

Validation of the PSA in McLean County, IL

Respectfully Submitted

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

November 9, 2020

Table of Contents

Executive Summary	3
Introduction	5
I. McLean County, the PSA, and Validation.....	5
a. McLean County	5
b. The PSA.....	6
c. PSA in McLean County	7
d. Validation of Risk Assessment Instruments	7
e. Data Available	9
i. Data Sources	9
ii. Data Limits.....	9
II. Findings	10
a. Outcome Definitions	11
b. Descriptive statistics	11
c. Traditional validation techniques	15
i. PSA scores and failure rates.....	15
ii. Bivariate correlations	20
iii. Area under the curve	25
d. Techniques Used Outside the Pretrial Context	28
i. Regression	28
ii. Balanced accuracy measures	32
e. Validation by Racial And Gender Groups	41
i. PSA scores and failure rates by race	41
ii. Moderated regression	45

Executive Summary

On January 11, 2016, McLean County, IL integrated the Public Safety Assessment (“PSA”) and the accompanying Decision Making Framework (“DMF”) into its pretrial processes. The PSA is a pretrial risk assessment instrument, a process that uses criminal history factors and age inputs to produce scores that classify an individual’s risk of misbehavior if released pretrial. Specifically, the PSA classifies individuals on risk of being arrested or cited for new criminal activity (“NCA”) and failure to appear (“FTA”) through two 1-6 integer scales, and on risk of new violent criminal activity (“NVCA”) through an on-off “flag.” The PSA scores are typically accompanied by the DMF, which incorporates the objective information from the PSA with community-specific determinations regarding local policy and values, state statutes, and jurisdictional resources to produce a release recommendation as well as (in locations that choose to use it this way) a supervision level to be imposed if the individual is released. The PSA scores rely on objective data, and the scoring system is the same in all jurisdictions. The DMF recommendation system can be different in each jurisdiction. The decision about whether to release or detain an individual, and the level of supervision accompanying any release, rests always with the Judge. The PSA does not produce a recommendation, and the DMF’s recommendation is not binding. The PSA was developed with support from Arnold Ventures, a Houston-based philanthropy, to reduce the burden placed on vulnerable populations at the frontend of the criminal justice system.

The Access to Justice (“A2J”) Lab was asked to conduct a validation study of the PSA in McLean 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 McLean’s use of the PSA from the Integrated Justice Information System, a McLean-administered data repository integrating criminal justice data from multiple county sources. The data addressed PSAs calculated between January 14th, 2016 and January 19th, 2020.

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

- There was moderate evidence for the overall validity of the PSA scales in McLean. Most validation techniques (simple plots, logistic regression, balanced accuracy measures, and area under the curve for NCA) provided either moderate or strong evidence for validity, while the remainder (correlations, area under the curve for other than NVCA) provided weak evidence of validity. No technique suggested invalidity.
- Analyses assessing potential differences in PSA operation by gender and race were inconclusive. There was some weak evidence suggesting differences in scale performance for race and gender groups, but also some weak evidence to suggest no such differences. With respect to gender (but not race), evidence occasionally showed

directionally inconsistent differences (i.e., one scale showed higher failure rates for men than women, while another scale showed the opposite).

- There was weak evidence to suggest that for FTA, 1-level scale increases did not correspond to similarly sized jumps in failure rates. For NCA, there was evidence to suggest that 1-level increases did correspond to similarly sized jumps in failure rates, except perhaps for minor departures in the 2-3 and 3-4 score transitions.

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 McLean County, Illinois. This report analyzes data with respect to PSA calculations made in McLean for felony and misdemeanor arrests from January 14th, 2016 to January 19th, 2020 as well as corresponding rates of failure to appear (“FTA”), new criminal activity (“NCA”), and new violent criminal activity (“NVCA”) among those released over the same time period.

In brief, validation of risk assessment instruments consists of comparing the classifications individuals (as of particular arrest events) receive on 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 McLean 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.

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

The A2J Lab is appreciative to the McLean County Courts and the McLean County Adult Court Services whose assistance made this report possible.

I. McLean 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 McLean 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 McLean County. Section D discusses the nature of validation of risk assessment instruments as applied to McLean 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. McLean County

McLean County, Illinois is located in the central portion of Illinois and has a population of 171,517 people.¹ The racial makeup of the county is 83% White with an African-American population of 8.4%.²

¹ U.S. Census Bureau (2020). *QuickFacts McLean County, Illinois*. Retrieved from <https://www.census.gov/quickfacts/mcleancountyillinois> (last visited Nov. 5, 2020).

² *Id.*

b. The PSA

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

The PSA is a pretrial risk assessment instrument (“PRAI”) 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

The PSA was developed with support from Arnold Ventures, a Houston-based philanthropy, to reduce the burden placed on vulnerable populations at the frontend of the criminal justice system.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 McLean 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 McLean County

PSA Implementation

This section discusses pretrial processes and the PSA in McLean County. It briefly describes the PSA's implementation history, including the classes of arrests to which McLean County applied the PSA.

McLean implemented the PSA-DMF System January 11, 2016 for both misdemeanors and felonies.⁷ When an individual was arrested in McLean, the police officer brought the defendant to the McLean County Jail located in the McLean County Law and Justice Center.⁸ The PSA-DMF System Report was created by staff of the McLean County Pretrial Unit known as assessors. The current charge factor of the PSA-DMF System was based on the arresting officer's recommended charges.⁹ Defendants charged with select misdemeanor offenses could bond out early according to a bail schedule.¹⁰ Defendants charged with either a felony or a misdemeanor not on the bail schedule went to a bond hearing.¹¹ The bond hearings were done by video.¹²

The Office of the State Attorney charged the defendant before the bond hearing. After the defendant had been officially charged, Pretrial reviewed whether the State Attorney charges varied from those the arresting office recommended. If so, and if those changes might have affected the PSA-DMF System scores, Pretrial attempted to recalculate the PSA.¹³ Ordinarily in such situations, Pretrial had enough time to update the PSA scores. In rare instances, when the State Attorney charged a defendant right before a bond hearing, Pretrial did not have an opportunity to review the change before the bond hearing.¹⁴

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.

⁶ See <https://www.psapretrial.org/about#jurisdictions-united-states> (last visited Feb. 19, 2020).

⁷ Matthew Stubenberg, Memo, "McLean A2J Lab Meeting," Memorializing Conversation on July 30, 2020 (on file with the Access to Justice Lab).

⁸ Matthew Stubenberg, Memo, "McLean A2J Lab Meeting," Memorializing Conversation on October, 13 2020 (on file with the Access to Justice Lab).

⁹ Stubenberg Memo, supra note 6.

¹⁰ Id.

¹¹ Id.

¹² Id.

¹³ Id.

¹⁴ Stubenberg Memo, supra note 7.

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.¹⁵ 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 Judges 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 McLean County.¹⁶

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 Judge incarcerates an individual, which would then eliminate (or greatly reduce) the possibility of an FTA or N(V)CA. Despite this fact, the validation study we report here, like all previous PRAI validation studies of which we are aware, 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 McLean County Judges react to that

¹⁵ 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).

¹⁶ With AV’s support, the A2J Lab is pursuing RCTs in four jurisdictions in the United States.

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 Judge has access to information other than the PSA that helps the Judge 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

McLean's criminal justice stakeholders shared an integrated data repository called the Integrated Justice Information System (IJIS). The McLean court system managed the repository, which contained PSA data from Pretrial, case information from the court, and jail release data from the Sheriff's Department. Calculating the main N(V)CA and FTA information was straightforward due to the integrated nature of the data system. The majority of necessary information was in a single table where each instance represented an outcome of a specific charge in a case stemming from a unique PSA generating event. This main data source contained roughly 300,000 charges-outcomes. These charge-outcome instances stemmed from 7,796 unique court cases, each of which was attached to a unique PSA. The Lab aggregated data from the main table to the case level utilizing a case number identifier. The only meaningful disagreement between charges in the same case concerned disposition dates, since different charges were often disposed of on different dates. In these instances, we considered the date the last charge was disposed of as the case disposition date. Disposition dates, filing dates, arrest dates, release dates, arrestee information, PSA inputs, PSA outputs, and initial hearing results were all obtained from this single data source. FTA information and the warrant information necessary to determine if a bench warrant was issued were obtained from a separate table. We joined these two sources of information using a common case number and individual ID number found in both data sources. Individual PSA instances were filtered out on the basis of jail release date information. For 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 6,345 entries. Each entry was a single PSA assessment attached to a specific case that featured at least one day of pretrial release.

ii. Data Limits

The data received was limited to records provided by the McLean County departments identified previously. McLean officials and the A2J Lab explored the possibility of obtaining statewide arrest data from the Illinois State Police (ISP). The request for statewide arrest data from ISP was submitted in May 2020 but as of October 10, 2020, the data request had not yet been approved.

II. Findings

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

A conclusion that a tool is valid, at least partially, 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.

One obvious 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.

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 Judges 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 Judges 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 charge that carries the potential of incarceration as a sentence, is observed during a case's pretrial period, *i.e.*, from the initial bail hearing until case disposition.

NVCA- An NVCA event is observed if a new arrest event, with an associated charge that carries the potential of incarceration and is considered a violent charge, is observed during a case's pretrial period.

FTA- An FTA event is observed if the court records indicate a missed court event during a case's pretrial period that resulted in the issuance of a bench warrant. This event must be attached to the original PSA case number, *i.e.*, it occurred in the case from which the PSA originated.

b. Descriptive statistics

The study population consists of 6345 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 4707 unique individuals charged with either misdemeanors or felonies over the period of January 14, 2016 to January 20, 2020. Individuals were recorded with six separate racial category identifiers; however, less than 6% were categorized as something other than Black or White. For the purposes of readability, we condensed the original racial categories to 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, had White individuals as the largest racial group, with 3309 PSA instances indicating a White individual, or about 52.2% of the study population. PSA instances indicating a Black individual comprised a 10% smaller portion of the population at 42%, or 2662 PSA instances. 5.7% of the study population were PSA instances that indicated an individual of a different racial group. In terms of gender distribution (Figure 2), 73.7% 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 31.7 years old at time of arrest

and a median age of 28 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 81% of all PSA instances had an arrestee observe at least 1 day of pretrial release (the other 19% 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 (0.846 vs. 0.781), while female arrestees had statistically and substantively significantly higher rates of pretrial release than male arrestees (0.906 vs. 0.785). 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 but lower FTA rates than female arrestees (although the FTA difference is insignificant ($p=0.3$), while White arrestees observed lower failure rates across all outcomes than Black arrestees. Overall, 15.5% and 18.1% of the study population observed either an NCA or FTA event, respectively. Only 5.5% of PSA instances observed an NVCA failure during the relevant pretrial release period.

Figure 1: Distribution of Individual Race Across PSA Instances

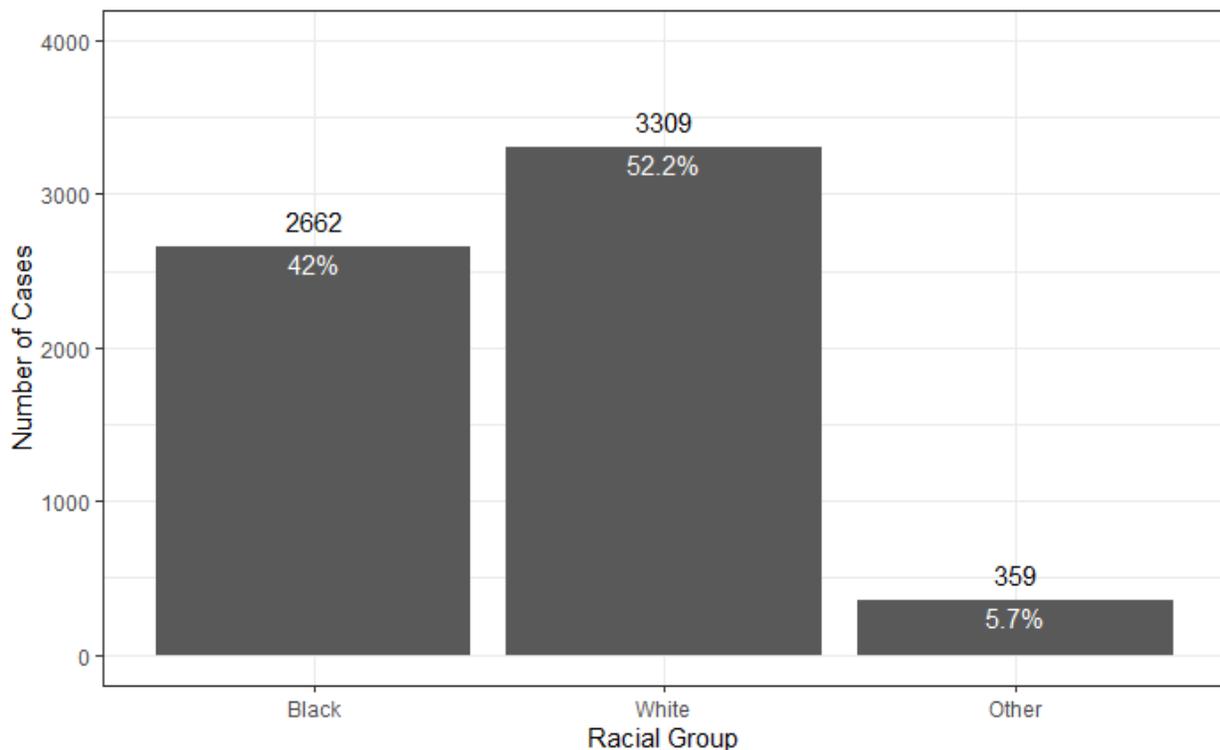


Figure 1 displays the distribution of racial categories for individuals. Each portion of the chart indicates the percentage of unique PSA submissions that listed the relevant Race category for the individual. Thus, individuals could appear here more than once. The initial data obtained from McLean County contained six separate racial categories; however, two of the categories, White

and Black, represented 94.2% of all cases. White individuals were the modal category, representing the majority of individuals who received a PSA assessment at about 52% of the study population.

