at this threshold, the PSA has a forecasting true positive rate of about 20 percent, thus exhibiting a forecasting false positive rate of about 80 percent. This latter finding presents as an extreme error rate for predicting failure.

In contrast, the NPVs for predicting success for low-risk defendants were far more accurate. At Cut-point A, 86–91 percent of low-risk defendants appeared for their court dates. Thus, the forecasting true negative rate at this threshold is appreciable at about 90 percent.

Cut-point B predicted only high-risk fail. At Cut-point B, the PPVs marginally improved to 27–28 percent. Yet, this also left large false positive error rates of over 70 percent. In other words, the high-risk label is incorrect in about five out of seven cases. But raising the threshold to distinguish the high-risk classification exclusively, is at the expense of slightly lowering the forecasting true negative rates of 84–88 percent (compared to 86–91% at Cut-point A). Hence, when the tool predicts all those in the low- and medium-risk bins to succeed, it exhibits a forecasting false negative rate less than 20 percent.

Overall, these results with forecasting accuracy mean that the PSA functions quite poorly in predicting failure (PPVs) yet has high accuracy rates when predicting success for appearing in court (NPVs). Indeed, the error rates of predicting failure to appear are large, regardless of the cutoff, with false positive rates over 70 percent. In contrast, forecasting false negative rates are much lower at less than 20 percent. These findings are not overly surprising, considering the relatively low base rates of failure in these samples, at less than 20 percent. It is simply harder to forecast an infrequently occurring outcome.²⁵⁶

256. Garrett & Monahan, *supra* note 19, at 456.

We then consider the perspective whereby a tool is well-calibrated using the E/O test of whether its predicted and observed failure rates are similar. A review of results at both cut-points is relevant. The observed failure to appear base rates in the samples ranged from 15 to 19 percent, and are applicable to both cut-points for the E/O test. As shown in Table 3, clearly Cut-point B (in which only high-risk classifications are predicted to fail) is preferable as it is better calibrated for all samples. The Cut-point B expected rates were 5–24 percent, which more closely align with the observed rates of 15–19 percent. Using the lower threshold of Cut-point A (where both medium and high risk are predicted failures), the expected failures rates of 52–64 percent would mean the tool was considerably overcalibrated across samples in seriously overpredicting failure. In sum, the predicted rates of failure that isolate predictions of failure to those scoring as high risk is preferred as being better calibrated to actual results in all samples.

b. Differential Calibration Validity. Differential validity on calibration is next considered. The Hosmer-Lemeshow goodness-of-fit test reflected unequal calibration ability. The H-L test indicated a good calibration model fit for one sample, was equivocal for the second sample, yet tested as a bad model fit for the third.

The existence of differential validity in calibration was confirmed by chi-square analyses indicating statistically significant differences among samples at each cut-point ($p < .001$) for overall accuracy rates, PPVs, NPVs, predicted failure rates, and observed failure rates. The PSA's calibration does not in practice achieve consistent results across samples.

In terms of a tool being well-calibrated under the E/O test when the predicted failure rate is equivalent to the observed failure rate, it was shown earlier that Cut-point B is optimal. Still, variations in performance were found. By predicting all high-risk defendants to fail, the PSA was well-calibrated for Kentucky with the predicted and observed failure rates being equal. But, in the Illinois sample, the PSA seriously underpredicted in that the predicted rate of 5 percent is far below its observed failure rate of 17 percent at that site. In New Mexico, the opposite occurred in that the predicted failure rate of 24 percent overpredicted as this site's observed failure rate was 19 percent.

The H-L test results might appear to conflict with the E/O test. For example, for the Kentucky sample, the H-L test was equivocal on the

calibration properties, but the E/O test at Cut-point B found perfect alliance. The difference is that the two measures are evaluating the tool from different angles. The H-L test computes expected versus observed rates in multiple subgroups and then combines results into a single score. The E/O test statistic, instead, compares the (expected and observed) rates in the sample as an undivided whole. Thus, these measures are simply evaluating different calibration properties of the dataset.

