Validation of the PSA in Harris County, TX

Respectfully Submitted

D. James Greiner
Matthew Stubenberg
Ryan Halen
Access to Justice Lab
Harvard Law School

November 6, 2020

Table of Contents

Executive Summary 4

Introduction 6

I. Harris County, the PSA, and Validation 6

a. Harris County 7

b. The PSA 8

c. Pretrial Processes and PSA in Harris County 9

d. Validation of Risk Assessment Instruments 11

e. Data Available 13

i. Data Sources 13

ii. Data Limits 14

II. Findings 16

a. Outcome Definitions 17

b. Descriptive statistics 17

c. Traditional validation techniques 22

i. PSA scores and failure rates 22

ii. Bivariate correlations 28

iii. Area under the curve 34

d. Techniques Used Outside the Pretrial Context 37

i. Regression 37

ii. Balanced accuracy measures 42

e. Validation by Racial And Gender Groups 53

i. PSA scores and failure rates by race 53

Appendices 64

Appendix A Felony Bail Schedule 64

Appendix B: Risk Based Release Protocol 66

Appendix C: Additional Figures 67

Appendix D: Note on FTA Identification Issues and Calculations 69

Executive Summary

On July 28, 2017, Harris County, TX integrated the Public Safety Assessment ("PSA") and the accompanying Decision Making Framework ("DMF") into its pretrial processes. The PSA is a pretrial risk assessment instrument, a process that uses criminal history factors and age inputs to produce scores that classify an individual's risk of misbehavior if released pretrial. Specifically, the PSA classifies individuals on risk of being arrested or cited for new criminal activity ("NCA") and failure to appear ("FTA") through two 1-6 integer scales, and on risk of new violent criminal activity ("NVCA") through an on-off "flag." The PSA scores are typically accompanied by the DMF, which incorporates the objective information from the PSA with community-specific determinations regarding local policy and values, state statutes, and jurisdictional resources to produce a release recommendation as well as (in locations that choose to use it this way) a supervision level to be imposed if the individual is released. The PSA scores rely on objective data, and the scoring system is the same in all jurisdictions. The DMF recommendation system can be different in each jurisdiction. The decision about whether to release or detain an individual, and the level of supervision accompanying any release, rests always with the magistrate. The PSA was developed with support from Arnold Ventures, a Houston-based philanthropy, to reduce the burden placed on vulnerable populations at the frontend of the criminal justice system.

The Access to Justice ("A2J") Lab was asked to conduct a validation study of the PSA in Harris County. In a validation study of a risk assessment instrument, researchers deploy statistical techniques to assess the strength of the relationship between the instrument's risk categories and the occurrence rates of the outcomes about which the instrument purports to provide classifying information. Other researchers have completed validations studies of the PSA's risk categories, and this report contributes to this body of knowledge.

The A2J Lab analyzed data on Harris' use of the PSA from the Harris County District Clerk, the Harris County Justice Administration Department, and Harris County Pretrial Services. The data addressed PSAs calculated between July 16, 2017 and December 31, 2019.

A top-level summary of the A2J Lab's findings is as follows:

- Despite challenges in the implementation of the FTA scale, as well as challenges in coding FTA outcomes, there was moderate evidence that the PSA was overall valid in Harris County. Some validation techniques (e.g., simple correlations, area under the curve, balanced accuracy) provided weak evidence of validity, others (e.g., simple plots, logistic regression) provided stronger evidence of validity. No technique suggested invalidity.
- There was no substantial evidence to suggest that the PSA scales performed differently for different racial and gender groups. Although some techniques showed statistically significant differences in PSA performance across demographic groups, the differences were substantively small and/or directionally contradictory (i.e., one scale showed higher failure rates for blacks than whites, while another scale showed the opposite).

- For NCA and FTA, 1-level scale increases generally corresponded to similarly sized jumps in failure rates, except for the FTA increase from 5-6, where the evidence available did not allow firm conclusions, a phenomenon likely due to the aforementioned difficulty in classifying prior missed court appearances as FTAs.

The A2J Lab is grateful for the opportunity to work on this project.

Introduction

This report discusses the Access to Justice ("A2J") Lab's findings with respect to the validation study we conducted on the use of the Public Safety Assessment ("PSA") in Harris County, Texas. This report analyzes data with respect to PSA calculations made in Harris for felony and misdemeanor arrests from July 27, 2017 to December 31, 2019, as well as corresponding rates of failure to appear ("FTA"),[1] new criminal activity ("NCA"), and new violent criminal activity ("NVCA") among those released for the time period of July 28, 2017 to January 1, 2020. Validation of risk assessment instruments generally consists of comparing the classifications individuals (as of particular arrest events) receive from an instrument's risk scores to the subsequent rates of the failure events corresponding to the risk scores. Here, the A2J Lab deployed several statistical techniques to compare the scores Harris County assigned to individuals on the PSA's scales to the corresponding FTA, NCA, and NVCA rates, understanding that under Arnold Ventures ("AV") definitions, none of these three failures can occur with respect to individuals while they are incarcerated.

This report proceeds in two parts. Part I addresses Harris County and its experience with the PSA, along with the nature of validation and the data available. Part II describes the A2J Lab's findings.

The A2J Lab is appreciative to Harris County Pretrial Services, the Justice Administration Department, the Harris County Sheriff's Office, the Harris County District Clerk, and the Harris County Courts whose assistance made this report possible.

I. Harris County, the PSA, and Validation

This Part provides the background needed to understand the findings in Part II. It consists of five sections. Section A describes Harris County, including a brief discussion of its criminal justice system as well as the impact of Hurricane Harvey from late August of 2017 to the present. Section B briefly describes the PSA. Section C discusses the implementation of the PSA in Harris County. Section D discusses the validation of risk assessment instruments as applied to Harris County's deployment of the PSA, including limits inherent in the validation of any pretrial risk assessment instrument ("PRAI"). Section E describes the available data.

[1] As discussed below, Harris County experienced difficulty in ascertaining when the presence of a warrant in an individual's record indicated an FTA or something else, such as a rescheduled court date, or a new arrest or charge. Moreover, Harris judges issued different types of warrants, and across judges a particular warrant type might be used in different ways. As a result, measuring FTA involved some inferences. Harris County officials assisted the A2J Lab in defining the circumstances under which the Lab might conclude that a particular combination of warrant type and values of other variables might be counted as an FTA, and using that information, the Lab constructed the two FTA measurements discussed below. Future investigation, likely conducted by researchers other than those at the A2J Lab, may uncover other, potentially more accurate, methods to infer FTA counts based on available data.

a. Harris County

This section briefly describes Harris County, its demographics, and its experiences with the aftermath of Hurricane Harvey.

Harris County Overview

Harris County is the most populous county in Texas with almost five million people.[2] The county is home to the city of Houston, Texas' most populous city and the fourth most populous city in the United States.[3] The County is racially diverse with approximately 40% Hispanic, 30% white, and 20% black residents.[4]

Hurricane Harvey

On August 25, 2017, less than a month after Harris County implemented the PSA,[5] Hurricane Harvey arrived.[6] The hurricane affected almost every aspect of the state, impacting a third of the state's population, killing 94 people, and causing billions of dollars in damage.[7] The hurricane's impact on Harris County in particular was devastating, as it dumped more than four feet of water on Houston and submerged almost a third of the county.[8] It took years for Harris County to return to approximate normalcy, and some sectors still remain affected.[9] The impact on the criminal justice system was severe. The main Courthouse in Houston was significantly damaged and did not reopen for almost a year.[10] After reopening, the courthouse operated at reduced capacity for a significant period of time.[11]

[2] U.S. Census Bureau (2020). *QuickFacts Harris County, Texas.* Retrieved from https://www.census.gov/quickfacts/harriscountytexas (last visited Sept. 4, 2020).

[3] U.S. Census Bureau (2020). *Annual Estimates of the Resident Population….* Retrieved from https://www2.census.gov/programs-surveys/popest/tables/2010-2019/cities/totals/SUB-IP-EST2019-ANNRNK.xlsx. (last visited Sept. 4, 2020).

[4] U.S. Census Bureau (2020). *QuickFacts Harris County, Texas.* Supra note 2.

[5] Matthew Stubenberg, Memo, "Harris County Pretrial Memo," Memorializing Conversation on July 31, 2020 (on file with the Access to Justice Lab).

[6] Parraga, Marianna. "Funding Battle Looms as Texas Sees Harvey Damage at up to $180 Billion." *Reuters*, Thomson Reuters, 4 Sept. 2017, www.reuters.com/article/us-storm-harvey/texas-governor-says-harvey-damage-could-reach-180-billion-idUSKCN1BE0TL.

[7] Texas Department of State Health Services Hurricane Harvey Response After-Action Report, May 30, 2018, https://sk75w2kudjd3fv2xs2cvymrg-wpengine.netdna-ssl.com/wp-content/uploads/2018/08/Texas-DSHS-Hurricane-Harvey-AAR-FINAL.pdf.

[8] Kevin Sullivan, Arelis R. Hernandez and David A. Fahrenthold. *Harvey Leaving Record Rainfall, at Least 22 Deaths behind in Houston*. 30 Aug. 2017, www.chicagotribune.com/nation-world/ct-hurricane-harvey-flooding-houston-20170829-story.html.

[9] Stubenberg Memo, supra note 5.

[10] *Harris County Reopens Part Of Courthouse After Sustaining Damage From Hurricane Harvey*, Jun. 4, 2018. https://www.houstonpublicmedia.org/articles/news/2018/06/04/289086/harris-county-courthouse-reopens-after-sustaining-damage-from-hurricane-harvey/

[11] Stubenberg, supra note 5.

b. The PSA

This section briefly describes the PSA for persons unfamiliar with its operation.

The PSA is a PRAI that magistrates may use when deciding whether to release or detain an individual before trial. The PSA takes as inputs data on the individual's criminal history, current charge, and age. These inputs (some in combination) are assigned an initial set of integer weights. Those integer weights are further processed to produce two risk scores that can take on values of 1-6, with higher numbers signaling higher risk. The first score classifies individuals on risk of being arrested or cited for new criminal activity ("NCA") if released pending disposition. The second 1-6 scale classifies individuals on risk of failure to appear ("FTA") at the case's court hearings. The PSA also flags individuals to signal an elevated risk of being arrested for new violent criminal activity ("NVCA") before disposition; the flag operates as a 0-1 variable.[12]

The PSA was developed with support from Arnold Ventures, a Houston-based philanthropy, to reduce the burden placed on vulnerable populations at the frontend of the criminal justice system.[13] AV and the developing researchers sought to construct a PRAI that (i) did not require inputs from an expensive and legally concerning interview with the individual, and (ii) produced risk categories informative in any jurisdiction in the United States. Validation studies, in which researchers assess whether the PSA's risk categories correspond to differences in released individuals' misbehavior rates, have been completed in several other jurisdictions,[14] and this report contributes to that body of research.

The PSA scores are typically accompanied by the Decision Making Framework ("DMF"), which incorporates the objective information from the PSA with community-specific determinations regarding local policy and values, state statutes, and jurisdictional resources to produce a release recommendation as well as (in locations that choose to use it this way) a supervision level to be imposed if the individual is released. The PSA scores rely on objective data, and the scoring system is the same in all jurisdictions. The DMF recommendation system can be different in each jurisdiction. The decision about whether to release or detain an individual, and the level of supervision accompanying any release, rests always with the magistrate. The PSA does not produce a recommendation, and the DMF's recommendation is not binding.

This validation report focuses on the PSA scores and the corresponding failure rates. It does not examine the Harris County DMF.

[12] A complete discussion of the PSA's inputs, initial integer weights, and processing of those weights into 1-6 FTA and NCA risk categories is available at https://advancingpretrial.org/psa/factors/ (last visited Sept. 4, 2020).

[13] Support for the assertions in this paragraph appear in https://www.psapretrial.org/about/background (last visited Feb. 19, 2020), which provides a more detailed discussion of the PSA's features and development, as well as links for additional information.

[14] The Access to Justice Lab is currently pursuing validation efforts in three other counties.

Dozens of jurisdictions have implemented the PSA-DMF System, including three entire states and several large cities.[15]

c. Pretrial Processes and PSA in Harris County

This section discusses Pretrial Processes and the PSA in Harris County. It briefly describes the PSA's implementation history, including the classes of arrests to which the PSA was applied.

The PSA-DMF System was implemented in Harris County on July 28th, 2017 for all felonies and class A/B misdemeanors.[16] At this time, when a defendant was arrested in Harris, the police officers consulted with the DA's Office who tentatively signed off on the charges. The defendant was brought to a local holding facility or the Inmate Processing Center.[17] After the construction of the Joint Processing Center in early 2019 all defendants were brought there. At this point, a staffer for Harris County Pretrial Services ("Pretrial") manually calculated a preliminary set of PSA calculations accompanied by a set of recommendations stemming from the corresponding DMF.[18] Shortly after, the DA finalized the charges and Pretrial finalized its PSA-DMF-System Report.

Prior to February 2019, individuals charged with one of a handful of particular charges[19] had to be seen by a magistrate at a 15.17 bail hearing, there being no opportunity to be released on bond according to the bail schedule for cases involving such charges. For individuals otherwise charged, Pretrial used the PSA-DMF scores and recommendation, in conjunction with the severity of the charge according to a bail schedule, to determine the individual's bond amount, if any, as follows.[20] In felony cases, if the Report detailed an FTA or NCA score of 6 or if the NVCA Flag was triggered, the individual could not bail out according to the bail schedule before seeing a magistrate for a bail hearing.[21] In misdemeanor cases, if the FTA or NCA score was a five or a six or the NVCA flag was triggered, there was no presumption of personal bond. If the PSA-DMF System scores were below those thresholds and the individual was not held on an ineligible charge, the individual had the option of paying a bond and being released. The bond amount depended on the severity of the charge as well as the PSA-DMF System score. The

[15] See https://advancingpretrial.org/psa/about/#jurisdictions-united-states (last visited Sept. 4, 2020).

[16] Matthew Stubenberg, Memo, "Harris County Pretrial Memo," Memorializing Conversation on August 11th, 2019 (on file with the Access to Justice Lab).

[17] Matthew Stubenberg, Memo, "Comments Made to Draft Report," Memorializing comments made to the draft report. (on file with the Access to Justice Lab).

[18] Id.

[19] The list of charges can be seen in Felony Bail Schedule which appears as Appendix A. They include a capital felony, any first degree felony, if the defendant was on bail for any felony charge, unlawful possession of a firearm by a felon, etc.

[20] Stubenberg Memo, supra note 14.

[21] Note that, as described below, only an extremely small fraction of individuals received either an FTA score or an NCA score of 6.

bond amount generally ranged from a personal release bond to $50,000, depending on the charge severity and the PSA-DMF scores.

For individuals who did not bond out, a hearing was held, and magistrates had access to the PSA-DMF System Report in deciding whether each defendant should be held until trial, be released on a personal release bond, be released with conditions, or have the bond amount reduced. In the vast majority of these hearings, magistrates followed the recommendation of the bail schedule.[22]

Pretrial and the Harris County Community Supervision and Corrections Department ("Probation") also used the PSA-DMF System Report to determine the level of oversight for released individuals.[23] The process was mechanical and formulaic, and did not vary with an individual individual's particular circumstances.[24]

O'Donnell Lawsuit

In May of 2016, Maranda O'Donnell, who in 2016 had been arrested on a misdemeanor charge and held because she could not afford the $2500 bond mandated by the bail schedule, brought an action in federal district court alleging that the Harris County's pretrial release process was unconstitutional.[25] The complaint alleged the strict adherence to the bail schedule, which was based only on offense without regard to ability to pay or risk of flight, was unconstitutional.[26]

The federal district court certified the class and granted a preliminary injunction. Subsequently, the parties jointly submitted a new set of procedures to govern the pretrial release making process in Harris County, subsequently encapsulated in Local Rule 9. Harris County implemented a modified Local Rule 9.1 in February of 2019, which made changes to the misdemeanor pretrial release and detention processes. The default for individuals arrested on a misdemeanor became released on a personal (*i.e.*, unsecured) bond as soon as practicable after arrest unless the individual was charged with a narrow range of offenses. Pretrial ceased preparing PSA-DMF System Reports for misdemeanor cases. Public defenders represented individuals at 15.17 hearings, and the magistrate considered information about an individual's ability to pay before imposing monetary bail.[27]

The lawsuit concluded with a consent decree issued on November 21st, 2019, which required Harris County to implement additional changes to its misdemeanor pretrial release process and

[22] Stubenberg Memo, supra note 14.

[23] At the time of this writing we are attempting to confirm the extent of Probation's use of the PSA-DMF System Report. A copy of the Misdemeanor Supervision Schedule prior to Local Rule 9.1. appears as Appendix B.

[24] Stubenberg Memo, supra note 14.

[25] ODonnell v. Harris Cty., 892 F.3d 147 (5th Cir. 2018).

[26] Complaint at 1. ODonnell v. Harris County, Texas et al, 4:16CV01414. https://www.clearinghouse.net/chDocs/public/CJ-TX-0010-0002.pdf.

[27] See Local Rule 9.1 section 2. A copy can be found at https://hccla.org/wp-content/multiverso-files/829_56990d05d6719/AdminOrder-MISD.pdf.

mandated improved data collection and transparency.[28] As noted above, one result of the O'Donnell lawsuit was more defendants with misdemeanor cases were automatically released before a 15.17 hearing negating the need for a PSA-DMF System Report. Therefore, pretrial halted calculation of the PSA-DMF System for all misdemeanors on February 16th, 2019.[29]

Hurricane Harvey Impact on PSA Outcomes

As noted above, Hurricane Harvey's effect on the Harris County criminal justice system was severe. Many people lost their homes or moved in with friends and family. Even after the courthouse resumed operation, there was difficulty in notifying defendants of dates and locations of rescheduled hearings, and the extent of the damage made transportation difficult, rendering some defendants unable to attend hearings. It became difficult for judges to distinguish between defendants who did not appear because of the hurricane and those who did not appear in an attempt to avoid prosecution. The hurricane likely also affected the new criminal activity rate, given the state resources consumed by response.

Pre-hurricane data systems also posed challenges. There was no dedicated field documenting FTA for judges.[30] The issuance of a bond forfeiture warrant exclusively indicated an FTA..[31] Other types of warrants might also have indicated an FTA, but might also have corresponded to other offenses such as a violation of a bond..[32] Warrant practices also varied from judge to judge, making FTA difficult to determine.[33]

d. Validation of Risk Assessment Instruments

This section discusses the validation of risk assessment instruments, what validation does and does not do, and the limits of validation techniques.

As noted above, validation studies focusing on the PSA have been completed in several jurisdictions. These studies have generally found that the PSA is valid under the techniques they used, although they have noted challenges with the data available in each jurisdiction.[34]

[28] A copy of the Consent Decree can be found on the Harris County Justice Administration Department's website at http://jad.harriscountytx.gov/Portals/70/documents/Consent%20Decree%2011.21.19-2.pdf?ver=2020-01-28-125545-333

[29] Matthew Stubenberg, Memo, "Harris County Pretrial Memo," Memorializing Conversation on August 11th, 2019 (on file with the Access to Justice Lab).

[30] Matthew Stubenberg, Memo, "Harris County Pretrial Memo," Memorializing Conversation on July 31, 2020 (on file with the Access to Justice Lab).

[31] Id.

[32] Id.

[33] Id.

[34] See, e.g., DeMichele, M, Baumgartner, P, Wenger, M, Barrick, K, Comfort, M. Public safety assessment: Predictive utility and differential prediction by race in Kentucky. Criminal Public Policy. 2020; 19: 409– 431.; DeMichele, Matthew DeMichele, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. "The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky." Available at SSRN 3168452 (2018).

Ordinarily, the finding of validity meant that individuals classified into higher PSA risk categories and who were released subsequently "failed," meaning they experienced FTA or NCA or NVCA under applicable definitions, at higher rates than individuals classified into lower PSA risk categories who were subsequently released. This report deploys other measurement techniques addressing whether the instrument's classifications correspond to the frequency of the outcomes upon which the instrument focuses. Part II describes these more complicated techniques.

All validation techniques share certain limits. First, validation provides no information on whether a jurisdiction is better or worse off using a risk assessment instrument as opposed to not using one. An instrument might be valid as measured by various statistical techniques, but its classifications might not correspond to a community values, or magistrates who access its classifications might not use them well (or at all), or judicial decisions informed by the instrument's classifications may not be markedly different from those made without such information, or a community might react unfavorably to the instrument for reasons apart from its validity. These and other questions must be answered to determine whether a community experiences the adoption of a risk assessment instrument positively. Some of these questions can be answered with a well-run randomized control trial ("RCT"); the A2J Lab did not conduct an RCT in Harris County.[35]

Second, validation of PRAIs in particular, and of most risk assessments in general, is limited by the fact that if the instrument classifies cases well, and if decision makers use the instrument's classifications well, the data observed could make it appear that the instrument classifies poorly. The reason is that when a valid instrument accurately classifies a case as presenting a high risk of failure, and a decision maker reacts to that classification by taking aggressive action to prevent failure, the aggressive action often does what it was designed to do, *i.e.*, reduces or eliminates the chance of failure. In the case of a PRAI such as the PSA, a high risk score, along with other available information, could make it more likely that a magistrate incarcerates an individual, which would then eliminate (or greatly reduce) the possibility of an FTA or N(V)CA.

Despite this fact, the validation study we report here, like all previous PRAI validation studies of which we are aware, analyzes only the failure rates of released individuals; we are unaware of established and principled statistical techniques that would allow us to do otherwise. The result is that if the PSA classifies individuals well, and if Harris County magistrates react to that classification by incarcerating a greater fraction of high-risk individuals, then more high-risk individuals were effectively removed from the data that the A2J Lab used for this validation, potentially culling all but the (comparatively) less risky individuals within the high-risk category. Particularly when, as could be true in first appearance hearings, the magistrate has access to information other than the PSA that helps the magistrate classify the individual's risk of misbehavior, this fact could make the PSA appear less valid than it actually is.

[35] With AV's support, the A2J Lab is pursuing RCTs in four jurisdictions in the United States.

Third, some of the off-the-shelf statistical techniques used in previous PRAI validation studies and deployed below have difficulty assessing the validity of risk assessments as applied to rare events. This is a common problem with classification techniques generally in statistics and related fields, such as epidemiology. The problem is well-understood but nevertheless difficult to solve. It may affect some of the NVCA results discussed in Part II.

e. Data Available

i. Data Sources

This subsection describes the sources of the data comprising the analysis dataset. The data used in the analysis originated from three primary sources:

1) Court data provided by the Harris County District Clerk,
2) Jail data provided by the Harris County Justice Administration Department, and
3) PSA data provided by Harris County Pretrial.

The A2J Lab did not have direct access to any of these data sources and relied on the three sources identified above to write queries to pull only the data pertinent to the validation study. A2J Lab was able to combine the three data sources using a combination of common identifiers, as follows.

Pretrial provided data on 167,299 PSA instances assessed between July 16, 2017 and December 31, 2019. Of these instances, 127 were duplicates or amended PSAs, resulting in 167,172 unique PSAs. Each PSA was attached to a unique DA Log Number, the only identifier linking the PSA data to the court data. Of the initial 167,172 PSAs attached to unique DA Log Numbers, 30,447 do not appear in the associated court data. We understand that this is due either to dismissals by the DA or to record expungements. Either way, there were no further records to analyze in these cases, and we dropped from the analysis dataset.

The remaining 136,852 entries represented PSAs with attached case data, including disposition dates. The court data provided a second necessary identifier in the form of a unique Case Number, which allowed joining the FTA information from the court data to booking information from the jail data. The jail data included release dates and release reasons. Using this information a number of additional entries were excluded from the analysis dataset, including cases in which individuals were released into the authority of another agency or treatment program, and cases in which release dates coincided with case disposition dates. In the first set, a lack of access to other institutional data meant that the A2J Lab was unable to trace incarceration at other institutions and therefore could not assess what if any post-arrest time the individual spent released from custody. In the second set, cases that were disposed of on the same date as, or prior to, release from jail, there was no pretrial period, and thus no possibility of failure events under AV definitions. These exclusions resulted in an analysis data set of

61,603 entries. Each entry was a single PSA assessment attached to a specific case that featured at least one day of pretrial release.

The A2J Lab pulled the following information from the three sources identified above.

1) Court Data: The court data provided by the District Clerk contained information related to the final charges, disposition, disposition date, and whether a bench warrant was issued.
2) Jail Data: The data provided by the Jail included when an individual was booked into and released from the jail as well as the charges at arrest. The A2J Lab used this dataset to help the NCA/N(V)CA rate.
3) Pretrial Data: The data provided by Pretrial Services contained the PSA/DMF system inputs and PSA/DMF system scores necessary for calculating the PSA/DMF system report. This dataset also contained timestamp information related to when the PSA calculation process began and ended.

The Pretrial data provided PSA inputs and outputs as well as the starting point for integrating and joining more data to each PSA instance. The court data provided case disposition information, including dates, which, when combined with the PSA assessment date, provided the date range of the pretrial window; the jail data provided the dates in this range during which an individual was released. The court data additionally provided separate FTA instances with attendant dates. If an FTA instance occurred during a relevant date range, we counted it as an observed FTA failure. If a subsequent PSA entry was created during the relevant date range, it was counted as an NCA instance.[36] An additional PSA input field for current violent offence indicated whether the charge that initiated the PSA was violent. Combining this information in the same manner as the NCA information produced NVCA outcomes.

ii. Data Limits

The data received was limited to records provided by the Harris County departments identified previously. Harris County officials and the A2J Lab explored the possibility of obtaining statewide arrest data from the Department of Public Safety (DPS). The A2J Lab continues to pursue this option.

As noted above, another data limit was the absence of a field dedicated recording FTA. While a field does exist in the Court's database to collect the attendance of different parties in the

[36] Calculating NCA and NVCA from the PSA data was done prior to any filtering processes, i.e. matches were made on the full pretrial PSA assessment list. PSAs are generated at jail booking and persist regardless of how quickly the resulting case is disposed, meaning that all potential jailable arrests are captured by the full PSA assessment list. The list of PSA events were additionally checked against the jail entry data and no additional case numbers were present in the jail data.

Note that PSAs were generated post-arrest, regardless of when the charged offense occurred. This approach did not distinguish offenses that occurred during the release period (which were NCAs for PSA purposes) from those that occurred prior to the release period (which were not NCAs for PSA purposes). Comment by Dennis Potts on Report Draft (September 3, 2020). The A2J Lab has requested data on dates of alleged offenses so as to address this problem.

Courtroom, the field was not reliable as it was populated only when Court staff had the opportunity to do so, and usage of the field varied between courtrooms.[37] The A2J Lab inferred when an FTA occurred based on whether a warrant was issued. There were a number of possible warrant types a judge could issue; however, only on one type of warrant, a bond forfeiture warrant, exclusively[38] indicated with an FTA. According to Harris County officials, other warrant types, including an Order of the Court warrant, a Bond Surrender warrant, or an Alias Issued warrant, were possible, and these may have indicated an FTA but were not conclusive. Staff for the Harris County District Clerk had estimated 50 to 60% of FTAs resulted in a bond forfeiture warrant with the other 40 to 50% resulting in one of the other warrant types.[39] These facts create tension between the accuracy and the completeness of any FTA outcome measure. Relying on only bond forfeiture creates the most accurate FTA measure as all bond forfeiture warrants are FTAs; however, relying on all warrant types creates a more complete FTA measure at the cost of accuracy. We utilize two separate FTA measures to address this tension. The first, Base FTA, calculates FTA only based on Bond Forfeiture warrants, while the second, FTA+, calculates FTA on the basis of all warrant types.[40] As part of the O'Donnell consent decree, the Harris County Court system will start to capture a specific FTA field.