Figure 2: Distribution of Individual Gender Across PSA Instances

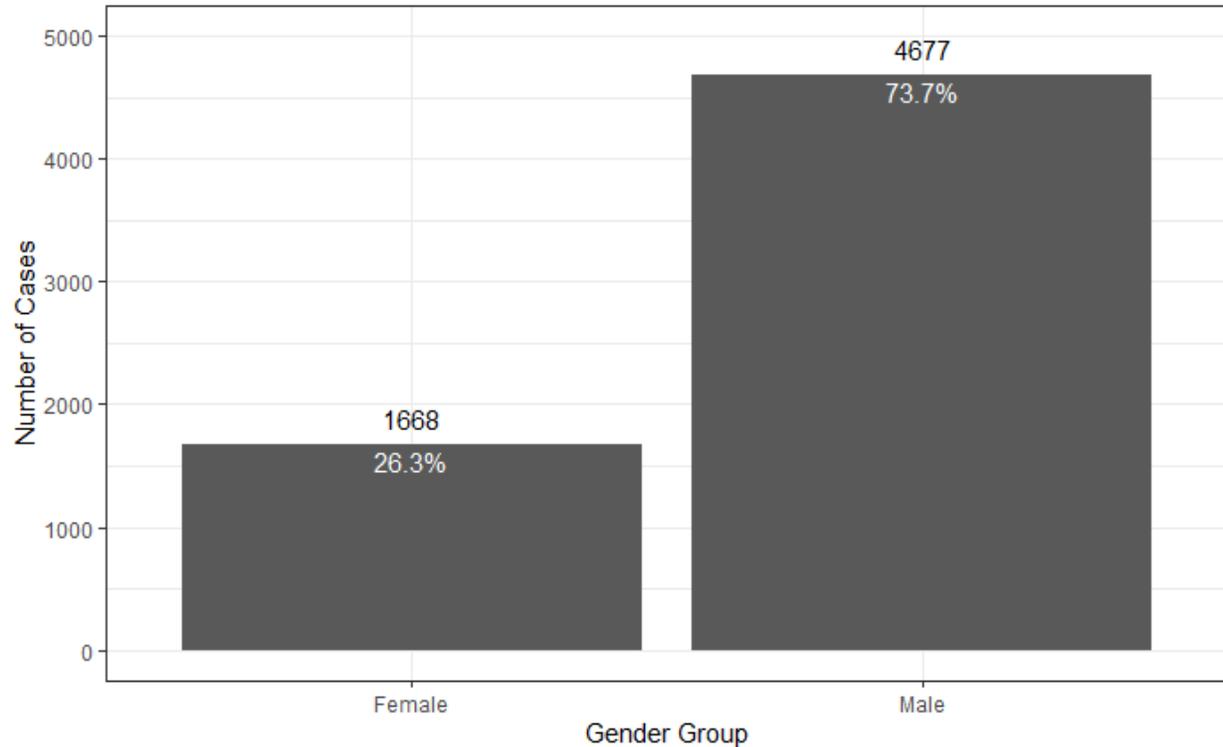


Figure 2 displays the distribution of gender categories within the study population. Female individuals represent just over a quarter of the total study population at 26.3%, which made male individuals the overwhelming majority.

Figure 3: Distribution of individual Age Across PSA Instances

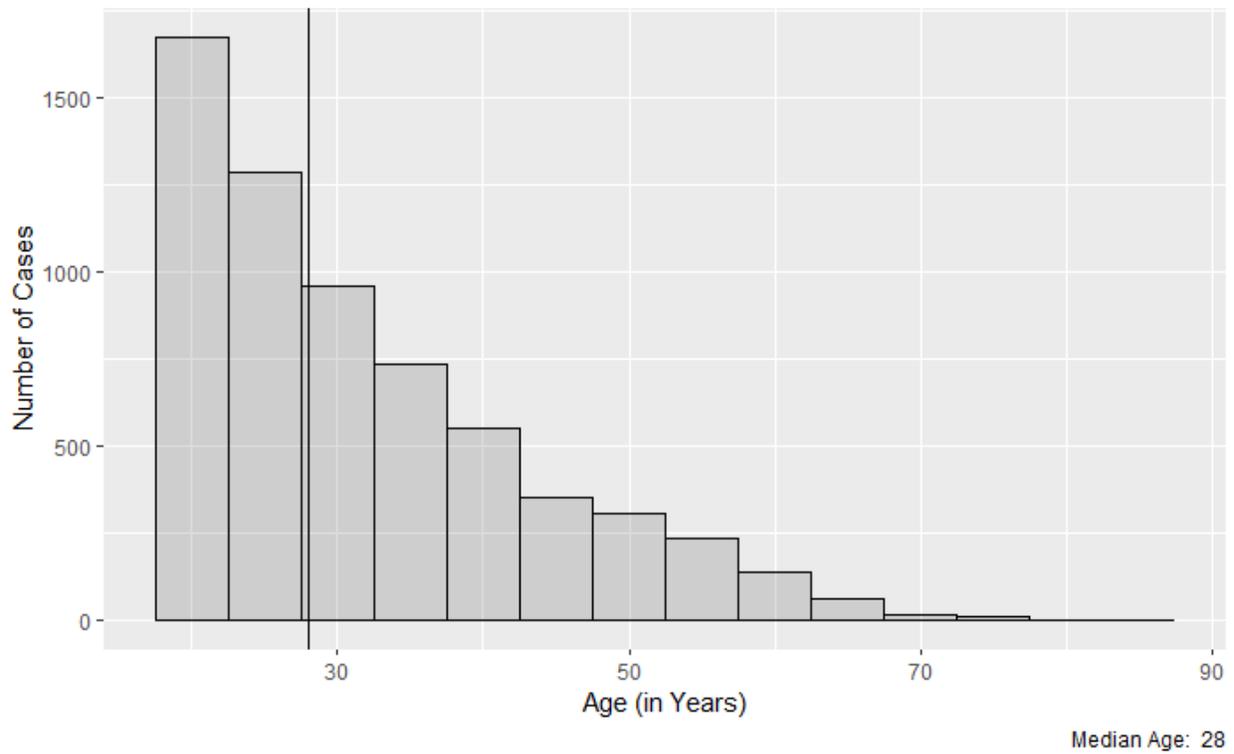


Figure 3 plots the distribution of individual age within the study population. This figure indicates a higher fraction of younger individuals. The mean age was slightly above 31 years old, with a median age of 28 years old. Thus, half of the study population existed within a 10-year age range of 18-28, while the rest occupied a 55-year age range, from 29 to 84.

Table 1: Summary of Failure Rates by Demographic Group

Group	# of PSAs	Released (N)	NCA Fail Rate	NVCA Fail Rate	FTA Fail Rate
Overall	7796	6345	0.155	0.055	0.181
Black	3886	3036	0.156	0.06	0.194
White	3926	3324	0.145	0.048	0.147
Female	1841	1668	0.112	0.039	0.192
Male	5955	4677	0.171	0.061	0.177

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 6.6% higher than Black arrestees (84.6% vs. 78.0%), while Female arrestees had at least 1 day of pretrial release at a rate about 12.1% higher than Male arrestees (90.6% vs. 78.5%). 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 rates but lower FTA rates 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. Overall, the PSA for the most part appears 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 was overall valid, and provides some but not conclusive evidence that the PSA was uniformly valid, more so for the NCA scale than the FTA scale. Key details are as follows:

- N(V)CA and FTA measures showed consistent increases in failure rates as scores increased, with the exception of FTA scores of 1-2 and 5-6. These results provide

strong evidence that all three PSA scales were, for the most part, overall valid according to this statistical technique.

- NCA failure rate increases across scores did not differ significantly, i.e. failure rate differences between NCA scores of 1 and 2 were statistically similar to differences between NCA scores of 3 and 4. This fact provides evidence the NCA scale was uniformly valid under this technique.
- There were notable variations in the FTA failure rates at steps 3-4 and 4-5. Although these differences were not statistically significant, when examined with other evidence, this analysis provides evidence that the FTA PSA scale was not uniformly valid.

The failure rate for an event, be it a N(V)CA or an FTA, is defined as the proportion of cases that observed at least one of the relevant events during the appropriate time frame. The goal of the failure rate analysis is to assess whether there are statistically significant differences in the rate of failures across consecutive levels of the relevant risk score scale.¹⁷ We use a 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 validate under the uniform validation framework, the magnitude of the differences in failure rates between each paired score comparison should not differ significantly for either the NCA or FTA risk score scale. The following figures plot the overall failure rates for each relevant PSA Risk Assessment score across each of the three outcome events: N(V)CA/FTA. They show that under this definition, the PSA is overall valid for each consecutive score comparison. The PSA's NCA scale, but not the FTA scale, appears uniformly valid with respect to all comparisons; the FTA scale shows much smaller increases in risk for consecutive scores at the lower end of the scale, compared to consecutive score comparisons at the higher end of the scale.

¹⁷ For studies that adopt this approach in whole or part, see:

- DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. "The public safety assessment: A re-validation and assessment of predictive 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

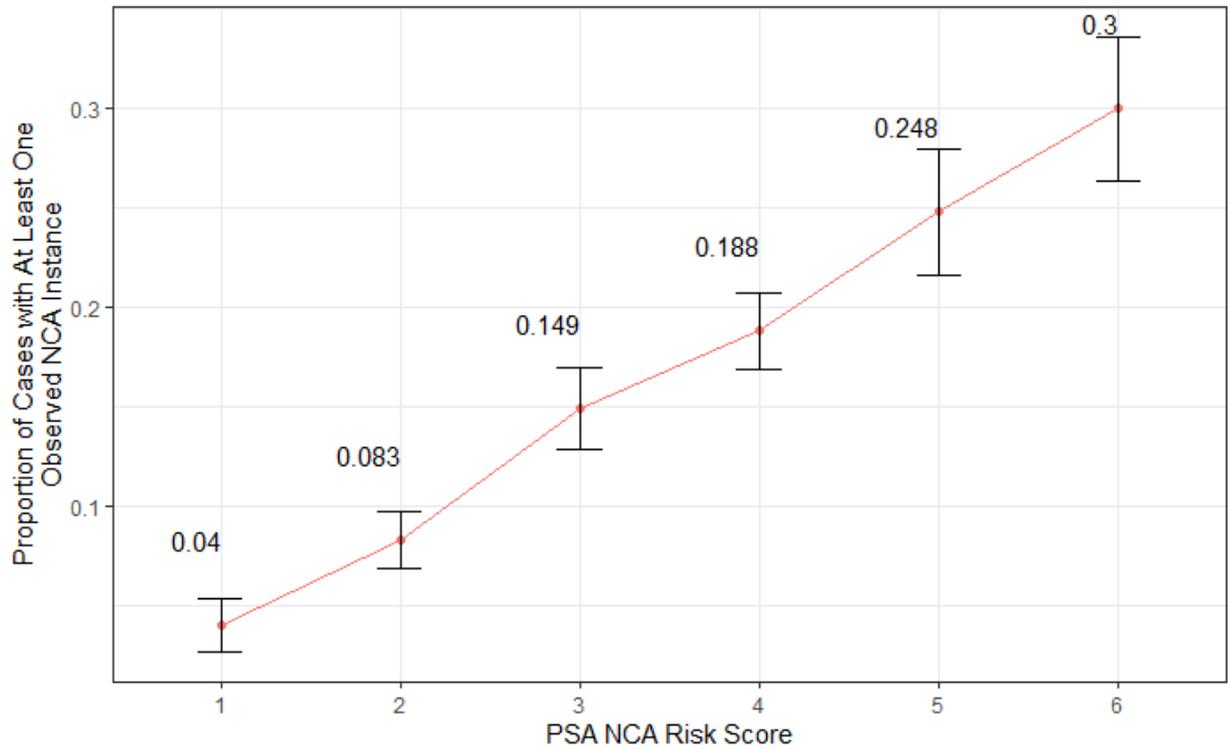


Figure 4 shows the relevant failure rates and associated 95% confidence intervals for NCA by risk score category. A valid risk assessment tool should show significant increases in failure rate at each subsequent level of the associated risk score. The lack of overlap between confidence intervals and the next score failure rate for consecutive paired scores indicates that each subsequent increase in the PSA NCA risk score was associated with a significant increase in failure rates. All differences are statistically significant. A one-unit change in risk score level was associated with a failure rate increase of roughly 5%. The increases for each subsequent score increase were fairly uniform and consistent. Overall, this figure provides evidence supporting the overall and uniform validity of the PSA with respect to NCA outcomes.

Figure 5: NVCA Failure Rates

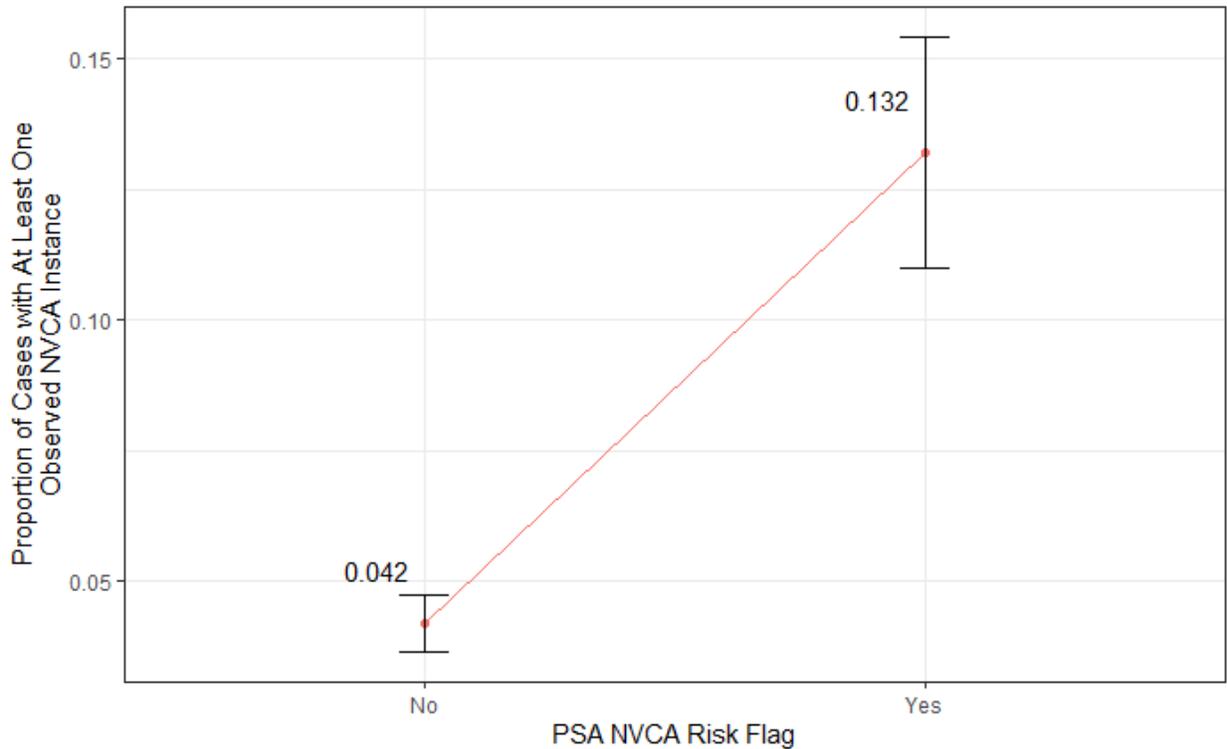


Figure 5 shows the relevant failure rates and associated 95% confidence intervals for NVCA by presence of the NVCA Risk Flag. An overall valid risk assessment tool should show significant increases in failure rate when a binary prediction flag is present, which could be indicated by no overlap between the confidence intervals. The presence of the PSA NVCA risk flag was associated with a significant increase in failure rates of 9 percentage points. The difference was statistically significant, and thus provides evidence supporting the validity of the PSA with respect to NVCA outcome.

