DAMNED LIES & CRIMINAL SENTENCING USING EVIDENCE-BASED TOOLS

JOHN LIGHTBOURNE[†]

ABSTRACT

The boom of big data and predictive analytics has revolutionized business. eHarmony matches customers based on shared likes and expectations for romance, and Target uses similar methods to strategically push its products on shoppers. Courts and Departments of Corrections have also sought to employ similar tools. However, the use of data analytics in sentencing raises a host of constitutional concerns. In State v. Loomis*, the Wisconsin Supreme Court was faced with whether the use of an actuarial risk assessment tool based on a proprietary formula violates a defendant's right to due process where the defendant could not review how the various inputs were weighed. The opinion attempts to save a constitutionally dubious technique and reads as a warning to lower courts in the proper use of predictive analytics. This article explores certain equal protection and due process arguments implicated by* Loomis.

INTRODUCTION

On October 5, 2016, Eric Loomis' council petitioned the United States Supreme Court for certiorari hoping to get a definitive answer to a constitutionally, and morally, troubling question: can a sentencing judge use actuarial risk assessments to help decide an offender's sentence when we do not know what variables the formula uses?[1] These assessments are commonly referred to as evidence-based sentencing tools and sold as software suites used by departments of correction across the United States. These assessments output an estimate of an individual's future risk of recidivism based on statistical models built from data of convicted individuals gathered from correctional institutions. Different assessments use different variables supposedly correlated with a risk of future criminality. Some examples include age at first arrest and whether the

[†] Duke University School of Law, J.D. expected May 2018; M.A. in Applied Economics, The University of Alabama, May 2015; B.S. in Economics and Finance, The University of Alabama, May 2015.

[1] Loomis v. Wisconsin, No. 16-6387 (U.S. Oct. 5, 2016), *available at* http://www.scotusblog.com/wp-content/uploads/2017/02/16-6387-cert-petition.pdf.

individual has ties to other criminals.[2] The defendant's specific information, gathered through face-to-face interviews and public information gathered by the state's Department of Corrections, is entered into the software which then outputs an estimate of an individual's risk of recidivating based on regression modeling.[3] The tools have recently come under scrutiny, partially due to the former Attorney General Eric Holder's public warning about their use in 2014.[4] Proponents argue that they are scientific[5] and statistically unbiased,[6] and that judges should use all tools at their disposal when making sentencing decisions.[7] Critics dispute these claims and question their constitutionality.[8]

This article does not evaluate all of the constitutional arguments that surround the use of evidence-based sentencing or their possible reinforcement of bias in the criminal justice system. Instead, this article considers whether the specific risk tool used in *Loomis* violates the Equal Protection Clause of the Constitution for its use of gender specific norming groups and the Due Process Clause because of its use of gender norming

---

[2] NORTHPOINTE, PRACTITIONER'S GUIDE TO COMPAS CORE 27 (2015), http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf.

[3] Part I *infra* explains in more detail how the process and models work.

[4] *See* Devlin Barrett, *Holder Cautions on Risk of Bias in Big Data Use in Criminal Justice*, WALL ST. J. (Aug. 1, 2014, 2:10 PM), http://www.wsj.com/articles/u-s-attorney-general-cautions-on-risk-of-bias-in-big-data-use-in-criminal-justice-1406916606 ("Criminal sentences, he said, 'should not be based on unchangeable factors that a person cannot control, or on the possibility of a future crime that has not taken place.'").

[5] Richard G. Kopf, *Federal Supervised Release and Actuarial Data (Including Age, Race, and Gender): The Camel's Nose and the Use of Actuarial Data at Sentencing*, 27 FED. SENT. R. 207, 207 (2015) ("We must ask ourselves whether we wish to follow science, understanding that it may open up 'terrifying vistas of reality,' or whether we will 'flee from this deadly light into the peace and safety of a new dark age.'").

[6] NORTHPOINTE, *supra* note 2, at 15; *see also* Sam Corbett-Davies, Emma Pierson, Avi Feller & Sharad Goel, *A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually Not That Clear.* WASH. POST: MONKEY CAGE (October 17, 2016, https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.6f0b4aa73a1b (explaining in layman's terms the disputed "bias" in COMPAS risk assessments).

[7] *See, e.g.*, State v. Loomis, 881 N.W.2d 749, 765 (Wis. 2016), *petition for cert. filed*, No. 16-6387 (U.S. Oct. 5, 2016); Malenchik v. State, 928 N.E.2d 564, 572 (Ind. 2010).

[8] *See, e.g.*, Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 821 (2014).

groups and proprietary nature. After explaining how tools like the Correctional Offender Management and Profiling Alternative Sanctions assessment (COMPAS) used in *State v. Loomis*[9] are developed, this article concludes that it violates both equal protection and due process to use these tools during sentencing due to their reliance on group based averages and denial of any proper opportunity to contest the tool's accuracy due to the algorithm's proprietary nature.

