Juliet Bu
ECON 970
December 10, 2025
Final Paper

Evaluating Judicial Override Behavior and Economic Efficiency of Algorithmic Risk Assessments in Pre-Trial Detention Decisions

Introduction

The use of algorithmic decision making tools in the criminal justice system has been around since as early as the 1960s, starting as an actuarial tool for prediction based on static, individual factors linked to higher recidivism rates (Solow-Niederman et al. 710, 2019). Concerns about their use have accompanied their increasing popularity as these tools have evolved to incorporate dynamic factors and, recently, leverage machine learning technologies. In 2014, then U.S. Attorney General Eric Holder raised concerns with the use of algorithms in criminal justice system decisionmaking, stating "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice" (ProPublica, 2014). His concerns have only become more relevant in recent years. With huge improvements in artificial intelligence and an increasingly prolific culture around technology and efficiency in governance, it seems inevitable that the criminal justice system will continue moving towards the use of algorithms to assist in judicial decisionmaking around defendant outcomes. Already, as states like California and Pennsylvania move to abolish cash bail, they have turned to algorithms to assess defendants' risk of new offenses and failures to appear (CA SB 36, 2019; PA Code § 305.1, 2020). Algorithms like the Pretrial Safety Assessment (PSA), which is the focus of this paper, are often adopted under the veneer of science-based objectivity, despite a lack of transparency in the algorithms' calculation processes and scrutiny of how algorithms may perpetuate human biases from its data sources (Hamilton 164, 2021; Copp et al. 336, 2022). The role of algorithms in the criminal justice system begs the question of how judges utilize algorithmic risk assessment outputs in their pretrial detention decisions and what the implications of these behaviors are for accuracy and economic efficiency.

To answer these questions, my paper uses administrative court data from a randomized PSA provision experiment conducted in Dane County, Wisconsin, between 2017-2018. The original experiment was conducted by the Harvard Law School Access to Justice Lab, led by Renee Danswer and James Greiner. My analysis first focuses on judge override behavior, exploring whether judges systematically override for certain groups and how judges' release and override decisions change for different groups following the provision of the PSA output. To answer my first question, I run a logit regression to assess if there are any defendant demographic variables that are significantly correlated with a judge override in favor of the defendant—that is, when the algorithm recommends detention but the judge releases. I find that gender is the only variable significantly correlated with favorable judge override: judges are 11.2 percentage points more likely to favorably override the algorithm for female defendants compared to male defendants. To assess whether PSA provision systematically changes judges' release decisions for certain groups, I run several logit regressions on gender, race, and a gender-race interaction term, finding no significant changes in judge decisions following the provision of the PSA score. I do, however, find significant changes in judicial override behavior following the provision of the PSA score. Finally, to evaluate the accuracy of algorithm-assisted judicial decision making, I compare the incidence of failures to appear (FTA), new criminal activity (NCA), or new violent criminal activity (NVCA) for defendants for whom the judges released when they were provided the PSA output to the incidence of FTA/NCA/NVCA for defendants for whom the judges released without the PSA output. I find a weak, non-statistically significant increase in the accuracy of judicial decisions once the PSA is provided.

In this paper, I will begin by defining my research scope and the limitations of my research, followed by an overview of the current literature in the field of algorithm-assisted judicial decision making. Then, I will overview my research methods and analysis, present my findings, and conclude with the policy implications of my research.

Scope and Limitations

While there have been analyses conducted on the Dane County dataset before by Imai et al. in 2021 and Ben-Michael et al. in 2025, my analysis focuses specifically on judicial override of the algorithm and what factors drive override. While Imai et al. have studied the question of how the PSA affects judicial decision making for different groups, I reassess these claims by using a logit regression instead of the authors' causal effect estimation methodology. Similarly, Ben-Michael et al. used the dataset to assess relative accuracy of human-alone, human and algorithm, and algorithm-alone decision-making, which I expand upon by further drawing conclusions about the false negative rates for judges using the algorithm compared to judges without. Notably, my findings diverge with Imai et al. and Ben-Michael et al., challenging the robustness of their findings. Additionally, I contribute a novel analysis of what factors may drive a judge using the algorithm to systematically override the algorithm in favor of releasing a defendant, and whether this propensity for defendant-favorable disagreement changes with the provision of the PSA score. Building upon prior literature on economic efficiency, I also use accuracy as a proxy to assess the economic efficiency of algorithm-assisted judicial decision making, whereas efficiency has previously been considered in the literature in relation to detention and crime rate reductions. I define relative accuracy by comparing the percentage of defendants released by the judge who had either an FTA/NCA/NVCA. While there are a number of ways to evaluate economic efficiency, including assessing decreases in detention rates or cost savings, my paper uses accuracy as a proxy for gains in economic efficiency from PSA adoption. From an economics perspective, one goal of the criminal justice system should be to reduce recidivism and therefore the cost of crime to society. If judges using algorithms are more accurate than without, this outcome would suggest that adopting algorithms helps to reduce the crime rate, improving economic efficiency.

