DOI: 10.1111/1745-9133.12481

CRIMINOLOGY
& Public Policy

ORIGINAL ARTICLE

RISK ASSESSMENT

Public safety assessment

Predictive utility and differential prediction by race in Kentucky

Matthew DeMichele | **Peter Baumgartner** | **Michael Wenger** | **Kelle Barrick** | **Megan Comfort**

RTI International

Correspondence
Matthew DeMichele, Center for Courts amd Corrections Research, 3040 E. Cornwallis Road, Research Triangle Park, NC 27709-2194.
Email: mdemichele@rti.org

Funding information
Laura and John Arnold Foundation, Grant/Award Number: 0215129

Research Summary: We assess the predictive validity and differential prediction by race 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 front end of the criminal justice system. The growing and disparate use of incarceration is one of the most pressing social issues facing the United States. The implementation of risk assessments has provided fuel for both sides of the reform debate with proponents arguing that the use of these assessments offers a policy mechanism to alleviate populations and bias. Risk assessment critics, however, argue that the use of the assessments exacerbates bias and does not improve decision-making. By examining a statewide data set 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. The differences we found indicate the PSA scores for failure to appear (FTA) are moderated by race, but these differences do not lead to disparate impact.

Policy Implications: We point to data limitations and the need for localized risk assessment studies, and we emphasize that risk assessments are decision-making tools that require ongoing refinement. Risk assessment developers, opponents, and proponents would do better to focus on

the reality of risk assessments as probabilistic models. The results of these assessments cannot predict with certainty, and they are not inherently biased. Rather, criminologists and policy makers need to understand the uncertainty that comes with any predictive model.

Pretrial populations are a large and growing contributor of mass incarceration. According to the Bureau of Justice Statistics (BJS), the proportion of jail populations that are unconvicted has increased from 50% in 1985 to nearly 65% in 2017 (Zeng, 2019). Although jail populations have declined to approximately 745,000 in 2017 since their peak in 2008 from slightly more than 785,000 (Zeng, 2019), BJS estimated that nearly 95% of the growth in jail populations since 2000 was a result of the increase in the proportion of those held in jails that are unconvicted (Minton & 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 & Ackerman, 2014). During pretrial, individuals are legally innocent and have a right to be released, but many jails are filled with pretrial populations because judges have the ability to detain individuals as a result of concerns of flight or safety (United States v. Salerno, 1985).

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 on criminal history (Gottfredson & Gottfredson, 1988; Spohn & Holleran, 2000), but these decisions are often made quickly, with limited information, and rely on predetermined bond schedules. 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 reliance on financial conditions for pretrial release has “almost from its inception, been the subject of dissatisfaction” (Ares, Rankin, & Sturz, 1963, p. 67). The nature of these concerns has been 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, 1954). Pretrial detention is associated with a higher likelihood of conviction and with longer terms of incarceration, and it has the potential to destabilize families (Sacks & Ackerman, 2014).[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 & Gottfredson, 1985; Mayson, 2018; United States v. Salerno, 1985). Furthermore, the legal rules for pretrial release allow for judges to consider extralegal factors such as employment, community ties, and marital status when deciding whether to release someone (Goldkamp & 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 (Pretrial Justice Institute, 2017). Pretrial risk assessments are developed to identify the likelihood that defendants will remain crime free and that they will return to court. The Pretrial Justice Institute (2017) estimated that approximately 25% of jurisdictions use an actuarial risk assessment, which is an increase from just 10% in 2013. These tools are emerging among vocal opposition about predictive utility and whether they contribute to racial disparities, with critics arguing that risk assessments are reliant on group-based patterns that will lead to unfair treatment of people of color (e.g., Harcourt, 2008, 2015; Starr, 2014, 2015), not to mention general opposition by the bail bonds industry.

Others, however, suggest that risk assessments can be used to structure criminal justice decisions, increase objectivity and fairness, and have the potential to reduce incarcerated populations (Cooprider, 2009; Flores, Bechtel, & Lowenkamp, 2016; Skeem & Lowenkamp, 2016). We address the extent to which prediction bias exists in a pretrial risk assessment in the current article by analyzing predictive validity and differential prediction by race of the Public Safety Assessment (PSA).

We investigate two primary research objectives using a data set from a statewide pretrial services agency in Kentucky ($N = 164{,}597$). First, we assess the overall predictive validity of the PSA. Second, we assess differential validity and predictive bias between Black and White defendants. Although there are many studies about risk assessment development and validation, there are fewer published studies within the criminal justice literature in which the potential for predictive bias by race has been assessed (for exceptions, see Cohen & Lowenkamp, 2019; Flores et al., 2016; Skeem & Lowenkamp, 2016).

First, we briefly describe the emergence and use of risk assessments within the criminal justice system and the pretrial system specifically. Second, we describe the development of the PSA and how the PSA is used. Third, we describe our methods and provide descriptive statistics and predictive utility measures, as well as test for predictive bias. In the end, we suggest that criminologists need to develop a better understanding of the drivers of pretrial failures and to move away from searching for a statistical silver bullet.[2] Instead, the field needs to develop normative standards of fairness and disparate impact and to develop a broader understanding of the capabilities of risk assessments by more fully acknowledging the uncertainty inherent in probabilistic tools.

1 | RISK ASSESSMENT AND PRETRIAL RISK ASSESSMENT

The use of risk assessments in the criminal justice system is not new. They have been used at least since 1928 when Burgess developed a parole risk assessment to help the Illinois paroling authority make release decisions. Since Burgess's time, the use of risk assessments has increased across criminal justice systems 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 and sentencing phases (Kleiman, Ostrom, & Cheesman, 2007), but judges are concerned about replacing their decision-making with statistical models (Chanenson, 2003).

Pretrial risk assessments were developed to reduce potential disparate impacts related to bail. In 1961, the Vera Institute of Justice found that using a risk assessment instrument increased release rates and improved court appearance rates compared with 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 incorporated Vera's instrument into their pretrial processes, and some jurisdictions created their own risk assessments.

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, & Harrell, 2003). Virginia developed a pretrial risk assessment instrument that comprised nine factors with six 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, & Spruance, 2016). Lowencamp, Lemke, and Latessa (2008)) validated an instrument among a sample of 342 adult defendants on pretrial release in several pretrial agencies across two states. They found FTA and a new arrest during pretrial were related to eight items (i.e., age at first arrest, history of FTA, FTA within 2 years, prior jail incarcerations, employment status, drug use, drug-related problems, and residential stability).

Some pretrial risk assessment instruments have been developed by several states, the District of Columbia, the federal court system, and approximately three dozen jurisdictions in approximately 15 states (Mamalian, 2011). The instruments are primarily reliant on measures of criminal history, but they also tend to include community ties, residential stability, substance abuse, employment and education, and age. These factors are 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, 2015; Starr, 2014). 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. They 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, p. 85). The results of Bechtel et al.'s meta-analysis and other research highlight the gap in knowledge about the different forms of pretrial failures. They demonstrated that potentially different factors are related to failure to appear versus a new arrest. An impetus for the current research is to assess validity and predictive bias of a popular pretrial risk assessment with the hopes of contributing to risk assessment and pretrial research more broadly.