I. HOW COMPAS SCORES ARE CALCULATED

A. Brief Overview of Actuarial Risk Assessment

Actuarial risk assessments can be calculated in many different ways, including using models relying on multiple regression, decision tree modeling, and simple summations inspired by the work of Lloyd B. Ohlin and Ernest Burgess.[10] The tool used in COMPAS is likely based on a regression formula, meaning that the weights that COMPAS attributes to a defendant's various characteristics is based on group averages. This conflicts with the jurisprudence that group based stereotypes, even if some statistical evidence can be posited in support of them, violate equal protection doctrine.[11]

In a standard linear regression model, some target variable is assumed to have a linear relation to one or more explanatory variables multiplied by some coefficient and an error term. The model finds the combination of weights given the inputted variables to minimize the error term. Solving for the coefficients shows that each weight is a function of the values of both the target variable as well as the different explanatory variables across all sampled observations. This shows mathematically that the coefficient is based on group data and averages.

---

[9] State v. Loomis, 881 N.W.2d 749, 753 n.10 (Wis. 2016), *petition for cert. filed*, No. 16-6387 (U.S. Oct. 5, 2016).

[10] Eric Silver & Lisa L. Miller, *A Cautionary Note on the Use of Actuarial Risk Assessment Tools for Social Control*, 48 CRIME & DELINQUENCY 138, 139 (2002). Burgess suggested a simple linear summation where each variable's weight was equal to one. *See* DON M. GOTTFREDSON & HOWARD N. SNYDER, NAT'L CTR. FOR JUVENILE JUSTICE, THE MATHEMATICS OF RISK CLASSIFICATION: CHANGING DATA INTO VALID INSTRUMENTS FOR JUVENILE COURTS 18 (2005), https://www.ncjrs.gov/pdffiles1/ojjdp/209158.pdf (providing examples of "Burgess-type models").

[11] J.E.B. v. Alabama *ex rel*. T.B., 511 U.S. 127, 141 n.11 (1994) ("Even if a measure of truth can be found in some of the gender stereotypes . . . that fact alone cannot support discrimination on the basis of gender . . . ."); *see also* Craig v. Boren, 429 U.S. 190, 204, 209 (1976) (noting the "normative philosophy" against "loose-fitting generalities" behind the equal protection clause).

Logistic regression is similar to linear regression and is meant for categorical targets–meaning a variable that either occurs or does not, like whether someone recidivates.[12] Risk assessments can then use the predicted coefficients to provide the weights for the risk score.[13]

B. The Use of Norming Groups

The use of "norming groups" is common in certain areas like standardized testing in schools.[14] Put simply, a norming group is a random sample collected from the larger sample of observations used to form the model.[15] Essentially, the raw score provided by the model for any one individual is compared to the raw scores of a selected sample, or norming group, in order to gain insight into the individual observation in relation to a representative group. Without going into detail about the various methods of probability sampling,[16] it is important to note that "there are [multiple] possible norm groups for any test" depending on what variables are used to select a norming group.[17] Further, the rankings will differ depending on what norm group is chosen. Thus, "the composition of the norm group is [crucial when] interpret[ing]" the ranking.[18]

C. How COMPAS Calculates Risk Scores

When using COMPAS, the defendant's information is collected through a combination of face-to-face interviews with a member from the state's Department of Corrections and information from the defendant's criminal file.[19] Since COMPAS is proprietary, the weights assigned to

---

[12] DANIEL T. LAROSE, DATA MINING METHODS AND MODELS 155 (2006).

[13] *See* GOTTFREDSON & SNYDER, *supra* note 10, at 20, 21 (explaining how the predicted coefficients might be transformed and used in risk assessments).

[14] For a general overview of norming with test scores, see generally Maximo Rodriguez, Norming and Norm-Referenced Test Scores (Jan. 29, 1997) (unpublished manuscript), http://files.eric.ed.gov/fulltext/ED406445.pdf. *See also* W. L. Bashaw, *Assessing Learner Performance*, *in* INSTRUCTIONAL DESIGN: PRINCIPLES AND APPLICATIONS, 151, 151–72 (Leslie J. Briggs, Kent L. Gustafson & Murray H. Tillman eds., 2nd ed. 1991).

[15] For example, COMPAS' norming groups comprise observations from a sample of over 30,000 assessments made between 2004–2005 at various "prison, parole, jail and probation sites." NORTHPOINTE, *supra* note 2, at 11.

[16] For explanations of sampling methods, see ELAZAR J. PEDHAZUR & LIORA PEDHAZUR SCHMELKIN, MEASUREMENT, DESIGN, AND ANALYSIS: AN INTEGRATED APPROACH 320–26 (1991).

[17] Rodriguez, *supra* note 14, at 4 (citing F. BROWN, PRINCIPLES OF EDUCATIONAL AND PSYCHOLOGICAL TESTING (2d ed. 1976)).

