THE BIASED ALGORITHM: EVIDENCE OF DISPARATE IMPACT ON HISPANICS

Melissa Hamilton[*]

I. INTRODUCTION

Automated risk assessment is all the rage in the criminal justice system. Proponents view risk assessment as an objective way to reduce mass incarceration without sacrificing public safety. Officials thus are becoming heavily invested in risk assessment tools—with their reliance upon big data and algorithmic processing—to inform decisions on managing offenders according to their risk profiles.

While the rise in algorithmic risk assessment tools has earned praise, a group of over 100 legal organizations, government watch groups, and minority rights associations (including the ACLU, NAACP, and Electronic

[*] Senior Lecturer of Law & Criminal Justice, University of Surrey School of Law; J.D., The University of Texas at Austin School of Law; Ph.D, The University of Texas at Austin (criminology/criminal justice).

Frontier Foundation) recently signed onto "A Shared Statement of Civil Rights Concerns" expressing unease with whether the algorithms are fair.[1] In 2016, the investigative journalist group ProPublica kickstarted a public debate on the topic when it proclaimed that a popular risk tool called COMPAS was biased against Blacks.[2] Prominent news sites highlighted ProPublica's message that this proved yet again an area in which criminal justice consequences were racist.[3] Yet the potential that risk algorithms are unfair to another minority group has received far less attention in the media or amongst risk assessment scholars and statisticians: Hispanics.[4] The general disregard here exists despite Hispanics representing an important cultural group in the American population considering recent estimates reveal that almost 58 million Hispanics live in the United States, they are the second largest minority, and their numbers are rising quickly.[5]

This Article intends to partly remedy this gap in interest by reporting on an empirical study about risk assessment with Hispanics at the center. The study uses a large dataset of pretrial defendants who were scored on a widely-used algorithmic risk assessment tool soon after their arrests. The report proceeds as follows. Section II briefly reviews the rise in algorithmic risk assessment in criminal justice generally, and then in pretrial contexts more specifically. The discussion summarizes the ProPublica findings regarding the risk tool COMPAS after it analyzed COMPAS scores comparing Blacks and Whites.

---

[1] Ted Gest, *Civil Rights Advocates Say Risk Assessment may "Worsen Racial Disparities" in Bail Decisions*, THE CRIME REPORT (July 31, 2018), https://thecrimereport.org/2018/07/31/civil-rights-advocates-say-risk-assessment-may-worsen-racial-disparities/.

[2] *See infra* Section II.C.

[3] *E.g.*, Li Zhou, *Is Your Software Racist?*, POLITICO (Feb. 7, 2018, 5:05 AM), https://www.politico.com/agenda/story/2018/02/07/algorithmic-bias-software-recommendations-000631; Ed Yong, *A Popular Algorithm is no Better at Predicting Crime than Random People*, THE ATLANTIC (Jan. 18, 2018); Max Ehrenfreund, *The Machines that Could Rid Courtrooms of Racism*, WASHINGTON POST (Aug. 18, 2016), https://www.washingtonpost.com/news/wonk/wp/2016/08/18/why-a-computer-program-that-judges-rely-on-around-the-country-was-accused-of-racism/?noredirect=on&utm_term=.ce854f237cfe; NPR, *The Hidden Discrimination in Criminal Risk-Assessment Scores* (May 24, 2016), https://www.npr.org/2016/05/24/479349654/the-hidden-discrimination-in-criminal-risk-assessment-scores.

[4] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 428 (2016).

[5] Antonio Flores, *How the U.S. Hispanic Population is Changing*, PEW RESEARCH (Sept. 18, 2017), http://www.pewresearch.org/fact-tank/2017/09/18/how-the-u-s-hispanic-population-is-changing/.

Section III discusses further concerns that algorithmic-based risk tools may not be as transparent and neutral as many presume them to be. Insights from behavioral sciences literature suggest that risk tools may not necessarily incorporate factors that are universal or culturally-neutral. Hence, risk tools developed mainly on Whites may not perform as well on heterogeneous minority groups. As a result of these suspicions, experts are calling on third parties to independently audit the accuracy and fairness of risk algorithms. The study reported in Section IV responds to this invitation. Using the same dataset as ProPublica, we offer a range of statistical measures testing COMPAS' accuracy and comparing outcomes for Hispanics versus non-Hispanics. Such measures address questions about the tool's validity, predictive ability, and the potential for algorithmic unfairness and disparate impact upon Hispanics. Conclusions follow.

II. ALGORITHMIC RISK ASSESSMENT

Risk assessment in criminal justice is about predicting an individual's potential for recidivism in the future.[6] Predictions have long been a part of criminal justice decisionmaking because of legitimate goals of protecting the public from those who have already been identified as offenders.[7] Historically, risk predictions were generally based on gut instinct or the personal experience of the official responsible for making the relevant decision.[8] Yet, advances in behavioral sciences, the availability of big data, and improvements in statistical modeling have ushered in a wave of more empirically-informed risk assessment tools.

A. The Rise of Algorithmic Risk Assessment in Criminal Justice

The "evidence-based practices movement" is the now popular term to describe the turn to drawing from behavioral sciences data to improve offender classifications.[9] Scientific studies on recidivism outcomes are benefiting from the availability of large datasets (i.e., big data) tracking offenders post-release to statistically test for factors which that correlate with recidivism.[10] Risk assessment tool developers use computer modeling to

[6] Melissa Hamilton, *Risk and Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 232 (2015).

[7] 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).

[8] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 556 (2015).

[9] Faye S. Taxman, *The Partially Clothed Emperor: Evidence-Based Practices*, 34 J. CONTEMP. CRIM. JUST. 97, 97-98 (2018).

[10] Kelly Hannah-Moffat, *Algorithmic Risk Governance: Big Data Analytics, Race and*

combine factors of sufficiently high correlation and to weight them accordingly with increasingly complex algorithms.[11] Broadly speaking, "[d]ata-driven algorithmic decision making may enhance overall government efficiency and public service delivery, by optimizing bureaucratic processes, providing real-time feedback and predicting outcomes."[12] With such a tool in hand, criminal justice officials can more consistently input relevant data and receive software-produced risk classifications.[13] Dozens of automated risk assessment tools to predict recidivism are now available.[14] They are popular.

The utility of risk instruments has attracted energetic support from reputable policy centers, namely the Justice Center of the Council of State Governments,[15] the Justice Management Institute,[16] the Center for Effective Public Policy,[17] the Vera Institute,[18] and the Center for Court Innovation.[19] News headlines and academic literature have also been expounding upon the benefits generated by the government's use of big data to predict the future risk posed by individuals.[20] Algorithmic risk assessment tools offer the ability

Information Activism in Criminal Justice Debates, THEORETICAL CRIMINOLOGY (forthcoming 2018).

[11] An algorithm refers to "computation procedures (which can be more or less complex) drawing on some type of digital data ("big" or not) that provide some kind of quantitative output (be it a single score or multiple metrics) through a software program." Angéle Christin, *Algorithms in Practice: Comparing Web Journalism and Criminal Justice*, BIG DATA & SOC'Y 1, 2 (July-Dec. 2018).

[12] Bruno Lepri et al., *Fair, Transparent, and Accountable Algorithmic Decision-Making Processes*, PHIL. & TECH. 1, 1 (forthcoming 2018), www.nuriaoliver.com /papers/Philosophy_and_Technology_final.pdf.

[13] J. Stephen Wormith, Automated Offender Risk Assessment, 16 CRIMOLOGY & PUB. POL'Y 281, 285 (2017).

[14] Daniele Kehl et al., *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing* 3, DASH (2017), https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_ 2.pdf?sequence=1.

[15] COUNCIL OF STATE GOV'T, LESSONS FROM THE STATES: REDUCING RECIDIVISM AND CURVING CORRECTIONS COSTS THROUGH JUSTICE REINVESTMENT 6-7 (2013), http://csgjusticecenter.org/jr/publications/lessons-from-the-states.

[16] MAREA BEEMAN & AIMEE WICKMAN, THE JUST. MGMT. INST., RISK AND NEEDS ASSESSMENT 3 (2013).

[17] Ctr. for Effective Pub. Pol'y, *Behavior Management of Justice0Involved Individuals* 19 (2015), https://s3.amazonaws.com/static.nicic.gov/Library/029553.pdf.