2 | FAIRNESS: DEFINITION AND MEASUREMENTS

Fairness and bias with risk assessments are some of the biggest issues facing criminal justice policy makers and stakeholders. A heated debate has emerged about the fair use of risk assessments to inform decisions about pretrial release or detention, with these debates often at the center of discussions about broader pretrial reform. Criminologists have been slow to offer ideas about how to assess fairness and bias (Eckhouse, Lum, Conti-Cook, & Ciccolini, 2019). Simply put, what is predictive bias? How does a judge or a pretrial officer know whether a risk assessment increases bias in his or her jurisdiction? Additionally, although likely improving error in human decision-making alone, policies around risk assessments should be focused on their probabilistic nature.

Differential prediction in the criminal justice system received a lot of attention as a result, in part, of a *ProPublica* article in which the authors emphatically stated that the COMPAS risk assessment equated to "machine bias" (Angwin, Larson, Mattu, & Kirchner, 2016). This article set off something of a firestorm within the criminal justice research and practitioner communities because their analysis "turned up significant racial disparities" (Angwin et al., 2016). The *ProPublica* article, of course, was not the first to highlight the potential for predictive bias with risk assessments. Although the authors of the *ProPublica* article provided an insightful analysis, they overlooked other concerns such as how do risk assessments compare with decision-making as usual (i.e., mass incarceration was built without pretrial risk assessments) and whether or in what ways pretrial risk assessments can ease mass incarceration.

Risk assessment proponents argue that risk assessments "can scaffold efforts to unwind mass incarceration without compromising public safety" (Skeem & Lowenkamp, 2016, p. 705). Currently, there is little evidence that the use of a pretrial risk assessment will reduce jail populations, and in a recent impact study, Stevenson (2018) found no change in Kentucky after full implementation of the PSA. Based on the findings from initial studies, it is unlikely that risk assessments alone are going to unwind mass incarceration because mass incarceration was built through a constellation of policy changes that have unfolded since the 1970s, and few of these policies included pretrial risk assessments. Instead, there is a pretrial culture rooted in bond schedules, high detention rates for the poor and people of color, and detention as the status quo (Koepke & Robinson, 2018), and undoing such an interlocking set of policies will take sustained reform efforts that go far beyond risk assessments. Risk assessment proponents envision assessments as one tool within a broader suite of reforms that can expedite pretrial case processing, improve release decisions, and decrease bias related to errors in human decision-making.

Before risk assessments can be incorporated into a wider set of pretrial reforms, research is needed to understand the extent to which assessments are accurate or biased. Flores et al. (2016), p. 45) analyzed the same data as *ProPublica* using a different methodology, and they came to a nearly opposite conclusion of "no evidence of racial bias." Flores et al. claimed that *ProPublica* "strayed from their own code of ethics in that they did not present the facts accurately" and that the article provided "misleading information about the reliability, validity, and fairness" of risk assessments (Flores et al., 2016, p. 45). There is a need for moral and ethical investigations of pretrial risk assessments to consider fairness empirically and to compare to human decision-making.

Criminologists and data scientists have begun to study statistical fairness (Berk, Heidari, Jabbari, Kearns, & Roth, 2017) and the different results between Angwin et al. and Flores et al. fit what Kleinberg, Mullainathan, and Raghavan (2016) referred to as the "inherent tradeoffs in the fair determination of risk scores" as they, and other data scientists, have put forth several formal definitions of fairness (Berk et al., 2017; Chouldechova, 2017; Corbett-Davies & Goel, 2018). Investigating algorithms to assess fairness is a new area of research for criminology that has produced several proposed measures of fairness, including the following:

- **Error rate balance**, which is met when false-positive and false-negative rates are equal across groups at some threshold (e.g., Black and White defendants are incorrectly labeled "high risk," equal false-negative/positive rates by group)
- **Predictive parity**, which is met when the likelihood of recidivism/outcome for a particular threshold is equal across groups (e.g., Black and White defendants are correctly classified at a specific threshold)
- **Calibration**, which is met when an instrument provides similar probabilities for recidivism/outcome for any score regardless of group membership (i.e., a score of *X* has the same probability of outcome regardless of group membership)

This is not an exhaustive list of measures of statistical fairness, but these are the definitions most relevant for the PSA research (Romei & Ruggieri, 2013). Dwork, Hardt, Pitassi, Reingold, and Zemel (2012) viewed fair risk assignment as a constrained optimization problem in which the goal is to ensure public safety, avoid unnecessarily punishing people, and treat similar people similarly. And, in fact, the formal fairness definitions are all intended to achieve the same results of developing probability estimates that have the same effectiveness regardless of group membership (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018, p. 3). There are tradeoffs, however, as these definitions cannot be met simultaneously other than in rare situations (Chouldechova, 2017).

The legal scholarship critiquing sentencing risk assessments is instructive for developing pretrial risk assessments because it provides a foundation from which to reflect critically on the utility of potential factors used in the algorithm. Earlier pretrial risk assessments included factors related to criminal history, community ties, residential stability, substance abuse, employment and education, and age (Bechtel et al., 2011). The PSA, however, includes only criminal history factors, whether the current charge is violent, and age; it eschews factors related to prior arrests and instead uses prior conviction factors. These decisions were made to alleviate bias associated with socioeconomic characteristics and to increase fairness. These development decisions align with the less ardent critics of sentencing assessments that indicate that criminal history items are appropriate for legal decision-making (Starr, 2014). Skeem and Lowenkamp (2016) found that criminal history factors accounted for the majority of mean score differences between Black and White individuals.

2.1 | Public safety assessment

The PSA, which was created through investments made by the Laura and John Arnold Foundation (LJAF) using a large database of greater than 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 & Lowenkamp, 2013). The PSA was developed to identify the strongest predictors of failure to appear (FTA), new criminal activity (NCA), and new violent criminal activity (NCVA). Criteria for variable selection included the predictors needed to be related to the current charge (i.e., violent or not) or criminal history related, consistent with prior research findings, and gathered without a defendant interview (VanNostrand & Lowenkamp, 2013). The PSA intentionally leaves out demographic factors related to race/ethnicity and gender as well as socioeconomic variables such as 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. After the initial development of the PSA, researchers conducted validation analyses on a sample of more than 500,000 cases (i.e., validation sample) from jurisdictions in the Northeast, Southwest, Midwest, and two states (unpublished Luminosity training materials).[3] Initially, the PSA factors and weights were not released to the public and jurisdictions were required to sign nondisclosure agreements, but now the PSA is available to the public and jurisdictions can use a Web-based application to implement the PSA on their own (psapretrial.org).

The PSA is reliant 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 for release decisions. It also, however, has the potential to leave out information about individuals that judges might find beneficial. Second, LJAF created the PSA with intentions of creating a risk assessment that could be used by jurisdictions across the country. Many pretrial risk assessment instruments were not intended to be used outside of the jurisdiction in which they were developed.[4] 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 (http://www.arnoldfoundation.org/wp-content/uploads/Criminal-Justice-Data-Used-to-Develop-the-Public-Safety-Assessment-Final.pdf).[5]

In 1976, Kentucky became one of four states to ban commercial bail bonding services, and it is 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 noncriminal justice factors including housing and employment status, and the KPRA is completed with a defendant interview (Austin, Ocker, & Bhati, 2010). Kentucky's jail and prison population, similar to much of the country, grew throughout the 2000s and policy makers wanted 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 (Stevenson, 2018). 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. Stevenson (2018) reported further changes to the PSA in KY in mid-year 2014, and she did not find that jail populations or racial disparity decreased postimplementation of the PSA in Kentucky. The PSA is completed by pretrial officers or other relevant court personnel prior to first appearance. Pretrial officers use administrative data to conduct a thorough review of criminal history records. This article is the first pubic validation study of the PSA.