[18] *Id.*

[19] *Id.* at 754.

certain variables and the type of model are not publically available.[20] The practitioner's guide gives some insight into what variables the different "scales" include, even if it does not provide the coefficients for each variable.[21] The three "risk scales" are summations of various "criminogenic need scales" multiplied by weights corresponding to "the strength of the item's relationship to" recidivism.[22] For example, the practitioner's guide provides that the Violent Recidivism Risk Scale is calculated as: Risk Score=β1*age + β2*(age at first arrest) + β3*(history of violence) + β4*(vocation/education scale) + β5*(history of noncompliance).[23]

The items labeled "history of noncompliance," "history of violence," and "vocation/education" are different criminogenic need scales, meaning a separate equation can denote each one.[24] Again, those equations and weights are proprietary. The General Recidivism risk scale includes criminogenic needs scales for "prior criminal history, [affiliation with] criminal associates, drug involvement, and early indicators of juvenile delinquency problems."[25]

A calculated risk score can be compared to the scores of a normative group that has been grouped in "ascending order" and then "dividing these scores into ten equal sized groups."[26] This provides the practitioner with a decile rank, where a lower rank indicates the offender has similar attributes as individuals who have a lower risk of recidivism

---

[20] *Id.* at 761 (observing that Northpointe "does not disclose how the risk scores are determined or how the factors are weighed."). Surprisingly, the Wisconsin Supreme Court denied an amici brief from Northpointe to help explain COMPAS' accuracy and efficiency, despite neither the state nor defendant's counsel being able to answer questions regarding the tool. *Id*. at 774 (Abrahamson, J., concurring).

[21] NORTHPOINTE, *supra* note 2, at 27–29, 32–46.

[22] *See id.* at 29 (explaining how risk scores are calculated using the Violent Recidivism Risk Score as an example).

[23] *Id*. This provides a basic idea of what variables are considered in the summation and resembles the kinds of summations detailed in GOTTFREDSON & SNYDER, *supra* note 10.

[24] NORTHPOINTE, *supra* note 2, at 38, 39, 44 (providing a brief description of the different scales). For example, the vocation/education need scale is a "amalgam of education attainment, vocational skills, job opportunities . . . [and] good income" among other variables. *Id.* at 44.

[25] *Id*. at 27.

[26] *Id.* at 11. Currently, COMPAS provides eight different normative subgroups based on gender and prison or parole populations. *Id.*

and a higher rank indicates the offender shares similar attributes as individuals who have a higher risk of recidivism.[27]

However, the deciles are only informative in that they compare the offender to the chosen normative group.[28] If the norming group is not representative, the ranking will be inaccurate. This is where COMPAS incorporates gender–male offenders are compared to a male normative group and females to a female normative group.[29] Instead of automatically lowering an individual's risk of recidivism if female–or increasing it if male[30]–COMPAS removes gender as an independent variable. By comparing recidivism scores in gender normed groups, a man and woman with identical variables will receive identical risk scores but may still rank differently as to risk when compared to the norming group.

II. STATE ENDORSEMENTS OF EVIDENCE-BASED RISK ASSESSMENTS

Numerous states use evidence-based tools at sentencing, and a few even require judges consider them when making decisions.[31] Wisconsin severely limited the use of COMPAS assessments in Loomis.[32] The court held that sentencing judges cannot use the risk scores to determine whether someone is incarcerated, as opposed to parole, or for how long they are incarcerated.[33] Further, the judge must "explain the factors in addition to" the risk scores that "independently support the sentence imposed."[34] Finally, the judge must be provided with a list of warnings explaining the limitations of the risk score, including that we do not know how the model determines scores, that it may "disproportionately classify minorit[ies]"

---

[27] *See id.* ("It is important to note that decile scores can only be interpreted in a relative sense, and are always linked to the norm group.").

[28] For instance, if the normative group is male murderers with a history of violence, a lower decile may not *actually* indicate a lower risk of violence. *Id.* at 11.

[29] *Id.*

[30] There are numerous articles devoted to the fact that men recidivate at a higher rate than women. *See, e.g.*, Michael M. Wehrman, *Examining Race and Sex Inequality in Recidivism*, 5 SOCIOLOGY COMPASS 179, 179 (2011) ("The probability of recidivating is not a randomly distributed event; men are more likely than women to recidivate . . . .").

[31] *See, e.g.*, KY. REV. STAT. ANN. § 532.007(3)(a) (2016) (requiring judges to consider risk and needs assessments); OKLA. STAT. tit. 22, § 988.18(B) (2016) (requiring judges to refer to risk assessment if considering any kind of community punishment for felony offenders).

[32] State v. Loomis, 881 N.W.2d 749, 769–70 (Wis. 2016), *petition for cert. filed*, No. 16-6387 (U.S. Oct. 5, 2016).

[33] *Id.* at 769.

[34] *Id.* (emphasis added).

higher risk[35] and that COMPAS "was not developed for use at sentencing" since it is not meant to identify a "particular high-risk individual."[36] Despite these warnings, judges can consider the risk scores as "one of many factors" during their sentencing deliberations.[37] These limitations are important since defendants do not have the ability to, in this author's opinion, properly contest these risk assessments. The reports are given directly to the judge during sentencing, and although defendants have the right to contest any information that went into the model–unlikely since the defendant provided much of the information during the face-to-face interview–proving the assessment is imprecise without knowing how the models are formed is difficult.[38]

Indiana took a slightly different approach when it reviewed the risk assessment tool LSI-R–a close relative to COMPAS and what COMPAS likes to compare itself to.[39] The court gave great weight to social science research supporting the tool.[40] However, like in Wisconsin, the court continued to reiterate that the assessments were a "supplemental source of information" and not meant to decide sentence length.[41] The

---

[35] *Id.* An easy example of how COMPAS may classify minorities as higher risk can be seen with the variable regarding juvenile delinquency: certain minority groups have much higher juvenile incarceration rates than Caucasians. Joshua Rovner, *Racial Disparities in Youth Commitments and Arrests*, THE SENTENCING PROJECT (Apr. 1, 2016), http://www.sentencingproject.org/publications/racial-disparities-in-youth-commitments-and-arrests/ ("As of 2013, black juveniles were more than four times as likely to be committed as white juveniles . . . and Hispanic juveniles were 61 percent more likely."). Jeff Larson provides a more in depth analysis regarding COMPAS and the potential for racially biased misclassification. Jeff Larson, Julia Angwin, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Jeff Larson's follow up articles are available at https://www.propublica.org/site/author/jeff_larson.