The problems involved with identifying FTA instances also affected PSA generation. Prior FTA events are an input into PSA score calculation for both the FTA and NCA scales, although more heavily weighted in FTA calculation. Pretrial reported reluctance to conclude that a prior missed court date plus a warrant constituted a prior FTA. Pretrial also could not always tell whether a defendant had missed a court date, in part because missed court dates sometimes resulted in case resets as opposed to warrants.[41] The results of this reluctance are most evident at the upper end of the FTA scale in that a vanishingly small number of arrest events resulted in FTA scores of 5 or (especially) 6. Of the 61,603 PSA instances that made up the analysis dataset, only 153 cases (.25%) received an FTA score of 6. Similarly, because an NCA score of 6 was impossible without positive identification of FTAs in the past 2 years, only 1427 cases (2.3%) were assessed at an NCA Risk Score of 6.

Pretrial's (understandable) reluctance to conclude that a past warrant corresponded to an FTA was most visible at the upper ends of the FTA and NCA scores, but its effect extended to all score ranges. The presence or absence of prior FTAs also made a difference in whether an individual received a 1 or a 2, or a 2 or a 3, etc. As we discuss below, given the challenges inherent in the data, it is surprising that the PSA validates as well as it does.

[37] Matthew Stubenberg, Memo, "Harris County Court Memo," Memorializing Conversation on September 25, 2020 (on file with the Access to Justice Lab).
[38] For instance, a defendant may be in court but a Judge may order their bond revoked because of new information.
[39] This was an anecdotal observation not a scientific measure. See Email from LaShanda to Matthew Stubenberg Aug. 19, 2020 (on file).
[40] See Appendix Section D for a detailed description on the processes used to generate each FTA metric.
[41] Comment by Dennis Potts on prior draft of Report (September 3, 2020).

II. Findings

The logic of validating an assessment tool or instrument is clearest in the context of binary classification models, in which an algorithm translates data into one of two classifications, (i) high risk of an event's occurrence, or (ii) low risk of an event's occurrence. In this kind of binary risk classification, the two categories map directly onto two observed outcome categories (event occurred versus event did not occur). Validating a binary instrument means comparing these outcomes to the classifications. In the context of criminal justice, for example, a binary classification algorithm might attempt to classify risk of new criminal activity during the pretrial period. This set up generates two potential prediction categories: a positive classification (high risk that an NCA will be observed) and a negative classification (low risk that an NCA will be observed)

A conclusion that a tool is valid, at least partially, indicates that its classifications provided information concerning the relative occurrence of outcomes beyond the information available without the tool (or as measured against some other standard, such as a random 50/50 guess). Most standard validation metrics assume that the instrument consists of this kind of binary classification. Moreover, most instruments classify risk with respect to only one outcome.

The PSA is different, and those differences pose challenges. First, the PSA's FTA and NCA scores consist not of binary values but of 1-6 scales. Second, the PSA classifies with respect to three outcomes: FTA, NCA, and NVCA, with NVCA different from the first two in that it *is* on a binary (0-1) scale.

One response to these challenges is simple: compare the failure rates to the risk scores to see if the two tend to increase (or decrease) together, according to some statistical model. We implement this approach below. That is our first validation framework, and we label it "overall validity.".

The PSA's complexity allows for (or necessitates) other approaches, however, that we also pursue as well with respect to the FTA and NCA scales. For our second validation framework, which we label "uniform validity," we examine whether steps up from a lower to the next higher score correspond to roughly same increase in failure rates, *i.e.* whether the increase in risk when moving from a score of 1 to 2 is roughly the same as moving from a 3 to a 4. This framework provides information potentially useful to magistrates and practitioners, who might wish to know whether step increases signal equivalent risk increases.

Third, we assess what we label "equitable validity," which concerns whether the PSA validates equally for different subgroups defined by, for example, race and gender.

The remainder of the section proceeds in five subsections Subsection A provides rigorous definitions of FTA, NCA, and NVCA. Subsection B provides descriptive statistics. Subsection C provides the results of techniques traditionally used in the PRAI validation literature. Subsection

D provides the results of techniques used to validate risk assessment instruments outside of the pretrial context. Subsection E provides results of our validation by demographic group.

a. Outcome Definitions

We analyze the NCA, NVCA, and FTA scales separately.

NCA- An NCA event is observed if a new arrest event, with an associated charge that carries the potential of incarceration as a sentence, is observed during a case's pretrial period, *i.e.*, from the initial bail hearing until case disposition. (See above regarding date of offense.)

NVCA- An NVCA event is observed if a new arrest event, with an associated charge that carries the potential of incarceration and is considered a violent charge, is observed during a case's pretrial period. Within Harris County different agencies maintain different lists of violent charges that vary slightly. As part of the calculation of the PSA, pretrial services maintains its own distinct list of violent charges that was constructed when the PSA was adopted in Harris County. We use this list. (See above regarding date of offense.)

Base FTA- An FTA event is observed if the court records indicate a missed court event during a case's pretrial period that resulted in the issuance of a bench warrant. This event must be attached to the original PSA case number, *i.e.*, it occurred in the case from which the PSA originated. As noted above, given the difficulties in identifying FTA instances from court data in Harris County, this definition of FTA focuses only on warrants associated with bond forfeiture, which are the only warrants that always indicate an FTA.

FTA+- An FTA event is observed if the court records indicate a missed court event during a case's pretrial period that resulted in the issuance of a bench warrant. This event must be attached to the original PSA case number, *i.e.*, it occurred in the case from which the PSA originated. As noted above, given the difficulties in identifying FTA instances from court data in Harris County, this definition of FTA focuses on all warrant types that could potentially be an FTA; however, not all instances of these warrants indicate an FTA. Instead, dependent on warrant type, we used timing data related to the event date, filing date, execution date, and return date of the warrant to identify which warrant instances are potentially FTAs and which instances are not.[42]

b. Descriptive statistics

The study population consists of 61,603 unique PSA submissions that resulted in charges being filed in a case where the individual was released for at least one full day of their pretrial period. These cases represent 53,805 unique individuals charged with either misdemeanors or felonies over the period of July 27, 2017 to December 31, 2019. individuals were recorded with six

[42] For a more detailed description of the construction of both FTA outcome metrics, see Appendix Section D.

separate racial category identifiers; however, less than 5% were not categorized as either Black or White. For the purposes of readability, we condensed the original race categories to three: Black, White, and Other. For the purposes of analyses concerning equitability validity, we used only individuals categorized as either Black or White. The distribution of individual race was fairly even between Black and White (Figure 1). There were only about 3800, or 14%, more White individuals than Black individuals in the analysis dataset. In terms of gender distribution (Figure 2), 76.4% of the unique PSA assessments in the analysis dataset attached to male individuals. The age distribution tends young, with a mean age of 33.3 years old at time of arrest and a median age of 31 years old at time of arrest (Figure 3). Table 1 provides a brief summary of total PSAs, number of arrestees with at least 1 day of pretrial release, and failure rates for all PSA outcomes. These statistics are reported for both the overall sample as well as for each demographic group (Black arrestees, White arrestees, female arrestees, and male arrestees). Overall, roughly 57% of all PSA instances had an arrestee observe at least 1 day of pretrial release (the other 43% of PSA instances either had arrestees remain incarcerated during the entire pretrial period, or the relevant case was disposed of on the same day as the initial hearing). White arrestees had marginally but statistically significantly higher rates of pretrial release than Black arrestees (0.578 vs. 0.576), while female arrestees had statistically and substantively significantly higher rates of pretrial release than male arrestees (0.66 vs. 0.55). Differences in failure rates are analyzed in further detail later in the report, but overall group differences were significant, with Black arrestees observing higher overall NCA and NVCA rates but lower FTA and FTA+ rates than White arrestees, while female arrestees observed lower failure rates across all outcomes than male arrestees. Overall, 17.4% and 18.8% of the study population observed either an NCA or FTA event, respectively. Only 4.1% of PSA instances observed an NVCA failure during the relevant pretrial release period. Using the expanded definition of FTA+ increases the observed failure rate to 27.2%, overall.

Figure 1: Distribution of individual Race

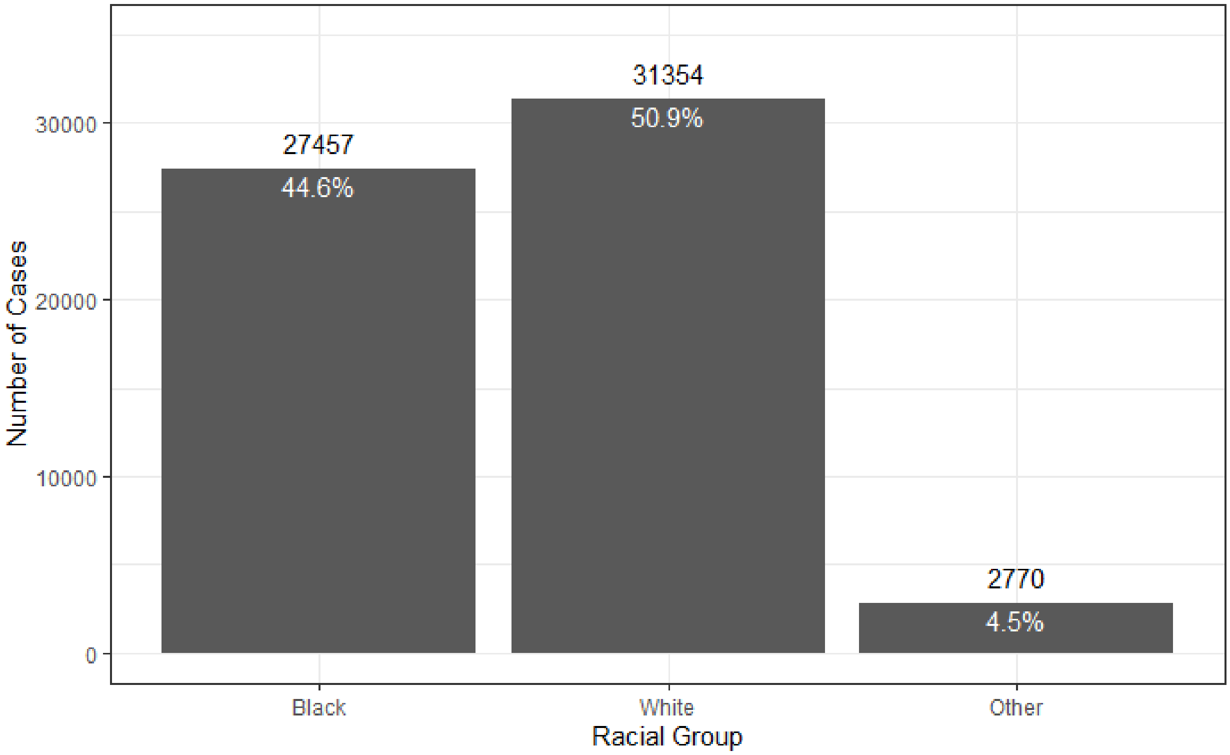


Figure 1 displays the distribution of racial categories for individuals. Each portion of the chart indicates the percentage of unique PSA submissions that listed the relevant Race category for the individual. The initial data obtained from Harris County contained six separate racial categories; however, two of the categories, White and Black, represented 95% of all cases. White individuals were the modal category, representing the majority of individuals who received a PSA assessment at about 51% of the study population.

Figure 2: Distribution of individual Gender

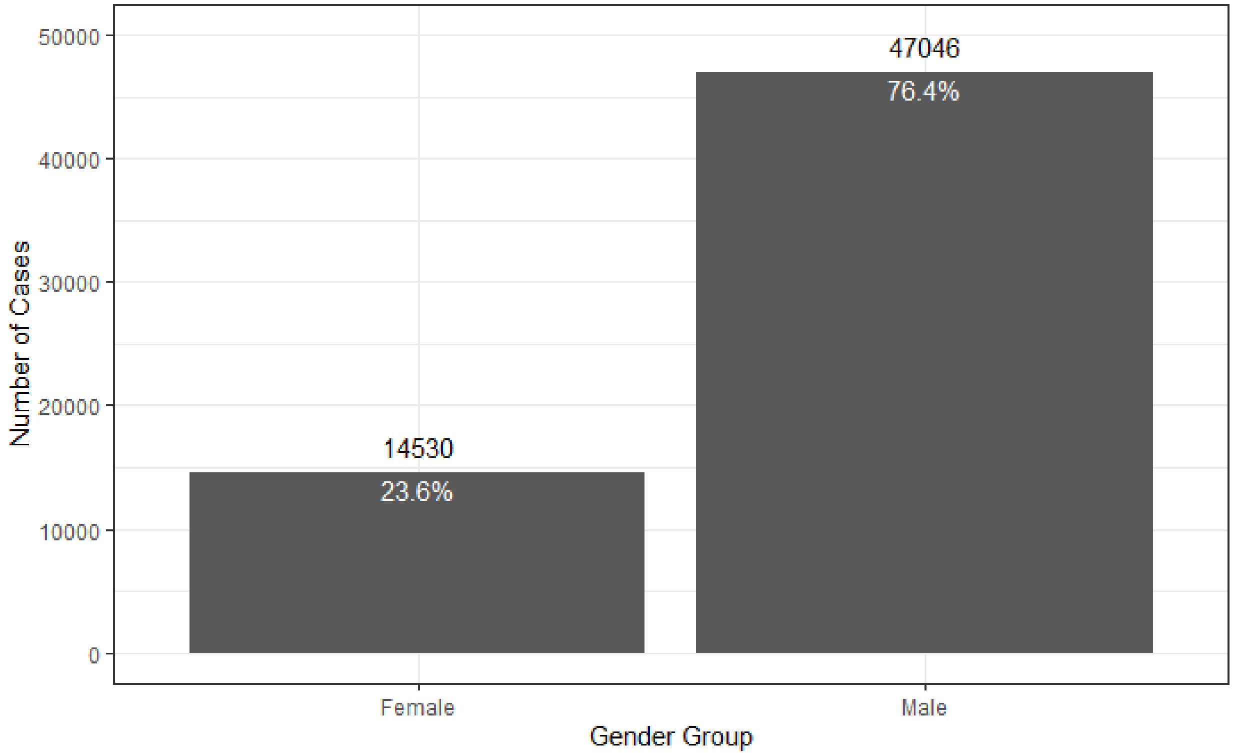


Figure 2 displays the distribution of gender categories within the study population. Female individuals represent just under a quarter of the total study population at 23.6%, which makes Male individuals the overwhelming majority of individuals with a PSA assessment. Out of the total study population of 61,576, there were 14,530 PSA assessments with a female individual and 47,046 PSA assessments with a male individual.

Figure 3: Distribution of individual Age

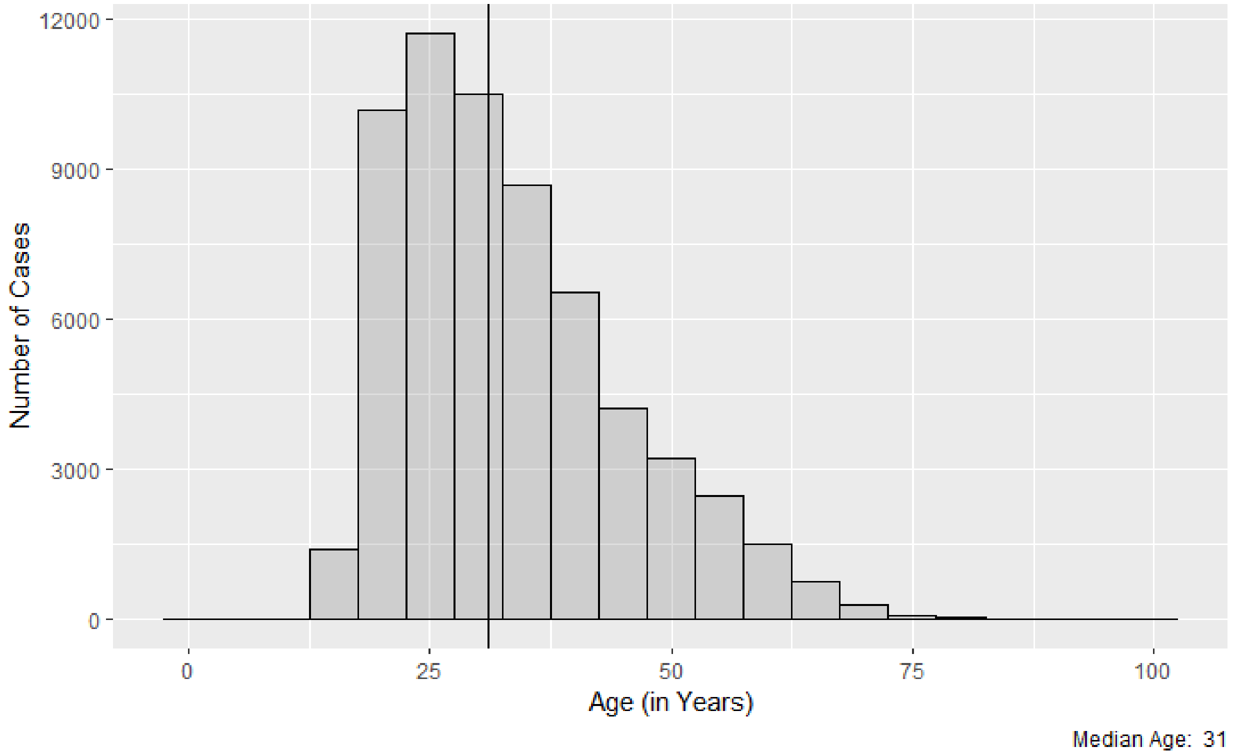


Figure 3 plots the distribution of individual age within the study population. This indicates a higher fraction of younger individuals. The mean age was slightly above 33 years old, with a median of 31 years old. This means that half of the study population fell within a 14-year age range, from 17-31, while the rest occupied a 54-year age range, from 31 to 85.

Table 1: Summary of Failure Rates by Demographic Group

Arrestees	# of PSAs	Released (N)	NCA Fail Rate	NVCA Fail Rate	FTA Fail Rate	FTA+ Fail Rate
Overall	107488	61603	0.174	0.041	0.188	0.272
Black	53335	30249	0.184	0.045	0.172	0.243
White	54198	31376	0.152	0.034	0.192	0.282
Female	22042	14557	0.135	0.027	0.173	0.253
Male	85493	47073	0.185	0.045	0.193	0.278

Table 1 reports total PSA counts, number of released arrestees (which is the study population), and failure rates for each of the main outcomes. Release rates differ significantly between paired demographic groups (Black arrestees/White arrestees and female arrestees/male arrestees) at the p<0.001 level. White arrestees had at least 1 day of pretrial release at a rate about 1% higher than Black arrestees (57.8% vs. 56.7%), while Female arrestees had at least 1 day of pretrial release at a rate about 11% higher than Male arrestees (66% vs. 55%). Likewise, all reported failure rates are significantly different across paired demographic groups. Black arrestees observed slightly higher NCA and NVCA rates but lower FTA and FTA+ rates than their White peers, while female arrestees observed lower NCA, NVCA, FTA, and FTA+ rates than their male peers.

c. Traditional validation techniques

This subsection provides the results of validation techniques traditionally used in the literature on PRAIs. Subsection 1 shows a raw comparison of PSA scores and failure rates. Subsection 2 discusses bivariate comparisons. Subsection 3 discusses the results of an area under the curve analysis. The PSA achieves overall validity with respect to commonly used benchmarks.

i. PSA scores and failure rates

This subsection reports the results of simple comparisons of failure rates across risk assessment score categories. This analysis provides easily interpretable and strong evidence that all three PSA are overall valid, strong evidence that the NCA scale is uniformly valid, and mixed evidence that the FTA scale is uniformly valid. Key details are as follows.

- N(V)CA and both FTA measures show consistent increases in failure rates as scores increase, with the exception of FTA scores of 5-6. These results provide strong evidence that all three PSA scales are, for the most part, overall valid according to this statistical technique.
- NCA failure rate increases across scores do not differ significantly, i.e. a failure rate differences between NCA scores of 1 and 2 are statistically similar to differences between NCA scores of 3 and 4. This fact provides evidence the NCA scale is uniformly valid under this technique.
- FTA failure rates at scores of 2 and 3 increase statistically significantly less than other FTA failure rate increases. Thus, the FTA scale appears uniformly valid with respect to most but not all of the step increases under this technique.

The failure rate for an event, be it a N(V)CA or an FTA, is defined as the proportion of cases that observed at least one of the relevant events during the appropriate time frame. The goal of the failure rate analysis is to assess whether there are statistically significant differences in the rate of failures across consecutive levels of the relevant risk score scale.[43] We use a difference

[43] For studies that adopt this approach in whole or part, see:

- DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. "The public safety assessment: A re-validation and assessment of predictive of proportions test between the consecutive comparison categories, *i.e.*, comparing failure rates for NCA score 1 to NCA score 2, 2 to 3, etc. These comparisons provide information on both overall and uniform validity. For the PSA to validate overall, each pairwise score comparison (1-2, 2-3, 3-4, 4-5, 5-6 for NCA/FTA and No-Yes for NVCA) should have significantly different failure rates, with the higher score category having a higher rate. For the PSA to validate under the uniform validation framework, the magnitude of the differences in failure rates between each paired score comparison should not differ significantly for either the NCA or FTA risk score scale. The following figures plot the overall failure rates for each relevant PSA Risk Assessment score across each of the three outcome events: N(V)CA/FTA. They show that under this definition, the PSA is overall valid with the exception of the transition from FTA 5 to FTA 6, for which statistical tests show no significant difference. We attribute to the vanishingly small number of cases receiving an FTA score of 6, which in turn stems from Pretrial's (understandable) difficulty in concluding that a defendant missed a court date, and its (understandable) reluctance to identify and conclude that prior missed court dates corresponded to FTAs. The PSA's NCA scale, but not the FTA scale, appears uniformly valid with respect to all comparisons; again, the difficulty in inputting reliable data on prior FTAs likely played a role.

utility and differential prediction by race and gender in kentucky." Available at SSRN 3168452 (2018).

- DeMichele, M, Baumgartner, P, Wenger, M, Barrick, K, Comfort, M. Public safety assessment: Predictive utility and differential prediction by race in Kentucky. Criminal Public Policy. 2020; 19: 409– 431.
- VanNostrand, Marie, and Gena Keebler. "Pretrial risk assessment in the federal court." Fed. Probation 73 (2009): 3.
- VanNostrand, Marie, and Christopher T. Lowenkamp. "Assessing pretrial risk without a defendant interview." Laura and John Arnold Foundation (2013).

Figure 4: New Criminal Activity Failure Rates by Risk Score

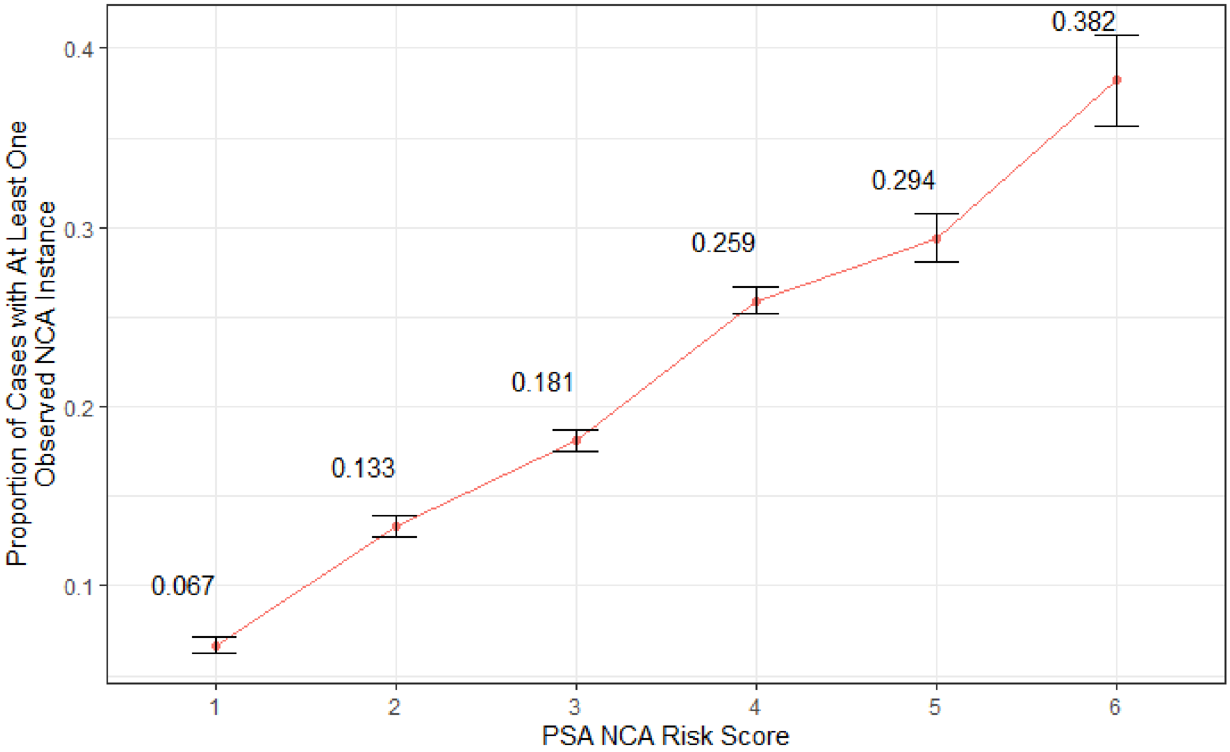


Figure 4 shows the relevant failure rates and associated 95% confidence intervals for NCA by risk score category. A valid risk assessment tool should show significant increases in failure rate at each subsequent level of the associated risk score. The lack of overlap between confidence intervals for consecutive paired scores indicates that each subsequent increase in the PSA NCA risk score is associated with a significant increase in failure rates. All differences are statistically significant. A one unit change in risk score level was observed to have an associated failure rate increase of roughly 6.5%. The increases for each subsequent score increase, with the exception of the change from a NCA risk score of 3 to 4, are fairly uniform and consistent. Overall, this figure provides evidence supporting the overall and uniform validity of the PSA with respect to NCA outcome.

Figure 5: New Violent Criminal Activity Failure Rates by Presence of NVCA Flag

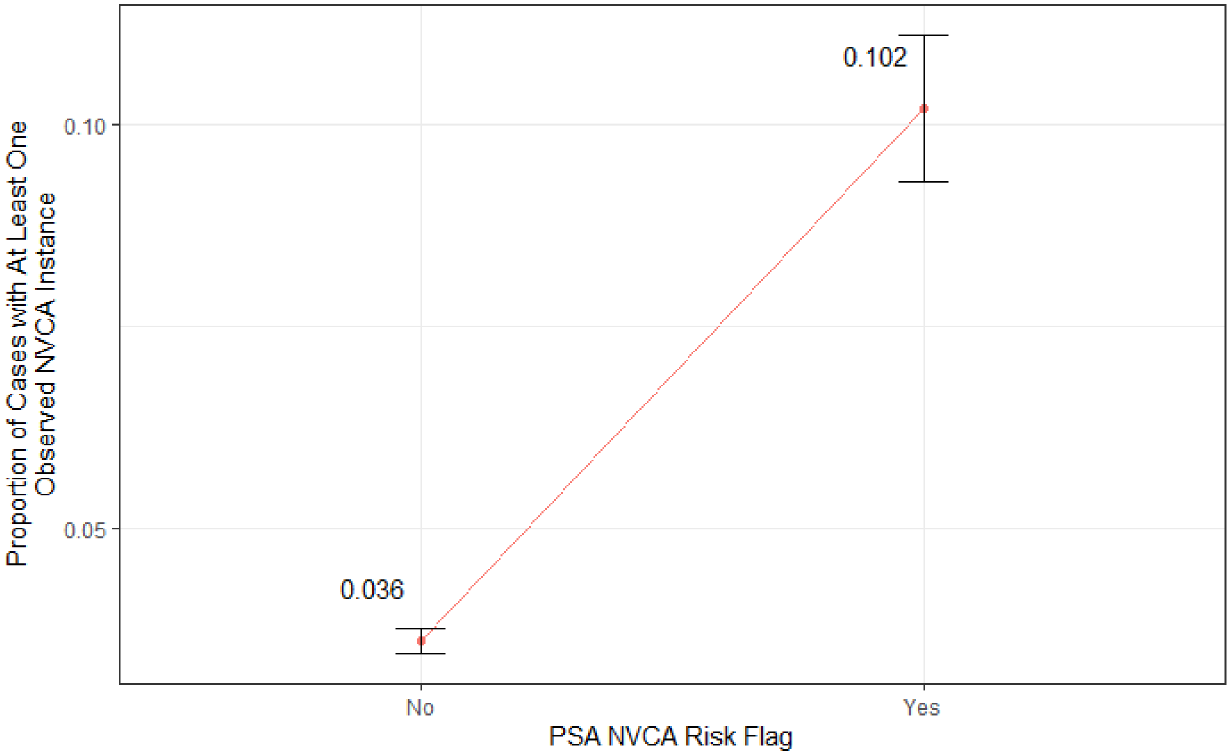


Figure 5 shows the relevant failure rates and associated 95% confidence intervals for NVCA by presence of the NVCA Risk Flag. An overall valid risk assessment tool should show significant increases in failure rate when a binary prediction flag is present, which could be indicated by no overlap between the confidence intervals. The presence of the PSA NVCA risk flag is associated with a significant increase in failure rates of 6.5 percentage points. The difference is statistically significant, and thus provides evidence supporting the validity of the PSA with respect to NVCA outcome.