Figure 6: FTA Failure Rates

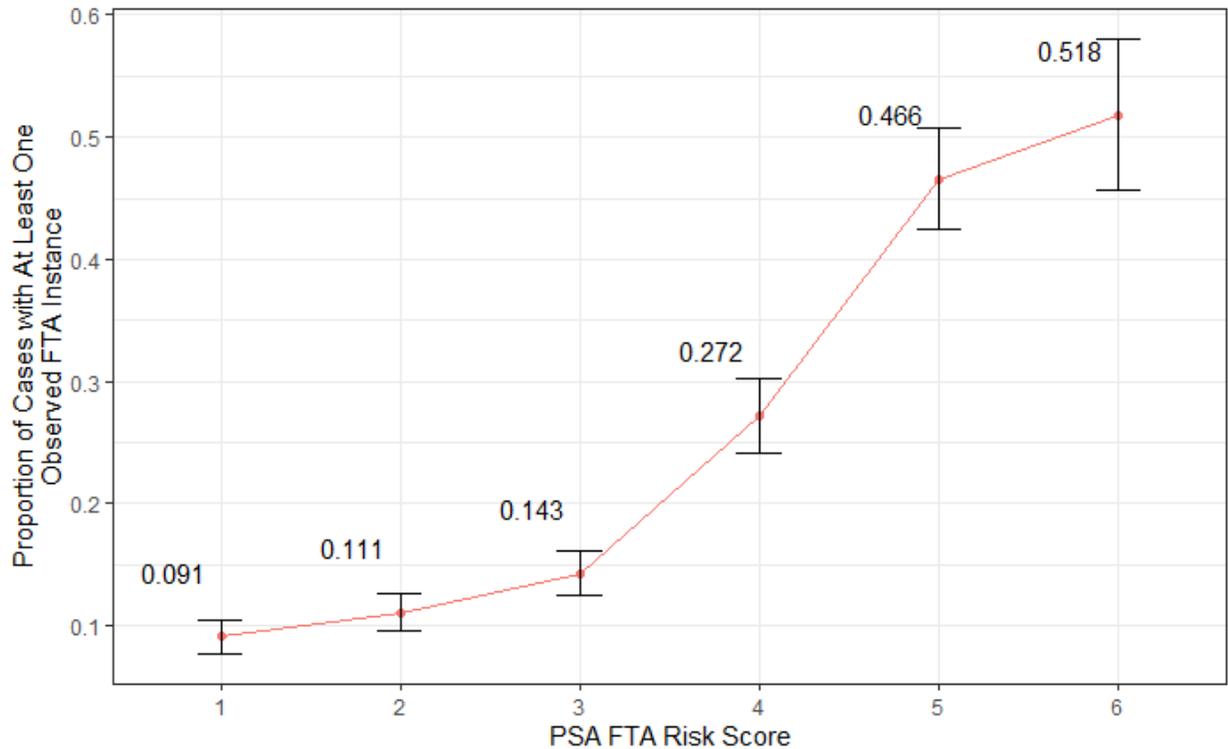


Figure 6 shows the relevant failure rates and associated 95% confidence intervals for FTA by risk score category. Only three (2-3, 3-4, and 4-5) transitions in FTA scores were associated with a statistically significant increase in failure rates. The differences for the 1-2 and 5-6 transitions had associated p-values of $p=0.056$ and $p=0.169$. A one unit change in risk score was associated with a variety of increases in failure rates, ranging from an increase of only 3.2% to an increase of 12.9%. Overall, this figure provides some evidence supporting the overall validity of the PSA for FTA, but provides evidence suggesting that the PSA was not uniformly valid.

Figures 4-6 demonstrate consistently increasing failure rates for each of the three PSA outcome events. Risk assessment scores for NCA, NVCA, and FTA all report higher failure rates for the higher score of each consecutive score pairing (or the single pairing for NVCA). NCA failure rates corresponded to a minimum of 4% for cases with risk scores of 1 and maximum failure rate of 30% for cases with risk scores of 6. FTA failure rates achieve minimum and maximum rates at scores of 1 and 6, 9.1% and 51.8% respectively. Cases with an NVCA flag present were associated with approximately three times the failure rate of cases without the flag present. 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 1-2 and 5-6 on the FTA scale. With respect to uniform validity, rough statistical comparisons¹⁸ suggest that each step increase in the NCA scores was associated with

¹⁸ 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 the intervals overlap with at least one other interval, however, this is due in part to the very large intervals over some of the consecutive score intervals. For example, the increase in failure rates from an FTA score of 4 to 5 exists on a

approximately equivalent increases in failure rates. For FTA scores, the differences in failure rates vary considerably and have large confidence intervals, suggesting that this score was not uniformly valid.

ii. Bivariate correlations

This subsection provides the results of bivariate comparisons, also known as correlations. This correlation analysis provides some, but weak, evidence that the PSA was overall valid. For NCA and NVCA, overall risk scores achieved larger correlation coefficients than any of their individual input factor coefficients, suggesting that input factors provide some non-overlapping predictive information that is preserved by the assessment’s calculation methods. Such was not the case for FTA, where one item had a larger coefficient than the overall scale. Key details are as follows.

- Each input factor across all PSA metrics was statistically significantly correlated with the relevant outcome in the expected direction.
- These correlations were generally small, ranging between 0.05 and 0.34.
- For NCA and NVCA, the largest correlations in magnitude for each PSA metric were associated with the overall risk score (or flag) and the relevant outcome metric. For FTA, the largest correlation was not the overall score but rather the presence of prior FTAs within the past two years.

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

Table 1: PSA Input Factors For Each Outcome Risk Score

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

confidence interval of 0.14 to 0.24, which is quite large. Ultimately, the large range in point estimates and confidence intervals for differences in failure rates for FTA do not provide evidence for uniform validity.

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	
<p>*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.</p>			

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.¹⁹ We use a common measure of correlation, a Pearson's 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 Pearson's r coefficient for each relevant PSA Risk Assessment input across the three outcome events: N(V)CA/FTA.

¹⁹ 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 Factor Input Correlations

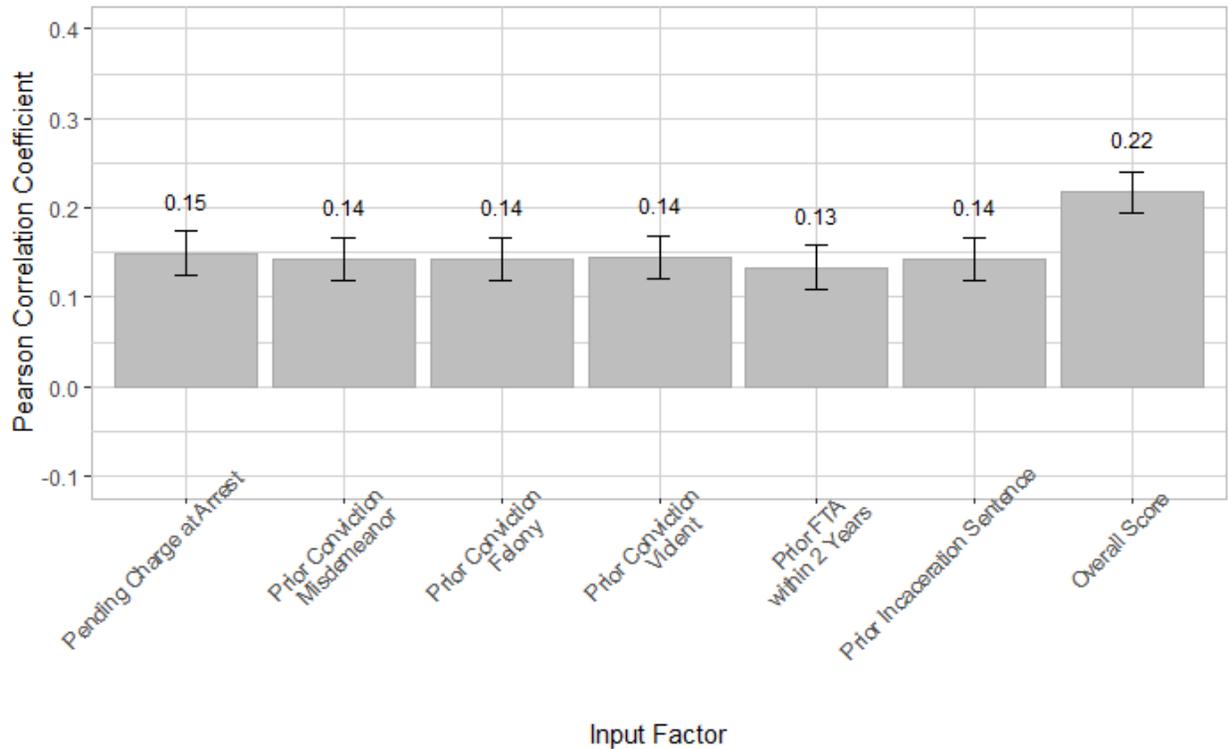


Figure 7 shows the Pearson r Correlation Coefficient and associated 95% confidence interval for each of the factors used in the calculation of the PSA NCA score, as well as the correlation of the overall score, with observed NCA events. The figure indicates that each input factor and the overall risk score is significantly correlated with observed NCA events in the appropriate direction. The overall risk score achieves a larger correlation coefficient than any individual factor coefficient, suggesting that input factors provide some non-overlapping predictive information that is preserved by the assessment's calculation methods. This figure provides evidence for the overall validity of the PSA with respect to NCA outcomes.

Figure 8: NVCA Factor Input Correlations

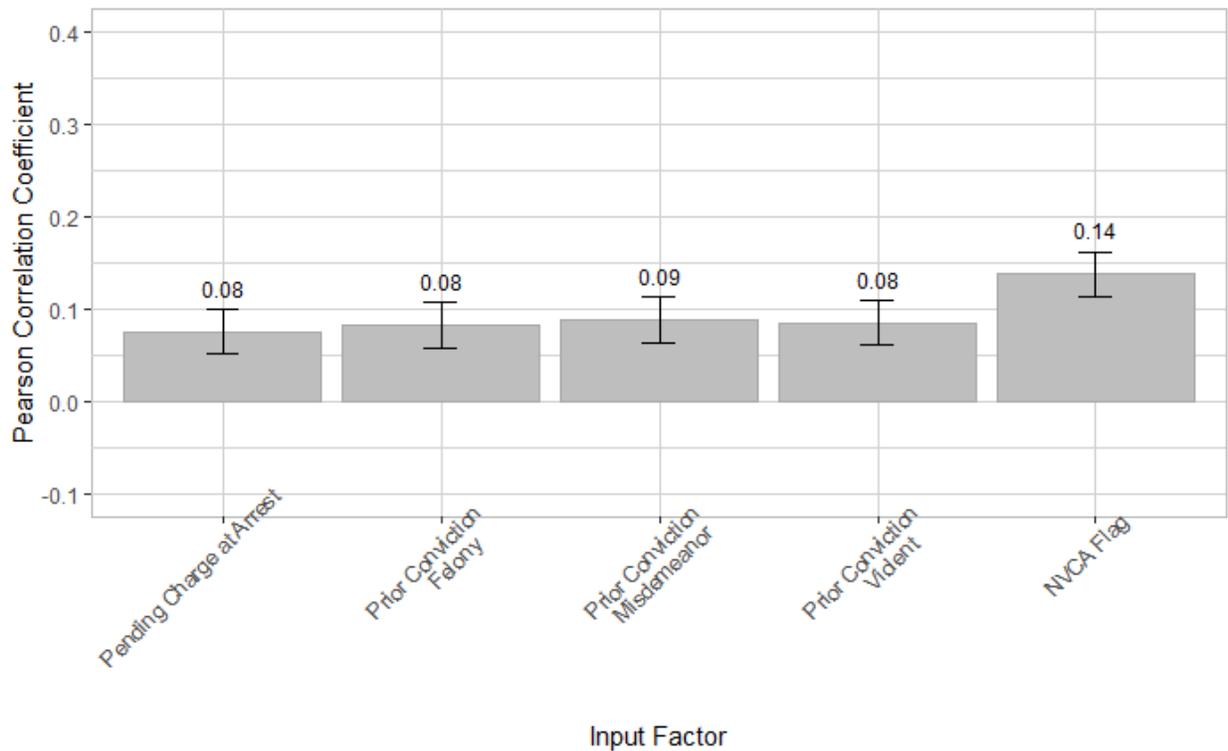


Figure 8 shows the Pearson r Correlation Coefficient and associated 95% confidence interval for each of the factors used in the calculation of the PSA NVCA risk flag, as well as the correlation of the presence of the risk flag, with observed NVCA events. The lack of overlap between the confidence intervals and 0 indicates that each input factor and the overall risk score is significantly correlated with observed NVCA events in the appropriate direction. The overall risk score achieves a larger correlation coefficient than any individual factor coefficient, suggesting that input factors provide some non-overlapping predictive information that is preserved by the assessment's calculation methods for NVCA. Overall, this figure provides evidence for the overall validity of the PSA with respect to NVCA outcomes.

Figure 9: FTA Factor Input Correlations

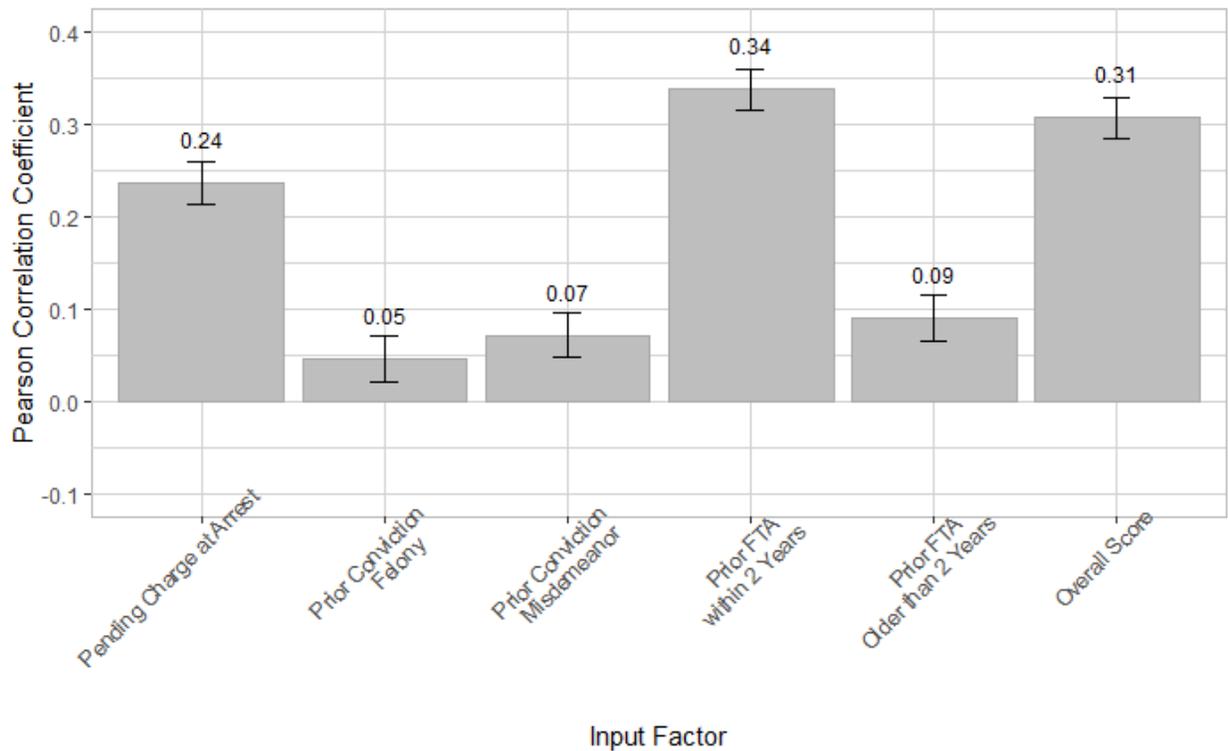


Figure 9 shows the Pearson r Correlation Coefficient and associated 95% confidence intervals for each of the five factors used in the calculation of the PSA FTA score, as well as the correlation of the overall score, with observed FTA events. Each input factor and the overall risk score is significantly correlated with observed FTA events in the appropriate direction. The overall risk score did not achieve the largest correlation coefficient; the correlation for Prior FTA within 2 Years was higher, suggesting that the risk score did not provide classifying information distinct from that of its inputs. Overall, this figure provides some, but weak, evidence for the overall validity of the PSA with respect to FTA outcomes, and some evidence to contract overall validity.

