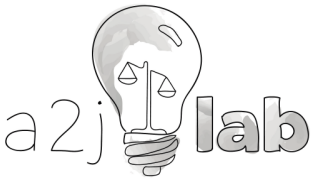
Access to Justice Lab | Center on the Legal Profession
Harvard Law School



Validation of the PSA in Kane County, IL

Respectfully Submitted

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

February 11, 2021

Executive Summary

On January 11, 2016, Kane County, IL integrated the Public Safety Assessment ("PSA") and accompanying Decision Making Framework ("DMF") into its pretrial processes. The PSA is a pretrial risk assessment instrument, a scoring system that uses criminal history 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 does not produce a recommendation, and the DMF's recommendation is not binding. Arnold Ventures, formerly the Laura and John Arnold Foundation, supported researchers to produce the PSA-DMF System in a project that concluded in 2013.

The Access to Justice ("A2J") Lab was asked to conduct a validation study of the PSA in Kane 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 validation studies of the PSA's risk categories, and this report contributes to this body of knowledge.

The A2J Lab analyzed data on Kane's use of the PSA from the Kane County Court Services, Kane County Circuit Clerk, and Kane County Sheriff's Department. The data addressed PSAs calculated between February 17th, 2016 to October 1st, 2019.

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

- There was evidence for the overall validity of the PSA scales in Kane. N(V)CA and FTA measures show increases in failure rates as scores increase, with the exception of NCA score transitions from 4-5 and 5-6. This provides evidence of moderate overall validity of the PSA, for NVCA and FTA, and weak overall validity, for NCA.
- The increases in the NCA and FTA scales were sometimes of markedly different magnitudes across score transitions, providing evidence against the uniform validity of the PSA.
- In general, there was no evidence suggesting equitable invalidity of the PSA. For both race and gender comparisons, various validation metrics showed statistically significant but directionally contradictory differences in classification gains. Some metrics showed greater classification power for Black arrestees, others showed greater classification

power for White arrestees, with the same being true for gender comparisons. There was also some, although less frequent, contradiction within a particular validation metric. Under these circumstances, we conclude that we have no evidence of equitable invalidity, although we also cannot confirm equitable validity.

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 it conducted on the use of the Public Safety Assessment ("PSA") in Kane County, Illinois. This report analyzes data with respect to PSA calculations made in Kane for felony and misdemeanor arrests from February 17th, 2016 to October 1st 2019, as well as corresponding rates of failure to appear ("FTA"), new criminal activity ("NCA"), and new violent criminal activity ("NVCA") among those released for the same time period. In brief, validation of risk assessment instruments consists of comparing the classifications individuals (as of particular arrest events) receive from an instrument's risk scores to the subsequent incidence rates of the failure events corresponding to the risk scores. Here, the A2J Lab deployed several statistical techniques to compare the scores Kane County assigned to individuals on the PSA's FTA, NCA, or NVCA 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.

The A2J Lab is appreciative of the Kane County Circuit Clerk, Kane County Court Services, Kane County Sheriff's Department, and AV, all of whose support made this report possible.

This report proceeds in two parts. Part I addresses Kane 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.

I. Kane 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 Kane County, including a brief discussion of the status of the criminal justice system from January of 2016 to the present. Section B briefly describes the PSA. Section C discusses the implementation of the PSA in Kane County. Section D discusses the nature of validation of risk assessment instruments as applied to Kane 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.

a. Kane County

Kane County is located in the north east of the state and has a population of roughly 500,000 people.[1] The county is racially diverse with 56% white, 32% hispanic, and 6% black.[2]

[1] U.S. Census Bureau (2020). *QuickFacts McLean County, Illinois.* Retrieved from https://www.census.gov/quickfacts/kanecountyillinois
[2] Id.

b. The PSA

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

The PSA is a PRA that judges 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.[3]

AV, formerly the Laura and John Arnold Foundation, supported researchers to produce the PSA in a project that concluded in 2013.[4] AV and the developing researchers sought to construct a PRAI that (i) did not require inputs from an expensive and potentially legally fraught 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,[5] and this report contributes to that literature.

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 judge. 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 Kane County DMF.

[3] 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://www.psapretrial.org/about/factors (last visited Feb. 19, 2020).

[4] 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.

[5] 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.[6]

c. PSA in Kane County

This section discusses pretrial processes and PSA in Kane 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 Kane County on January 11th, 2016. At implementation the PSA-DMF reports were only generated for cases with a felony charge. On February 1st, 2016, the program was expanded to include the creation of PSA-DMF reports for both felony and misdemeanor cases.

After an individual was arrested they would be brought to one of several locations depending on where in the county they were arrested. If the individual was arrested in the south, east, or west of the county their bond call appearance would be heard at the Kane County Judicial Center.[7] If the individual was arrested in the north of the county their bond call appearance would be heard at the Elgin Branch Count (EBC).[8] If the individual was suspected of committing a felony, the officer would consult the on-call Assistant State's Attorney at the time of arrest to determine the exact charges.[9] A PSA-DMF report would then be calculated by a Kane County Pretrial Assessor.[10]

Individuals who were charged with select misdemeanor offenses were permitted to bond out early.[11] Individuals charged with non eligible misdemeanors offenses or any felony charge were required to go to a bond hearing.[12] The bond hearings took place at the Kane County Judicial Center and the EBC.[13] Individuals arrested in the south, east, or west of the County would attend the hearing in person while defendants arrested in the northern part of the County appeared virtually from the local police department.[14]

Individuals who were ordered released on their own recognizance were immediately released without being booked into the Kane County Jail.[15] This was also true for individuals who were able to immediately pay the required bond amount imposed by the Judge.[16] This had an impact on the data and is discussed below.

---

[6] See https://www.psapretrial.org/about#jurisdictions-united-states (last visited Feb. 19, 2020).
[7] Matthew Stubenberg, Memo, "Kane A2J Lab Meeting," Memorializing Conversation on September 3, 2020 (on file with the Access to Justice Lab).
[8] Id.
[9] Id.
[10] Id.
[11] Id.
[12] Id.
[13] Id.
[14] Id.
[15] Id.
[16] Id.

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 validation techniques they used, although they have noted challenges with the data available in each jurisdiction.[17] 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 Kane County.[18]

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.

[17] 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).

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

Despite this fact, the validation study we report here, like all previous PRAI validation studies of which we are aware, analyzed 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 Kane 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.

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:

- Court data provided by the Kane County Circuit Clerk,
- Jail data provided by the Kane County Sheriff's Department, and
- PSA data provided by Kane County Court Services,

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. The A2J Lab was able to combine the three data sources using a combination of common identifiers, as follows.

Court Services provided data on 14,586 PSA instances assessed between February 17th 2016 and October 1st, 2019. Each of these PSA events was assigned a unique PSA ID that was associated with all court cases originating from the same event which generated the PSA. This PSA ID was present in the court data which listed all associated court cases, including disposition status and dates. Each PSA instance was joined with the court data for the relevant cases on the basis of this PSA ID. In the event multiple cases were associated with the same PSA ID, the beginning of the pretrial period was the day of the first initial hearing and the end of the pretrial period was the disposition date of the last court case. After matching on this criteria, 341 PSA events had no corresponding court cases. We understand that this is due to a

combination of dismissals by the DA, test PSAs, or record expungements.[19] Either way, there were no further records to analyze in these cases, and we dropped from the analysis dataset.

The remaining 14,245 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 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 13,094 entries. Each entry was a single PSA assessment attached to at least one specific case that featured at least one day of pretrial release.

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

1) Court Data: The court data provided by the Kane County Circuit 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.
3) Pretrial Data: The data provided by Court 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 resulting in the issuance of a bench warrant occurred during a relevant date range in a case associated with the initial PSA, it was counted as an observed FTA failure. If a subsequent PSA entry was created during the relevant date range, it was counted as an NCA instance.[20] An additional PSA input field for current violent offence indicated

[19] Matthew Stubenberg, Memo, "Kane A2J Lab Meeting," Memorializing Conversation on December 19, 2019 (on file with the Access to Justice Lab).
[20] 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

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 Kane County departments identified previously. Kane County officials and the A2J Lab explored the possibility of obtaining statewide arrest data from the Illinois State Police (ISP). The request to receive statewide arrest data from ISP was submitted in May 2020 but as of January 10, 2021, the data request has not been approved by the ISP.

The data was also limited because not all individuals who were arrested were booked into jail. Individuals who were arrested but were able to make bond according to the bail schedule, were released on their own recognizance, or were able to make bond immediately after the hearing were not officially booked into the jail.

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 classification 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, would indicate 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 0 or 1 scale.

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.

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. 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 the same increase in failure rates, *i.e.* whether the increase in risk when moving from a score of 1 to 2 is 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 examine what we label “equitable validation,” which concerns whether the PSA validates equally for different subgroups defined by, for example, race and gender. We pursue this analysis under the assumption that magistrates and practitioners will find such information useful.

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 alleged crime 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.

NVCA- An NVCA event is observed if a new arrest event, with an associated alleged crime that carries the potential of incarceration and is considered a violent charge, is observed during a case’s pretrial period. As part of the calculation of the PSA, pretrial services maintains its own distinct list of violent charges used specifically for the Current Violent Offense input field. We use the value of this field and not any specific list of charges.

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 a court case associated with the original PSA events PSA ID.

b. Descriptive statistics

The study population consists of 13,094 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 10,059 unique individuals charged with either misdemeanors or felonies over the period of February 17th, 2016 to October 1st, 2019. Individuals were recorded with five separate racial category identifiers; however, less than 1.5% were not categorized as either Black or White. For the purposes of readability, the original racial categories were condensed down 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, which can be viewed in Figure 1 below, showed more White arrestees. There were 4889, or over two times, more White individuals than Black individuals in the analysis dataset. In terms of gender distribution (Figure 2), 79.7%, or about 4/5ths, of the unique PSA assessments in the analysis dataset attached to male individuals. The age distribution, seen in Figure 3, 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. 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 92% of all PSA instances had the corresponding individual experience at least one day of pretrial release (the other 8% 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 statistically significantly higher rates of pretrial release than Black arrestees (0.93 vs. 0.89), while female arrestees had statistically significantly higher rates of pretrial release than male arrestees (0.94 vs. 0.91). Differences in failure rates are analyzed in further detail later in the report, but overall group differences were significant, with male arrestees observing higher overall NCA and NVCA rates than female arrestees (the corresponding FTA difference is statistically insignificant ($p=0.421$)), while White arrestees observed lower failure rates across all outcomes than Black arrestees. Overall, 16.6% and 22.3% of the study population observed either an NCA or FTA event, respectively. Only 7.1% of PSA instances observed an NVCA failure during the relevant pretrial release period.

Figure 1: Distribution of Individual Race

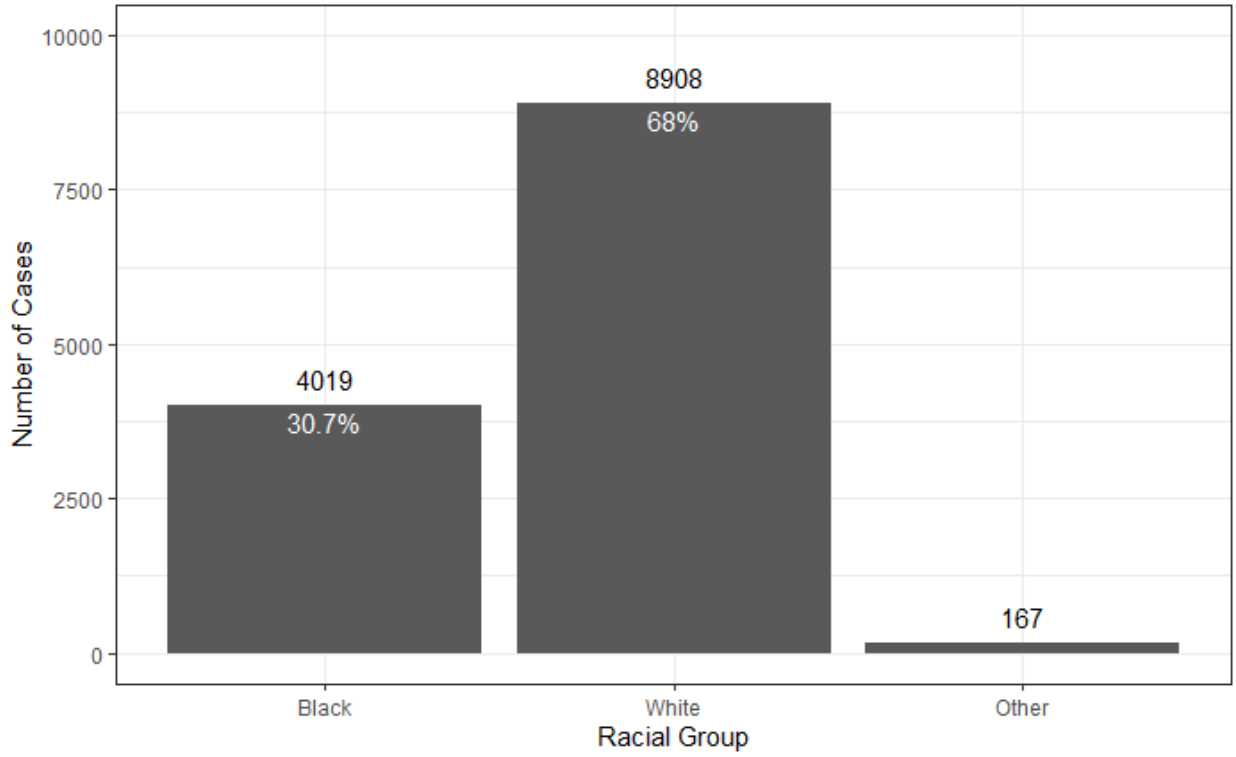


Figure 1 displays the distribution of racial categories for individuals. Each portion of the chart indicates the percentage of unique PSA submissions that listed the relevant Race category for the individual. The initial data obtained from Kane County contained five separate racial categories; however, two of the categories, White and Black, represented 98.7% of all cases. White individuals were the modal category, representing the majority of individuals who received a PSA assessment at about 68% of the study population.