There are several limitations of my research. Given the Dane County data is from 2017-2018, Dane County has since made several reforms to how the PSA is used by judges. While the PSA algorithm itself has not changed, since the randomized control trial (RCT), Dane County has shifted institutional attitudes towards how judges are using PSA output. In January 2024, Dane County received the results of the randomized control trial, acknowledging that the PSA had failed to achieve improvements in the criminal justice system that the County had hoped for (Dane County Community Justice Council 2, 2024). Since 2018, Dane County has entered into a partnership with Advancing Pretrial Policy and Research (APPR), who offered training to Dane County judges on responsible PSA use in April 2023. The APPR has also made several recommendations around presenting the PSA to judges that Dane County has signalled a willingness to adopt, including revising PSA output to not include cash bail amounts, and giving judges decision trees to standardize how the PSA is used in their pre-trial decision (Frank-Loron 10, 2023). These changes suggest that, at the very least, attitude shifts from criminal justice system administrators in Dane County could lead to pressures on judges to change their behavior following the 2017-2018 RCT findings. Changes in judicial behavior would undermine the generalizability of my findings. However, many jurisdictions in the United States use the PSA, many of which do not have the same reform-minded attitude as Dane County. Thus, even if my findings are no longer an accurate reflection of how Dane County judges interact with the PSA, my findings still hold some external validity in assessing algorithm-informed judicial decision making behaviors. Another limitation in my research lies in the data. Measurements of accuracy rely on whether a defendant has an FTA, NCA, or NVCA. FTA is an indication of whether the defendant went on to miss one of their future court dates. NCA and NVCA are indicators for whether the defendant, after release, was arrested for another criminal or violent criminal offense. One caveat with the NCA/NVCA is that both are based on arrests rather than convictions, meaning they are not entirely accurate measures of recidivism. It must be acknowledged that overpolicing of certain populations could overstate the perceived recidivism rates of individuals in my analysis, as my research does not have a way to distinguish between true crimes versus inaccurate arrests. Finally, it is important to note that my analysis focuses primarily on the PSA algorithm, which helps determine pretrial defendant release or detention; my analysis does not examine the use of other algorithms or the use of algorithms in other aspects of the criminal justice process, such as sentencing.

Literature Review

In comparing algorithm-informed versus human-alone judicial decisionmaking, the literature distinguishes the positive task of prediction from the normative task of decision, offering different assessments of whether accuracy in prediction can be equated to efficiency of a criminal justice system. Kleinberg et al.'s 2017 piece "Human Decisions and Machine Predictions" separates the task of prediction and decision, depicting how judges are both responsible for predicting a defendants' risk of new offenses and failures to appear and applying a normative standard for society's appetite for risk, releasing defendants whose perceived risk is below a certain socially-acceptable threshold. From an economics perspective, this threshold can be considered the point at which society achieves optimal deterrence of criminal activity. The authors construct a machine learning model to predict defendant risk, only considering information available to judges (e.g., current offense, prior criminal record), notably excluding race, ethnicity, and gender. In their study of case data from New York City between 2008-2013, Kleinberg et al. find that judges release many defendants that the algorithm identifies as high risk, and even stricter judges are not detaining those defendants who pose the greatest risk, suggesting that there can be reductions in detention that result in the same or a greater reduction in crime. They therefore find that at the same jailing rate as the New York judges, their predictive algorithm could reduce crime between 14.7%-24.7%, and at the same crime rate, their predictive algorithm could reduce jail rates between 18.5%-41.9% (Kleinberg et al 239-241, 2017). Kleinberg et al.'s work has been used to argue from an economics perspective the merits of algorithm-informed judicial decisionmaking in reducing crime rates and jail rates by giving judges a more accurate assessment of defendant risk. However, Kleinberg et al.'s piece notably does not account for algorithms currently used in the criminal justice system, such as the Pretrial Safety Assessment (PSA) and COMPAS, and the authors also do not consider how discrepancies in how judges choose to use the algorithm may affect their predicted improvements in crime rate and incarceration rates. Kleinberg et al. do, however, acknowledge that the algorithms' crime reductions could compromise normative values, like reducing racial disparity, which could be aggravated by algorithmic predictions (Kleinberg et al. 241, 2017).