2.2 | Study method

The analysis is a validation and test of predictive bias of the PSA using an historical data set including all individuals booked into jail in Kentucky between July 1, 2013 and December 30, 2014. The data set was made available to us by LJAF through a cloud-based repository as part of a larger study of the PSA and legal actor decision-making. The instrument development team (Luminosity) collected the data from Kentucky and processed the data sets to develop deidentified analytic files with the binary outcome variables (i.e., failure to appear, new criminal arrest, and new violent criminal arrest), risk factors, age, race, gender, and booking and release dates. The data set for the current study was collected as part of postdevelopment validation of the PSA.[6] Variable creation for each risk factor, that is, matching individuals to criminal history and determining the presence or absence of specific risk factors, was completed by the PSA developers.

2.2.1 | Pretrial outcome variables

We scored each case using the scoring criteria for the PSA that provides separate risk scores for each of the three outcomes listed as follows:

- **Failure to appear (FTA)**: Variable measuring whether a released individual missed any court date before the disposition of their case.
- **New criminal activity (NCA)**: Variable measuring whether a released individual was arrested for any offense prior to the disposition of their case.
- **New violent criminal activity (NVCA)**: Variable measuring whether a released individual was arrested for a violent offense prior to the disposition of their case.

2.2.2 | Scoring the PSA

The PSA is scored by applying the predefined scores to each factor, summing the scores, and converting the scores to scales that range from 1 to 6 for each outcome. Three FTA factors are binary indicators (i.e., pending charge, any prior conviction, and any FTA older than 2 years) scored as 0, 1, and prior FTA within past 2 years (0 = 0, 1 = 2, and 2+ = 4). These risk scores range from 0 to 7 and were converted into an FTA scale score according to the PSA instructions. The risk scores are converted into FTA scale scores as follows: 0 = 1, 1 = 2, 2 = 3, 3 and 4 = 4, 5 and 6 = 5, and 7 = 6.

The NCA risk score includes seven factors that are scored as follows: prior misdemeanor conviction (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 scores are converted into an NCA scales as follows: 0 = 1, 1 and 2 = 1, 3 and 4 = 3, 5 and 6 = 4, 7 and 8 = 5, 9-13 = 6.

The NVCA risk scores range from 0 to 7. Three factors are binary indicators (i.e., any pending charges, any prior convictions, and current offense is violent and ≤20 years old) 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 to 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 scale score of 5 and 6 receiving a violent flag.

2.2.3 | Data set and research questions

The analyses were focused on the released adult pretrial population with a validation data set containing 286,247 cases from Kentucky. The data set was restricted to all adult defendants booked into jail during the 18-month study period. Cases were removed if they were younger than 18 years of age at the time of booking (n = 679) or had booking dates after December 2014 (n = 45,299). Cases were defined as released if they had release dates and had disposition dates after their release dates. The validation released sample included N = 164,597 (68.5%) with n = 75,662 (31.5%) detained. The scoring rules were applied for each three outcome (i.e., FTA, NCA, and NVCA) for each caseto address the following research objectives:[7]

- ***Assess Overall Predictive Validity***: How accurately does the PSA predict each of the three outcomes (i.e., FTA, NCA, and NVCA)?
- ***Assess Predictive Validity by Race:*** How accurately does the PSA predict each of the three outcomes of interest by race?
- ***Assess Differential Prediction by Race:*** Does the PSA provide different results based on race? We expect to find that the PSA predicts equally well across race (i.e., race will not moderate the relationship between the PSA and failures).

2.2.4 | Analyses

First, we review the descriptive characteristics of the sample and report the presence of each factor (by outcomes) for the entire sample and by race. The results of these initial analyses demonstrate that a higher percentage of Black defendants have each of the risk factors relative to the entire sample and White defendants. The base rates by scale scores are presented and demonstrate a general similarity in failure rates by race across the scale scores.

Second, predictive validity is measured using Area Under the Curve (AUC) Receiver Operator Characteristics (ROC) estimates. AUCs are commonly used to evaluate risk assessment tools (Singh & 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 .5 referring to random chance and 1.0 referring to perfect prediction. The ROC score provides an 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. Desmarais, Johnson, and Singh (2016) have conducted meta-analyses and evaluations of criminal justice risk assessment instruments. They suggested that AUC values of .54 and below are poor, .55 to .63 are fair, and .64 to .7 are good, with values higher than .71 being excellent. Using these ranges, the ROC values for PSA for the three outcomes are in the good range.

The third research question is focused on whether there is evidence of predictive bias by race with the PSA. We use a moderator regression technique commonly cited in psychological studies and testing literature (e.g., Cleary, 1968; Sackett, Borneman, & Connelly, 2008). As a result, we can estimate four regression models to assess 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, & Lowenkamp, 2017, p. 193). Predictive bias is tested

TABLE 1 Descriptive statistics of Kentucky defendants

Variable	*n*	% or mean
Release Status		
Detained	75,662	31.5
Released	164,597	68.5
Released Cases		
Age	164,597	34
Sex		
Male	113,376	68.9
Female	50,592	30.7
Unknown	629	0.4
Race		
White	133,517	81.1
Black	27,656	16.8
Other	2,174	1.3
Unknown	1,250	0.8
Base Failure Rates		
FTA	24,293	14.8
NCA	17,512	10.6
NVCA	1,826	1.1

Note. Other race includes Asians, Native Americans, and "Other" races.

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 moderator regression technique is recommended by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2018). The approach is designed to test for whether the PSA scores are moderated or conditioned by race as they predict the outcomes (i.e., a given score on the PSA does not have the same meaning for Black and White defendants).

As a result of the large sample size, we follow recent practices and set more stringent statistical significance levels at $p < .001$. For example, Monahan et al. (2017) followed this approach with a data set of 7,350, which is much smaller than either of the data sets used here (Lin, Lucas, & Shmueli, 2013).

2.3 | Findings

The sample characteristics are reported in Table 1 and show that 68% ($n = 164{,}597$) of defendants were released in Kentucky (68%) prior to trial. The Kentucky defendants are 34 years old on average, 81% ($n = 133{,}517$) of the defendants are White and nearly 17% of the defendants are Black.[8] Nearly 70% of the defendants are male ($n = 113{,}376$). The overall base rates for the three outcomes are FTA rate of 14.8%, an NCA rate of 10.6%, and an NVCA rate of 1.1%. These sample descriptions are similar to what was reported in other studies of Kentucky's pretrial system. Austin et al. (2010) cited a 74% release rate for a 3-month validation study (July through September 2009), although they had lower FTA (8%) and NCA rates (7%). Stevenson (2018) did not report the overall release rate, but she reported that 77% of misdemeanors and 62% of felony cases were released prior to disposition. She also reported that 10% and 8% of misdemeanor and 13% and 8% of felony defendants had an FTA or an NCA, respectively.

2.3.1 | Risk factors by race

Risk assessments are routinely criticized for being inherently biased against people of color. Table 2 shows how the risk factors are distributed across the entire released sample and by race. The PSA includes a total of 11 factors in which each of the three outcomes are modeled with 4 (FTA), 5 (NVCA), and 7 (NCA) factors for each outcome. Reviewing the distribution of risk factors by race allows for us to understand what factors are driving any racial differences in scores. We find that a larger proportion of Black individuals have each risk factor compared with White individuals. The higher scores for Black individuals fit with Skeem and Lowenkamp's (2016) finding that mean differences in risk assessment scores were a result of higher criminal histories for Black probationers. The PSA factors are nearly all criminal history items. There are structural differences in which Black individuals live in communities that are more heavily policed, and thus, they are more likely to be stopped, arrested, and convicted than White individuals. More aggressive enforcement likely drives some differences in convictions, especially misdemeanor convictions (in which officers have more discretion). A greater proportion of Black individuals have prior FTAs, felony convictions, prior incarcerations, and prior convictions for violent crimes. These risk assessment scores alone are not evidence of predictive bias with the PSA as subgroup differences are a common feature of social phenomenon. It is important to remember that the validation sample includes only released individuals, and in Kentucky, judges are willing to release Black defendants with higher average risk scores, which could mean that judges are using alternative omitted variables to make release decisions (Kleinberg et al., 2018).