In sum, the evidence consistently presents a conclusion that the PSA exhibits differential calibration validity. Collectively, the foregoing results provide a statistical rendering of a popular risk tool's achievements in the field. It may be of interest to forensic scientists, stakeholders, and practitioners how these findings translate to inform criminal justice policies on the employment of algorithmic risk assessment practices to achieve desired reforms.

B. Policy Implications

The results suggest several policy-related matters of interest. On a positive note, the PSA is able to exhibit some level of utility in its risk bins outcomes. In terms of discriminatory ability, the risk binning strategy is positively correlated with failures to appear, and the tool is more likely than not to rank failures in a higher risk grouping than non-failures. As well, the rates of failure and the odds ratios of failure increase to a fair degree as the risk bins move from low to medium to high.

Another positive and meaningful result is that the PSA performs quite well in predicting success in court date compliance in terms of large accuracy rates (84% and better) for the combined low- and medium-risk classifications. 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 satisfied that individuals will likely appear for their court dates, the PSA FTA scale may thus be conceptualized as a strong tool to initially screen out risk, and thus release, those defendants who are not scored as high risk.

The remainder of the evidence from the results of these analyses requires more caution and suggests the need for relevant adjustments to improve the utility of risk tools to achieve reform goals. First, the results will be informative to stakeholders who might not be cognizant of what level of

probability is associated with each risk classification.²⁵⁷ Indeed, clinicians, statisticians, and/or practitioners share no consensus on what any risk bin means in statistical terms for the likelihood of failure.²⁵⁸ Indeed, a survey of judges who use the PSA in practice revealed that these judges actively wished to understand the relevant probabilities of appearance for defendants with certain risk scores.²⁵⁹ Associated failure to appear rates for the PSA in this study are these: low risk, 9–14 percent, to medium risk, 17–20 percent, to high risk, 28–29 percent. Importantly, there was no overlap in rates across risk bins. Thus, this information can be useful to allow stakeholders a rough idea of the percentile likelihood of failure by category and thereby to manage accordingly. One study, for instance, found that when users were not given base rates associated with the relevant risk levels, they tended to presume a greater likelihood of failure than was actually the case.²⁶⁰ To illustrate further, another survey observed that while some respondents thought that a 30 percent failure rate qualified as high risk, the responses fluctuated dramatically in that others put the relevant rate as high as 90 percent.²⁶¹

These findings are instructive for other reason. Simply put, the risk bins do not always mean the same thing across samples. The probabilities of failure in the high-risk bins are roughly equal. Yet, the medium bin reflects modest variations in failure rates. More problematic is the low-risk category, in which the probabilities across samples, at the extreme, vary by over 50 percent.²⁶² The risk bin assignments, therefore, just do not mean the same likelihood of failure, at least in the low and medium categories. This highlights the idea that risk bins might be better promoted as providing

257. Stevenson, *supra* note 6, 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.”).

258. Neller & Federick, *supra* note 246, at 142; 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).

259. Matthew DeMichele et al., *The Intuitive-Override Model: Nudging Judges Toward Pretrial Risk Assessment Instruments* (2018) (manuscript at 21–22), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3168500.

260. Ashley B. Batastini et al., *Does the Format of the Message Affect What Is Heard? A Two-Part Study on the Communication of Violence Risk Assessment Data*, 19 J. FORENSIC PSYCHOL. RES. & PRAC. 44, 56 tbl. 4 (2019).

261. Klingele, *supra* note 65, at 222.

262. The calculation of the percentage change takes the differential in percentages, divides by the original percent, and multiplies by 100 to derive the total changes in percentiles terms: $((14 - 9)/9) \times 100 = 56\%$.