Other works in the field have examined the practical divergence in positive versus normative values of using algorithms for risk assessment. In their 2023 piece, “Pretrial Risk Assessment on the Ground: Algorithms, Judgments, Meaning, and Policy,” Moore et al. audit a risk assessment algorithm used in Albuquerque, New Mexico, finding that high risk labels often correspond to extremely low rates of serious violent rearrest; rearrest for high-level felonies is extremely rare, even among those classified as risky (Moore et al. 10, 2023). Whereas Kleinberg et al. define risk solely in terms of the probability of a new arrest, which in itself is riddled with potential biases from systemic overpolicing of certain populations, Moore et al. question what kinds of risk matter for economic efficiency. A defendants’ risk of failing to appear is a different type of risk from a potential new violent criminal arrest, but an algorithmic risk score may not account for the overall rarity of rearrest for serious offenses in pretrial defendant populations. Between Kleinberg et al. and Moore et al., there is broad agreement that assessing algorithmic judicial decisionmaking requires considering both the accuracy of risk predictions and normative values of equality and which groups are most impacted by algorithm-informed judicial decision making. However, whereas Kleinberg et al. offer a more optimistic theoretical view on the potential of algorithms to reduce detention and crime rates, Moore et al. raise normative concerns about the overstatement of risk.

In considering normative values of equality and bias, Koepke and Robinson also raise the concept of “zombie predictions;” pretrial risk assessment tools are often trained on pre-bail reform data (Koepke and Robinson 1726, 2018). Specifically, Koepke and Robinson discuss how both PSA and COMPAS, the two most commonly used risk assessment tools, were built using older, multi-jurisdiction datasets often from higher-crime jurisdictions. As a result, the PSA and COMPAS may be inflating estimates of risk in modern criminal justice systems where initiatives like better pretrial services and changes in policing could potentially improve outcomes like FTA, NCA, and NVCA (Koepke and Robinson 1758, 2018). The use of pretrial risk assessment tools could be entrenching previous detention outcomes, creating a cycle where the inflated predictions of these tools cement outdated perceptions of defendant risk, resulting in an inefficient level of pretrial detention. Koepke and Robinson's piece primarily makes their claims based on a theoretical argument rather than empirical evidence. As such, they rely on the assumption that pretrial reforms do actually have a meaningful impact on FTA/NCA/NVCA outcomes and fail to provide empirical evidence that risk scores are overinflated and causing overincarceration. It is interesting to note, however, that Moore et al.'s subsequent 2023 study seems to partially validate Koepke and Robinson's argument. Koepke and Robinson's zombie predictions theory supports Moore et al.'s findings that high risk scores corresponded to extremely low rates of violent rearrests. Moore et al.'s study offers empirical evidence that risk assessment tools overstate the true risk of FTA/NCA/NVCA for defendants, and this observed overinflation of the risk scores could be explained by Koepke and Robinson's theory of zombie predictions.