[18] Memorandum from Vera Inst. of Just. to Delaware Just. Reinvestment Task Force, Oct. 12, 2011, at 1-2.

[19] MICHAEL REMPEL, CTR. FOR COURT INNOVATION, EVIDENCE-BASED STRATEGIES FOR WORKING WITH OFFENDERS 1-2 (2014).

[20] *E.g.,* Crysta Jentile & Michelle Lawrence, *How Government Use of Big Data can Harm Communities*, FORD FOUNDATION (Aug. 30, 2016), https://www.fordfoundation.org/ideas/equals-change-blog/posts/how-government-use-of-big-data-can-harm-communities/; Sony Kassam, *Legality of Using Predictive Data to Determine Sentences Challenged in Wisconsin Supreme Court Case*, A.B.A J. (June 27,

to reduce mass incarceration by diverting low risk defendants from prison, while targeting greater supervision and services to those at higher risk.[21]

Many parties presume that algorithmic risk assessment tools developed on big data represent a transparent, consistent, and logical method for classifying offenders.[22] The mathematical character of risk assessment suggests the ability to quantify the future and transport it into the present.[23] Evidence-based practices thereby present a welcome displacement of human instinct.[24] Risk assessment practices have been heavily oriented to back-end decisions, such as sentencing, early release decisions, and post-incarceration supervision.[25] More recent attention considers the potential benefits that automated risk assessment practices provide in pretrial settings.[26]

B. Pretrial Risk Assessment

Algorithmic risk assessment informs such pretrial decisions as deferred adjudication and bail.[27] The basic idea of risk-informed decisions for pretrial purposes has a longer trajectory, being first approved by the United States Supreme Court in a 1987 opinion. In *United States v. Salerno*, the Court found constitutional the practice of ordering pretrial detention based on an estimate of the individual defendant's future dangerousness.[28] Still, these predictions must be taken with care because of the potential consequences to individual rights. "Pretrial decision-making involves a fundamental tension between the court's desire to protect citizens from dangerous criminals, ensure that accused individuals are judged before the law, and minimize the amount of pretrial punishment meted out to legally innocent defendants."[29]

The terminology has changed since *Salerno* from the vagueness of "future

---

2016, 1:07 PM), http://www.abajournal.com/news/article/legality_of_using_predictive_data_to_determine_sentences_challenged_in_wisc; Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PENN. L. REV. 327, 407 (2015).

[21] Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings*, 13 PSYCHOL. SCI. 206, 206 (2016).

[22] 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, 725 (2011).

[23] M. Roffey & S.Z. Kaliski, *To Predict or not to Predict-That is the Question*, 15 AFR. J. PSYCHIATRY 227, 227 (2012).

[24] Alfred Blumstein, *Some Perspectives on Quantitative Criminology Pre-JQC: And then Some*, 26 J. QUANTITATIVE CRIMINOLOGY 549, 554 (2010).

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

[26] 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).

[27] MAREA BEEMAN & AIMEE WICKMAN, THE JUST. MGMT. INST., RISK AND NEEDS ASSESSMENT 3 (2013).

[28] United States v. Salerno, 481 U.S. 739, 751 (1987).

[29] THOMAS BLOMBERG ET AL., VALIDATION OF THE COMPAS RISK ASSESSMENT CLASSIFICATION INSTRUMENT 4 (2010).

dangerousness" to a more refined perspective of "risk assessment."[30] Then risk assessment has evolved over the last three decades from unstructured clinical decisions to the actuarial form involving algorithmic processing.[31]

To improve the fairness and effectiveness of pretrial decisions, behavioral science experts encourage officials to use more objective criteria, such as those offered by evidence-based risk tools.[32] Legal reformers generally welcome this practice as well. Risk assessment has become a foundation for the bail reform movement by offering a substitute to a long-standing dependence upon monetary bail.[33] Releasing more defendants who do not pose a substantial risk can alleviate the harms that money bail systems disproportionately wreak on poor and minority defendants.[34] At the same time, reducing the rate of pretrial detention prevents other negative consequences to individual offenders as studies consistently show that pretrial detention is correlated with a greater likelihood of a guilty plea, a longer sentence, job loss, family disruption, and violent victimization in jail.[35]

Notwithstanding the broad support and high hopes for algorithmic risk assessment practices in criminal justice, an investigative report publicized in 2016 called into question their objectivity and fairness.

C. The ProPublica Study

News journalists at ProPublica reported on statistical analyses the group had conducted involving a real dataset and a popular risk tool named COMPAS—the acronym for Correctional Offender Management Profiling for Alternative Sanctions. ProPublica investigators obtained through Freedom of Information Act requests the data on over 7,000 arrestees who were scored on COMPAS in a pretrial setting in a southern county of Florida.[36] ProPublica concluded COMPAS was racist in that its algorithm

---

[30] KIRK HEILBRUN, EVALUATION FOR RISK OF VIOLENCE IN ADULTS 708 (2009).

[31] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537 (2015).

[32] Arthur Rizer & Caleb Watney, *Artificial Intelligence can Make our Jail System More Efficient, Equitable and Just,* TEX. REV. L. & POLITICS 1, 5 (forthcoming 2018).

[33] Megan Stevenson, *Assessing Risk Assessment*, 102 MINN. L. REV. (forthcoming 2018), https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3139944_code2420348.pdf?abstractid=3016088&mirid=1.

[34] John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, WASH. L. REV. (forthcoming 2018), https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3142948_code2621669.pdf?abstractid=3041622&mirid=1.

[35] Laura I. Appleman, *Justice in the Shadowlands: Pretrial Detention, Punishment, & the Sixth Amendment*, 69 WASH. & LEE L. REV. 1297, 1320 (2012).

[36] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

produced a much higher false positive rate for Blacks than Whites, meaning that it overpredicted high risk for Blacks.[37]

COMPAS's corporate owner, Northpointe, quickly rejected such characterization.[38] After running their own statistical analyses on the same dataset ProPublica had compiled, Northpointe statisticians asserted that their results demonstrated COMPAS outcomes achieved predictive parity for Blacks and Whites.[39]

It turns out a rather simple explanation accounts for the dispute: contrasting measures of algorithmic fairness. ProPublica touted the false positive rate, while Northpointe preferred an alternative measure called the positive predictive value.[40] As will be addressed later below, these measures are not synonymous and offer distinct, sometimes conflicting, impressions of a tool's accuracy.[41]

III. AUDITING THE BLACK BOX

For purposes here, a "black-box" tool refers to an algorithmic risk instrument which is not transparent about what is input into the software program and/or how the outputs are generated and quantified.[42] This characterization is more probably appropriate in the case of an algorithmic instrument that is proprietary and its owner declines to reveal much information based on a claim of trade secrets.[43] COMPAS, for example, is proprietary and its corporate owner declines to reveal its algorithm, which likely is a reason for ProPublica's interest in auditing it.

A. Calls for Third Party Audits

ProPublica's study certainly brought the issue of algorithmic fairness to the forefront in the popular media.[44] Questions are being raised in the

---

[37] *Id*.

[38] Northpointe rebranded with the trade name equivant (lower case intended) in January 2017. Press Release, equivant, Courtview, Constellation & Northpointe Re-brand to equivant (2017), http://www.equivant.com/blog/we-have-rebranded-to-equivant.

[39] William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity* 2 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[40] Tafari Mbadiwe, *Algorithmic Injustice*, THE NEW ATLANTIC 1, 18 (Winter 2018), https://www.thenewatlantis.com/publications/algorithmic-injustice.

[41] *Infra* Section IV.F.

[42] ROBYN CAPLAN ET AL., ALGORITHMIC ACCOUNTABILITY: A PRIMER 2-3 (Apr. 18, 2018), https://datasociety.net/pubs/alg_accountability.pdf.

[43] Sarah Tan et al., Auditing Black-Box Models Using Transparent Model Distillation with Side Information, Paper presented at NIPS 2017 Symposium on Interpretable Machine Learning (2017), https://arxiv.org/pdf/1710.06169.