Figure 2: Distribution of Individual Gender

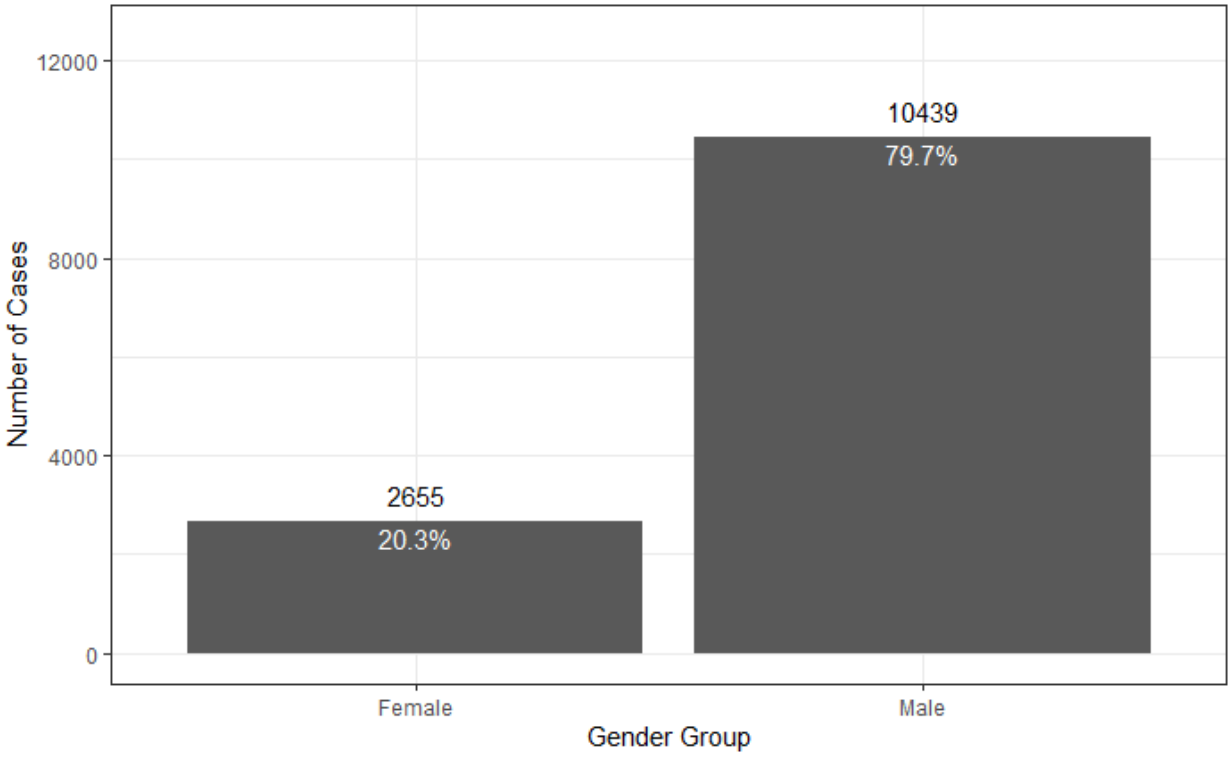


Figure 2 displays the distribution of gender categories within the study population. Female individuals represent just over a fifth of the total study population at 20.3%, which makes male individuals the overwhelming majority of individuals with a PSA assessment. Out of the total study population of 13,094, there were 2655 PSA assessments with a female individual and 10,439 PSA assessments with a male individual.

Figure 3: Distribution of Individual Age

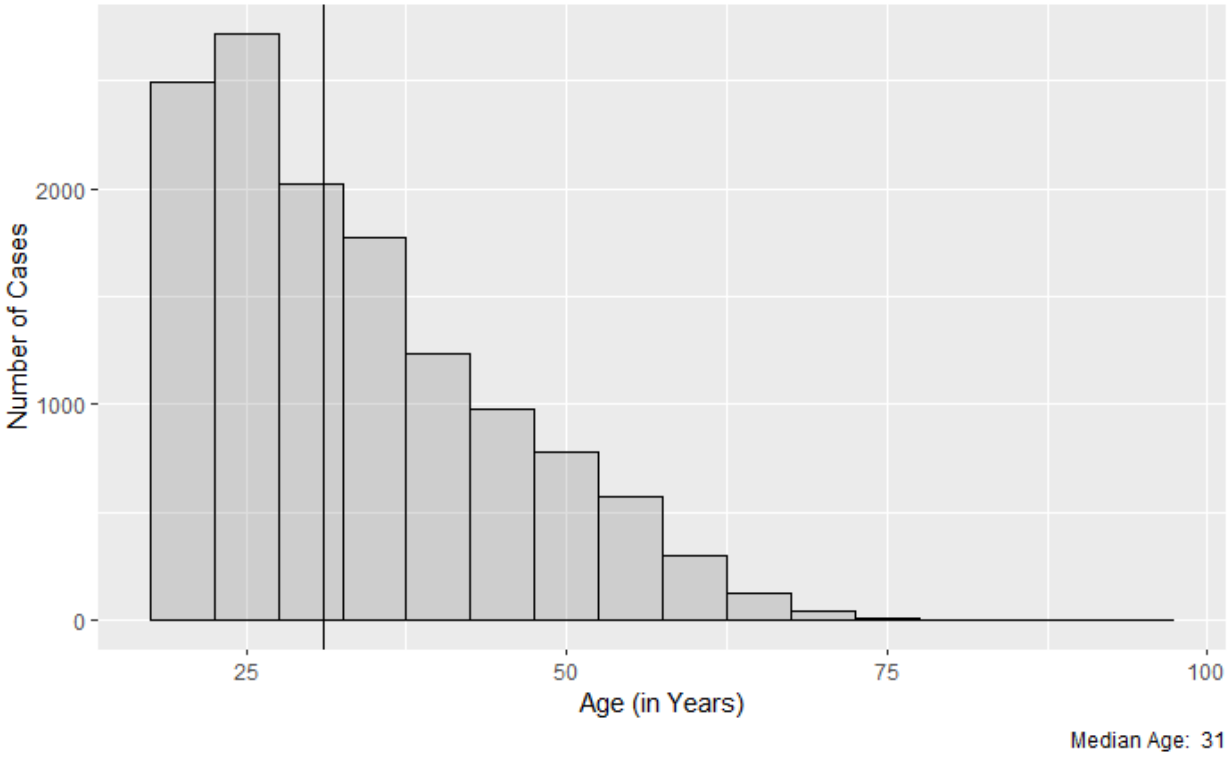


Figure 3 plots the distribution of individual age within the study population. This indicates a higher fraction of younger individuals. The mean age is slightly above 33 years old, with a median of 31 years old. This means that half of the study population exists with a 13 year age range: from 18-31, while the rest occupy a 62 year age range, from 32 to 94.

Table 1: Summary of Failure Rates by Demographic Group

Group	# of PSAs	Released (N)	NCA Fail Rate	NVCA Fail Rate	FTA Fail Rate
Overall	14245	13094	0.166	0.071	0.223
Black	4680	4186	0.191	0.083	0.242
White	9565	8908	0.153	0.065	0.21
Female	2818	2655	0.143	0.061	0.227
Male	11439	10439	0.172	0.074	0.222

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 4% higher than Black arrestees (93% vs. 89%), while female arrestees had at least 1 day of pretrial release at a rate about 3% higher than male arrestees (94% vs. 91%). Likewise, all reported failure rates are significantly different across paired demographic groups with the exception of gender based FTA failure rates, which are not statistically significant. Male arrestees observed higher NCA and NVCA than their female peers, while White arrestees observed lower NCA, NVCA, and FTA rates than their Black arrestee 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 for the most part appears overall valid 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 evidence that the PSA is mostly overall valid, and provides some but not conclusive evidence that the PSA is not uniformly valid. Key details are as follows:

- N(V)CA and FTA measures show consistent increases in failure rates as scores increase, with the exception of NCA scores of 4-5 and 5-6. This provides moderate, for NVCA and FTA, and weak, for NCA, evidence of overall validity of the PSA.
- Some FTA failure rate increases across scores do not differ significantly. For example, the failure rate difference between FTA scores of 1 and 2 is statistically indistinguishable from the corresponding difference between FTA scores of 3 and 4. Similarly, the NCA failure rate increase at scores of 4 and 5 is statistically significantly less than other NCA failure rate increases. In general, two distinct patterns emerge for failure rate increases in both NCA and FTA scores that indicate different behavior for lower score transitions than higher score transitions. This provides evidence against the uniform validity of the PSA.

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. A primary 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.[21] We use a

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

difference 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 be uniformly valid, 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: NCA, NVCA, and FTA. They show that under this definition, the PSA is overall valid with the exception of the transition from NCA 4 to NCA 5 and NCA 5 to NCA 6, for which statistical tests show no significant difference. Both the PSA's NCA scale and FTA scale exhibit distinct, but differing, patterns in failure rate increases for lower score transitions when compared to higher score transitions, which provides evidence against the uniform validity.

- 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).
- 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: NCA Failure Rates

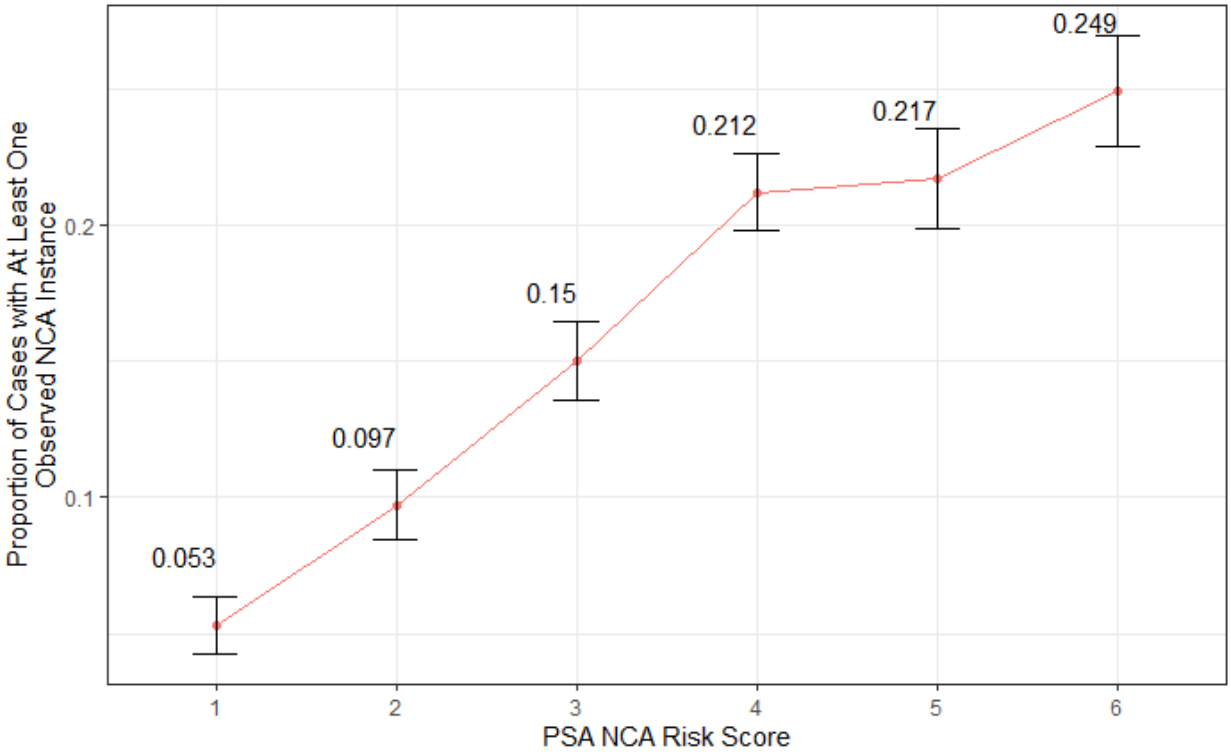


Figure 4 shows the relevant failure rates and associated 95% confidence intervals for NCA by risk score category. An overall valid risk assessment tool should show significant increases in failure rate at each subsequent level of the associated risk score. A lack of overlap between confidence intervals for consecutive paired scores indicates that each subsequent increase in the PSA NCA risk score is associated with a significant increase in failure rates. Differences between scores of 1-2, 2-3, and 3-4 are statistically significant; differences between scores of 4-5 and 5-6 are not significant. A one unit change in risk score level was observed to have an associated failure rate increase of roughly 3.1%. The jumps in failure rates associated with each score increase are fairly consistent at the 1-2, 2-3, and 3-4 transitions, but, the upper transitions indicate a distinct, and indeterminate, pattern. Overall, this figure provides weak evidence supporting overall validity but the presence of two distinctive patterns for lower NCA scores and higher NCA scores provides evidence against uniform NCA validity.

Figure 5: NVCA Failure Rates

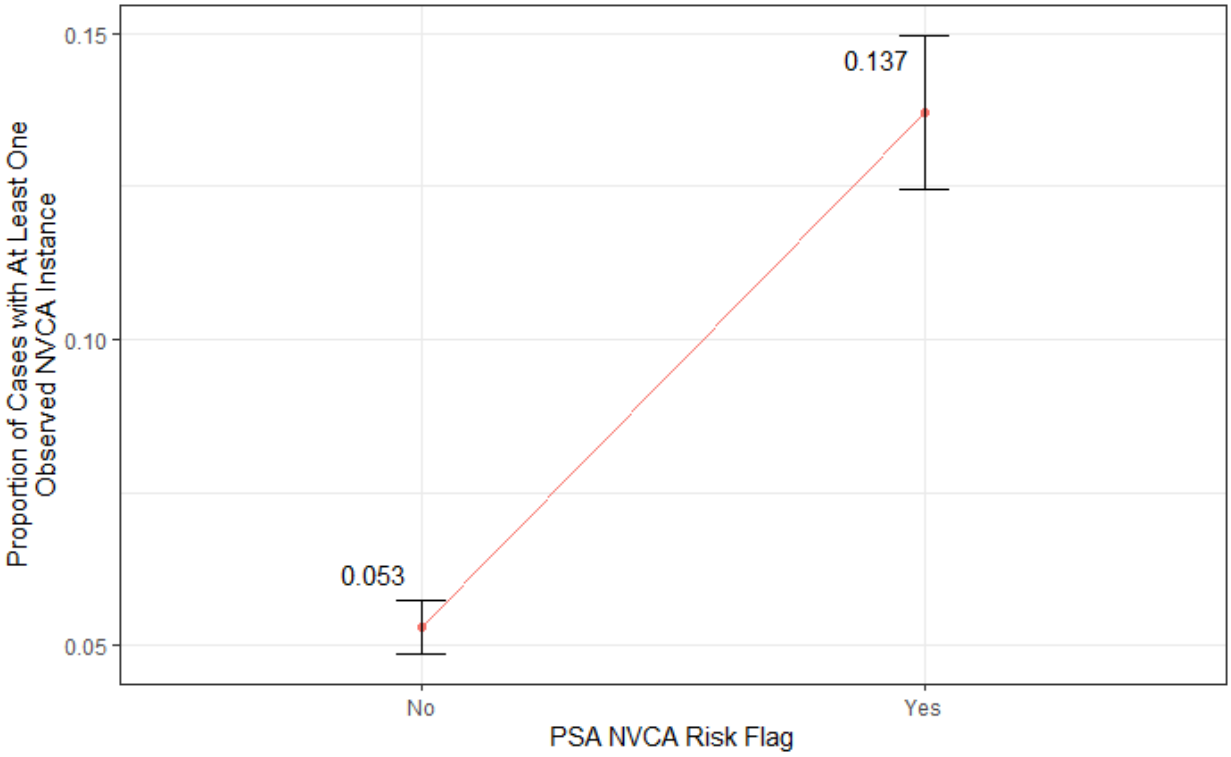


Figure 5 shows the relevant failure rates and associated 95% confidence intervals for NVCA by presence of the NVCA Risk Flag. An overall valid risk assessment tool should show significant increases in failure rate when a binary flag is present, which could be indicated by no overlap between the confidence intervals. Here, the presence of the PSA NVCA risk flag is associated with a significant increase in failure rates of 8.4 percentage points. The difference is statistically significant, and thus provides evidence supporting the validity of the PSA with respect to NVCA outcomes.