Figure 6: Base FTA Failure Rates by Risk Score

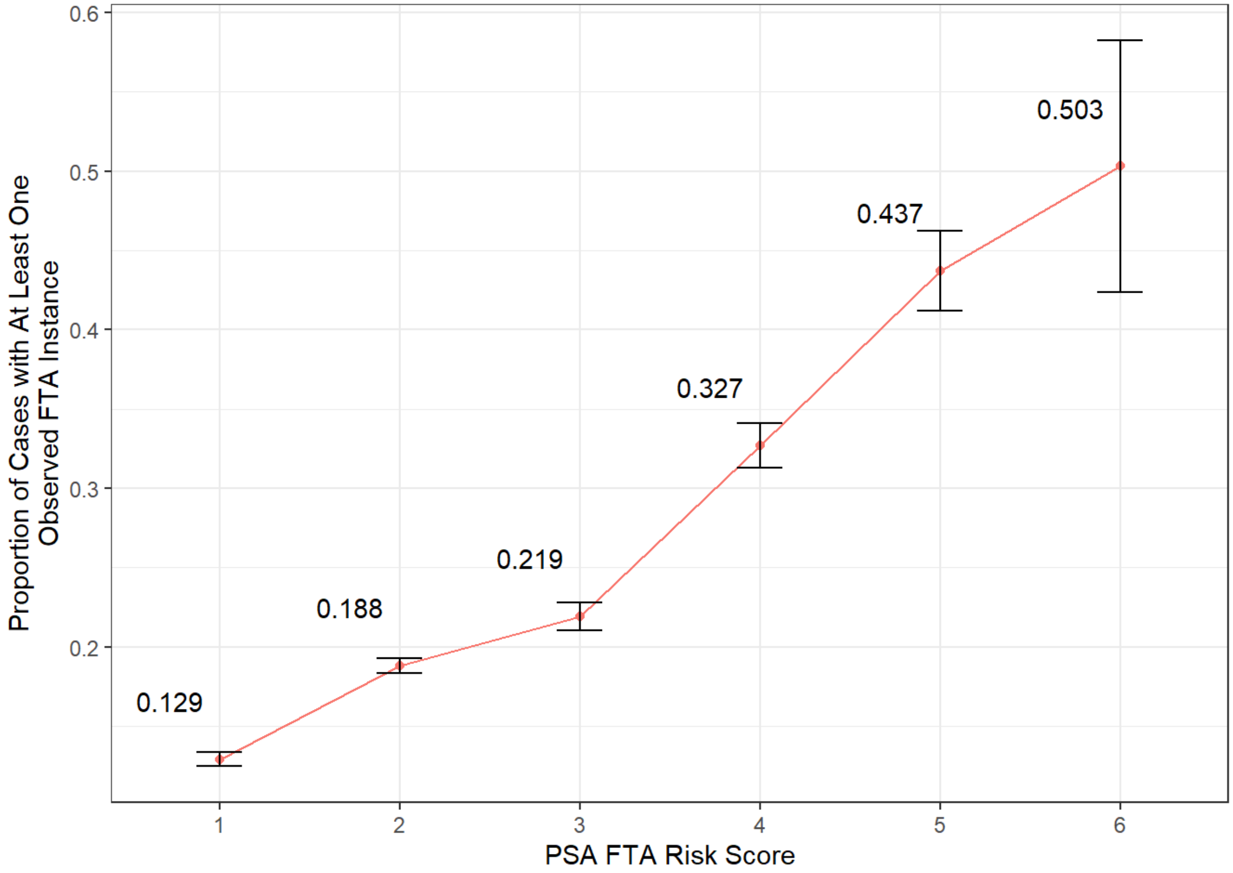


Figure 6 shows the relevant failure rates and associated 95% confidence intervals for Base FTA by risk score category. For the base FTA metric, each increase in the PSA FTA risk score is associated with a statistically significant increase in failure rates, with the exception of an increase from an FTA risk score of 5 to 6. As noted above, this is likely due to the vanishingly small fraction of cases receiving a risk classification of 6. A one unit change in risk score is associated with a variety of increases in failure rates, ranging from an increase of only 3.1% to an increase of 10.8%. Overall, this figure provides evidence supporting the overall validity of the PSA for FTA at risk scores of 5 or below but raises some questions as to uniform validity.

Figure 7: FTA+ Failure Rates By Risk Score

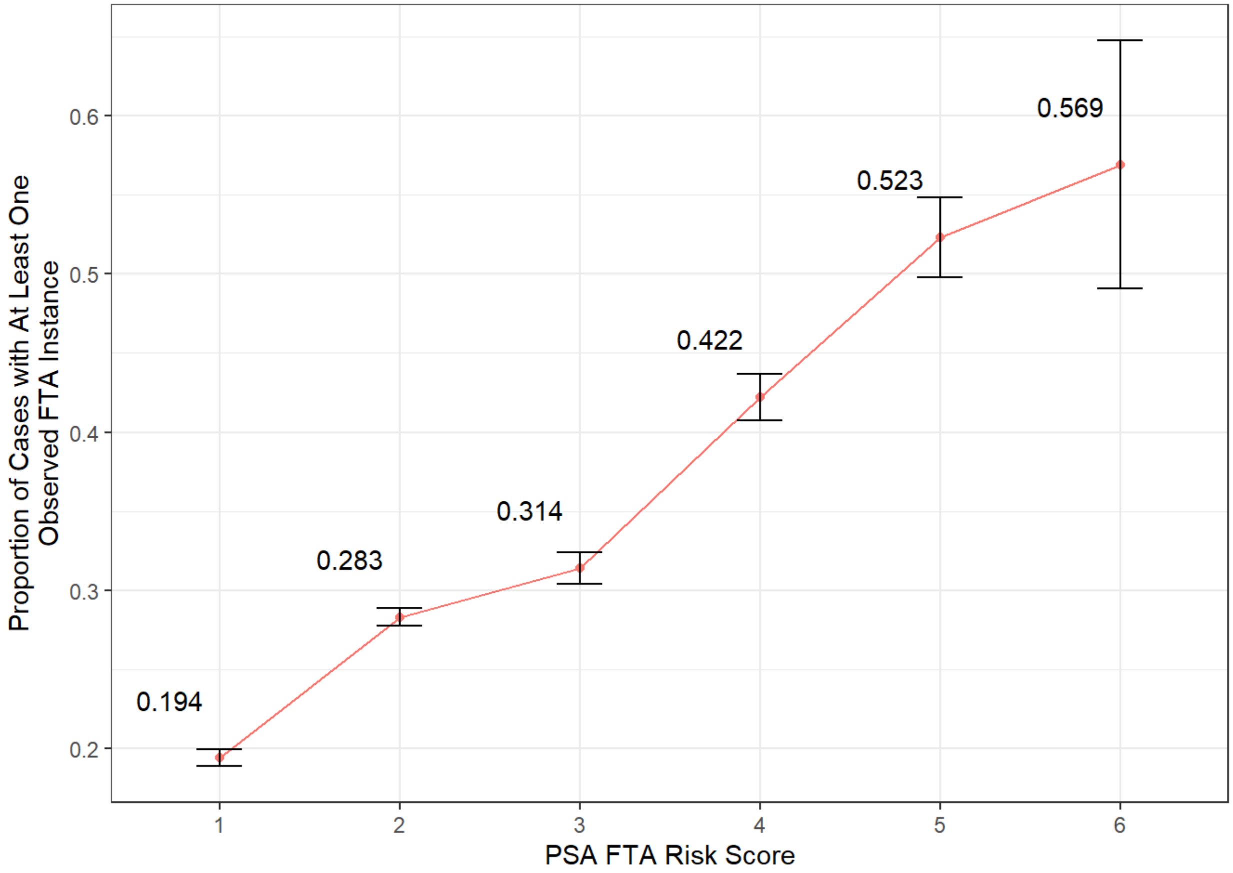


Figure 7 shows the relevant failure rates and associated 95% confidence intervals for FTA+ by risk score category. For the FTA+ metric, each increase in the PSA FTA risk score is associated with a statistically significant increase in failure rates, with the exception of an increase from an FTA risk score of 5 to 6. As noted above, this is likely due to the vanishingly small fraction of cases receiving a risk classification of 6. The FTA+ failure rates are uniformly higher than the baseline FTA rates, which is to be expected given that FTA+ includes all warrant types. These increases range from around 7%-10%. A one unit change in risk score is associated with a variety of increases in failure rates, ranging from an increase of only 3.1% to an increase of 10.8%. Overall, this figure provides evidence supporting the overall validity of the PSA for FTA at risk scores of 5 or below but raises some questions as to uniform validity.

Figures 4-7 demonstrate consistently increasing failure rates for each of the three PSA outcome events. Risk assessment scores for NCA, NVCA, and FTA all report higher failure rates for the higher score of each consecutive score pairing (or the single pairing for NVCA). NCA failure rates corresponded to a minimum of 6.7% for cases with risk scores of 1 and maximum failure rate of 38.2% for cases with risk scores of 6. Base FTA failure rates achieve a similar minimum and maximum at scores of 1 and 6, with rates of 12.9% and 50.3% respectively. The equivalent FTA+ rates are 19.4% and 56.9%. Cases with an NVCA flag present were observed with approximately three times the failure rate of cases without the flag present, with failure rates of 3.6% and 10.2%, respectively. All score transitions (*i.e.*, 1 to 2, 2 to 3, etc. for the FTA and NCA scales, 0 to 1 for NVCA) corresponded to statistically significant differences except, as noted above, for the transition from 5-6 on the FTA scale for both outcome constructions. With respect to uniform validity, as the figures above suggest, rough statistical comparisons[44] suggest that each step increase in the NCA and FTA scores are associated with approximately equivalent increases in failure rates except for the FTA transition from 2 to 3, which has a lower increase, and the transition from 5-6, which is statistically insignificant but whose potential range is large.

ii. Bivariate correlations

This subsection provides the results of bivariate comparisons, also known as correlations. This correlation analysis provides moderate to strong evidence that the PSA scales are overall valid. In particular, the overall risk score achieves a larger correlation coefficient than any individual factor coefficient, suggesting that input factors provide some non-overlapping predictive information that is preserved by the assessment's calculation methods. Key details are as follows:

- Each input factor across all PSA metrics is statistically significantly correlated with the relevant outcome in the expected direction (positive for all factors except Age at Current Arrest, which is negative, as expected).
- These correlations are overall small, ranging between 0.03 and 0.21.
- The largest correlations in magnitude for each PSA metric are associated with the overall risk score (or flag) and the relevant outcome metric, indicating that input factors provide some non-overlapping predictive information that is preserved by the assessment's calculation.

The PSA risk scores are composite measures based on nine separate input variables. Not every input is used for each score. The table below reports which scores are calculated from each of the nine separate inputs.

Table 1: PSA Input Factors For Each Outcome Risk Score

Input	NCA Risk Score	NVCA Risk Flag	FTA Risk Score
Age at Current Arrest	X	X**	
Pending Charge at Time of Current Offense	X	X	X
Prior Misdemeanor Conviction	X	X*	X*
Prior Felony Conviction	X	X*	X*

[44] We examined whether the 95% confidence intervals for the failure rate increases for each step increase overlapped with the interval for any other step increase. All do with the exception of the FTA 2-3 transition, which overlaps with the interval for no other FTA step increase.

Prior Violent Conviction	X	X	
Prior FTA in the Past 2 Years	X		X
Prior FTA older than 2 Years			X
Prior Sentence to Incarceration	X		
Current Violent Offense		X	
*These variables are used in a joint ‘OR’ manner where either a prior misdemeanor or a prior felony conviction is considered a prior conviction. **This variable is only used in a joint “AND’ manner with prior violent conviction.			

Validation by input correlations examines whether the PSA’s inputs are meaningfully related to the relevant outcomes. Under this validation technique, each of the items used to construct the relevant PSA risk scores should correlate in a statistically significant way to the relevant outcomes.[45] We use a common measure of correlation, a Pearsons r coefficient, and the corresponding significance test that the reported coefficient is significantly different from zero. As a secondary analysis, we also examine the magnitude of the coefficient. These tests allow us to evaluate the overall validity of the PSA. The following figures plot the overall Pearsons r coefficient for each relevant PSA Risk Assessment input across the three outcome events: N(V)CA/FTA.

[45] Input correlations are more often used during the initial construction phase of building a PRAI, but they are still useful in the context of validation. For relevant examples of correlations used in an PRAI assessment capacity, see Bechtel, Kristin, Christopher T. Lowenkamp, and Alex Holsinger. "Identifying the predictors of pretrial failure: A meta-analysis." Fed. Probation 75 (2011): 78; DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. "The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky." Available at SSRN 3168452 (2018).

Figure 8: NCA Input Factor Correlations with Observed NCA Events

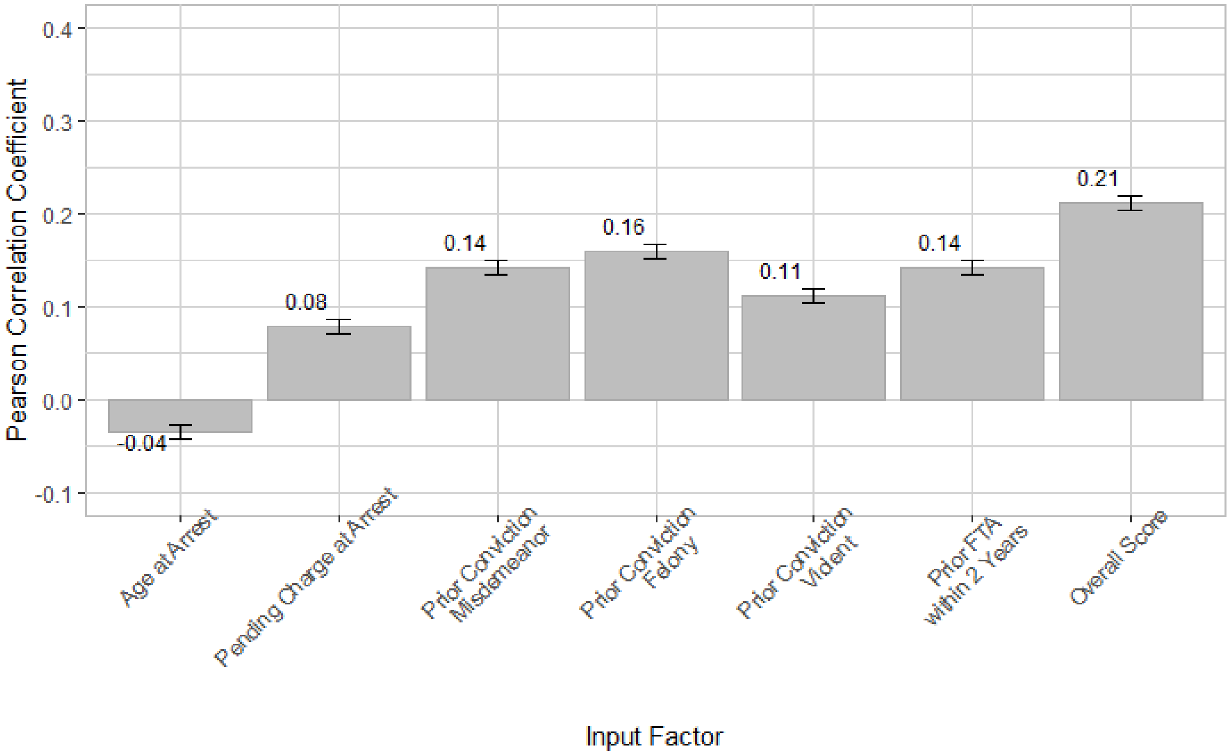


Figure 8 shows the Pearson Correlation Coefficient and associated 95% confidence interval for each of the six factors used in the calculation of the PSA NCA score, as well as the correlation of the overall score, with observed NCA events. The figure indicates that each input factor and the overall risk score is significantly correlated with observed NCA events in the appropriate direction. The overall risk score achieves a larger correlation coefficient than any individual factor coefficient, suggesting that input factors provide some non-overlapping predictive information that is preserved by the assessment's calculation methods. This figure provides evidence for the overall validity of the PSA with respect to NCA outcomes.

Figure 9: NVCA Input Factor Correlations with Observed NVCA Events

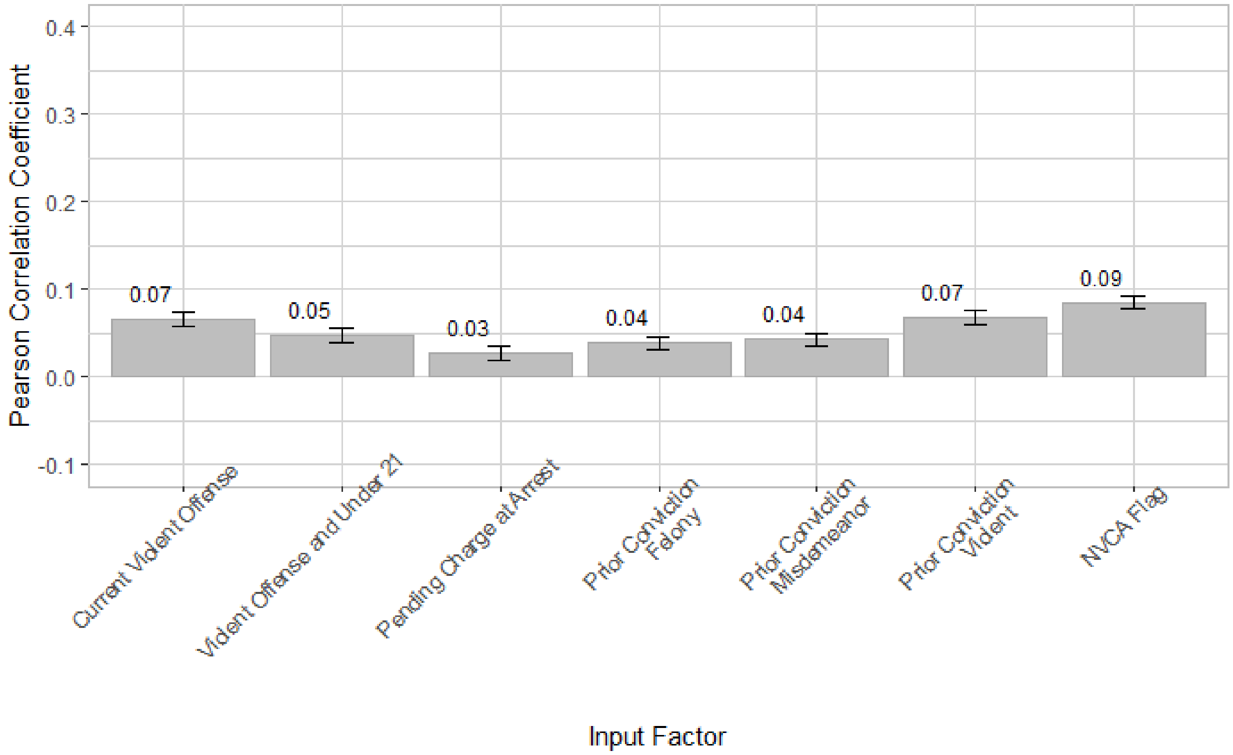


Figure 9 shows the Pearson Correlation Coefficient and associated 95% confidence interval for each of the six factors used in the calculation of the PSA NVCA risk flag, as well as the correlation of the presence of the risk flag, with observed NVCA events. The lack of overlap between the confidence intervals and 0 indicates that each input factor and the overall risk score is significantly correlated with observed NVCA events in the appropriate direction. The overall risk score achieves a larger correlation coefficient than any individual factor coefficient, suggesting that input factors provide some non-overlapping predictive information that is preserved by the assessment's calculation methods for NVCA. Overall, this figure provides evidence for the overall validity of the PSA with respect to NVCA outcomes.

Figure 10: Base FTA Input Factor Correlations with Observed FTA Events

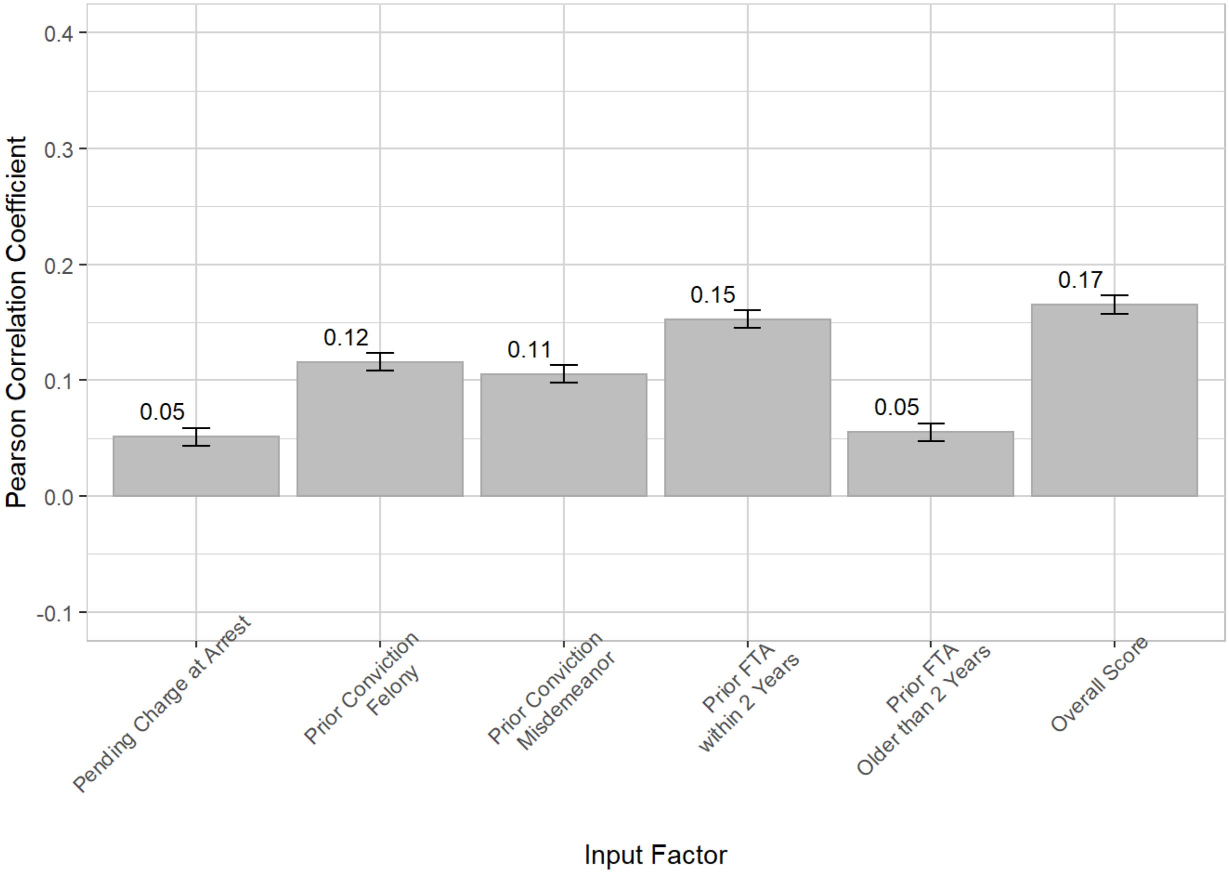


Figure 10 shows the Pearson Correlation Coefficient and associated 95% confidence intervals for each of the five factors used in the calculation of the PSA FTA score, as well as the correlation of the overall score, with observed Base FTA events. Each input factor and the overall risk score is significantly correlated with observed Base FTA events in the appropriate direction. The smallest and largest factor correlation coefficients are obtained for Prior FTAs older than 2 Years and within 2 Years, respectively. The overall risk score achieves a larger correlation coefficient than any individual factor coefficient for almost all outcome constructions, suggesting that input factors provide some non-overlapping predictive information that is preserved by the assessment's calculation methods. Overall, this figure provides evidence for the overall validity of the PSA with respect to Base FTA outcomes.

Figure 11: FTA+ Input Factor Correlations with Observed FTA Events

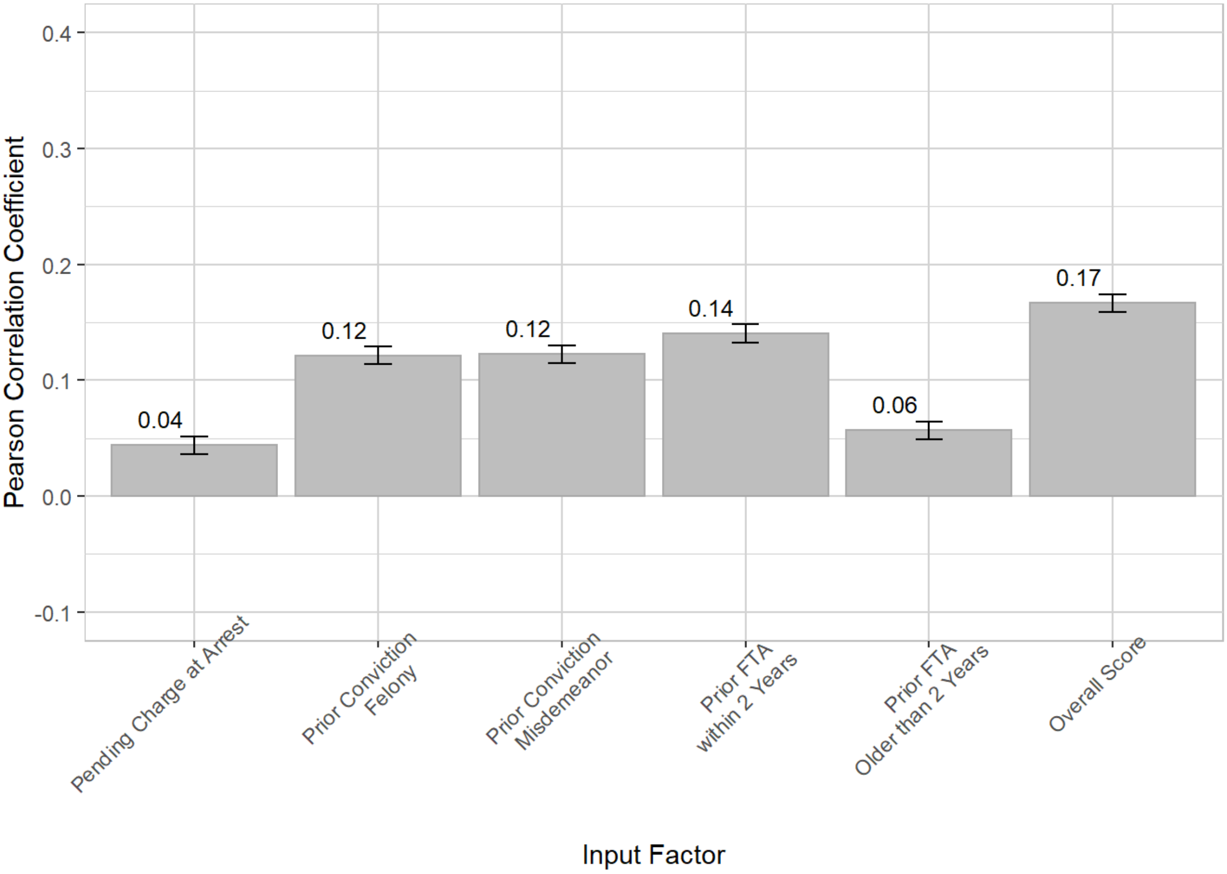


Figure 11 shows the Pearson Correlation Coefficient and associated 95% confidence intervals for each of the five factors used in the calculation of the PSA FTA score, as well as the correlation of the overall score, with observed FTA+ events. Each input factor and the overall risk score is significantly correlated with observed FTA+ events in the appropriate direction. The smallest and largest factor correlation coefficients are obtained for Prior FTAs older than 2 Years and within 2 Years, respectively. The overall risk score achieves a larger correlation coefficient than any individual factor coefficient for almost all outcome constructions, suggesting that input factors provide some non-overlapping predictive information that is preserved by the assessment's calculation methods. Each factor correlation for the FTA+ outcome construction is within 1/100th of the equivalent correlation for the Base FTA measure. Overall, this figure provides evidence for the overall validity of the PSA with respect to FTA+ outcomes.

