

The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky

Matthew DeMichele
Senior Research Sociologist
Center for Justice, Safety, and Resilience
3040 E. Cornwallis Road
Research Triangle Park, NC 27709-2194
919-541-6452
mdemichele@rti.org

Peter Baumgartner
Research Data Scientist
3040 E. Cornwallis Road
Research Triangle Park, NC 27709-2194
919-541-5807
pbaumgartner@rti.org

Michael Wenger
Research Data Scientist
3040 E. Cornwallis Road
Research Triangle Park, NC 27709-2194
mwenger@rti.org

Kelle Barrick
Research Sociologist
Center for Justice, Safety, and Resilience
3040 E. Cornwallis Road
Research Triangle Park, NC 27709-2194
919-541-6854
kbarrick@rti.org

Megan Comfort
Senior Research Sociologist
Behavioral and Urban Health Program
San Francisco, California
415-848-1375
mcomfort@rti.org

Shilpi Misra
Public Health Analyst
Center for Justice, Safety, and Resilience
RTI International, Hill 401
Phone: 919-485-2771
Email: smisra@rti.org

Abstract

In this paper, we assess the predictive validity and differential prediction by race and gender of one pretrial risk assessment, the Public Safety Assessment (PSA). The PSA was developed with support from the Laura and John Arnold Foundation (LJAF) to reduce the burden placed on vulnerable populations at the frontend of the criminal justice system. The growing and disparate use of incarceration is one of the most pressing social issues facing the U.S. Although it has received less attention, pretrial populations are a large and growing contributor of mass incarceration. The pretrial phase is often said to be the most consequential in the criminalizing process because it is related a higher likelihood of conviction, longer terms of incarceration, and has the potential to destabilize families. Recognizing the inherent challenges in pretrial release decisions, there has been increased development and use of pretrial risk assessments. Pretrial risk assessments are developed to identify the likelihood that defendants will remain crime free and that they will return to court. There have been several critiques of risk assessments, but none have assessed differential validity or prediction using pretrial outcomes. Using a statewide dataset from Kentucky ($n = 164,597$) we found the PSA to have predictive validity measures in line with what are generally accepted within the criminal justice field. We applied a regression modeling approach commonly used to assess bias in test instruments (e.g., cognitive and employment testing), and found some instances of differential prediction by race. These differences suggest that the PSA scores to predict failure to appear (FTA) are moderated by race, with no significant differences found for new crimes and new violent crimes between black and white defendants. The findings show differential prediction for new violent criminal arrests between male and female defendants, similar to what was found by Skeem et al. (2016). In the end, we point to data limitations that weaken external validity, point to areas for future research, and suggest that risk assessments are not silver bullets, but rather decision-making tools that require ongoing refinement.

Introduction

The growing and disparate use of incarceration is one of the most pressing social issues facing the U.S. After maintaining relatively stable incarceration rates for much of the 20th century, incarceration rates increased dramatically starting in the mid-1970s and have risen to well-over 700 per 100,000 adults incarcerated in prisons and jails (Carson, 2015). Mass incarceration has received attention from the media, policymakers, and researchers that has revealed several collateral consequences related to incarceration growth. The bulk of these studies and commentary focus on sentenced and convicted populations, which has resulted in a lack of understanding about pretrial processes and outcomes (Demuth, 2003).

Although it has received less attention, pretrial populations are a large and growing contributor of mass incarceration. According to the Bureau of Justice Statistics, the proportion of jail populations that are unconvicted has increased from 50 percent in 1985 to nearly 63 percent in 2014 (Minton and Zeng, 2015). Over this period, jail populations grew from about 256,000 to nearly 745,000, with BJS estimating that nearly 95 percent of the growth in jail populations since 2000 was due to the increase in the proportion of those confined in jails being held pretrial (Minton and Zeng, 2015).

The pretrial phase is often said to be the most consequential in the criminalizing process because it is related to several legal and personal outcomes (Sacks and Ackerman, 2012). During the period before trial, individuals are legally innocent and have a right to be released, but many jails are filled with pretrial populations. Judges make decisions about the release or detention of someone on a regular basis. For the most part, pretrial release decisions are based on the seriousness of the crime and prior criminal history (Gottfredson and Gottfredson, 1988; Spohn and Holleran, 2000), but these decisions are often made quickly and with limited information to make the most effective

decisions (Karnow, 2008). Pretrial release decisions are especially challenging because judges grapple with balancing public safety and protecting the community with the inherent rights of the accused.

The process of pretrial release and the reliance on financial conditions of release have “almost from its inception, been the subject of dissatisfaction” (Ares, Rankin, and Sturz, 1963: 67). The nature of these concerns has focused on the fairness by which pretrial release decisions are made and the potential for disparate treatment of the poor and vulnerable (e.g., Beeley, 1927; Foote, 1957). Pretrial detention is associated with a higher likelihood of conviction, longer terms of incarceration, and has the potential to destabilize families (Sacks and Ackerman, 2012).¹ The speed by which pretrial release decisions are made often results in legal actors having incomplete information and a high amount of discretion in which two criteria are the basis for release decisions: public safety and likelihood of returning to court (Goldkamp and Gottfredson, 1985; *United States v. Salerno*, 1985). Further, the legal rules for pretrial release allows judges to consider extralegal factors such as employment, community ties, and marital status when deciding whether to release someone (Goldkamp and Vilcica, 2009). These challenges to pretrial release are compounded by the reliance on financial conditions or bail as a requirement of release, with bail having an enduring history of negative impacts for the poor and communities of color (e.g., Ares et al., 1963; Demuth, 2003).

Recognizing the inherent challenges in pretrial release decisions, there has been increased development and use of pretrial risk assessments (Mamalian, 2011). Pretrial risk assessments are developed to identify the likelihood that defendants will remain crime free and that they will return to court. In general, there has been a slow adoption of risk assessment for pretrial release decisions,

¹ The association between pretrial release and detention with future crime is still being determined. Recent research by Dobbie, Goldin, and Yang (2017) did not find an association in Miami-Dade County between pretrial release or detention with committing a crime within four years. In two other studies, Gupta, Hansman, and Frenchman (2016) and Heaton, Mayson, and Stevenson (2017: 672) found in Philadelphia and Pittsburgh, and Harris County, Texas, respectively, that pretrial detention was associated with a 6-9% and 22% increase in crime within one year of release, respectively.

with an estimated 10 percent of jurisdictions using actuarial risk assessments (Clarke and Henry, 2007). And, these tools are emerging within a chorus of concerns about predictive utility and whether they contribute to racial disparities, with critics arguing that risk assessments rely on group based patterns that will unfairly treat people of color (e.g., Hannah-Moffat, 2015; Harcourt, 2008, 2015; Starr, 2014, 2015). Others, however, suggest that risk assessments structure criminal justice decisions, increase objectivity and fairness, and they have the potential to reduce incarcerated populations (Cooprider, 2009; Flores, Bechtel, and Lowenkamp, 2016; Skeem and Lowenkamp, 2016). The extent to which prediction bias exists in risk assessments is an empirical question that we address in the current paper. In this paper, we assess the predictive validity and differential prediction by race and gender of one pretrial risk assessment, the Public Safety Assessment (PSA).

The PSA was developed with support from the Laura and John Arnold Foundation (LJAF) to reduce the burden placed on vulnerable populations at the frontend of the criminal justice system (Lowenkamp, VanNostrand, and Holsinger, 2013; VanNostrand and Lowenkamp, 2013). The PSA is intended to assist judges and court professionals to quickly and accurately classify individuals for release or detention. The PSA includes prediction models for three outcomes during the pretrial phase: failure to appear (FTA), new criminal activity (NCA), and new violent criminal activity (NVCA). The PSA received widespread attention (Dewan, 2015) and was quickly adopted by several jurisdictions in California, North Carolina, Ohio, and Arizona. Following these initial implementation sites, the PSA continues to be implemented in numerous state and local jurisdictions, and, as of Winter 2018, there were over 38 state and local jurisdictions using the PSA. On any given day, thousands of pretrial release decisions are informed by the PSA. To date, however, the PSA has not been examined by external validation to assess overall validity or

predictive bias (i.e., differential prediction) by race and gender.² LJAF initially did not reveal the factors, weights, or scoring procedures to the public. Pilot jurisdictions were provided training and technical assistance and full disclosure of the factors, weights, and scoring procedures, and they were required to sign nondisclosure forms. More recently, LJAF has published the factors, weights, and scoring procedures on their website.³

We conduct analyses to investigate the following three primary research objectives using a dataset from a statewide pretrial services agency in Kentucky (n =164,597). First, we assess the overall predictive validity of the PSA. Second, we assess for differential validity and predictive bias between black and white defendants. Third, we conduct the same analyses to determine accuracy and assess if bias existed between male and female defendants. Although there are many studies about risk assessment development and validation, there are fewer published studies within the criminal justice literature that assess the potential for predictive bias by gender (e.g., Skeem, Monohan, and Lowenkamp, 2016; Van Voorhis, Salisbury, Wright, and Bauman, 2010; Walters and Lowenkamp, 2016) or race (Flores, Bechtel, and Lowenkamp, 2016; Skeem and Lowenkamp, 2016).

This paper is arranged in the following order. First, we briefly describe the emergence and use of risk assessments within the criminal justice system and the pretrial system specifically. Within this discussion, we review the general critiques of predictive bias in risk assessments, and highlight the more recent studies that have empirically studied this issue using methods commonly applied in organizational/industrial psychology and testing fields (e.g., Cleary, 1968; Sackett, Schmitt, Ellingson, and Kabin, 2001). Third, we describe the development of the PSA and how the PSA is used. Next, we describe our methods and provide the descriptive statistics, predictive utility

² Megan Stevenson (2017) has recently published an impact study showing that the PSA did not contribute to long term reductions in jail population, nor did the PSA exacerbate racial disparity in pretrial detention. However, her research did not include a validation of the PSA, as we provide here.

³ More information about the PSA can be found here: <http://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf>

measures, and test for predictive fairness. The PSA, in Kentucky, is found to provide what is considered by criminal justice researchers as a good level of overall predictive utility (ROC = 0.64-0.66), slightly weaker predictive utility for black defendants (ROC = 0.61-0.63) and females (ROC = 0.63-0.65) (Desmarais and Singh, 2013).⁴ But, we do not find these results to exacerbate disparate treatment by race and gender.⁵ Finally, we conclude by suggesting that pretrial risk assessment needs to develop a better understanding of the drivers of pretrial failures and to move away from searching for a statistical silver bullet.⁶ Instead, the field needs to develop normative standards (similar to what exist in other fields) of fairness and disparate impact, and to develop empirical standards of what merits a qualified empirical assessment of risk.

Risk Assessment and Pretrial Risk Assessment

The use of risk assessments in the criminal justice system is not new. These instruments have been used at least since the late 1920s when Burgess developed a parole risk assessment to help the Illinois paroling authority make release decisions. Since Burgess's time, there has been an increased development and use of risk assessments across the criminal justice systems (Harcourt, 2010) as probation and parole professionals use them to inform case plans, and other instruments are used to inform the supervision of domestic violence or sex offenders. More recently, there has been a push to introduce risk assessments at the pretrial (VanNostrand, 2003) and sentencing phases (Kleiman, Ostrom, and Cheesman, 2007).

⁴ The Demarais and Singh (2013: 12) AUC range standards are slightly lower than the AUC ranges commonly used to assess predictive utility when considering the fair range. They define 0.55-0.63 as fair, 0.64-0.70 as good, and 0.71-1.00 as excellent. Rice and Harris's (2005) AUC standards of 0.56, 0.64, and 0.71 as small, medium, and large effects are more commonly used across the social sciences. The Demarais and Singh (2013) ranges are used because they were derived from analysis of criminal justice risk assessments.

⁵ KY's pretrial population is about 17% black, whereas KY's overall black population is around 8%.

⁶ This is a point made most forcefully by several data scientists, with Berk et al. (2017: 34) arguing that criminologists and statisticians cannot alone decide what are the best risk assessments. Rather, stakeholders must weigh-in about what are the acceptable error rates and expected level of accuracy. Further, stakeholders need to agree and commit to the use of risk assessments in a consistent way, or move away from them.

Risk assessment development at the pretrial phase began as an opportunity to reduce potential disparate impacts related to bail. The first pretrial risk assessment instrument was developed in 1961 by the Vera Institute of Justice as part of the Manhattan Bail Project. Through this experiment, Vera showed that using a risk assessment instrument increased release rates and improved court appearance rates compared to relying on a charge based bail schedule. The Vera risk assessment included information about an individual's employment status, community/familial ties, criminal history, and associations. Jurisdictions slowly began to incorporate the Vera instrument into their pretrial processes and some jurisdictions created their own risk assessments. In this section, we provide a brief background about pretrial risk assessments, but do not fully cover all pretrial assessments in depth (for more information, see Mamalian, 2011).

Pretrial processes differ across the country, but most jurisdictions rely on the severity of the charged offense and criminal history when making determinations of bail.⁷ Pretrial risk assessments, for the most part, include the nature of the current charge (e.g., is it violent or a felony), but the argument for them is that the models include other non-charge items. The District of Columbia developed a pretrial risk assessment tool that included 22 items to measure criminal history, demographics, current criminal charges, and drug involvement (Winterfield, Coggeshall, and Harrell, 2003). Virginia developed a pretrial risk assessment instrument that uses nine factors with six of the factors measuring criminal history – charge type, pending charges, outstanding warrants, criminal history, prior failure to appear, and prior violent convictions. The remaining three factors assess residential stability, employment, and drug use (Danner, VanNostrand, and Spruance, 2016). The PSA differs from these instruments because the only non-criminal justice or behavioral item

⁷ Frase, Roberts, Hester, and Mitchell (2015) provide a thorough review of the use of criminal history to make sentencing decisions. They show how state sentencing guidelines are predicated on criminal history (e.g., often aggregated to create a prior criminal history score) and current charge (e.g., often weighted to develop an offense gravity score). Pretrial release decisions have yet to be standardized as such to make release decisions.

included is young age, whereas the Virginia and D.C. instrument include items that are clearly measuring socio-economic status (for a critique of this practice, see Starr, 2014).