[44] *See e.g.*, Li Zhou, *Is Your Software Racist?*, POLITICO (Feb. 7, 2018, 5:05 AM), https://www.politico.com/agenda/story/2018/02/07/algorithmic-bias-software-recommendations-000631.

scientific and policy communities that even "reasonable algorithms" may fail to result in fair and equitable treatment of diverse populations.[45] There is even a new scientific literature on the topic named FATML, or "fairness, accountability and transparency in machine learning."[46] There is some overlap among the FATML goals:

> Fairness can . . . be related to the notion of transparency – the question of how much we are entitled to know about any automated system that is used to make or inform a decision that affects us. Hiding the inner workings of an algorithm from public view might seem preferable, to avoid gaming the system. But without transparency, how can decisions be probed and challenged?[47]

Importantly, no formal mechanism in the law or in the sciences exists to consistently enforce any form of algorithmic accountability.[48] Thus, observers call for third party auditing, such as exemplified by ProPublica's efforts, to engage in any form of scientific inquiry that may be possible to reveal information about the empirical validity and fairness of black-box tools.[49] Such data will be useful to legal practitioners and policymakers in considering—or reevaluating—the use of automated risk assessment to inform criminal justice decisions which carry significant consequences for individuals.[50]

Moreover, despite the many advantages of algorithmic assessment, risk profiling may fail to alleviate all of the harms of mass incarceration as some "scholars are suspicious that contemporary extensions of risk assessment and risk reduction will likely only reproduce, or may even exacerbate, the injustices of contemporary criminal justice policy under a more 'objective' guise."[51] For instance, a White House Report on Big Data from the Obama administration lauded the public benefits of big data in criminal justice, but also promoted academic research into big data systems "to ensure that people are treated fairly."[52]

---

[45] OSONDE OSABA & WILLIAM WELSER IV, AN INTELLIGENCE IN OUR IMAGE: THE RISKS OF BIAS AND ERRORS IN ARTIFICIAL INTELLIGENCE 19 (2017) (Rand Corporation publication), https://www.rand.org/pubs/research_reports/RR1744.html.

[46] Harsh Gupta, *Constitutional Perspectives on Machine Learning* 4 (Dec. 17, 2017), https://osf.io/preprints/socarxiv/9v8js/download?format=pdf.

[47] Sofia Olhede & Patrick Wolfe, *When Algorithms go Wrong, Who is Liable?*, 14 SIGNIFICANCE 8, 9 (2017).

[48] *See* Robyn Caplan et al., *Algorithmic Accountability: A Primer* 10 (Apr. 18, 2018), https://datasociety.net/pubs/alg_accountability.pdf.

[49] *See Id*.

[50] Jennifer Skeem et al., *Gender, Risk Assessment, and Sanctioning*, 40 LAW & HUM. BEHAV. 580, 590 (2016).

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

[52] Executive Office of the President, Big Data: A Report on Algorithmic Systems,

A particularly acute focus is the concern that potential unfairness will likely fall mostly upon already beleaguered minority groups. "The use of big data analytics might impose disproportionate adverse impacts on protected classes, even when organizations do not intend to discriminate and do not use sensitive classifiers like race and gender."[53] Moreover, cross-disciplinary sharing is necessary because of difficulties in translation where data scientists are often not trained in law and policy, while civil rights experts in turn may not have statistical expertise.[54] In sum, many are just realizing that big data analytics can create civil rights problems,[55] such that the "patina of fairness" that otherwise seems to attach to big data algorithms may be unjustified.[56] Thus, to the extent that justice decisions may bring negative consequences upon defendants, it is particularly advisable to study whether and how an algorithmic tool disparately impacts protected groups.[57]

B. Black-Box Tools and Ethnic Minorities

The ProPublica study was concerned with Black minorities. This paper focuses on the ethnic minority group of Hispanics. This Article follows the tradition of the United States Census Bureau in classifying Hispanics as an ethnicity rather than a race.[58]

Importantly, Hispanics comprise a significant proportion of the American population, making them a reasonable population to analyze. Plus, reasons exist to suspect that an algorithm may not assess an ethnic minority group very well.

"[A] transparent, facially neutral algorithm can still produce discriminatory results."[59] Even if a particular tool is shown to perform well on its training sample(s), it is not advisable to simply transport that tool across to new populations and settings because of the potential for risk-relevant differences in offenders and the availability of rehabilitation-oriented services that can undermine the tool's performance.[60] Unfortunately, officials

---

Opportunity, and Civil Rights 22-23 (May 2016), https://obamaWhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

[53] Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUM. L. REV. 67, 79 (2017).

[54] Solon Barocas et al., *Big Data, Data Science, and Civil Rights* 6 (2017), https://arxiv.org/pdf/1706.03102.

[55] *Id*. at 1.

[56] Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1034 (2017).

[57] Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67, 67 (2017).

[58] U.S. Census Bureau, *Race and Ethnicity*, https://www.census.gov/mso/www/training/pdf/race-ethnicity-onepager.pdf.

[59] Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1024 (2017).

[60] Sarah L. Desmarais et al*., Performance of Recidivism Risk Assessment Instruments in*

often disregard this advisory. Recent experience informs that "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."[61]

Another reason for suspicion is that, despite their disproportionate presence in criminal justice statistics, minorities tend to be underrepresented in testing or validation samples for most risk assessment tools.[62] Yet it is unreasonable to assume risk assessment tools will perform as well for culturally diverse minority groups.[63] A risk assessment process that presumes that risk tools are somehow universal, generic, or culturally-neutral may well be flawed on the following grounds:

> The over or under estimation of risk that can ensue from this process is entirely plausible given (a) the potential omission of meaningful risk items specific to minority populations, (b) the inclusion of risk factors that are more relevant to White offenders, and (c) variation in the cross-cultural manifestation and expression of existing risk items.[64]

For example, a risk tool may yield unequal results for cultural minorities if it fails to incorporate or otherwise consider their unique "behavioral practices and expectations, health beliefs, social/environmental experiences, phenomenology, illness narratives, deviant conduct, and worldview."[65] It is not surprising, then, when risk tools are originally normed on largely White samples that at least some studies show risk tools provide more accurate predictions for Whites than other groups.[66]

Regrettably, relatively few validation studies on ethnic minorities exist.[67] Two recent validation studies that have included an ethnic minority group may be telling. The data reported with revalidation studies of the federal Pretrial Risk Assessment and the Post Conviction Risk Assessment tools show that each tool overpredicted or underpredicted recidivism for Hispanics

---

U.S. Correctional Settings, 13 PSYCHOL. SERV. 206, 207 (2016).

[61] Seth J. Prins & Adam Reich, *Can we Avoid Reductionism in Risk Reduction?*, 22 THEORETICAL CRIMINOLOGY 258, 260 (2018) (internal citations omitted).

[62] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 428 (2016).

[63] *Id*.

[64] *Id*. at 429.

[65] *Id*.

[66] Jay P. Singh et al., *Comparative Study of Violence Risk Assessment Tools: A Systematic Review and Metaregression Analysis of 68 Studies Involving 25,980 Participants*, CLINICAL PSYCHOL. REV. (2011); Jay P. Singh & Seena Fazel, *Forensic Risk Assessment: A Metareview*, 37 CRIM. JUST. & BEHAV. 965, 978 (2010).

[67] T. Douglas et al., *Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data*, 42 EUR. PSYCHIATRY 134, 135 (2017).

for at least a few of the recidivism outcomes tested.[68] Despite the seeming importance of these findings, the authors do not expound upon these results.

With respect specifically to Hispanic Americans, academics have suggested that risk assessment tools may not perform well if they fail to "consider the centrality of family, acculturation strain, religiosity, gender role expectations, and culturally stoic responses to adversity" unique to this particular cultural group.[69]

C. Legal Challenges

The legal and big data communities are beginning to realize that there may be legal implications in terms of disparate impact if algorithmic bias exists.[70] One commentator, while acknowledging the potential for unfairness in risk assessment-led decisions in criminal justice, has suggested that the practice at least is better than uninformed judgments about dangerousness and that the adversarial process will encourage attorneys to challenge the tools, which may yield improvements to the science underlying them.[71] Yet, so far there is very little evidence of such confrontations in the courts. An exception is a 2016 case, in which a defendant protested the black-box nature of COMPAS (the same tool in ProPublica's research and used in the study presented herein).