Figure 6: FTA Failure Rates

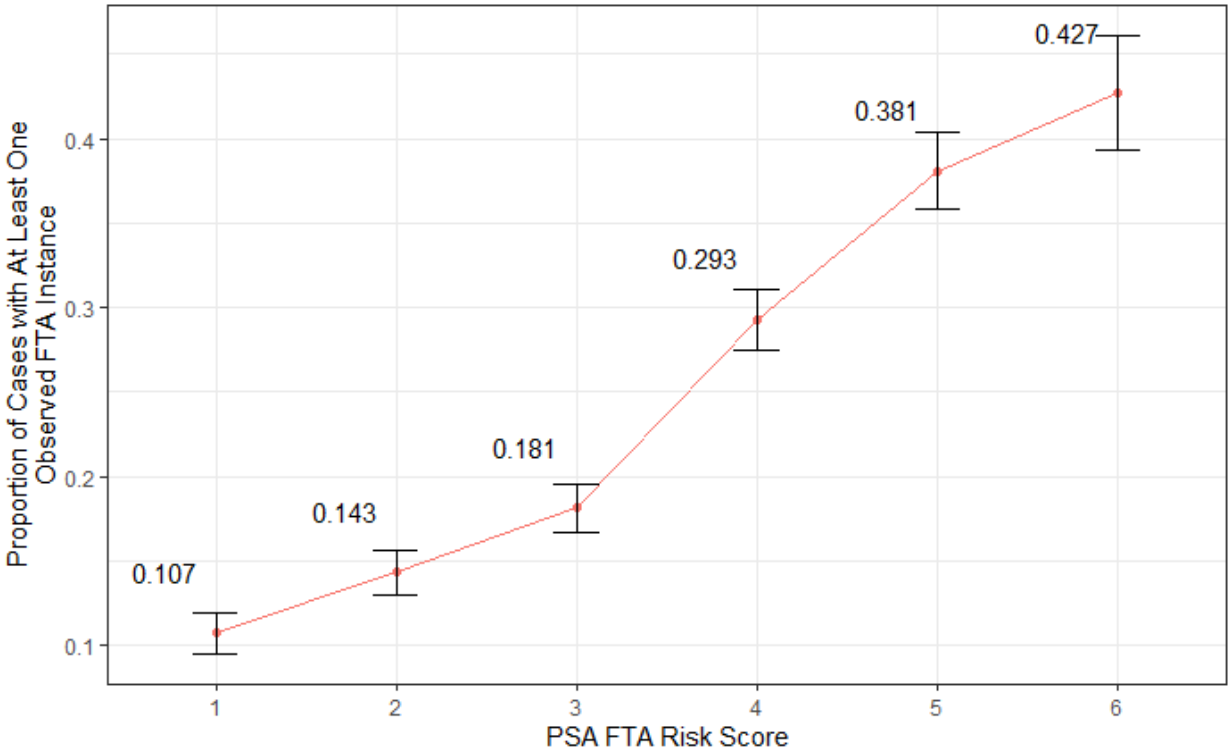


Figure 6 shows the relevant failure rates and associated 95% confidence intervals for FTA by risk score category. Here, each increase in the PSA FTA risk score is associated with a statistically significant increase in failure rates. A one unit change in risk score is associated with a variety of increases in failure rates, ranging from an increase of only 3.6% to an increase of 10.2%. This figure provides evidence supporting the overall validity of the PSA for FTA, but the presence of distinctive patterns for lower FTA score transitions and higher FTA score transitions provides evidence against uniform validity.

Figures 4-6 demonstrate 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 5.3% for cases with risk scores of 1 and maximum failure rate of 24.9% for cases with risk scores of 6. FTA failure rates achieve a similar minimum and maximum at scores of 1 and 6, with rates of 10.7% and 42.8% respectively. Cases with an NVCA flag present were observed with approximately 2.5 times the failure rate of cases without the flag present, with failure rates of 5.3% and 13.7%, 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 4-5 and 5-6 on the NCA scale.

With respect to uniform validity, rough statistical comparisons[22] suggest that each step increase in the NCA and FTA scores was associated with statistically indistinguishable increases in

[22] 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 did with the exception of the NCA 4-

failure rates except for the NCA transitions from 4 to 5 and from 5-6, which had increases that were not statistically significant. Despite the overlap in confidence intervals for score transitions, the figures above suggest two distinct patterns at work in both the NCA and FTA calculations. In both instances, scores at the lower end of the relevant scale exhibit different patterns from those at the higher end, but these patterns are reversals of each other across the two outcome types. For NCA, failure rate increases tend to be larger for lower score transitions and lower for higher score transitions, while for FTA, failure rate increases tend to be lower for lower score transitions and higher for higher score transitions. These results provide some evidence against the uniform validity of the PSA.

ii. Bivariate correlations

This subsection provides the results of bivariate comparisons, also known as correlations. This correlation analysis provides some but not strong evidence that the PSA is overall valid. In particular, the overall risk score achieves at least as large a correlation coefficient as 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 was statistically significantly correlated with the relevant outcome in the expected direction. This provides weak evidence in support of overall validity.
- These correlations were generally modest, ranging between 0.04 and 0.25.
- For NCA and NVCA, the largest correlations in magnitude for each PSA metric corresponded to the overall risk score (or flag), as opposed to any particular input, except for FTA, where the largest correlation was shared between the overall score and the presence of prior FTAs within the past two years. This result also provides evidence in support of overall validity.

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.

5 score transition. The NCA score transition for 4-5 only overlapped with the NCA score transition of 5-6, but this is due more to the large confidence interval around the 5-6 transition because of the small number of cases that observed a score of 5 or 6.

Table 2: 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*
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.[23] 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.

[23] 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 7: NCA Input Factor Correlations with Observed NCA Events

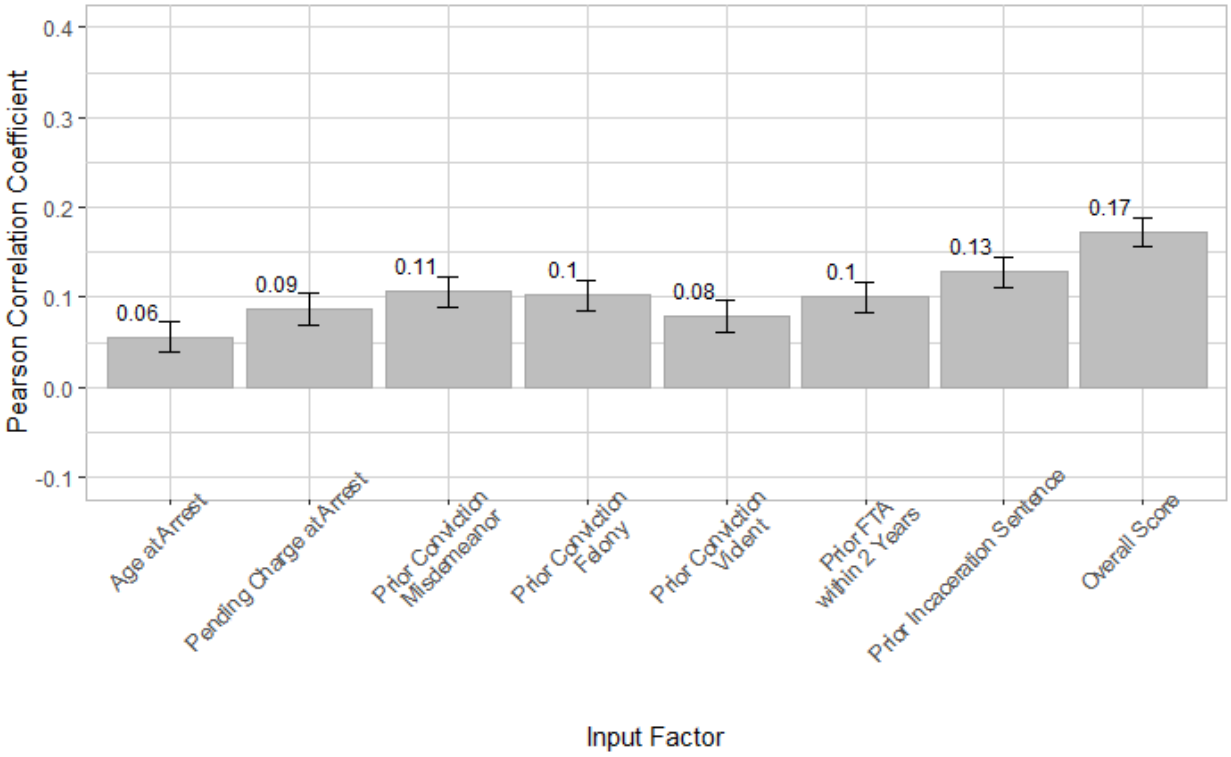


Figure 7 shows the Pearson Correlation Coefficient and associated 95% confidence interval for each of the six factors used in the calculation of the PSA NCA score, as well as the correlation of the overall score, with observed NCA events. The figure indicates that each input factor and the overall risk score is statistically 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 8: NVCA Input Factor Correlations with Observed NVCA Events

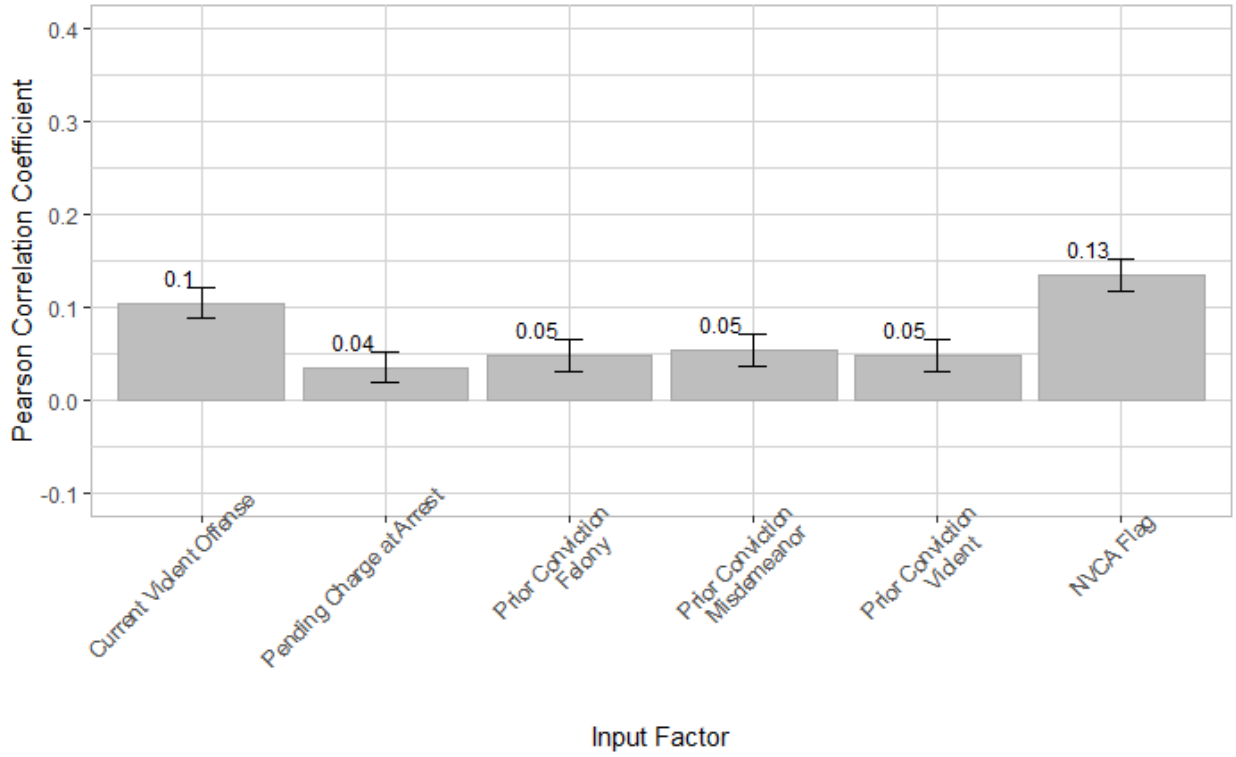


Figure 8 shows the Pearson Correlation Coefficient and associated 95% confidence interval for each of the six factors used in the calculation of the PSA NVCA risk flag, as well as the correlation of the presence of the risk flag, with observed NVCA events. The lack of overlap between the confidence intervals and 0 indicates that each input factor and the overall risk score is statistically 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 9: FTA Input Factor Correlations with Observed FTA Events

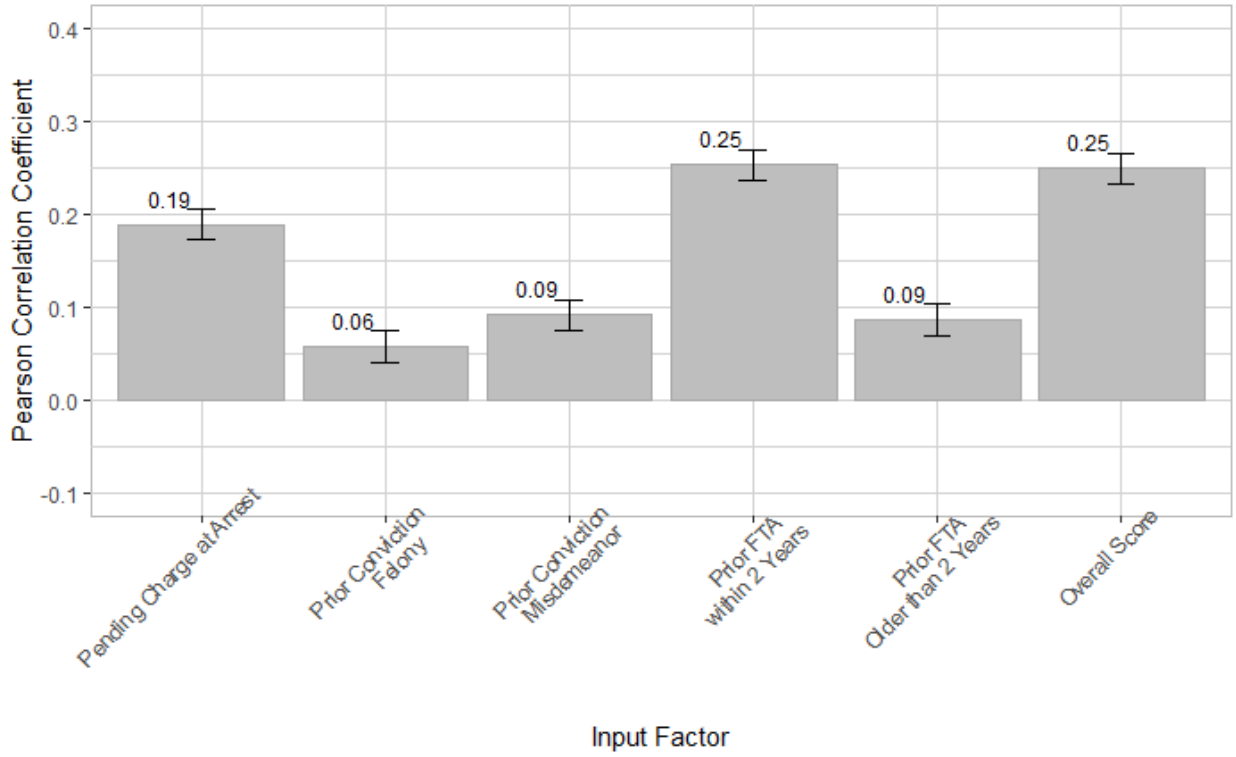


Figure 9 shows the Pearson Correlation Coefficient and associated 95% confidence intervals for each of the five factors used in the calculation of the PSA FTA score, as well as the correlation of the overall score, with observed FTA events. Each input factor and the overall risk score is statistically significantly correlated with observed FTA events in the appropriate direction. The smallest and largest factor correlation coefficients are obtained for Prior Felony Conviction and Prior FTA within 2 Years, respectively. The overall risk score achieves a larger correlation coefficient than any individual input factor coefficient except with respect to Prior FTA within 2 Years, suggesting that this predictor alone functions as well as the score calculation. This figure provides weak evidence for the overall validity of the PSA with respect to FTA outcomes.