Further, Lowenkamp, Lemke, and Latessa (2008) developed and validated an instrument. They studied a sample of 342 adult defendants on pretrial release in several pretrial agencies across two states to investigate the relationship between 64 items and pretrial failures. These analyses identified eight items (i.e., age at first arrest, history of FTA, FTA within two years, prior jail incarcerations, employment status, drug use, drug-related problems, and residential stability) to create a risk assessment score that was significantly related to both FTA and a new arrest.

More recently, pretrial risk assessment development has been advanced by researchers within the Administrative Office of the U.S. Courts, Office of Probation and Pretrial Services. Lowenkamp and Whetzel (2009; VanNostrand and Keebler, 2009) provided details about the development and validation of the Pretrial Services Risk Assessment (PTRA) created for the federal system. Developing the PTRA included reviewing existing pretrial risk assessment instruments and federal pretrial populations to learn that the federal “population of defendants differed enough from that of other pretrial services populations (for example, only federal courts address immigration charges) to warrant development of a tool using federal data”⁸ (Cadigan, Johnson, and Lowenkamp, 2012: 6). The Federal PTRA was developed using a sample of 565,178 defendants in the federal court system and began by assessing the relationship with over 70 potential predictors of FTA and NCA with most of these factors providing some measure of criminal history (i.e., felony convictions, prior FTAs, pending cases) and current offense (i.e., type, felony or misdemeanor), with other measures of

⁸ Skeem and Lowenkamp (2016) and Walters and Lowenkamp (2016) reported a similar finding that federal populations differed enough from state and local jurisdictions to warrant separate analyses. It is worth mentioning that the PSA was developed with nearly 1.5 million cases in which approximately 900,000 of them came from the Federal system.

age at interview, level of education, employment status, home ownership, and substance abuse (Lowenkamp and Whetzel, 2009).⁹

There are pretrial risk assessment instruments developed by six states, the District of Columbia, the federal court system, and about three dozen jurisdictions in approximately 15 states (Mamalian, 2011). The instruments rely mostly on measures of criminal history, but also tend to include community ties, residential stability, substance abuse, employment and education, and age. These factors are specifically at the heart of the controversy regarding using pretrial risk assessment because critics argue that the poor, people of color, and the most vulnerable are further penalized as these items do not have anything to do with an individual's criminal offense even if they are correlated with future crimes (Harcourt, 2010; Starr, 2014). Summarizing the general state of knowledge within pretrial risk assessment, Bechtel, Lowenkamp, and Holsinger (2011) conducted a meta-analysis of pretrial risk assessment instruments in which they found several significant but weak correlations with risk factors and outcomes. Their meta-analysis found that "risk items with the strongest correlations that were also in the expected direction are primarily static indicators, such as prior convictions, prior felonies, prior misdemeanors, prior failure to appear, and juvenile arrests" (Bechtel et al., 2011: 85). This finding suggests that empirical research and ethical critiques are beginning to align to suggest that pretrial risk assessments should only include factors directly related to one's criminal behavior. Of course, this is not to say that such an approach will eliminate any bias or disparate impact due to over-enforcement and punishment (i.e., different enforcement patterns) of people of color, but it does address some of the critiques that risk assessment scores are so correlated with race that they are merely a proxy.¹⁰

⁹ These non-criminal justice factors are left out of the official score of the PTRAs as Bureau of Prison researchers were conducting future research to determine their utility. Cohen and Bechtel (2017) analyzed the non-scored items from the PCRA and found insignificant improvements in predictions, and hence recommended removing them.

¹⁰ Bernard Harcourt (2010) draws a stark line in the sand about risk being a proxy for race because of different enforcement and long-term patterns of bias against people of color by criminal justice systems. He argues that criminal history items are suspect when it comes to making sentencing decisions. Other critics (e.g., Hannah-Moffat, 2013; Starr,

The prior pretrial risk assessment research demonstrates that court professionals can use a relatively small set of measures to classify defendants by risk. Relying on risk of failure to appear and the commission of new crimes has been shown to reduce failures and reliance on cash bond (Coopridner, 2009; Levin, 2006), reduce jail populations (Mahoney, Beaudin, Carver, Ryan, and Hoffman, 2001), and support less restrictive conditions (Toberg, Yezer, Tseng, and Carpenter, 1984). More recently, however, Stevenson (2017) conducted research in Kentucky using a pretrial dataset from July 2009 through 2016, and she found that despite initial reductions in jail populations using the PSA those jail populations increased over time. She also found slight increases in FTA and NCA rates, but did not find any increase in racial bias due to the use of the PSA. Although research is unclear about whether or not pretrial risk assessments contribute to lowering pretrial/jail populations, it is less ambiguous that pretrial detention has negative impacts for the detained and their families.

Critique of Risk Assessments: Is Risk a Proxy for Race?

Risk assessments are challenged on the basis that they introduce systematic bias into criminal justice decisions by unfairly punishing certain subpopulations. The bulk of these critiques have come from legal scholars suggesting that criminal history (due to an ongoing legacy of over enforcement of communities of color) is a correlate for race (Harcourt, 2010) or that factors related to socio-economics (due to entrenched institutionalized forms of oppression and exclusion) are approximate measures of poverty and race (Starr, 2014, 2015). These critiques frame risk assessment as using variables that correlate so heavily with race and poverty that they merely provide a way of using scientific discourse and methods to hide differential prediction or bias (for a critique, see McIntyre and Baradaran, 2013). One argument is that risk assessments are unfair because they include factors

2014) focus on the non-criminal justice related items, and suggest that – despite differential enforcement – criminal history and current offense are appropriate factors to consider at sentencing.

such as race, gender, age, education, and residential stability. The PSA includes – as do several other assessments – one of these measures to account for age using a young offender factor. Tonry (2014) is a vocal critic of the use of age to make sentencing decisions because he argues that youths have yet to fully develop cognitively, which puts them at a lower level of moral culpability. The critique against risk assessment goes beyond the inclusion of socio-economic factors,¹¹ but rather Harcourt (2010) argues that the legacy of unfair criminal justice practices (and general forms of racism) makes it (nearly) impossible to remove bias from most risk factors (e.g., criminal history).

To address these criticisms, Monahan, Skeem, and Lowenkamp (2017), Skeem, Monahan, and Lowenkamp (2016), and Skeem and Lowenkamp (2016) have conducted studies to assess the predictive utility and predictive fairness of the federal system’s Post Conviction Risk Assessment (PCRA) by age (young vs. older), gender, and race (black vs. white). The federal datasets used to assess differential prediction and disparate treatment with the PCRA – similar to many criminal justice datasets – were characterized by subgroup mean differences (i.e., there are differences in recidivism conditioned by age, race, and gender). In each of their analyses, they use a moderator regression technique commonly cited in psychological studies and testing literature (e.g., Cleary, 1968; Sackett, Borneman, and Connelly, 2008). This approach estimates four regression models to assess what is referred to as calibration to understand the extent to which “...a given score will have the same meaning regardless of group membership (e.g., an average risk score of X will relate to an average recidivism rate of Y for all relevant [sub] groups)” (Monahan, Skeem, and Lowenkamp, 2016: 193). Predictive bias is tested by assessing the extent to which subgroups have similar (i.e., not significantly different) intercepts and slopes (i.e., they possess similar regression lines). The

¹¹ Harcourt provides a thorough review of the historical development of risk assessments – mostly used for parole release decisions – to show that many of those instruments included direct factors such as race, place of birth, and parent’s nationality. Readers should also see the more contemporaneous treatment of race in sentencing risk assessment by Kleiman, Ostrom, and Cheesman (2007) in which they used regression models that included race as a control, but removed it from the instrument. Gender has been used until recently in many assessments.

moderator regression technique is recommended by the Standards for Educational and Psychological Testing (AERA, American Psychological Association, & NCME, 2014) and has been used widely by psychometricians and organizational scholars (e.g., Cleary, 1968). Calibration – knowing that people with similar scores are treated similarly – fits with utilitarian notions of fairness with the equal administration of law and justice delivered without favor.

Skeem et al. provided the first applications of the moderator regression approach to criminological literature to assess predictive fairness using the PCRA.¹² Regarding gender, they found that the PCRA strongly predicts recidivism for both genders, but overpredicts for women. That is, women, on average, received higher scores due to the influence of male scores driving estimates higher (Skeem et al., 2016). They suggested that excluding gender as a risk variable – as the PCRA does – can lead to over punishing women (i.e., seeing them as riskier) because the predicted probabilities are heavily influenced by the higher offending patterns of males.

For race, they found that the PCRA strongly predicts recidivism for black and white individuals and they found “little evidence of test bias...across groups” (Skeem and Lowenkamp, 2016: 680). They did observe that black individuals were more likely (than whites) to have higher PCRA scores due to higher criminal history scores, with criminal history mediating the relationship between race and recidivism.¹³ They indicated that, although they did not find test bias, the race group differences in “...some applications [of the PCRA] could create disparate impacts” for African-Americans (Skeem and Lowenkamp, 2016: 680).

In their study assessing predictive bias by age, Monahan et al. (2017) found that the PCRA overestimates recidivism rates for older individuals and underestimates recidivism for younger

¹² Flores, Bechtel, and Lowenkamp (2016) used a similar approach with the COMPAS, not the PCRA.

¹³ There is no way to assess the causes for these higher criminal history scores. That is, whether these differences fit Harcourt and others' critique that higher criminal history scores reflect differential enforcement (e.g., over policing and enforcement of people of color), or if these differences are higher simply due to people of color having higher criminal propensity.

individuals. They did not find differences in the slopes (form) of the relationships between age and the PCRA scores with re-arrest, but they did find that arrest increases with decreasing (younger) age. Although age was not found to moderate the relationship between PCRA scores and arrest, "...age adds small, but significant incremental utility (differences in intercept) to the PCRA in predicting both arrest and violent arrest" (Monahan et al., 2017: 196). The PCRA – and the PSA – include measures of age, and Monahan et al. (2017: 200) offer suggestions for ways to improve the use of age on risk assessments (i.e., change the age categories, alter the weights, adjust PCRA score interpretations for different age groups).

Fairness: Definition and Measurements

Skeem et al.'s research is an important contribution to the criminological literature and increase our understanding of the potential utility and predictive fairness for different subgroups when using risk assessments. To date, we are unaware of a similar approach applied to pretrial defendants with pretrial outcomes. The closest application is that of authors critiquing one study finding disparate impacts for communities of color using the COMPAS at pretrial. ProPublica (Angwin, Larson, Mattu, and Kirchner, 2016) analyzed a dataset of pretrial defendants from Broward County, Florida and posited that the COMPAS¹⁴ resulted in classification errors that negatively impact black defendants. Flores, Bechtel, and Lowenkamp (2016) critiqued this research on several methodological and substantive grounds, and reanalyzed the Angwin et al. (2016) dataset.¹⁵ In the reanalysis, Flores et al. (2016) did not find meaningful differences attributable to prediction bias using the moderator regression approach recommended by the Standards for

¹⁴ Much of the critique about risk assessments that have surfaced regarding the COMPAS have focused on the lack of transparency regarding the contents of the COMPAS, the research behind the instrument, and dissemination of ongoing assessments. On the surface, then, transparency is a first step toward improving pretrial risk assessments.

¹⁵ The datasets were not identical as Flores et al. (2016) reduced the dataset somewhat as they focused on the differences between blacks and whites only, whereas Angwin et al. (2016) included other races/ethnicities.

Educational and Psychological Testing. Although Flores et al. (2016) used a pretrial sample, the outcomes were not related to the pretrial period as is the current study.

Critically, there are several competing definitions of fairness measures, formalized by Kleinberg et al. (2016) and further illustrated on the COMPAS instrument data by Chouldechova (2017). The lack of formalization can be viewed as the source of disagreement between ProPublica and Flores et al. (2016), as the ProPublica analysis assessed error rate balance (i.e., equal false positive and false negative rates across races) and the moderator regression approach assesses calibration (i.e., showing that a score X has the same meaning for racial groups). Kleinberg et al. (2016) showed that the fairness measures are mathematically incompatible, that is, it is impossible to satisfy all definitions if base rates differ among populations. Furthermore, Berk et al. (2017) push this message even further by positing accuracy and fairness are conflicting goals, stating that “if there is a policy preference, it should be built into the algorithm.” Policy preferences and desired goals for any risk instrument need to be thought out, agreed upon, and articulated from the beginning, with some acknowledged tradeoff between accuracy and fairness. If, as Skeem and Lowenkamp (2016) argued, that risk assessments can “unwind mass incarceration,” then such a policy preference should be stated and the risk assessment developed to maximize reductions in incarcerated populations (Berk et al., 2017: 14)

Assessing risk assessment predictive utility and fairness are essential tasks. Although criminologists are studying the potential for disparate impact related to risk assessment in the post-conviction context (e.g., Oliver, Stockdale, and Wormith, 2013), we are unaware of a similar application in the pretrial context.