Figures 7-9 each show significant, positive correlations between the various factor inputs to the PSA scores and the relevant PSA outcome. Each of these correlations was in the same direction as the rule which translates them into the relevant risk score. For each PSA outcome event, all input correlation coefficients were significantly different from 0 at the $p < 0.001$ level. The significance levels of these findings were in line with the expectations of the overall validation framework. The magnitude of the correlation coefficients, all of which are below 0.35, would not generally be considered strong in most social science disciplines. For the NCA and NVCA scales, but not for the FTA scale, the overall score corresponded to the largest correlation coefficient, suggesting that the former two scores provided classifying information above that provided by the individual inputs. Overall the item-based correlation measures provide some, but not strong, evidence of overall validation.

iii. Area under the curve

This subsection discusses the results of the area under the curve (“AUC”) analysis. The AUC analysis shows mostly moderate evidence of the overall validity of the PSA and mixed evidence for equitable validity. Key details are as follows:

- NCA and FTA outcomes show moderate AUC scores indicating some gain in classifying power from the risk score, while the NVCA flag has an AUC score that indicates weak gain 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 individual counterparts.
- There are no significant differences in AUC scores across gender demographic subgroups for any of the PSA outcome metrics.

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.²⁰ 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 McLean County data, that would mean classifying all individuals as low risk for NCA, NVCA, and FTA. More accurate and informative models should provide greater distance between the ROC curve and the hypothetical no-information 45 degree line. Standard practice in this area is to assess this gain in information by measuring the area under the ROC curve (“AUC”), which quantifies the difference between the predictive gain of the model under the ROC curve and the baseline performance of the no-information line. ROC curves do not have direct analogies to classification models with multiple categories, such as the PSA NCA and FTA scores. The AUC measurement does, however, generalize to such multi-category classification settings.

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

²⁰ 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.

measurement indistinguishable from 0.50. The following benchmarks are sometimes used: an AUC measurement less than 0.54 indicates no evidence of validity.²¹ An AUC measurement between 0.55 and 0.63 indicates some, but not strong, evidence of validity. AUC measurements between 0.64 and 0.70 indicate moderate evidence, and a measurement greater than 0.70 indicates strong evidence. To the extent that there is no significant difference in AUC measures across either racial or gender pairings, we conclude that the PSA provides equivalent gains in predictions for each group within the pairing. The figure below plots the AUC measures for each of the three outcome event constructions: NCA, NVCA, and FTA.

Figure 10: AUC Score by PSA Outcome

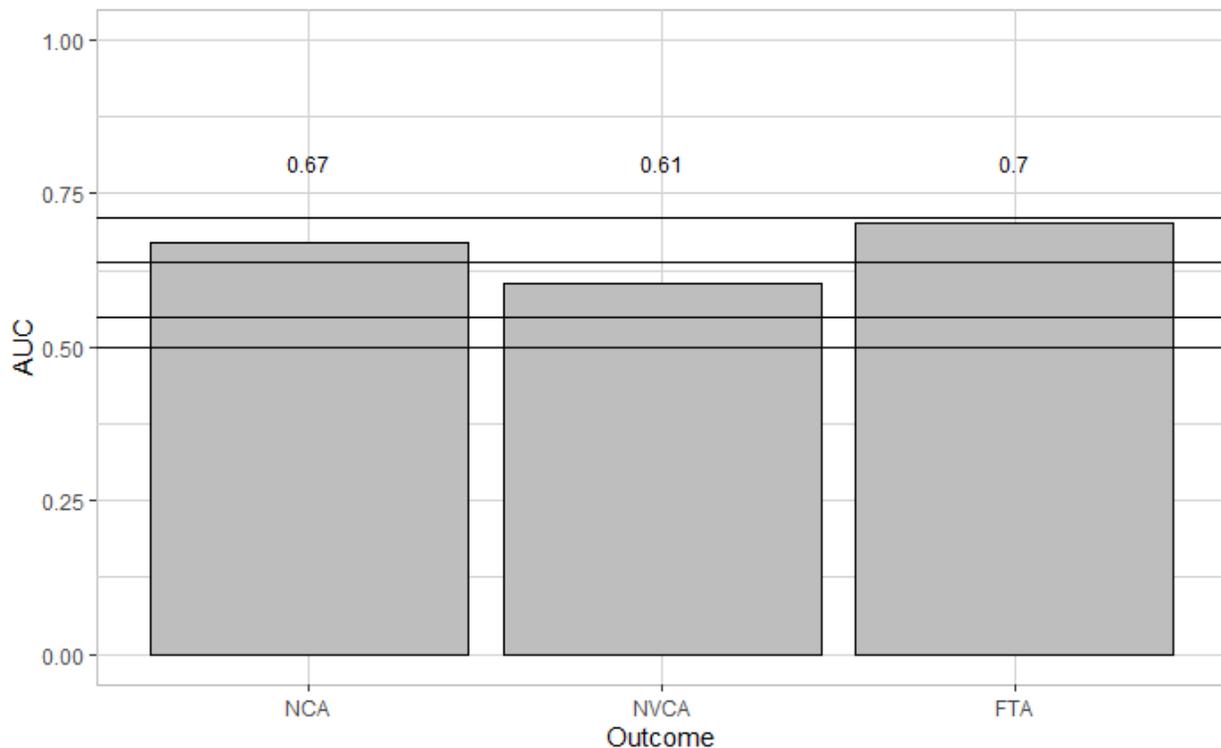


Figure 10 shows the area under the curve values for each outcome construction. AUC values range from 0 to 1, and in the case of a multi-outcome predictive assessment tool, like the PSA, are best understood as the probability that a randomly selected case with an observed failure for a relevant outcome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

The AUC metrics show positive gains above the random chance threshold of 0.50 for each of the outcome constructions under all three PSA outcome events. The AUC metric values for

²¹ 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).

NCA, NVCA, and FTA are .67, .61, and .70, respectively, indicating moderate, weak, and moderate evidence of overall validity.

Figure 11: AUC Score by PSA Outcome Across Demographic Groups

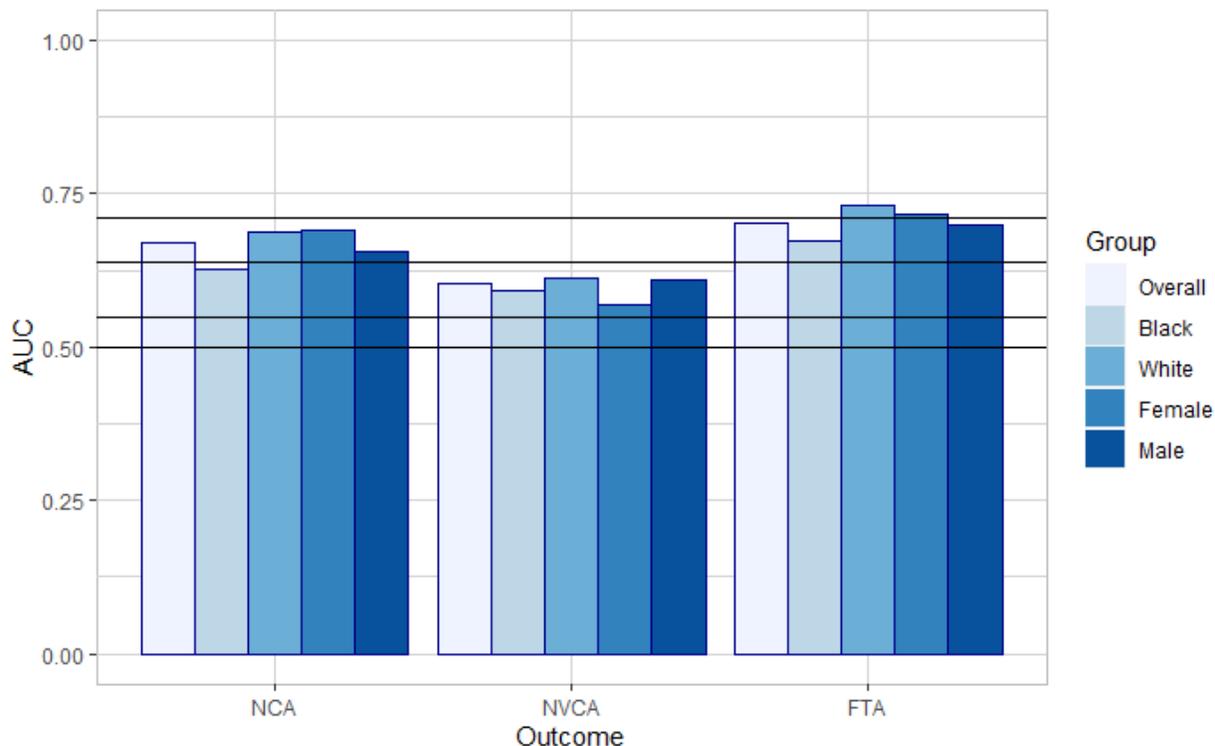


Figure 11 shows the area under the curve values for each outcome construction across the four main demographic groups of analysis as well as for the overall study population. AUC values range from 0 to 1, and in the case of a multi-outcome predictive assessment tool, like the PSA, are best understood as the probability that a randomly selected case with an observed failure for a relevant outcome will have a higher corresponding risk score than a randomly selected case with no observed failure for a relevant outcome. Across all three outcome metrics, there are no significant differences in AUC measurements across gender group pairings. Significant differences exist for racial group pairings for both NCA and FTA. For both of these outcomes, the relevant risk score provides significantly more predictive gains for cases with White individuals than for Black individuals. The divergence in findings for gender and race based comparisons provides mixed evidence for equitable validation.

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 11. We find little evidence of a difference in validity with respect to gender group pairings. Significant differences exist for NCA and FTA risk scores across racial group pairings. NCA and FTA risk scores provided meaningfully higher gains in predictive power for cases with White individuals than cases with Black individuals. Ultimately, the presence of significant differences for racial groups but lack of significant differences for gender groups provides mixed evidence for 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 weak evidence that the PSA is uniformly valid. Key details are as follows.

- PSA risk scores/flags have significantly positive coefficients, indicating that increases in risk scores are statistically significantly associated with increases in the probability of observing a relevant outcome.
- Moving from the minimum to the maximum risk score is associated with a 5x increase in the probability in observing an NCA and a 6x increase in observing an FTA.
- The presence of the NVCA Flag is associated with a 3x increase in observing an NVCA.
- Higher order interaction terms are statistically significant but substantively small, indicating marginal differences in the coefficient for risk scores at different levels of the relevant scale.

A logistic regression framework provides an off-the-shelf²² method for assessing the overall validity of the PSA.²³ The following figure plots predictive probabilities of observing at least one 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.

²² 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.

²³ For other validation studies that have utilized a regression framework, see:

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

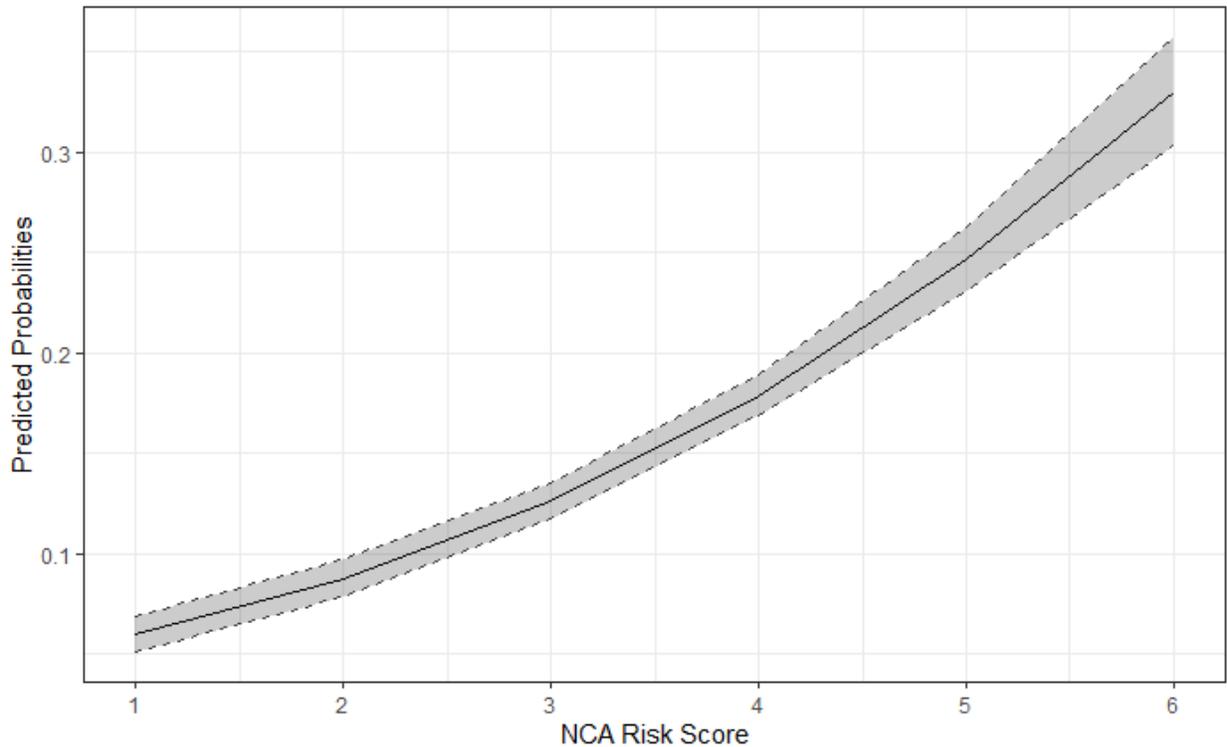


Figure 12 reports predicted probabilities and 95% confidence intervals for observing an NCA event obtained from the bivariate regression model with only the relevant PSA risk score scale as the regressor. The PSA NCA risk score has a significant, positive coefficient, indicating that higher NCA risk scores are significantly associated with a higher probability of an observed NCA failure. A one unit increase in the NCA risk score is associated with a 50% increase in the odds of observing an NCA failure versus not observing an NCA failure. This estimate exists on a confidence interval from a 44% increase in the odds ratio to a 58% increase in the odds ratio. This figure provides support for the overall validity of the PSA with respect to NCA outcomes. For uniform validity, the confidence intervals and predicted probability estimates increase at slightly steeper rates at the higher end of the NCA scale, providing weak evidence of uniform validity.