2.3.2 | Predictive utility

In table 3, we show that individuals are more likely to experience each outcome as the scores increase. For FTAs, between 7% and 10% of defendants with scores of one and two had an FTA during their pretrial period, whereas between 26% and 32% of those with a score of five and six had an FTA. NCAs have a similar pattern with between 4% and 7% of defendants with scores of one or two arrested for a new crime during pretrial. NCA rates increase as the scale score increases such that more than a fourth of those with an NCA score of six are rearrested for a new crime during pretrial. The PSA provides a binary indicator flag for NVCAs, and three times the proportion of those with a violent flag were rearrested for a violent crime during pretrial relative to those without the flag. The AUCs for each outcome are in the good range. The FTA AUC (.646) has the lowest AUC of the three outcomes, NCAs (AUC = .650) are slightly higher, and higher still for NVCAs (AUC = .664). These AUCs are similar to the findings from Danner et al. (2016)) research on the Virginia pretrial risk assessment but lower than what Skeem and Lowenkamp (2016) reported for the PCRA. Austin et al. (2010), in their validation of the Kentucky pretrial risk assessment, did not include AUC scores.

Table 4 presents the failure rates by PSA scale score for Black and White individuals. In the case of FTAs, Black individuals with PSA scores of 1 and 2 (the lowest scores) have significantly higher FTA rates than White individuals, but these differences do not remain as the scale scores increase. For FTA scale scores of three or higher, the FTA rates are nearly identical by race within each score, but Black individuals do have higher FTA base rates. There is a similar pattern with NCA rates, with the exception that White individuals with lower NCA scale scores have higher failure rates ($p < .001$) than Black individuals. These differences disappear as the scale scores increase and there is no difference in NCA base rates by race.

TABLE 2 PSA distribution of risk factors for released defendants and by race

PSA Risk Factor		Percentage of Released (*n*)	Percentage of Released Black Defendants (*n*)	Percentage of Released White Defendants (*n*)	Scales Incorporating Risk Factor
Pending Charge	Yes	19.0 (31,294)	20.8 (5,745)	18.9 (25,202)	FTA, NCA, NVCA
	No	81.0 (133,303)	79.2 (21,911)	81.1 (108,315)	
Prior FTA in Past 2 Years	Two or More	13.0 (21,347)	18.0 (4,990)	12.1 (16,187)	FTA, NCA
	One	17.6 (28,993)	20.3 (5,607)	17.1 (22,890)	
	No	69.4 (114,257)	61.7 (17,059)	70.7 (94,440)	
Prior FTA Older than 2 Years	Yes	41.3 (67,972)	48.3 (13,362)	40.4 (53,963)	FTA
	No	58.7 (96,625)	51.7 (14,294)	59.6 (79,554)	
Any Prior Conviction	Yes	74.5 (122,545)	79.2 (21,899)	74.1 (98,962)	FTA, NVCA
	No	25.5 (42,052)	20.8 (5,757)	25.9 (34,555)	
Prior Misdemeanor Conviction	Yes	72.8 (119,875)	77.1 (21,324)	72.6 (96,907)	NCA
	No	27.2 (44,722)	22.9 (6,332)	27.4 (36,610)	
Prior Felony Conviction	Yes	29.2 (48,034)	39.8 (10,998)	27.5 (36,755)	NCA
	No	70.8 (116,563)	60.2 (16,658)	72.5 (96,762)	
Prior Violent Conviction	Three or More	4.0 (6,643)	6.8 (1,893)	3.5 (4,722)	NCA, NVCA
	One to Two	17.8 (29,322)	24.8 (6,852)	16.6 (22,212)	
	No	78.1 (128,632)	68.4 (18,911)	79.8 (106,583)	
Prior Sentence to Incarceration > 14 days	Yes	32.4 (53,288)	41.0 (11,336)	31.1 (41,588)	NCA
	No	67.6 (111,309)	59.0 (16,320)	68.9 (91,929)	
Current Age	< = 22 Years	16.2 (26,720)	21.5 (5,934)	15.1 (20,130)	NCA
	> = 23 Years	83.8 (137,877)	78.5 (21,722)	84.9 (113,387)	
Current Violent Offense & ≤ 20 Years Old	Yes	1.4 (2,263)	2.4 (666)	1.2 (1,560)	NVCA
	No	98.6 (162,334)	97.6 (26,990)	98.8 (131,957)	
Current Violent Offense	Yes	14.6 (23,986)	16.9 (4,672)	14.2 (18,897)	NVCA
	No	85.4 (140,611)	83.1 (22,984)	85.8 (114,620)	

Notes. Black and White individuals will not add to the total *N released* as a result of the exclusion of records with “Other” and “Unknown” race. Percentages may not add up to 100% as a result of rounding.

TABLE 3 PSA failure rates by scale scores for FTA, NCA, NVCA

Risk Scale	***n***	**Failure to Appear (FTA)**	**New Criminal Activity (NCA)**	**New Violent Criminal Activity (NVCA)**
FTA Scale Score				
One	2,186	7.5%	—	—
Two	4,171	9.7	—	—
Three	5,287	13.9	—	—
Four	5,901	19.8	—	—
Five	5,163	26.5	—	—
Six	1,585	32.1	—	—
Base FTA Rate	*24,293*	*14.8*	—	
NCA Scale Score				
One	834	—	3.9%	—
Two	3,575	—	6.8	—
Three	4,499	—	10.9	—
Four	4,769	—	15.1	—
Five	2,513	—	19.7	—
Six	1,322	—	26.3	—
Base NCA Rate	*17,512*	*10.6*		
NVCA Binary Indicator				
Violent Flag (5–6)	325	—	—	3.0%
No Violent Flag (1–4)	1,501	—	—	1.0
Base NVCA Rate	*1,826*			*1.1*
AUC, PSA Total		0.646	0.650	0.664
99.9% CI		(0.641–0.658)	(0.642–0.656)	(0.641–0.683)

Notes. — = not applicable; AUC = area under the ROC curve; CI = confidence interval.

Significantly larger proportions of Black defendants are arrested for a violent crime among scores of 1–4 and nearly double the base rate for new violent arrests. The NVCA scores are collapsed into the binary NVCA flag with scores of 1–4 equal to no flag and scores of 5–6 equal to a violent flag. The findings from these bivariate analyses show that Black defendants with low NVCA scores (without the flag) have a significantly higher rate of NVCAs relative to that for White defendants with similar scores but no significant differences between White and Black defendants with the violent flag.

In Table 4, we assess differential predictive validity to determine to what extent the PSA is more accurate at predicting pretrial outcomes for White or Black defendants. The AUCs in Table 4 show that there are significant differences in the predictive validity between Black (AUC = .612) and White (AUC = .655) defendants. The FTA AUCs demonstrate that PSA FTA scale scores are not as accurate at classifying FTAs for Black defendants as they are for White defendants. The PSA FTA scale scores, at best, are a fair predictor of FTAs for Black defendants. There are no significant differences in validity between White and Black defendants for the NCA or NVCA scales—other than the AUC for NVCAs for Black defendants, all of the AUCs are in the good range.