Public Safety Assessment

The PSA, which was created through investments made by LJAF using a large database of over 1.5 million cases drawn from more than 300 U.S. jurisdictions, with analysis conducted on

750,000 suitable cases to examine the predictive validity of hundreds of risk factors (VanNostrand and Lowenkamp, 2013). The PSA was developed to identify the strongest predictors of FTA, NCA, and NCVA. Criterion for variable selection were that the predictors needed to be related to the current charge (i.e., violent or not) or criminal history related,¹⁶ consistent with prior research, and gathered without a defendant interview (VanNostrand and Lowenkamp, 2013). Following the initial development of the PSA, researchers conducted validation analyses on a sample of over 500,000 cases (i.e., validation sample) from jurisdictions in the Northeast, Southwest, Midwest, and two states (unpublished Luminosity training materials).¹⁷

The PSA differs from many pretrial risk assessments in three important ways. First, the PSA relies on administrative records only and can be completed without conducting an interview with the defendant. This is a nontrivial issue because forgoing the interviews is expected to allow for assessing more defendants in less time, which has the potential to provide quicker arraignment/first appearance and less time until release decision. Second, LJAF created the PSA with intentions of creating a risk assessment that could be used by jurisdictions across the country. Many of the pretrial risk assessment instruments were not intended to be used outside of the jurisdiction in which they were developed.¹⁸ Third, the PSA includes the ability to predict the likelihood of a future new violent criminal act during the pretrial phase (something other pretrial assessments do not include). The Foundation has released a brief description of the methods used to develop the PSA

¹⁶ The PSA intentionally leaves out critical demographic factors related to race/ethnicity and gender as well as socio-economic variables such residential stability, educational attainment, and employment. These items were excluded to reduce potential for predictive bias for the poor and communities of color. Young age is the one demographic variable included in the PSA.

¹⁷ LJAF has developed an ongoing pretrial research arm that is not fully described here. Readers are encouraged to visit LJAF's website to read more detailed information about the research used to develop the PSA and ongoing validation efforts. <http://www.arnoldfoundation.org/initiative/criminal-justice/crime-prevention/public-safety-assessment/>

¹⁸ There are notable exceptions such as the Ohio Pretrial Risk Assessment that has been adopted by pretrial agencies in Indiana. Additionally, it is common within the criminal justice system for agencies to forego development of a localized instrument due to cost restraints and to simply adopt an assessment developed in another jurisdiction. But, of course, universal backend assessments exist (e.g., the Level of Service Inventory).

(<http://www.arnoldfoundation.org/wp-content/uploads/Criminal-Justice-Data-Used-to-Develop-the-Public-Safety-Assessment-Final.pdf>).¹⁹

Use of the Public Safety Assessment

Kentucky is often recognized as a leader in pretrial services and, in 1976, they became one of four states to ban commercial bail bonding services, which made them one of only a few states to have statewide pretrial services. Kentucky pretrial services incorporated the Vera risk assessment tool in 1976, and implemented a new Kentucky Pretrial Risk Assessment (KPRA) in 2006. The KPRA included several criminal history factors, prior FTA, and non-criminal justice factors including housing and employment status, and included an interview with defendants (see Austin, Ocker, and Bhati, 2010). Kentucky's jail and prison population, similar to much of the country, grew throughout the 2000s and policymakers were looking for ways to reduce the burden on the criminal justice system. In July, 2011, House Bill 463 went into effect in Kentucky to mandate the use of a validated risk assessment tool to measure a person's flight risk and threat to public safety (for a review, see Stevenson, 2017). In July 2013, Kentucky became the first jurisdiction to use the PSA, with LJAF researchers (e.g., Luminosity) conducting ongoing research and modifying the PSA as needed (with Stevenson, 2017 reporting changes adopted in KY in mid-year 2014).

The PSA is completed by pretrial officers or other relevant court personnel prior to first appearance. Pretrial officers use administrative data and conduct a thorough review of criminal history records. The risk assessment instrument includes a total of nine factors to develop three

¹⁹ The instrument development team – led by Drs. Marie VanNostrand and Christopher Lowenkamp - processed these datasets to identify the predictors of each of the three outcome variables. They used a series of statistical techniques (e.g., logistic regression, contingency tables) that produced hundreds of effect sizes. The effect sizes were averaged, and were restricted to variables that were at least one standard deviation above the mean effect size. Further analyses were conducted to identify the best effect sizes and operationalization in which each predictor variable had at least a 5 percent increase in likelihood of failure to appear or new criminal activity. The new violence criminal activity flag used a variable selection criterion of doubling the probability of failure when the item was included in a model (this paragraph is adapted from unpublished materials by Luminosity).

prediction models (one for each outcome). Below are the three outcomes and each of the factors included in the predictive models:

- Failure to appear (FTA): pending charge at time of arrest, prior conviction, prior failure to appear within two years, and prior failure to appear longer than two years.
- New criminal activity (NCA): pending charge at time of arrest, prior misdemeanor conviction, prior violent convictions prior, felony conviction, prior failure to appear within two years, prior sentence to incarceration, young age at current arrest.
- New violent criminal activity (NVCA): pending charge at the time of arrest, prior conviction, prior violent conviction, current offense violent, and current offense violent * young age at current arrest.

The presence or absence of each factor adds a specific value to the overall risk score, which is then scaled down to separate FTA and NCA scales that range from 1 to 6, and a new violent criminal activity flag (i.e., binary indicator of yes/no).²⁰ The NVCA flag is used to offer stakeholders an indicator of an elevated risk of violence.

The FTA and NCA scale scores are converted into recommendations for each defendant through a decision-making framework. The decision-making framework provides policy-based guidance that can range from release on own recognizance, various levels of supervision, and recommended not for release. The specific way the risk assessment instrument is completed varies to fit each jurisdiction's standard operating practices and courtroom culture.

Current Study Methods

The data used for this validation study were provided to us by LJAF. Using the dataset and documentation given from LJAF, we constructed what data scientists often refer to as a tidy dataset (Wickham 2014). That is, the data we were provided with allow for the analysis of the PSA factors using the risk factors, gender, and race, and each of the three outcomes. This dataset was collected

²⁰ The NVCA raw scores are converted to a 6-point scale score prior to being collapsed into the binary flag.

as part of post-development validation of the PSA, and provided to the current authors. The unit of analysis are cases in the KY pretrial system.²¹ Variable creation for each of the risk factors, i.e. matching individuals to criminal history and determining the presence or absence of specific risk factors, was completed by the PSA developers prior to the authors receipt of the dataset. The dataset was collected and processed by Luminosity as part of their ongoing PSA research and development. Given the risk factors for each outcome, we scored each case using the scoring criteria adhering to the PSA scoring system.

The Kentucky “validation” dataset contains 286,247 cases from Kentucky. Building on the criteria originally developed by Luminosity for case inclusion with the following conditions:

- 1) A booking date before January 1, 2015 (removed 45,299 cases)
- 2) An age at booking of at least 18 (removed 679 cases)

This processing resulted in a dataset of 240,219 cases. Following that, the release and detention status of the case was calculated as follows:

- 1) If a case had a missing release and disposition date, it is “detained”
- 2) If a case had a missing release date and a present disposition date, it is “detained”
- 3) If a case had both release and disposition date, and the release date happened after the disposition date, it is “detained”
- 4) Cases with none of the above definitions of “detained” were considered “released”

In table 1, we report that 164,597 (68.5%) individuals were released and 75,662 (31.5%) were detained based on the above definitions.

With the existing data and calculated scores, we formed an analytic dataset containing the following variables:

- An indicator of whether that individual was originally released or detained
- All the risk factors across the 3 PSA models
- Actual outcomes (FTA, NCA, NVCA) for each case
- PSA scores for each outcome given the risk factors
- Race coded for Black and White defendants

²¹ The unit of analysis are cases, which means that any individual could be in the dataset multiple times for multiple arrests. This is the same unit of analysis used by the PSA developers, and used by Stevenson (2017, footnote 191, pg. 37).

- Gender coded for Male and Female defendants

These datasets are used to address the following research objectives:²²

- *Assess Overall Predictive Validity*: How accurately does the PSA predict each of the three outcomes? The PSA relies on extensive research and prior knowledge of pretrial failures, which leads us to expect moderate to strong predictive validity.
- *Assess Predictive Validity by Race and Gender*: How accurately does the PSA predict each of the three outcomes of interest by race and gender?
- *Assess Differential Prediction by Race and Gender*: Does the PSA provide different results based on race and gender? We expect to find that the PSA predicts equally well across race and gender (i.e., race and gender will not moderate the relationship between the PSA and failures).

Analysis

Sample Description

Our first research aim is to address the overall predictive utility of the PSA across the three pretrial outcomes. Pretrial studies require identifying individuals that are booked into jail but have been released into the community. Table 1 shows that in Kentucky 68 percent (n = 164,597) of defendants were released. Individuals were booked between July 1, 2013 to December 30, 2014. The analyses focus on those that were released.

²² We follow recent practices and set more stringent statistical significance levels at $p < .001$ due to the large sample sizes used for both jurisdictions. For example, Monahan et al. (2017) followed this approach with a dataset of 7,350, which is much smaller than either of the datasets used here.

Table 1. Distribution of the Samples: Release Status, Gender, Race, and Base Failure Rates

	Kentucky N (%)
Release Status	
Detained	75,662(31.5)
Released	164,597(68.5)
Released Cases	
Sex²³	
Male	113,376 (68.9)
Female	50,592 (30.7)
Race	
Black	27,656 (16.8)
White	133,517 (81.1)
Other	3,424 (2.1)
Base Failure Rates	
FTA	24,293 (14.8)
NCA	17,512 (10.6)
NVCA	1,826 (1.1)

In table 1, we report the racial and gender distributions of the released sample, with 81 percent (n = 133,517) of the cases are white, and nearly 17 percent black.²⁴ Nearly, 70 percent of the cases are men (n = 113,376). Kentucky has an FTA rate of 14.8 percent, an NCA rate of 10.6 percent, and an NVCA rate of 1.1 percent.²⁵

²⁴ According to the US Census, Kentucky has an overall population of 4,454,189, 85 percent white non-Hispanic, 8 percent black. <https://www.census.gov/quickfacts/KY>

²⁵ Our sample description is similar to what is reported by Austin et al. (2010) with a 74% release rate for a 3-month validation study (July through September 2009), although they had lower FTA rates (8%) and NCA (7%). Stevenson (2017) does not report the overall release rate, but she reports that 77% of misdemeanors and 62% of felony cases were released prior to disposition. She also reports that 10% and 8% of misdemeanor and 13% and 8% of felony defendants with an FTA or NCA, respectively.

FTA Factors, Failure Rates, and Scores

The PSA includes four factors to predict FTA. In table 2 and figure 1, we show that the factors are related to higher FTA rates in Kentucky. Table 2 shows that 19 percent of individuals had a pending charge at the time of their arrest, 30 percent had one or more prior FTAs within the past 2 years, and nearly 75 percent had at least one prior conviction.

Table 2. Number and Percent with FTA Risk Factors

Factor		Number	Percentage
Pending Charge	Yes	31,294	19.0
	No	133,303	81.0
Prior FTA in Past 2 Years	Two or More	21,347	13.0
	One	28,993	17.6
	No	114,257	69.4
Prior FTA Older than 2 Years	Yes	67,972	41.3
	No	96,625	58.7
Any Prior Conviction	Yes	122,545	74.5
	No	42,052	25.5

In figure 1, we include the risk factors and their associated FTA rates, and a bootstrapped 95% confidence interval is displayed at the end of each bar. This figure illustrates that all of the FTA risk factors are associated with higher FTA rates. For example, about 13 percent of cases without a pending charge had a new FTA, whereas about 22 percent of those with a charge pending at the time of their arrest had a new FTA. This bivariate pattern holds for the other factors, with increasing FTA rates for defendants with the risk factor.

Figure 1. FTA Rates by FTA Factors

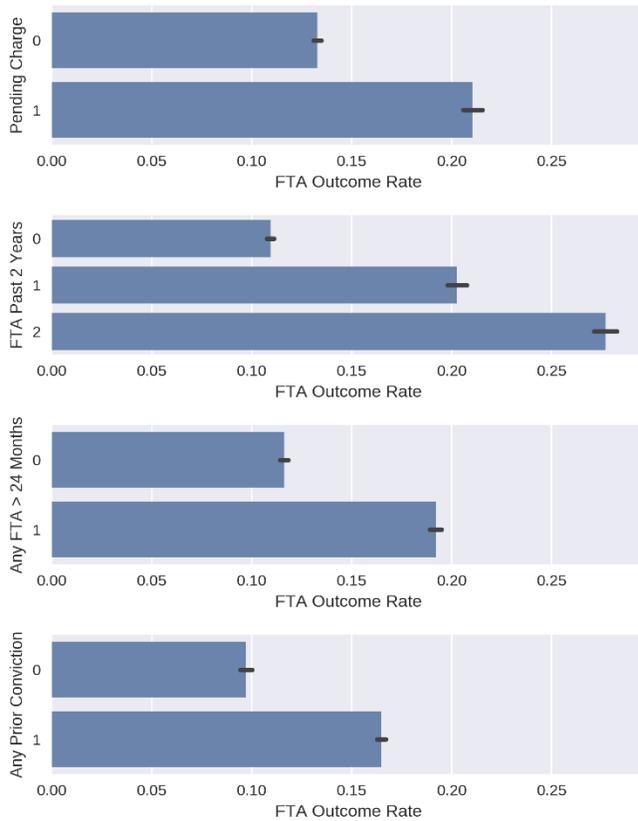


Table 3 includes Pearson’s correlation coefficients (r) for each of the FTA risk factors and a new FTA. The correlations are small, but positive, indicating a positive relationship for each risk factor with a new FTA. Number of FTAs in the past 2 years ($r = .172$) shows the strongest correlation, with the overall model having an $r = .188$.²⁶

Table 3. Pearson Correlations between Risk Factors and FTA

Variable	Correlation with Outcome
Pending Charge	0.086*
FTA Past 2 Years	0.172*
Any FTA > 24 Months	0.105*
Any Prior Conviction	0.083*
PSA FTA Score	0.188*

* = $p < .001$

²⁶ Following Cohen (1988: 79), correlation coefficients between .10 and .29 are considered small, .30 and .49 are considered moderate, and those of .50+ are strong. The significant tests assess whether the correlation coefficients are greater than zero.