Several authors also compare human-alone decision making to algorithmic decision making. In "Evaluating Algorithmic Risk Assessment," Hamilton criticizes traditional validation studies of the PSA that focus on whether failure scores increase with risk scores, instead running a cross-validation study of PSA outputs from three jurisdictions, concluding that local context significantly affects the PSA's predictive performance (Hamilton 207, 2021). While Hamilton attributes local discrepancies in PSA risk-level classifications to miscalibration of the PSA to different localities, Hamilton's study does not fully account for confounding, jurisdiction specific variables that could drive differences in baseline average FTA across jurisdictions reflected in the PSA output. For example, different jurisdictions may have different funding for pretrial services or court infrastructure that could cause systematically higher or lower FTA rates which are authentically picked up by the PSA's predictions and not due to algorithmic miscalibration. Similar to Hamilton, Stevenson's paper, "Assessing Risk Assessment in Action," also assesses the PSA but instead focuses on the impacts of PSA adoption and passage of HB 463 in Kentucky, which mandated judges to consider the risk assessment. Stevenson found no noticeable improvement in detention, releases, FTA, pretrial crime, or racial disparities after PSA adoption, suggesting no changes in accuracy or economic efficiency derived from adopting the PSA (Stevenson 346 & 369, 2018). The external validity of Stevenson's study, however, is questionable; Kentucky is one of few states that has used risk assessment tools since the 1970s, meaning the attitudes towards risk assessment tools could explain why PSA adoption did not change outcomes. By contrast, adoption of the PSA in a state that went from cash bail to the PSA might cause a more drastic shift in outcomes, as judges in those states might incorporate risk assessment outputs differently than Kentucky.

Notably, in considering judicial behavior, Stevenson finds that, despite HB 463's mandate of a statutory presumption of release for low and moderate risk defendants, judges instead used their discretion to preserve the status-quo and ignore the statutory presumption in over two-thirds of cases. Stevenson also found that, despite initial changes in judicial behavior that temporarily increased non-monetary releases, after 2-3 years, bail patterns returned to pre-reform levels (Stevenson 308-309, 2018). A 2021 study by Imai et al. affirms Stevenson's findings, suggesting Stevenson's results have some external validity to other jurisdictions. Imai et al. conducted a randomized control trial of PSA usage, randomly revealing PSA outputs to some judges in Dane County, Wisconsin, to estimate causal effects of algorithm-assisted decision making. The authors found that, even though PSA provision to judges may create more judicial leniency towards female defendants, there is overall little impact on judge's decisions (Imai et al. 186, 2021).

Imai et al.'s study makes several assumptions subject to critique. In using an instrumental variable regression, Imai et al. assume that PSA provision only affects outcomes through judicial decision and has no direct effect on defendants. However, a defendant's PSA score could inadvertently alter attorneys' treatment of defendants or downstream supervision conditions that could increase a defendants' propensity for FTA/NCA/NVCA. Imai et al. also assume that all information used by judges to make decisions is available in the dataset, but there are a number of factors, like the identity of the lawyer or defendant in-court demeanor, unobserved in the dataset that influence judicial decisions (Imai et al. 173-175, 2021). Nevertheless, Imai et al.'s findings agree with Stevenson's conclusions from Kentucky, suggesting some consistency in the lack of impact of the PSA on judicial decisions.

In contrast to Imai et al. and Stevenson, Green and Chen find that risk assessments systematically alter how people perceive risk, providing risk scores to laypeople and finding that score provision made participants prioritize risk over other normative criminal justice system goals, increasing racial disparities in detention by 1.9%, a statistically significant change, despite improvements in risk prediction (Green & Chen 3, 2021). An obvious weakness of Green and Chen's piece, however, is that they use a sample of laypeople rather than judges to extrapolate to the impact of the PSA on judicial decisionmaking. Green and Chen may have some merit in arguing that providing a risk assessment increases the salience of mitigating risk as a priority of the criminal justice system at the expense of normative values from an individual perspective, but aggregate trends in judicial behavior as shown by Imai et al. and Stevenson do not show significant influence over judicial decisionmaking.