In *Loomis v. Wisconsin*, the defendant claimed the use of COMPAS in his sentencing proceeding constituted a due process violation because the software owner's claims of trade secrets prevented him from challenging the tool's validity or its algorithmic scoring.[72] The Wisconsin Supreme Court found against Loomis, but with some caveats.[73] The majority ruled that any use of COMPAS must come with certain written cautions, including warnings regarding the proprietary (hence, secretive) nature of the tool and

---

[68] Thomas A. Cohen & Christopher Lowenkamp, *Revalidation of the Federal Pretrial Risk Assessment Instrument (PTRA): Testing the PTRA for Predictive Biases* 23 (Feb. 2018), https://www.researchgate.net/publication/322958782_Revalidation_of_the_Federal_Pretrial_Risk_Assessment_Instrument_PTRA_Testing_the_PTRA_for_Predictive_Biases; Christopher T. Lowenkamp, *PCRA Revisited: Testing the Validity of the Federal Post Conviction Risk Assessment (PCRA)*, 12 PSYCHOL. SERV. 149 tbl. 5 (2015).

[69] Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 498 (2016).

[70] Osonde Osaba & William Welser IV, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*, RAND 19 (2017) https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf.

[71] David E. Patton, *Guns, Crime Control, and a Systemic Approach to Federal Sentencing*, 32 CARDOZO L. REV. 1427, 1456-58 (2011).

[72] State v. Loomis, 881 N.W.2d 749, 753 (Wisc. 2016).

[73] *Id*. at 769.

the potential that it may disproportionately classify minorities as high risk.[74] The court seemed marginally concerned with the fact that COMPAS had never been validated on a Wisconsin-based population. The court suggested that more or less weight might be put on a COMPAS score if such a localized validation study were to be conducted.[75]

In contrast, a very recent decision by the Canadian Supreme Court found the lack of a relevant validation study dispositive. In *Ewert v. Canada*, the defendant, an indigenous minority, successfully protested the use of a different risk assessment tool (i.e., not COMPAS) because of the nonexistence of any validation studies of the tool specifically on indigenous groups.[76] The Canadian Supreme Court, ruling in Ewert's favor, determined that without evidence that the particular tool was free of cultural bias, it was unjust to use it on indigenous inmates.[77] The justices observed that "substantive equality requires more than simply equal treatment" as treating groups identically may itself produce inequalities.[78]

Despite the divergence of the rulings in *Loomis* and *Ewert*, both courts supported the relevance of validation studies appropriate to the underlying population(s) on which the tool is to be used. According to the 100 plus groups of civil rights and defense counsel organizations in their "Shared Statement of Civil Rights Concerns," the stakes are significant. "If the use of a particular pretrial risk assessment instrument by itself does not result in an independently audited, measurable decrease in the number of people detained pretrial, the tool should be pulled from use until it is recalibrated to cause demonstrably decarceral results."[79] Moreover, this Shared Statement encourages defense attorneys to inspect any risk tool to review how it was created, the input factors used, weights assigned to the factors, thresholds for risk bins, and outcome data used in its development.[80]

The next Section responds to the call for third-party auditing by academics to evaluate how a tool may perform on a minority group.

---

[74] *Id.*

[75] *Id.*

[76] Ewert v. Canada, 2018 S.C.R. 30 (S.C.C. June 13, 2018).

[77] *Id*. at ¶63.

[78] *Id*. at ¶54.

[79] African American Ministers in Action et al., *The Use of Pretrial "Risk Assessment" Instruments: A Shared Statement of Civil Rights Concerns* 5 (2018), http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf.

[80] *Id*. at 7.

IV. EVALUATING ALGORITHMIC FAIRNESS WITH ETHNICITY

This section reports on a study of COMPAS, a popular algorithmic risk assessment tool, employing the same real-world dataset that ProPublica compiled for its evaluation of outcomes for Blacks versus Whites.[81] The data and the tool will be briefly addressed. Then a variety of empirical methods assess the accuracy, validity, and predictive ability of the tool and further to determine whether algorithmic unfairness or disparate impact appear present.

A. The Samples and the Test

The primary dataset includes individuals arrested in Broward County, Florida who were scored on a COMPAS risk scale soon after their arrests in 2013 and 2014.[82] Notably, Broward County is among the top twenty largest American counties by population, thus improving the potential for a large and diverse sample set. The pretrial services division of the Broward County Sheriff's Department has been using COMPAS since 2008 to inform judicial determinations concerning pretrial release.[83] This study uses two subsets of data, one of which tracks general recidivism (n=6,172) and the other violent recidivism (n=4,020). The follow-up recidivism period is two years.

COMPAS is a software application widely used across correctional institutions, offering a general recidivism risk scale and a violent recidivism risk scale.[84] The general recidivism risk scale contains about two dozen items related to age at first arrest, age at intake, criminal history, drug problems, and vocational/educational problems (e.g., employment, possessing a skill or trade, high school grades, suspensions).[85] The violent recidivism scale differs in that in lieu of criminal history generally it uses a history of violence and instead of drug problems, it incorporates factors related to a history of non-compliance (e.g., previous parole violations, arrested while on probation).[86]

The COMPAS algorithms produce outcomes as decile scores of 1-10 with

---

[81] COMPAS is the "most successful and popular application of machine learning." J. Stephen Wormith, *Automated Offender Risk Assessment*, 16 CRIMINOLOGY & PUB. POL'Y 281, 285 (2017) (citing developers using a decision tree model to educate the tool).

[82] *See generally* Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. ProPublica has generously made the data available for other researchers to access. https://github.com/propublica/compas-analysis.

[83] THOMAS BLOMBERG ET AL., VALIDATION OF THE COMPAS RISK ASSESSMENT CLASSIFICATION INSTRUMENT 15-16 (2010).

[84] EQUIVANT, COMPAS CLASSIFICATION 2 (2017), http://equivant.volarisgroup.com/assets/img/content/Classification.pdf.

[85] NORTHPOINTE, COMPAS CORE NORMS FOR ADULT INSTITUTIONS 80 tbl. 3.41 (Feb. 11, 2014), https://epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-WIDOC_DAI_norm_report021114.pdf.

[86] NORTHPOINTE, PRACTITIONER'S GUIDE TO COMPAS CORE 28, 29 (2015).

higher deciles representing greater predicted risk. COMPAS then subdivides decile scores into three, ordinal risk bins: low (deciles 1-4), medium (deciles 5-7), and high (deciles 8-10).

These analyses generally followed the methodology of the ProPublica researchers in terms of defining what acts comprise general or violent recidivism, excluding cases with missing data, and excluding cases where the individuals were not scored on COMPAS in a timely manner.[87]

B. Differential Validity Measures

Validity, while a technical term, simply means the extent to which a test properly reflects the concept it is designed to reflect.[88] For purposes here, validity asks whether COMPAS adequately measures recidivism risk. The term *differential validity* applies when test validity varies between groups.[89]

Multiple measures related to validity are available to judge the diagnostic and prognostic capabilities of an assessment tool. *Discrimination* reflects how well the tool distinguishes higher risk from lower risk (i.e., relative accuracy).[90] *Calibration* concerns how accurate the tool statistically estimates the outcome of interest (i.e., absolute predictive accuracy).[91]

A few descriptive statistics of the samples are provided in Figure 1.

Figure 1. Descriptive Statistics (means or proportions)

	General Recidivism Sample		Violent Recidivism Sample	
Factor	*Hispanic*	*Non-Hispanic*	*Hispanic*	*Non-Hispanic*
Recidivism Rate	.37	.46***	.10	.17***
Decile Score	3.38	4.51***	2.64	3.33***
No. Prior Offenses	2.10	3.35***	1.70	2.52***
Age	35.02	34.49	36.14	35.70
Females	.16	.19	.17	.21
n	509	5663	355	3665

Notes: * $p<.05$, ** $p<.01$, *** $p<.001$.

Figure 1 shows that general and violent recidivism rates are substantially

[87] *See generally* Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[88] MICHAEL G. MAXFIELD & EARL BABBIE, RESEARCH METHODS IN CRIMINAL JUSTICE AND CRIMINOLOGY 109 (2nd ed. 1998).

[89] Christopher M. Berry et al., *Can Racial/Ethnic Subgroup Criterion-to-Test Standard Deviant Ratios Account for Conflicting Differential Validity and Differential Prediction Evidence for Cognitive Ability Tests?*, 87 J. OCCUPATIONAL & ORG. PSYCHOL. 208, 209 (2014).

[90] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 11 (2017).