The four factors used in the FTA scale range from 0-7 points. Three of the factors are binary indicators (no = 0, yes = 1), except for prior FTA within past 2 years which is scaled as 0 = 0, 1 = 2, and 2+ = 4. These weights are converted into an FTA scale score that ranges from 1-6. The score conversions are as follows: 0 = 1, 1 = 2, 2 = 3, 3 and 4 = 4, 5 and 6 = 5, and 7 = 6. In table 4, we report the proportion of each sample by their FTA score and failure rates. The rate of FTAs increases with each increase in the FTA score. In Kentucky, FTAs range from about 7.5 percent to 32 percent within each of the scores, with scores of 1-3 below the average overall FTA rate, and scores of 5-6 are approaching or exceeding twice the overall FTA rate.

Table 4. Proportion of each sample by their FTA score and failure rates

FTA Scale Score	Kentucky N	Kentucky FTA %
1	2,186	7.5
2	4,171	9.7
3	5,287	13.9
4	5,901	19.8
5	5,163	26.5
6	1,585	32.1
Total	24,293	14.8

NCA Factors, Failure Rates, and Scores

In table 5, we report the number and proportion of the cases with each risk factor for the NCA scale. The NCA scores include two of the FTA risk factors (i.e., pending charge, prior FTA within 2 years), and the NCA scores include more detail about an individual’s criminal history. NCA scores provide information about misdemeanor convictions (72.8%), felony convictions (29.2%), and violent convictions (21.8%). Additionally, about one-third of the defendants have been incarcerated in the past, and 16 percent are 22 years of age or younger.

Table 5. Number and Percent with New Criminal Activity Risk Factors

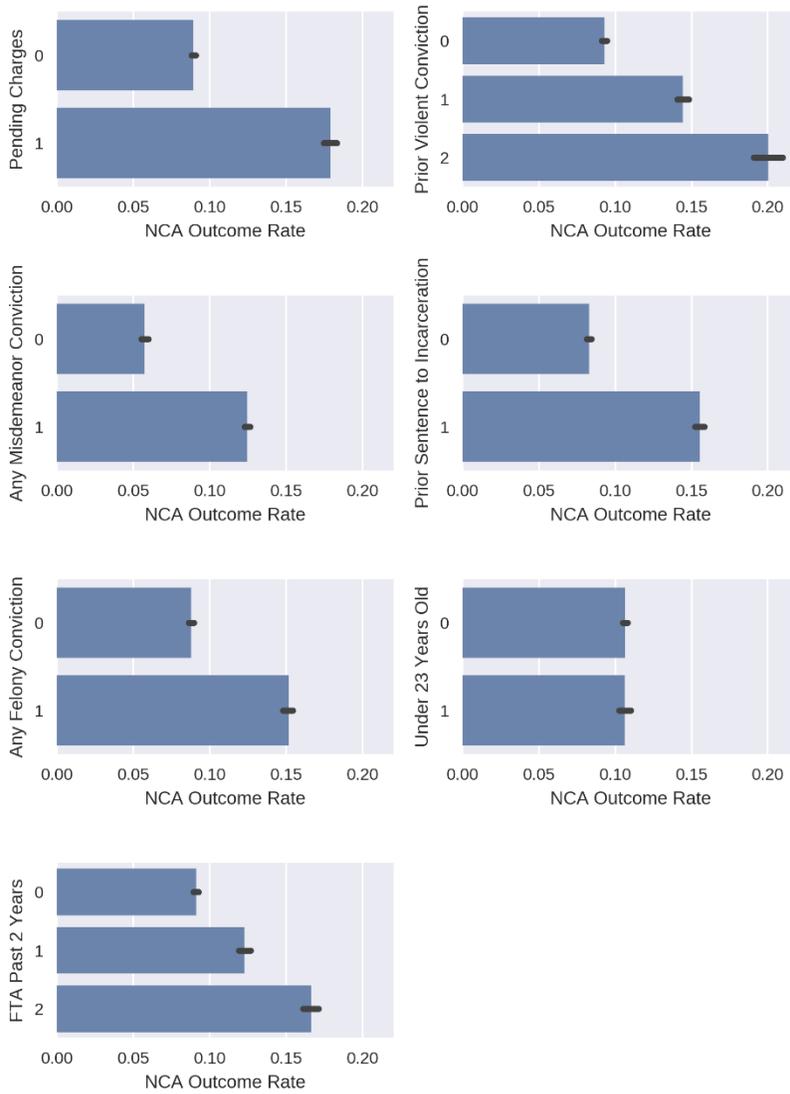
Factor		Number	Percentage
Pending Charge	Yes	31,294	19.0
	No	133,303	81.0
Prior Misdemeanor Conviction	Yes	119,875	72.8
	No	44,722	27.2
Prior Felony Conviction	Yes	48,034	29.2
	No	116,563	70.8
Prior FTA in Past 2 Years	Two or More	21,347	13.0
	One	28,993	17.6
	No	114,257	69.4
Prior Violent Conviction	Three or More	6,643	4.0
	One to Two	29,322	17.8
	No	128,632	78.1
Prior Sentence to Incarceration > 14 days	Yes	53,288	32.4
	No	111,309	67.6
Current Age	<= 22 Years	26,720	16.2
	>= 23 Years	137,877	83.8

The associations between the risk factors and outcome rates are presented in Figure 2.

Whether someone is under 22 years of age does not differentiate between individuals that have an NCA during pretrial as 12 percent of those under and over 22 years of age have an NCA.

Defendants with 2 or more violent convictions (20%) have more than twice the NCA rate as defendants without any violent conviction (9%).

Figure 2. NCA Rate for NCA Factors



In table 6, we report the correlation coefficients for the NCA factors with new NCAs. Similar to FTAs, the results provide positive yet weak coefficients. The individual risk factors do not have strong associations with NCAs, with pending charge and prior sentence to incarceration (> 14 days) having the strongest association ($r = 0.11$).

Table 6. Pearson Correlation Coefficients between Risk Factors and NCA

Variable	Correlation with Outcome
Pending Charge	0.114*
Prior Misdemeanor conviction	0.097*
Prior Felony conviction	0.094*
FTA Past 2 Years	0.084*
Prior Violent conviction	0.089*
Prior sentence to incarceration	0.110*
Under the Age of 23 years	0.000
PSA NCA Score	0.171*

* = p < .001

The seven NCA factors range between 0 and 13 with the following weights applied: Prior misdemeanor (No = 0, Yes = 1), Prior felony conviction (No = 0, Yes = 1), Pending charge (No = 0, Yes = 3), Prior incarceration sentence (No = 0, Yes = 2), Prior violent convictions (0 = 0, 1 or 2 = 1, 3+ = 2), Prior FTA in past 2 years (0 = 0, 1 = 1, 2+ = 2), and Age at current arrest (23+ = 0, 21 and 22 = 2, 20 or younger = 2). These weights are converted into an NCA scale score that ranges from 1 to 6. The conversions are as follows: 0 = 1, 1 and 2 = 1, 3 and 4 = 3, 5 and 6 = 4, 7 and 8 = 5, 9-13 = 6.

In table 7, we report the NCA scale scores and their associated NCA rates. The NCA rate increases as the NCA scale scores increase. There is nearly a seven-fold increase, 3.9 percent vs. 26.3 percent, in the rate of NCAs between a scale of 1 versus 6. NCA scale scores between 1-2 are below the overall NCA rate, and an NCA score of 3 is slightly above the base rate (10.9 vs. 10.6). NCA scores of 5 and 6 are nearly fifty percent larger than the NCA base rate, and a score of 6 is nearly 2.5 times the base rate.

Table 7. Proportion of each sample by their NCA score and failure rates

Score	Kentucky NCA N	Kentucky NCA %
1	834	3.9
2	3,575	6.8
3	4,499	10.9
4	4,769	15.1
5	2,513	19.7
6	1,322	26.3
Total	17,512	10.6

NVCA Factors, Failure Rates, and Scores

In table 8, we report the number and percent of defendants with each NVCA risk factor.

The NVCA risk scale incorporates any prior conviction, prior violent convictions, and pending charges, with two factors not included in either the FTA or NCA scales. The new factors are current offense is violent (14.6 percent) and whether the current offense is violent and the defendant is 20 years old or younger (1.4 percent).²⁷

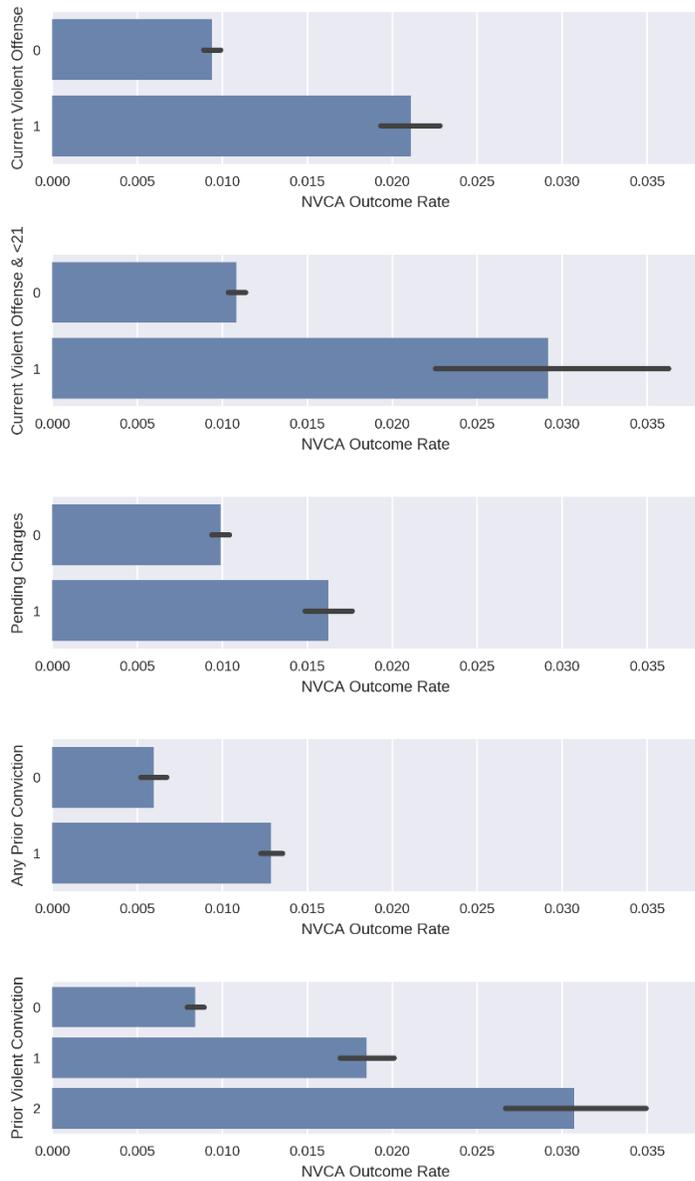
Table 8. New Violent Criminal Activity Outcome by New Violent Criminal Activity Risk Factors

Factor		Number	Percentage
Current Violent Offense	Yes	23,986	14.6
	No	140,611	85.4
Current Violent Offense & ≤ 20 Years Old	Yes	2,263	1.4
	No	162,334	98.6
Pending Charge	Yes	31,294	19.0
	No	133,303	81.0
Any Prior Conviction	Yes	122,545	74.5
	No	42,052	25.5
Prior Violent Conviction	Three or More	6,643	4.0
	One to Two	29,322	17.8
	No	128,632	78.1

²⁷ Our analyses only include individuals 18 years of age and older as we are focused on adults.

Figure 3 displays the NVCA rates for each NVCA risk factor with 95% bootstrapped confidence intervals. The confidence intervals appear large due to the low number of individuals with the an NVCA.

Figure 3. NVCA Rate for NVCA Factors



In table 9, we report the correlation coefficients for the NVCA risk factors and NVCA. The coefficients are positive, but very weak. The correlation coefficients do not approach 0.10 for

NVCAs by risk factors, with the strongest association between prior violent conviction ($r = 0.053$) - not current violent offense or the current violent offense and young age factor.

Table 9. Pearson Correlation Coefficients between Risk Factors and NVCA

Variable	Correlation with Outcome
Current Violent Offense	0.039*
Current Violent Offense & <21	0.020*
Pending Charges	0.024*
Any Prior Conviction	0.029*
Prior Violent Conviction	0.053*
NVCA Score	0.067*
Violent Flag	0.048*

* = $p < .001$

The NVCA risk scores range from 0 to 7. Three of the factors are binary indicators measured as 0,1; current violent offense is a binary factor measured 0,2; and prior violent conviction is measured as follows: 0 = 0, 1 and 2 = 1, and 3+ = 2. These scores are converted into a scale score ranging from 1-6 as follows: 0 = 1, 1 = 2, 2 = 3, 3 = 4, 4 = 5, 5 or above = 6. The NVCA scale scores are used to create a binary indicator, with defendants with an NVCA score of 5 and 6 receiving a violent flag to suggest they have a higher likelihood of committing a violent crime during their pretrial release. Table 10 show that few defendants ($n = 325$) that received a violent flag have an NVCA rate three times as large as those without the flag. New violent crimes have a base rate of 1.1 percent overall, 1 percent of those without a violent flag are arrested for a violent crime, whereas 3 percent of those with the violent flag are arrested for an NVCA.

Table 10. New Violent Criminal Activity Outcome by Violent Flag

	Kentucky NVCA N	Kentucky NVCA %
Violent Flag (5-6)	325	3.0
No Violent Flag (1-4)	1,501	1.0
Total	1,826	1.1

Predictive Utility

We assess the predictive utility (i.e., accuracy) of each of the three models using Area Under the Curve (AUC) Receiver Operator Characteristics (ROC) estimates. AUCs are commonly used to evaluate risk assessment tools (Singh and Falzer, 2010) because they are not influenced by base rates and allow for making comparisons across models and groups (Swets 1988). The ROC scores range from 0 to 1.0 with 0.5 referring to random chance and 1.0 referring to perfect prediction. The ROC score provides a rather intuitive interpretation as it reports the likelihood that when randomly selecting a case that had one of the outcomes, that case would have a higher score on the PSA than a randomly selected case that did not have one of the outcomes.