One proxy measure of efficiency gains from algorithmic-assisted judicial decisions is increases in predictive accuracy of pre-trial outcomes, measured by whether judges' decisions to release a defendant results in an FTA/NCA/NVCA. In a recent paper titled "Does AI help humans make better decisions?," Ben-Michael et al. differentiate between human-alone, human and AI, and AI-alone decision making, using the same Dane County dataset as Imai et al. to find that risk assessment recommendations do not improve the accuracy of judges' decisions to impose cash bail and replacing judges with predictive algorithms produces even worse classification performance (Ben-Michael et al. 3, 2025). Ben-Michael's findings contradict the theoretical accuracy gains argued by Kleinberg, who found that algorithm-alone decisions could decrease detention and crime rates. Angelova et al. similarly conduct an empirical study, leveraging quasi-random case assignment and focusing on judicial override of algorithmic recommendations. The authors find that 90% of judges underperform the algorithm when they override its recommendations, and the 10% of judges that outperform the algorithm in their overrides are stronger at determining and incorporating legitimately relevant private information about the case in their decisions (Angelova et al. 1-3, 2023). Some examples of legitimately relevant private information that the authors allude to include aggravating factors flagged by pretrial services, like mental illness or the defendant posing a threat to others, details about the victims of the defendants' offenses, and history of substance abuse (Angelova et al. 30, 2023). These factors are not included in the algorithm, but are still potentially predictive of a defendants' failure to appear or likelihood to commit a new offense, suggesting that when judges use these legitimately relevant variables to complement the PSA, there are improvements in judicial decision making. Both Angelova et al. and Ben-Michael et al. use accuracy as the metric of good judicial decision making without considering whether judges' overrides might be serving normative goals beyond accuracy. However, in considering goals of economic efficiency in a criminal justice system—optimal deterrence to minimize the combined costs to society of crime and detention—accuracy of judicial release decisions are an important measure to consider the relative economic costs of judge-alone versus algorithm-assisted decision making.

Several authors have also conducted empirical studies to estimate the impact of algorithm-assisted judicial decision making on the frequency of non-financial release and detention decisions. In their article "Pretrial risk assessment instruments in practice," Copp et al. assess the use of a pretrial risk assessment instrument in a large southeastern county, finding little evidence that instrument adoption improved detention efficiency or increased nonfinancial release because judges frequently deviated from algorithmic recommendations, imposing harsher conditions than the algorithm and especially for black and brown defendants. (Copp et al. 346-347, 2022). In contrast to Kleinberg's theoretical argument for efficiency gains from risk assessments, driven by a more accurate ability to detain the highest risk offenders and release lower risk offenders, Copp et al. actually find that use of the predictive tool increased detention of misdemeanor defendants and decreased detention of felony defendants, suggesting the predictive tool did not increase the efficiency of pretrial detention rates (Copp et al. 344, 2022). Copp et al.'s study is weaker than those of Angelova et al. and Ben-Michael et al. because it is observational in nature and can thus not lend itself to any robust causal conclusions. Additionally, Copp et al. only focus on one southeastern county, limiting the generalizability of their results on the impacts of adopting a predictive tool.

In fact, other studies have found more optimistic outcomes following the adoption of predictive risk assessments. In their 2018 working paper “The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes,” Sloan et al. use a regression discontinuity design to evaluate the impacts of Travis County, Texas adopting a risk score policy. They found a 4.5–7.5 percentage-point increase in non-financial bonds and a 7–10 percentage-point reduction in pretrial detention, seemingly driven by poor defendants (Sloan et al. 4, 2018). The disparities between Sloan et al. and Copp et al.’s findings could be explained by the fact that Copp’s study was observational whereas Sloan studied changes in outcomes immediately after a policy change. Sloan et al. note that the adoption of the risk tool was a part of a policy initiative to reduce pretrial detention, suggesting judges in the county were already primed to meaningfully utilize the risk scores to increase nonfinancial release outcomes (Sloan et al. 9, 2018). On the other hand, Copp et al.’s study observed judges who were already adapted to using the risk assessment score, causing them to systematically override the algorithm and rely on monetary release conditions. Between Sloan et al. and Copp et al., there are divergences in the empirical findings of whether algorithms can increase reliance on non-financial release and reduce pretrial detention. While Imai et al., Ben-Michael et al., Copp et al., and Stevenson all find negligible changes in outcomes between judge-only and algorithm-assisted decision making, Sloan et al. and Angelova et al. suggest there could be improvements in algorithm-assisted judicial decision making.