Figures 8-11 each show positive correlations between the various factor inputs to the PSA scores and the relevant PSA outcome, with the exception of the Age factor, which is negatively correlated (as expected). Each of these correlations is in the same direction as the rule which translates them into the relevant risk score. For each PSA outcome event, all input correlation coefficients are significantly different from 0 at the p < 0.001 level. The significance levels of these findings are in line with the expectations of the overall validation framework. The magnitude of the correlation coefficients, all of which are below 0.3, would not generally be considered strong in most social science disciplines. We speculate that this fact might be due to the nature of the PSA's treatment of its inputs, in which input values are often binned or dichotomized and then translated into a one or two unit additive. In this way a portion of the information in the raw form of the inputs is lost. The statistical significance results provide evidence for overall validation;

again, the magnitudes are below 0.3. Overall the item-based correlation measures provide some, but not strong, evidence of overall validation.

iii. Area under the curve

This subsection discusses the results of the area under the curve ("AUC") analysis. The AUC analysis shows weak to moderate evidence of the overall validity of the PSA scales. Key details are as follows:

- NCA and FTA outcomes show weak to moderate AUC scores indicating some gain in predictive power from the risk score, while the NVCA flag has an AUC score that indicates no or weak gain in predictive power.
- There are no significant differences in AUC scores across demographic subgroups for any of the PSA outcomes, providing strong evidence of equitable validity.

One of the most commonly used diagnostic tools for evaluating the performance of binary classification, or binary outcome, models is the Receiver Operating Characteristic ("ROC") curve, which plots the trade-off in a model's sensitivity at different thresholds of considering a case under one predictive category versus another.[46] In other words, ROC curves examine the difference between the true positive (an observation classified as high risk later corresponds to a failure) rate and the false positive (an observation classified as high risk later does not correspond to a failure) rate at different thresholds of making a positive prediction. A binary classification model that provided no inherent increase in information would appear as a straight 45 degree line that indicated no change from a sensitivity value of 0.50. Essentially, that means that the risk assessment instrument performs no better than having a model that classifies all observations into the most commonly observed outcome; in the Harris County data, that would mean classifying all individuals as low risk for NCA, NVCA, and FTA. More accurate and informative models should provide greater distance between the ROC curve and the hypothetical no-information 45 degree line. Standard practice in this area is to assess this gain in information by measuring the area under the ROC curve ("AUC"), which quantifies the difference between the predictive gain of the model under the ROC curve and the baseline performance of the no-information line. ROC curves do not have direct analogies to classification models with multiple categories, such as the PSA NCA and FTA scores. The AUC measurement does, however, generalize to such multi-category classification settings.

In the case of the PSA, the AUC measurement provides the probability that a randomly selected case that observed a failure (*i.e.*, observed at least one NCA, NVCA, or FTA event under the relevant outcome construction definition) had a higher score than a randomly selected case that did not observe a failure. As in the binary classification case, an assessment tool that provides no additional useful information, and thus fails to overall validate, would have an AUC measurement indistinguishable from 0.50. The following benchmarks are sometimes used: an

[46] See Huang, Jin, and Charles X. Ling. "Using AUC and accuracy in evaluating learning algorithms." IEEE Transactions on knowledge and Data Engineering 17, no. 3 (2005): 299-310, for a discussion on the connections between ROC, AUC, and accuracy measures for assessing classifier models.

AUC measurement less than 0.54 indicates no evidence of validity.[47] An AUC measurement between 0.55 and 0.63 indicates some, but not strong, evidence of validity. AUC measurements between 0.64 and 0.70 indicate moderate evidence, and a measurement greater than 0.70 indicates strong evidence. To the extent that there is no significant difference in AUC measures across either racial or gender pairings, we conclude that the PSA provides equivalent gains in predictions for each group within the pairing. The figure below plots the AUC measures for each of the four outcome event constructions: NCA, NVCA, Base FTA, and FTA+.

Figure 12: Area Under the Curve Values by Outcome Construction

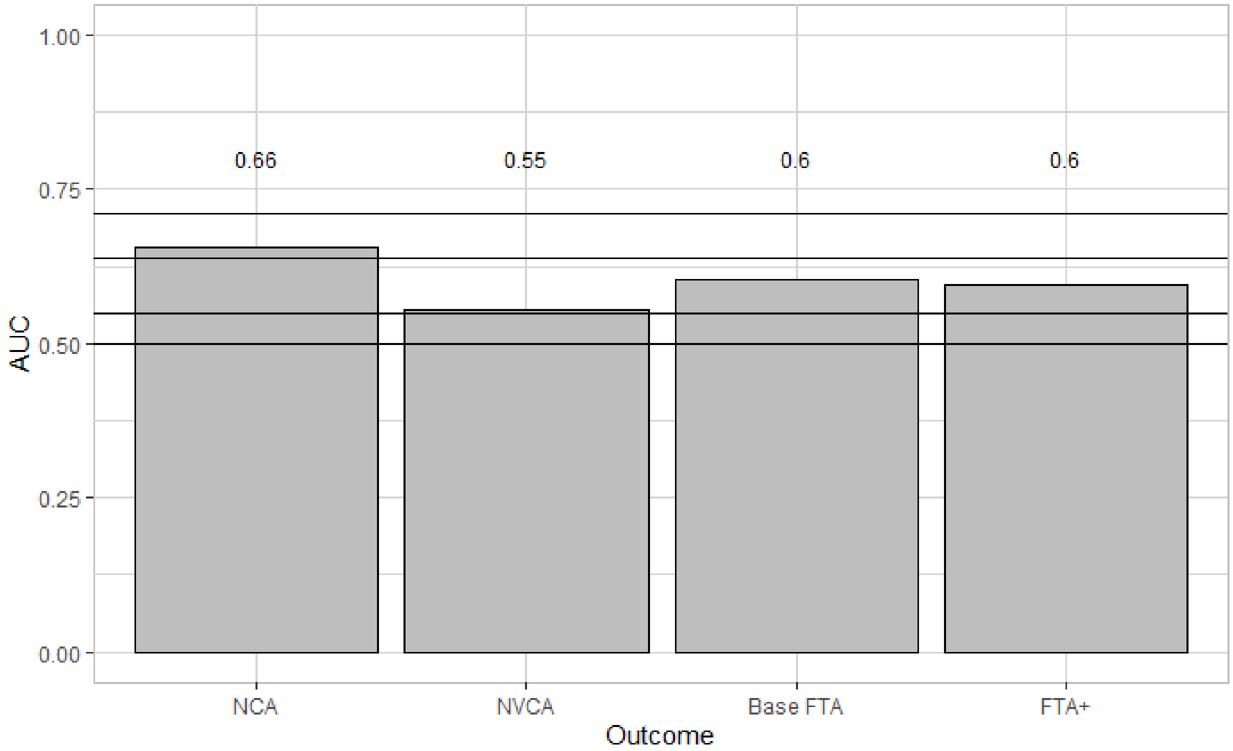


Figure 12 shows the area under the curve values for each outcome construction. AUC values range from 0 to 1, and in the case of a multi-outcome predictive assessment tool, like the PSA, are best understood as the probability that a randomly selected case with an observed failure for a relevant outcome will have a higher corresponding risk score than a randomly selected case with no observed failure for a relevant outcome. For the purposes of this study we rely on the following cutoffs for evaluating the strength of evidence provided by an AUC measurement: 0.5 - 0.54: no evidence, 0.55 - 0.63: weak evidence, 0.64 - 0.7: moderate evidence, > 0.7: strong evidence. The results reported in the above figure provide borderline weak evidence that the PSA works better than chance at classifying NVCA events, weak evidence the PSA works better than chance at classifying FTA events (for both outcome constructions), and moderate evidence that the PSA works better than chance at classifying NCA events.

[47] DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. "The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky." Available at SSRN 3168452 (2018); Desmarais, Sarah L., and Jay P. Singh. "Risk assessment instruments validated and implemented in correctional settings in the United States." Lexington, KY: Council of State Governments (2013).

The AUC metrics show positive gains above the random chance threshold of 0.50 for each of the outcome constructions under all three PSA outcome events. For the NCA outcome constructions, the overall AUC metrics are 0.66, which represents moderate gains in predictive power. For NVCA, the overall AUC metric is 0.55, which represents borderline weak gains in predictive power. For both FTA outcomes, the overall AUC metric is 0.60, indicating weak gains in predictive power. The figures for the NCA outcomes correspond to moderate evidence of overall validity, but all outcome constructions provide some evidence for overall validity.

Figure 13: Area Under the Curve Values by Outcome Construction Across Demographic Subgroups

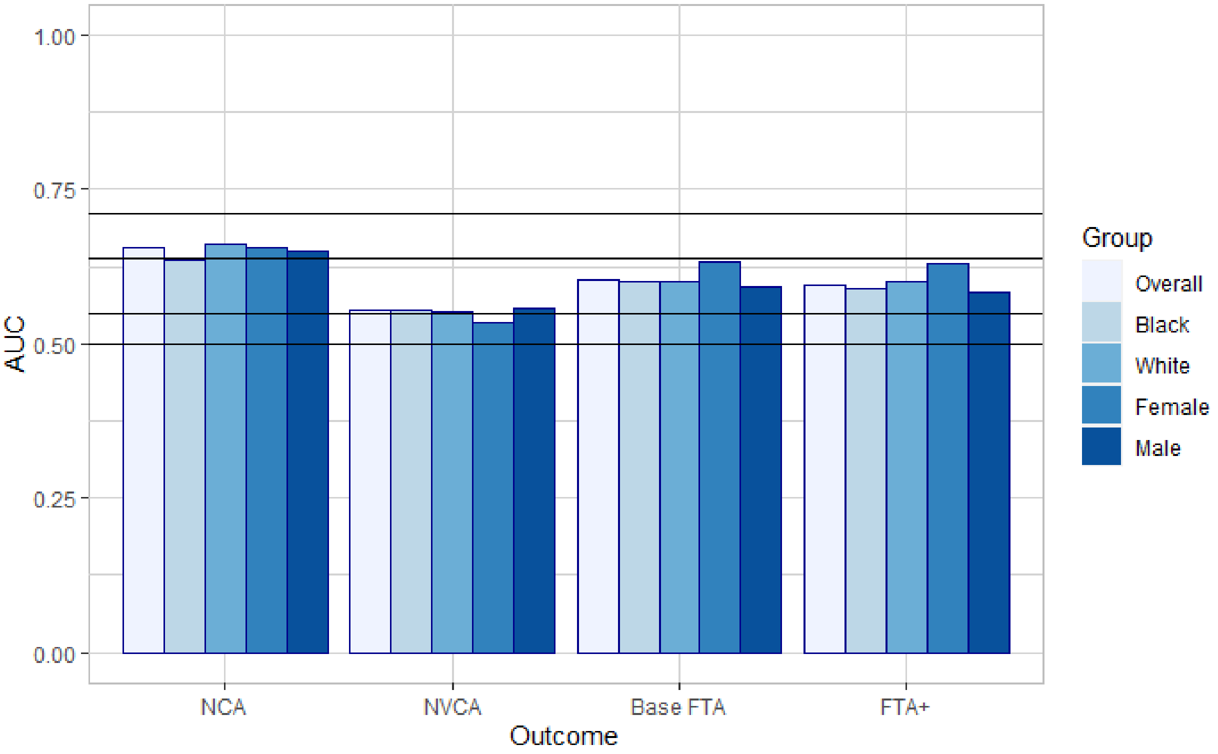


Figure 13 shows the area under the curve values for each outcome construction across the four main demographic groups of analysis as well as for the overall study population. AUC values range from 0 to 1, and in the case of a multi-outcome predictive assessment tool, like the PSA, are best understood as the probability that a randomly selected case with an observed failure for a relevant outcome will have a higher corresponding risk score than a randomly selected case with no observed failure for a relevant outcome. For the purposes of this study we rely on the following cutoffs for evaluating the strength of evidence provided by an AUC measurement: 0.5 - 0.54: no evidence, 0.55 - 0.63: weak evidence, 0.64 - 0.7: moderate evidence, > 0.7: strong evidence. The results reported in the above figure provide borderline weak evidence that the PSA works better than chance at classifying NVCA events, weak evidence the PSA works better than chance at classifying FTA events (for both metrics), and moderate evidence that the PSA works better than chance at classifying NCA events.

The AUC metric can also be used to evaluate whether the PSA equitably validates. AUC metrics can be calculated on demographic subgroup populations specifically, and these measures can be used to test for significance in the difference between racial and gender comparison AUC metrics, which are shown in Figure 13. We find little evidence of a difference in the validity with respect to racial or gender groups. Black and White individuals did differ significantly for NCA outcomes constructions only. Male and female individuals differed significantly for NVCA and FTA outcomes; however, neither set of differences was large. The AUC metrics differ by at most 0.03 and in no outcome construction did the differences in AUC metrics cross any evaluative thresholds.

d. Techniques Used Outside the Pretrial Context

i. Regression

This subsection provides the results of a logistic regression analysis. This analysis provides strong evidence that the PSA is overall valid, and weak evidence that the PSA is uniformly valid. Key details are as follows.

- PSA risk scores/flags have significantly positive coefficients, indicating that increases in risk scores are statistically significantly associated with increases in the probability of observing a relevant outcome.
- Moving from the minimum to the maximum risk score is associated with a 5x increase in the probability of observing an NCA and a 3x increase in observing an FTA, again suggesting overall validity.
- The presence of the NVCA Flag is associated with a 3x increase in observing an NVCA, providing strong evidence of overall validity.

A logistic regression framework provides an off-the-shelf[48] method for assessing the overall validity of the PSA.[49] The following figure plots predictive probabilities of observing at least one

[48] Because instances of NCA, NVCA, or FTA failure can be dichotomized and reported as a binary outcome (where 1 indicates one or more of the relevant events observed under a specific outcome construction, and 0 indicates no observed relevant events), we can estimate the relationship between a PSA risk assessment score and the relevant outcome in this fairly standard statistical format. A bivariate logistic regression, with the risk assessment score regressed on the relevant outcome, will provide an exponentiated coefficient estimate of the relationship between the risk score and the odds ratio of observing at least one relevant event failure relative to not observing a relevant event failure. The extent that this exponentiated coefficient is significantly larger than 1 provides evidence for the overall validity of the PSA, with a larger magnitude indicating stronger evidence. An additional regression is computed that includes a higher order risk assessment term to test uniform validity. To the extent this coefficient is significantly different from one, this indicates that lower levels of the risk assessment score provide different magnitude of effects than higher levels of the risk assessment score. An insignificant coefficient on this ‘self-interaction’ term would provide evidence that the PSA uniformly validates.

[49] For other validation studies that have utilized a regression framework, see:

- Bechtel, Kristin, Alexander M. Holsinger, Christopher T. Lowenkamp, and Madeline J. Warren. "A meta-analytic review of pretrial research: Risk assessment, bond type, and interventions." American Journal of Criminal Justice 42, no. 2 (2017): 443-467.
- Desmarais, Sarah L., Samantha A. Zottola, Sarah E. Duhart Clarke, and Evan M. Lowder. "Predictive Validity of Pretrial Risk Assessments: A Systematic Review of the Literature." Criminal Justice and Behavior (2020): 0093854820932959.

relevant outcome event for each of the outcome events across relevant risk assessment scores obtained from a bivariate logistic regression model where the main outcome event regressed only on the relevant risk score scale.

Figure 14: NCA Birvariate Predicted Probabilities

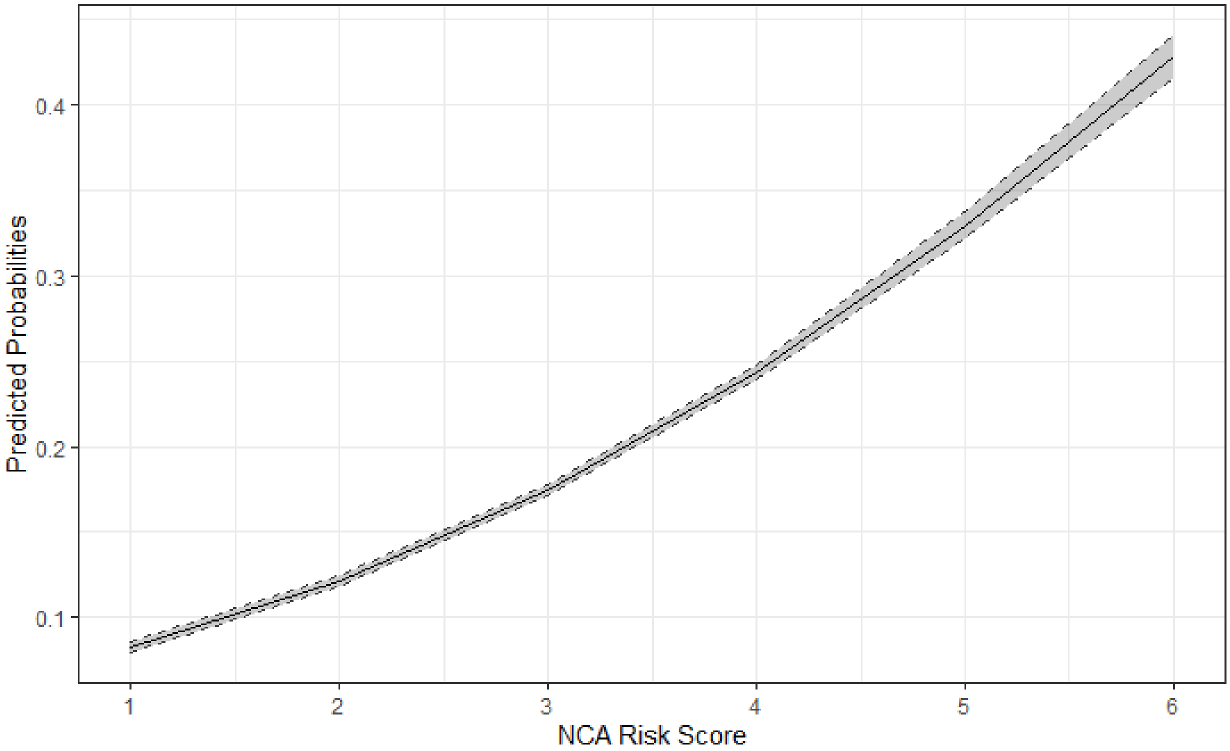


Figure 14 reports predicted probabilities and 95% confidence intervals for observing an NCA event obtained from the bivariate regression model with only the relevant PSA risk score scale as the regressor. The PSA NCA risk score has a significant, positive coefficient, indicating that higher NCA risk scores are significantly associated with a higher probability of an observed NCA failure. A one unit increase in the NCA risk score is associated with a 53% increase in the odds of observing an NCA failure versus not observing an NCA failure. This estimate exists on a confidence interval from a 50% increase in the odds ratio to a 55% increase in the odds ratio. Thus, this figure provides support for the overall validity of the PSA with respect to NCA outcomes.

- DeMichele, M, Baumgartner, P, Wenger, M, Barrick, K, Comfort, M. Public safety assessment: Predictive utility and differential prediction by race in Kentucky. Criminal Public Policy. 2020; 19: 409– 431.

Figure 15: NVCA Bivariate Predicted Probabilities

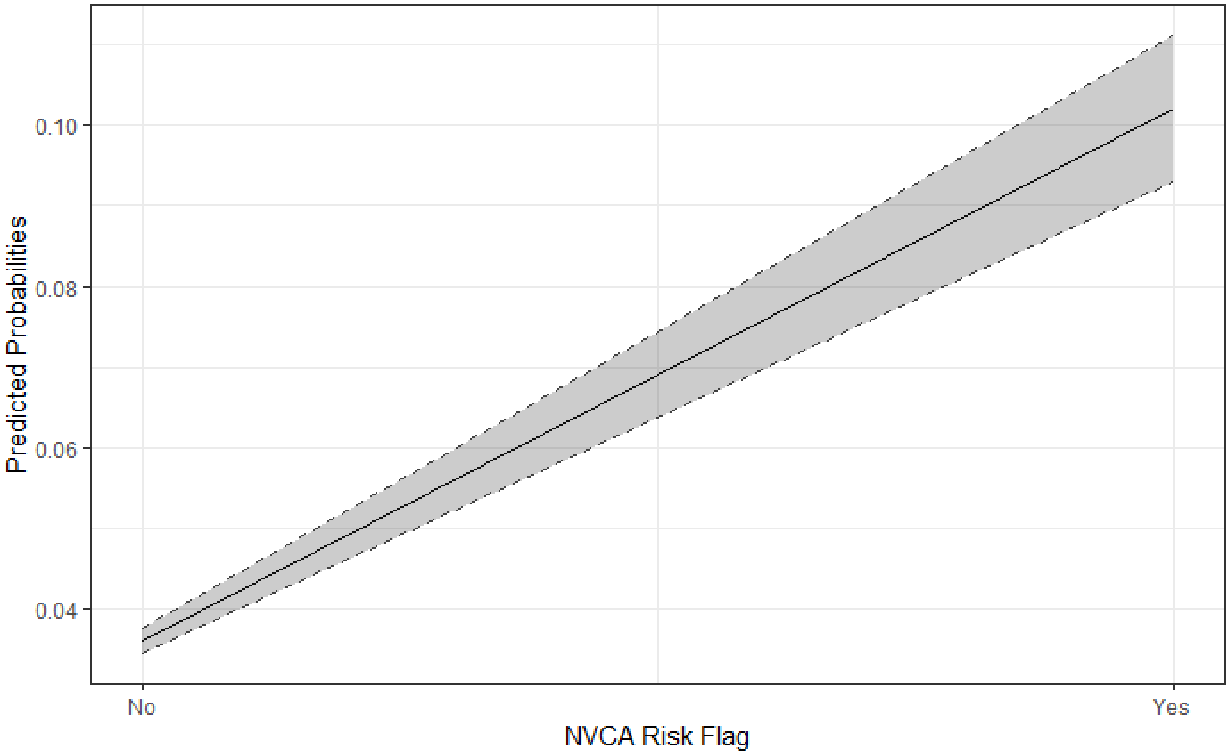


Figure 15 reports predicted probabilities and 95% confidence intervals for observing an NVCA event obtained from the bivariate regression model with only the relevant PSA risk score scale as the regressor. The PSA NVCA risk flag has a significant, positive coefficient, indicating that the presence of the NVCA Risk Flag is significantly associated with a higher probability of an observed NVCA failure. The presence of the NVCA Risk Flag is associated with a 204% increase in the odds of observing an NVCA failure versus not observing an NVCA failure. This estimate exists on a confidence interval from a 172% increase in the odds ratio to a 238% increase in the odds ratio. Thus, this figure provides support for the overall validity of the PSA with respect to NVCA outcomes.

Figure 16: Base FTA Bivariate Predicted Probabilities

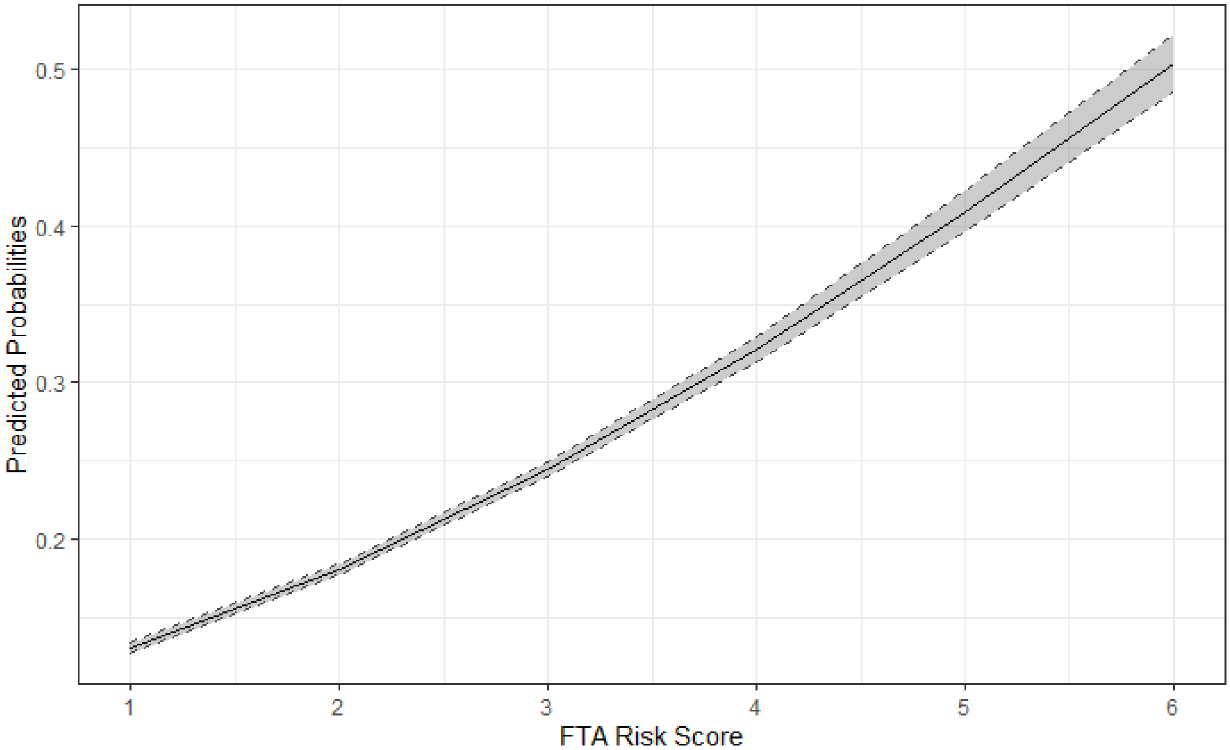


Figure 16 reports predicted probabilities and 95% confidence intervals for observing a Base FTA event obtained from the bivariate regression model with only the relevant PSA risk score scale as the regressor. The PSA FTA risk score has a significant, positive coefficient, indicating that higher FTA risk scores are significantly associated with a higher probability of an observed FTA failure. A one unit increase in the FTA risk score is associated with a 46% increase in the odds of observing an FTA failure versus not observing an FTA failure for the Base FTA outcome construction. This estimate exists on a confidence interval from a 44% increase in the odds ratio to a 49% increase in the odds ratio. Thus, this figure provides support for the overall validity of the PSA with respect to FTA outcomes.