[36] *Loomis*, 881 N.W.2d at 769; *see also* NORTHPOINTE, *supra* note 2, at 29 ("An individual's level of risk is estimated based on . . . offenders with similar characteristics."). The court also included a warning that COMPAS had not been validated for the state of Wisconsin. *Loomis*, 881 N.W.2d at 764. However, a validation on a subgroup of Wisconsin inmates or parolees would do nothing to qualm the constitutional concerns of considering the offender's gender.

[37] *Loomis*, 881 N.W.2d at 769.

[38] This is discussed further at *infra* notes 57–58 and accompanying text.

[39] Malenchik v. State, 928 N.E.2d 564, 574–75 (Ind. 2010) (holding it was not discriminatory to consider a LSI-R report because the information used by the report was required to be presented to the judge by statute); NORTHPOINTE, *supra* note 2, at 20 (comparing COMPAS scales to the "gold standard," LSI-R).

[40] *Malenchik*, 928 N.E.2d at 574–75.

[41] *Id.* at 573, 575.

court endorsed the use of risk assessment tools as a way to "design a probation program for the offender" and suspend sentences for low risk individuals.[42] States' limitations on risk assessments seem like an implicit acknowledgment of their questionable constitutionality.[43]

III. DUE PROCESS CHALLENGES TO THE USE OF COMPAS' PROPRIETARY FORMULAE

The Supreme Court has long recognized that criminal justice theory has evolved beyond treating every identical offense with the same punishment.[44] In *Williams v. New York*, the Court emphasized the role of judges in crafting individualized sentences by "draw[ing] on information concerning every aspect of a defendant's life."[45] It cautioned that the Due Process Clause should not be viewed as barring a judge from considering "out-of-court" information.[46] Instead, considering outside sources of information–like a probation report–simply allows judges to make a "more enlightened and just sentence."[47]

The Court has recognized several limitations to a sentencing judge's discretion,[48] including "sentencing . . . must satisfy the requirements of the Due Process Clause."[49] This generally involves ensuring the defendant is given the opportunity to contest, or explain, the evidence used against him and that the judge is unbiased.[50] Lower courts

---

[42] *Id.* at 573.

[43] *See* State v. Loomis, 881 N.W.2d 749, 757 (Wis. 2016), *petition for cert. filed*, No. 16-6387 (U.S. Oct. 5, 2016) (noting that the court imposed limitations on the use of COMPAS "must [be] observe[d] in order to avoid potential due process violations").

[44] Williams v. New York, 337 U.S. 241, 247 (1949).

[45] *Id.* at 250.

[46] *Id.* at 251.

[47] *Id.* at 250, 251.

[48] *See* Apprendi v. New Jersey, 530 U.S. 466, 490 (2000) ("[A]ny fact that increases the penalty for a crime beyond the prescribed statutory maximum must be submitted to a jury, and proved beyond a reasonable doubt."); Gardner v. Florida, 430 U.S. 349, 362 (1977) (holding that the presentence investigation report must be disclosed if it is considered by a judge who imposes a death sentence despite a jury recommendation of life in prison).

[49] *Gardner*, 430 U.S. at 357 (plurality opinion). The specific limitations imposed by the due process clause was controversial for the judges. For instance, Justice White believed the Due Process Clause was merely the "vehicle by which the . . . Eighth Amendment" applies. *Id.* at 364 (White, J., concurring).

[50] *See, e.g.*, *Apprendi*, 530 U.S. at 490 (holding the defendant has a right to contest facts a judge relied on to increase the defendant's sentence above the statutory maximum); *In re* Murchison, 349 U.S. 133, 136 (1955) ("A fair trial in a fair tribunal is a basic requirement of due process."); United States v. Gambino-

have expanded on the Supreme Court's jurisprudence to prohibit the consideration of "factors that could lead to unwarranted discrimination."[51]

However, the Court has continuously restated that judges have wide discretion "taking into consideration various factors relating both to offense and offender" choosing a sentence "within the range prescribed by statute."[52] Further, although the plurality and concurrence in *Gardner* were conflicted in how the due process clause applies in sentencing, neither believed that all the procedural rights guaranteed at trial apply during sentencing.[53]

In general, due process protections during sentencing are more procedural than substantive. Unlike *Gardner*, Loomis provided the information used in the report and was able to see what variables went into the risk assessment. He could contest the truth of those variables during sentencing.[54] Of course, without seeing how the risk assessment weighs the different variables, the ability to contest the inputs is a small comfort. The Wisconsin Supreme Court interpreted Loomis' due process challenge regarding the tool's proprietary nature and conceded that Loomis had the right to be sentenced based on "accurate information."[55]

The Wisconsin Supreme Court interpreted "accurate information" to mean that the risk needs assessment must be statistically accurate.[56] To

---

Zavala, 539 F.3d 1221, 1228 (10th Cir. 2008) (stating that judicial bias violates due process).