[91] *Id*.

lower for Hispanics, that is at 20 and 41 percent lower than non-Hispanics, respectively. Decile score outcomes on both scales for Hispanics were also substantially lower. Lower mean decile scores for Hispanics can partly be explained by their also having lower mean numbers of prior offences as each scale includes factors that count criminal history.

The table also includes a comparison on age and gender as these factors have been shown in previous studies to be strong predictors of recidivism.[92] Here, test comparison results between groups of the average age of offenders and their gender makeup are not statistically significant. Thus, these results generally indicate that any differential validity between these groups is not obviously a result of age or gender disparities.

Figure 2 includes statistics to explore the degree (i.e., the strength) of the relationship between COMPAS scores and recidivism for each scale and compares Hispanics with non-Hispanics.

Figure 2. COMPAS Failure Rates

	General Recidivism Sample		Violent Recidivism Sample	
	Hispanic	*Non-Hispanic*	*Hispanic*	*Non-Hispanic*
Risk Level				
Low	.30	.32	.09	.11
Medium	.55	.55	.08	.27***
High	.57	.75***	.46	.49
AUC	.637	.714***	.641	.722**
r	.238	.372**	.148	.319***

Notes: * $p<.05$, ** $p<.01$, *** $p<.001$.

The tabular statistics in Figure 2 contain a host of relevant information about COMPAS' performance.

1. Validity of Risk Bins

Three rows display actual rates of recidivism by offenders classified in each of the COMPAS risk bins of low, medium, and high. The results show both intergroup and intragroup disparities. Significant intergroup differences in recidivism rates occur in the high risk classification for the general recidivism risk scale (57% for Hispanics versus 75% for non-Hispanics) and in the medium risk bin for the violent recidivism scale (8% for Hispanics versus 27% for non-Hispanics). Thus, these two risk bins perform disparately regarding reoffending based on ethnicity.

For the non-Hispanics group, actual recidivism rates operate in a

[92] *See generally* David E. Olson et al., *Comparing Male and Female Prison Releasees across Risk Factors and Postprison Recidivism*, 26 WOMEN & CRIM. JUST. 122 (2016) (listing studies).

monotonic fashion as risk bins rise (low-medium-high) with substantively increasing rates of recidivism on each scale, thereby exemplifying favorable performance. Contrastingly, this is not the case for Hispanics. Highlighted within Table 2, one can observe that with the general recidivism scale, the recidivism rate for Hispanics is only slightly higher in the high risk bin than the medium bin, with the difference not statistically significant. Then with the violent recidivism scale, the rate for Hispanics unexpectedly decreases from the low to the medium bins. Hence, these measures confirm differential validity in that COMPAS fails to perform equivalently for these groups. Indeed, the three risk bin strategy collapses for Hispanics. (A two risk bin scheme appears more appropriate for Hispanics whereby the medium and high risk bins could be combined for general recidivism while a low and medium risk bin combination might better fit the data for violent recidivism.)

2. The Area Under the Curve

The next row in Figure 2 involves a metric called the area under the curve (AUC)—it is derived from a statistical plotting of true positives and false positives across a risk tool's rating system.[93] More specifically, an AUC is a discrimination index that represents the probability that a randomly selected recidivist received a higher risk classification than a randomly selected non-recidivist.[94] AUCs range from 0-1.0 with 0.5 indicating no better accuracy than chance and a 1.0 meaning perfect discrimination.[95] For this statistic, COMPAS decile scores are utilized, rather than the ordinal bins.

Risk assessment scholars often refer to AUCs of .56, .64, and .71 as the thresholds for small, medium, and large effect sizes, respectively.[96] Using these suggestions, it would mean that COMPAS operates effectively for both groups, albeit at disparate magnitudes. The AUCs would represent moderate effect sizes for Hispanics and large effect sizes for non-Hispanics. Notwithstanding, agreement on the strength of AUCs is not universal.[97] A more conservative conceptualization is that AUCs between .60 and .69 are poor, from .70 to .79 are fair, .80 to .89 are good, and over .90 are excellent.[98]

---

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

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

[95] Martin Rettenberger et al., *Prospective Actuarial Risk Assessment: A Comparison of Five Risk Assessment Instruments in Different Sexual Offender Subtypes,* 54 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 169, 176 (2010).

[96] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 12 (2017).

[97] Jay P. Singh, *Five Opportunities for Innovation in Violence Risk Assessment Research*, 1 J. THREAT ASSESSMENT & MGMT. 179 181 (2014).

[98] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 9 (2017).

Based on these standards, the AUCs for Hispanics would be judged as poor and for non-Hispanics as fair.

Still, the AUC has serious limitations as discussed elsewhere and thus cannot present a holistic portrait of a tool's abilities.[99] Unfortunately, the AUC is too commonly misinterpreted as measuring calibration accuracy; but a higher AUC does not mean more accurate prospective prediction.[100] Further, the AUC cannot calculate how well an instrument selects those at high risk.[101] For example, the AUC could be very high even if no recidivists were ranked as high risk. To use a hypothetical, the AUC for COMPAS would actually reflect perfect accuracy (AUC=1.0) where all recidivists were classified Decile 2 and all non-recidivists as Decile 1 (i.e., all were classified as low risk), with very little distinction considering the decile scale ranges from 1 to 10.

Importantly, substantial group differences indicated by corresponding *z*-tests for AUCs will indicate the presence of differential discrimination accuracy in terms of degree.[102] Here, the AUCs indicate that COMPAS is a weaker discriminatory tool for Hispanics than non-Hispanics as the AUC measures on both the general and violent recidivism tools are lower for Hispanics, and with statistical significance (*z*-tests of the differences in proportions indicate levels of p=.000 and p=.001 for general and violent recidivism, respectively). These results on the AUCs are another indicator that COMPAS exhibits differential accuracy in its degree of discrimination utility on both scales relative to Hispanic ethnicity.

3. Correlations

The final row in Figure 1 contains correlation coefficients representing another barometer of the strength of COMPAS decile scores relative to recidivism. Like the AUC, correlations test discrimination ability.[103] Separate

---

[99] Jay P. Singh, *Predictive Validity Performance Indicators in Violent Risk Assessment*, 31 BEHAV. SCI. & L. 8, 16-18 (2013); Stephane M. Shepherd & Danny Sullivan, *Covert and Implicit Influences on the Interpretation of Violence Risk Instruments*, 24 PSYCHIATRY, PSYCHOL. & L. 292, 294 (2017).

[100] Jay P. Singh, *Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer*, 31 BEHAV. SCI. & L. 8, 16 (2013). *See e.g.,* Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*," 80 FED. PROBATION 38, 41 (2016) (incorrectly referring to the AUC of COMPAS using the Broward County data as indicating "predictive accuracy").

[101] Jay P. Singh, *Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer*, 31 BEHAV. SCI. & L. 8, 17 (2013).

[102] Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses*, 80 FED. PROBATION 38, 41 (2016).

[103] L. Maaike Helmus & Kelly M. Babchishin, *Primer on Risk Assessment and the Statistics Used to Evaluate its Accuracy*, 44 CRIM. JUST. & BEHAV. 8, 11 (2017).

statistical runs (not shown in Figure 2) find that each of the four individual correlation coefficients are statistically significant at p=.005 or below. These individual correlations thus signify that the scales achieve some statistical strength relative to recidivism in both groups. Notwithstanding, correlation coefficients can signify differential validity as well.[104] Here, comparing correlation statistics shows that COMPAS scales are more weakly associated with general and violent recidivism for Hispanics, and that these differences are statistically significant. In sum, this result adds to the cumulative evidence of differential validity for COMPAS regarding Hispanics.

C. Test Bias

A well calibrated instrument is one that is also "free from *predictive bias* or *differential prediction*."[105] Differential prediction demonstrates group disparities in predictive ability.

Researchers examining group bias in psychological testing in education have standardized a methodology to empirically study it, with the endorsement of the American Psychological Association.[106] Group bias represents test bias, which refers to the existence of systematic errors in how a test measures members of any group.[107] This gold standard for evaluating test bias involves a series of nested models of regression equations involving the test, the group(s) of interest, and an interaction term (test*group) as predictors of test outcomes.[108] Basically, a regression analysis is a statistical method to evaluate the relationship between one or more predictors with a response (outcome) variable.[109] Then, an interaction term refers to the product of two predictor variables (here, test and group) to determine whether the effect on the outcome of either predictor is moderated by the presence of

---

[104] Christopher M. Berry et al., *Can Racial/Ethnic Subgroup Criterion-to-Test Standard Deviant Ratios Account for Conflicting Differential Validity and Differential Prediction Evidence for Cognitive Ability Tests?*, 87 J. OCCUPATIONAL & ORG. PSYCHOL. 208, 209 (2014).