Figure 13: Bivariate Regression Predicted Probabilities for NVCA

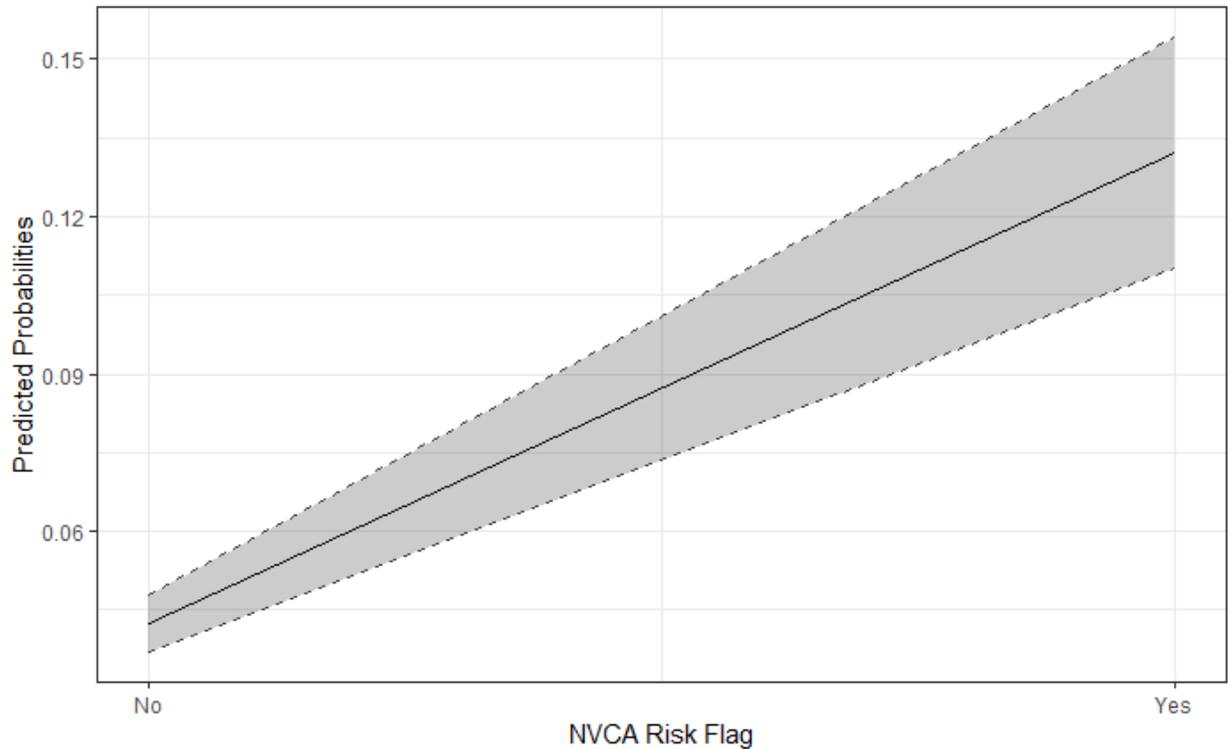


Figure 13 reports predicted probabilities and 95% confidence intervals for observing an NVCA event obtained from the bivariate regression model with only the relevant PSA risk score scale as the regressor. The PSA NVCA risk flag has a significant, positive coefficient, indicating that the presence of the NVCA Risk Flag is significantly associated with a higher probability of an observed NVCA failure. The presence of the NVCA Risk Flag is associated with a 243% increase in the odds of observing an NVCA failure versus not observing an NVCA failure. This estimate exists on a confidence interval from a 171% increase in the odds ratio to a 332% increase in the odds ratio. Thus, this figure provides support for the overall validity of the PSA with respect to NVCA outcomes.

Figure 14: Bivariate Regression Predicted Probabilities for FTA

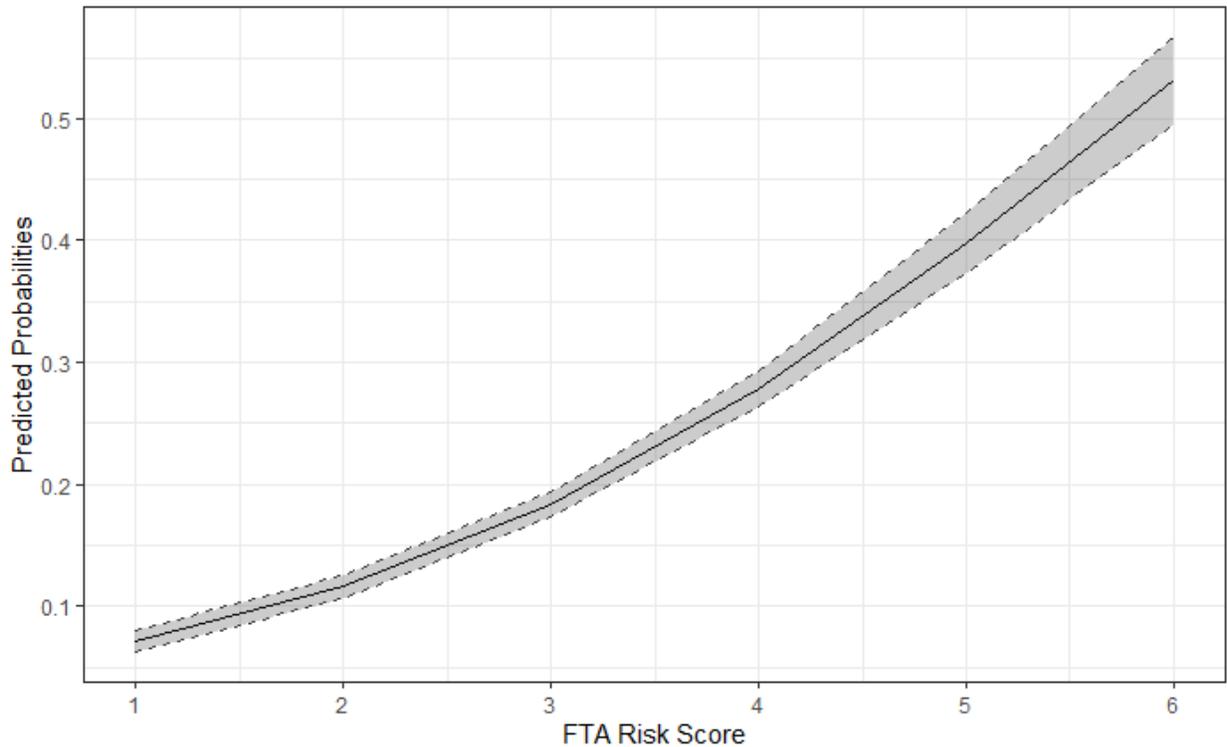


Figure 14 reports predicted probabilities and 95% confidence intervals for observing an FTA event obtained from the bivariate regression model with only the relevant PSA risk score scale as the regressor. The PSA FTA risk score has a significant, positive coefficient, indicating that higher FTA risk scores are significantly associated with a higher probability of an observed FTA failure. A one unit increase in the FTA risk score is associated with a 71% increase in the odds of observing an FTA failure versus not observing an FTA failure for the Base FTA outcome construction. This estimate exists on a confidence interval from a 64% increase in the odds ratio to an 80% increase in the odds ratio. This figure provides support for the overall validity of the PSA with respect to FTA outcomes. For uniform validity, the confidence intervals and predicted probability estimates themselves increase at slightly shaper rates at the higher end of the FTA scale, providing weak evidence of uniform validity.

Figures 12-14 show that the predicted probabilities across each relevant risk score level significantly increase along the risk score scale. For the NCA model, these ranged from a minimum of a 6% predicted chance of observing an NCA outcome at an NCA risk score of 1 to a maximum of 33% 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 8% at an FTA score of 1 and a maximum predicted probability of 53% at an FTA score of 6. For the NVCA model, having the NVCA flag present resulted in a predicted probability of 13% while not having the flag present is associated with a predicted probability of 4%. The standard error regions around these probability estimates indicate that the differences between the predicted probabilities was 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 was 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 1.50 with a 95% confidence interval of (1.44, 1.58), indicating that a one unit increase in NCA risk score was associated with a 50% 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 3.43 with a 95% confidence interval of (2.71, 4.32), indicating that the presence of the NVCA Flag was associated with a 243% 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 1.71 with a 95% confidence interval of (1.64, 1.80), indicating that a one unit increase in FTA risk score is associated with a 71% increase in the odds ratio of observing at least one case-specific FTA resulting in the issuance of a bench warrant. The significance and magnitude of the exponentiated coefficient estimates provides strong evidence for the overall validity of the PSA.

Evaluating uniform validity with a logistic regression is also possible through the inclusion of a higher order 'self-interaction' term. This term consists of interacting the risk assessment scale score with itself (squaring it), which allows the model to estimate a differential relation of the scale score on the outcome observations at higher levels of the scale score. The significance of the higher order term will indicate whether the association between the risk score and the relevant observation changes with different scores, indicating that the PSA risk score implies different increases of risk at different points of the score scale. In the NCA and FTA outcome models, the higher order coefficient (estimated from a logistic regression model including the risk score and the higher order risk score as IVs and the relevant outcome as DV) was significant at the $p < 0.001$ level. The exponentiated higher order coefficients were 0.94 and 1.06, for the NCA and FTA models respectively. This indicates that higher levels of the NCA score scale were associated with smaller increases in the probability of observing an NCA event, and higher levels of the FTA score scale were associated with larger increases in the probability of observing an FTA event. In both cases, the change in the odds ratio of observing the event, represented by the exponentiated coefficients, of -6% and 6%, were fairly minor. However, given the significance, the logistic regression analysis provides weak evidence that the PSA does not uniformly validate.

ii. Balanced accuracy measures

This section reports the results of a balanced accuracy analysis. The balanced accuracy measures provided moderate evidence of overall validity of the PSA. They also provided some mixed, but overall weak, evidence that the PSA was not equitably valid. 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 2 and 3, which showed modest gains above 0.60 for both NCA and FTA.

- Meaningful racial differences existed in a minority of NCA hypothetical threshold cases and a majority of FTA hypothetical thresholds.
- There were no meaningful gender differences in any of the PSA outcome metrics.

Accuracy is a commonly used assessment technique in machine learning. Accuracy is based on a confusion matrix.²⁴ One constructs a confusion matrix by dividing each case/observation either into a positive/high risk category or into a negative/low risk category. One then classifies each observation in the positive/high risk category as “true” or “correct” if a failure (here, an FTA or N(V)CA) occurs, and “false” or “incorrect” if no failure occurs. Correspondingly, one classifies each negative/low risk observation as true/correct if no failure occurs, and false if a failure occurs. One calculates the so-called “Accuracy metric” by adding together the number of true positives and true negatives, then dividing by the total number of cases, thus yielding a proportion of ‘correct’ classifications.²⁵

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).²⁶ As noted above, while this condition is true for the NVCA Flag, it is not true for the FTA and NCA scores, which both have six risk categories and only two observed outcome categories. The second issue is that the PSA does not make a discrete classification, but instead attempts to classify the level of risk of an individual by an ordinal scale. To address these issues for FTA and NCA, we implement five separate thresholds, meaning risk scores of 1, 2, 3, 4, and 5, for which a score above the threshold represents a positive classification and a score at or below the threshold represents a negative classification. We then calculate each accuracy metric for each of the five hypothetical thresholds for FTA and NCA and the one hypothetical threshold for NVCA.

There is an additional challenge. Accuracy, when used as a diagnostic statistic, is most useful when there is a balance in observed outcome categories, *i.e.* the number of observed positive and negative outcome cases is roughly equal. This is due to the fact that standard practice is to compare Accuracy with a theoretical “no information rate,” which is calculated by taking the number of correct predictions a model would make by simply assigning all cases the most common category (which is the classification or “guess” one would make if one had no

²⁴ 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.

²⁵ 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.

²⁶ 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 McLean County data are not equal or uniform across outcomes. As previously discussed, across FTA, NCA, and NVCA, at least 80% of cases corresponded to no failure of any kind.

For this reason, we show below not the raw Accuracy metric but instead what researchers call the “Balanced Accuracy” statistic.²⁷²⁸ 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.²⁹ 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.

²⁷ 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.

²⁸ 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.

²⁹ 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 Metric for NCA Scores

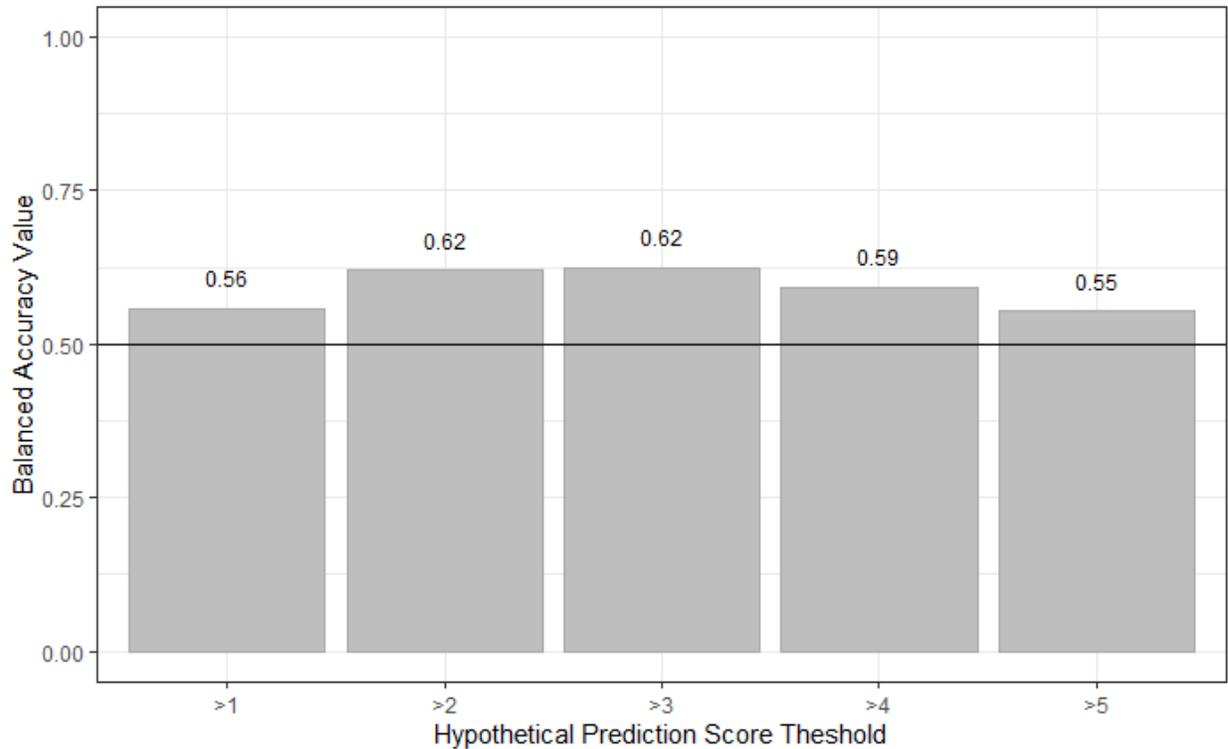


Figure 15 reports Balanced Accuracy measures for NCA outcomes using NCA score thresholds of 1, 2, 3, 4, and 5. The figure above shows that all 5 of the hypothetical prediction thresholds obtained balanced accuracy measures meaningfully higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, increased distinguishing power beyond classifying cases with no information beyond outcome distribution. Overall, this figure provides moderate evidence supporting the validity of the PSA with respect to NCA outcomes. The 2 and 3 thresholds classified with greater Balanced Accuracy, with metric values above 0.6, suggesting a modest gain in classification information.