Assessing the strength or degree of utility for the each PSA scale 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 comprises four regression models used in the following

TABLE 4 PSA failure to appear, new criminal activity, and new violent criminal activity by race

	Failure to Appear (FTA)				**New Criminal Activity (NCA)**				**New Violent Criminal Activity (NVCA)**			
	White		**Black**		**White**		**Black**		**White**		**Black**	
Risk Scale	***n***	%	***n***	%	***n***	%	***n***	%	***n***	%	***n***	%
FTA Scale Score					—	—	—	—	—	—	—	—
One*	1,623	6.7%	406	11.4%	—	—	—	—	—	—	—	—
Two*	3,333	9.3	722	11.8	—	—	—	—	—	—	—	—
Three	4,243	13.8	967	14.3	—	—	—	—	—	—	—	—
Four	4,740	20.0	1,104	19.7	—	—	—	—	—	—	—	—
Five	3,988	26.5	1,132	26.1	—	—	—	—	—	—	—	—
Six	1,195	32.4	381	31.0	—	—	—	—	—	—	—	—
Base FTA Rate	*19,122*	*14.3*	*4,712*	*17.0**	—	—	—	—	—	—	—	—
NCA Scale Score												
One	—	—	—	—	746	4.0%	68	3.4%	—	—	—	—
Two*	—	—	—	—	3,098	7.1	421	5.8	—	—	—	—
Three*	—	—	—	—	3,843	11.4	620	8.9	—	—	—	—
Four	—	—	—	—	3,708	15.2	1,033	14.8	—	—	—	—
Five	—	—	—	—	1,946	20.0	558	18.7	—	—	—	—
Six	—	—	—	—	935	26.5	384	25.7	—	—	—	—
Base NCA Rate	—	—	—	—	*14,276*	*10.7*	*3,084*	*11.1*	—	—	—	—
NVCA Scale Score												
One*	—	—	—	—	—	—	—	—	101	.4%	37	.9%
Two*	—	—	—	—	—	—	—	—	345	.6	107	1.1
Three*	—	—	—	—	—	—	—	—	359	1.1	129	1.8
Four*	—	—	—	—	—	—	—	—	287	2.0	123	2.9
Five	—	—	—	—	—	—	—	—	153	2.6	58	3.1
Six	—	—	—	—	—	—	—	—	76	3.5	36	4.7
Base NVCA Rate	—	—	—	—	—	—	—	—	*1,321*	*.9*	*490*	*1.7**
AUC, PSA		0.655		0.612*		0.647		0.659		0.666		0.631
99.9% CI		(0.649–0.661)		(0.598–0.626)		(0.639–0.654)		(0.643–0.676)		(0.641–0.690)		(0.589–0.672)

Notes. — = not applicable; AUC = area under the ROC curve; CI = confidence interval.

$^{*}p < .001$.

sequence. First, we estimate a model with only the subgroup of interest (i.e., race). Second, a model is fit with only the test score (i.e., PSA score for each outcome separately). Third, we use a model to estimate both the subgroup and the PSA score. Fourth, in our final model, we include the subgroup, PSA score, and the interaction of the subgroup and the PSA score, which allows us to estimate the main effects of each variable separately before testing to see whether the PSA by race interaction terms is significant. The interaction term tests to what extent the likelihood of a pretrial failure is a matter of how race moderates the PSA scores such that scores will have different meanings for each racial group. Risk assessments are intended to classify individuals so that a score *X* means the same regardless of subgroups.

Table 5 presents the odds ratios and confidence intervals for the four regression models for FTA, NCA, and NVCA. For FTA, there are consistent significant effects for race, FTA score, and the interaction term. These differences show 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 be seen in Figure 1 in which we plot the predicted probabilities for an FTA by FTA score for White and Black defendants from Model 4 in Table 5. Figure 1 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.

The odds ratios for the interaction term shows differential prediction by race. The relationship between the FTA scores and FTAs are moderated by race. Figure 1 shows flatter slopes and higher intercepts for Black defendants. In Table 5, Model 3 shows the general predictive utility of the FTA scale scores; with each 1-point increase in the scale score, there is a 47% increase in the odds of a defendant experiencing an FTA. These effects, however, are moderated by race such that in Figure 1 (results from Model 4) the intersecting lines show that Black defendants' FTA rates are underestimated. Figure 1 shows that a Black defendant with a score of 1 has the same predicted probability of an FTA as a White defendant with a score of 2 but that these differences fade as risk scores increase with no differences between White and Black defendants with a score of 4 and White defendants with a score of 5 or 6 having a higher likelihood of an FTA than Black defendants with the same score (for a similar result in the testing literature, see Houston & Novick, 1987, p. 319).

Table 5 includes the results for similar regression models testing for differences between Black and White defendants for the NCA and NVCA outcomes. For NCA we found a much different pattern in which in two of the three models, White defendants are predicted to be ∼14% and 28% more likely to have a new arrest for any crime during their pretrial period. Although we did not find that that race moderates the relationship between the PSA and a new crime, these results did show that White defendants in Kentucky are more likely to be arrested after controlling for the NCA scale score. Figure 2 plots the predicted probabilities for an NCA by race (using Model 4) and shows that, although the lines are parallel and not significantly different, White defendants are more likely to be arrested pretrial across each scale score. These findings fit with the actual NCA failure rates (Table 4) in which White defendants have significantly higher rates of NCAs than do Black defendants for NCA scores of 2 and 3, and slightly greater proportions of Whites arrested for a new crime at every NCA scale score.

In Table 5, we report the four logistic regression models testing for race differences on the NVCA scale. [9] There are significant main effects for race and for the NVCA scales to predict future violent arrests during pretrial. The findings from the models do not demonstrate that the PSA is moderated by race, but Black defendants are significantly more likely to be arrested for a new violent crime during their pretrial period. Figure 2, however, plots the predicted probabilities of a new violent criminal arrest (using model 4) by NVCA scale scores for Black and White defendants. The lines are parallel and show nearly identical predicted probabilities for White and Black defendants by scale score. For instance,

TABLE 5 Logistic regressions models testing the predictive fairness of the PSA by Race for FTA, NCA, and NVCA

Failure to Appear (FTA)												
	Model 1			**Model 2**			**Model 3**			**Model 4**		
		99.9% CI			**99.9% CI**			**99.9% CI**			**99.9% CI**	
Model variables	**Odds Ratio**	**Lower**	**Upper**	**Odds Ratio**	**Lower**	**Upper**	**Odds Ratio**	**Lower**	**Upper**	**Odds Ratio**	**Lower**	**Upper**
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
Log-likelihood	–67468.			–59274.			–59254.			–59220.		
Model Pseudo R^2	0.001			0.043			0.043			0.043		
New Criminal Activity (NCA)												
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	0.032
Log-likelihood	–49254.			–47227.			–47210.			–47209.		
Model Pseudo R^2	0.000			0.040			0.041			0.041		
New Violent Criminal Activity (NVCA)												
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
Log-likelihood	–8643.0			–8407.9			–8377.9			–8375.0		
Model Pseudo R^2	0.006			0.032			0.036			0.036		

Notes. — = not applicable; AUC = area under the ROC curve; CI = confidence interval.

$^*p < .001$.