relative predictions and not as salient for absolutist likelihood estimates.²⁶³

Although the observed rates of failure in the high-risk group were roughly equal, a second critical matter emerges. The PSA's high-risk bin for failure to appear does not isolate to the degree of likelihood of failure that stakeholders might be expecting. Less than 30 percent of defendants classified at high risk failed to appear for a court date. This means that over 70 percent of individuals who were classified as "high risk" of failing to appear were instead successful, keeping all of their court dates. Although no normative consensus exists in the criminal justice risk field on the percentile probability that ought to qualify as high risk,²⁶⁴ this finding at least raises the question of whether stakeholders would regard a less than 30 percent probability as justifying such a consequential status label.²⁶⁵ "[S]igmatic high-risk label[s] . . . translate . . . into action-directives," likely entailing incarceration.²⁶⁶ This means that if the PSA's designation of "high risk" triggers a presumption of pretrial detention, a significant percentage of individuals might be detained unnecessarily. This discovery should inform relevant public debate as to what normatively should qualify as sufficiently risky to justify incarceration for purposes of ensuring individuals comply with court dates.²⁶⁷

Consequently, the base rates of failure in the high-risk bin across sites at less than 30 percent raises the possibility that the tool does not yet isolate a discrete group of defendants who truly pose a far greater risk of failure, such as a more likely than not threshold, which would signify a failure rate of greater than 50 percent. One option is to recalibrate the tool to achieve a high-risk grouping associated with an escalated predicted rate of failure.²⁶⁸

263. Klingele, *supra* note 65, at 219.

264. 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); Oleson et al., *supra* note 258, at 55.

265. Robinson et al., *supra* note 156, at 9 (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.").

266. Stevenson, *supra* note 6, at 336.

267. Stevenson & Mayson, *supra* note 51, at 39.

268. Hamilton et al., *supra* note 90, at 560 ("The key for cut point score selection is to optimize the category demarcations in such a way that the 'high' category captures a substantial portion of those committing the outcome in question and the other 'non-high' categories capture less but still remain relevant.").

Alternatively, efforts may be justified to modify or add predictive risk factors to better distinguish such a group associated with a substantially elevated probability of failing.

This query is even more salient in that the 30 percent rate does not necessarily implicate absconders. The tool defined failure to appear as having missed any single court date. Individuals who complied with all of their court dates except for one (no matter how inconsequential that particular hearing) were counted as failures. Thus, the number could include those who did not appear on schedule because they faced unexpected transportation problems, were wary of losing jobs by taking off work to attend, became unable to obtain appropriate childcare, were sidelined with health issues of their own or family members, or simply forgot.²⁶⁹ In sum, the tool basically misses predicting the person to whom judges likely are trying to isolate: the true flight risk who never returns to face justice.²⁷⁰

The third policy issue regards the findings that the PSA FTA scale performs with extremely high error rates in predicting failure to appear (as opposed to success), and deserves further circumspection. Weak accuracy performance for higher risk attributions signifies that the PSA tool should not be the only data point in decisions *not to* release. Thus, when the tool predicts medium or high risk of failure to appear, the large error rates signify that judges should have less confidence and thus look for more information to supplement the PSA outcomes before determining the individual should be detained in order to assure their court appearances.

No scientific or legal reason appears evident to preclude officials from using a risk tool in a manner that places more emphasis on predictions in one direction than another.²⁷¹ The empirical evidence here strongly supports the utility of the PSA as a “screen out” instrument for success, while at the same time demonstrating that it substantively fails as a “screen in” instrument for likely failures to appear.

269. See Brook Hopkins et al., *Principles of Pretrial Release: Reforming Bail Without Repeating its Harms*, 108 J. CRIM. L. & CRIMINOLOGY 679, 682 (2019); Koepke & Robinson, *supra* note 41, at 1765.

270. Lauryn P. Gouldin, *Defining Flight Risk*, 85 U. CHIC. L. REV. 677, 725 (2018).

271. Hester, *supra* note 125, at 253 (noting “most pertinent risk information may . . . be low-risk information”); Robinson et al., *supra* note 156, at 10–11 (suggesting to practicing judges “risk assessment tools can provide a first screening by which to identify a large group of defendants for immediate and automatic release”).

Fourth, a notable and strong conclusion from this study is that the PSA is not as generalizable across the country as its developers propose. The results were consistent in confirming the presence of differential validity whereby the tool fails to perform substantially similarly across these three diverse jurisdictions. These findings thus raise concerns about the true global accuracy of this tool when used in the field.