Figure 16: Balanced Accuracy Metric for NVCA Scores

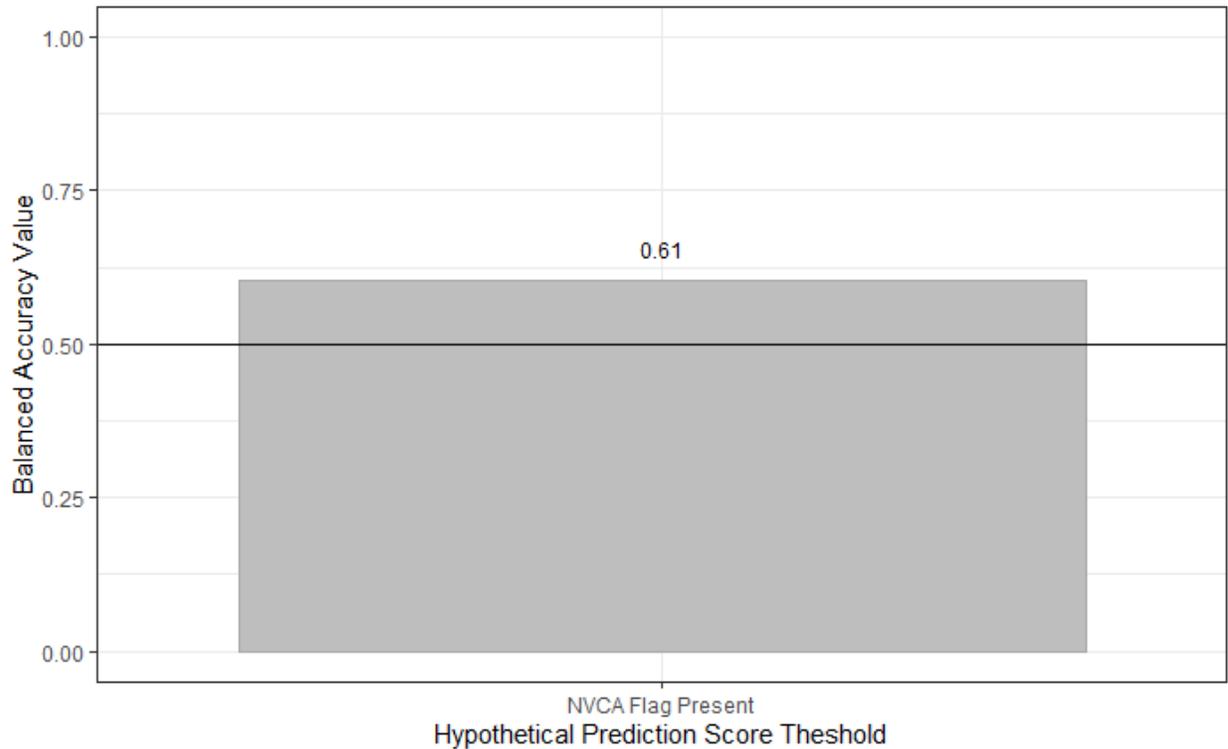


Figure 16 reports the Balanced Accuracy measure for NVCA outcomes. The figure above shows that under this hypothetical prediction rule, the PSA NVCA Risk Flag obtained a balanced accuracy measure slightly higher than 0.6, indicating that the PSA provided a modest increase in distinguishing power beyond classifying cases on limited information. Overall, this figure provided moderate evidence supporting the validity of the PSA with respect to NVCA outcomes.

Figure 17: Balanced Accuracy Metric for FTA Scores

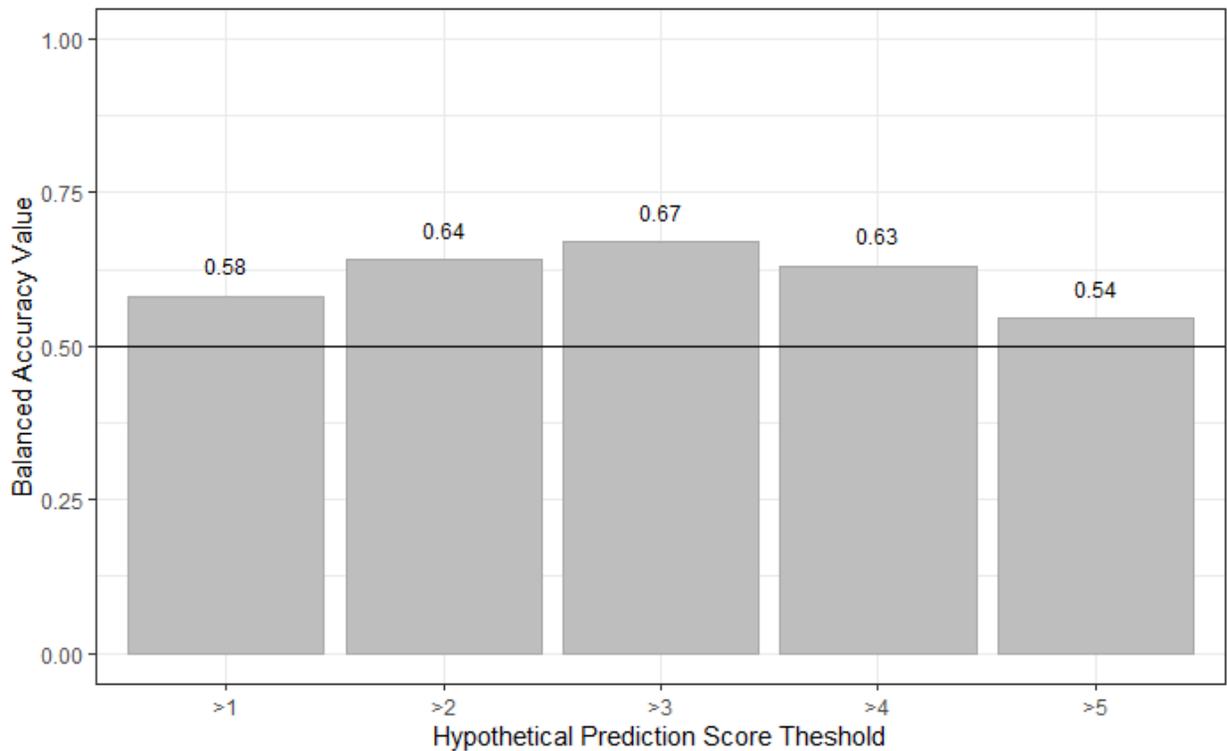


Figure 17 reports Balanced Accuracy measures for the FTA outcome construction for FTA score thresholds of 1, 2, 3, 4, and 5. The figure above shows that all 5 of the hypothetical prediction thresholds obtained balanced accuracy measures meaningfully higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, provided modest increases in classifying power. Overall, this figure provided moderate 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 0.625 with a minimum of 0.554 under the NCA Score > 5 threshold. FTA achieved its maximum Balanced Accuracy metric under the FTA Score > 3 threshold of 0.67 with a minimum of 0.545 under the FTA Score > 5 threshold. The NVCA outcome Balanced Accuracy metric was 0.605.

Figure 18: Balanced Accuracy Metric for NCA Scores by Demographic Group

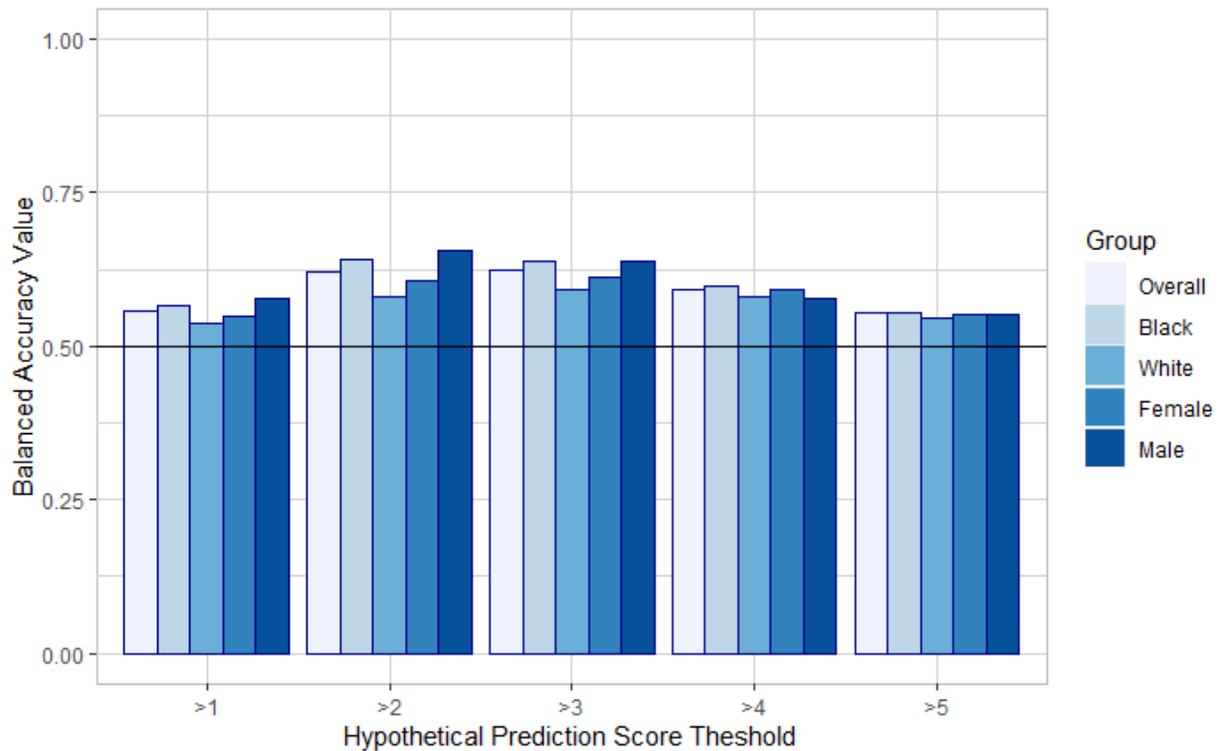


Figure 18 reports Balanced Accuracy measures for NCA outcomes. The figure above showed that all 5 of the hypothetical prediction thresholds obtained balanced accuracy measures higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, provided increases in predictive power beyond randomly classifying cases. These findings were consistent for both the overall study population as well as for each of the main study demographic groups. Meaningful gender and racial differences existed for two of the five hypothetical threshold rules at >2 and >3. Overall, this figure provides evidence supporting the validity of the PSA with respect to NCA outcomes and some, but weak, indication of meaningful differences in predictive power across race and gender groups.

Figure 19: Balanced Accuracy Metric for NVCA Scores by Demographic Group

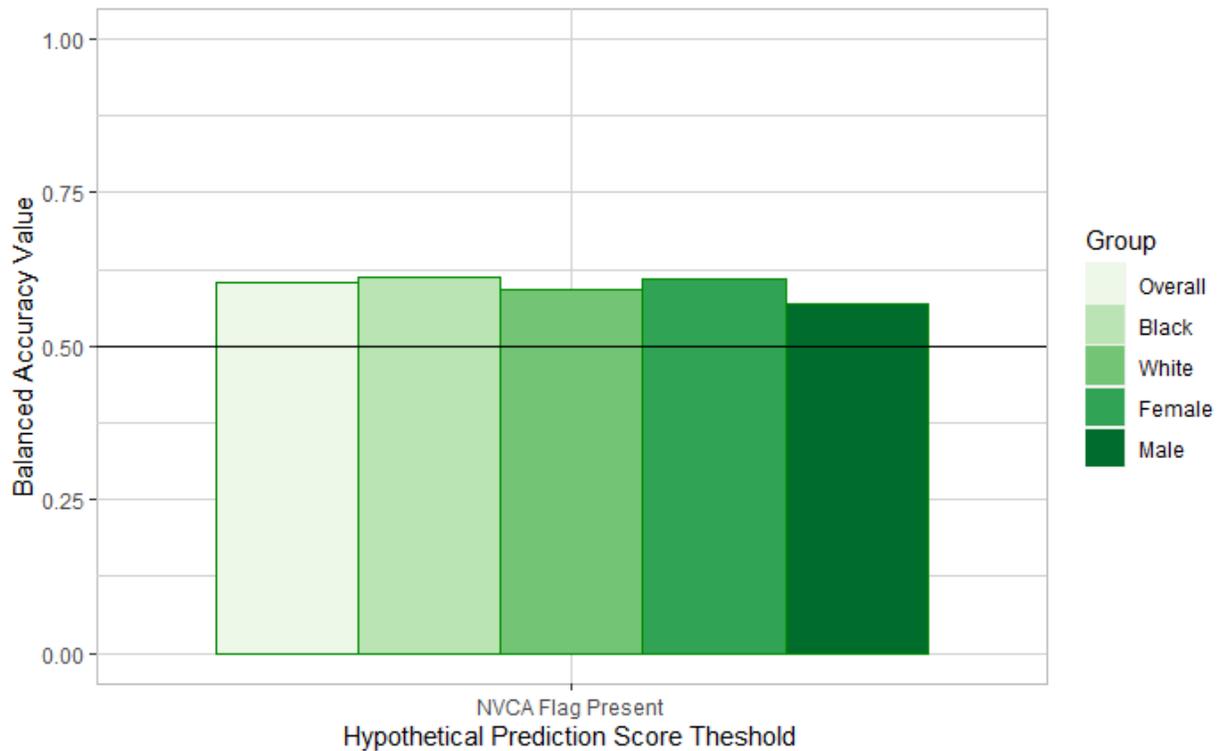


Figure 19 reports Balanced Accuracy measures for NVCA outcomes. The figure above shows modest distinguishing power gains under the hypothetical prediction threshold of classifying cases on the basis of the presence of the NVCA Flag. These findings are consistent for both the overall study population as well as for each of the main study demographic groups. Overall, this figure provides evidence supporting the validity of the PSA with respect to NVCA outcomes and no indication of meaningful differences in predictive power across race or gender groups.