[105] Alexandria Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 154 (2017) (emphasis in original).

[106] Nathan R. Kuncel & Davide M. Kleiger, *Predictive Bias in Work and Educational Settings, in* THE OXFORD HANDBOOK OF PERSONNEL ASSESSMENT 462, 463 (Neil Schmitt ed., 2012) (confirming endorsements also from the National Council on Measurement in Education and the American Educational Research Association).

[107] Adam W. Meade & Michael Fetzer, *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*, 12 ORG. RES. METHODS 738, 738 (2009).

[108] Jeanne A. Teresi & Richard N. Jones, *Bias in Psychological Assessment and Other Measures, in* APA HANDBOOK OF TESTING AND ASSESSMENT IN PSYCHOLOGY 139, 144 (vol. 1 2013).

[109] RONET D. BACHMAN & RAYMOND PATERNOSTER, STATISTICS FOR CRIMINOLOGY AND CRIMINAL JUSTICE 675 (1997).

the other (i.e., changes its strength or direction of influence).[110]

This nested models method detects group differences in the form of the relationship between the test and the outcome in terms of intercepts and slopes[111] in order to reveal differential prediction.[112] The rule of thumb in the psychological assessment field is that a significant group difference in either the intercept or the slope represents that a single regression equation for the groups combined will predict inaccurately for one or both groups, and therefore a separate equation for each group must be considered.[113] Unequal intercepts or slopes also signify *disparate impact*.[114] A system that treats persons unfairly suggests disparate impact may be present without requiring evidence of any discriminatory intent.[115] Selected researchers in criminal justice have recently begun to apply this methodological practice of nested models to evaluate group bias in recidivism risk tools.[116]

The nested model structure here utilized variables labeled as Hispanic (coded as Hispanic=1, non-Hispanic=0), the COMPAS decile score, and an interaction between them as Hispanic*COMPAS decile score. A four model structure is employed with the outcome variable being recidivism. Model 1 tests only the Hispanic variable; Model 2 tests just COMPAS Decile score; Model 3 includes both the Hispanic and COMPAS decile score variables; and Model 4 retains Hispanic and COMPAS decile score while adding the interaction term. The general recidivism model utilizes the COMPAS general recidivism scale while the violent recidivism model relies upon the COMPAS violent recidivism scale.

Figures 3 and 4 provide the relevant logistic regression equation results for general recidivism and violent recidivism, respectively. *Logistic* regression applies when investigating an association between one or more predictor variables with an outcome of interest that is dichotomous in nature

---

[110] JAMES JACCARD, INTERACTION EFFECTS IN LOGISTIC REGRESSION 12 (2001).

[111] Jennifer L. Skeem & Christopher T. Lowenkamp, *Race, Risk, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 692 (2016).

[112] Adam W. Meade & Michael Fetzer, *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*, 12 ORG. RES. METHODS 738, 740 (2009).

[113] Cecil R. Reynolds & Lisa A. Suzuki, *Bias in Psychological Assessment: An Empirical Review and Recommendations, in* HANDBOOK OF PSYCHOLOGY 82, 101 (Irving B. Weiner ed., 2003).

[114] Adam W. Meade & Michael Fetzer, *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*, 12 ORG. RES. METHODS 738, 741 (2009).

[115] Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109. 121-22 (2017).

[116] Jennifer Skeem et al., *Gender, Risk, Assessment, and Sanctioning: The Cost of Treating Women Like Men*, 40 LAW & HUM. BEHAV. 580, 585 (2016).

(e.g., recidivist=yes/no).[117] The logistic coefficients have been translated in the tables as odds ratios for interpretive purposes.

Figure 3. Logistic Regression Predicting the Odds of General Recidivism

Predictor	Model 1	Model 2	Model 3	Model 4
Hispanic	.686***	---	.914	1.284
Decile	---	1.323***	1.322***	1.331***
Hispanic*Decile Interaction	---	---	---	.911*
Constant	.861	.239	.242	.235
-2LL	8490.49	7650.88	7650.11	7644.43
χ^2	15.93	855.53	856.30	861.99
n=6,172				

Notes: * $p<.05$, ** $p<.01$, *** $p<.001$. Coefficients represent odds ratios.

Model 1 indicates that Hispanics are significantly less likely to engage in any act of recidivism. Model 2 supports the overall utility of COMPAS for the groups combined in that the odds of general recidivism increases 32% for every one unit increase in COMPAS decile score. In Models 3 and 4, when adding the decile score, Hispanic ethnicity is no longer statistically significant. While the odds are less for Hispanics in Model 3, such that there is evidence of overprediction for them, it is not statistically significant (p=.380). However, the interaction term in Model 4 is statistically significant (p=.015). This means that Hispanic ethnicity significantly moderates the relationship between COMPAS decile score and general recidivism. As the interaction term is below one, the regression slope is less steep for Hispanics, revealing that as the COMPAS score increases, it has weaker influence on predicting recidivism for Hispanics. In other words, COMPAS does not predict as strongly for Hispanics and thus is not as valid a predictor for that group.[118] Further, unequal slopes reflect differential test prediction and thus present as disparate impact.

[117] FRED C. PAMPEL, LOGISTIC REGRESSION 1 (2000).

[118] *See* Christopher M. Berry, *Differential Validity and Differential Prediction in Cognitive Ability Tests*, 2 ANN. REV. ORG. PSYCHOL. & ORG. BEHAV. 435, 443 (2015).

Figure 4. Logistic Regression Predicting the Odds of Violent Recidivism

Predictor	Model 1	Model 2	Model 3	Model 4
Hispanic	.540***	---	.678*	1.033
Decile	---	1.379***	1.375***	1.384***
Hispanic*Decile Interaction	---	---	---	.891
Constant	.202	.056	.058	.057
-2LL	3550.99	3207.91	3203.36	3201.18
χ^2	13.00	356.08	360.63	362.80
n=4,020				

Notes: * $p<.05$, ** $p<.01$, *** $p<.001$. Coefficients represent odds ratios.

With respect to the COMPAS violent risk scale, there are some consistencies with the general risk scale just discussed. In Figure 4, Model 1 shows that Hispanics are significantly less likely to commit a violent offense. The utility of the COMPAS violent recidivism scale generally is supported in Model 2. For every one unit increase in the violent decile score, the odds of violent recidivism increases by 38%.

The results here diverge from the regression models for general recidivism in two ways. The Hispanic variable in Model 3 remains statistically significant (p=.04), meaning that there are discrepancies in the intercepts of the regression lines for violent recidivism. The lower intercept means overprediction of Hispanics on the violent recidivism scale.[119] But as the Model 4 interaction is not significant (p=.143), there is no detectable variation in the slopes.

In sum, the nested regression models reveal differential predictive validity in COMPAS based on Hispanic ethnicity, though the form varies. For general recidivism, there is asymmetry in the slopes, while for the violent recidivism scale the problem is unequal intercepts. In both cases, though, differential predictive validity is shown and COMPAS can be challenged as presenting unfair and biased algorithmic results based on Hispanic ethnicity. Test bias also symbolizes that while COMPAS scores may have some meaning within groups, comparisons across groups are problematic.[120]

One may wonder if these regression models are too simplistic. For example, other researchers have found that gender, age, and criminal history are statistically significant factors in recidivism risk.[121] Because of this,

[119] *See* Christopher M. Berry, *Differential Validity and Differential Prediction in Cognitive Ability Tests*, 2 ANN. REV. ORG. PSYCHOL. & ORG. BEHAV. 435, 443 (2015).

[120] *See* Adam W. Meade & Michael Fetzer, *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*, 12 ORG. RES. METHODS 738, 738 (2009).