Figure 17: FTA+ Bivariate Predicted Probabilities

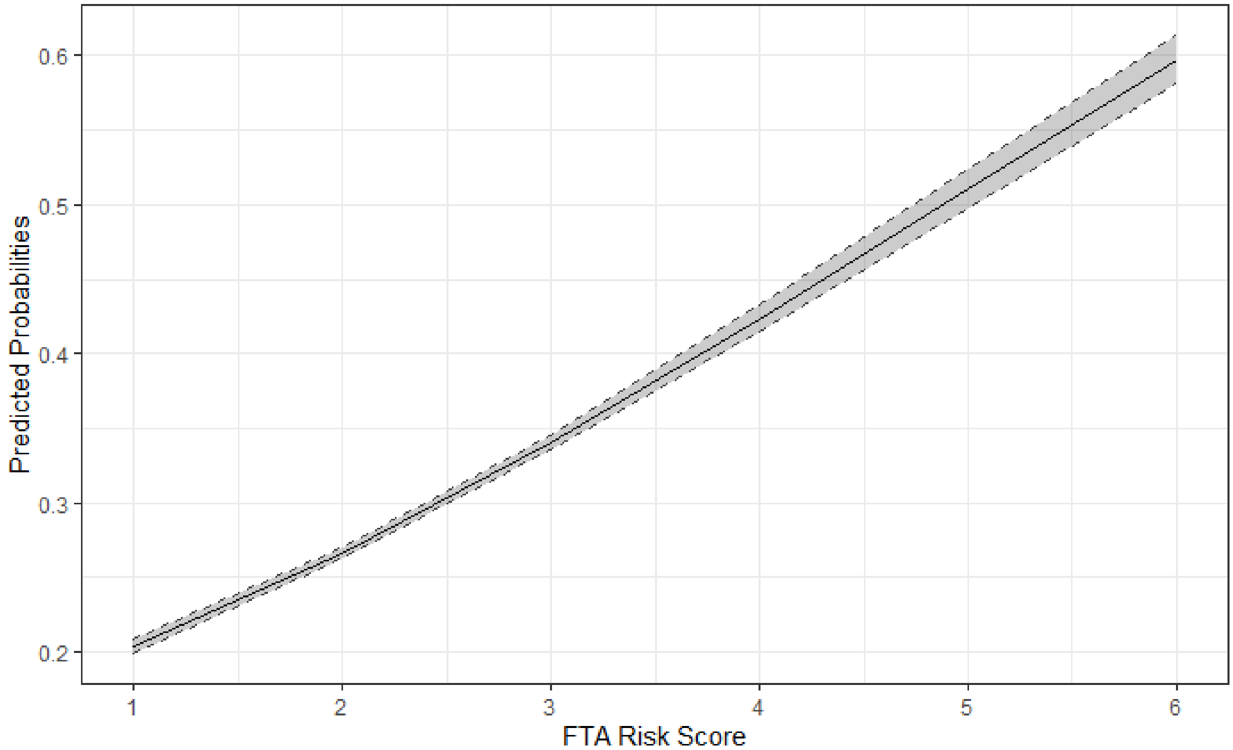


Figure 17 reports predicted probabilities and 95% confidence intervals for observing an FTA+ event obtained from the bivariate regression model with only the relevant PSA risk score scale as the regressor. The PSA FTA risk score has a significant, positive coefficient, indicating that higher FTA risk scores are significantly associated with a higher probability of an observed FTA failure. A one unit increase in the FTA risk score is associated with a 42% increase in the odds of observing an FTA failure versus not observing an FTA failure for the FTA+ outcome construction. This estimate exists on a confidence interval from a 40% increase in the odds ratio to a 45% increase in the odds ratio. Thus, this figure provides support for the overall validity of the PSA with respect to FTA outcomes.

Figures 14-17 show that the predicted probabilities across each relevant risk score level significantly increase along the risk score scale. For the NCA model, these range from a minimum of an 8.3% predicted chance of observing an NCA outcome at an NCA risk score of 1 to a maximum of 42.9% at an NCA score of 6. For the base FTA model, these range from a minimum predicted probability of observing at least one FTA event of 13.1% at an FTA score of 1 and a maximum predicted probability of 50.3% at an FTA score of 6. The equivalent predicted probabilities for the FTA+ outcome construction are a 20.3% chance and a 59.7% chance. For the NVCA model, having the NVCA flag present results in a predicted probability of 10.2% while not having the flag present is associated with a predicted probability of 3.6%. The standard error regions around these probability estimates indicate that the differences between the predicted probabilities is significant. These probabilities are generated from simple bivariate logistic regression models with the relevant risk score as the independent variable and the related observed outcome as the dependent variable. The exponentiated coefficient estimate for the risk score scale is significantly greater than one across all outcome models, indicating that increases in the associated PSA risk score is associated with increases in observed instances of failed outcome observations.

In the case of the bivariate NCA model, the exponentiated coefficient estimate for the NCA risk score scale is 1.53 on a 95% confidence interval of (1.50, 1.55), indicating that a one unit increase in NCA risk score is associated with a 53% increase in the odds ratio of observing at least one NCA event during the pretrial period. For the bivariate NVCA model, the exponentiated coefficient estimate for the the presence of the NVCA Flag is 3.04 on a 95% confidence interval of (2.72, 3.38), indicating that the presence of the NVCA Flag is associated with a 204% increase in the odds ratio of observing at least one NVCA event during the pretrial period. For the Base FTA bivariate model, the exponentiated coefficient estimate for the FTA risk score scale is 1.46 on a 95% confidence interval of (1.44, 1.49), indicating that a one unit increase in FTA risk score is associated with a 46% increase in the odds ratio of observing at least one case-specific FTA resulting in the issuance of a bench warrant. For the FTA+ bivariate model, the exponentiated coefficient estimate for the FTA risk score scale is 1.42 on a 95% confidence interval of (1.40, 1.45), indicating that a one unit increase in FTA risk score is associated with a 42% increase in the odds ratio of observing at least one case-specific FTA resulting in the issuance of a bench warrant. The significance and magnitude of the exponentiated coefficient estimates provides strong evidence for the overall validity of the PSA.

Evaluating uniform validity with a logistic regression is also possible through the inclusion of a higher order ‘self-interaction’ term. This term consists of interacting the risk assessment scale score with itself (squaring it), which allows the model to estimate a differential relation of the scale score on the outcome observations at higher levels of the scale score. The significance of the higher order term will indicate whether the association between the risk score and the relevant observation changes with different scores, indicating that the PSA risk score implies different increases of risk at different points of the score scale. In the NCA, Base FTA, and FTA+ outcome models, the higher order coefficient (estimated from a logistic regression model including the risk score and the higher order risk score as IVs and the relevant outcome as DV) is significant at the $p<0.01$ level. The exponentiated higher order coefficients are 0.95 and 1.03, for the NCA and FTA models respectively. This indicates that higher levels of the NCA score scale are associated with smaller increases in the probability of observing an NCA event, and higher levels of the FTA score scale are associated with larger increases in the probability of observing an FTA event. In both cases, the change in the odds ratio of observing the event, represented by the exponentiated coefficients, of -5% and 3%, are fairly minor. However, given the significance, the logistic regression analysis does provide weak evidence that the PSA does not uniformly validate.

ii. Balanced accuracy measures

This section reports the results of a balanced accuracy analysis. The balanced accuracy measures provide some but mostly weak evidence of overall validity of the PSA. It also provides evidence that the PSA is equitably valid. Key details are as follows.

- Balanced accuracy metrics across all hypothetical score thresholds show some gain in predictive power above the 0.50 threshold; these gains are largest for threshold scores of 2 and 3, but are still relatively weak. The NCA score thresholds of 2 and 3 show moderate gains in predictive power.
- There are no significant differences in balanced accuracy metrics across demographic subgroups for any of the PSA outcomes, suggesting that the PSA scales are equitably valid.

Accuracy is a commonly used assessment technique in machine learning. Accuracy is based on a confusion matrix.[50] One constructs a confusion matrix by dividing each case/observation either into a positive/high risk category or into a negative/low risk category. One then classifies each observation in the positive/high risk category as “true” or “correct” if a failure (here, an FTA or N(V)CA) occurs, and “false” or “incorrect” if no failure occurs. Correspondingly, one classifies each negative/low risk observation as true/correct if no failure occurs, and false if a failure occurs. One calculates the so-called “Accuracy metric” by adding together the number of true positives and true negatives, then dividing by the total number of cases, thus yielding a proportion of ‘correct’ classifications.[51]

Two factors complicated the use of an Accuracy-based metric for validating the PSA. First, Accuracy-based metrics, and the confusion matrices upon which they are based, are built on the assumption that there are only two classifications (high versus low risk) and two outcomes (true/correct versus false/correct).[52] As noted above, while this condition is true for the NVCA Flag, it is not true for the FTA and NCA scores, which both have six risk categories and only two observed outcome categories. The second issue is that the PSA does not make a discrete classification, but instead attempts to classify the level of risk of an individual by an ordinal scale. To address these issues for FTA and NCA, we implement five separate thresholds, meaning risk scores of 1, 2, 3, 4, and 5, for which a score above the threshold represents a positive classification and a score at or below the threshold represents a negative classification. We then calculate each accuracy metric for each of the five hypothetical thresholds for FTA and NCA and the one hypothetical threshold for NVCA.

There is an additional challenge. Accuracy, when used as a diagnostic statistic, is most useful when there is a balance in observed outcome categories, *i.e*. the number of observed positive

[50] For a discussion of the Confusion Matrix, its application to PRAI studies, with a focus on fairness concerns, see: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in criminal justice risk assessments: The state of the art." Sociological Methods & Research (2018): 0049124118782533.

[51] Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in criminal justice risk assessments: The state of the art." Sociological Methods & Research (2018): 0049124118782533.; Daskalaki, Sophia, Ioannis Kopanas, and Nikolaos Avouris. "Evaluation of classifiers for an uneven class distribution problem." Applied artificial intelligence 20, no. 5 (2006): 381-417.

[52] One can generalize such matrices to a risk assessment context in which there the number of risk classifications and the number of outcomes are the same. But this generalization also does not fit the FTA and NCA scales because they have six classifications and two outcomes.

and negative outcome cases is roughly equal. This is due to the fact that standard practice is to compare Accuracy with a theoretical “no information rate,” which is calculated by taking the number of correct predictions a model would make by simply assigning all cases the most common category (which is the classification or “guess” one would make if one had no classifying information available at all). When the number of cases across the different classification categories is equal or uniform, this no information rate is smallest, and that provides the best comparison. As the distribution of cases across observed outcome categories diverges from equal/uniform, the accuracy of the no information guess improves, making any risk score assessed by the Accuracy metric look worse regardless of how well it performs. The Harris County data are not equal or uniform across outcomes. As previously discussed, across FTA, NCA, and NVCA, at least 80% of cases corresponded to no failure outcome.

For this reason, we show below not the raw Accuracy metric but instead what researchers call the “Balanced Accuracy” statistic.[53][54] Balanced Accuracy also comes from machine learning. It corrects for imbalance across outcome categories by calculating accuracy not on an overall basis (total correct classifications divided by total classifications) but by averaging accuracy across outcome categories.[55] That raises a problem in that the no information rate becomes irrelevant, so researchers instead use a series of ranges and thresholds similar in structure to those used for the area under the curve measurement. Balanced Accuracy metrics less than 0.5 represent a loss of information, while those above 0.5 represent at least some gain in predictive accuracy. Additional thresholds above 0.5 differ throughout the literature, but in a general sense, values around 0.5 show no meaningful gain in predictive accuracy, values between 0.6 and 0.7 indicate a modest gain in predictive accuracy, and values above .70 represent a major gain in predictive accuracy.

Calculation of the Balanced Accuracy metric proceeds in the same way as the Accuracy metric, with threshold values (1, 2, 3, 4, or 5) used to construct a prediction rule that translates an NCA or FTA risk score into discrete binary predictions. The Balanced Accuracy metric can be used to evaluate the PSA for both overall and equitable validity by analyzing the metric for the overall study population as well as subgroup comparisons. The figure below plots the Balanced Accuracy metric for each of the three outcome events: NCA, NVCA, and FTA.

[53] Elazmeh, William, Nathalie Japkowicz, and Stan Matwin. "Evaluating misclassifications in imbalanced data." In European Conference on Machine Learning, pp. 126-137. Springer, Berlin, Heidelberg, 2006.

[54] Mohr, Johannes, Sambu Seo, and Klaus Obermayer. "A classifier-based association test for imbalanced data derived from prediction theory." In 2014 International Joint Conference on Neural Networks (IJCNN), pp. 487-493. IEEE, 2014.

[55] Specifically, the metric is the sum of category correct predictions divided by total category predictions, then divided by number of outcome categories.

Figure 18: Balanced Accuracy Measurements for NCA By Hypothetical Prediction Score Thresholds

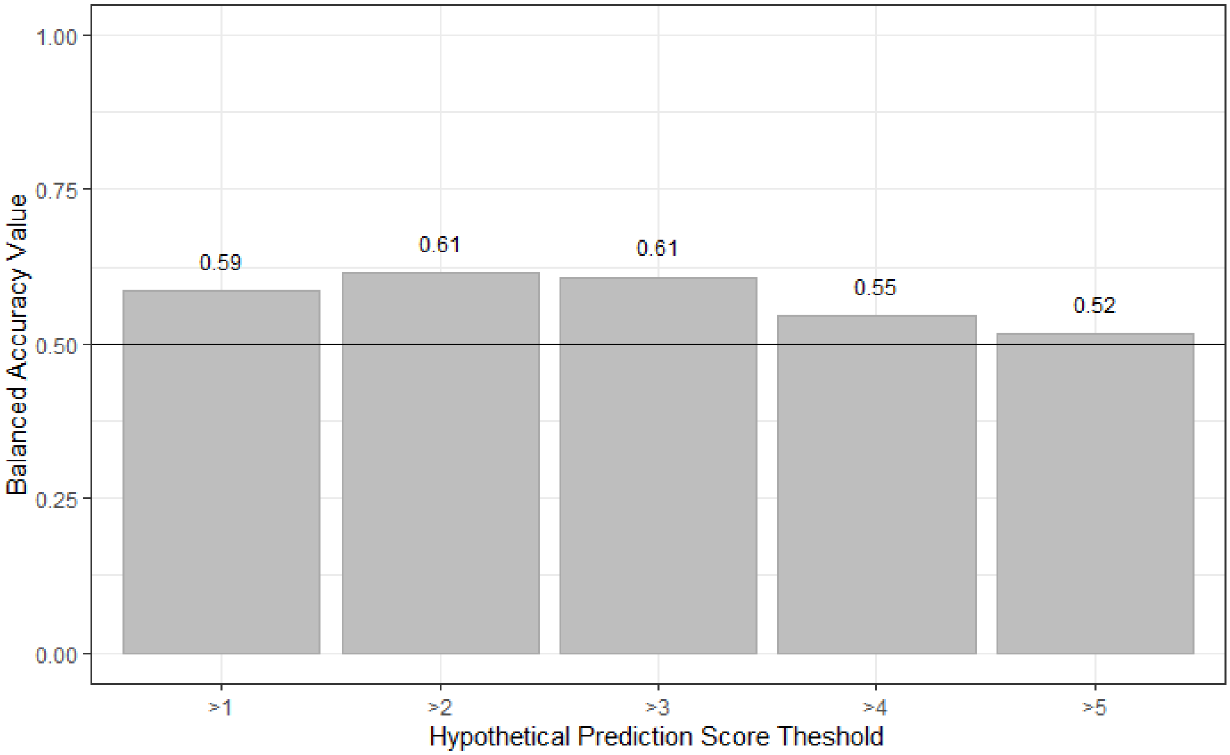


Figure 18 reports Balanced Accuracy measures for NCA outcomes using NCA score thresholds of 1, 2, 3, 4, and 5. The figure above shows that 4 of the 5 hypothetical prediction thresholds obtain balanced accuracy measures higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, increases predictive power beyond classifying cases with no information beyond outcome distribution. These findings are consistent for both the overall study population as well as for each of the main study demographic groups (see Appendix C for relevant demographic based figures). Overall, this figure provides some evidence supporting the validity of the PSA with respect to NCA outcomes. The 2 and 3 thresholds classify with greater Balanced Accuracy, with metric values above 0.6, suggesting a modest gain in classification information.

Figure 19: Balanced Accuracy Measurements for NVCA Outcome Constructions By Possible Prediction Score Threshold

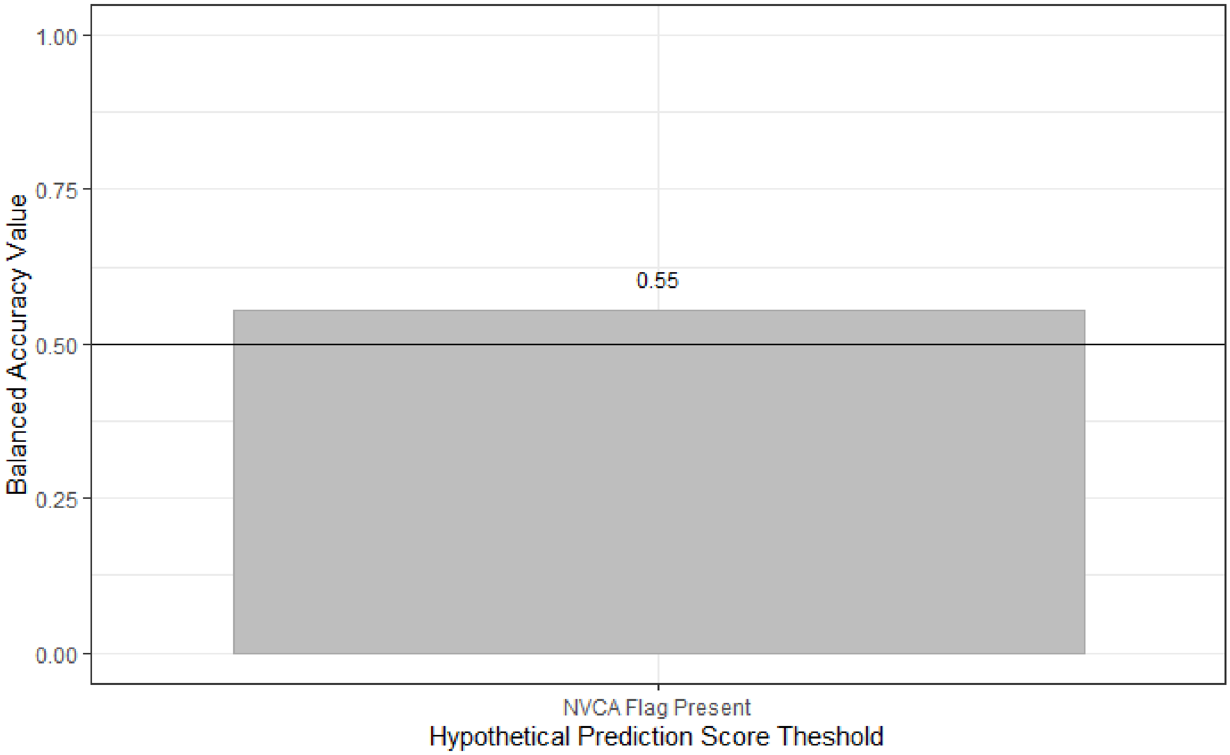


Figure 19 reports Balanced Accuracy measures for NVCA outcomes. The figure above shows that under this hypothetical prediction rule, the PSA NVCA Risk Flag obtains a balanced accuracy measure slightly higher than 0.5, indicating that the PSA provides a marginal increase in predictive power beyond classifying cases on limited information. Overall, this figure provides some, but weak, evidence supporting the validity of the PSA with respect to NVCA outcomes.

Figure 20: Balanced Accuracy Measurements for FTA Outcome Constructions By Possible Prediction Score Thresholds

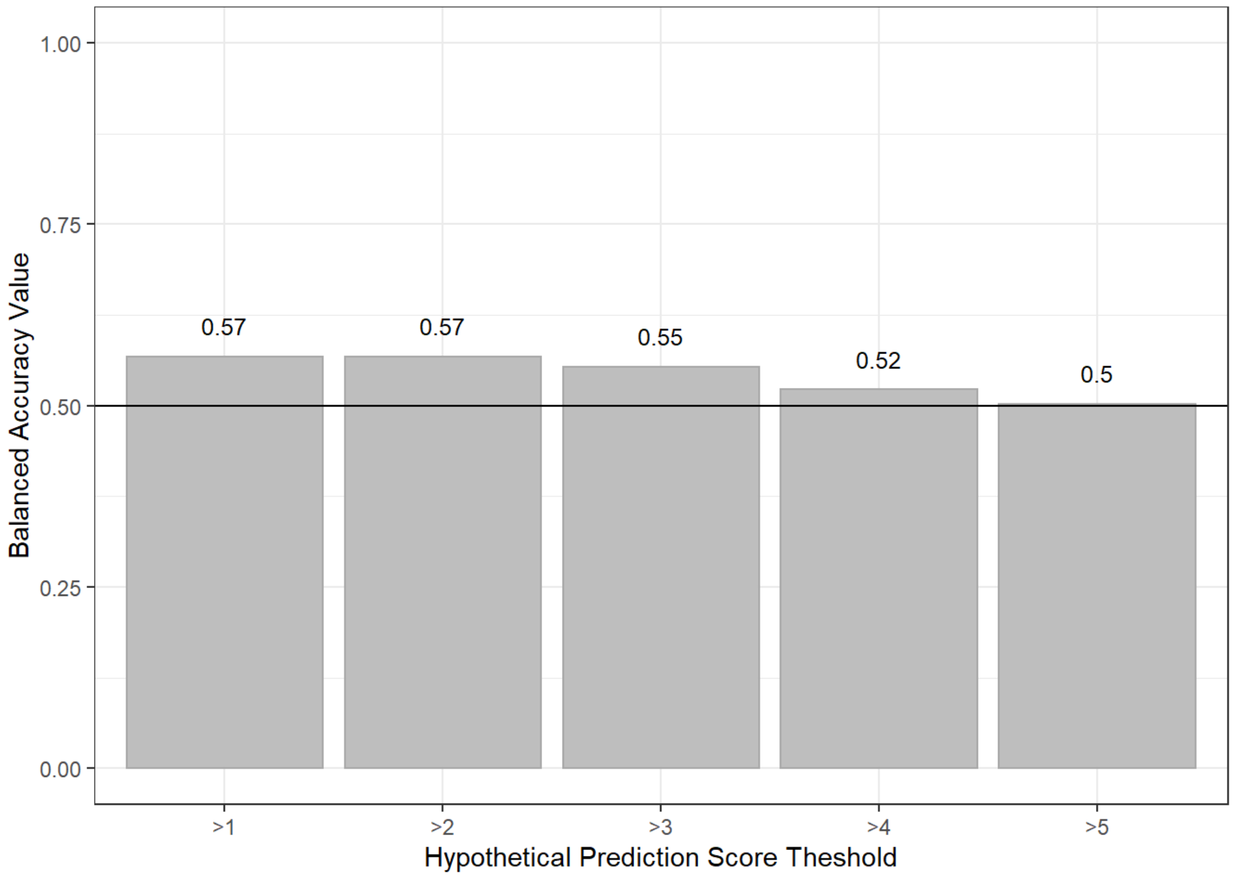


Figure 20 reports Balanced Accuracy measures for the Base FTA outcome construction for FTA score thresholds of 1, 2, 3, 4, and 5. The figure above shows that three of the five hypothetical prediction thresholds obtain balanced accuracy measures slightly higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, provides marginal increases in classifying power. Overall, this figure provides some, but weak, evidence supporting the validity of the PSA with respect to FTA outcomes.

Figure 21: Balanced Accuracy Measurements for FTA+ Outcomes By Possible Prediction Score Thresholds

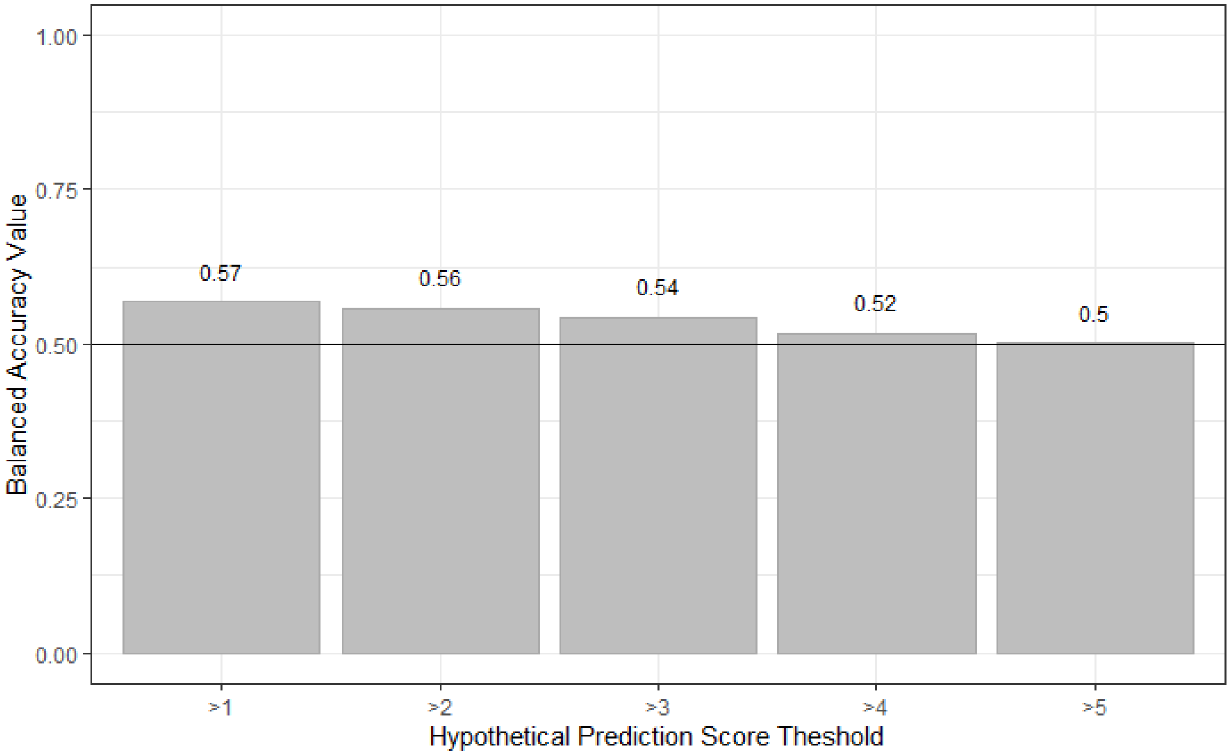


Figure 21 reports Balanced Accuracy measures for the FTA+ outcome construction for FTA score thresholds of 1, 2, 3, 4, and 5. The figure above shows that three of the five hypothetical prediction thresholds obtain balanced accuracy measures slightly higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, provides marginal increases in classifying power. A few of the values are 1/100th less than their Base FTA counterparts, but with respect to the evaluative threshold, the findings are identical between the FTA+ and Base FTA outcome constructions. Overall, this figure provides some, but weak, evidence supporting the validity of the PSA with respect to FTA outcomes.

Figures 18-21 show each hypothetical threshold rule for calculating the Balanced Accuracy metric across all outcome constructions, with the NCA and FTA calculations for all possible threshold values (1, 2, 3, 4, and 5). All classifications achieve some classification gain, as evidenced by Balanced Accuracy scores above 0.5. Some exceed the 0.6 value, suggesting modest classification gains. For NCA outcomes, the Balanced Accuracy metric achieves its maximum under the NCA Score >2 threshold of 0.614 with a minimum of 0.517 under the NCA Score > 5 threshold. Base FTA achieves its maximum Balanced Accuracy metric under the FTA Score > 1 threshold of 0.567 with a minimum of 0.503 under the FTA Score > 5 threshold (these values are the same under the FTA+ outcome construction). The fact that the PSA scores show lower Balanced Accuracy values with the threshold of 5 is unsurprising, given the understandable reluctance of Harris Pretrial to conclude that an individual had prior FTA, resulting in a vanishingly rare number of individuals' being assigned either an NCA or an FTA score of 6. The NVCA outcome Balanced Accuracy metric is 0.539.