[51] *See* Carissa Byrne Hessick & F. Andrew Hessick, *Recognizing Constitutional Rights at Sentencing*, 99 CA. L. REV. 47, 55 (2011).

[52] *Apprendi*, 530 U.S. at 481 (emphasis omitted). Despite acknowledging procedural and substantive protections during sentencing over time, *Williams v. New York* has never been formerly overturned. *See, e.g.*, *Apprendi*, 530 U.S. at 481; United States v. Mills, 446 F. Supp. 2d 1115, 1124 (C.D. Cal. 2006) (acknowledging that although "*Williams*'s holding may be rendered questionable" because of subsequent decisions, it was never "explicitly overruled").

[53] *Gardner*, 430 U.S. at 358 n.9 (plurality opinion) ("The fact that due process applies does not, of course, implicate the entire panoply of criminal trial procedural rights.").

[54] State v. Loomis, 881 N.W.2d 749, 761 (Wis. 2016), *petition for cert. filed*, No. 16-6387 (U.S. Oct. 5, 2016) (stating Loomis' assessment was based on "his answers to questions and publically available data" that he "had the opportunity to verify"). Loomis also had the opportunity to argue that "other factors or information demonstrate" the risk score's "inaccuracy." *Id.* at 761–62.

[55] *Id.* This is derivative of the right acknowledged in *Townsend v. Burke*, 334 U.S. 736, 741 (1948) (holding that sentencing based on "materially untrue" assumptions of criminal history violates due process).

[56] *Loomis*, 881 N.W.2d at 762–64 (explaining various validation studies of COMPAS Core).

the court's credit, it provided studies that were critical of COMPAS as well as state validation studies that approved of its accuracy.[57] The court recognized these scholarly disagreements in limiting the tool's use.[58]

However, statistical accuracy should not be the measure of accuracy courts focus on. The validity measurements that these tools rely on, called area under the curve, relies on the ratio of false positives to false negatives. Essentially, the area under the curve indicates the likelihood that a randomly chosen observation is correctly listed as either higher probability or lower probability than another observation. The industry accepted standard is *ROC = .70*, meaning a defendant is correctly classified only 70% of the time.[59] In other words, there is a 70% chance that any randomly selected higher-risk individual is classified as higher risk than a randomly selected low-risk individual. Inversely, there is a 30% chance that a lower risk individual will be ranked higher than our actual high risk individual. Because our criminal justice system is premised on the theory that "it is far worse to convict an innocent man than to let a guilty man go free,"[60] basing sentencing decisions off a tool that incorrectly labels individuals at these rates is unsettling at best.

Sentencing is not merely a part of the criminal justice system, but the precise point where one's liberty is infringed.61 Although one has been convicted at sentencing, our criminal justice system still provides protections for defendants through acknowledgment of due process protections. If the criminal justice system prefers type II errors over type I, then sentences should reflect the same sentiment that objectively less risky defendants should not be subject to overly severe punishment. Without knowing how the tool weighs variables, defendants cannot

---

[57] *Id.*

[58] Northpointe strongly disputes the critique of COMPAS put out by Jeff Larson and ProPublica. However, both Northpointe and Jeff Larson have valid points. Their back and forth perfectly sums up the consequences of judging a statistical tool's accuracy with statistics such as ROC ratios. *Compare* NORTHPOINTE, *supra* note 2, at 14–16 (arguing that similar AUC scores for different ethnic groups is evidence of validity), *with* Larson et al., *supra* note 35 (arguing that COMPAS resulted in significant racial disparities).

[59] NORTHPOINTE, *supra* note 2, at 17 ("By convention an AUC of 0.70 is regarded as good in criminal justice settings."); *see* J.A. Hanley, *Receiver Operating Characteristic (ROC) Curves*, WILEY STATSREF: STATISTICS REFERENCE ONLINE, Sept. 29, 2014, at 1, http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat05255/pdf (explaining how ROC curves are calculated).

[60] *In re* Winship, 397 U.S. 358, 372 (1970) (Harlan, J., concurring).

[61] *See* Hessick & Hessick, *supra* note 51, at 49 ("Sentencing is the process through which the state deprives those convicted of crimes of their liberty. Thus, the recognition of constitutional rights at sentencing is paramount.").

properly defend themselves against the tool's prediction. Discrediting the tool would require providing contrary statistical and sociological studies explaining why correlations acknowledged in the model are flawed or miscalculated in order to argue that they are falsely ranked high. Although the private company has an interest in its proprietary formula, that interest should not outweigh the public's interest in a fair and effective criminal justice system if judges choose to use these tools at sentencing.

IV. EQUAL PROTECTION CHALLENGES TO THE USE OF GENDER NORMING

The Supreme Court has never explicitly held that sentencing judges could not consider factors like race or gender.[62] However, lower courts have recognized that considering these factors would be unconstitutional[63] although these kinds of claims have historically been limited, or ignored, during sentencing.[64] Whether risk assessment tools that use these variables would be upheld is unclear.[65]

---

[62] *See, e.g.*, Dodakian v. United States, no. 14-cv-01188 (AJN)(SN), 2015 WL 11144511, at *12 (S.D.N.Y. Aug. 14, 2015) ("Although the Supreme Court . . . ha[s] never held that gender discrimination in imposing a criminal sentence violates the Equal Protection Clause, it follows from the progression of equal protection [case law] that it does."). Language in the Court's denial of certiorari regarding *Buck v. Thaler* suggests the Court believes providing to a jury during the penalty phase of a capital trial evidence that certain races "are statistically more likely" to offend would violate the defendant's constitutional rights. *See* Buck v. Thaler, 132 S. Ct. 32, 33 (2011) (Alito, J., concurring in denial of certiorari).