These significant differences in various measures of classification accuracy and of predictive ability suggest the presence of site effect. The results confirm that retooling by site is justifiable to tune the PSA's performance to the characteristics and challenges of a particular site's local population, body of criminal laws, and criminal justice practices.²⁷² Indeed, the tool could benefit from recalibration by site, with the results here suggesting the need for adjustments to correct for over-calibration in one jurisdiction and under-calibration in another. Relatedly, the PSA FTA scale predictors and its outcome are all crime-related. It could be that the retooling needs to take into account law-specific characteristics of the site that may affect these measurements of crime. Differences in legislation, criminal justice policies, policing and prosecutorial practices, and judicial norms may skew the PSA variables from location to location.²⁷³ In addition, the retooling might seek to incorporate additional factors on an environmental level that may predict geographical variations in offending that are peculiar to that site.²⁷⁴

It is notable here that the Kentucky dataset showed better ordinal risk bin performance overall, was better at predicting success (i.e., higher NPVs), and was well-calibrated with the E/O test at the higher cut-point. These results are not entirely unforeseeable considering PSA developers worked with Kentucky officials as the first implementation site and made relevant changes, thus potentially refining the tool to the Kentucky experience. This observation is consistent with, and confirms, the expectation that a tool will generally perform better with data on which it was trained. Hence, this study verifies the value of cross-validation research to provide much-needed evidence for algorithmic proficiency as a tool is rolled out across jurisdictions.

272. Hamilton et al., *supra* note 90, at 553 (warning adoption of an “off the shelf” tool may cause issues, and it is thus “important to understand the concepts of adjusting items used, weights, and cut-point placement to address local population issues”).

273. See Campagna et al., *supra* note III (discussing risk tool performance depending on “how justice is conducted differently in different settings”).

274. *Id.*

Nonetheless, considering such ties between the developers and the state's experience with the tool, it might appear curious that the Kentucky sample did not perform more accurately than the other datasets in predicting failure to appear as indicated by PPVs. The forecasting true positive rates were worse than the others, though to a relatively small degree. This finding, though, may be explained by the fact that the overall base rate of failure was the lowest in Kentucky. PPVs are particularly sensitive to base rates,²⁷⁵ often fluctuating in tandem with the rise or fall in base rates.²⁷⁶ Hence it is understandable that this state's lower observed failure rate would result in lower PPVs.

C. Study Limitations

As is expected with empirical research, study limitations should be mentioned. These constraints are largely due to this study being a secondary data analysis where we did not direct or control the collection of data. We thought a trade-off here was reasonable in context. The study limitations presented below are real, but in the world of criminal justice risk assessment, the lack of transparency is strong. Tool owners and criminal justice agencies are notoriously protective of the secrecy of their datasets and practices.²⁷⁷ They rarely grant access to independent researchers, typically rejecting even open records requests, as was done to researchers who attempted to obtain information on the field operation of the PSA.²⁷⁸ Hence, the opportunity to conduct this study at all still appears valuable to provide information points that are not otherwise publicly known. The datasets used herein were not intentionally released so that third parties would be able to mine them to perform validation analyses. Instead, each of them was published as evidence to justify their proponents' arguments that the use of the PSA in the particular jurisdiction was appropriate and yielded productive results.²⁷⁹ We were fortunate that those releases provided sufficient data points to permit the robust analyses offered herein. Reliance

275. Singh, *supra* note 245, at 12.

276. Seena Fazel et al., *Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24,827 People: Systematic Review and Meta-Analysis*, 345 *BMJ* e4692 (2012).

277. Garrett & Monahan, *supra* note 19, at 458.

278. See *supra* note 178 and accompanying text.

279. See *supra* notes 181, 189, and 208 and accompanying text.

upon secondary datasets such as these may at this point provide the best available evidence to contribute to the debate about the efficacy of algorithmic risk assessment. With these potential benefits in mind, we concede the following caveats.

The units of analysis are cases rather than individuals. Thus, it is possible that certain defendants may be represented in any dataset more than once if they were released, rearrested, and released again in the time periods studied. Still, this issue is not uncommon in pretrial risk studies because of the limits of available datasets.²⁸⁰ For privacy concerns, institutions may decline to provide personally identifiable data that would otherwise enable researchers to distinguish individual defendants and thus limit redundancy.