Figure 22: Balanced Accuracy Measurements for NCA Across Demographic Subgroups

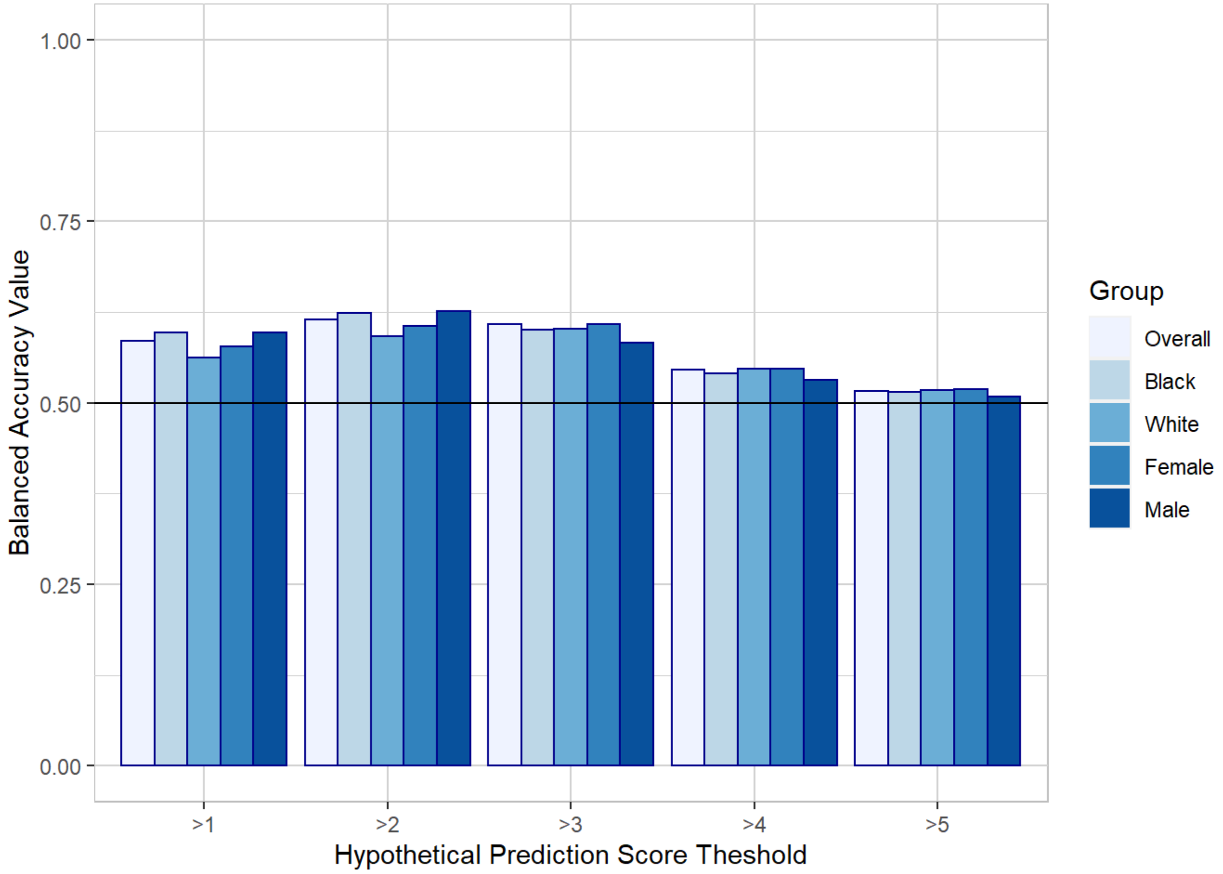


Figure 22 reports Balanced Accuracy measures NCA outcomes. The figure above shows that 4 of the 5 hypothetical prediction thresholds obtain balanced accuracy measures higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, provides increases in predictive power beyond randomly classifying cases. These findings are consistent for both the overall study population as well as for each of the main study demographic groups. Overall, this figure provides evidence supporting the validity of the PSA with respect to NCA outcomes and no indication of meaningful differences in predictive power across race or gender groups.

Figure 23: Balanced Accuracy Measurements for NVCA Outcomes Across Demographic Subgroups

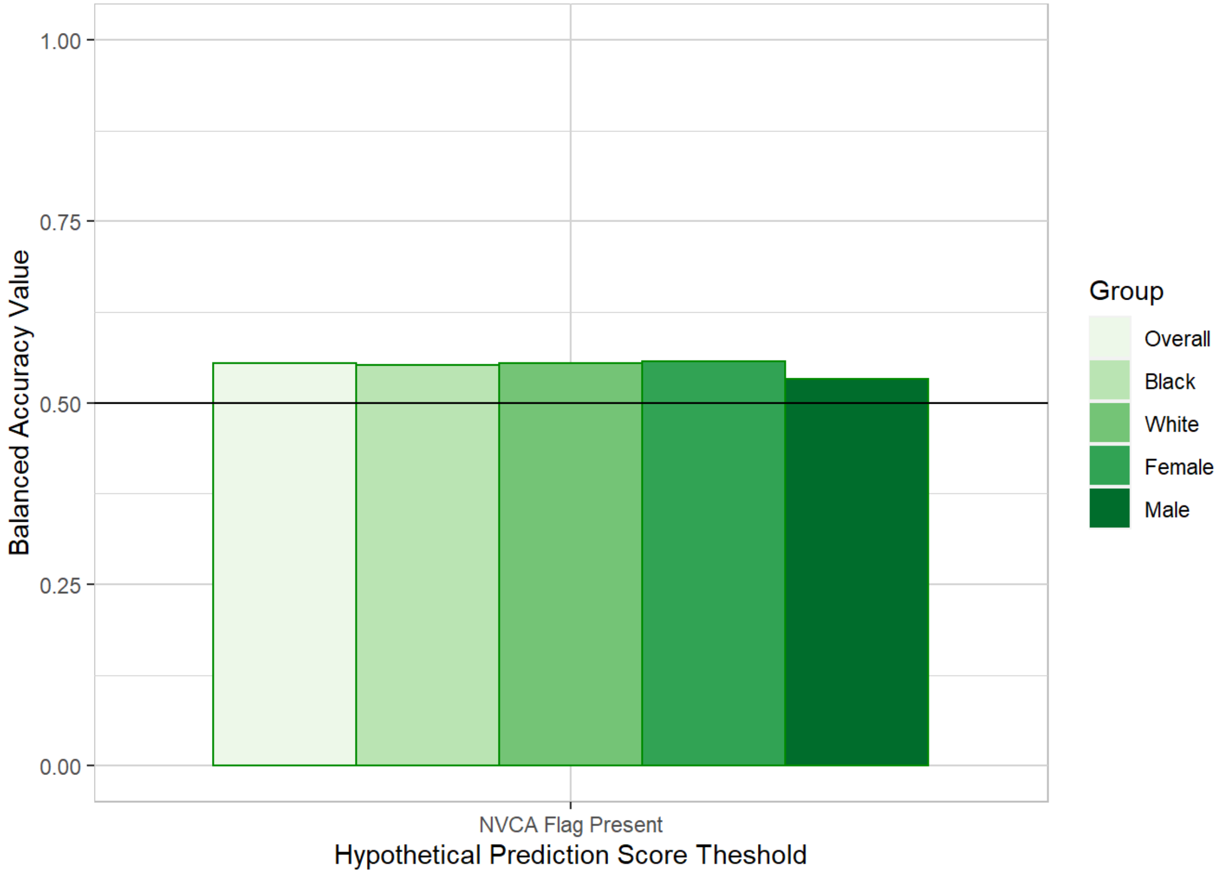


Figure 23 reports Balanced Accuracy measures NVCA outcomes. The figure above shows very little predictive power is gained under the hypothetical prediction threshold of classifying cases on the basis of the presence of the NVCA Flag. These findings are consistent for both the overall study population as well as for each of the main study demographic groups. Overall, this figure provides evidence supporting the validity of the PSA with respect to NVCA outcomes and no indication of meaningful differences in predictive power across race or gender groups.

Figure 24: Balanced Accuracy Measurements for Base FTA Outcomes Across Demographic Subgroups

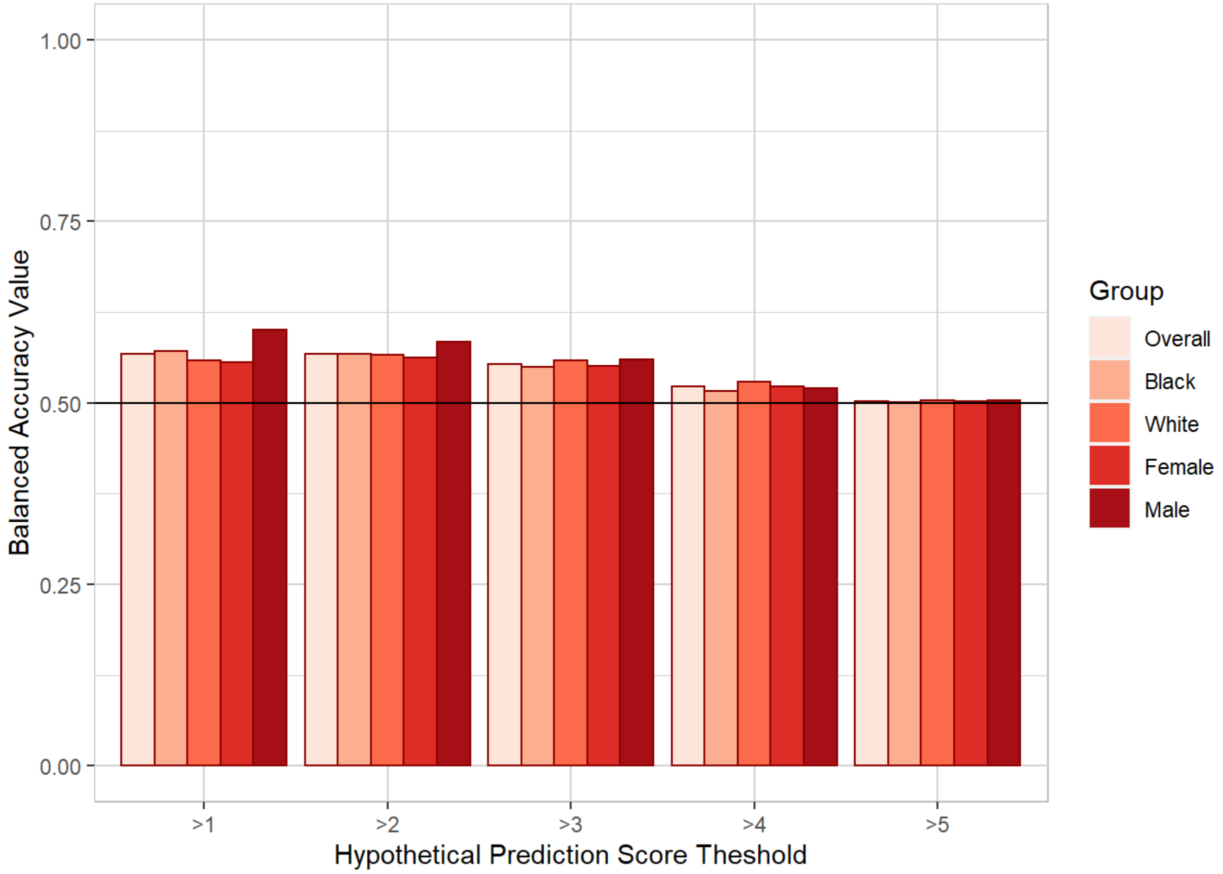


Figure 24 reports Balanced Accuracy measures the Base FTA outcome construction. The figure above shows that 4 of the 5 hypothetical prediction thresholds obtain balanced accuracy measures higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, provides increases in predictive power beyond randomly classifying cases. These findings are consistent for both the overall study population as well as for each of the main study demographic groups. Overall, this figure provides evidence supporting the validity of the PSA with respect to Base FTA outcomes and no indication of meaningful differences in predictive power across race or gender groups.

Figure 25: Balanced Accuracy Measurements for FTA+ Outcomes Across Demographic Subgroups

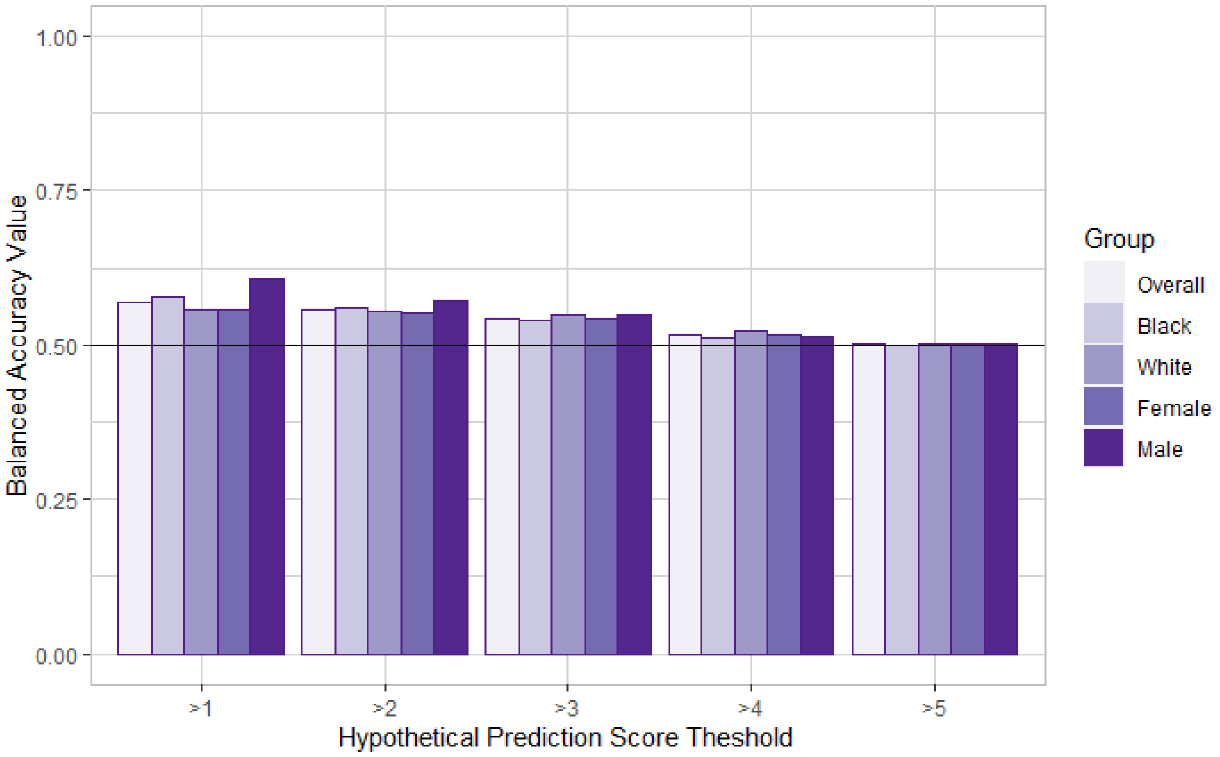


Figure 25 reports Balanced Accuracy measures for the FTA+ outcome Construction. The figure above shows that 4 of the 5 hypothetical prediction thresholds obtain balanced accuracy measures higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, provides increases in predictive power beyond randomly classifying cases. These findings are consistent for both the overall study population as well as for each of the main study demographic groups. Overall, this figure provides evidence supporting the validity of the PSA with respect to FTA+ outcomes and no indication of meaningful differences in predictive power across race or gender groups.

The Balanced Accuracy metric can additionally be used to evaluate the PSA under the equitable validity framework in much the same way as the Area Under the Curve analysis. By comparing paired subgroup population values for Balanced Accuracy, we can evaluate whether the PSA provides differential gains in predictive power for different subgroup populations. For NCA outcomes, the maximum difference in Balanced Accuracy across racial groups is 0.035; the relevant maximum differences for NVCA outcomes and Base FTA outcomes are 0.002 and 0.02, respectively. Comparing across gender groups, the maximum differences are 0.025, 0.024, and 0.032 for NCA, NVCA, and Base FTA outcomes, respectively. Each of these maximum differences represents minor differences in Balanced Accuracy metrics. Overall, the Balanced Accuracy metric provides some evidence of equitable validity for the PSA.

e. Validation by Racial And Gender Groups

i. PSA scores and failure rates by race

This subsection reports the results of a comparison by race and gender of PSA scores and corresponding failure rates. There are a few statistically significant differences by race and gender, but those differences are either in inconsistent directions or are substantively small. The analysis provides some evidence that the PSA is equitably valid, and no evidence to the contrary. Key details are as follows.

- There exists significant differences in failure rates across racial demographic groups for all NCA risk scores, the NVCA flag, and four FTA risk scores. But
  - racial group differences in N(V)CA/FTA failure rates are directionally mixed, with White arrestees observing higher FTA failure rates but lower N(V)CA failure rates than their Black arrestee counterparts; and
  - these group differences are small in magnitude, ranging between 1-3 percentage points for racial group differences and 1-5 percentage points for gender group differences.
- There exists significant differences in failure rates across gender demographic groups for four NCA risk scores, the NVCA flag, and one of the FTA risk scores. Female arrestees observe lower rates of failure in each of these instances.

Differences in failure rates for each PSA score category are additionally calculated across study demographic groups: Black individuals, White individuals, male individuals, and female individuals. Statistically significant differences in classification failure rates across demographic groups indicate that the same risk score relays different information depending on the demographic of the individual. We again use differences of proportion tests to analyze the statistical difference between failure rates for relevant demographic subpopulation comparisons (race and gender) at fixed risk score levels. Few or no reported significant differences would provide strong evidence for equitable validity.[56] [57] The following figures plot outcome failure rates by relevant risk score across study demographic groups for each of the main outcome events.

[56] DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. "The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky." Available at SSRN 3168452 (2018).

[57] DeMichele, M, Baumgartner, P, Wenger, M, Barrick, K, Comfort, M. Public safety assessment: Predictive utility and differential prediction by race in Kentucky. Criminal Public Policy. 2020; 19: 409–431.

Figure 26: NCA Failure Rates Across Risk Score by Study Demographic Group

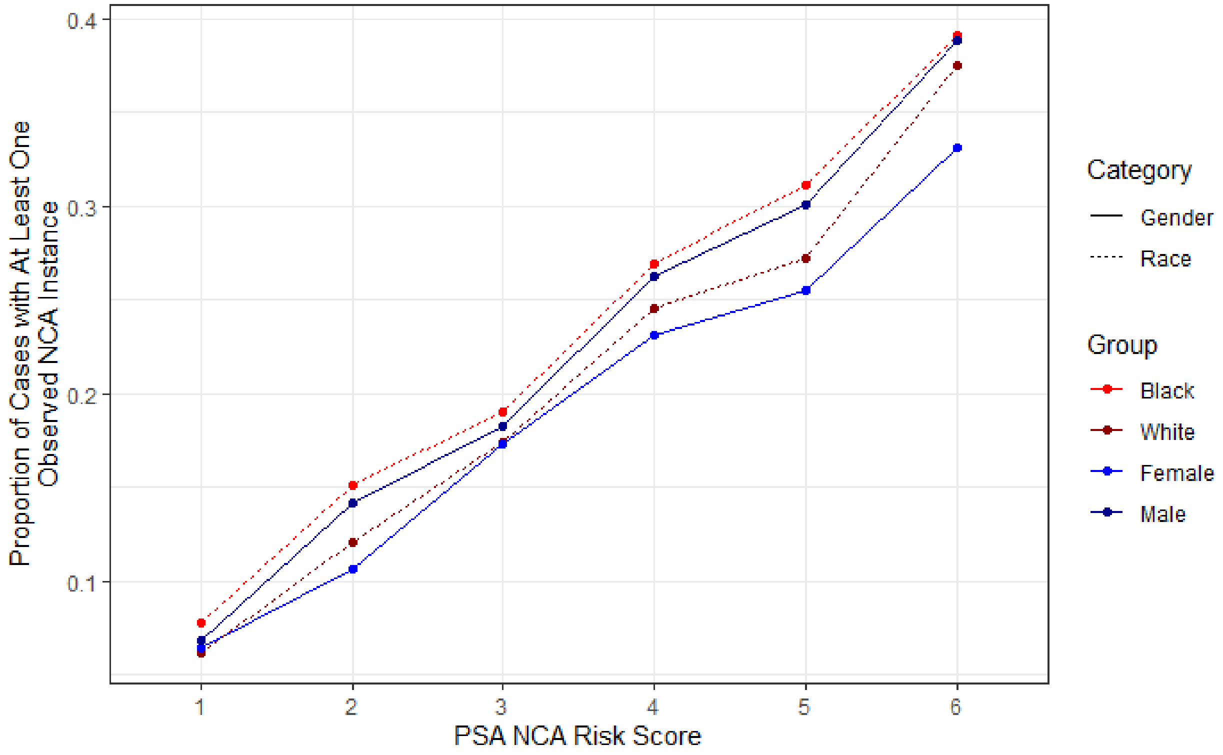


Figure 26 reports the observed failure rate for NCA outcomes across PSA NCA Risk Score categories by the four main study demographic groups: Black individuals, White individuals, Male individuals, and Female individuals. Line types and colors differ across category comparisons (Race and Gender). These results allow us to gauge the overall validity of the PSA across demographic subgroups as well as assess any differential impact between racial subgroup pairings and gender subgroup pairings. For NCA outcomes, there exist significant differences in observed failure rates between Black individuals and White individuals for each level of the NCA risk score (as well as overall observed failure rates) with the exception of NCA scores of 6, which do not show significant differences in failure rates between racial subgroups. At each level of NCA risk score (with the exception of scores of 6), Black individuals have observed failure rates 1-2 percentage points higher than their White individual counterparts. With respect to gender comparisons, there exist significant differences in observed failure rates between Male and Female individuals only for NCA risk score categories of 2, 4, and 5 (as well as overall observed failure rates). These differences range from 3 to 5.5 percentage points with Female individuals observing lower failure rates than their Male individual counterparts in each instance. This figure provides support both for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates), as well as for significant differences in observed failure rates for both racial and gender subgroups with respect to NCA events.

Figure 27: NVCA Failure Rates Across Risk Flag Presence By Study Demographic Group

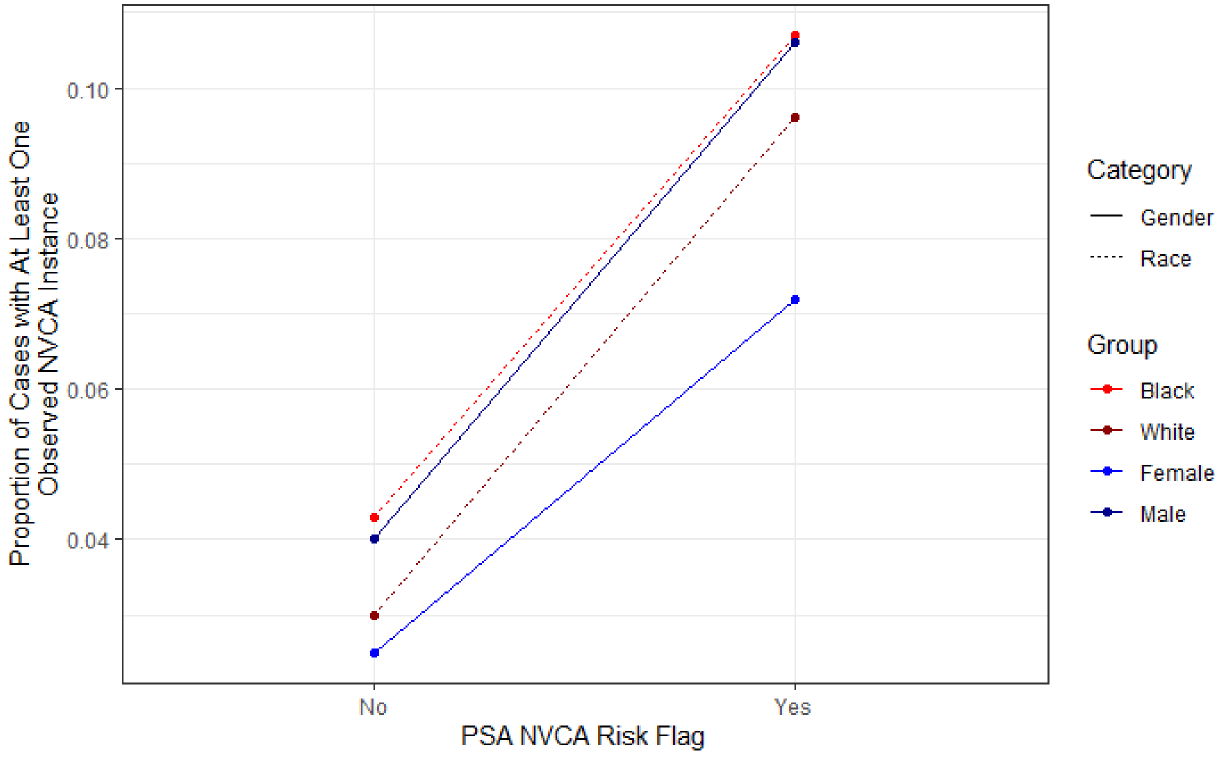


Figure 27 reports the observed failure rate for NVCA outcomes across categories of PSA NVCA Risk Flag presence by the four main study demographic groups: Black individuals, White individuals, Male individuals, and Female individuals. Line types and colors differ across category comparisons (Race and Gender). These results allow us to gauge the overall validity of the PSA across demographic subgroups as well as assess any differential impact between racial subgroup pairings and gender subgroup pairings. For NVCA outcomes, there exist significant differences in observed failure rates between Black individuals and White individuals only when the NVCA Risk Flag is not present. For cases without the NVCA risk flag, Black individuals have observed failure rates 1-2 percentage points higher than their White individual counterparts. With regards to gender comparisons, there exist significant differences in observed failure rates between Male and Female individuals both when the NVCA risk flag is present and when it is not. These differences range from 1.5 to 4 percentage points with Female individuals observing lower failure rates than their Male individual counterparts in each instance. This figure provides support both for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates), as well as for significant differences in observed failure rates for both racial and gender subgroups with respect to NVCA events.