Findings on accuracy and outcomes of algorithm use lend themselves to various takeaways about the economic efficiency of algorithmic-assisted judicial decision making. Overall, the literature suggests that accuracy improvements from algorithm-assisted judicial decision making are modest or context-dependent. Hamilton’s cross-validation study of the PSA challenges the one-size-fits-all marketing narrative that its creators sell, instead arguing that PSA performance varies across jurisdictions (Hamilton 49 & 52, 2021). Stevenson's study of Kentucky following the adoption of the PSA found that the tool did not substantially change FTA/NCA/NVCA volumes, showing that any potential accuracy gains from the PSA are not translating into actual outcomes (Stevenson et al. 358-359, 2018). Imai et al. and Ben-Michael et al. confirm that PSA outputs do not increase judges' classification accuracy, which Angelova et al. partially explain by identifying how most judges' overrides of the PSA reduce accuracy relative to a high-performing algorithm (Imai et al. 168, 2021; Ben-Michael et al. 7, 2025; Angelova et al. 2, 2023). There are, however, variations in the literature as to the true accuracy gains of pure algorithmic risk prediction, though in practice, this paper will focus on judicial-assisted decision making.

Economic efficiency can also be framed in terms of detention costs. Kleinberg et al. argue for the theoretical pareto gains in both lower detention rates and lower crime rates by citing the potential of algorithms to give an accurate ranking of the highest and lowest risk defendants, allowing judges to release more low risk offenders while detaining only the highest risk offenders (Kleinberg et al. 239-241, 2017). Empirical evidence, however, suggests that risk algorithms may cause overdetention. Moore et al. showed that higher risk scores largely predict minor offenses, and that the marginal serious felony prevented by additional detention is extremely rare (Moore et al. 13, 2023). Given that the average cost of incarcerating an individual per day at the federal level was $120 in 2024, Moore's findings suggest that the PSA may lead to excessive and inefficient detention costs that exceed the deterrence effect of that expenditure on detention (Wills, 2024). While this per day cost figure varies widely by jurisdiction due to different detention systems and municipal resources, there is still a cost to detaining individuals that may not be justified by Moore's findings that higher risk scores predict merely minor offenses.

In contrast, Sloan et al.'s regression discontinuity analysis found a moderate increase in non-financial release conditions that reduced pretrial detention without raising violent recidivism rates, a result consistent with efficiency gains from algorithm-informed decision making (Sloan et al. 2018, 4). Notably, Sloan et al.'s analysis did find a mild increase in non-violent recidivism, leaving an open normative question about society's trade off preference between different types of recidivism and detention costs.

Another normative challenge not fully captured by focus on economic efficiency is fairness between identity groups. Copp et al. found that use of predictive risk tools worsened racial disparities, as judges were more likely to override the algorithm recommendation in a more punitive way for Black and Brown defendants, reducing fairness (Copp et al. 346-347, 2022). Green and Chen corroborate the increase in racial disparity from algorithmic-informed decision making, finding that risk tools made study participants more risk averse in racially patterned ways (Green & Chen 3, 2021). Imai et al. similarly found that judges' decisions became slightly less fair by gender when provided with algorithmic risk assessment outputs (Imai et al. 186, 2021). Beyond the accuracy of algorithmic decision making, economic efficiency, as defined in this paper, does not account for potential normative values of fairness and equal treatment across identity groups. Biased data fed to predictive algorithms, generated by prosecutorial and policing systems riddled with patterns of discriminatory behavior against certain minority groups lead algorithms to account for different base rates in risk across different demographic groups. Therefore, the use of algorithmic-assisted decision making may perpetuate or even exacerbate human biases in the pretrial detention or release decision.