Figures 7-9 each show positive, statistically significant correlations between the various factor inputs to the PSA scores and the relevant PSA outcome, including the Age factor, which is based on the calculated score input of the age category not on age itself. Each of these correlations is in the expected/appropriate direction. For each PSA outcome, all input correlation coefficients are significantly different from 0 at the $p < 0.001$ level. 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 available from the raw form of the inputs is lost. In general, 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 moderate to weak evidence of the overall validity of the PSA and weak evidence

against the equitable validity of the PSA. With respect to equitable validity, however, the AUC results contradict those from the Balance Accuracy Measurement ("BAM") technique discussed below, so we draw no firm inferences. Key details are as follows:

- NCA and NVCA outcomes indicate weak gain in classifying power from the risk score, while the AUC analysis of FTA indicates moderate gains in classifying power.
- There are significant differences in AUC scores across racial demographic subgroups for both NCA and FTA outcomes, with higher AUC scores for White individuals indicating stronger classifying gains for cases with individuals of that racial group relative to their Black counterparts. These differences, standing, provide weak evidence against equitable validity, but because they directionally contradict the results from the BAM technique (see below), we draw no firm inferences from them.
- Statistically significant gender differences exist for NCA outcomes, with female arrestees observing higher AUC scores than male arrestees. These differences cross the weak-moderate evaluative threshold and provide weak evidence against equitable validity. On equitable validity, however, these results are directionally inconsistent with those from the BAM technique, so we draw no firm inferences from them.

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.[24] 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 Kane 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

[24] 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.

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 AUC measurement less than 0.54 indicates no evidence of validity.[25] An AUC measurement between 0.55 and 0.63 indicates weak 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, and FTA.

Figure 10: Area Under the Curve Values by Outcome Construction

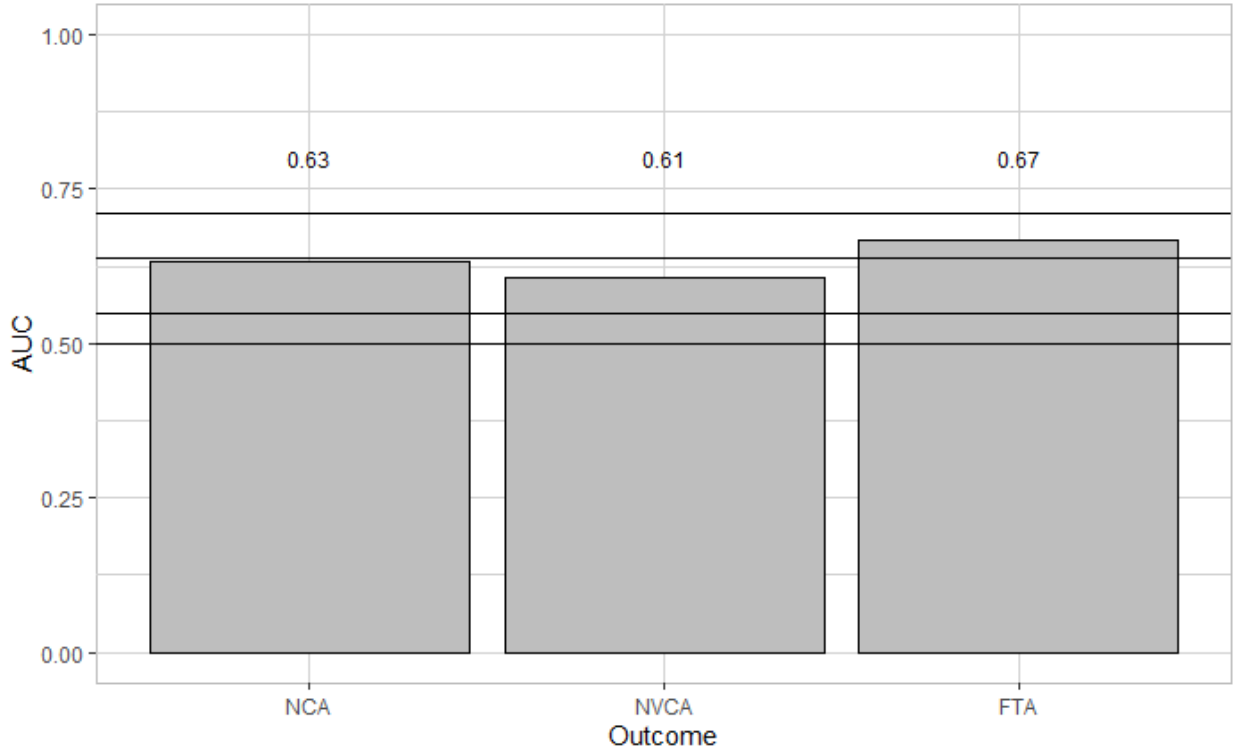


Figure 10 shows the area under the curve values for each outcome construction. AUC values range from 0 to 1, and in the case of a multi-outcome predictive assessment tool, like the PSA, are best understood as the probability that a randomly selected case with an observed failure for an outcome will have a higher corresponding risk score than a randomly selected case with no observed failure for that outcome. 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 weak evidence that the PSA works better than chance at classifying NCA and NVCA events and moderate evidence that the PSA works better than chance at classifying FTA events.

[25] 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 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 metric is 0.63, which represents weak, bordering on moderate, gain in predictive power. For NVCA, the overall AUC metric is 0.61, which represents a weak gain in predictive power. For FTA outcomes, the overall AUC metric is 0.67, indicating a moderate gain in predictive power. Thus, the AUC metric provides weak to moderate evidence of overall validity.

Figure 11: Area Under the Curve Values by Outcome Construction Across Demographic Subgroups

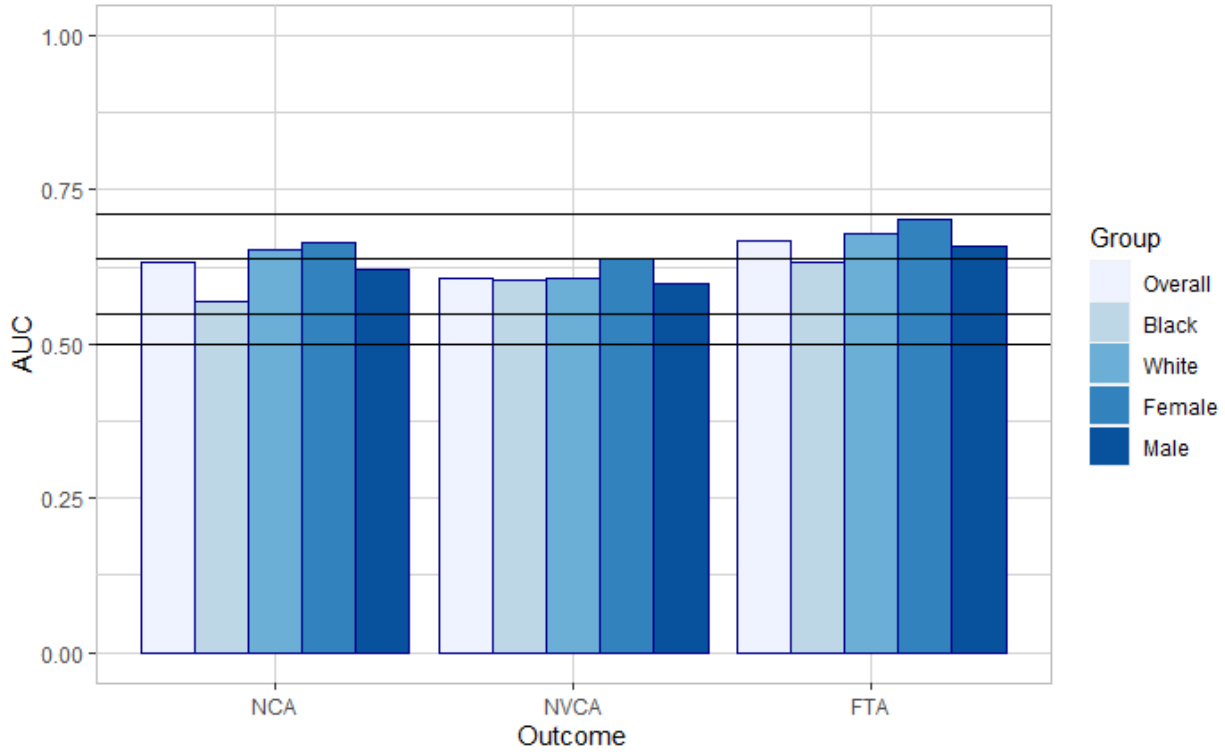


Figure 11 shows the area under the curve values for each outcome construction across the four main demographic groups of analysis as well as for the overall study population. See previous graph captions for explanations of the AUC metric and corresponding thresholds. Across demographic subgroup pairings, the figures indicate differential performance across evaluative thresholds for gender and racial pairings on NCA and for racial pairings on FTA. The AUC values suggest the PSA provides better predictive performance for White arrestees (moderate gains) than Black arrestees (weak gains) for both NCA and FTA outcomes, while for gender, female arrestees observe moderate gains for NCA while male arrestees observe weak gains.

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 evidence of a difference in the validity with respect to racial and gender groups. Black and White individuals differed

significantly for NCA and FTA outcomes constructions. The PSA showed moderate gains in predictive power for White arrestees but only weak gains in predictive power for Black arrestees. Male and female individuals differed significantly for NCA outcomes, with the PSA providing moderate gains in predictive power for female arrestees and weak gains for male arrestees. The AUC metrics differ by at most 0.04 to 0.09. This provides weak evidence against the equitable validity of the PSA.

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 mixed evidence on uniformly valid. Key details are as follows:

- PSA risk scores/flags have statistically significantly positive coefficients, indicating that increases in risk scores are statistically significantly associated with increases in the probability of observing a relevant outcome. This result supports the overall validity of the PSA.
- Moving from the minimum to the maximum risk score is associated with a 6.5x increase in the probability of observing an NCA and a 4x increase in observing an FTA, again suggesting overall validity.
- The presence of the NVCA Flag is associated with a 2.5x increase in observing an NVCA, providing evidence of overall validity.
- Including a self interaction term in the NCA and FTA models results in a significant higher order coefficient for only the NCA model, providing mixed results with respect to uniform validity.

A logistic regression framework provides an off-the-shelf[26] method for assessing the overall validity of the PSA.[27] The following figure plots predictive probabilities of observing at least one

[26] 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.

[27] 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.

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 was regressed on only the relevant risk score scale.

Figure 12: NCA Birvariate Predicted Probabilities

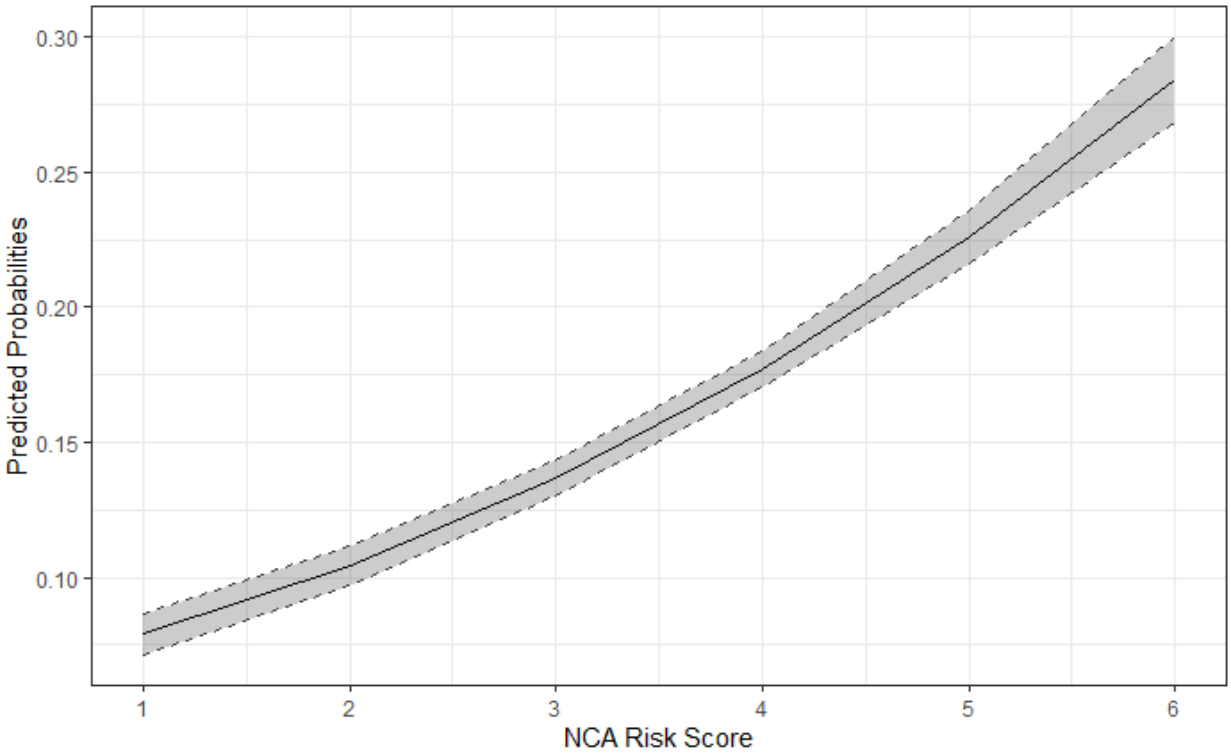


Figure 12 reports predicted probabilities and 95% confidence intervals for observing an NCA event obtained from the bivariate regression model with the relevant PSA risk score as the only regressor. The PSA NCA risk score has a significant, positive coefficient, indicating that higher NCA risk scores are statistically significantly associated with a higher probability of an observed NCA failure. A one unit increase in the NCA risk score is associated with a 36% (95% confidence interval 32% to 40%) increase in the odds ratio of observing an NCA failure versus not observing an NCA failure. Thus, this figure provides support for the overall validity of the PSA with respect to NCA outcomes.