[63] *See, e.g.*, United States v. Taveras, 585 F. Supp. 2d 327, 336 (E.D.N.Y. 2008) (holding that the court could not rely on defendant's "race, ethnicity or national origin alone" as sufficient to implicate defendant in a gang, an "aggravating factor" for sentencing); Williams v. Currie, 103 F. Supp. 2d 858, 863 (M.D.N.C. 2000) ("[I]nvidious gender discrimination [during sentencing] violates the Equal Protection Clause of the Fourteenth Amendment."); Hessick & Hessick, *supra* note 51, at 55 ("[C]ourts have forbidden consideration of race, national origin or gender at sentencing.").

[64] Hessick & Hessick, *supra* note 51, at 84 ("[T]hese courts have nevertheless allowed consideration of those factors on the ground that any limitation on the information a judge could consider would impair the sentencing judge's ability to arrive at the 'correct' sentence.").

[65] Although Federal Sentencing Guidelines state that race and gender are "not relevant in [determining] a sentence," U.S. SENTENCING COMM'N, UNITED STATES SENTENCING COMMISSION GUIDELINES MANUAL 2016 § 5H1.10 (2016), comments accompanying a tentative second draft of the Model Penal Code's sentencing guidelines suggest that while including race in risk assessments may be unconstitutional, including gender is likely not. MODEL PENAL CODE:

The use of data norming with representative subgroups thus poses an even more complex question of constitutionality. Gender norming removes the bias that would otherwise be included had a model merely included gender as a variable. This author is not certain that the typical equal protection arguments derived from cases like Craig v. Boren66 or United States v. Virginia67 would succeed. Part A lays out these arguments in detail and Part B concludes that intermediate scrutiny may invalidate risk assessment tools that use gender norming precisely because the concept behind norming is that individuals will act like the gender based norming group.

A. Equal Protection Case Law and Gender Stereotypes

Classifications based on gender are subject to intermediate scrutiny, meaning they must "serve[] important governmental objectives" and the "means employed are substantially related to the achievement of those objectives."[68] Absent an "exceedingly persuasive" justification[69] by the state meeting these requirements, the overt classification violates the Equal Protection Clause.

Protecting the citizenry through appropriate criminal sentencing is likely sufficient to be an important governmental objective.[70] The question becomes whether stereotyping an individual based on group averages is substantially related to the achievement of those objectives.

The Supreme Court has historically decried attempts to justify gender-based classifications using statistical generalizations. For instance, in Craig v. Boren, Oklahoma argued that statistical inferences based on "random roadside survey[s]" and "analysis of arrest statistics" supported a ban on men buying a low alcohol beer instead of women because men were more likely to drink and drive.[71] The Court criticized the studies[72]

SENTENCING § 6B.09 (AM. LAW INST., Tentative Draft No. 2, 2011). The reasoning for this discrepancy is unclear. If including race triggers heightened scrutiny, so would gender.

[66] Craig v. Boren, 429 U.S. 190 (1976).

[67] United States v. Virginia, 518 U.S. 515 (1996).

[68] *Id.* at 523 (quoting Mississippi Univ. for Women, 458 U.S. 718, 724 (1982)).

[69] *Id.* at 533.

[70] *See* Schall v. Martin, 467 U.S. 253, 264 (1984) ("The legitimate and compelling state interest in protecting the community from crime cannot be doubted." (internal quotations omitted)). *But see* Williams v. Currie, 103 F. Supp. 2d 858, 863 (M.D.N.C. 2000) (noting North Carolina had not provided any "important governmental objective to support discriminating . . . based on . . . gender" during sentencing).

[71] *Craig*, 429 U.S. at 200–03.

[72] *Id.* at 202 n.14.

and the lack of connection to the age-sex interaction the State sought to end.[73] However, the flawed studies were likely irrelevant. In general, "proving broad sociological propositions by statistics . . . inevitably is in tension with the normative philosophy that underlies the Equal Protection Clause."[74] Thus, "loose-fitting generalities" based on statistics will not persuade the Court.[75] The Court reiterated the impermissibility of gender classifications, despite statistical evidence supporting the classification, in J.E.B. v. Alabama ex rel. T.B.[76]

Virginia expanded on Craig, stating that the "justification must be genuine" and "not rely on overbroad generalizations about the different talents, capacities, or preferences of males and females."[77] The State's expert witnesses claimed that women and men thrived in different kinds of school environments, reflecting opinions on "typically male or typically female tendencies," that would entail the end of Virginia Military Institute's adversarial system if women were allowed to attend.[78] The Court reiterated that courts should "take a hard look" at these sorts of generalizations[79] and that the conclusion that VMI would have to adopt another learning method was unjustified,[80] reflecting the kind of "self-fulfilling prophec[ies] . . . routinely used to deny rights or opportunities [to women]."[81] The Court has consistently rejected "group tendencies as a proxy for individual characteristics" in gender-based equal protection jurisprudence.[82]