The datasets did not allow for a fixed or consistent follow-up period that might have strengthened analyses. Nonetheless, such a limitation may not be too problematic considering two meta-analyses of criminal justice risk assessment tool performance found that, overall, variations in the length of follow-up did not moderate effect sizes across studies.²⁸¹ Because of our reliance upon secondary data analysis, we could not control for varying circumstances of release (e.g., restrictions, programming, notifications) that may correlate with failure to appear.

Then, no interrater reliability scores were available. This limitation is not overly concerning either, considering that the PSA uniquely uses a small number of predictors, such predictors rely upon official records, and the simple scoring instructions minimize subjective bias. Further, few predictive validity reports include any information on interrater reliability,²⁸² and for this reason, this study is not contrary to industry practices.

Finally, it is also possible that PSA outcomes are plagued by predictive biases for certain sociodemographic subgroups within samples. One dataset did not include relevant information in order to monitor subgroup biases, and in any event, it was beyond the scope of this particular study to delve further into additional research questions such as exploring evidence for disparate impact or discrimination. Future investigation of subgroup differential validity and bias is certainly warranted.

280. DeMichele et al., *supra* note 160, at 18.

281. Fazel et al., *supra* note 276; Singh et al., *supra* note 250, at 509 tbl. 9.

282. Desmarais et al., *supra* note 104, at 14.

CONCLUSIONS

The goal of this study was not to elicit evidence in order to recommend whether the PSA should be used to inform judicial decisions on pretrial detention.²⁸³ That type of value-laden conclusion is best left to policy-makers with motivations to ensure defendants comply with court appearance dates, legal responsibilities to honor defendants' rights, and safety obligations to the public at large.²⁸⁴ Instead, the results herein are intended to provide data and context on how a popular risk tool performs in field settings and how consistent (or inconsistent) its performance stands across sites.

This research expands the existing knowledge base on the validity of risk tools by reporting on results regarding how a popular instrument performs in the field in three jurisdictions. We used techniques similar to other risk assessment validation studies, and supplemented them with those less frequently employed in order to bolster the analyses and offer various perspectives with which to judge a risk assessment tool's performance. Findings of mixed results on discriminative and calibration abilities confirm that reliance upon one or a few performance metrics may offer an overly limited and skewed perspective.

Systematic reviews across sites, such as the one offered here, hold promise for assisting to understand the potential applicability and generalizability of a purportedly "national" prediction model.²⁸⁵ By demonstrating differential validity across samples, the results herein confirm warnings that off-the-shelf risk assessment tools may not perform equally well in local sites. We hope this study encourages local officials to properly validate any tool on their own populations before and after implementation to ensure it performs at what they consider reasonable levels. If not, then potentially

283. Skeptics acknowledging evidence that risk tools may not perform very well still warn this may be the best option available. Douglas et al., *supra* note 17, at 135. "Risk assessment, compared with what?" Skeem & Lowenkamp, *supra* note 29, at 706.

284. To be sure, algorithmic risk assessment is a political beast. Barabas, *supra* note 56 ("In the context of the U.S. Penal System, the term AI has emerged as a contested vehicle for 'evidence-based reform,' the most recent instance in a long history of state efforts to wield statistics as a means of depoliticizing state violence and reasserting legitimacy amid significant social change and unrest.").

285. Thomas P.A. Debray et al., *A Guide to Systematic Review and Meta-Analysis of Prediction Model Performance*, 356 *BMJ* 1, 1 (2017).

appropriate adjustments to classification and calibration properties can improve the utility of an algorithmic scoring model to inform such important decisions as pretrial release.

In the end, despite differential performance across sites, the hopeful last reflection is that there exists an easily scorable and license-free tool that appears to be capable of screening out a significant number of pretrial defendants who, if released, remain very likely to appear for all of their court dates. The utility of such a tool to help in the reform efforts to reduce reliance upon incarceration and on money bail confirms some of the promise of the evidence-based practices movement.