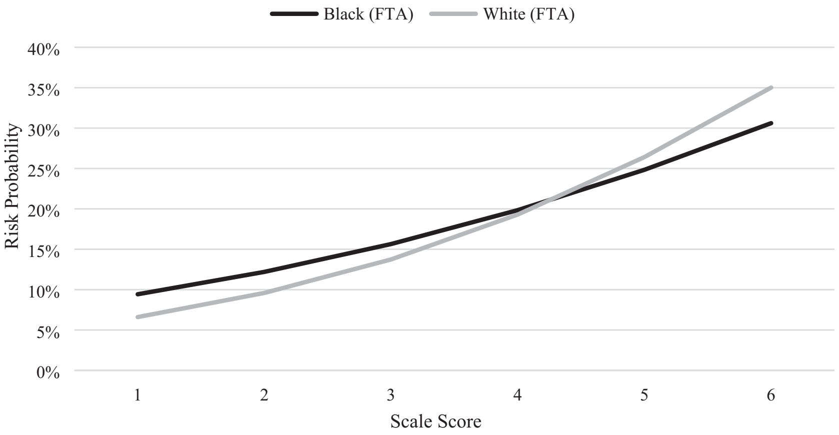

FIGURE 1 Predicted probabilities of failure to appear by PSA FTA score between Whites and Blacks

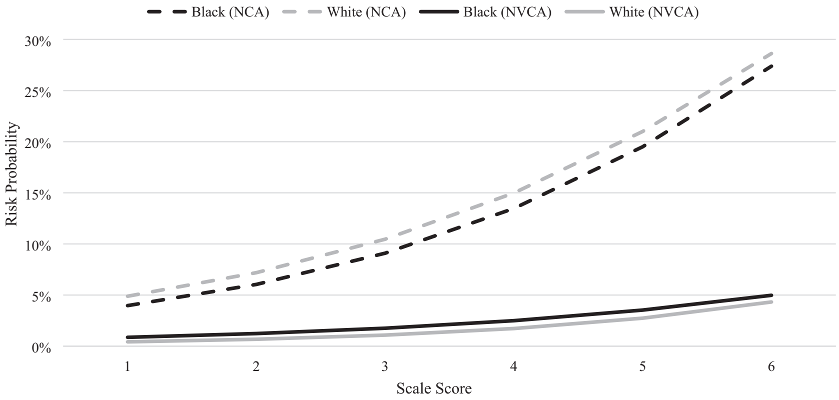

FIGURE 2 Predicted probabilities of pretrial new criminal activity and new violent criminal activity by PSA NCA and NVCA scores between Whites and Blacks

5.0% of Black defendants with a 6 on the NVCA scale score are expected to have an NVCA, whereas ∼4.9% of White defendants with a 6 are expected to have an NVCA.

3 | CONCLUSION

In this article, we provide the first public validation of the PSA that includes tests of predictive validity and predictive bias by race. The PSA has been adopted by dozens of jurisdictions, and it is contributing to thousands of release decisions each day. The findings indicate that when drawing two random cases from the data set, one of which had the pretrial outcome and the other did not, between 64% and 66% of the time the case with the pretrial outcome would have a higher score than the successful case. The results reported here showed that the PSA meets standards for criminal justice risk assessments (Desmarais et al., 2016) and findings reported elsewhere with a federal pretrial risk assessment (Cohen & Lowenkamp, 2019).

The FTA scale score was significantly more predictive for White defendants (AUC = .655) than for Black defendants (AUC = .612). This is a large and significant difference ($p < .001$) demonstrating that the FTA scale is a *fair* predictor of pretrial success for Black defendants but shows *good* performance for White defendants. After reviewing the public safety outcomes, we did not find a significant difference (p = .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 = .666) than for Black defendants (ROC = .631). Although the difference in NVCA accuracy is large, it is not significant (p = .015). Our findings show that the PSA has general parity for predicting new arrests across races but that there are significant differences by race for FTAs.[10]

The final research question is focused on whether there is predictive bias by race with the PSA. To answer this question, we used a moderator regression modeling approach (Cleary, 1968; Sackett et al., 2008; SIOP, 2018) that is commonly used to test for race and gender bias for several cognitive (e.g., ACT and GRE) and employment tests, and recently this approach has been applied in several criminal justice risk assessment studies (Cohen & Lowenkamp, 2019; Skeem & Lowenkamp, 2016). The regression approach is followed to test for intercept and slope differences and is well established for testing bias in psychometric scales (e.g., United States v. City of Erie, 2005). We build on that work and follow the definition for predictive bias issued by SIOP (2018).

There are intercept differences for FTAs, NCAs, and NVCAs, as well as slope differences for FTAs. 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 slope differences. The slope differences for FTAs indicate that the FTA scores are moderated by race. The FTA model in Kentucky has both intercept differences (Table 5 comparing Model 2 with Model 3)[11] and slope differences (comparing Model 3 with Model 4). These effects are clearly seen in Figure 1 in which the lines cross one another around an FTA score of 4.

The differences in the slopes show that the FTA scale is not well calibrated. Overall, Black defendants have higher mean FTA scale scores (3.17) than do White defendants (2.38) and that Black defendants have higher FTA base rates. Although the FTA scores are not calibrated by race, the nature of error is that the PSA underpredicts the likelihood for Black defendants (i.e., resulting in higher false negatives) with low scale scores to miss court. Assessing pretrial risks requires considering the different costs related to different types of errors such that jurisdictions need to decide how concerned they are with false negatives for FTAs (i.e., someone misses court) versus false negatives for violent crimes (i.e., a violent crime is committed; Berk et al., 2017). Jurisdictions must grapple with the social costs related to decision errors (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017) because holding someone in jail comes with a host of costs to the individual, his or her families, and the system. Pretrial detention is routinely found to be related to a host of legal consequences (Corbett-Davies et al., 2017; Dobbie, Goldin, & Yang, 2018; Gupta, Hansman, & Frenchman, 2016; Heaton, Mayson, & Stevenson, 2017) for individuals such that jurisdictions may want to consider whether detaining people as a result of a risk of missing a court date is justified. These considerations are important especially because most people are successful during pretrial with nearly 85%, 90%, and 99% of released individuals attending all court dates and not arrested for a new crime or a violent crime, respectively.

3.1 | Limitations and future research

The findings should be interpreted with an understanding of the limitations and weaknesses of this study. The analyses are based on one statewide pretrial release population. Kentucky is unique in ways that may weaken generalization and external validity. Kentucky is a small rural state with ~40% 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 they serve to highlight the need for ongoing research about the drivers of FTAs. A further limitation of the study is that we had only a limited data set. The data set allowed for conducting a validation and test of bias by race but there are at least three central limitations of these data. First, the data set does not include detailed information about the conditions of release or whether someone was being actively monitored by pretrial services. Second, the data set lacks important variables related to case processing factors (e.g., did defendants have a defense attorney) and time at risk was not included. Third, these data are at the case and not at the individual level, and we did not have access to a unique identifier to collapse multiple cases for each individual.