[121] John Monahan et al., *Age, Risk Assessment, and Sanctioning*, 41 LAW & HUM. BEHAV. 191 (2017); Jennifer Skeem et al., *Gender, Risk, Assessment, and Sanctioning: The Cost of Treating Women Like Men*, 40 LAW & HUM. BEHAV. 580 (2016); Nicholas Scurich

separate regression models (not offered herein) were run to include controls for gender, age, and number of prior counts. The results were substantially the same: unequal slopes for the general recidivism scale and unequal intercepts for the violent recidivism scale.

D. Differential Prediction via E/O Measures

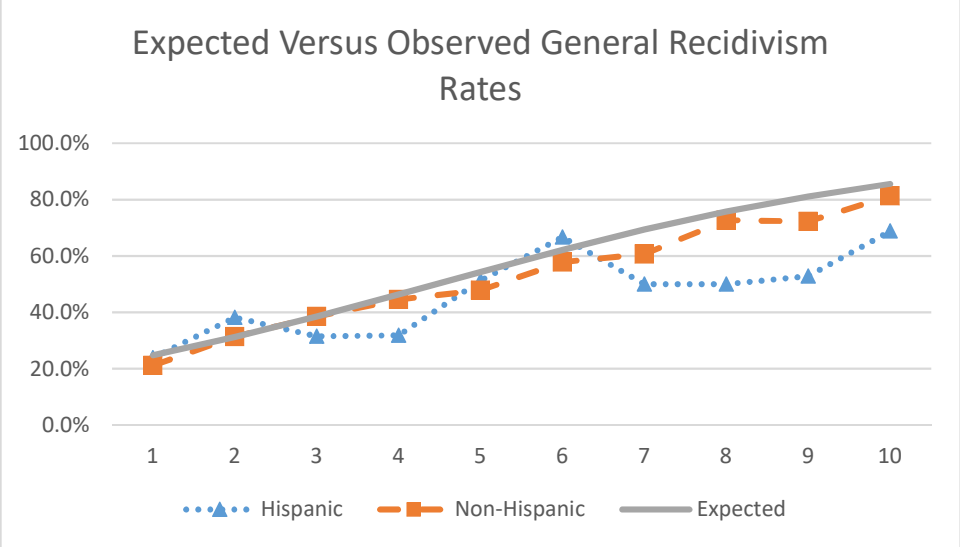
An alternative perspective on differential prediction looks to whether expected rates of recidivism are the same as observed recidivism rates in action, and that this is the case irrespective of group membership.[122] Expected rates are generated from logistic regression models whereas observed rates are the actual recidivism statistics. We can show the differences by plotting expected rates of recidivism against the observed rates at every COMPAS decile score (with expected versus observed referred to herein as "E/O").

For each COMPAS scale, a single expected rate line was computed using Model 2 statistics for general and violent recidivism from Figures 3 and 4, respectively. The single expected rate line represents the expected recidivism rate at each decile score, regardless of group. Two lines of observed rates of recidivism are presented (one for Hispanics and the other for non-Hispanics). Figures 5 and 6 provide visual bar graphs of these E/O plots for general recidivism and violent recidivism, respectively.

& John Monahan, *Evidence-Based Sentencing: Public Openness and Opposition to Using Gender, Age, and Race as Risk Factors for Recidivism*, 40 LAW & HUM. BEHAV. 36, 37 (2016).

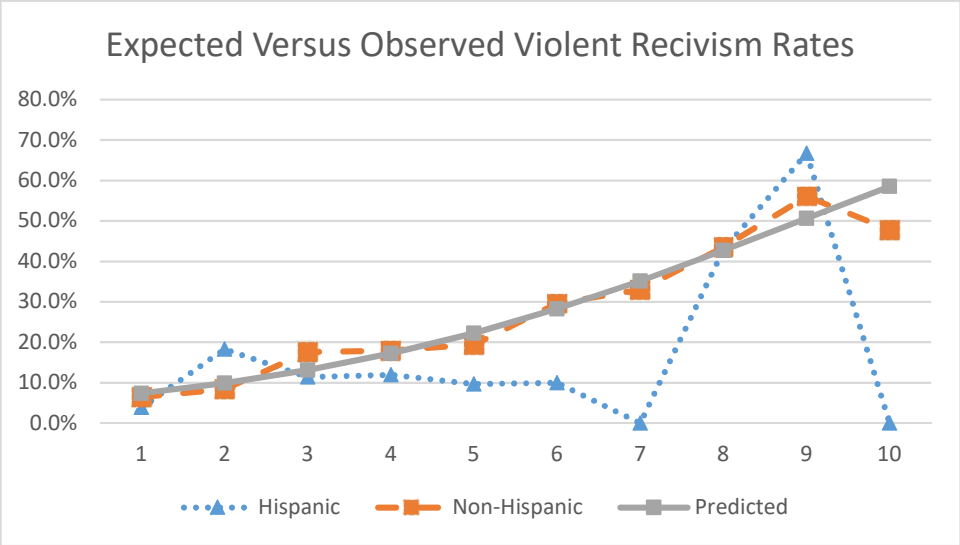
[122] Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments* 18-19 (2017), https://crim.sas.upenn.edu/sites/crim.sas.upenn.edu/files/Berk_FairJuvy_1.2.2018.pdf.

Figure 5. E/O General Recidivism Plots by Ethnicity



The solid line represents expected recidivism rates by decile regardless of grouping. The observed plotting for general recidivism outcomes for non-Hispanics moderately tracks the single, expected rates slope. The observed plotting for Hispanics, though, does not present a similar shape and, indeed, fails to reflect a consistent linear increase in actual recidivism rates as decile scores rise. This plotting visually represents differential prediction for COMPAS in the general recidivism scale. Figure 5 also graphically depicts accuracy errors for Hispanics.

Figure 6. E/O Violent Recidivism Plots by Ethnicity



The plotting in Figure 6 follows the same patterns just discussed for general recidivism rates. Though, the discrepancies are more dramatic. Actual recidivism rates for Hispanics oddly decline from decile 2 to a 0% recidivism rate at decile 7, then increase substantially in deciles 8 and 9, before plummeting back to 0% in decile 10. Hence, COMPAS violent decile score does not have a linear relationship to violent recidivism for Hispanics, thus undermining its performance for this group. This visual plotting also confirms differential prediction for COMPAS in its violent recidivism scale regarding Hispanic ethnicity.

So far, comparisons of risk bin outcomes, AUCs, and correlation coefficients reflect differential validity, while logistic regressions and E/O plots present differential prediction for COMPAS. Some argue that these are not the only appropriate measures of a test's ability to classify individuals, to predict risk, or to reflect unfairness.[123]

E. Mean Score Differences

It was shown earlier that average COMPAS decile scores on the general and violent recidivism scales significantly varied by Hispanic ethnicity. Yet it is too simplistic to assert that mean score differences signify test bias (even though it violates the algorithmic fairness notion of statistical parity[124]) as

[123] Richard W. Elwood, *Calculating Probability in Sex Offender Risk Assessment*, 62 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 1262, 1264 (2018).

[124] Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, Conf. on Innovations in Theoretical Computer Science 11 (2017), https://arxiv.org/abs/1609.05807.

there may be some risk-relevant differences between groups.[125] Still, closely-related fairness indices are indicative of *differential calibration*—group variances in calibration—and potentially disparate impact. The algorithmic fairness concept of "balance for the positive class" refers to requiring that the mean test score for those in the positive class—here, meaning recidivists—be the same across groups.[126] Correspondingly, "balance for the negative class" requires equal mean test score for those in the negative class—i.e., non-recidivists. A test is well-calibrated with respect to these definitions of algorithmic fairness then if a particular score is associated with the same likelihood of the outcome occurring, regardless of group membership.[127] Figure 7 presents mean COMPAS scores comparing recidivists and non-recidivists on both the general and violent recidivism scales.

Figure 7. Mean Decile Scores

	General Recidivism Scale		Violence Recidivism Scale	
	Recidivists	Non-Recidivists	Recidivists	Non-Recidivists
Hispanic	4.17	2.92***	3.54	2.54**
Non-Hispanic	5.65	3.53***	5.03	2.98***

Notes: * $p<.05$, ** $p<.01$, *** $p<.001$.