- 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.
- 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 13: NVCA Bivariate Predicted Probabilities

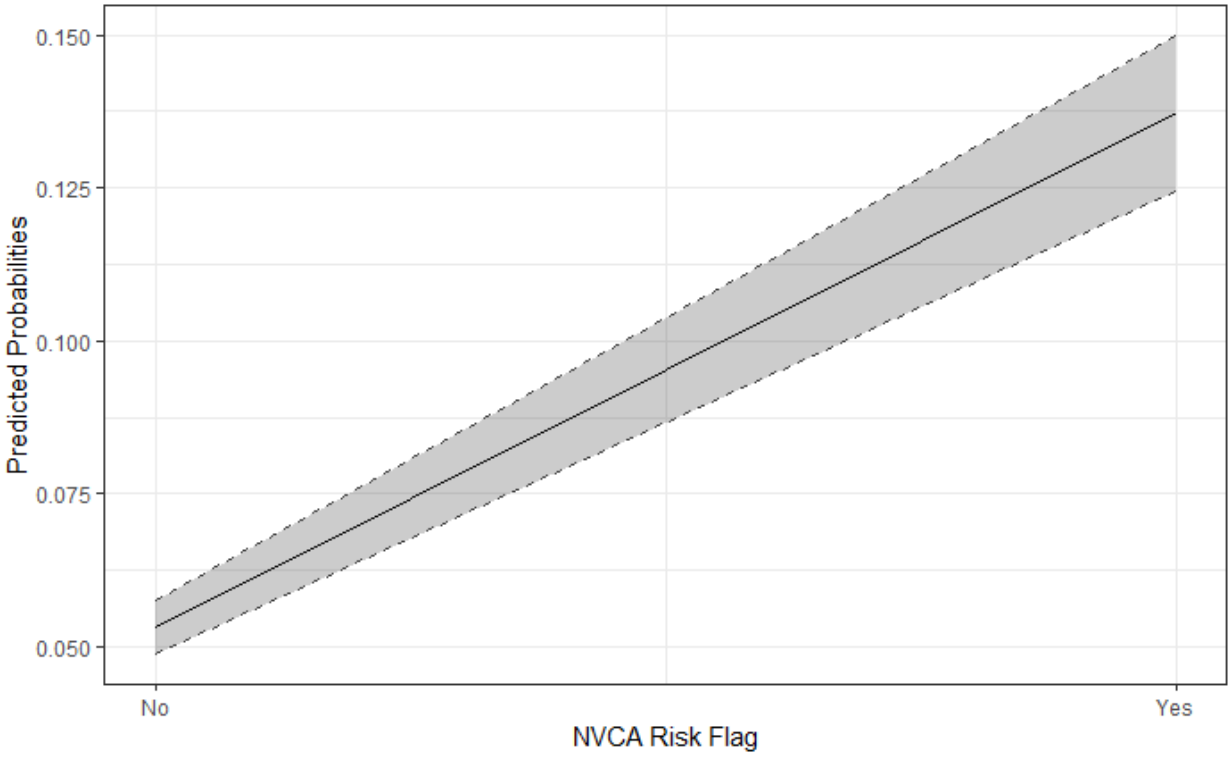


Figure 13 reports predicted probabilities and 95% confidence intervals for observing an NVCA event obtained from the bivariate regression model with the presence/absence of the NVCA flag as the only regressor. The flag has a statistically significant, positive coefficient, indicating that its presence is significantly associated with a higher probability of an observed NVCA failure. The presence of the flag is associated with a 183% (95% confidence interval of 146% to 224%) increase in the odds ratio of observing an NVCA failure versus not observing an NVCA failure. Thus, this figure provides support for the overall validity of the PSA with respect to NVCA outcomes.

Figure 14: FTA Bivariate Predicted Probabilities

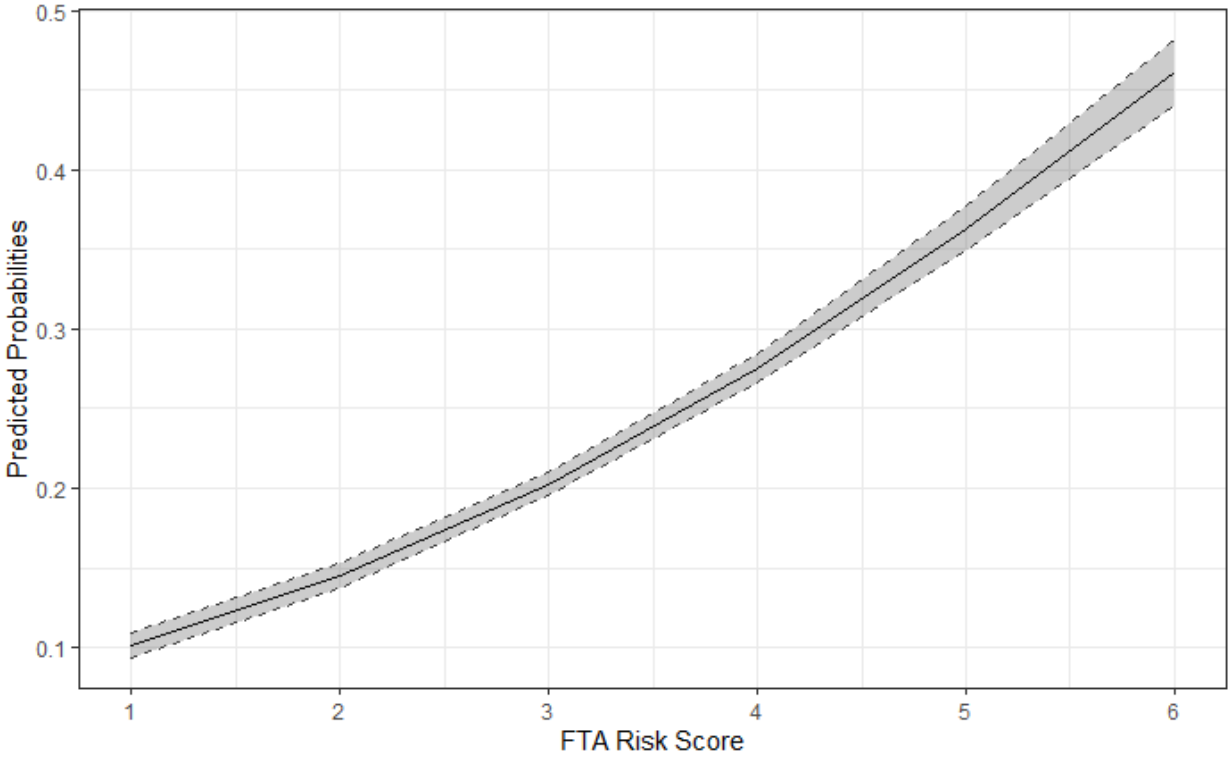


Figure 14 reports predicted probabilities and 95% confidence intervals for observing an FTA event obtained from the bivariate regression model with the FTA score as the only regressor. The FTA risk score has a statistically significant, positive coefficient, indicating that higher FTA risk scores are significantly associated with a higher probability of an observed FTA failure. A one unit increase in the FTA risk score is associated with a 50% (95% confidence interval of 46% to 54%) increase in the odds ratio of observing an FTA failure versus not observing an FTA failure for the FTA outcome construction. Thus, this figure provides support for the overall validity of the PSA with respect to FTA outcomes.

Figures 12-14 show that the predicted probabilities across each risk score level significantly increase along the risk score scale. For NCA, these ranged from a minimum of a 7.9% predicted chance of observing an NCA outcome at an NCA risk score of 1 to a maximum of 28.4% at an NCA score of 6. For the FTA model, these ranged from a minimum predicted probability of observing at least one FTA event of 10.2% at an FTA score of 1 and a maximum predicted probability of 46.1% at an FTA score of 6. For the NVCA model, having the NVCA flag present resulted in a predicted probability of 5.3% while not having the flag present was associated with a predicted probability of 13.7%. The standard error regions around these probability estimates indicated that the differences between the predicted probabilities were significant. These probabilities were 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 was significantly greater than one across all outcome models, indicating that increases in the associated PSA risk score were 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 was 1.36 on a 95% confidence interval of (1.32, 1.40), indicating that a one unit increase in NCA risk score was associated with a 36% 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 presence of the NVCA Flag was 2.83 on a 95% confidence interval of (2.46, 3.24), indicating that the presence of the NVCA Flag was associated with a 183% increase in the odds ratio of observing at least one NVCA event during the pretrial period. For the FTA bivariate model, the exponentiated coefficient estimate for the FTA risk score scale was 1.50 on a 95% confidence interval of (1.46, 1.54), indicating that a one unit increase in FTA risk score was associated with a 50% 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 indicates 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 outcome model, the higher order coefficient (estimated from a logistic regression model including the risk score and the squared risk score) was significant at the $p<0.01$ level. The exponentiated higher order coefficient was 0.93, which is less than one, indicating that higher levels of the NCA score scale were associated with smaller increases in the probability of observing an NCA event. The change in the odds ratio of observing the event, represented by the exponentiated coefficient, of -7%, was fairly minor. However, given the statistical significance of the squared term, the logistic regression analysis provides weak evidence that the PSA does not uniformly validate with regards to NCA. For the FTA model, the higher order coefficient was not statistically significant ($p = 0.85$), which provides evidence in support of the uniform validity of the PSA with respect to FTA outcomes.

ii. Balanced accuracy measures

This section reports the results of a balanced accuracy analysis. The balanced accuracy measures provide moderate evidence of overall validity of the PSA. They also provide weak evidence that the PSA is not equitably valid, but these results directionally contradict those from the AUC analysis. Therefore, we draw no firm inferences regarding the equitable validity of the PSA from these results. Key details are as follows:

- Balanced accuracy metrics across all hypothetical score thresholds showed some gain in classification power above the 0.50 threshold. These gains were largest for threshold scores of 3, which showed modest gains above 0.60 for NCA, NVCA, and FTA. These results provide moderate support for the overall validity of the PSA.
- Racial differences existed in a majority of NCA and FTA hypothetical threshold cases, which provides weak evidence against equitable validity. Gender differences existed in a

majority of NCA and minority of FTA hypothetical threshold cases, which provides weak evidence against equitable validity. Again, however, these results should be compared to those from the AUC metric, and we draw no firm conclusions from them.

Accuracy is a commonly used assessment technique in machine learning. Accuracy is based on a confusion matrix.[28] One constructs a confusion matrix by dividing each case/observation either into a positive/high 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.[29]

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).[30] 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 numbers of observed positive and negative outcome cases are 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

[28] 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.

[29] 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.

[30] 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.

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 Kane County data are not equal or uniform across outcomes. As previously discussed, across FTA, NCA, and NVCA, at least 78% of cases have no observed no failure.

For this reason, we show below not the raw Accuracy metric but instead what researchers call the "Balanced Accuracy" statistic.[3132] 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.[33] 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.

[31] 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.
[32] 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.
[33] Specifically, the metric is the sum of category correct predictions divided by total category predictions, then divided by number of outcome categories.

Figure 15: Balanced Accuracy Measurements for NCA By Hypothetical Prediction Score Thresholds

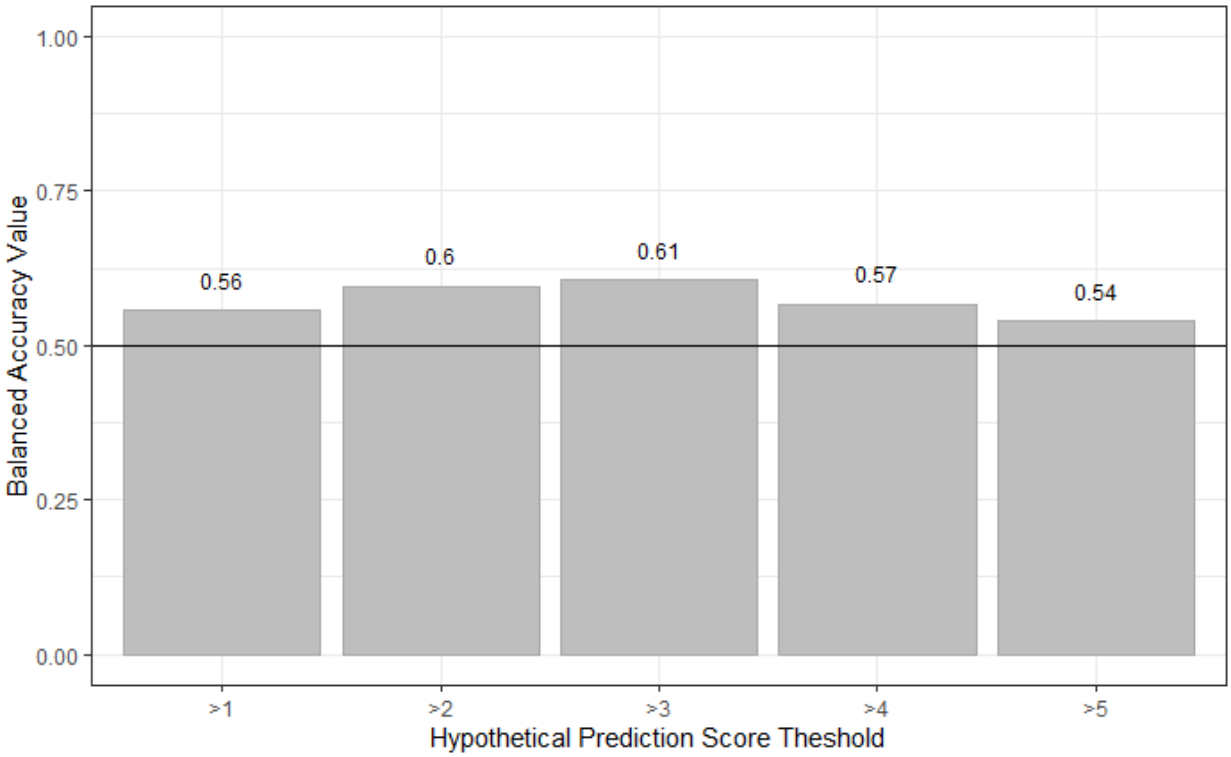


Figure 15 reports Balanced Accuracy measures for NCA outcomes using NCA score thresholds of 1, 2, 3, 4, and 5. The figure above shows that 2 of the 5 hypothetical prediction thresholds obtained balanced accuracy measures higher than 0.6, indicating that the PSA, under these hypothetical prediction rules, increases predictive power beyond classifying cases with no information beyond outcome distribution. This figure provides modest evidence supporting the overall validity of the PSA with respect to NCA outcomes.