Table 11. Area Under the Curve Receiver Operator Characteristics

	FTA AUC	NCA AUC	NVCA AUC
Kentucky	0.646	0.650	0.664

* = $p < .001$

Recently, Demarais and colleagues (Demarais and Singh, 2013; Demarais, Johnson, and Singh, 2016) have conducted meta-analyses and evaluations of criminal justice risk assessment instruments. They suggested that AUC values of 0.54 and below are poor, 0.55 to 0.63 are fair, and 0.64 to 0.7 are good, with values higher than 0.71 being excellent. Using these ranges, the ROC values for PSA for the three outcomes are in the good range. The FTA ROC (0.646) reaches the lower bound of what is considered good, they are slightly stronger for NCAs (AUC = 0.650), and stronger for NVCAAs (AUC = 0.664).²⁸

The PSA and Race and Gender

In this section of the paper, we assess to what extent the PSA scale scores exhibit differential utility and predictive bias by race and gender. We begin by presenting the base failure rates for black

²⁸ These ROC scores are similar to Danner et al.'s (2016) research on the Virginia pretrial risk assessment, but lower than what Skeem and Lowenkamp (2016) reported for the PCRA. Austin et al.'s (2010) validation of the Kentucky pretrial risk assessment does not include ROC scores.

and white defendants and male and female defendants for FTAs, NCAs, and NVCA. The Kentucky defendants have significant differences in FTA and NVCA rates between black and white defendants. Black defendants, relative to white defendants, have a nearly 20 percent higher FTA rate, little difference in NCA rates, and almost double the NVCA rate ($p < .001$). Male defendants, relative to females, have no difference in FTA rates, significantly higher NCA rates ($p < .001$), and higher NVCA rates ($p < .001$).

Table 12: Base Failure Rates by Race and Gender

	Kentucky Failure N	Kentucky Failure Rates (%)
Race		
FTA		
Black	4,712	17.0*
White	19,122	14.3
NCA		
Black	3,084	11.1
White	14,276	10.7
NVCA		
Black	490	1.7*
White	1,321	0.9
Gender		
FTA		
Male	16,767	14.8
Female	7,468	14.8
NCA		
Male	12,590	11.1*
Female	4,896	9.7
NVCA		
Male	1,493	1.3*
Female	332	0.7

* $p < .001$

Predictive Utility by Race: FTA, NCA, and NVCA in Kentucky

We begin by assessing the strength of the associations between race and each of the pretrial outcomes. Table 13 presents the FTA scale scores and failure rates by race. In Kentucky, black defendants have significantly ($p < 0.001$) higher FTA rates for scores of 1 and 2. It appears that the rates of FTAs are potentially underestimated for black defendants at lower scores. The PSA shows

parity between races in outcome rates at higher levels of the scale, but is challenged in assessing truly low risk black defendants by assigning them a score at parity with the white outcome rate. These statistically significant differences do not exist for any of the higher FTA scores.

Table 13: FTA Rates by Race and FTA scores

	Black % FTA (n)	White % FTA (n)
1*	11.4 (406)	6.7 (1,623)
2*	11.8 (722)	9.3 (3,333)
3	14.3 (967)	13.8 (4,243)
4	19.7 (1,104)	20.0 (4,740)
5	26.1 (1,132)	26.5 (3,988)
6	31.0 (381)	32.4 (1,195)

*p < .001

Table 14 presents the associations between NCAs and the NCA risk scores for black and white defendants. The Kentucky sample reveals statistically significant differences between black and white defendants ($p < .001$) with scores of 2 and 3. The Kentucky sample is composed of 80 percent white defendants and the base rates for NCAs are nearly identical for white defendants (10.7 percent) and black defendants (11.1 percent). It is important to point out that the FTA and NCA scale scores are used together to inform release decisions (as they are used in a matrix format similar to sentencing guidelines grids).

Table 14: NCA Rates by Race and NCA scores

	Black % NCA	White % NCA
1	3.4 (68)	4.0 (746)
2*	5.8 (421)	7.1 (3,098)
3*	8.9 (620)	11.4 (3,843)
4	14.8 (1,033)	15.2 (3,708)
5	18.7 (558)	20.0 (1,946)
6	25.7 (384)	26.5 (935)

* p < .001

In table 15, we report the NVCA rates for the NVCA scale scores for white and black defendants. The NVCA scores are collapsed into the binary NVCA flag in with scores of 1-4 equal to no flag and scores of 5-6 equal to a violent flag. These bivariate analyses show that black defendants with low NVCA scores (without the flag) have a statistically significant higher rate of

NVCAs relative to white defendants with similar scores. The NVCA scores show a similar pattern to FTA, except extending discrepancies in outcome parity to low and medium risk defendants. Additionally, though the high levels of the scale show no significant differences, the pattern of higher outcome rates for blacks continues. Clearly, the prevalence and rate of a violent arrest during pretrial is highly unlikely, NVCA rates increase for both racial categories as the NVCA scores increase.

Table 15: NVCA Rates by Race and NVCA score

	Black % NVCA	White % NVCA
1*	0.9 (37)	0.4 (101)
2*	1.1 (107)	0.6 (345)
3*	1.8 (129)	1.1 (359)
4*	2.9 (123)	2.0 (287)
5	3.1 (58)	2.6 (153)
6	4.7 (36)	3.5 (76)

* p < .001

Although these descriptive statistics provide an important understanding about some of the underlying nuances of failures rates by racial groups, we use AUC ROC scores to assess the predictive utility of the risk scores for each racial group. Table 16 reports the ROC scores for each of the pretrial outcomes by race. The PSA is predicting within the fair to good range across race. In Kentucky, the PSA is a significantly stronger predictor for white defendants for FTAs ($p < 0.001$) than black defendants. The ROC (0.612) for black defendants falls within the fair range, whereas the ROC for white defendants (0.655) for white defendants falls within the good range (e.g., Desmarais et al., 2013). NCA and NVCA scores show some differences in ROC values by race, but these differences are not statistically significant.

Table 16: AUC Scores for FTA, NCA, and NVCA by Race

FTA		NCA		NVCA	
Kentucky					
Black	White	Black	White	Black	White
0.612	0.655	0.659	0.647	0.631	0.666
p-value:	<0.001*	p-value:	0.023	p-value:	0.015

* $p < 0.001$

Differential Prediction: Testing for Predictive Fairness by Race

The PSA ROC scores show that PSA has provide fair to good accuracy, with significant differences found between black and white defendants for FTAs. Assessing the strength or degree of utility for the each of the PSA scales by race is different than assessing the form or shape of the relationship between race and the PSA scores with pretrial outcomes (Arnold, 1982). The moderated regression approach uses four regression models in the following sequence. First, a model is estimated with only the subgroup of interest (i.e., race, gender). Second, a model is fit with only the test score (i.e., PSA score for each outcome separately). Third, a model estimates both the subgroup and the PSA score. Fourth, a final model includes the subgroup, PSA score, and the interaction of the subgroup and the PSA score. Building to the final full model allows for estimating the main effects of each variable separately before testing to see if the PSA by race interaction terms are significant. The interaction term tests to what extent the likelihood of a pretrial failure is a matter of how race and the PSA scores operate together – i.e., values of the PSA scores have different meanings (effect) on odds of failure for black and white defendants (or for male and female defendants).

Table 17 presents the odds ratios and confidence intervals for the four regression models for FTA. There are consistent significant results for race, FTA score, and the interaction term. These differences suggest that the association between a new FTA and a given score on the PSA are not the same for white and black defendants. This relationship can best be explained with figure 4 in which we plotted the predicted probabilities for an FTA by FTA score for white and black defendants from model 4. Figure 4 presents nonparallel lines and intersection of the race-specific lines with the predicted probabilities of FTA by race for each PSA score. The FTA score by race

interaction demonstrates that the effect of race is different at different levels of the PSA score for black and white defendants.

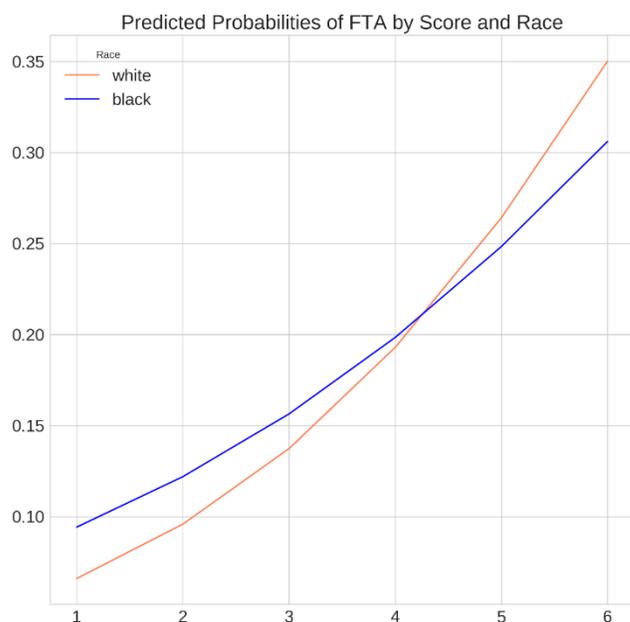
Table 17: Logistic Regression Models Testing Predictive Fairness of the PSA by Race for FTA

Kentucky								
	Model 1		Model 2		Model 3		Model 4	
	OR	CI	OR	CI	OR	CI	OR	CI
Race(white)	0.814*	0.768-0.863			0.915*	0.862-0.972	0.604*	0.513-0.071
FTA Score			1.471*	1.446-1.496	1.468*	1.443-1.493	1.335*	1.284-1.387
FTA Score*Race							1.125*	1.077-1.174
Constant	0.205*	0.195-0.217	0.051*	0.048-0.055	0.055*	0.051-0.06	0.07*	0.067-0.09
Model Pseudo R2	0.001		0.043		0.043		0.043	

* p<0.001

The odds ratios for the interaction term affirm differential prediction by race. The relationship between the FTA scores and FTAs are moderated by race. Figure 4 shows flatter slopes and higher intercepts for black defendants. Model 3 demonstrates general predictive utility of the FTA scale scores to suggest that for each 1-point increase in the scale there is a 47 percent increase in the odds of a defendant experiencing an FTA. These effects, however, are moderated by race such that in model 4 the intersecting race lines suggest that black defendants' FTA rates are underestimated (see race differences in table 13 that shows higher false negatives for black defendants), with a slight overestimation of white defendant failures at the higher scale scores (for a similar result in the testing literature, see Houston and Novick, 1987: 319).

Figure 4: Predicted Probabilities of FTA by FTA Score for White and Black Defendants



In table 18, we present results of similar regression models testing for differences in the form of the relationship between black and white defendants and NCA rates. Race is significant in models 3 and 4, but not in model 1, and the NCA score is significant and positive in all three models. The NCA by race interaction term is not significant and suggests equal slopes for white and black defendants on NCAs.

Table 18: Logistic Regression Models Testing Predictive Fairness of the PSA by Race for NCA

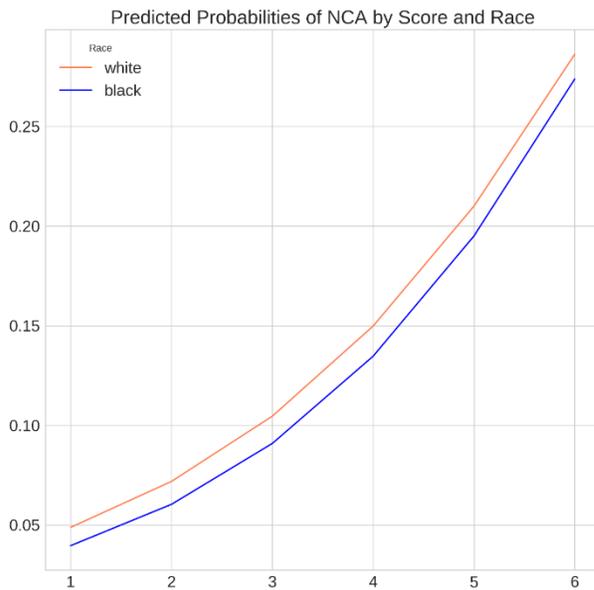
Kentucky								
	Model 1		Model 2		Model 3		Model 4	
	OR	CI	OR	CI	OR	CI	OR	CI
Race(white)	0.954	0.89-1.022			1.143*	1.065-1.228	1.283*	1.036-1.589
NCA Score			1.509*	1.478-1.54	1.517*	1.486-1.548	1.556*	1.481-1.636
NCA Score*Race							0.969	0.918-1.023
Constant	0.126*	0.118-0.134	0.033*	0.02-0.03	0.029*	0.026-0.032	0.027*	0.022-.032

Model Pseudo R ²	0.000		0.040		0.041		0.041	
-----------------------------	-------	--	-------	--	-------	--	-------	--

* p<0.001

The results from model 4 in table 18 are displayed in Figure 5 as predicted probabilities by race for each NCA scale score.

Figure 5: Predicted Probabilities of NCA by NCA Score for White and Black Defendants



white defendants have small but significantly larger odds of failure (model 3). This finding fits with the actual NCA failure rates (table 14) in which white defendants have significantly higher rates of NCA than black defendants for NCA scores of 2 and 3, and small insignificant higher rates for a score of 6. Nonetheless, the interaction term in table 18 confirms the equal slopes assumption seen in Figure 5. The predicted probabilities are rather similar in shape and there are small differences between the regression lines for white and black defendants.

In table 19, we report the four logistic regression models testing for race differences on the NVCA scale.²⁹ There are significant main effects for black defendants and for the NVCA scales to predict future violent arrests during pretrial. The interaction term, however, is insignificant and does not contribute to the model (no change in R²) finding that race is not moderating the relationship between the NVCA scales and new violent arrests.