Figure 28: Base FTA Failure Rates Across Risk Score By Study Demographic Group

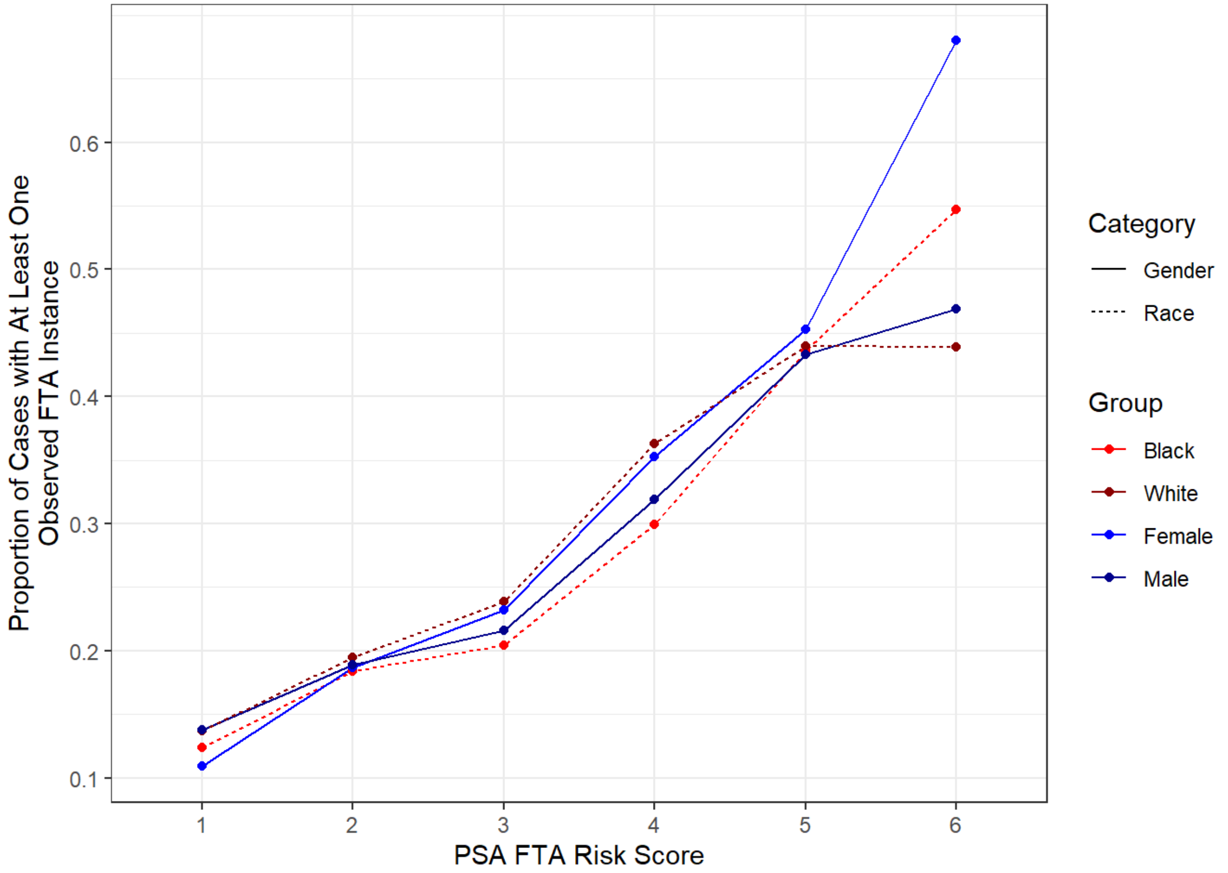


Figure 28 reports the observed failure rate for the Base FTA construction across PSA FTA Risk Score categories by the four main study demographic groups: Black individuals, White individuals, Male individuals, and Female individuals. Line types and colors differ across category comparisons (Race and Gender). These results allow us to gauge the overall validity of the PSA across demographic subgroups as well as assess any differential impact between racial subgroup pairings and gender subgroup pairings. There exists significant differences in observed failure rates between Black individuals and White individuals for FTA scores of 1, 2, 3, and 4. At these levels of FTA risk score, Black individuals have observed failure rates 1.3-4.5 percentage points lower than their White individual counterparts. There additionally exists significant differences in observed failure rates between Male and Female individuals only for the FTA risk score category of 1. These differences are about 3 percentage points, with Female individuals observing lower failure rates than their Male individual counterparts. This figure provides support both for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates), as well as for significant differences in observed failure rates for racial subgroups only at the middle of the FTA risk scale and for gender subgroups only at the lowest end of the FTA risk scale.

Figure 29: FTA+ Failure Rates Across Risk Score By Study Demographic Group

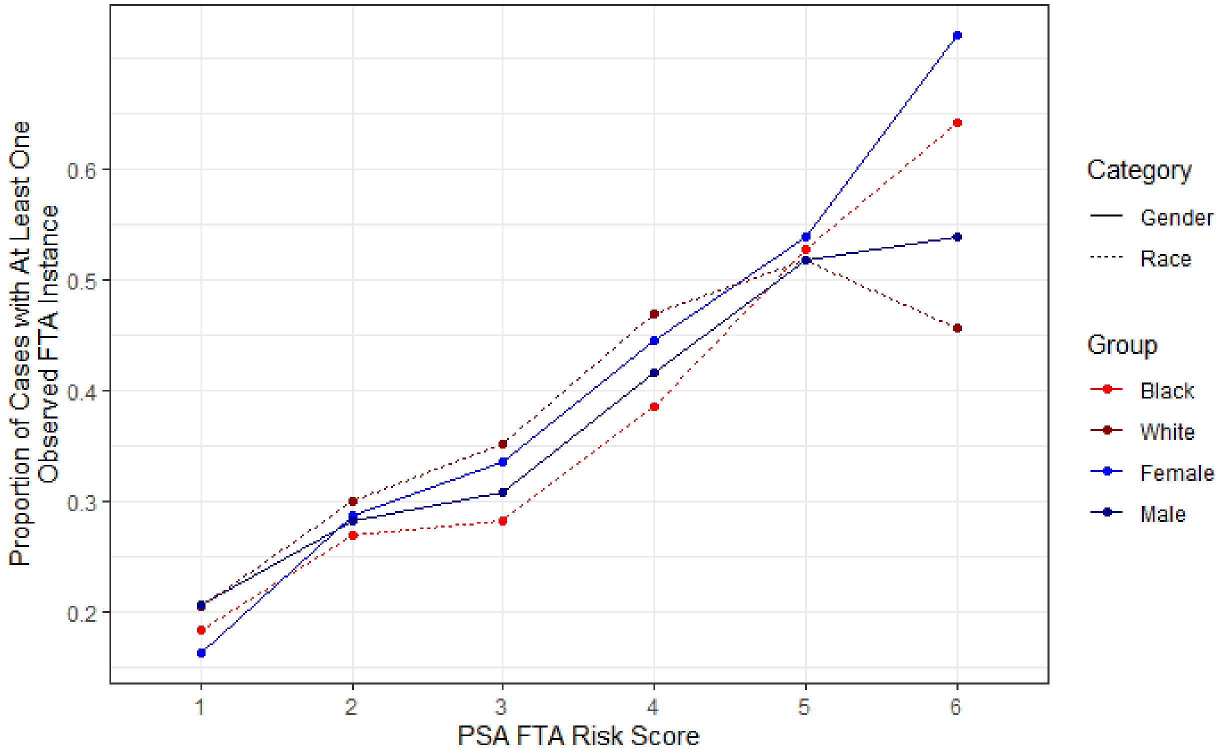


Figure 29 reports the observed failure rate for the FTA+ construction across PSA FTA Risk Score categories by the four main study demographic groups: Black individuals, White individuals, Male individuals, and Female individuals. Line types and colors differ across category comparisons (Race and Gender). These results allow us to gauge the overall validity of the PSA across demographic subgroups as well as assess any differential impact between racial subgroup pairings and gender subgroup pairings. There exists significant differences in observed failure rates between Black individuals and White individuals for FTA scores of 1, 2, 3, and 4. At these levels of FTA risk score, Black individuals have observed failure rates 2-8.3 percentage points lower than their White individual counterparts. There additionally exists significant differences in observed failure rates between Male and Female individuals only for the FTA risk score category of 1. These differences are about 4.3 percentage points, with Female individuals observing lower failure rates than their Male individual counterparts. This figure provides support both for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates), as well as for significant differences in observed failure rates for racial subgroups only at the middle of the FTA risk scale and for gender subgroups only at the lowest end of the FTA risk scale.

Figures 26-29 break out the Failure Rate by PSA Risk Score category analysis by demographic subgroups, allowing for an evaluation of the equitable validity of the PSA. The figures indicate that while failure rates tend to move similarly across demographic subgroups, there is meaningful separation. Additionally, cases with FTA scores of six exhibit significant divergence from one another; however, this is likely due, again, to the very small number of cases assessed at an FTA risk score of 6. For NCA outcomes, there exist significant ($p<0.05$) racial group differences in failure rates at NCA scores of 1, 2, 3, 4 and 5 while significant gender group differences in failure rates exist at NCA scores of 2, 4, and 5. For NVCA outcomes significant racial group differences in failure rates exist for cases with no violence flag and for gender group differences in failure rates for both cases with and without the violence flag. For Base FTA outcomes, significant racial group differences in failure rates exist for cases with FTA risk scores of 1, 2, 3, and 4 and for gender group differences in failure rates for cases with FTA scores of 1. These differences are also significant for the FTA+ outcome construction, but larger in magnitude. There are a large number of scores that imply statistically different rates of failure for either race or gender pairs: all levels of NCA scores except 6, both categories of the NVCA Flag, and FTA scores of 1,2, 3, and 4. However, despite the statistical significance of these differences, most of the subgroup specific failure rates are still near one another, differing often only by 1 or 2 percentage points and at most by 4.5 percentage points. However, this differs under the FTA+ outcome construction, where racial differences in the middle of the FTA scale (3 or 4) are around 6.8 and 8.3 percentage points, which is substantive. Moreover, with regards to race, there is no consistent pattern with respect to difference in classifying information. White failure rates were lower than corresponding Black rates with respect to NCA and NVCA, but higher with respect to FTA. Ultimately, the subgroup paired comparison of PSA score specific failure rates provide no meaningful evidence that the PSA does not equitably validate; significant differences exist, but they are either substantively small or inconsistent in direction.

2. Moderated regression

This section provides the results of a moderated regression analysis to assess equitable validity. This analysis shows some statistically significant differences across racial groups, but the size of those differences is small. Thus, this analysis provides some evidence of, and little evidence to contract, equitable validity. Key details are as follows:

- Each of the PSA risk scores/flags show significant, positive correlations with the probability of observing a relevant outcome of roughly similar magnitudes to the bivariate regression.
- The interaction between race and risk score, which would indicate whether the predictive meaning of the risk score changes significantly across racial groups, is significant only for NCA scores. Further, the NCA score - race interaction is substantively small, with a point estimate of about a 4% decrease in the odds of a Black arrestee observing an NCA relative to a White arrestee.

Moderated regression provides a way of jointly testing the base classification power of the PSA risk score on the relevant outcome as well as the classification power accounting for potential moderating effects of important demographic variables.[58] In simpler terms, we fit a model with just the PSA scores and assess how well the scores relate to failure outcomes. Then, we fit a model with the PSA scores and other variables, especially demographic variables, and examine whether using all of these variables results in a stronger relationship to failure outcomes. If so, then we have some evidence that the PSA classifications may operate differently by

[58] DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. "The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky." Available at SSRN 3168452 (2018).

demographic group.[59] We focus on race, as opposed to gender, due to the fact that the racial distinctions appeared to correspond to greater differences in the previous section. The following figures plot predicted probabilities obtained from the various outcome events regressed under PSA risk score scales plus race variables for each of the main outcome events.

Figure 30: Predicted NCA Probabilities from Moderated Regression Model

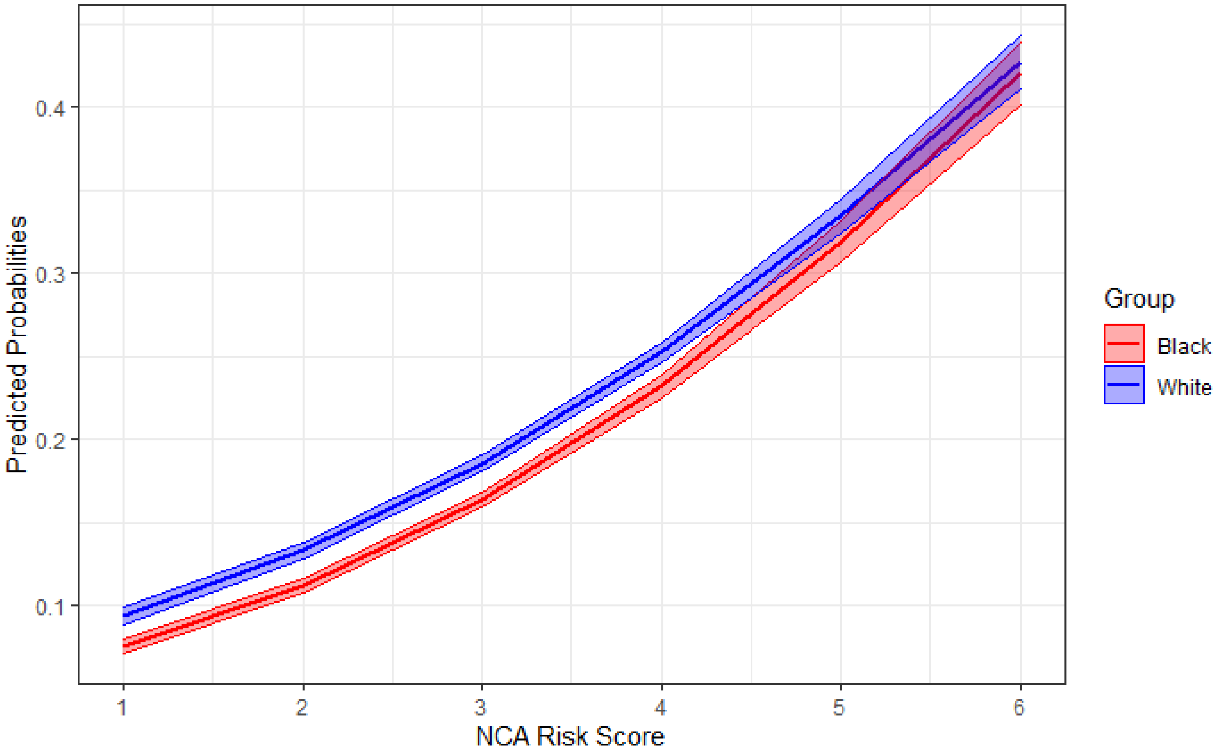


Figure 30 reports predicted probabilities and associated 95% confidence intervals for observing an NCA event obtained from the moderated regression model with both PSA score and race variables. The PSA NCA risk score has a significant, positive coefficient, indicating that higher NCA risk scores are significantly, independently associated with a higher probability of an observed NCA failure. A one unit increase in the NCA risk score is associated with a 55% increase in the odds of observing an NCA failure

[59] More specifically, the moderated regression framework proceeds in four steps: the first model regresses only the hypothesized moderating variable on the outcome; the second model regresses only the risk score variable on the outcome; the third model regresses both the hypothesized moderating variable as well as the risk score variable on the outcome variable; and the fourth model regresses the hypothesized moderating variable, the risk score variable, and an interaction between the two on the outcome variable. By evaluating the risk assessment score coefficient across these separate models, we can determine the impact of including a potentially moderating variable, such as race, on how the assessment score relates to the relevant outcome. Evaluating the basic value of the risk score can be done by analysing the size and significance of the risk score coefficient in models 2 and 3, while the potential moderating effects can be gauged by analyzing the significance of the interaction coefficient in model 4. Analyzing the risk score coefficient in models 2 and 3 replicates the analysis in Section III.C.1. Instead, this section focuses on evaluating overall and equitable validity by evaluating the model estimates from model 4. The estimated coefficients from the risk score and interactive term can provide evidence as to whether the PSA scores provide meaning information about the occurrence of relevant outcomes within the context of additionally knowing racial demographic data and whether this information is meaningful moderated by membership in a racial demographic group.

versus not observing an NCA failure. This estimate exists on a confidence interval from a 51% increase in the odds ratio to a 59% increase in the odds ratio. The interaction term is also significant (0.96 odds ratio on a 95% CI of 0.93 - 0.99), indicating that a one unit increase in the NCA scale is associated with between a 1% and 7% decrease in the odds ratio of observing an NCA event for Black individuals relative to their White individual peers. Functionally, this means that when taking into account racial categories, Black individuals have lower predicted probabilities for observing an NCA than their White peers when only looking at NCA scores, which themselves are overall statistically significantly predictive of observed NCA events. The shaded confidence interval regions illustrate the moderating impact of race: at lower levels of the NCA Risk Score Scale, there is no overlap in predictive probabilities for Black and White individuals, indicating that the scores are meaningfully different between the groups. However, at the upper ends, the confidence interval regions overlap, indicating that the predicted probabilities cannot be statistically distinguished from one another. Overall, this figure provides support for both the overall validity of the PSA and for the inference that there are statistically significant differences in predictive strength across racial subgroups, although differences are of modest size.

Figure 31: Predicted NVCA Probabilities from Moderated Regression Model

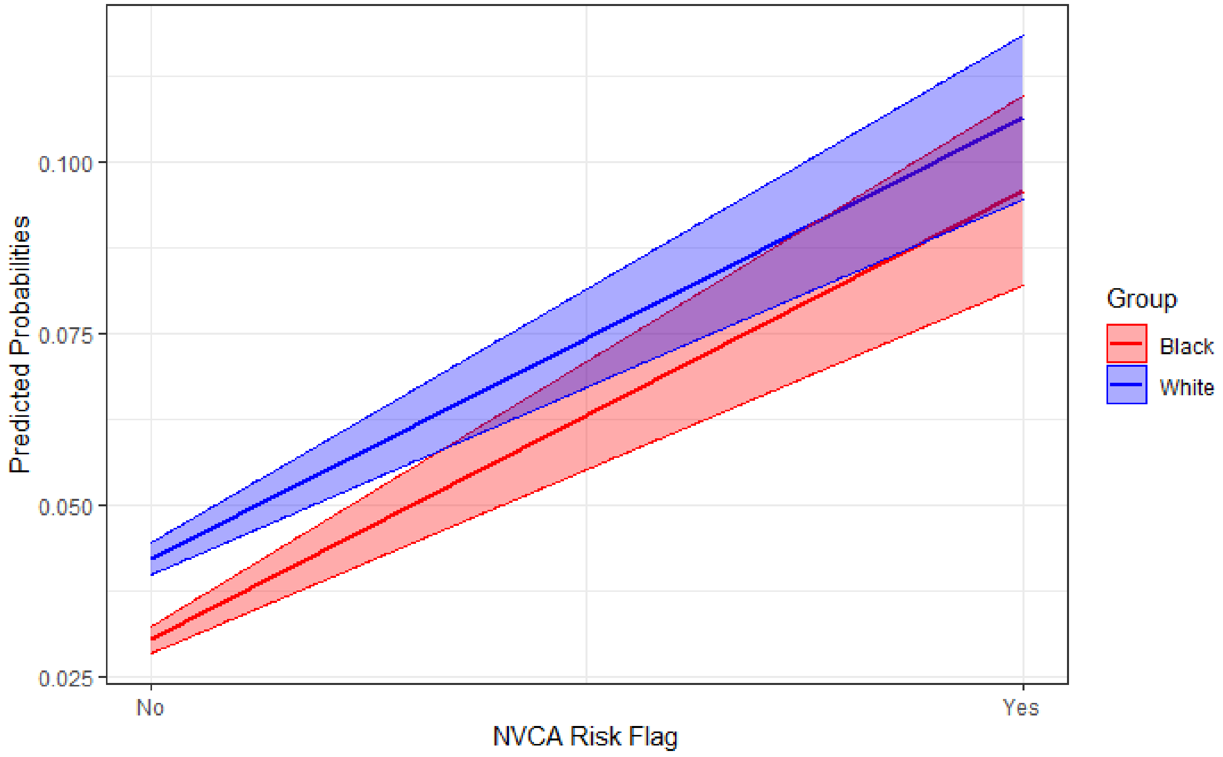


Figure 31 reports predicted probabilities and associated 95% confidence intervals for observing an NVCA event obtained from the moderated regression model with both PSA score and race variables. The PSA NVCA risk flag has a significant, positive coefficient, indicating that the presence of a risk flag is significantly, independently associated with a higher probability of an observed NVCA failure. The presence of an NVCA risk flag is associated with a 237% increase in the odds of observing an NVCA failure versus not observing an NVCA failure. This estimate exists on a confidence interval from a 183% increase in the odds ratio to a 300% increase in the odds ratio. The interaction term is not significant. Functionally, this means that when taking into account racial categories, Black and White individuals have statistically similar predicted probabilities for observing an NVCA. The shaded confidence regions indicate a statistically meaningful difference in predicted probabilities only when the NVCA Flag is not present;

however, the predicted probabilities are indistinguishable when the NVCA Risk Flag is present. Given that the only category asserting 'risk' under the NVCA construction is the Risk Flag present category, this does not provide strong support for the risk flag itself conveying differential information. Overall, this figure provides support for both the overall validity of the PSA, and we see no evidence of racial differences, for the NVCA outcome.

Figure 32: Predicted Base FTA Probabilities from Moderated Regression Model

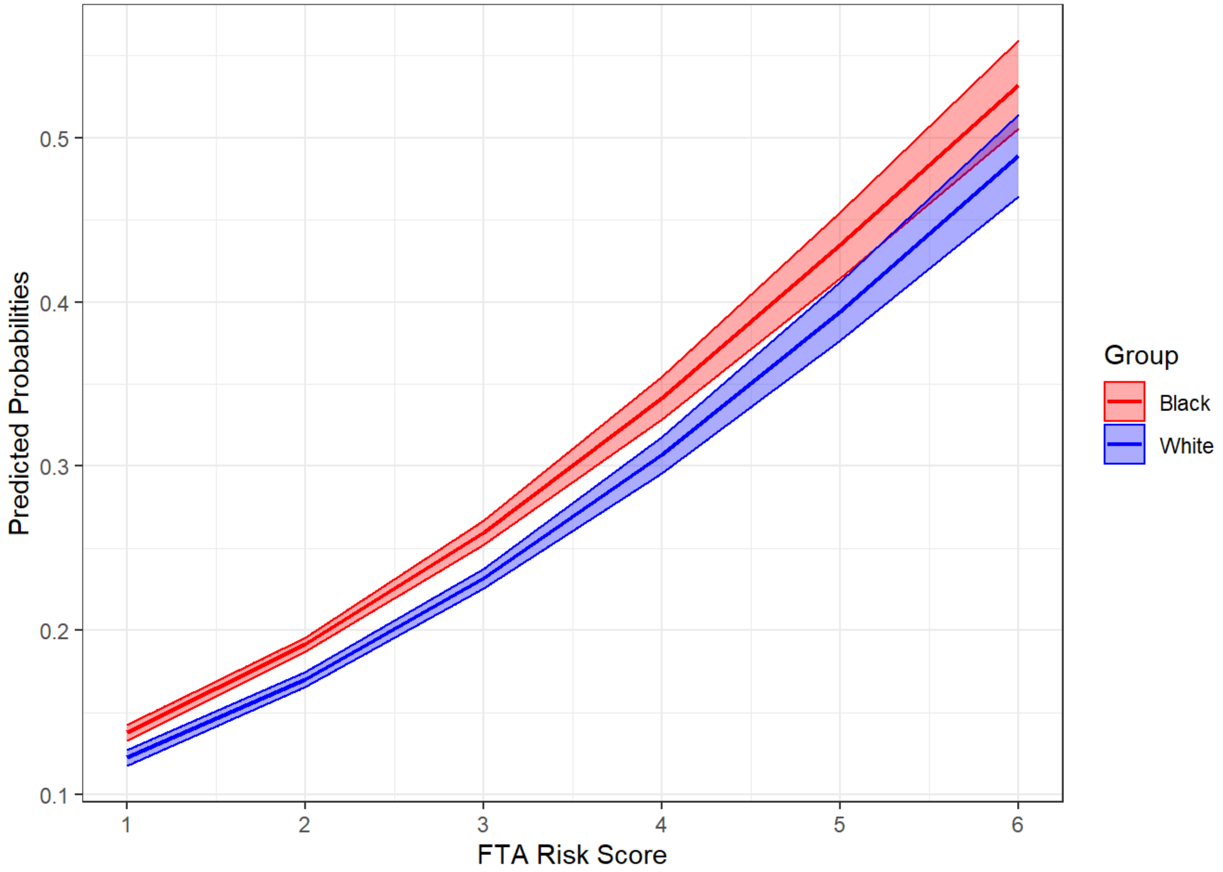


Figure 32 reports predicted probabilities and associated 95% confidence intervals for observing a Base FTA event obtained from the moderated regression model with both PSA score and race variables. The PSA FTA risk score has a significant (at the p<0.001 level), positive coefficient, indicating that higher FTA risk scores are significantly, independently associated with a higher probability of an observed FTA failure. A one unit increase in the FTA risk score is associated with a 48% increase in the odds of observing an FTA failure versus not observing an FTA failure. This estimate exists on a confidence interval from a 44% increase in the odds ratio to a 52% increase in the odds ratio. The interaction of PSA and race is not statistically significant, suggesting that there is no evidence of racial subgroup differences with respect to the FTA outcome. The shaded confidence interval region shows some statistically significant separation of scores by racial group for the lower scales of the FTA Risk Score Scale, but the magnitude of the separation is so minor as to provide little evidence of meaningful difference in information conveyed by the FTA Risk Score. Overall, this figure provides support for the overall validity of the PSA with respect to FTA outcomes and suggests little evidence of racial differences.