The PSA is used in dozens of jurisdictions, and our results provide information about the instrument's performance in Kentucky. Research is needed throughout the jurisdictions using the PSA to assess accuracy and evaluate predictive bias using local data. For instance, jurisdictions that provide court text reminders, have implemented deflection policies, or no longer detain misdemeanants for inability to pay bail will have different patterns and distributions for the risk factors and outcomes than will jurisdictions lacking these policies or than the same jurisdiction before implementing such reforms (Koepke & Robinson, 2018). Currently, pretrial and bail reforms are being implemented widely, and these changes can have meaningful impacts on the actual risks of pretrial failure for defendants. New risk-mitigating policies implemented by state and local governments will likely alter defendant risks, and jurisdictions implementing the PSA as part of a broader suite of pretrial reforms will likely have different future patterns of pretrial outcomes. Just as these within-site changes can reduce the predictive validity of an instrument developed with historic data, differences in local context may affect the validity (and value) of tools developed with data from other jurisdictions. Effectively using risk assessments requires developing a systematic and ongoing approach to validation to understand the relationship between factors, weights, and cutoffs, as well as to evaluate the potential for instrument bias.

3.2 | Discussion

Pretrial decisions affect millions of people every year, with the extent of these consequences distributed unevenly across racial and economic lines. For nearly 100 years, criminologists have routinely found pretrial detention to disproportionately, unfairly, and unnecessarily jail poor people, people of color, and the most vulnerable people in our communities. The PSA alone cannot, of course, be used to fix all these problems. The problems with our pretrial justice system go far beyond the use of a risk assessment, but risk assessments have the potential to be part of more ambitious pretrial reforms. The current problems with pretrial detention are not new. Instead, jurisdictions continue to struggle with slow case processing times, limited defense counsel, and the inability of many people to afford even seemingly small bail amounts.

The PSA was developed to speed arraignment, improve identification of low- versus high-risk defendants, and decrease pretrial incarceration. The PSA is a short instrument comprising criminal justice factors (e.g., criminal history and current offense violent) and young age and provides 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. 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 (Corbett-Davies et al., 2017; Harcourt, 2015).

The current round of controversy about risk assessments seems to be from groups with similar goals and desired outcomes. That is, risk assessments are developed with the hopes of reducing jail and prison populations and of decreasing racial and ethnic disparities. Proponents suggest that risk assessments remove conscious and unconscious forms of human bias that pervades the criminal justice system. There is nothing inherent in risk assessments, however, that will reduce jail populations, make prison populations less racially disparate, or otherwise reform the criminal justice system (Stevenson, 2018). Risk assessments are, essentially, probabilistic models, and as such, the classification recommendations always have a certain amount of presumed error. The well-known line from George Box is instructive here: "All models are wrong, but some are useful."

There is a lot of emphasis placed on risk assessments as developers often argue how well they predict outcomes, but we should take Box's statement seriously and focus more on the uncertainty implicit in risk assessments. A key goal for risk assessments is for them to be useful to decision makers, not to replace human decision makers. Risk assessments need to be optimized on fairness as well as on accuracy that meet the needs and standards set by local communities and stakeholders. The results of statistical analysis can only provide estimates of the likelihood of outcomes associated with certain subpopulations, but these findings cannot tell policy makers and stakeholders what they should do with that information (Berk et al., 2017).

CONFLICTS OF INTEREST

All authors have no conflict of interest to declare.

ENDNOTES

[1] The association between pretrial release and detention with future crime is still being determined. Dobbie, Goldin, and Yang (2017) recently did not find an association in Miami-Dade County between pretrial release or detention with committing a crime within 4 years. Gupta, Hansman, and Frenchman (2016) and Heaton, Mayson, and Stevenson (2017) in two other studies found in Philadelphia and Pittsburgh, and Harris County, Texas, respectively, that pretrial detention was associated with a 6% to 9% and 22% increase in crime within 1 year of release, respectively.

[2] This is a point made most forcefully by several data scientists, with Berk, Heidari, Jabbari, Kearns, and Roth (2017, p. 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. Furthermore, stakeholders need to agree and commit to the use of risk assessments in a consistent way or move away from them.

[3] 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 (https://www.arnoldventures.org/work/pretrial-justice/).

[4] 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 as a result of cost restraints and to adopt an assessment developed in another jurisdiction. But, of course, universal backend assessments exist (e.g., the Level of Service Inventory).

[5] There is yet to be a fully detailed methodological document made public, but instrument development team used a series of statistical techniques (e.g., logistic regression and contingency tables) that produced hundreds of effect sizes. The effect sizes were averaged and were restricted to variables that were at least 1 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% 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).

[6] The unit of analysis are cases, which means that any individual could be in the data set multiple times for multiple arrests. This is the same unit of analysis used by the PSA developers, as well as used by Stevenson (2018, footnote 191, p. 37) and is commonly found in pretrial studies (e.g., Dobbie et al., 2017).

[7] We follow recent practices and set more stringent statistical significance levels at $p < .001$ as a result of the large sample sizes used for both jurisdictions. For example, Monahan, Skeem, and Lowenkamp (2017) followed this approach with a data set of 7,350, which is much smaller than either data set used here.

[8] According to the U.S. Census, Kentucky has an overall population of 4,454,189, 85% White non-Hispanic, 8% Black (https://www.census.gov/quickfacts/KY).

[9] We do not fully address in this article that, although logistic regression is not as susceptible to problems stemming from unbalanced groups sizes as is linear regression, estimation difficulties do arise in cases of rare events as a result of 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) provided a critique of estimating logistic regression models with rare events and suggested that, despite large sample sizes, when event occurrence is lower than 5%, there could be instability in the models.

[10] There were differences in correlation coefficients between for FTAs and NCAs by race. The FTA $r = 0.15$ and $r = 0.19$ and NCA $r = 0.18$ and $r = 0.16$, Black and White defendants, respectively. There were no differences for NVCA, $r = 0.06$ for Black and White defendants.

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

ORCID

Matthew DeMichele https://orcid.org/0000-0002-6534-474X

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington: AERA Publications.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. There is software that is used across the country to predict future criminals. And it is biased against Blacks. *ProPublica*, Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ares, C., Rankin, A., & Sturz, H. (1963). *The Manhattan Bail Project: An interim report on the pre-trial use of pre-trial parole*. New York: Vera Foundation.

Arnold, H. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior and Human Performance*, *29*, 143–174.

Austin, J., Ocker, R., & Bhati, A. (2010). Kentucky pretrial risk assessment instrument validation. *The JFA Institute*, Retrieved from https://www.pretrial.org/download/risk-assessment/2010%20KY%20Risk%20Assessment%20Study%20JFA.pdf

Bechtel, K., Lowenkamp, C., & Holsinger, A. (2011). Identifying the predictors of pretrial failure: A meta-analysis. *Federal Probation*, *75*(2), 78–87.

Beeley, A. L. (1927). *The Bail system in Chicago*. Chicago: Univ. of Chicago Press.

Berk, R. A., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). *Fairness in criminal justice risk assessments: The state of the art* (Working paper). Retrieved from https://arxiv.org/pdf/1703.09207.pdf

Burgess, E. W. (1928). Factors determining success or failure on parole. In A. A. Bruce, A. J. Harno, E. W. Burgess, & E. W. Landesco (Eds.), *The working of the indeterminate sentence law and the parole system in Illinois* (pp. 221–234). Springfield: State Board of Parole.

Chanenson, S. (2003). Sentencing and data: The not-so-odd couple. *Federal Sentencing Reporter*, *16*(1), 1–7.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163.

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., & Lowenkamp, C. (2019). Revalidation of the federal PTRA: Testing the PTRA for predictive bias. *Criminal Justice and Behavior*, *46*(2), 234–260.

Cooprider, K. (2009). Pretrial risk assessment and case classification: A case study control. *Federal Probation*, *73*(1), 1–14.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. Retrieved from https://arxiv.org/pdf/1808.00023.pdf

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806). New York: Association for Computing Machinery.