B. Applying Intermediate Scrutiny to COMPAS' Use of Norming Subgroups

Tools that include gender as an independent variable lead to inequalities in sentencing that generally disfavor men.[83] However, using

---

[73] *Id.* at 203 n.16.
[74] *Id.* at 204.
[75] *Id.* at 209.
[76] J.E.B. v. Alabama *ex rel.* T.B., 511 U.S. 127, 161 n.11 (1994) ("We have made abundantly clear in past cases that gender classifications that rest on impermissible stereotypes violate the Equal Protection Clause, even when some statistical support can be conjured up for the generalization.").
[77] United States v. Virginia, 518 U.S. 515, 533 (1996).
[78] *Id.* at 541, 542 (internal quotations omitted).
[79] *Id.* at 541 (internal quotations omitted).
[80] *Id.* at 542 n.12.
[81] *Id.* at 543 (citation and internal quotations omitted).
[82] Starr, *supra* note 8, at 827.
[83] *See id.* at 837 (noting that risk assessments "produce higher risk estimates, other things equal, for subgroups whose members are already disproportionately incarcerated").

norming groups may tend to disfavor women. Since men do recidivate more often, we would expect the average man in a prison norming group to have more prior convictions than the average woman in a similarly situated female norming group, and the risk scores are compared to these averages.[84] Thus, a man with fewer prior convictions will likely be considered lower risk but a woman with identical prior convictions may be riskier compared to the norm group since we would expect the average woman in the subgroup to have less prior convictions.

Does this violate the Equal Protection Clause? There is a facial classification: men are only compared to men and women are only compared to women. The evidence-based sentencing tool generalizes that an offender's gender affects whether the individual recidivates in the future. Although COMPAS does not assign a number value to an individual's gender like other tools, the norming process does implicate stereotypes of gender-based behavior based on a defendant only being compared to a norming group of his or her gender.

Further, the norming process does not determine whether someone is objectively risky. Instead, it merely suggests that the person is more or less risky than another in the gender-based norming group. Is a woman that places in the upper decile of the female norming group riskier than a man who scored in the lower deciles? If the two had similar characteristics it would be nonsensical to say that a woman is riskier, and therefore requires more restrictive sentencing.

Under intermediate scrutiny, the State has the burden to show the classification is "substantially related to" achieving an "important governmental objective[.]"[85] While maintaining an effective criminal justice system and reducing recidivism is arguably an important objective, there are alternate ways that do not rely on a gender-based classification using gender-based group averages. A judge could be presented with an evidence-based tool that omits gender completely. These models would not be as accurate as those that included gender, but if the point is to just

---

[84] Northpointe acknowledges this kind of counterintuitive pattern is possible and should be reviewed carefully. NORTHPOINTE, *supra* note 2, at 30–31. Of course, the inverse may be true since women are generally sentenced less severely than their male counterparts. *See* Nancy Gertner, *Women Offenders and the Sentencing Guidelines*, 14 YALE J.L. & FEMINISM 291, 292 (2002) ("[B]oth before and after the enactment of the Guidelines, women offenders have been treated more leniently than male offenders."). In which case, the women represented in the norming group may have riskier characteristics, because otherwise they would not have been sentenced to prison. Again, the scores are only descriptive in the context of the norming group.

[85] United States v. Virginia, 518 U.S. 515, 533 (1996).

show a spectrum of behavior and factors that correlate to increased riskiness, a judge could still get a similar sense of a person's propensity for recidivism as compared to other individuals. After all, the judge knows the defendant's gender; why is it then necessary to either bump up or down an individual's risk score because of it? Additionally, past discrimination in the criminal justice system may be incorporated in statistical models.[86] Although statisticians seek to control for this, without seeing the model and its treatment of raw data, observers to the criminal process cannot be sure these discriminations are not merely self-perpetuating through the use of statistical modeling. In short, the proprietary nature of the model and seeming acceptance of false positives and negatives merely reinforces public distrust in the criminal process.

The State's argument would hinge on the necessity of including gender to make the model more effective.[87] But if the State was that concerned about statistical accuracy, race should be included as well in order to compensate for any bias in regard to race. However, most, including the ALI, believe this would be unconstitutional.[88] Why would gender be different? It is unlikely that enhancing statistical accuracy for these reports would pass as an "exceedingly persuasive justification" for having the gender-based classification. For these reasons, the use of gender norming is likely not substantially related to the goals of criminal sentencing, especially since general reports on an individual's riskiness can be produced without quantifying the effect of the person's gender.

It seems more likely that the wide acceptance by trial judges is due to efficiency considerations. The prospect that a judge can receive one report that quantifies all the information scrawled across multiple reports and criminal files is enticing. But without knowing how the model is built, the possibility for impermissible discrimination based on gender in providing estimates of riskiness is unconstitutional and reflects another kind of generalization based on statistical evidence outlawed by J.E.B. Although the technique is less offensive than including gender as an independent variable, it is based on the theory that whether a person

---

[86] Jeff Larson's critique of COMPAS illustrates this rather well. *See supra* note 58.