Figure 33: Predicted FTA+ Probabilities from Moderated Regression Model

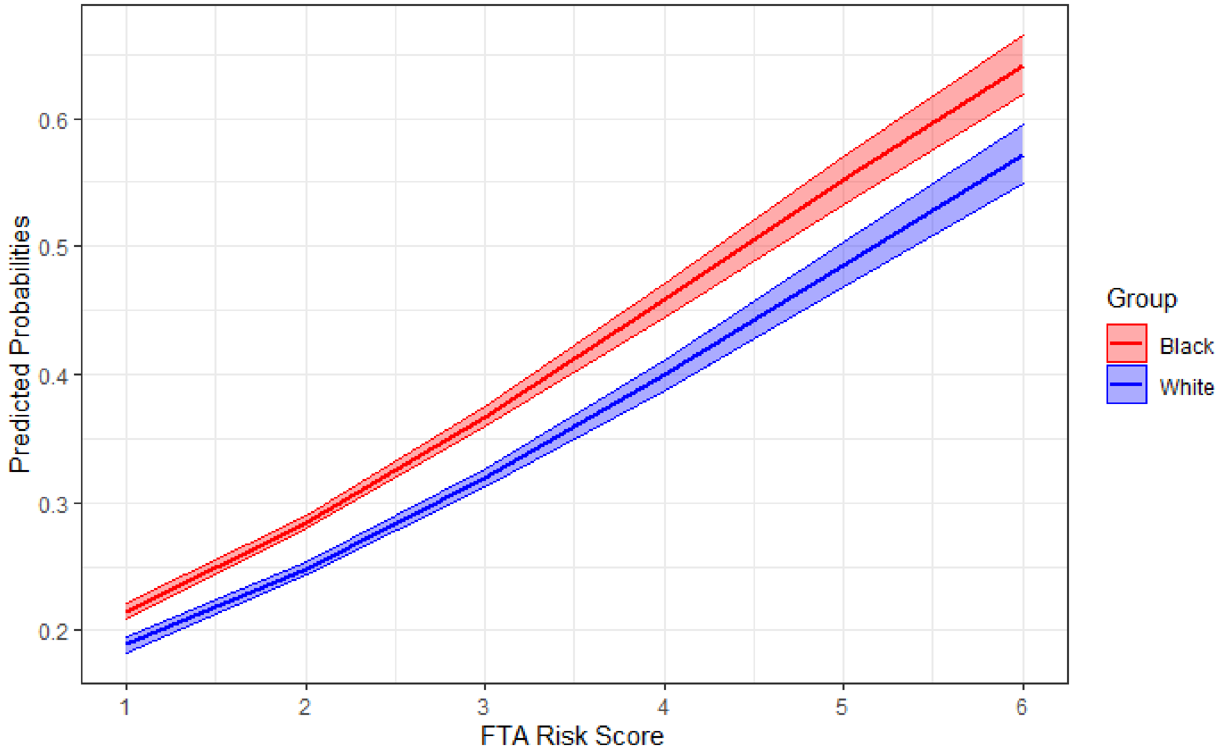


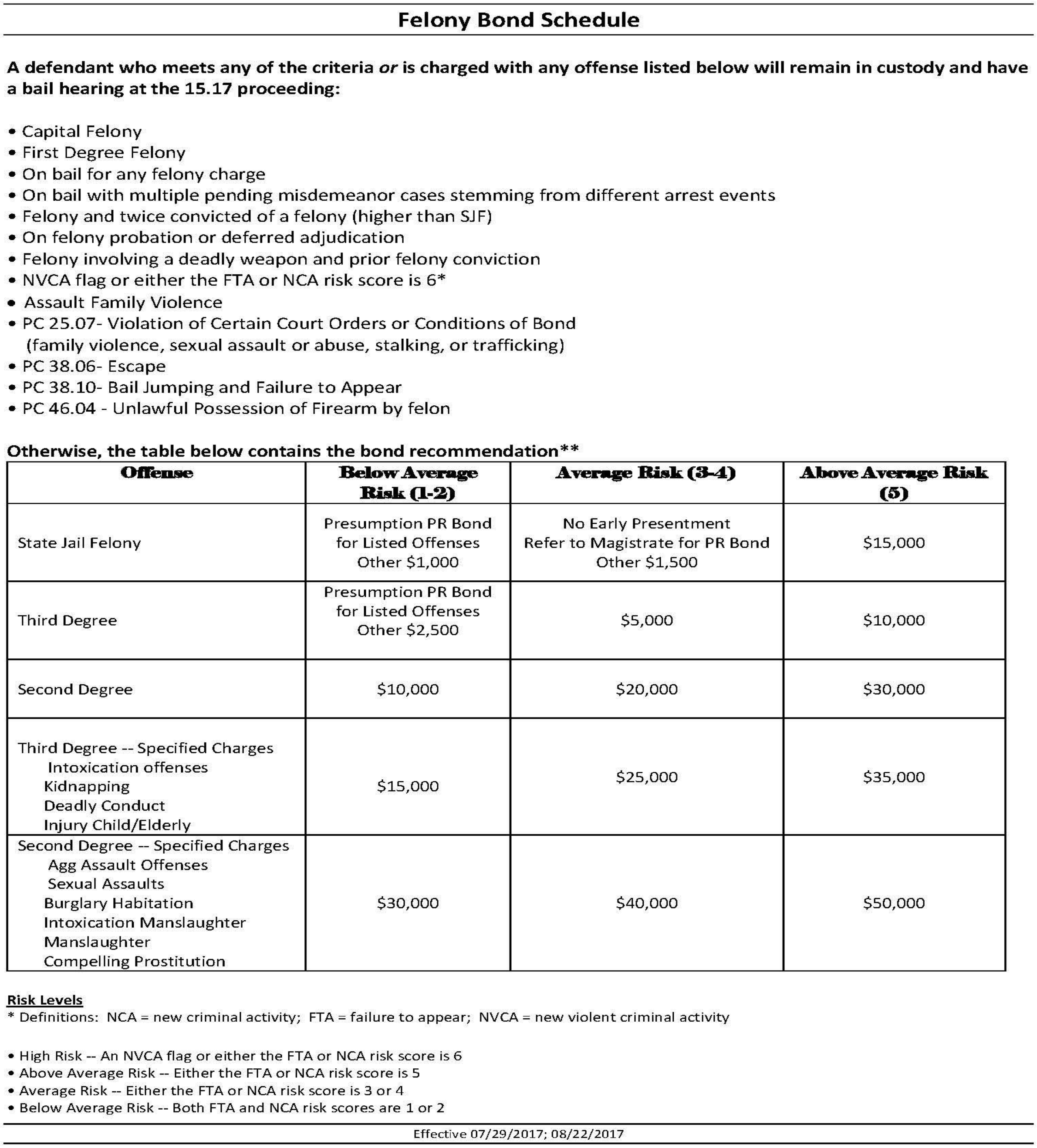
Figure 33 reports predicted probabilities and associated 95% confidence intervals for observing an FTA+ event obtained from the moderated regression model with both PSA score and race variables. The PSA FTA risk score has a significant (at the p<0.001 level), positive coefficient, indicating that higher FTA risk scores are significantly, independently associated with a higher probability of an observed FTA failure. A one unit increase in the FTA risk score is associated with a 46% increase in the odds of observing an FTA failure versus not observing an FTA failure. This estimate exists on a confidence interval from a 42% increase in the odds ratio to a 50% increase in the odds ratio. The interaction of PSA and race is not statistically significant, suggesting that there is no evidence of racial subgroup differences with respect to the FTA outcome. The shaded confidence interval region shows separation of scores by racial group along the FTA Risk Score Scale, but the magnitude of the separation is so minor as to constitute little evidence of meaningful difference in information conveyed by the FTA Risk Score. Overall, this figure provides support for the overall validity of the PSA with respect to FTA outcomes and suggests little evidence of racial differences.

The predicted probabilities shown in Figures 30-33 indicate the same overall increasing pattern (monotonicity) that defined the bivariate logistic regression predicted probabilities discussed earlier. The predicted probabilities here are obtained from the model of the moderated regression framework that includes the relevant risk assessment score scale, a racial group indicator, and the interaction term between the two as independent variables. The consistency of this trend indicates both evidence for the overall validity of the PSA as well as the fact that any moderating effect of race on the PSA risk scores is not significant enough to overwhelm information obtained through utilizing the scores. For the NCA, NVCA, and Base FTA models, the exponentiated coefficients under the moderated regression framework are statistically equivalent to the estimates under the bivariate logistic regression model, *i.e.*, their confidence intervals overlap. For the NCA model, the exponentiated coefficient on the NCA Score Scale is 1.55 on a 95% confidence interval of (1.51, 1.59), while the bivariate estimate was 1.53. For the Base FTA model, the exponentiated coefficient on the FTA Score Scale is 1.48 on a 95% confidence interval of (1.44, 1.52), while the bivariate estimate was 1.46. For the FTA+ model, the exponentiated coefficient on the FTA Score Scale is 1.46 on a 95% confidence interval of (1.42, 1.50), while the bivariate estimate was 1.42. For NVCA, the magnitude of the difference in the exponentiated coefficients between the moderated regression and the bivariate logistic regression appears larger, but the confidence intervals still overlap. The exponentiated coefficient estimate for the presence of the NVCA Flag is 3.37 on a confidence interval of (2.83, 4.00), while the bivariate exponentiated estimate was 3.04. The moderated regression framework ultimately provides the same strong evidence of overall validity for the PSA.

The primary benefit of the moderated regression framework for the purposes of this study is its ability to provide insight as to whether the PSA equitably validates. To the extent that the interaction term is significant, this indicates that information provided by the relevant PSA risk scale score statistically changes when moving from individuals of one racial group to another. For NCA, the interaction term is borderline significant at the $p < 0.05$ level with an exponentiated coefficient estimate of 0.96 on a confidence interval of (0.94, 0.99). This indicates that for each level of the NCA Score Scale the odds ratio of observing at least one NCA event during the pretrial period is about four percent lower for Black individuals than the corresponding odds ratio for the same NCA score for White individuals. This four percent moderating effect of race is about 1/13th the size of the effect of a one unit increase in the NCA score. For both NVCA and FTA (both constructions), the interaction term representing the moderating effect of race on the relevant PSA risk score scale is not statistically significant. Ultimately, when considering the small magnitude and inconsistency of results, the moderated regression framework provides no evidence that the PSA does not equitably validate.

Appendices

Appendix A Felony Bail Schedule



Felony Bond Schedule

A defendant who meets any of the criteria *or* is charged with any offense listed below will remain in custody and have a bail hearing at the 15.17 proceeding:

- Capital Felony
- First Degree Felony
- On bail for any felony charge
- On bail with multiple pending misdemeanor cases stemming from different arrest events
- Felony and twice convicted of a felony (higher than SJF)
- On felony probation or deferred adjudication
- Felony involving a deadly weapon and prior felony conviction
- NVCA flag or either the FTA or NCA risk score is 6*
- Assault Family Violence
- PC 25.07- Violation of Certain Court Orders or Conditions of Bond (family violence, sexual assault or abuse, stalking, or trafficking)
- PC 38.06- Escape
- PC 38.10- Bail Jumping and Failure to Appear
- PC 46.04 - Unlawful Possession of Firearm by felon

Otherwise, the table below contains the bond recommendation**

Offense	Below Average Risk (1-2)	Average Risk (3-4)	Above Average Risk (5)
State Jail Felony	Presumption PR Bond for Listed Offenses Other $1,000	No Early Presentment Refer to Magistrate for PR Bond Other $1,500	$15,000
Third Degree	Presumption PR Bond for Listed Offenses Other $2,500	$5,000	$10,000
Second Degree	$10,000	$20,000	$30,000
Third Degree -- Specified Charges Intoxication offenses Kidnapping Deadly Conduct Injury Child/Elderly	$15,000	$25,000	$35,000
Second Degree -- Specified Charges Agg Assault Offenses Sexual Assaults Burglary Habitation Intoxication Manslaughter Manslaughter Compelling Prostitution	$30,000	$40,000	$50,000

Risk Levels

* Definitions: NCA = new criminal activity; FTA = failure to appear; NVCA = new violent criminal activity

- High Risk -- An NVCA flag or either the FTA or NCA risk score is 6
- Above Average Risk -- Either the FTA or NCA risk score is 5
- Average Risk -- Either the FTA or NCA risk score is 3 or 4
- Below Average Risk -- Both FTA and NCA risk scores are 1 or 2

Effective 07/29/2017; 08/22/2017

Presumption of PR Bond

Credit/Debit Card Abuse 32.31(d)(SJF)
Criminal Nonsupport 25.05
Evading 38.04(b)(1)(a)
False Alarm/Report 42.06(b)(SJF)
Interference w/ Emerg Call 42.062(c) (w/ prior)
Prostitution 43.02(c)(2) and (c-1)(2)
Possession/Delivery/Manufacture (SJF)
Delivery of Marijuana less than 5 lb (SJF)
Possession of Marijuana less than 5 lb (SJF)
Possession of Marijuana 5-50 lbs (3rd degree)
Criminal Mischief (28.03)(b)(4)
False Stmt to Obtain Property/Credit 32.32(c)(4)
Fraudulent Transfer of Motor Vehicle 32.34(f)(1)
Unauth. Use Motor Vehicle 31.07(b)
Graffiti 28.08(b)(4)

Insurance Fraud 35.02(c)(4)
Money Laundering 34.02(e)(1)
Theft of Service 31.04(e)(4)
Theft 31.03 (e)(4)
Trademark Counterfeiting 32.23(e)(4)
Bigamy 25.01(e) (3rd degree)
Hindering Sec'd Creditors 32.33(d)(4) and (e)(4)
Interference with Railroad Property 28.07(e)(3)
Misapplication of Fiduciary Property 32.45(c)(4)
Unauthorized Use of Telecomm Services 33A.02(b)(3)
Forgery 32.21 (d)(SJF)
Forgery 32.21(e) (3rd degree)
Tampering with Evidence 37.09(c)(3rd degree)
Securing Execution of Document by Deception 32.46(b)(4)
Illegal Recruitment of Athlete 32.441(e)(4)

**** Code of Criminal Procedure art. 17.03**

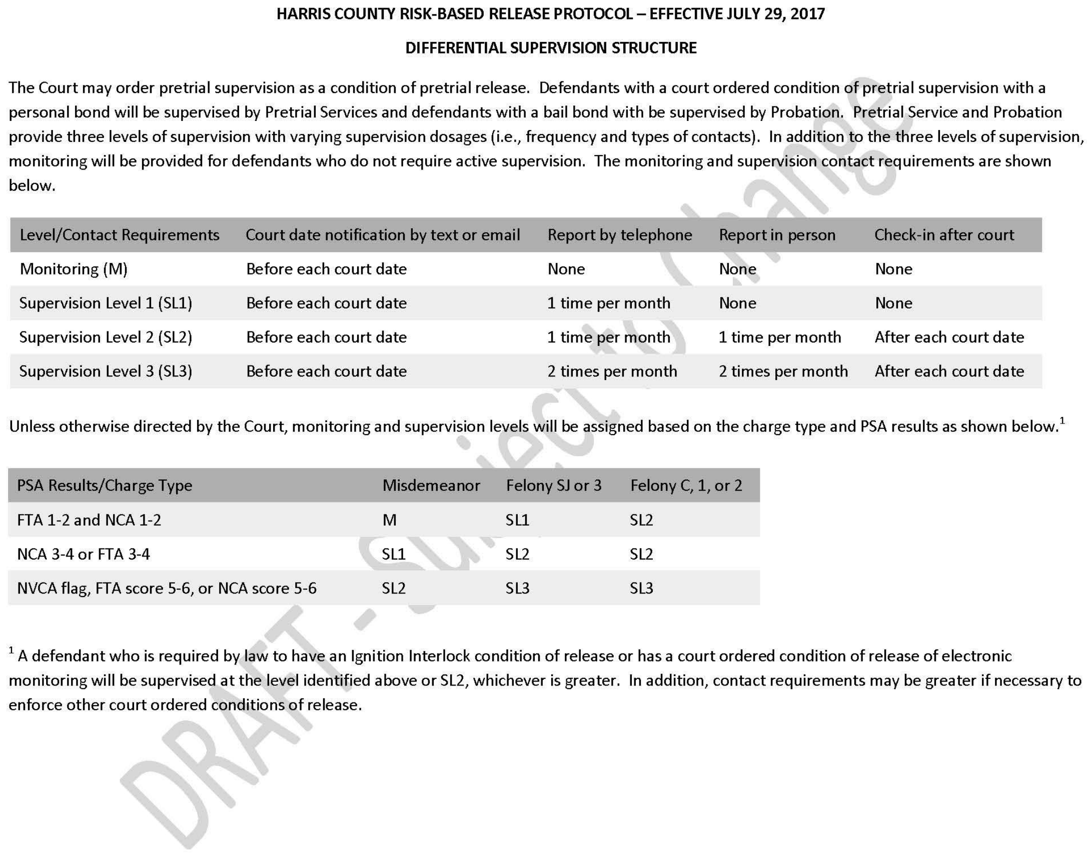
(b) Only the court before whom the case is pending may release on personal bond a defendant who

(1) is charged with an offense under the following sections of the Penal Code
(A) Section 19.03 (Capital Murder);
(B) Section 20.04 (Aggravated Kidnapping);
(C) Section 22.021 (Aggravated Sexual Assault);
(D) Section 22.03 (Deadly Assault on Law Enforcement or Corrections Officer, Member or Employee of Board of Pardons and Paroles, or Court Participant);
(E) Section 22.04 (Injury to a Child, Elderly Individual, or Disabled Individual);
(F) Section 29.03 (Aggravated Robbery);
(G) Section 30.02 (Burglary);
(H) Section 71.02 (Engaging in Organized Criminal Activity);
(I) Section 21.02 (Continuous Sexual Abuse of Young Child or Children); or
(J) Section 20A.03 (Continuous Trafficking of Persons);

(2) is charged with a felony under Chapter 481, Health and Safety Code, or Section 485.033, Health and Safety Code, punishable by imprisonment for a minimum term or by a maximum fine that is more than a minimum term or maximum fine for a first degree felony; or

(3) does not submit to testing for the presence of a controlled substance in the defendant's body as requested by the court or magistrate under Subsection (c) of this article or submits to testing and the test shows evidence of the presence of a controlled substance in the defendant's body.

Appendix B: Risk Based Release Protocol



HARRIS COUNTY RISK-BASED RELEASE PROTOCOL – EFFECTIVE JULY 29, 2017

DIFFERENTIAL SUPERVISION STRUCTURE

The Court may order pretrial supervision as a condition of pretrial release. Defendants with a court ordered condition of pretrial supervision with a personal bond will be supervised by Pretrial Services and defendants with a bail bond with be supervised by Probation. Pretrial Service and Probation provide three levels of supervision with varying supervision dosages (i.e., frequency and types of contacts). In addition to the three levels of supervision, monitoring will be provided for defendants who do not require active supervision. The monitoring and supervision contact requirements are shown below.

Level/Contact Requirements	Court date notification by text or email	Report by telephone	Report in person	Check-in after court
Monitoring (M)	Before each court date	None	None	None
Supervision Level 1 (SL1)	Before each court date	1 time per month	None	None
Supervision Level 2 (SL2)	Before each court date	1 time per month	1 time per month	After each court date
Supervision Level 3 (SL3)	Before each court date	2 times per month	2 times per month	After each court date

Unless otherwise directed by the Court, monitoring and supervision levels will be assigned based on the charge type and PSA results as shown below.[1]

PSA Results/Charge Type	Misdemeanor	Felony SJ or 3	Felony C, 1, or 2
FTA 1-2 and NCA 1-2	M	SL1	SL2
NCA 3-4 or FTA 3-4	SL1	SL2	SL2
NVCA flag, FTA score 5-6, or NCA score 5-6	SL2	SL3	SL3

[1] A defendant who is required by law to have an Ignition Interlock condition of release or has a court ordered condition of release of electronic monitoring will be supervised at the level identified above or SL2, whichever is greater. In addition, contact requirements may be greater if necessary to enforce other court ordered conditions of release.

Appendix C: Additional Figures

Figure 34: FTA Failure Rates by Risk Score (Any Case FTA)

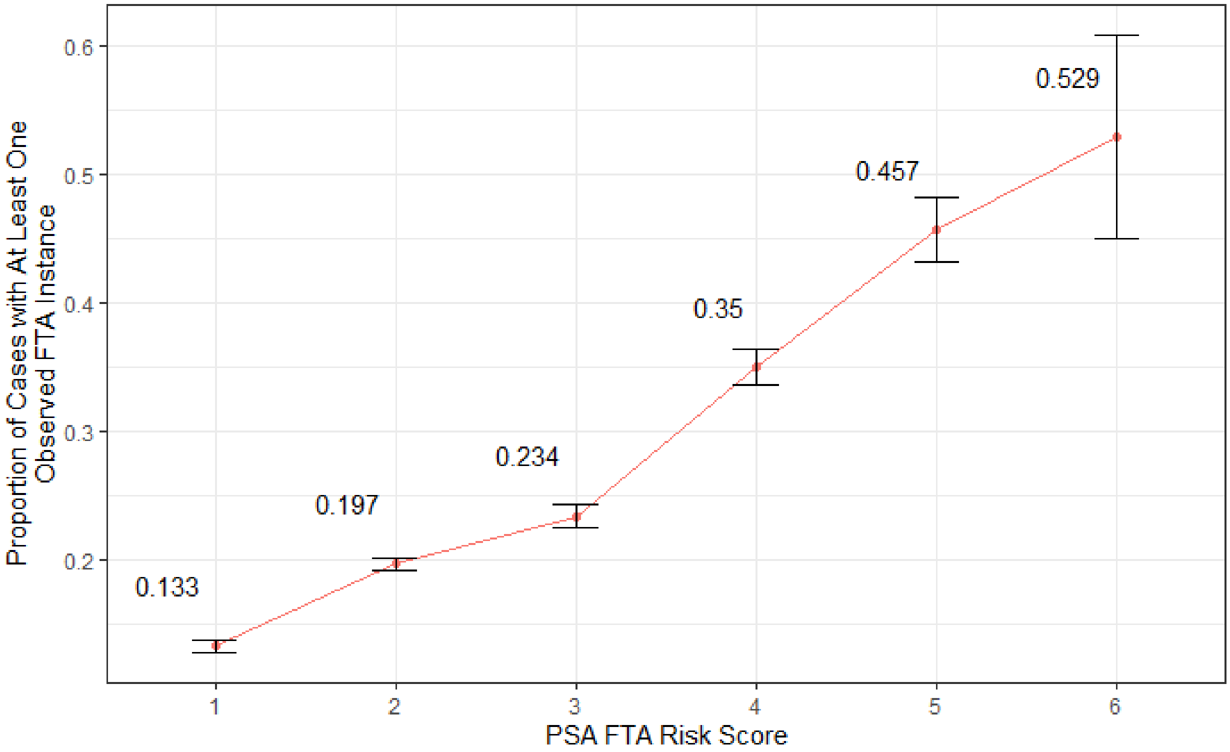


Figure 34 shows the relevant failure rates and associated 95% confidence intervals for FTA by risk score category for the Any Case FTA outcome construction. Under this construction, each increase in the PSA FTA risk score is associated with a statistically significant increase in failure rates, with the exception of an increase from an FTA risk score of 5 to 6. As noted above, this is likely due to the vanishingly small fraction of cases receiving a risk classification of 6. The findings of this figure are ultimately in line with the Base FTA construction in Figure 6; this figure provides evidence supporting the overall validity of the PSA for FTAs at risk scores of 5 or below but raises some questions as to uniform validity.

Figure 35: FTA Input Factor Correlations with Observed FTA Events (Any Case FTA)

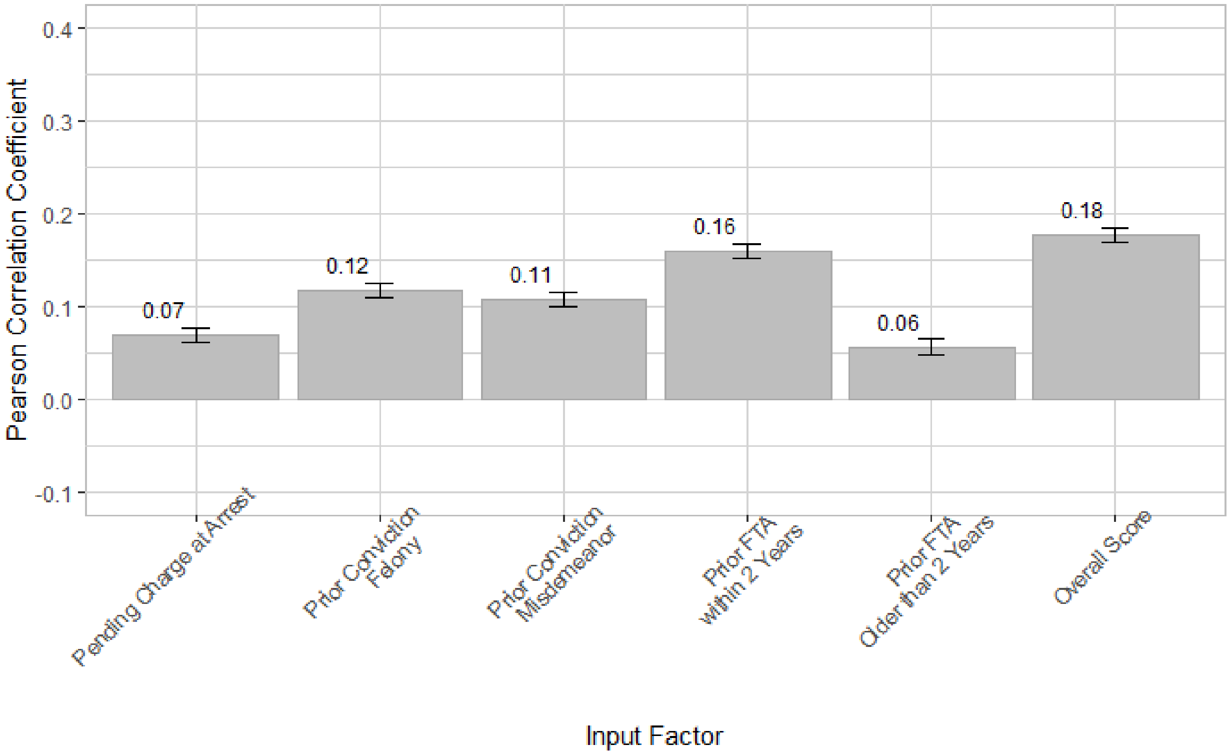


Figure 35 shows the Pearson Correlation Coefficient and associated 95% confidence intervals for each of the five factors used in the calculation of the PSA FTA score, as well as the correlation of the overall score, with observed FTA events under the Any Case FTA construction. Under this construction each input factor and the overall risk score is significantly correlated with observed FTA events in the appropriate direction, which is the same conclusion as Figure 9. Overall, this figure provides evidence for the overall validity of the PSA with respect to FTA outcomes.

Appendix D: Note on FTA Identification Issues and Calculations

As discussed in Section I.E.2 there exists complications in calculating accurate FTA outcome counts due to the lack of a dedicated FTA warrant field in Harris County's warrant data. Successful identification of prior instances of FTA events, where bench warrants are issued in response to a failed court appearance, are an important input in PSA score calculations, playing a role in calculating both FTA and NCA risk scores. Harris County PSA assessors lack clearly defined and centrally located data fields to identify past instances of FTAs by individuals when assessing new PSAs. Harris County PSA assessors have access to a database of warrants which would contain all potential warrants related to FTAs, but no dedicated field exists in this database that can identify which warrants are related to FTAs and which are not. Instead, assessors would need to investigate each warrant item across different databases. While a bond forfeiture warrant would always represent an FTA, other warrant types only might indicate an FTA event. The difficulty around correctly identifying FTA instances results in informal norms and practices among the assessors to determine which warrant instances should be pursued with investigation given the time constraints each individual assessor works under. The most immediate result of the difficulty around FTA identification is that assessors err on the side of not judging a warrant as an FTA related warrant when assessing potential FTA events under conditions of uncertainty. This likely results in an undercounting of potential FTA input events, which would explain the very low count of cases with FTA risk scores of 6.

The issues surrounding the identification of FTAs also directly impact the A2J Labs efforts to construct accurate FTA outcome counts. The A2J Lab used data obtained from the Harris County District Clerk as our starting point for constructed FTA outcome metrics. The data provided by the clerk's office contained a list of warrants related to bond forfeiture. This data had a unique case number identifier which was used to link these events to the relevant court case and PSA assessment (see Section I.E.1). An additional date field was used to determine whether these events occurred during the valid pretrial window. Thus, to construct our Base FTA measurement, where an FTA must have occurred in the PSA originating case and resulted in the issuance of a bench warrant, the following process was used: for an initializing PSA assessment, the attached defendant ID was used to gather all bond forfeiture warrants attached to the same defendant ID; all cases occurring outside the relevant pretrial period (assessment date to case disposition date) were filtered out; all bond forfeiture warrants with different case numbers were filtered out; finally, the number of unique dates in the remaining bond forfeiture warrants were calculated as the number of Base FTAs. The same process, with relevant changes in the filtering conditions, was used to calculate the alternative FTA construction of Any Case FTA. The main issue is that the warrant types are limited only to a specific type of warrant: the bond forfeiture warrant. Any FTA that resulted in a warrant of a different type is not present in our current analysis. Furthermore, judges and practitioners interested in missed court events that did not result in the issuance of a warrant cannot infer validity as to this understanding of FTA (although different from Arnold Ventures definition) from these data.

The A2J Lab received additional FTA information beyond the bond forfeiture warrant data supplied in the initial distribution of data from Harris County. This additional data contained entries for potential FTA warrants beyond (but also including) bond forfeiture warrants. The data describing these warrants included the following fields:

- Case Number- Indicated the court case the warrant was issued for
- Warrant Type- Abbreviated form of the type of warrant issued
- Warrant Activity Reason- Longer form of the type of warrant issued
- Document Filing Date- Date the warrant was filed
- Date Warrant Executed- Date the warrant was executed
- Date Warrant Returned- Date the warrant was returned

The Case Number field was the only field provided that could be used to link these warrant entries to other information obtained from Harris County. The Warrant Type field was used to identify different warrant types to apply different logics for identifying FTAs. The various date fields were used to construct the relevant logics for identifying FTAs for each relevant warrant type. The following logics were used to identify instances in the warrant data that indicated an FTA:

1. Any Alias Capias Issued warrant for a Bond Forfeiture is an FTA instance
    a. Indicated when the warrant type field was ACF
2. Any C87AI warrant issued and returned executed on the same day
    a. Indicated when warrant type field was 87A and all three date fields were equivalent
3. Any other Alias Capias Issued warrant with an open return date or executed on a later date
    a. Indicated when warrant type field was ACS or ACC and when the return date was a null value or the executed date was later than the filed date

These additional logics were used to construct the FTA+ measurement. The Base FTA measurement is represented by logic one, which captures only Bond Forfeiture warrants. Bond Forfeiture warrants always indicate an FTA. Logics 2 and 3 capture additional FTA data for warrants that sometimes indicate an FTA. Given these rules, we can understand that the FTA+ metric is a better indicator of the true rate of FTA observed within the study population, while the Base FTA rate is a better representation of the FTA instances represented by the FTA risk score calculated off the inputs of the pre-trial assessors, due to the difficulty of implementing the additional logics required to capture FTA instances beyond Bond Forfeiture warrants.