Danner, M., VanNostrand, M., & Spruance, L. (2016). *Race and gender neutral pretrial risk assessment, release recommendations, and supervision: VPRAI and PRAXIS Revised*. New York: Luminosity.

Demuth, S. (2003). Racial and ethnic differences in pretrial release decisions and outcomes: A comparison of Hispanic, Black, and White felony arrestees. *Criminology*, *41*(3), 873–908.

Desmarais, S. L., Johnson, K., & Singh, S. (2016). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services*, *13*(3), 206–222.

Dobbie, W., Goldin, J., & Yang, C. (2018). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, *108*(2), 201–240. https://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.20161503

Dwork, C., Hardt, T., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations of Theoretical Computer Science* (pp. 214–226). https://arxiv.org/abs/1104.3913

Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, *46*(2), 185–209.

Flores, A., Bechtel, K., & 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.

Goldkamp, J. S., & Gottfredson, M. R. (1985). Policy guidelines for bail: An experiment in court reform. Philadelphia: Temple University Press.

Goldkamp, J. S., & Vilcica, E. R. (2009). Judicial discretion and the unfinished agenda of American bail reform: Lessons from Philadelphia's evidence-based judicial strategy. *Studies in law, politics, and society volume : New perspectives on crime and criminal justice* (Vol. 47, pp. 115–157). Bingly: Emerald Publishing.

Gottfredson, Michael R., & Gottfredson, D. (1988). *Decision making in criminal justice: Toward the rational exercise of discretion* (2nd ed.). New York: Plenum Press.

Gupta, A., Hansman, C., & Frenchman, E. (2016). The heavy costs of high bail: Evidence from judge randomization. *The Journal of Legal Studies*, *45*(2), 471–505.

Harcourt, B. E. (2008). *Against prediction: Profiling, policing, and punishing in an actuarial age*. Chicago: University of Chicago Press.

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, *27*, 237–243.

Heaton, P., Mayson, S., & Stevenson, M. (2017). The downstream consequences of misdemeanor pretrial detention. *Stanford Law Review*, *69*(3), 711–794.

Houston, W., & Norvick, M. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement*, *24*(4), 309–320.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163.

Kleiman, M. A. R., Ostrom, B., & Cheesman, F. (2007). Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. *Crime & Delinquency*, *53*, 106–132.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *Quarterly Journal of Economics*, *133*, 1.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. ArXiv. https://arxiv.org/pdf/1609.05807.pdf

Koepke, Logan, J., & Robinson, D. G. (2018). *Danger Ahead: Risk Assessment and the Future of Bail Reform*, Retrieved from https://ssrn.com/abstract=3041622 or https://doi.org/10.2139/ssrn.3041622

Lin, M., Lucas, H. C., Jr., & Shmueli, G. (2013). Research commentary—Too big to fail: Large samples and the p-value problem. *Information Systems Research*, *24*, 906–917.

Lowencamp, C. T., Lemke, R., & Latessa, E. (2008). The development and validation of a pretrial screening tool. *Federal Probation*, *72*, 2–9.

Mamalian, C. (2011). *State of the science of pretrial risk assessment*. Washington: Pretrial Justice Institute.

Mayson, S. (2018). Dangerous Defendants. *Yale Law Journal*, *127*, 490–550. https://digitalcommons.law.uga.edu/fac_artchop/1154

Minton, T., & Zeng, Z. (2015). *Jail inmates at midyear 2014*. Washington: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice.

Monahan, J., Skeem, J., & Lowenkamp, C. (2017). Age, risk assessment, and sanctioning: Overestimating the old, underestimating the young. *Law and Human Behavior*, *41*(2), 191–201.

Pretrial Justice Institute. (2017). *The state of pretrial justice in America*. Washington: Author.

Romei, A., & Ruggieri, S. (2013). A multidisciplinary survey on discrimination analysis. *Knowledge Engineering Review*, *29*, 582–638.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*, 215–227.

Sacks, M., & Ackerman, A. R. (2014). Bail and sentencing: Does pretrial detention lead to harsher punishment? *Criminal Justice Policy Review*, *25*(1), 59–77.

Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. *Criminal Justice & Behavior*, *37*, 965–988.

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

Skeem, J., Monahan, J., & Lowenkamp, C. T. (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. (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). Retrieved from http://www.siop.org/_principles/principles.pdf

Spohn, C. C., & Holleran, D. (2000). The imprisonment penalty paid by young, unemployed black and Hispanic male offenders. *Criminology*, *38*, 281–306.

Starr, S. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*, 803–872.

Starr, S. (2015). The new profiling: Why punishing based on poverty and identity is unconstitutional and wrong. *Federal Sentencing Reporter*, *27*, 229–236.

Stevenson, M. (2018). Assessing risk assessment in action. *University of Minnesota Law Review*, 303.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

United States v. City of Erie, 352 F. Supp. 2d 1105, 2005 U.S. Dist. Lexis 33397 (W.D. Pa. 2005).

United States v. Salerno. 481 U.S. 739 (1985).

VanNostrand, M., & Lowenkamp, C. T. (2013). *Assessing pretrial risk without a defendant interview*. Houston: Laura and John Arnold Foundation. https://nicic.gov/assessing-pretrial-risk-without-defendant-interview

Winterfield, L., Coggeshall, M., & Harrell, A. (2003). *Development of an Empirically-Based Risk Assessment Instrument: Final Report*. Washington: Urban Institute Justice Policy Center.

Zeng, Z. (2019). *Jail inmates in 2017*. Washington: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice.

AUTHOR BIOGRAPHIES

Matthew DeMichele is a Senior Research Sociologist in RTI's Applied Justice Research Division. He is the Director of the Center for Courts and Corrections Research and has conducted criminal justice research on correctional population trends, risk prediction, terrorism/extremism prevention, and program evaluation. His research has recently been published in *Crime & Delinquency, American Sociological Review*, and the *Probation Journal*.

Peter Baumgartner is a data scientist focused on the use of machine learning, natural language processing, and algorithms. He has built software to help law enforcement agencies with arrest warrant service and evaluated the impact of the technology on the agency. Mr. Baumgartner completed

a successful pilot study of a novel data collection methodology involving machine learning models that identify arrest related deaths in news articles. He also assists in ongoing work evaluating the effect of pretrial risk assessment tools and their improvement.

Michael Wenger is a data scientist in the Center for Data Science at RTI International. Mr. Wenger uses his expertise in data visualization, data management, and machine learning to help inform decision making in public health, social science, and environmental applications.

Kelle Barrick is a research criminologist in the Center for Policing Research and Investigative Science at RTI International. She focuses broadly on producing empirical research to inform the improvement of processes and programs across the criminal justice system. Her current efforts include enhancing our understanding of and strengthening the response to human trafficking; demonstrating the value and utility of incident-based crime data; and assessing forensic evidence collection, submission, and analysis techniques. Her recent research has been published in *Criminology and Public Policy, Justice Quarterly, Journal of Criminal Justice,* and *Policing: An International Journal*.

Megan Comfort is a Senior Research Sociologist in the Youth, Violence Prevention, and Community Justice Program at RTI International. Her work focuses on the intersection of legal system involvement, family relationships, and public health. She is the author of *Doing Time Together: Love and Family in the Shadow of the Prison* (University of Chicago Press, 2008).

How to cite this article: DeMichele M, Baumgartner P, Wenger M, Barrick K, Comfort M. Public safety assessment: Predictive utility and differential prediction by race in Kentucky. *Criminol Public Policy*. 2020;19:409–431. https://doi.org/10.1111/1745-9133.12481