Figure 20: Balanced Accuracy Metric for FTA Scores by Demographic Group

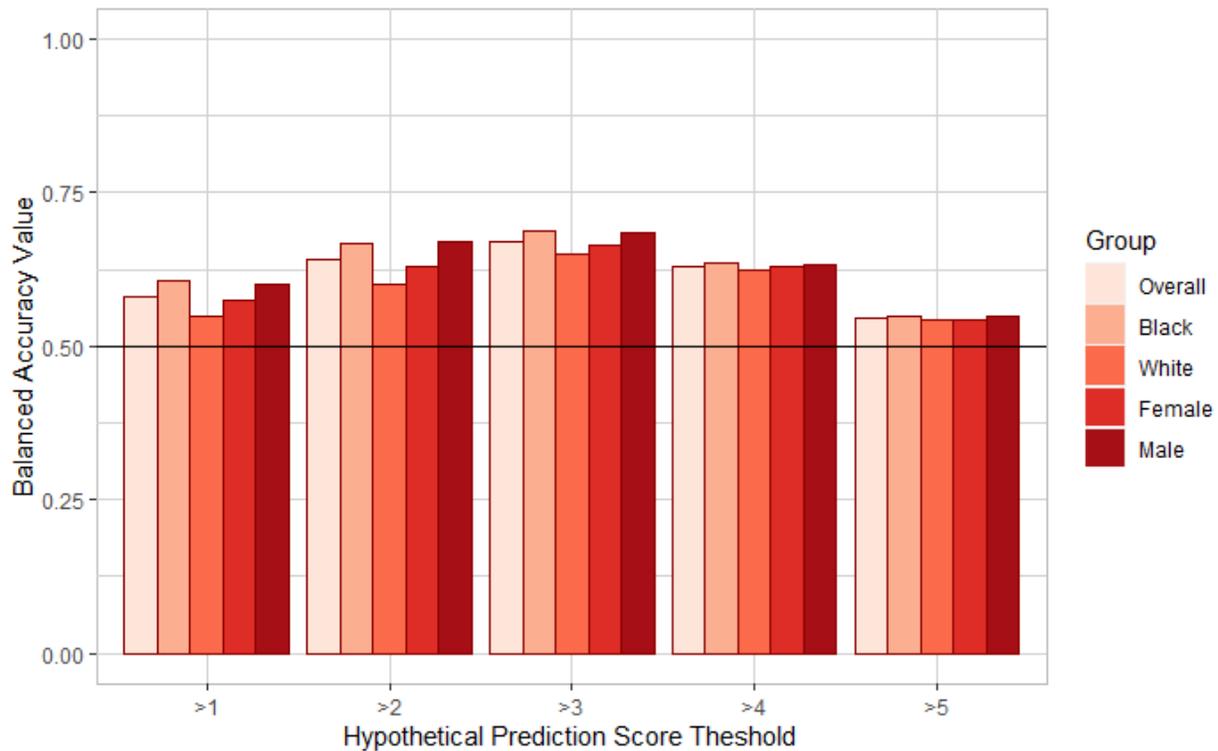


Figure 20 reports Balanced Accuracy Measures for FTA outcomes. The figure above shows that all 5 of the hypothetical prediction thresholds obtained balanced accuracy measures higher than 0.5, indicating that the PSA, under these hypothetical prediction rules, provided increases in predictive power beyond randomly classifying cases. These findings were consistent for both the overall study population as well as for each of the main study demographic groups. Meaningful gender and racial differences existed for three of the five hypothetical threshold rules at >1, >2, and >3. Overall, this figure provides evidence supporting the validity of the PSA with respect to FTA outcomes and some, but weak, indication of meaningful differences in classification power across race or gender groups.

The Balanced Accuracy metric can additionally be used to evaluate the PSA under the equitable validity framework in much the same way as the Area Under the Curve analysis. By comparing paired subgroup population values for Balanced Accuracy, we can evaluate whether the PSA provides differential gains in predictive power for different subgroup populations. For NCA outcomes, the maximum difference in Balanced Accuracy across racial groups was 0.059; the relevant maximum differences for NVCA outcomes and FTA outcomes were 0.019 and 0.067, respectively. Comparing across gender groups, the maximum differences were 0.02, 0.018, and 0.029 for NCA, NVCA, and FTA outcomes, respectively. Each of these maxima represents modest differences in Balanced Accuracy metrics for racial pairings and minor differences for gender pairings. Overall, the Balanced Accuracy metric provided some, but weak evidence that the PSA does not obtain equitable validity.

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 were statistically significant differences by race and gender, but those differences were few in number and inconsistent in direction. The analysis provided some, but weak evidence that the PSA was equitably valid. Key details were as follows.

- Significant differences existed in failure rates across racial demographic groups for FTA risk scores of 1 and 2 and for the no-NVCA-flag classification, but for no other comparison.
- Significant differences existed in failure rates across gender demographic groups for NCA risk scores of 2 and 5, for the no-NVCA-flag condition, and for one of the FTA risk scores. Female individuals demonstrated lower rates of failure for N(V)CA, but higher rates for FTA. Thus, the majority of hypothetical threshold cases indicate no gender differences and what differences do exist are inconsistent in direction.
- 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.

We calculated differences in failure rates for each PSA score category across study demographic groups. Statistically significant differences in classification failure rates across demographic groups would indicate that the same risk score relay different information depending on the demographic of the individual. We again used differences of proportion tests to analyze the 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.³⁰ The following figures plot outcome failure rates by relevant risk score across study demographic groups for each of the main outcome events.

³⁰ 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.

Figure 21: NCA Failure Rates by Demographic Group

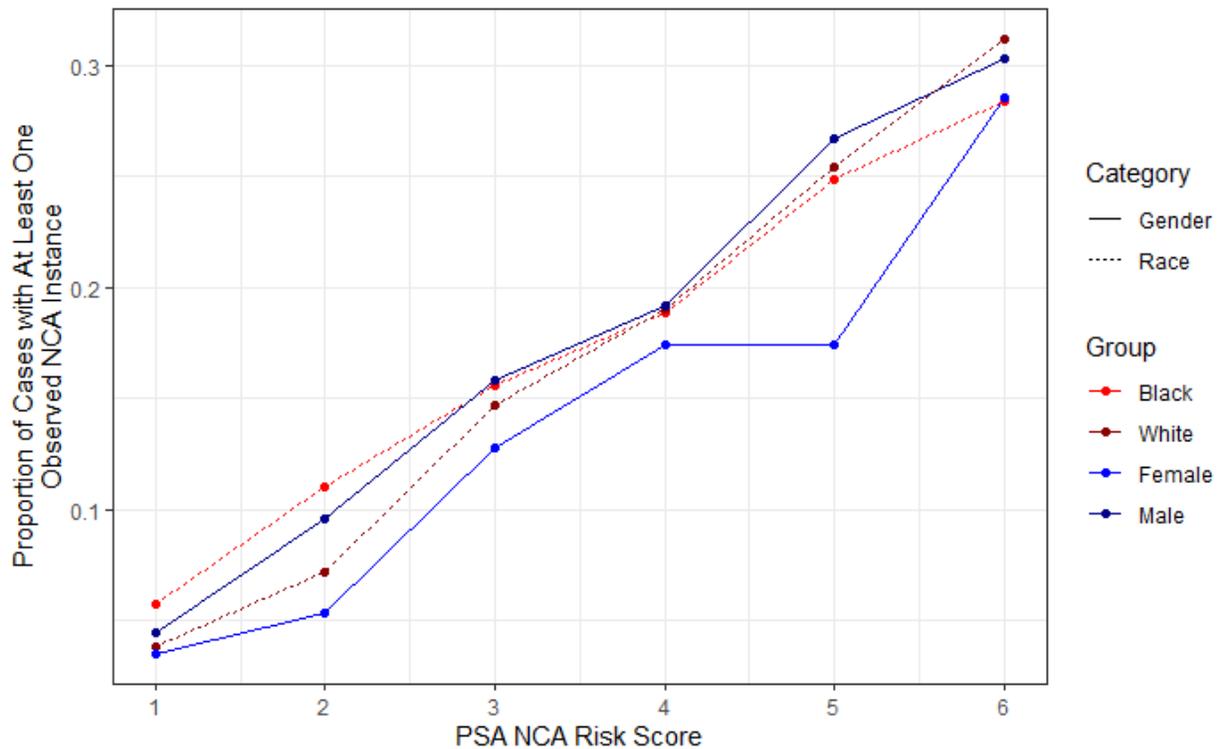


Figure 21 reports the observed failure rate for NCA outcomes across PSA NCA Risk Score categories by the four main study demographic groups. Line types and colors differ across category comparisons (Race and Gender). For NCA outcomes, there existed no significant differences in observed failure rates between Black individuals and White individuals for any level of the NCA risk score (as well as overall observed failure rates). With respect to gender comparisons, significant differences existed in observed failure rates between Male and Female individuals only for NCA risk score categories of 2 and 5 (as well as overall observed failure rates). These differences ranged from 4.3 to 9.3 percentage points with Female individuals observing lower failure rates than their Male individual counterparts in each instance. This figure provides support both for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates), as well as for significant differences in observed failure rates for gender, but not race, subgroups with respect to NCA events.

Figure 22: NVCA Failure Rates by Demographic Group

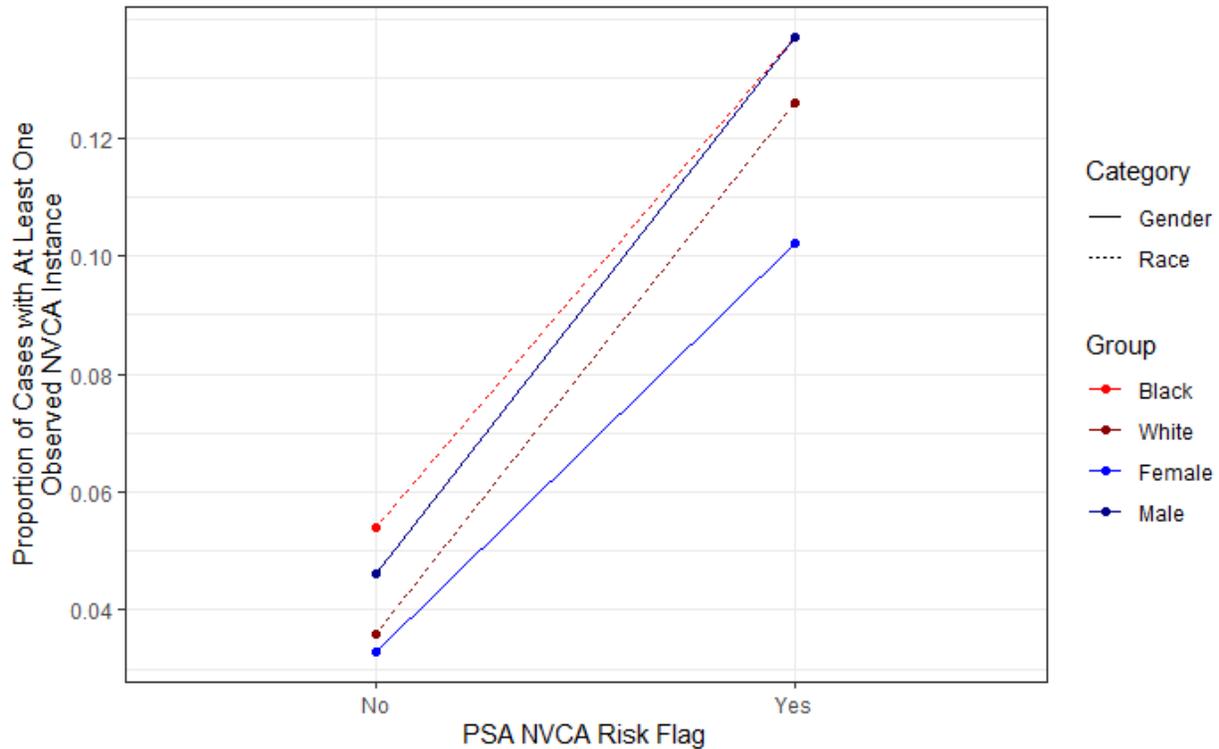


Figure 22 reports the observed failure rate for NVCA outcomes across categories of PSA NVCA Risk Flag presence by the four main study demographic groups. Line types and colors differ across category comparisons (Race and Gender). For NVCA outcomes, significant differences existed in observed failure 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, significant differences existed in observed failure rates between Male and Female individuals when the NVCA risk flag was not present. This difference was 1.3 percentage points with Female individuals experiencing lower failure rates than their Male individual counterparts. This figure provides support for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates), as well as some, but weak, evidence for significant differences in observed failure rates for both racial and gender subgroups with respect to NVCA events when the NVCA flag is not present.

Figure 23: FTA Failure Rates by Demographic Group

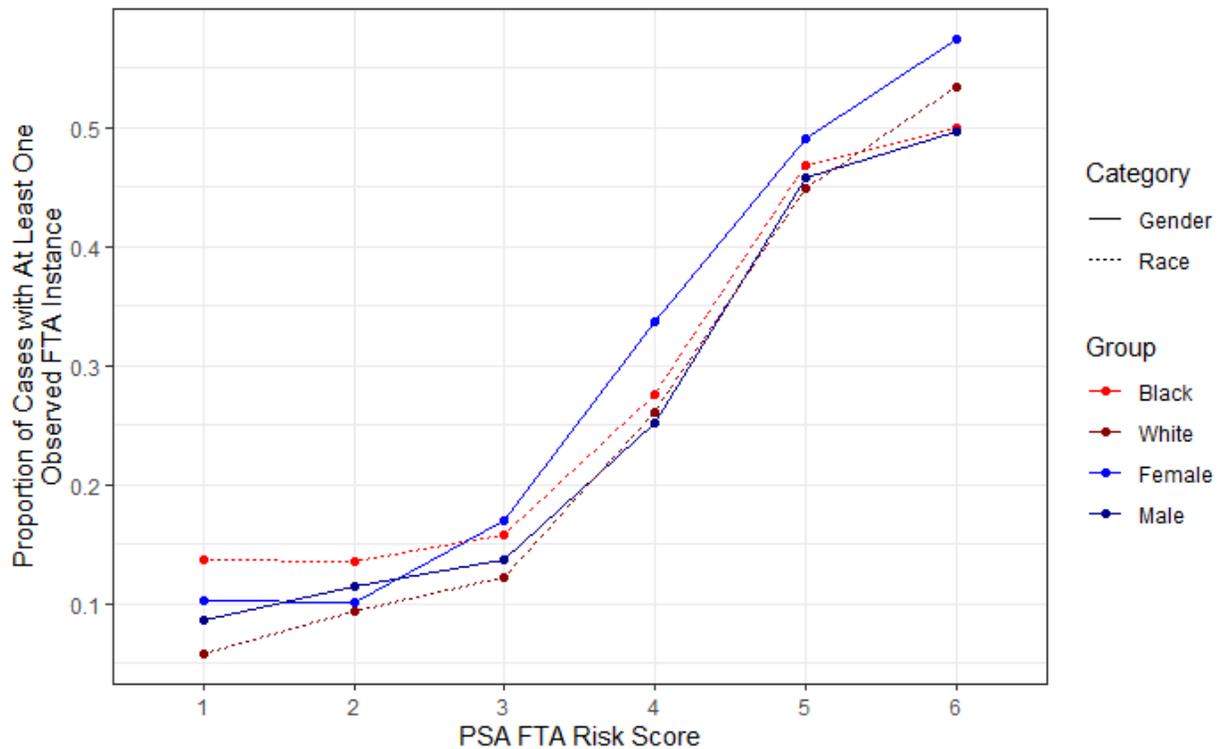


Figure 23 reports the observed failure rate for FTA outcomes across PSA FTA Risk Score categories by the four main study demographic groups. Line types and colors differ across category comparisons (Race and Gender). Significant differences existed in observed failure rates between Black individuals and White individuals for FTA scores of 1 and 2. At these levels of FTA risk score, Black individuals had observed failure rates 5.2-7.2 percentage points higher than their White individual counterparts. Significant differences existed in observed failure rates between Male and Female individuals only for the FTA risk score category of 4. This difference is about 10.3 percentage points, with Female individuals observing higher failure rates than their Male individual counterparts. This figure provides support for the overall validity of the PSA (higher risk scores are associated with higher observed failure rates), as well as some, but weak, evidence for significant differences in observed failure rates for racial and gender subgroups.