Figure 16: Balanced Accuracy Measurements for NVCA Outcome Constructions By Possible Prediction Score Threshold

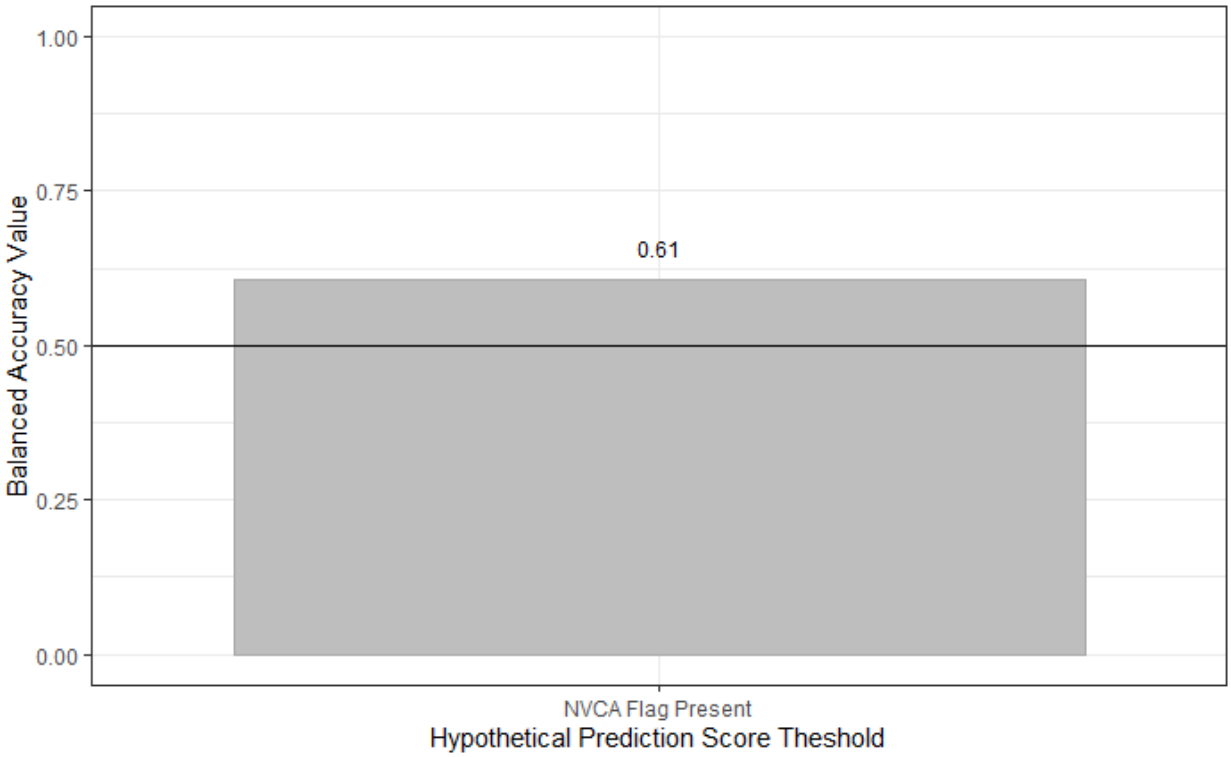


Figure 16 reports Balanced Accuracy measures for NVCA outcomes. The figure above shows that under this hypothetical prediction rule, the PSA NVCA Risk Flag obtained a balanced accuracy of 0.61, indicating that it provides a substantive increase in predictive power beyond classifying cases on limited information. Overall, this figure provides moderate evidence supporting the validity of the PSA with respect to NVCA outcomes.

Figure 17: Balanced Accuracy Measurements for FTA Outcome Constructions By Possible Prediction Score Thresholds

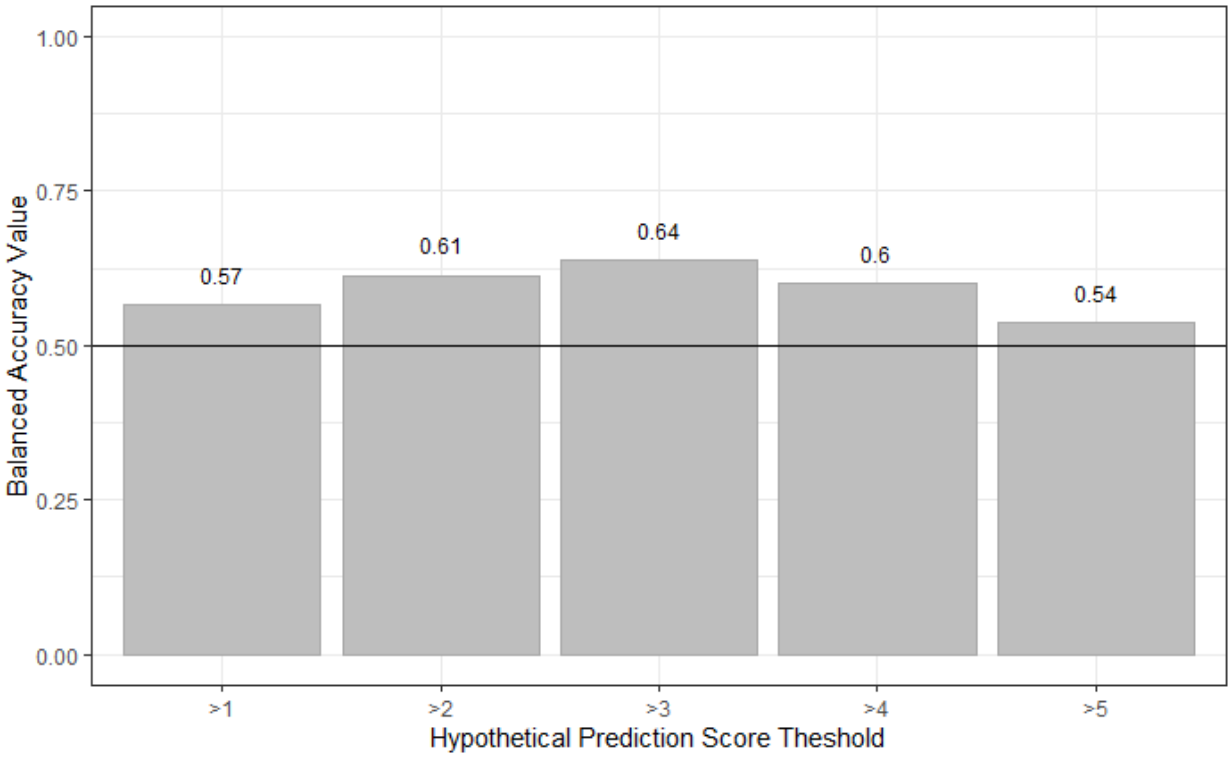


Figure 17 reports Balanced Accuracy measures for the FTA outcome construction for FTA score thresholds of 1, 2, 3, 4, and 5. The figure above shows that three of the five hypothetical prediction thresholds obtain balanced accuracy measures of 0.6 or higher, indicating that the PSA, under these hypothetical prediction rules, provides modest increases in classifying power. This figure provides some evidence supporting the validity of the PSA with respect to FTA outcomes.

Figures 15-17 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 achieved some classification gain, as evidenced by Balanced Accuracy scores above 0.5. Some exceeded the 0.6 value, suggesting modest classification gains. For NCA outcomes, the Balanced Accuracy metric achieved its maximum under the NCA Score > 3 threshold of 0.606 with a minimum of 0.539 under the NCA Score > 5 threshold. The FTA score also achieved its maximum Balanced Accuracy metric under the FTA Score > 3 threshold of 0.638 with a minimum of 0.537 under the FTA Score > 5 threshold. The NVCA outcome Balanced Accuracy metric is 0.608. The maximum values at each of these outcomes represent gains in predictive accuracy and even the minimum values indicate borderline marginal gains. These results provide moderate evidence of the overall validity of the PSA.

Figure 18: Balanced Accuracy Measurements for NCA Across Demographic Subgroups

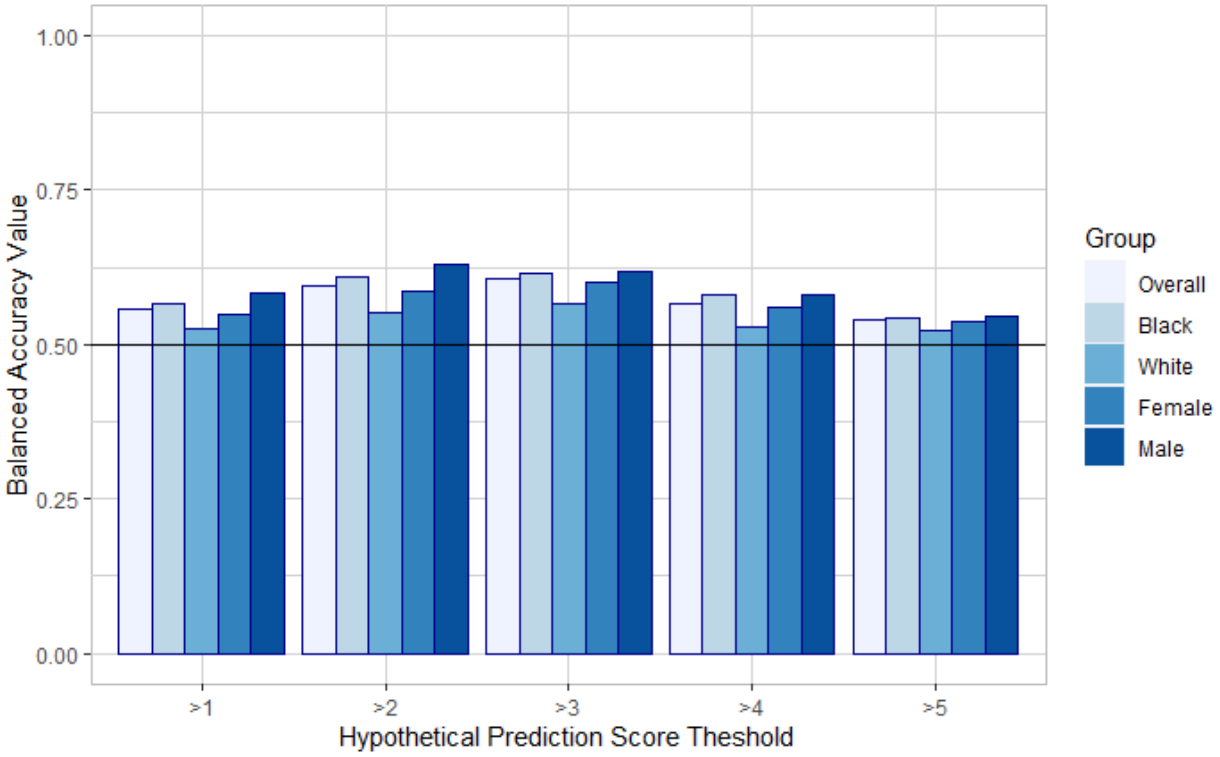


Figure 18 reports Balanced Accuracy measures for NCA outcomes. See above for a discussion of the overall figures. There are differences in the racial group paring under all but the >5 hypothetical thresholds, with the balanced accuracy metric indicating better accuracy for Black arrestees than White arrestees. These values range between 0.039 and 0.060. This figure provides evidence supporting the overall validity of the PSA with respect to NCA outcomes but some evidence against the equitable validity of the PSA with respect to racial groups.

Figure 19: Balanced Accuracy Measurements for NVCA Outcomes Across Demographic Subgroups

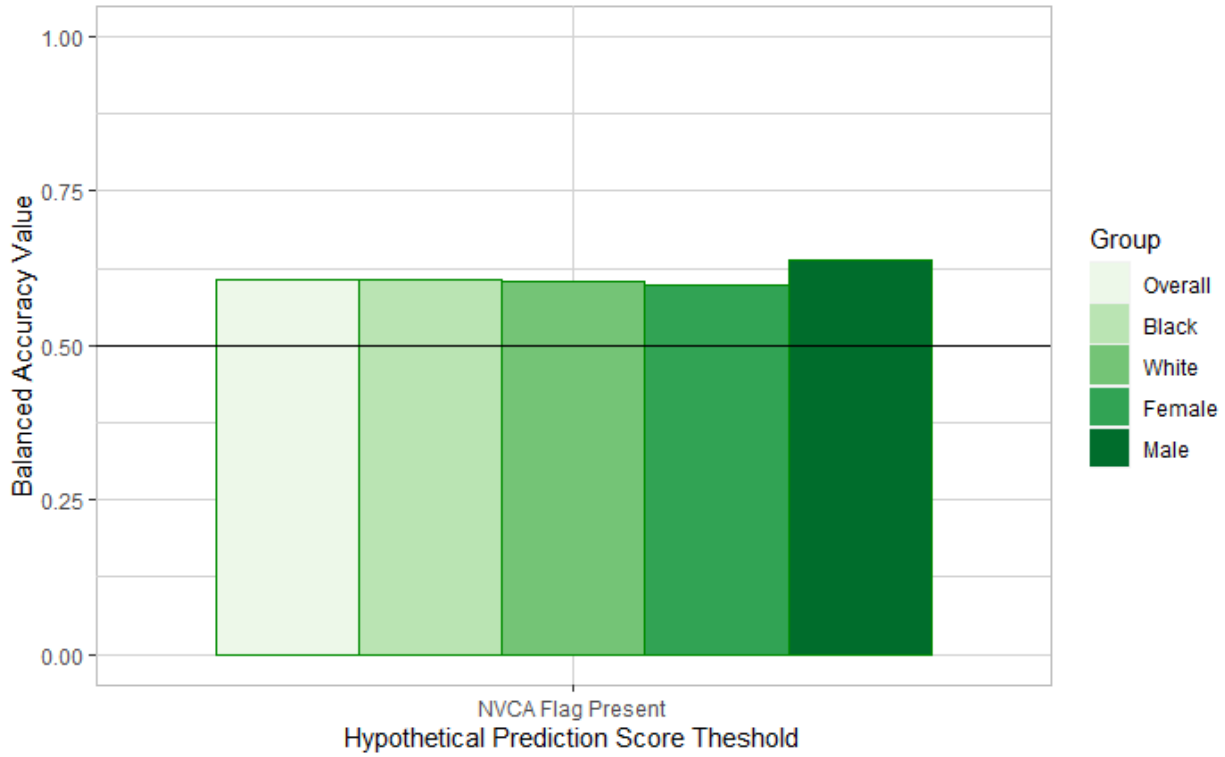


Figure 19 reports Balanced Accuracy measures NVCA outcomes. The figure above shows values near the .6 threshold for modest gains in classification power, but exceeding that value only for male individuals. These findings are consistent for both the overall study population as well as for each of the main study demographic groups. This figure provides some evidence supporting the overall validity of the PSA with respect to NVCA outcomes with no indication of meaningful differences in predictive power across racial groups and minor differences (0.041) across gender groups.