All four comparisons of mean decile score differences between groups are statistically significant at the level of $p=.005$ or below. Mean scores for recidivist and non-recidivist Hispanics are substantially lower than non-Hispanics. Practical differences exist as well. Mean scores for Hispanics exhibit much less variance between the recidivists and non-recidivists in both scales, being approximately one decile apart. The means for the non-Hispanics are at least two deciles apart, thus signifying greater mean score differences that distinguish non-Hispanic recidivists from non-recidivists in scoring to a greater degree. Overall, the metrics in Figure 7 signify a failure in balances for the positive and negative classes and thus again indicate algorithmic unfairness and disparate impact of COMPAS scoring based on Hispanic ethnicity.

F. Classification Errors

Additional computations measuring algorithmic accuracy and fairness are popular in the behavioral sciences and machine learning literatures. A few of

[125] Russell T. Warne et al., *Exploring the Various Interpretations of "Test Bias,"* 20 CULTURAL DIVERSITY & ETHNIC MINORITY PSYCHOL. 570, 572 (2014).

[126] Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, Conf. on Innovations in Theoretical Computer Science 2 (2017), https://arxiv.org/abs/1609.05807 (emphasis in original).

[127] Alexandria Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 154 (2017).

these can provide further insight into possible disparate impact which can occur when accuracy and error rates between groups are unequal.[128]

The first two are the true positive rate (TPR) and the true negative rate (TNR), representing a high risk and a low risk discrimination metric, respectively.[129] The TPR is alternatively titled sensitivity in the field of statistics and represents the nonerror rate for the recidivists.[130] The TNR is alternatively titled specificity and represents the nonerror rate for the non-recidivists.[131]

The TPR and TNR are *retrospective* in nature. Two metrics which more appropriately measure *prospective* predictive accuracy, and thereby are more important to practitioners who are interested in the predictive validity of risk tools, are the positive predictive value (PPV) and negative predictive value (NPV).[132] The PPV represents the probability that a high risk score will be correct, i.e., the proportion of high risk predictions who were recidivists.[133] The NPV then is the proportion of those classified as low risk who did not recidivate. The PPV is a high risk calibration measure while the NPV is a low risk calibration measure.[134]

These calculations require that the sample be divided into two groupings: one representing individuals predicted to be recidivists and the other non-recividivists. Regarding COMPAS, researchers typically opt to lump together COMPAS' low and medium risk bins into one group, merge the medium and high risk bins together, or for better measure offer both.[135] Figures 8 and 9 include the two different cutpoints for dichotomizing the COMPAS general and violent recidivism scales, respectively.

---

[128] Geoff Pleiss et al., *On Fairness and Calibration* 2, arxiv.org/pdf/1709.02012.pdf.

[129] Jay P. Singh, *Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer*, 31 BEHAV. SCI. & L. 8, 9 (2013).

[130] Kristian Linnet et al., *Quantifying the Accuracy of a Diagnostic Test or Marker*, 58 CLINICAL CHEMISTRY 1292, 1296 (2012).

[131] *Id*.

[132] Jay P. Singh, *Predictive Validity Performance Indicators for Violence Risk Assessment: A Methodological Primer*, 31 BEHAV. SCI. & L. 8, 12 (2013).

[133] ROBERT H. RIFFENBURGH, STATISTICS IN MEDICINE 254 tbl. 15.13 (2013).

[134] Jay P. Singh, *Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer*, 31 BEHAV. SCI. & L. 8, 11 fig. 1 (2013).

[135] Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses*, 80 FED. PROBATION 38, 42 (2016).

Figure 8. Measures for General Recidivism

	Low v. Medium/High		Low/Medium v. High	
Measure	Hispanic	Non-Hispanic	Hispanic	Non-Hispanic
Discrimination				
TPR	.42	.63***	.14	.31***
TNR	.81	.69***	.94	.91
Calibration				
PPV	.56	.63**	.57	.75***
NPV	.70	.68	.65	.61

Notes: * $p<.05$, ** $p<.01$, *** $p<.001$.

In Figure 8, five of the eight comparisons are statistically significant, again confirming differential validity and differential prediction. With respect to the True Positive Rate, the tool is significantly better at discriminating the recidivists from the non-recidivists for non-Hispanics at both cutpoints. Indeed, the general recidivism TPRs for non-Hispanics are approximately 50% and 120% better than for Hispanics. The low TPRs for Hispanics are actually quite poor from quantitative and qualitative perspectives. For example, at the higher cutpoint (low/medium versus high) the TPR is only 14%, meaning the test is wrong at least eight out of ten times for Hispanics.

Likewise, the PPV differences show that COMPAS is a better predictor of recidivism for non-Hispanics at both cutpoints. For example, at the higher cutpoint, of those that were predicted at high risk of general recidivism, 57% of Hispanics were reoffenders while 75% of non-Hispanics were correctly classified. The PPVs for Hispanics at both cutpoints are also considered poor, and indicate overprediction.[136]

Interestingly one need not here choose between the preference of ProPublica for the TPR or Northpointe (the owner of COMPAS) regarding the PPV, or which cutpoint to use. COMPAS on any of those measures performs far worse at predicting recidivists for Hispanics. The results thus suggest algorithmic unfairness.

In terms of identifying non-recidivists, one of the four measures is statistically significant. The tool performs modestly better at identifying non-recidivists among Hispanics.

[136] *See* James C. DiPerna et al., *Broadband Screening of Academic and Social Behavior, in* UNIVERSAL SCREENING IN EDUCATIONAL SETTINGS 223, 235, 239 (R.J. Kettler et al. eds., 2014).

Figure 9. Measures for Violent Recidivism

	Low v. Medium/High		Low/Medium v. High	
Measure	Hispanic	Non-Hispanic	Hispanic	Non-Hispanic
Discrimination				
TPR	.29	.54***	.14	.21**
TNR	.81	.77*	.98	.96
Calibration				
PPV	.14	.32***	.45	.49
NPV	.91	.89	.91	.86**

Notes: * $p<.05$, ** $p<.01$, *** $p<.001$.

Similar results are shown for violent recidivism in Figure 9. COMPAS generally performs much better from discrimination and calibration perspectives on classifying violent recidivists for non-Hispanics, with one caveat of the PPV at the higher cutoff where the PPV was still higher for non-Hispanics but not to a statistically significant degree. The violent recidivism TPRs for non-Hispanics are 86% and 50% better at the lower and higher cutpoints, respectively. The low PPVs for Hispanics again support the tendency of the tool to overclassify them. At the lower cutpoint, the predictions of violent recidivism for Hispanics were wrong at least 8 out of 10 times. Then, like the general recidivism scale, the TPRs and PPVs at both cutpoints indicate COMPAS achieves weaker accuracy at selecting recidivists for Hispanics. One need not choose between the TPR or PPV here either: either definition of algorithmic fairness indicates unfairness to Hispanics.

Like the general recidivism scale, the tool performs modestly better at selecting violent non-recidivists among Hispanics.

G. Limitations

Several limitations of this study should be mentioned. The single site limits generalization of results. This study relied upon archival data, and it is thereby possible for there to have been systematic errors in data collection that are not observable on secondary data analysis. Recidivism outcomes were from official records and thus will not include undetected crimes. The dataset did not include interrater reliability scores that would confirm the dependability of COMPAS scoring across evaluators and over time. Then, it would have been preferable to control for aspects of supervision as pretrial services/conditions may moderate reoffending rates, but such an option is also not available in this secondary data analysis.

V. CONCLUSIONS

Algorithmic risk assessment holds promise in informing decisions that can reduce mass incarceration by releasing more prisoners through risk-based selections that consider public safety. Yet caution is in order whereby presumptions of transparency, objectivity, and fairness of the algorithmic process may be unwarranted. Calls from those who heed such caution for third party audits of risk tools led to the study presented herein. This study rather uniquely focused on the potential of unfairness for Hispanics.

Using multiple definitions of algorithmic unfairness, results consistently showed that COMPAS, a popular risk tool, is not well calibrated for Hispanics. The statistics presented evidence of differential validity and differential predictive ability based on Hispanic ethnicity. The tool fails to accurately predict actual outcomes in a linear manner and overpredicts risk for Hispanics. Overall, there is cumulative evidence of disparate impact. It appears quite likely that factors extraneous to those scored by the COMPAS risk scales related to cultural differences account for these results.[137] This information should inform officials that greater care should be taken to ensure that proper validation studies are undertaken to confirm that any algorithmic risk tool used is fair for its intended population and subpopulations.

[137] *See* Adam W. Meade & Michael Fetzer, *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*, 12 ORG. RES. METHODS 738, 741 (2009).