Table 19: Logistic Regression Models Testing Predictive Fairness of the PSA by Race for NVCA

Kentucky								
	Model 1		Model 2		Model 3		Model 4	
	OR	CI	OR	CI	OR	CI	OR	CI
Race(white)	0.554*	0.465-0.66			0.636*	0.533-0.76	0.435*	0.274-0.689
NVCA Score			1.578*	1.53-1.64	1.555*	1.468-1.647	1.431*	1.281-1.598
NVCA Score*Race							1.121	0.985-1.275
Constant	0.018*	0.016-0.021	0.003*	0.003-0.004	0.005*	0.004-0.006	0.006*	0.004-0.009
Model Pseudo R ²	0.006		0.032		0.036		0.036	

* p<0.001

Figure 6 is used to demonstrate the relationships reported in table 19. We provide the predicted probabilities of failure plotted by their NVCA scale score for black and white defendants. These graphs demonstrate that, although there are differences by race that contribute to different intercepts, the slopes (forms) are similar.

²⁹ We do not fully address in this paper, is that, although logistic regression is not as susceptible to problems stemming from unbalanced groups sizes as is linear regression, but estimation difficulties do arise in cases of rare events due to the relative overabundance of zeros (no failure) relative to 1 (failure). This is likely more of an issue for the NVCA regression models because the base rates are so low. King and Zeng (2001) provide a critique of estimating logistic regression models with rare events and suggest that, despite large sample sizes, when event occurrence is lower than 5 percent there could be instability in the models.

Figure 6: Predicted Probabilities of NVCA by NVCA Score for White and Black Defendants

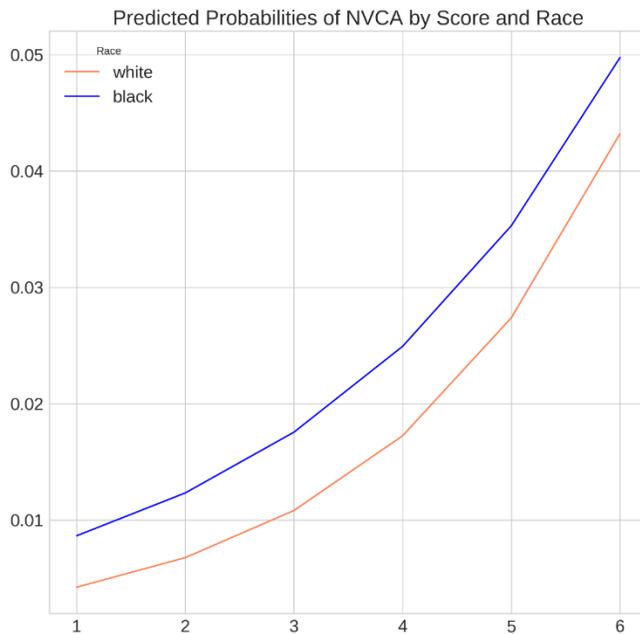


Figure 6 demonstrates that there are higher predicted probabilities for black defendants compared to white defendants in Kentucky. At each point on the NVCA scale score, black defendants have higher predicted odds of being arrested for a new violent crime relative to white defendants (OR = 0.60, 40 percent lower odds for white defendants). These differences, however, are not resulting from a moderating effect, with an insignificant interaction term, reduction in the race odds ratio (model 4), and little movement in model fit.

Predictive Utility by Gender

In this section of the paper, we report the same analyses used above to assess the extent to which the PSA scores vary in strength and form of the associations with FTA, NCA, and NVCA for gender. Tables 20 and 21 provide the FTA and NCA rates by gender. The FTA rates increase as the FTA scores increase, with no statistically significant differences in FTAs between males and females.

Table 20: FTA Rates by Gender and PSA Score

	Female % FTA	Male % FTA
1	7.1 (771)	7.7 (1,395)
2	10.1 (1,259)	9.5 (2,892)
3	13.9 (1,481)	13.9 (3,802)
4	20.2 (1,809)	19.7 (4,083)
5	26.8 (1,650)	26.4 (3,509)
6	34.4 (498)	31.1 (1,086)

* p<0.001

Table 21 demonstrates that the NCA rates also increase with each 1-point increase in the NCA scores for both males and females. There are no significant differences in new arrest rates between males and females, and with each 1-point increase in the NCA scale score there are increases in the NCA rates.

Table 21: NCA Rates by Gender and Score

	Kentucky Female % NCA	Kentucky Male % NCA
1	4.3 (365)	3.6 (464)
2	7.1 (1,299)	6.7 (2,266)
3	10.9 (1,395)	10.9 (3,097)
4	14.8 (1,102)	15.2 (3,664)
5	19.7 (546)	19.7 (1,966)
6	24.5 (189)	26.6 (1,133)

*p < .001

In table 22, we present the NVCA failure rates for male and female defendants by their NVCA scale scores. The NVCA rates for male and female defendants generally increase with each 1-point increase in the NVCA scale. There is one exception to this pattern in which there is a slightly higher rate of NVCA's for women with an NVCA score of 5 (2.4) compared to scores of 6 (2.2). The sample sizes for each NVCA score are rather small (e.g., n = 8), and both of these categories would be included in the NVCA violent flag. Females with NVCA scale scores of 4 points and higher exceed the overall NVCA base rate (overall base rates of 1.1 in Kentucky, see table 1). The same is true for male defendants, but at scale score of 3.

Table 22: NVCA Rates by Gender and Score

	Female % NVCA	Male % NVCA
1	0.3 (37)	0.5 (103)
2*	0.5 (103)	0.8 (354)
3*	0.7 (80)	1.5 (411)
4	1.6 (68)	2.3 (344)
5	2.4 (36)	2.8 (177)
6	2.2 (8)	4.0 (104)

*p < .001

Table 23 reports the AUC ROC scores for each of the pretrial outcomes by gender. The PSA is predicting within the fair to good range by gender. The ROCs do not differ much from those estimated for the overall samples or the race estimates. The ROCs remain in the fair to good range.

Table 23: AUC Scores for FTA, NCA, and NVCA by Gender

FTA		NCA		NVCA	
Male	Female	Male	Female	Male	Female
0.642	0.655	0.653	0.637	0.654	0.657
p-value: 0.016		p-value: <0.001*		p-value: 0.898	

* p<0.001

There appears to be little evidence of differences in predictive utility between male and female defendants for FTA and NVCA, but the NCA model has a statistically significant ($p < 0.001$) higher validity for men (ROC = 0.653) than women (ROC = 0.637). Next, we estimate four logistic regression models testing for interaction effects for gender and each of the PSA scale scores with each of the pretrial outcomes. We follow the same procedures used above for the tests by race.

Differential Prediction: Testing for Predictive Fairness by Gender

In this section of the paper, we provide the regression models in the same order as were presented above for race. In tables 24 through 26 and figures 6 through 9, we present the findings for differential prediction using a gender by PSA scale score for each interaction term to determine if gender moderate the effect of the PSA score. These analyses demonstrate, for the most part, that

males are predicted to have higher rates of failures across the three outcomes. However, none of the interaction terms reach statistical significance to suggest that the main effect of gender is moderating the effect of the PSA scores to predict failures. Figures 6 through 9 provide graphical evidence to support the tables.

Table 24 shows that gender does not have a main effect. Figure 7 demonstrates parallel slopes for FTA for male and female defendants. These models show that 1-point increases in the FTA scale score are associated with a 46 to 48 percent greater odds of a defendant having an FTA.

Table 24: Logistic Regression Models Testing Predictive Fairness of the PSA by Gender for FTA

	Model 1		Model 2		Model 3		Model 4	
	OR	CI	OR	CI	OR	CI	OR	CI
Gender(male)	1.002	0.954-1.053			0.975	0.928-1.026	1.038	0.91-1.184
FTA Score			1.465*	1.441-1.49	1.465*	1.441-1.49	1.483*	1.441-1.528
FTA Score* Gender							0.982	0.947-1.1.018
Constant	0.173*	0.166-0.181	0.052*	0.049-0.056	0.053*	0.049-0.057	0.051*	0.046-.057
Model Pseudo R ²	0.000		0.042		0.042		0.042	

*p < 0.001

Figure 7: Predicted Probabilities of FTA by FTA Score for Male and Female Defendants

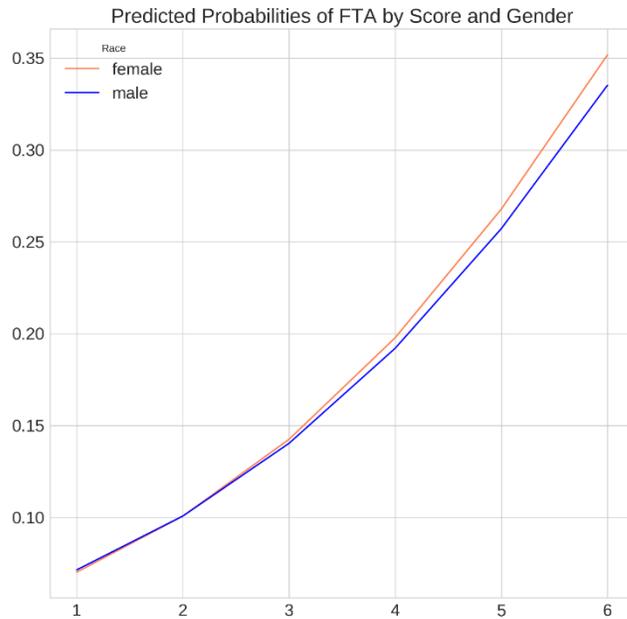


Table 25 shows gender main effects (OR = 1.16) for NCAs, but this effect does not remain when the other variables are entered in the model. There are significant ($p < 0.001$) improvements in prediction with the NCA scores, and no interaction effects.

Table 25: Logistic Regression Models Testing Predictive Fairness of the PSA by Gender for NCA

Kentucky								
	Model 1		Model 2		Model 3		Model 4	
	OR	CI	OR	CI	OR	CI	OR	CI
Gender(male)	1.166*	1.1-1.236			0.985	0.928-1.046	0.928	0.791-1.089
NCA Score			1.515*	1.484-1.547	1.516*	1.486-1.548	1.496*	1.438-1.556
NCA Score* Gender							1.019	0.972-1.067
Constant	0.107*	0.102-0.113	0.033*	0.03-0.035	0.033*	0.030-0.036	0.034*	0.030-.039
Model Pseudo R ²	0.000		0.041		0.041		0.042	

* $p < 0.001$

Figure 5 shows no differences in the predicted probabilities of an NCA for male and female defendants, and a linear trend with increasing NCA scores and NCAs.

Figure 8: Predicted Probabilities of NCA by NCA Score for Male and Female Defendants

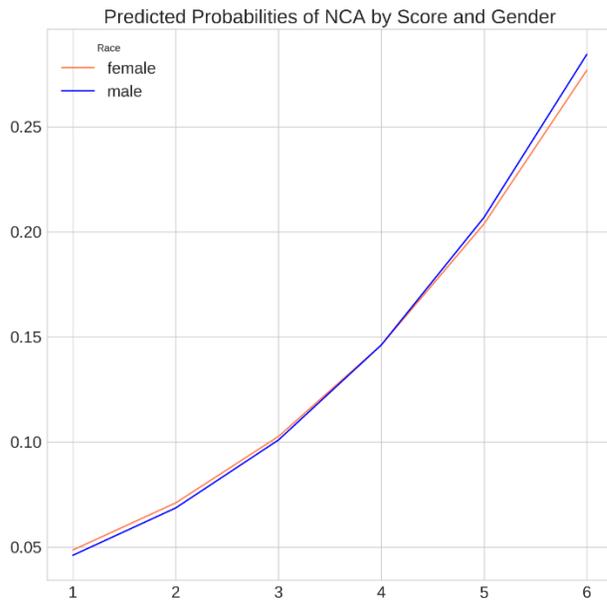


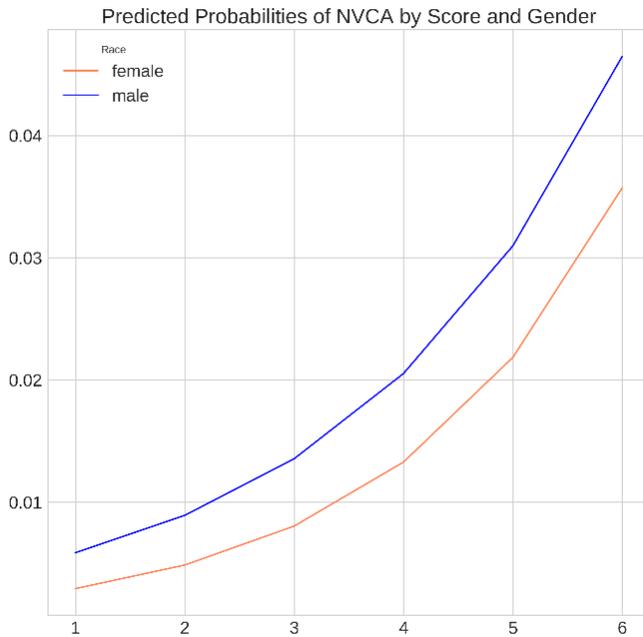
Table 26 shows results for NVCA for male and female defendants. There are significant main effects for violent arrests, with men have about twice the odds of an NVCA during pretrial compared to women. The NVCA score has significant main effects (see Model 2, OR 1.58), with a 1-point score associated with roughly a 58 percent higher odds of violent arrest. The interaction term shows that gender does not moderate the NVCA scores with new violent arrests. Figure 9 supports the findings these findings.