[87] Although this raises an interesting question: can the state proffer as evidence what is essentially on trial? The question is whether the use of gender-based statistics is appropriate. Can the state proffer alternate statistical analysis to support it? This kind of justification seems at odds with *Craig* and *J.E.B.* as discussed above at *supra* Part IV.A. For this reason, the State may wish to argue that gender is merely one part of the equation that is necessary for the model to effectively weigh other variables that are not protected characteristics.

[88] For further discussion, see *supra* note 65.

chooses to break the law again is partially based on one's gender. This is an "overbroad generalization" like the kind lambasted in Craig.[89]

However, the limits placed on the tools used by the Wisconsin Supreme Court may save it from constitutional invalidation.[90] The group of defendants most likely to be "harmed" by gender norming would be women who have similar criminogenic characteristics as male counterparts but due to the make-up of the norming group may be considered higher risk where men would be considered lower risk compared to the male norming group.[91] But these defendants still have to show discrimination–and not receiving a lax sentence is likely not an appropriate harm in a noncapital context since it is well established that defendants do not have a right to a reduced sentence so long as it is within the given statutory range.[92] It is hard to contemplate a situation where a defendant would have standing to claim discrimination when judges can defend their sentencing decisions by claiming their decision was based on other factors.

---

[89] Craig v. Boren, 429 U.S. 190, 204 (1976) ("[P]roving broad sociological propositions by statistics is a dubious business, and one that inevitably is in tension with the normative philosophy that underlies the Equal Protection Clause."). *But see* Schall v. Martin, 467 U.S. 253, 265 (1984) ("[T]he harm to society generally may even be greater . . . given the high rate of recidivism among juveniles."). The Court will not entertain basing generalizations and stereotypes based on statistics for protected classes like race or gender whereas it may for non-protected classes like age. *See* Buck v. Thaler, 132 S. Ct. 32, 33 (2011) (Alito, J., concurring in denial of certiorari) (noting that testimony stating a person's race makes one statistically more likely to commit future crime would be a basis for the sentence's reversal).

[90] *But see* Dawinder S. Sidhu, *Moneyball Sentencing*, 56 B.C. L. REV. 671, 726–27 (2015) ("[T]he understandable preference for the positive use of actuarial instruments does not eliminate the very real possibility that these instruments may be used in both directions."); Starr, *supra* note 8, at 840 ("There is no persuasive reason to believe access to risk predictions would only tend to reduce sentences rather than to also increase them in some cases.").

[91] Because men have higher rates of recidivism than women it is possible that female norming groups would be composed of individuals who have lower rates of recidivism than their male counterparts. For further discussion on gendered differences in recidivism, see *infra* note 93 and accompanying text. *See also* NORTHPOINTE, *supra* note 2, at 29–30 (describing another scenario where a defendant would receive a counterintuitive risk score).

[92] Gardner v. Florida, 430 U.S. 349, 358 (1977).

V. THE PARADOX IN BALANCING CONSTITUTIONAL PROTECTIONS AND STATISTICAL ACCURACY

To further complicate matters, some psychological studies suggest that women and men are driven to crime in different gendered paths.[93] Thus, an accurate model that predicts riskiness should likely include some interaction term(s) between an offender's gender and criminogenic needs assessments. The theory would be that gender affects future decision-making, and illegal behavior–an assumption at odds with Virginia, which rejected the premise that gender could dictate whether a woman would want to engage in and fulfill the requirements of the adversarial system at VMI.[94] If studies have found race-based and gendered explanations of criminality, then a model that excludes these interactions is, by definition, not as accurate as it could be. It is unjust to sentence individuals based on a tool that constitutional protections require be less accurate than possible.

CONCLUSION

In conclusion, despite state attempts at protecting actuarial risk assessments from constitutional scrutiny, heightened scrutiny should invalidate sentencing judges' use of these tools because they rely on impermissible generalizations of gender. Further, defendants have due process rights to be sentenced on accurate information. If that means that defendants only have the right to be sentenced by a statistically valid tool, then the industry standard should be reevaluated and a defendant should have the right to contest the structure of the model itself. In order to effectively do so, the models cannot stay proprietary. A defendant should have the ability to provide evidence suggesting that the tools used against him or her are flawed–just as he or she would with any other piece of evidence. When the courts choose to use proprietary tools in sentencing against their stated use, the courts wrongly maintain the formula's proprietary nature at the expense of a defendant's right to a fair and effective criminal justice system.

---

[93] *See, e.g.*, Sarah Bennett, David P. Farrington & L. Rowell Huesmann, *Explaining Gender Differences in Crime and Violence: The Importance of Social Cognitive Skills*, 10 AGGRESSION & VIOLENT BEHAV. 263, 273 (2005) (suggesting that gendered differences in social cognition development may explain differences in "delinquent behaviors"). One example is studies have found that males "have lower self-control than females" due to even "ineffective" parents being "more likely to control their daughters than their sons." Brenda Sims Blackwell & Alex R. Piquero, *On the Relationships Between Gender, Power Control, Self-Control, and Crime*, 33 J. CRIM. JUST. 1, 3 (2005). This decreased development of self-control may help explain higher crime rates for men. *See id.* (providing examples of studies both supporting and disputing this theory).

[94] United States v. Virginia, 518 U.S. 515, 542 (1996).