Figure 20: Balanced Accuracy Measurements for FTA Outcomes Across Demographic Subgroups

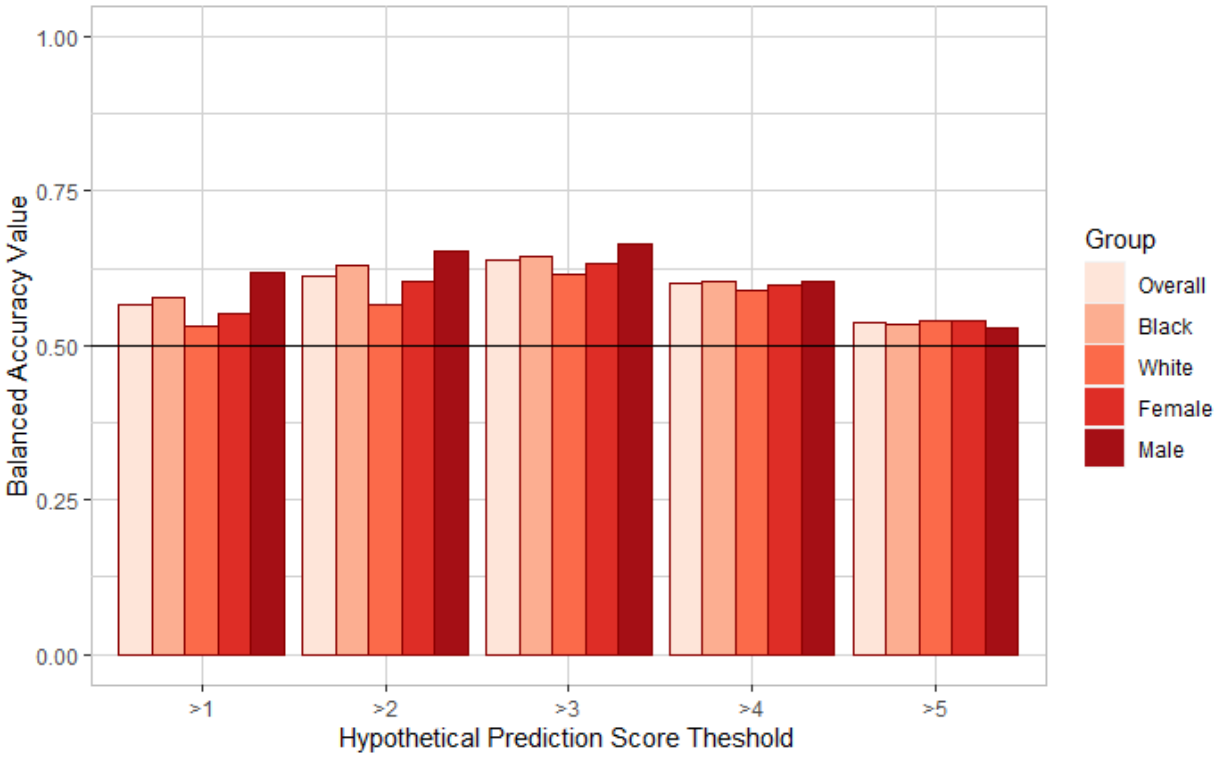


Figure 20 reports Balanced Accuracy measures for FTA outcomes. There appear to be substantive racial and gender differences for the >1, >2, and, to a lesser degree, >3 hypothetical threshold rules, with some as large as 0.062. Overall, this figure provides evidence supporting the validity of the PSA with respect to FTA outcomes and mixed evidence for 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 and FTA outcomes, the maximum difference in Balanced Accuracy across racial groups was 0.06 and 0.062, respectively, both of which were statistically significant; the relevant maximum differences for NVCA outcomes was 0.004. Comparing across gender groups, the maximum differences were 0.044, 0.041, and 0.066 for NCA, NVCA, and FTA outcomes, respectively. Each of these maximum differences suggests differential patterns by race and gender in Balanced Accuracy metrics. The Balanced Accuracy metric provides weak evidence against 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 and substantively significant differences by race and gender, and they are consistent with regard to racial categories. The analysis provides weak evidence that the PSA is not equitably valid. Key details are as follows:

- Significant differences existed in failure rates across racial demographic groups for FTA risk scores of 1, for the no-NVCA-flag classification, and for NCA scores of 2 and 4. This result provides weak evidence against equitable validity.
- Significant differences existed in failure rates across gender demographic groups for none of the NCA risk scores categories, for both of the NVCA-flag conditions, and for three of the FTA risk scores. Failure rate differences were inconsistent in direction for both NVCA and FTA. This result provides weak evidence against equitable validity.
- Racial group differences in N(V)CA/FTA failure rates were directionally consistent, with White individuals corresponding to lower FTA failure rates and lower N(V)CA failure rates than their Black individual counterparts. These results provided some, but weak, evidence against equitable validity.

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.[34] [35] The following figures plot outcome failure rates by relevant risk score across study demographic groups for each of the main outcome events.

[34] 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).

[35] 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 21: NCA Failure Rates by Demographic Group

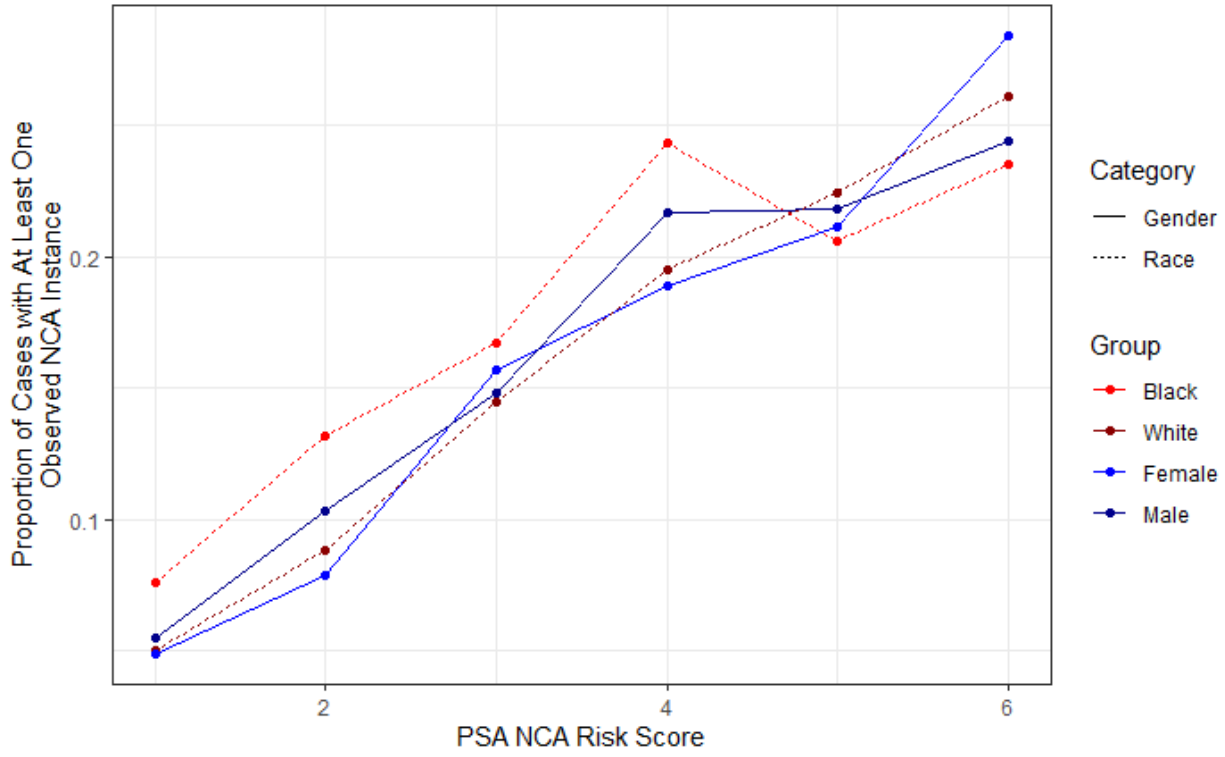


Figure 21 reports the observed failure rate for NCA outcomes across PSA NCA Risk Score categories by the four main study demographic groups: Black individuals, White individuals, male individuals, and female individuals. Line types and colors differ across category comparisons (race and gender). These 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 NCA risk scores of 2 and 4 (as well as overall observed failure rates). At these levels of NCA risk score, Black individuals had observed failure rates about 3.5% percentage points higher than their White individual counterparts. With respect to gender comparisons, there were no significant differences in observed failure rates between male and female individuals at any NCA risk. This figure provides weak support both for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates), as well as for equitable validity.

Figure 22: NVCA Failure Rates by Demographic Group

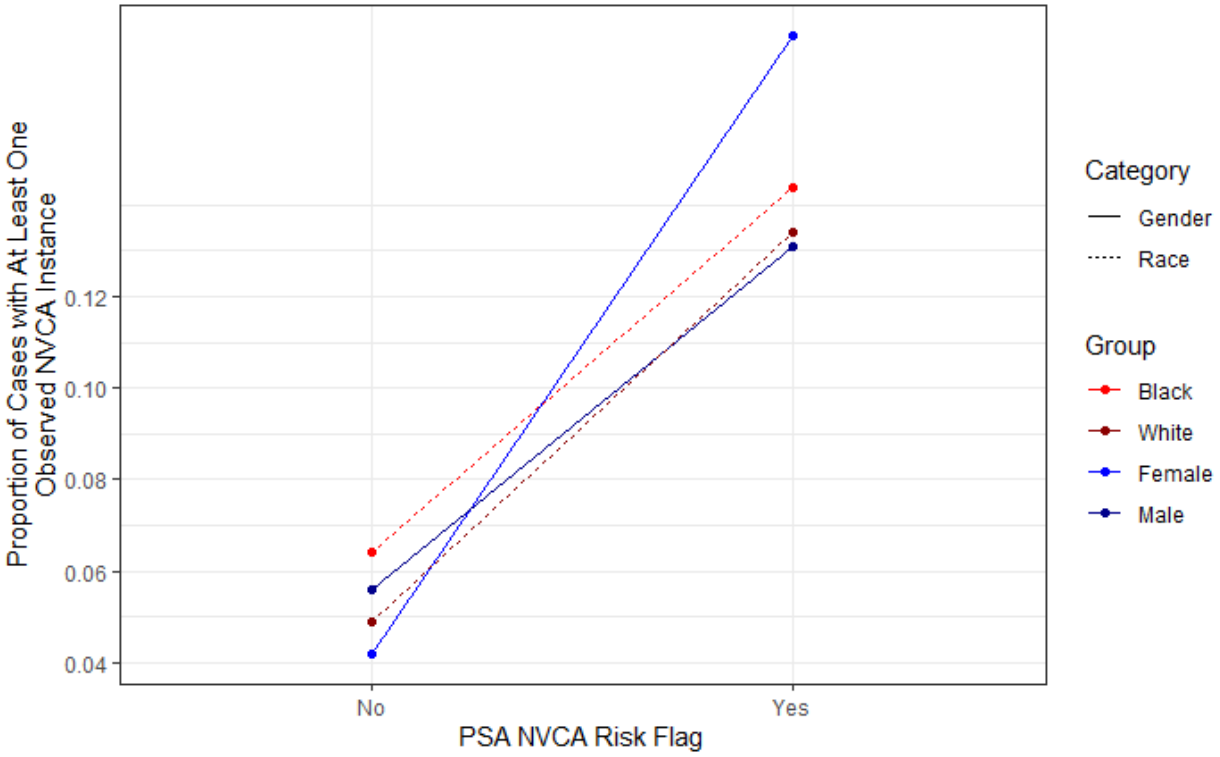


Figure 22 reports the observed failure rate for NVCA outcomes across categories of PSA NVCA Risk Flag presence by the four main study demographic groups: Black individuals, White individuals, male individuals, and female individuals. Line types and colors differ across category comparisons (race and 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 were significant differences in observed failure rates between Black individuals and White individuals only when the NVCA Risk Flag was not present. For cases without the NVCA risk flag, Black individuals had observed failure rates 1.4 percentage points higher than their White individual counterparts. With regards to gender comparisons, there were significant differences in observed failure rates between male and female individuals both when the NVCA risk flag was present and when it was not. When the flag was not present, female individuals observed failure rates 1.4% lower than their male peers. When the flag was present, male individuals experienced failure rates 4.6% lower than their female peers, a larger gap. This figure provides support both for overall validity of the PSA (higher risk scores are associated with higher observed failure rates), but mixed evidence regarding equitable validity.

Figure 23: FTA Failure Rates by Demographic Group

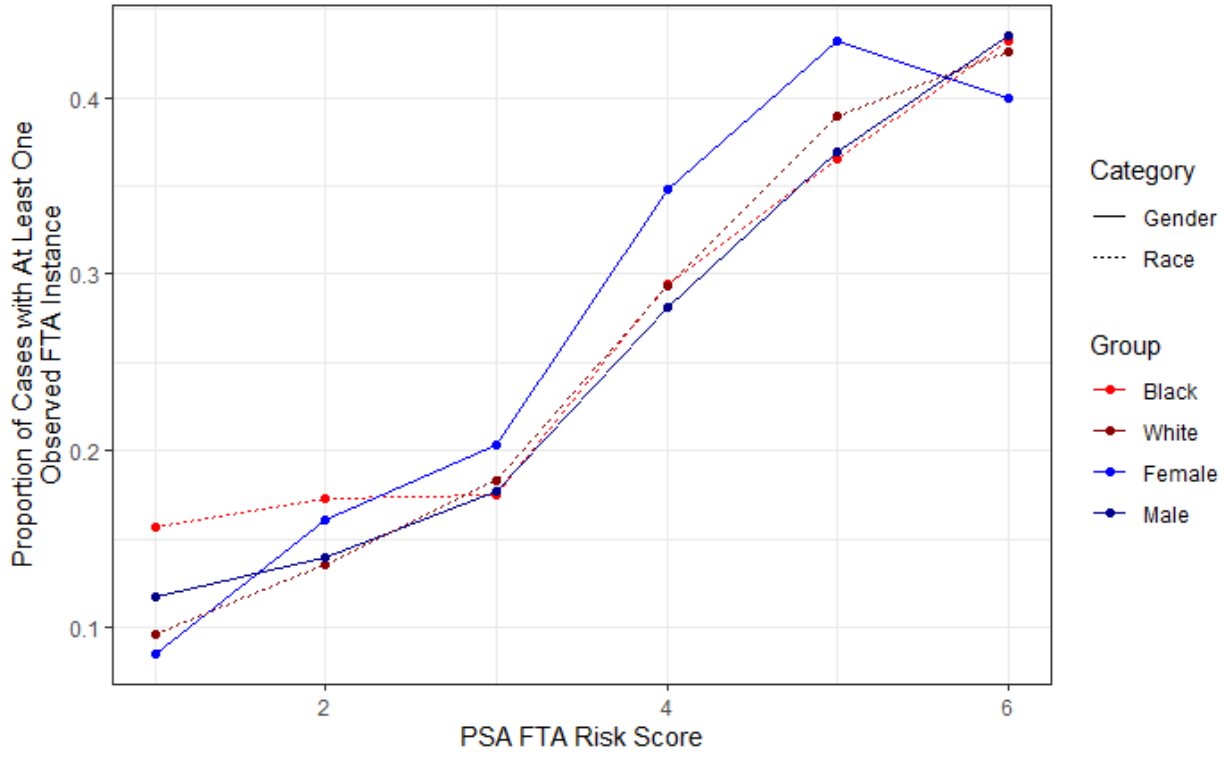


Figure 23 reports the observed failure rate for FTA outcomes across PSA FTA Risk Score categories by the four main study demographic groups: Black individuals, White individuals, male individuals, and female individuals. Line types and colors differ across category comparisons (race and gender). These 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 were significant differences in observed failure rates between Black individuals and White individuals for FTA scores of 1 and 2. At these levels of FTA risk score, Black individuals observed failure rates 3.3-5.2 percentage points higher than their White individual counterparts. There additionally were significant differences in observed failure rates between male and female individuals for the FTA risk score categories of 1, 4, and 5. These differences ranged from 3.2 to 6.7 percentage points, with female individuals observing lower failure rates for scores of 1 but higher failure rates for scores of 4 and 5. This figure provides support for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates) and weak evidence against equitable validity. While only two FTA score categories exhibited significant racial differences, these were somewhat large and consistent, while three categories supported gender differences that were also large but less consistent.