Table 26: Logistic Regression Models Testing Predictive Fairness of the PSA by Gender for NVCA

Kentucky								
	Model 1		Model 2		Model 3		Model 4	
	OR	CI	OR	CI	OR	CI	OR	CI
Gender(male)	2.02*	1.654-2.469			1.69*	1.387 - 2.079	2.17*	1.311-3.615
NVCA Score			1.58*	1.495-1.674	1.54*	1.459 - 1.636	1.65*	1.439-1.910
NVCA Score* Gender							0.919	0.797-1.074
Constant	0.007*	0.006-0.008	0.003*	0.003-0.004	0.002*	0.002 - 0.003	0.002*	0.001-.003
Model Pseudo R ²	0.007		0.03		0.04		0.04	

*p < 0.001

Figure 9: Predicted Probabilities of NVCA by NVCA Score for Male and Female Defendants



Discussion

Researchers often overlook the pretrial process despite the consequences for defendants, their families, and local criminal justice systems. Most of the recent research and commentary about risk assessments have focused on their use at sentencing, parole release (e.g., Harcourt, 2010; Starr, 2014; Tonry, 2014) and post-conviction (e.g., Skeem and Lowenkamp, 2016). Pretrial risk assessment was brought to light by ProPublica, igniting an intense debate about the potential for racial bias exacerbated by risk assessments. They claimed that different error rates between white and black defendants (e.g., more low risk black defendants were scored high risk) amounted to racial bias. Flores et al., (2016) responded to ProPublica by analyzing the same data to show that the COMPAS met statistical standards of calibration (i.e., a score of X had the same meaning for black and white defendants). Adding to this debate, several mathematicians, statisticians, and data scientists (e.g., Chouldecova, 2016; Corbett-Davies et al, 2017; Kleinberg et al, 2016) demonstrated that it is mathematically impossible to satisfy error rate balance and calibration across subgroups when the base rates differ, establishing that the debate was more about competing mathematical formulations of fairness than evidence or exoneration of bias. Although this debate is essential to move criminal justice risk assessments forward, these studies reveal little about pretrial risk assessments, since the authors did not study pretrial outcomes (e.g., FTA or NCA while awaiting trial) on a pretrial population.

The current paper is the first validation of the PSA and it includes the first tests of predictive validity and predictive bias by race and gender. The analyses are conducted using a pretrial release population and pretrial outcomes. The PSA has been adopted by dozens of jurisdictions and it is contributing to thousands of release decisions each day. The PSA has been well-received by stakeholders, believing its efficacy as a tool to speed arraignments, release people more quickly, and reduce jail populations. Further, the PSA has received much attention from journalists interested in

the use of pretrial risk assessments, but, again, there has yet to be a full study assessing accuracy and group based differences.

Predictive Validity: The PSA meets Validity Standards

Our study was motivated by three research questions. First, we assessed the support for overall predictive validity of the PSA. We found that the PSA meets standards for criminal justice risk assessments. First, bivariate statistics show that higher PSA scores are associated with higher outcome rates, these associations have small positive correlations for FTA ($r = 0.188$) and NCA ($r = 0.171$), and weak correlations for NVCA score ($r = 0.067$) and the NVCA violent flag ($r = 0.048$). Of the three PSA models, the NVCA model (ROC = 0.664) is the most accurate, with little difference between the FTA scale (ROC = 0.646) and the NCA scale (ROC = 0.650). All the AUC ROCs are within what Desmarais et al. (2016) defined as *good* based upon their review of risk assessments used to make criminal justice decisions. These findings indicate that when drawing two random cases from the dataset, one of which had the pretrial outcome and the other did not, between 64 and 66 percent of the time the case with the pretrial outcome would have a higher score than the successful case.

Differential Validity: FTA differences by Race and NCA difference by Gender

The second research question assesses whether the predictive accuracy of the PSA varies by race or gender. Considering predictive accuracy by race, we found the FTA scale to be significantly more predictive for white defendants (ROC = 0.655) than black defendants (ROC = 0.612). This is a large and significant difference ($p < 0.001$) demonstrating that the FTA scale is a fair predictor of pretrial success for black defendants but shows good performance for white defendants. Reviewing the public safety outcomes, we do not find a significant difference ($p = 0.023$) in the NCA scale to predict outcomes for black or white defendants, with the scale being slightly more accurate for black defendants. Conversely, the NVCA scale is more accurate at predicting violent arrests for white

defendants (ROC = 0.666) than black defendants (ROC = 0.631). Although the difference in NVCA accuracy is large, it is not significant ($p = 0.015$). Given these findings, the PSA does a significantly better job of predicting FTAs for white defendants.³⁰

Although racial bias related to the use of risk assessments has garnered wide spread attention, few researchers are studying the potential for predictive bias by gender. More concerning is that because risk assessments aggregate and average information about predicted probabilities of failure, the higher base rates for males are likely to drive up the expected failure rate for females (Skeem et al., 2017). Essentially, there is the potential for male defendants' characteristics to make female defendants appear riskier with traditional recidivism studies with longer follow-up periods. With the Kentucky pretrial dataset, this hypothesis is unlikely for FTAs since male and female defendants have the same base rate (F_TA = 14.8%). The F_TA scale's predictive accuracy did not significantly differ ($p = 0.016$) between male (ROC = 0.642) and female (ROC = 0.655) defendants, and the F_TA scale was a bit more predictive for females. Male defendants do have significantly higher NCA and NVCA base rates compared to female defendants, and there are significant differences ($p < 0.001$) in the predictive accuracy for NCAs. There is ample research demonstrating that men have greater criminal propensity, lengthier criminal histories, and are involved in far more violent crimes than women. Despite male defendants having nearly double the NVCA rate, the NVCA validity measures are nearly the same ($\delta=0.003$, $p = 0.898$) across genders.

Differential Prediction: Intercept differences by Race

The final research question is focused on whether there is predictive bias by race and gender with the PSA. To answer this question, we used a moderator regression modeling (e.g., Cleary, 1968) approach that is commonly used to test for race and gender bias for several cognitive (e.g., ACT,

³⁰ There were differences in correlation coefficients between for FTAs and NCAs by race. The F_TA $r = 0.15$ and $r = 0.19$ and NCA $r = 0.18$ and $r = 0.16$, black and white defendants, respectively. There were not differences for NVCA, $r = 0.06$ for black and white defendants.

GRE) and employment tests, and supported by the Standards for Educational and Psychological Testing (Standards) (AERA, American Psychological Association, & NCME, 2014). The regression approach tests for intercept and slope differences, and is well-established for testing bias in psychometric scales (e.g., United States v. City of Erie, 2005). This approach was most recently introduced to the criminological literature in three research papers testing for racial bias (Skeem and Lowenkamp, 2016), age bias (Monahan et al., 2017), and gender bias (Skeem et al., 2016) in the PCRA. We build on that work and follow the definition for predictive bias issued by the Standards (2003: 23) as “Slope and/or intercept differences between subgroups indicate predictive bias.”

We first address the findings related to predictive bias by race. The results show there is some level of predictive bias across all three outcomes. Specifically, there are intercept differences for FTAs, NCAs, and NVCAs, and slope differences for FTAs. These findings suggest that an average PSA score of X is not associated with an average FTA, NCA, or NVCA rate of Y for black and white defendants. Although the intercept differences are indicative of an incremental increase in outcomes by race after controlling for the influence of the PSA, these are less of a concern than finding there is a difference in the slopes. The slope differences for FTAs suggest that the FTA scores are moderated by race. The FTA model in Kentucky has both intercept differences (table 17 comparing model 2 to model 3)³¹ and slope differences (comparing model 4 to model 4). These effects are most clearly seen in figure 4 in which the plotted predicted probabilities of an FTA by FTA score for each race, and the lines cross one another around an FTA score of 4. These extreme differences in the form of the slopes suggests serious shortcomings for the FTA model to predict FTAs by race. The models show that white defendants initially have a lower predicted likelihood of an FTA (OR = 0.915, model 3), but due to the varying slopes, whites have a higher predicted

³¹ We assessed the intercept differences using a likelihood ratio test of the differences between model 2 and 3, and slope differences were assessed testing differences between model 3 and 4. All differences were assessed using $p < 0.001$.

probability than blacks for high scores on the scale (5-6). Overall, black defendants have higher mean FTA scores (3.17) than white defendants (2.38), and black defendants have higher FTA base rates. Taking this information, along with the differences in evidence for predictive validity, provides evidence that the FTA scores do not have the same meaning for white defendants and black defendants. It is possible that some of these issues reflect limitations of the Kentucky data (addressed below).

Although court appearance is a central concern for pretrial decisions, the PSA also provides scores to classify defendants according to likelihood of a new arrest. The pretrial window is usually no longer than 6 to 9 months, so defendants have a brief opportunity to commit a new crime. The two public safety outcome measures are any new arrest and any new violent arrest. The PSA provides more parity by race for the public safety outcomes than FTAs, but in both cases, we found intercept differences. For NCAs, we found that white defendants have a 14% larger odds of a new arrest than black defendants. This is interesting because base rates were slightly higher for black defendants (11.1) than for white defendants (10.7, ns). There is a strong relationship with the PSA scores and new crimes (OR = 1.517, $p < 0.001$), but the regression results are in line with what was found in the descriptive statistics in which white defendants have a consistently higher failure rate within each of the NCA scores ($p < 0.001$ for score of 2 and 3). Simply, black defendants are arrested for a new crime during pretrial at a lower rate than white defendants within the same NCA score. The intercept differences show that race adds incremental utility to the NCA scales to predict new crimes (i.e., after controlling for the NCA score, race still has an effect).

The second public safety outcome is what is viewed as the most serious – arrest for a violent crime. We find a similar pattern with NVCA as we did with NCAs as there are intercept differences, but no slope differences. The influence of race on NVCA scores to predict new violent arrests, however, shows that black defendants have a larger intercept. We find that race adds incremental

utility to the NVCA scores, with the relationship between NVCA scores and a violent arrest higher for black defendants than white defendants. The NVCA scales range from 1 to 6, but these scores are collapsed such that NVCA scores of 5 and 6 are combined into a violent flag to signal to judges that a defendant has a high probability of failure. Although we found that black defendants have higher rates of NVCA within each NVCA score, these differences become smaller (and insignificant) for score of 5 and 6 (i.e., suggesting the higher risk classifications are more accurate). It is difficult to draw concrete conclusions from these results due to the small sample sizes within cells, as there were 94 black and 229 white defendants. This is clearly an area needing further research to understand the racial patterning of violent arrests during pretrial.

Differential Prediction: Little Gender Differences

The final part of our analysis assessed for predictive bias by gender. We found the PSA to be free of predictive bias for FTAs and NCAs. But, we did find intercept differences showing that gender provides incremental utility to the relationship between the NVCA scores and a violent arrest. It is difficult to make too much of these findings, however, because the cell sizes for the higher NVCA scores are so small. There are 44 female defendants and 281 male defendants with an NVCA score of 5 or 6. The intercept differences in NVCA scales by gender are not entirely unsurprising, and they are in line with Skeem et al.'s (2016: 591) findings for post-conviction arrests in which they cautioned that the PCRA could discriminate against women because women that have what are considered high risk scores “do not present the same...risk of recidivism as do men who score within the same range on the instrument.” They suggested that the imbalance in the PCRA could be fixed with interpreting the “...scores in a gender-specific manner” (592). Although the authors did not elaborate on what is meant by a gender-specific interpretation, on its face, this advice seems fraught with potential problems as stakeholders are now being asked to conduct informal assessments when interpreting the risk assessment scores. Such advice seems antithetical to

the intention of risk assessments, and it would seem risk assessment developers need to further refine forecasting models, not inject more subjectivity. If practitioners are to make gender-specific interpretations, are they to make race, age, class, or other subgroup-specific interpretations?

Limitations and Future Research

These findings should be interpreted with an understanding of the limitations and weaknesses of our data and design. The analyses are based on one statewide pretrial release population. Kentucky is relatively unique in ways that may weaken generalization and external validity. Kentucky is a small rural state with about 40 percent lower black population than the nation (8% vs. 13%), and relatedly a smaller proportion of the sample were black than what is typically found in criminal justice research. These limitations do not nullify the importance of our results, but rather they serve to highlight the need for ongoing research about the drivers of FTAs and the patterns of new arrests and violent arrests during the pretrial phase, as such patterns could vary from what is found in longer follow-ups (e.g., what was used in the ProPublica article).

Further, the PSA is used in dozens of jurisdictions, so our results only provide information about the instrument's performance in Kentucky. Research is needed throughout the jurisdictions using the PSA to assess predictive validity, differential predictive validity, and prediction bias. The pretrial space is ripe for additional sociological and criminological research to understand not only the individual, familial, and system impacts of assessments, but also to understand the distinct patterning of behaviors during pretrial. For instance, how do mental health and substance abuse issues fuel FTAs? How are domestic violence charges associated with new violent arrests? This study, similar to pretrial research, only focuses on those released, but more research is needed to understand the composition of the detained population. And, importantly, more research is needed to understand how risk assessments are implemented, understood, and used by decision makers.

Conclusion

We found the PSA to have predictive validity in line with risk assessments used throughout the criminal justice system. The PSA scales are associated with increasing failure rates, something found across racial and gender subgroups. There are issues with predictive bias by race, but race does not moderate the relationship between the NCA or NVCA scales and new arrests. The FTA scale demonstrated both intercept and slope differences, indicating race moderates the relationship between scores and outcomes. We found little indication of predictive bias for FTAs or NCAs by gender, but there are differences for NVCAs.

To contextualize the current debate about risk assessments it is helpful to recall that throughout the 1990s and 2000s, criminologists were highly critical of risk assessments. Feely and Simon (1992) coined the term the *new penology* to refer to a new paradigm in punishment focused on probabilities, groups, and risk reduction. This new form of actuarial justice (Simon, 1994) replaced Enlightenment ideals of individual justice focused on moral culpability, individual blameworthiness, deterrence, and rehabilitation. Even the term mass incarceration was initially a critical term expanding insights from the new penology to demonstrate the dubious utility of actuarial justice for penal growth (Garland, 2001). These critiques of actuarial justice centered on two main issues. First, risk assessments and other actuarial techniques were necessary to create mass incarceration. Simply, mass incarceration was not possible without these technocratic instruments that enabled practitioners to screen, triage, and process more individuals in shorter amounts of time. Second, actuarial logic trained the focus of criminal justice policies, practices and stakeholders on the most vulnerable groups in society, which mostly includes people of color, but included other deviant groups (e.g., drug addicts, homeless, immigrants).