Figures 21-23 provide the failure rates by PSA Risk Score categories for demographic subgroups, allowing for an evaluation of the equitable validity of the PSA. The figures indicate that while failure rates tended to move similarly across demographic subgroups, there was some meaningful separation. For NCA outcomes, no significant ($p < 0.05$) racial group differences existed in failure rates at any NCA score, while significant gender group differences in failure rates existed at NCA scores of 2 and 5. For NVCA outcomes, significant racial group differences in failure rates existed for cases with no violence flag present for gender and racial subgroups. 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 4. Perhaps most importantly, with regards to gender, there was no consistent pattern with respect to difference in classifying information. Female failure rates were lower than corresponding male rates with respect to NCA and NVCA, but higher with respect to

FTA. Ultimately, the subgroup paired comparison of PSA score specific failure rates provided no meaningful evidence that the PSA does not equitably validate. Significant differences existed, but they are few in number and inconsistent in direction.

ii. Moderated regression

This section provides the results of a moderated regression analysis to assess equitable validity. This analysis showed some statistically significant differences across racial groups, but the size of those differences was small. Therefore, this analysis provides weak evidence 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.
- Interacting race and risk score, which would indicate whether the predictive meaning of the risk score changes significantly across racial groups, was significant only for FTA scores. The FTA score - race interaction was substantively small, with a point estimate of about a 12% decrease in the odds of a Black individual observing an FTA relative to a White individual.

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.³¹ 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.³² We focus on race, as opposed to gender, due to the fact that the racial distinctions have traditionally garnered heightened concern in the literature and from various

³¹ 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).

³² More specifically, the moderated regression framework proceeds in four steps: the first model regresses only the hypothesized moderating variable on the outcome; the second model regresses only the risk score variable on the outcome; the third model regresses both the hypothesized moderating variable as well as the risk score variable on the outcome variable; and the fourth model regresses the hypothesized moderating variable, the risk score variable, and an interaction between the two on the outcome variable. By evaluating the risk assessment score coefficient across these separate models, we can determine the impact of including a potentially moderating variable, such as race, on how the assessment score relates to the relevant outcome. Evaluating the basic value of the risk score can be done by 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.

stakeholders. 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 for NCA

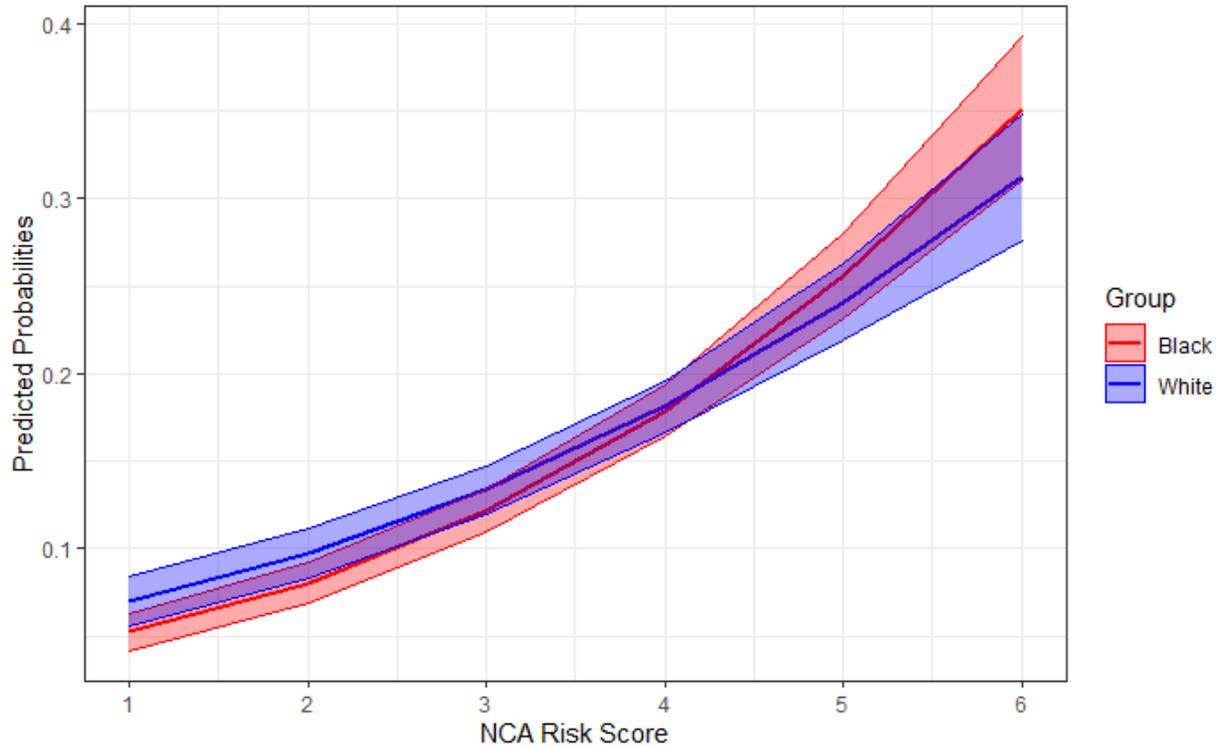


Figure 24 reports predicted probabilities and associated 95% confidence intervals for observing an NCA event obtained from the moderated regression model with both PSA score and race variables. The PSA NCA risk score had a significant, positive coefficient, indicating that higher NCA risk scores were 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 58% increase in the odds of observing an NCA failure versus not observing an NCA failure. This estimate corresponded to a confidence interval from a 47% increase in the odds ratio to a 69% increase in the odds ratio. The interaction term was insignificant (0.91 odds ratio on a 95% CI of 0.83 - 1.00). Functionally, this meant that when taking into account racial categories, Black individuals did not have significantly different predicted probabilities for observing an NCA than their White peers when looking only at NCA scores, which themselves were overall statistically significantly predictive of observed NCA events. The shaded confidence interval regions illustrate the lack of moderating impact of race. At all levels of the NCA Risk Score Scale, there is overlap in predictive probabilities for Black and White individuals, suggesting that the scores were not meaningfully different between the groups. Overall, this figure provides support for both the overall validity and equitable validity of the PSA.

Figure 25: Moderated Regression Predicted Probabilities for NVCA

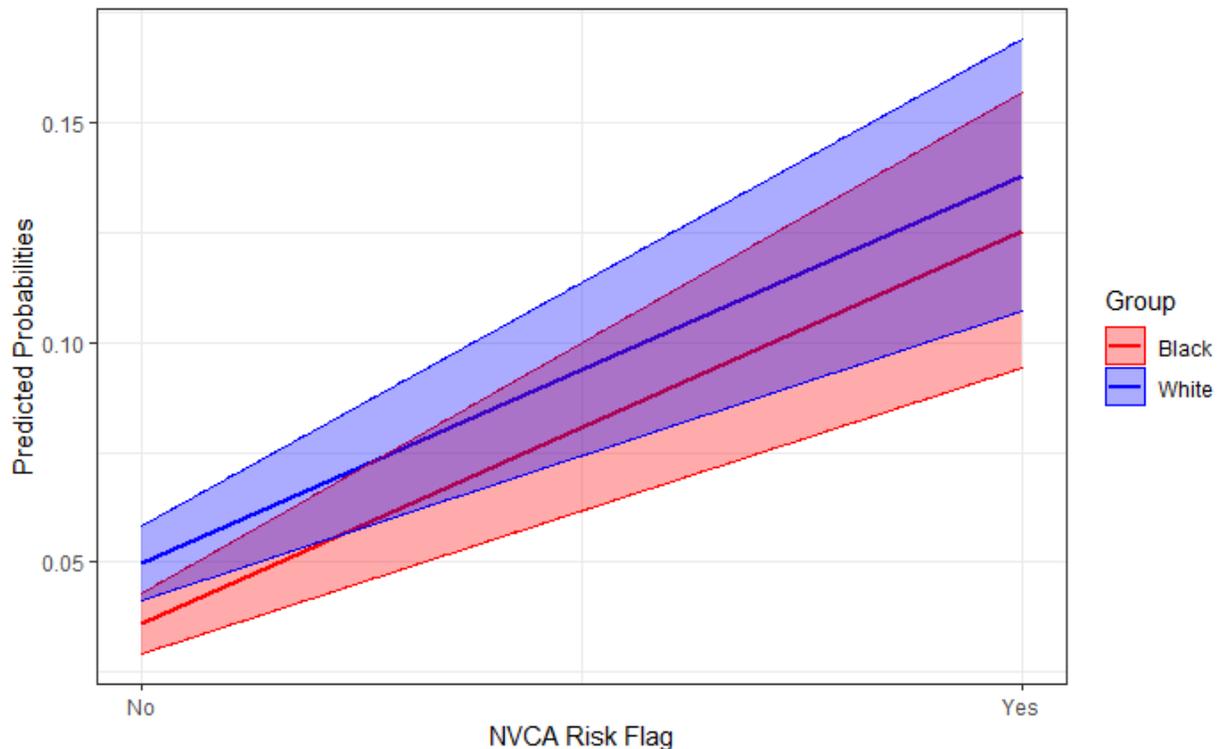


Figure 25 reports predicted probabilities and associated 95% confidence intervals for observing an NVCA event obtained from the moderated regression model with both PSA score and race variables. The PSA NVCA risk flag had a significant, positive coefficient, indicating that the presence of a risk flag was significantly, independently associated with a higher probability of an observed NVCA failure. The presence of an NVCA risk flag was associated with a 283% increase in the odds of observing an NVCA failure versus not observing an NVCA failure. This estimate corresponded to a confidence interval from a 170% increase in the odds ratio to a 439% increase in the odds ratio. The interaction term was not significant. Functionally, this meant that when taking into account racial categories, Black and White individuals had statistically similar predicted probabilities for observing an NVCA. The shaded confidence regions indicated no meaningful difference in predicted probabilities for both values of the NVCA Flag. Overall, this figure provides support for both the overall validity of the PSA as well as the equitable validity for the NVCA outcome.

Figure 26: Moderated Regression Predicted Probabilities for FTA

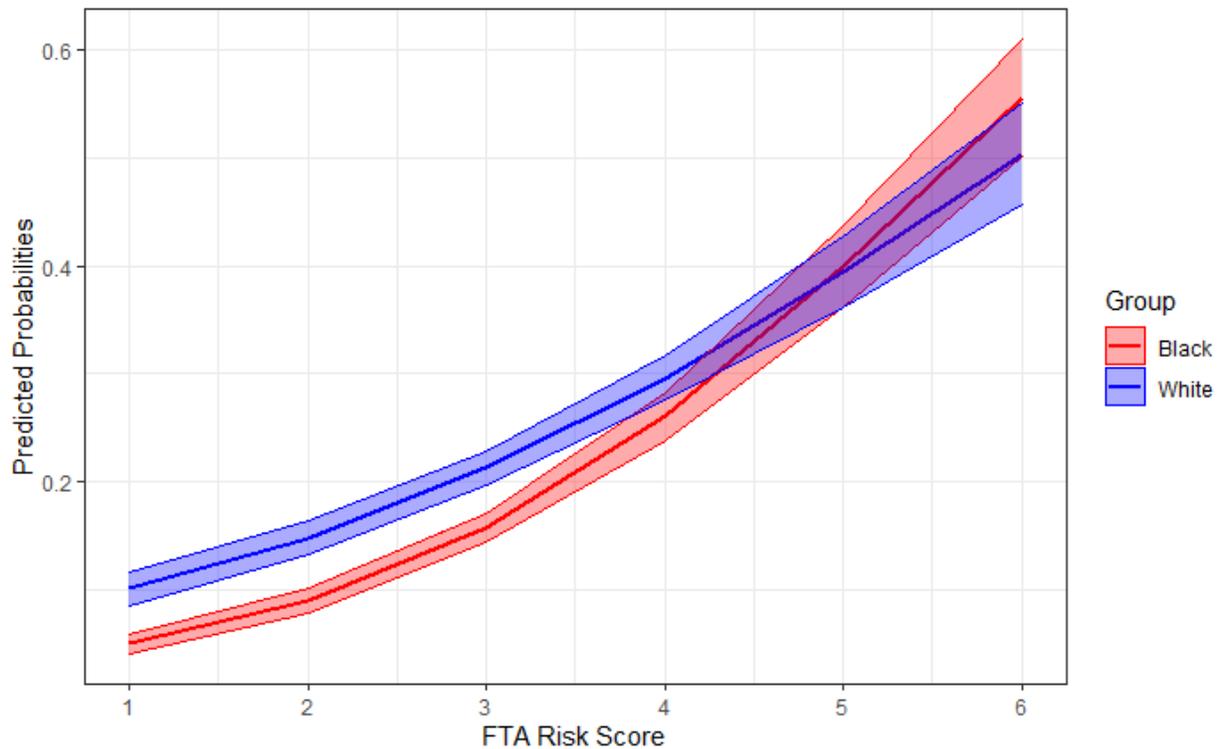


Figure 26 reports predicted probabilities and associated 95% confidence intervals for observing an FTA event obtained from the moderated regression model with both PSA score and race variables. The PSA FTA risk score had a significant (at the $p < 0.001$ level), positive coefficient, indicating that 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 an 88% increase in the odds of observing an FTA failure versus not observing an FTA failure. This estimate corresponded to a confidence interval from a 76% increase in the odds ratio to a 102% increase in the odds ratio. The interaction of PSA and race was statistically significant, suggesting that there was some evidence of racial subgroup differences with respect to the FTA outcome. The coefficient on the interaction term was 0.82, indicating that each level of the FTA score indicated an 18% lower odds ratio of observing an FTA for Black individuals compared to their White counterparts. 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. Overall, this figure provides support for the overall validity of the PSA with respect to FTA outcomes and provides 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 were obtained from the model of the moderated regression framework that included 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 indicated 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 was 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 were 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 was 1.58 with a 95% confidence interval of (1.47, 1.69), while the bivariate estimate was 1.50. For the FTA model, the exponentiated coefficient on the FTA Score Scale was 1.88 with a 95% confidence interval of (1.76, 2.02), while the bivariate estimate was 1.71 with a 95% confidence interval of (1.64-1.80). For NVCA, the magnitude of the difference in the exponentiated coefficients between the moderated regression and the bivariate logistic regression appeared larger, but the confidence intervals overlap. The exponentiated coefficient estimate for the presence of the NVCA Flag is 3.83 with a confidence interval of (2.70, 5.39), while the bivariate exponentiated estimate was 3.43. The moderated regression framework ultimately provided strong evidence of overall validity for the PSA.

The primary benefit of the moderated regression framework for the purposes of this study was its capacity to provide insight as to whether the PSA equitably validated. To the extent that the interaction term was significant, this indicated that information provided by the relevant PSA risk scale score statistically changed when moving from individuals of one racial group to another. For FTA, the interaction term was significant at the $p < 0.001$ level with an exponentiated coefficient estimate of 0.82 on a confidence interval of (0.75, 0.91). This indicated that for each level of the FTA Score Scale the odds ratio of observing at least one FTA event during the pretrial period was about 12 percent lower for Black individuals than the corresponding odds ratio for the same FTA score for White individuals. This twelve percent moderating effect of race was about 1/5th the size of the effect of a one unit increase in the FTA score. For both NVCA and NCA, the interaction term representing the moderating effect of race on the relevant PSA risk score scale was not statistically significant. Ultimately, when considering the small magnitude and inconsistency of results, the moderated regression framework provided evidence that the PSA equitably validated.