Figures 21-23 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 was separation. For NCA outcomes, there were significant ($p<0.05$) racial group differences in failure rates at NCA scores of 2 and 4 while no significant gender group differences in failure rates existed. For NVCA outcomes, significant racial group differences in failure rates existed for cases with no violence flag and for gender group differences in failure rates for both cases with and without the violence flag. For FTA outcomes, significant racial group differences in failure rates existed for cases with FTA risk scores of 1 and 2 and for gender group differences in

failure rates for cases with FTA scores of 1, 4, and 5. There were a number of scores that implied statistically different rates of failure for either race or gender pairs: NCA scores 2 and 4, both categories of the NVCA Flag, and FTA scores of 1,2, 4, and 5. Additionally, many of these differences in failure rates were of some magnitude, differing often by over 3 percentage points and as much as 6.7 percentage points. With regards to race, there was a consistent pattern with respect to difference in classifying information. White arrestee failure rates were lower than corresponding Black arrestee rates in each instance. Ultimately, the subgroup paired comparison of PSA score specific failure rates provides weak evidence that the PSA does not equitably validate. Significant differences existed in some number, but not in the majority, of score categories, many of which were potentially policy-relevant in size and consistent in direction.

ii. 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, some of which are small but are also consistent in how they affect each racial group. This analysis provides weak evidence of a failure of equitable validity. Key details are as follows:

- Each of the PSA risk scores/flags showed significant, positive correlations with the probability of observing a relevant outcome of roughly similar magnitudes to the bivariate regression case.
- Interactions of race and risk score, which indicates whether the predictive meaning of the risk score changes significantly across racial groups, was significant for NCA and FTA scores. The moderating effect of race on the score was a fourteen and nine percent decrease, for NCA and FTA, respectively, in the odds ratios for a Black arrestee of a given score relative to a White arrestee of the same score. The size and consistency of this effect provides weak evidence against equitable validity.

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.[36] 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 demographic group.[37] We focus on race, as opposed to gender, due to the fact that the racial

[36] 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).

[37] 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

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 24: Moderated Regression Predicted Probabilities by Race for NCA

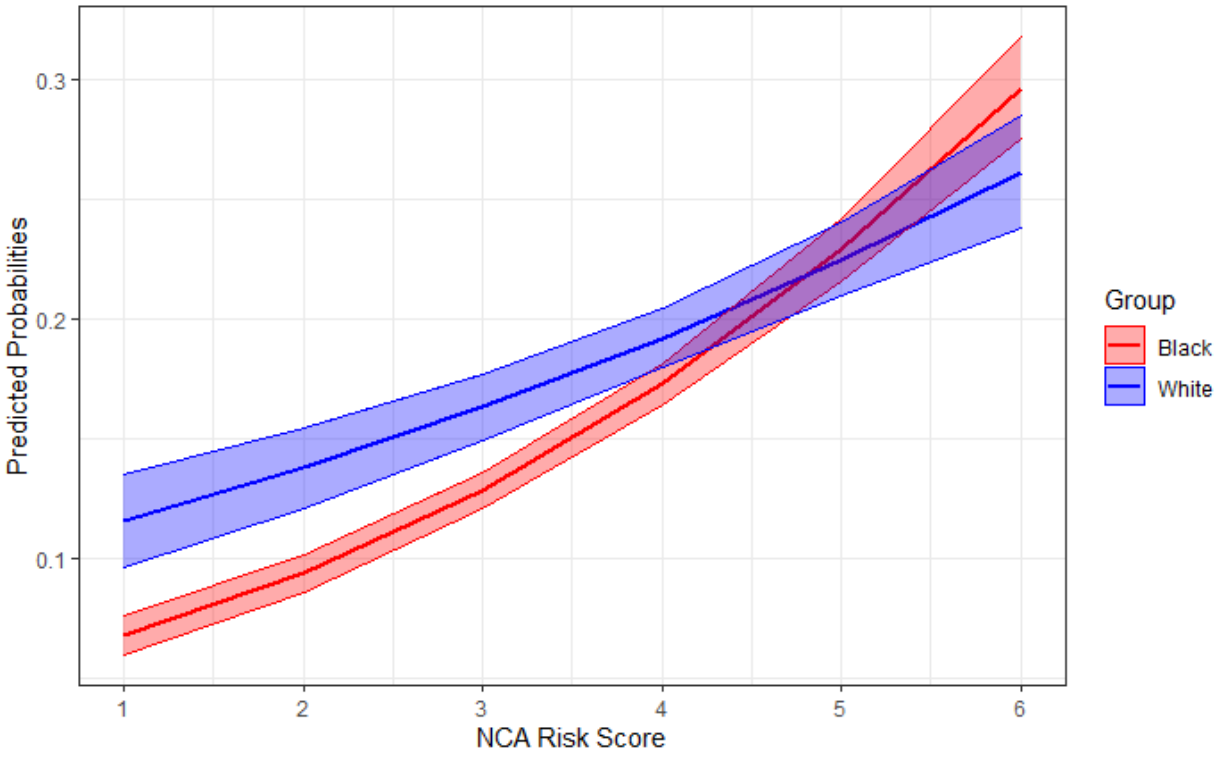


Figure 24 reports predicted probabilities and associated 95% confidence intervals for observing an NCA event obtained from the moderated regression model with both PSA score and race variables. The PSA NCA risk score had a significant, positive coefficient, indicating that higher NCA risk scores were significantly, independently associated with a higher probability of an observed NCA failure. A one unit increase in the NCA risk score was associated with a 42% (95% confidence interval of 37% to 48%) increase in the odds ratio of observing an NCA failure versus not observing an NCA failure. The interaction term was also significant (0.86 odds ratio on a 95% CI of 0.80 - 0.92), indicating that a one-unit increase in the NCA scale is associated with between an 8% and 20% 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 had lower predicted probabilities for observing an NCA than their White peers when only looking at NCA scores, which themselves were overall statistically significantly predictive of observed NCA events. The shaded confidence interval

---

assessment score relates to the relevant outcome. Evaluating the basic value of the risk score can be done by analyzing 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.

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 and substantively significant differences in predictive strength across racial subgroups.

Figure 25: Moderated Regression Predicted Probabilities by Race for NVCA

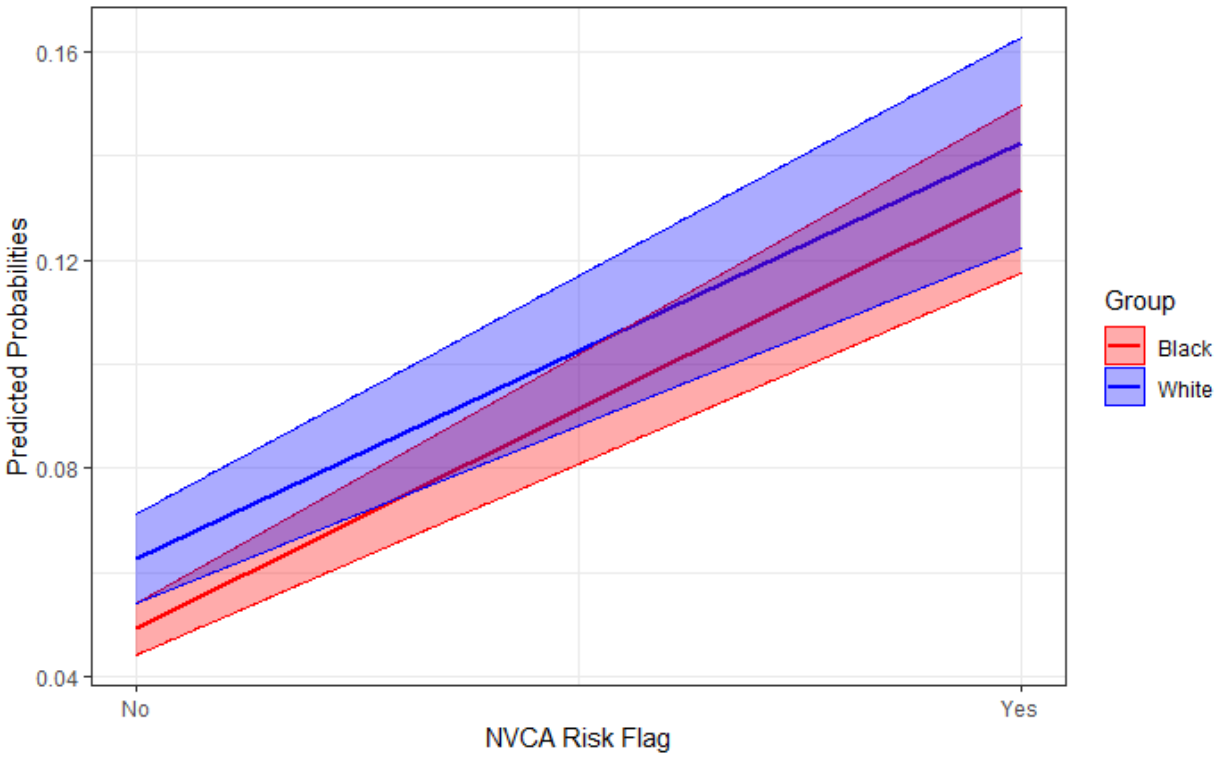


Figure 25 reports predicted probabilities and associated 95% confidence intervals for observing an NVCA event obtained from the moderated regression model with both PSA score and race variables. The PSA NVCA risk flag had a significant, positive coefficient, indicating that the presence of a risk flag is significantly, independently associated with a higher probability of an observed NVCA failure. The presence of an NVCA risk flag was associated with a 198% (95% confidence interval of 150% to 255%) increase in the odds of observing an NVCA failure versus not observing an NVCA failure. The interaction term was not significant. Functionally, this means that when taking into account racial categories, Black and White individuals had statistically similar predicted probabilities for observing an NVCA. The shaded confidence regions indicate no statistically meaningful differences in predicted probabilities. Overall, this figure provides support for the overall validity of the PSA, and we see no evidence of racial differences, for the NVCA outcome.

Figure 26: Moderated Regression Predicted Probabilities by Race for FTA

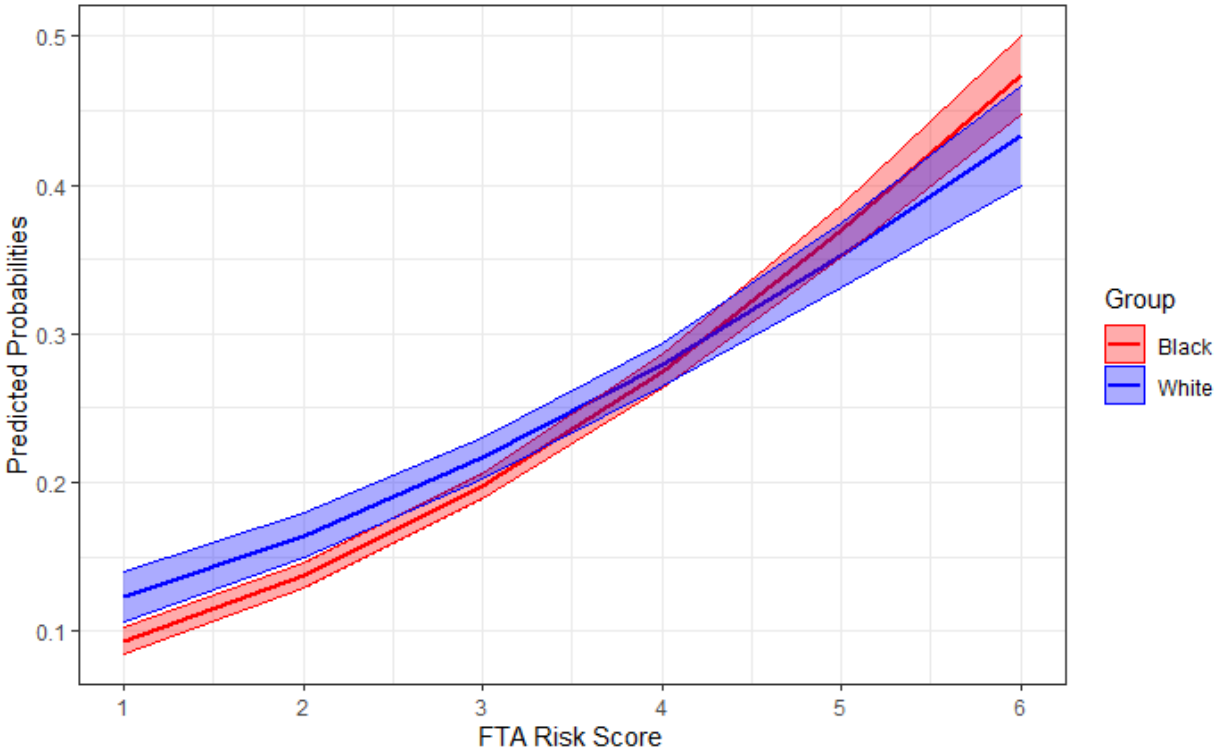


Figure 26 reports predicted probabilities and associated 95% confidence intervals for observing a FTA event obtained from the moderated regression model with both PSA score and race variables. The PSA FTA risk score had a significant (at the p<0.001 level), positive coefficient, indicating that higher FTA risk scores were significantly, independently associated with a higher probability of an observed FTA failure. A one unit increase in the FTA risk score was associated with a 54% (95% confidence interval of 49% to 60%) increase in the odds of observing an FTA failure versus not observing an FTA failure. The interaction term was also significant (0.91 odds ratio on a 95% CI of 0.86 - 0.97), indicating that a one unit increase in the FTA scale was associated with between a 3% and 14% decrease in the odds ratio of observing an FTA event for Black individuals relative to their White individual peers. Functionally, this means that when taking into account racial categories, Black individuals had lower predicted probabilities for observing a FTA than their White peers when only looking at FTA scores, which themselves were overall statistically significantly predictive of observed FTA events. 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 minor, and thus does not constitute much 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 24-26 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 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 1.42 on a 95% confidence interval of (1.37, 1.48), while the bivariate estimate was 1.36. For the FTA model, the exponentiated coefficient on the FTA Score Scale is 1.54 on a 95% confidence interval of (1.49, 1.60), while the bivariate estimate was 1.50. For NVCA, the exponentiated coefficient estimate for the presence of the NVCA Flag is 2.99 on a confidence interval of (2.57, 3.47), while the bivariate exponentiated estimate was 2.87. 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 significant at the $p < 0.001$ level with an exponentiated coefficient estimate of 0.86 on a confidence interval of (0.80, 0.92). 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 fourteen percent lower for Black individuals than the corresponding odds ratio for the same NCA score for White individuals. This fourteen percent moderating effect of race is about 1/3rd the size of the effect of a one unit increase in the NCA score. For FTA, the interaction term is significant at the $p < 0.01$ level with an exponentiated coefficient estimate of 0.91 on a confidence interval of (0.86, 0.97). This indicates that for each level of the FTA Score Scale the odds ratio of observing at least one FTA event during the pretrial period is about nine percent lower for Black individuals than the corresponding odds ratio for the same FTA score for White individuals. This nine percent moderating effect of race is about 1/6th the size of the effect of a one unit increase in the NCA score. For NVCA the interaction term representing the moderating effect of race on the relevant NVCA risk flag is not statistically significant. These moderating effects are large and consistent enough to provide weak evidence against the equitable validity of the PSA.