The current round of controversy about risk assessments has changed in ways that reflect greater awareness of social justice, law, and technological advances. The new wave of critiques come

from two primary camps. In one camp are legal scholars arguing there are legal and constitutional grounds (e.g., transparency, equal protection clause, disparate impact) to challenge risk assessments. The COMPAS, for example, is a proprietary instrument that is hidden from public view. Northpointe (now, Equivant), the company that owns the COMPAS, refuses to share the factors, weights, scaling, and detailed validations of the COMPAS. This lack of transparency is problematic. How are defendants expected to defend themselves when they are unaware of how their risks are defined? Central to these legal and constitutional challenges is that risk assessments do not include factors that are potentially correlated with hyper-policing of communities of color, higher rates of prosecution and sentencing, and cumulative disadvantage as people of color move through the sentencing process.

The Laura and John Arnold Foundation developed the PSA to speed arraignment, improve identification of low v. high risk defendants, and decrease pretrial incarceration. The PSA is a short instrument comprised of criminal justice related factors (e.g., criminal history, current offense violent) and young age. The PSA responds to legal critiques and judicial needs to understand failure to appear and public safety by providing separate scales for each outcome. These outcomes have different meanings for the justice system, the community, and stakeholders. The legal system has struggled with rooting out mistreatment of people based on class, race, and other statuses. The PSA does not include direct measures of ascribed status related to race, class, or gender, and the PSA does not include arrests or charges as risk factors. Rather, PSA developers recognized the potential for cumulative disadvantage as one moves through the system, and include prior convictions. Of course, removing ascribed statuses and focusing on convictions does not necessarily create a tool that is free of predictive bias, but it is an improvement over previous risk factors.

Interestingly, the controversy over risk assessments seems to be from groups wanting similar outcomes and goals. That is, risk assessments are being developed with the hopes of reducing jail

and prison populations and decreasing racial and ethnic disparities. Risk assessments are believed to remove conscious and unconscious forms of human bias, and provide a system to treat people fairly, regardless of any ascribed status. However, there is nothing inherent in risk assessments that will reduce jail populations, make prison populations less racially disparate, or otherwise reform the criminal justice system. Risk assessments are, essentially, probabilistic models, and, as such, they do not provide the correct answer in 100 percent of the cases. Instead, the well-known line from George Box is instructive: “all models are wrong, but some are useful.” Risk assessments should be seen less to unwind mass incarceration, and more as a decision-making tool. After all, much as James Forman (2017) depicted in his historical accurate, mass incarceration did not emerge spontaneously due to a single individual, group, or social issue. Rather, mass incarceration was assembled piecemeal over a forty-year period.” The overreliance on incarceration spreads throughout the fragmented agencies that make up our (non)system of criminal justice (Freed, 1969) as the police arrest, prosecutors charge, judges detain people pretrial and impose more sentences, and corrections officers incarcerate. Although risk assessments have been used for over a century, we are only now beginning to take the critiques seriously, which hopefully will lead to the development and implementation of more useful instruments.

References

- American Bar Association. (2002). *American Bar Association Criminal Justice Standards on pretrial Release. (3rd Ed.)*. Washington, D.C.: American Bar Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA Publications.
- Angwin, Larson, Mattu, and Kirchner, 2016
- Ares, C., Rankin, A., and Sturz, H. (1963). The Manhattan Bail Project: An interim report on the pre-trial use of pre-trial parole. *New York University Law Review*, 38: 67- 95.
- Arnold, H. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior and Human Performance*, 29, 143–174.
- Arnold, L. and Arnold, J. (2014). *Results from the first six months of the Public Safety Assessment – Court in Kentucky*. Laura and John Arnold Foundation, July.
- Arnold, J., & Arnold, L. (2015, March). Fixing justice in America. *Politico Magazine*. Retrieved from <http://www.politico.com/magazine/story/2015/03/criminal-justice-reform-coalition-for-public-safety-116057.html>
- Austin, J., Ocker, R., and Bhati, A. (2010). Kentucky pretrial risk assessment instrument validation. The JFA Institute. <https://www.pretrial.org/download/risk-assessment/2010%20KY%20Risk%20Assessment%20Study%20JFA.pdf>
- Bechtel, K., Lowenkamp, C., and Holsinger, A. (2011). Identifying the Predictors of Pretrial Failure: A Meta-Analysis. *Federal Probation*, 75 (2).
- Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. Working Paper found on January 5, 2018 here: <https://arxiv.org/pdf/1703.09207.pdf>
- Burgess, E.W. (1928). Factors determining success or failure on parole. In A. A. Bruce, E. W. Burgess, & A. J. Harno (Eds.), *The workings of the indeterminate sentence law and the parole system in Illinois* (pp. 221-234). Springfield, IL: State Board of Parole.
- Carson, Elizabeth. 2015. Prisoners in 2014. *Washington, DC: Bureau of Justice Statistics*. <http://www.bjs.gov/index.cfm?ty=pbdetail&iid=5387>.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115–124.
- Cohen, T. H., and Reaves, B. A. (2007). *State court processing statistics, 1990-2004: Pretrial release of felony defendants in state courts*. Washington, DC: U.S. Department of Justice.

- Danner, M., VanNostrand, M., and Spruance, L. (2016). *Race and gender neutral pretrial risk assessment, release recommendations, and supervision: VPRAI and PRAXIS Revised*. Luminosity.
- Demuth, S. (2003). Racial and Ethnic Differences in Pretrial Release Decisions and Outcomes: A Comparison of Hispanic, Black, and White Felony Arrestees. 41(3) *Criminology* 873.
- Desmarais, S. L., and Singh, J P. (2013). *Risk assessment instruments validated and implemented in correctional settings in the United States*. Lexington, Kentucky: Council of State Governments.
- Desmarais, Sarah, Kiersten Johnson, and Jay Singh. 2016. Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services*.
- Dewan, S. (2015). Judges replacing conjecture with formula for bail. *New York Times*, July 26.
- Dobbie, Will, Jacob Goldin, and Crystal Yang, "The effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," NBER Working Paper July 2016.
- Flores, A., Bechtel, K., and Lowenkamp, C. (2016). False positives, false negatives, and false analyses: A rejoinders to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *Federal Probation*, 80(2): 38-46.
- Foote, C. (1954). Compelling appearance in court: Administration of bail in Philadelphia. *University of Pennsylvania Law Review*, 102, 1031-1079.
- Forman, J. (2017). *Locking up our own: Crime and punishment in black America*. New York: Farrar, Straus, and Giroux.
- Frase, R., Roberts, J., Hester, R., and Mitchell, K. (2015). *Criminal history enhancements sourcebook*. University of Minnesota Law School, Robina Institute of Criminal Law and Criminal Justice.
- Freed, D. (1969). The Nonsystem of Criminal Justice. In *Law and Order Reconsidered: Report of the Task Force eds. James J. Campbell. Joseph R. Sahid. and Daniel P. Stang.* (Washington. D.C.: U. S. Court Printing Office) p. 265.
- Gottfredson, Michael R., and Don Gottfredson. 1988. *Decision Making in Criminal Justice: Toward the Rational Exercise of Discretion*, 2nd ed. New York: Plenum Press.
- Gottfredson, Stephen D., and G. Roger Jarjoura. 1996. Race, gender, and guidelines based decision making. *Journal of Research in Crime and Delinquency* 33:49–69.
- Gottfredson, S.D., & Moriarty, L.J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency*, 52, 178-200.
- Gupta, Arpit, Christopher Hansman, and Ethan Frenchman. The Heavy Costs of High Bail: Evidence from Judge Randomization," Working Paper August 2016.

- Harcourt, Bernard. 2008. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago, IL: University of Chicago Press.
- Harcourt, Bernard. 2015. Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter* 27:237–43.
- Heaton, Paul, Sandra Mayson, and Megan Stevenson, "The Downstream Consequences of Misdemeanor Pretrial Detention," Working Paper July 2016.
- Houston, W. and Norvick, M. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement*, 24: 4: 309-320.
- Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., and Robinson, C. R. (2011). The construction and validation of the Federal Post Conviction Risk Assessment (PCRA). *Federal Probation*, 75, 16–29.
- King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9: 137-163.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016, September 19). *Inherent Trade-Offs in the Fair Determination of Risk Scores*. *arXiv [cs.LG]*. Retrieved from <http://arxiv.org/abs/1609.05807>
- Kleiman, Matthew, Brian Ostrom, and Fred Cheesman. 2007. Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. *Crime & Delinquency*, 53:106–32.
- Lin, M., Lucas, H. C., Jr., and Shmueli, G. (2013). Research commentary—Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24, 906 –917.
- Lowenkamp, C.T., Lemke, R., and Latessa, E. (2008). The development and validation of a pretrial screening tool. *Federal probation*, 72, 2-9.
- Lowenkamp, C. T., Holsinger, A. M., & Cohen, T. H. (2015). PCRA revisited: Testing the validity of the federal Post Conviction Risk Assessment (PCRA). *Psychological Services*, 12, 149–157. <http://dx.doi.org/10.1037/ser0000024>
- Lowenkamp, C. T., and Whetzel, J. (2009). The development of an actuarial risk assessment instrument for U.S. Pretrial Services. *Federal Probation*, 73(3), 33-36.
- Lowenkamp, C., VanNostrand, M., and Holsinger, A., (2013). *Investigating the Impact of Pretrial Detention on Sentencing Outcomes*, Laura and John Arnold Foundation.
- Mahoney, B., Beaudin, B.D., Carver, J.A., Ryan, D.B., & Hoffman, R.B. (2001). *Pretrial services programs: Responsibilities and potential* (NIJ Issues and Practices). Washington, D.C: United States Department of Justice.
- Mamalian, C. (2011). *State of the science of pretrial risk assessment*. Pretrial Justice Institute.
- Minton, T. and Zeng, Z. (2015). *Jail inmates at midyear 2014*. Bureau of Justice Statistics

Monahan, John, and Jennifer L. Skeem. 2014. Risk redux: The resurgence of risk assessment in criminal sentencing. *Federal Sentencing Reporter* 26:158–66.

Monahan, John, and Jennifer L. Skeem. 2016. Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology* 12:489–513.

Olver, Mark, Keira Stockdale, and J. Stephen Wormith. 2009. Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice and Behavior* 36:329–53.

Piquero, A. R., and Brame, R. W. (2008). Assessing the race-crime and ethnicity-crime relationship in a sample of serious adolescent delinquents. *Crime and Delinquency*, 54, 390–422. <http://dx.doi.org/10.1177/0011128707307219>

Rice, M. E., and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*, 29, 615–620. <http://dx.doi.org/10.1007/s10979-005-6832-7>

Sackett, P. R., Borneman, M. J., and Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, 63, 215–227. <http://dx.doi.org/10.1037/0003-066X.63.4.215>

Sackett, P. R., and Bobko, P. (2010). Conceptual and technical issues in conducting and interpreting differential prediction analyses. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 3, 213–217. <http://dx.doi.org/10.1111/j.1754-9434.2010.01226.x>

Sacks, M. and Ackerman, A. (2012). Bail and sentencing: Does pretrial detention lead to harsher punishment. *Criminal Justice Policy Review*, 25(1): 59-77.

Siddiqi, Q. (2005). *An evaluation of the new pretrial Release Recommendation System in New York City: phase ii of the post-implementation Research*. New York, NY: New York City Criminal Justice Agency.

Singh, Jay, and Seena Fazel. 2010. Forensic risk assessment: A metareview. *Criminal Justice and Behavior* 37:965–88.

Skeem, J. L., Scott, E., & Mulvey, E. P. (2014). Justice policy reform for high-risk juveniles: Using science to achieve large-scale crime reduction. *Annual Review of Clinical Psychology*, 10, 709–739.

Skeem, J. L., and Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54, 680–712.

Skeem, J., Monahan, J., and Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and Human Behavior*, 40, 580–593.

Society for Industrial and Organizational Psychology. (2003). Principles for the validation and use of personnel selection procedures, 4th ed. Retrieved from <http://www.siop.org/principles/principles.pdf>

- Spohn, C. and Holleran, D. (2000). The imprisonment penalty paid by young, unemployed black and Hispanic male offenders. *Criminology*, 38, 281-306.
- Starr, Sonja. 2014. Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review* 66:803–72.
- Starr, Sonja. 2015. The new profiling: Why punishing based on poverty and identity is unconstitutional and wrong. *Federal Sentencing Reporter* 27:229–36.
- Stevenson, M. (2017). Assessing risk assessment in action.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240(4857), 1285–1293.
- VanNostrand, M. (2003). *Assessing Risk among pretrial Defendants in Virginia: the Virginia pretrial Risk assessment instrument*. Virginia Department of Criminal Justice Services.
- Tonry, M. (2014). Legal and ethical issues in the prediction of recidivism. *Federal Sentencing Reporter*, 26(3): 167-176.
- VanNostrand, M. & Keebler, G. (2009). *Pretrial risk assessment in the Federal Court*. Washington D.C.: U.S. Department of Justice.
- VanNostrand, M., & Keebler, G. (2009). Pretrial risk assessment in the federal court. *Federal Probation*, 73(2), 3-29.
- VanNostrand, M. and Lowenkamp, C. (2013). *Assessing Pretrial Risk without a Defendant Interview*, Laura and John Arnold Foundation.
- Walters, G. D., and Lowenkamp, C. T. (2016). Predicting recidivism with the Psychological Inventory of Criminal Thinking Styles (PICTS) in community-supervised male and female federal offenders. *Psychological Assessment*, 28, 652–659.
- Western, B. (2006). *Punishment and inequality in America*. New York, NY: Russell Sage.
- Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, Articles*, 59(10), 1–23.
- Winterfield, L., Coggeshall, M., and Harrell, A. (2003). *Empirically-based risk assessment instrument: Final report*. Urban Institute Justice Policy Center.