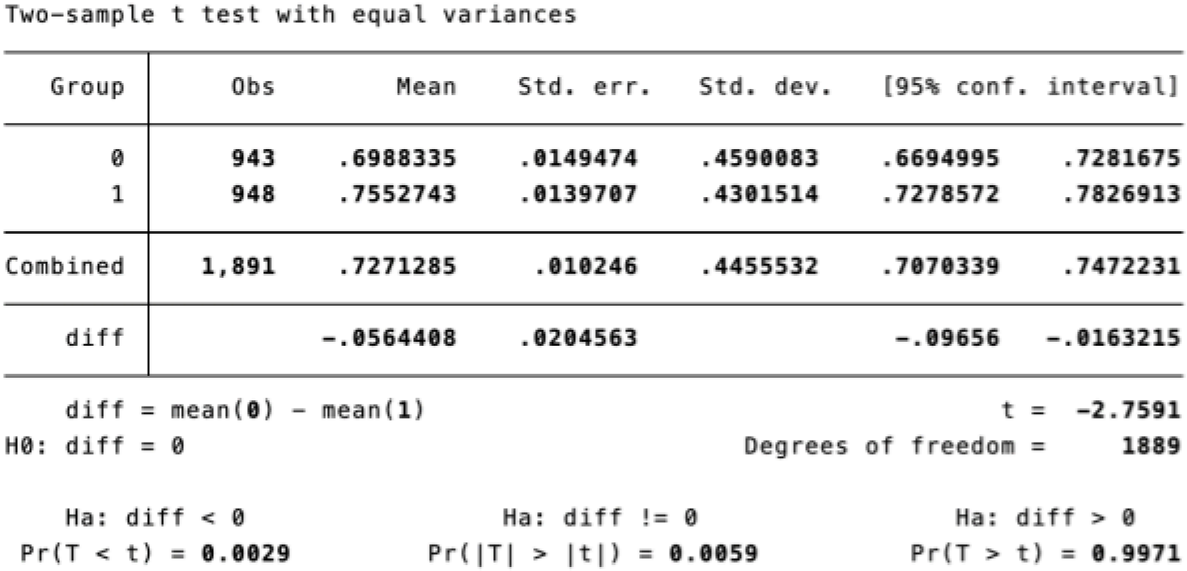
Research Methods and Analysis

My analysis uses data of defendants from Dane County, Wisconsin between 2017-2018. I first cleaned the data, re-coding several variables used for my analysis. I re-coded the judge decision indicator variable, D, from a three-value variable (0 for release, 1 for cash bail, 2 for detention) to a binary variable, with 0 indicating non-monetary release and 1 indicating cash bail or detention. I re-coded this variable to match the DMF variable, which represents the PSA's Decision Making Framework. The DMF variable indicates the algorithm's ultimate recommendation for a client's pre-trial detention or release. A value of 0 indicated unconditional release, and a value of 1 indicated a cash bail or detention recommendation. By re-coding my judicial decision variable, I could assess agreement with and override of the PSA. I created a new variable, judge_override, which took on a value of 1 if the judge overrode the algorithm in favor of the defendant—that is, the algorithm recommended bail or detention, but the judge released the defendant. I also merged the PSA output data with datasets containing FTA/NCA/NVCA for the same defendants, from which I was able to create a anyFlag variable to indicate whether a defendant had a failure to appear, new criminal activity, or new violent criminal activity. I then used this anyFlag variable to evaluate the accuracy of judicial decision making before and after PSA use.

For my analysis, I ran several logit regressions, which was the preferred methodology due to the binary nature of the variables I was regressing on. Using a logit model ensured predicted probabilities stayed within [0, 1] and allowed for meaningful interpretation of my interaction terms through marginal effects. I used a logit regression to determine if there were any demographic variables significantly correlated with a judges' decision to override the PSA. I also used a logit regression to determine if there was any significant difference in judges' probability to detain for different demographic groups following the provision of the PSA. To interpret the approximate probability effect from the logit regression coefficients, I had to determine a value of P where $\Delta P = b * P(1-P)$. P is the probability of judge override, also described as the proportion of times that the judge_override variable took on a value of 1 in the dataset. The b value was the coefficient on the variable of interest, which varied depending on the regression run. Because approximately a quarter of defendants were detained, I set $P = 0.25$ for logit regressions where judges' detention decision was the outcome variable. For logit regressions on judges' probability of overriding the algorithm, I set $P = 0.17$, as judges' overrode the algorithm in approximately 17% of cases. Thus, to determine the difference in override probabilities between different groups, I plugged in the coefficient on my variable of interest so that $\Delta P = b * (0.17)(0.83)$. To determine the change in probability of detention for a given demographic group, I plugged in the coefficient of the interaction term so that $\Delta P = b * (0.25)(0.75)$. Acknowledging that the judge_override variable is not a true override when the PSA is not provided, I note that, in cases where the PSA is not provided, judge_override still indicates a discrepancy between judge and algorithm decision which is more favorable for the defendant. Comparing the change in the favorable disagreement between algorithm and judge—that is, when the algorithm recommends detention but the judge chooses to release—still allows me to deduce whether PSA provision is systematically changing the likelihood that the judges' decision for defendants of a certain demographic group is more lenient than the algorithm recommends. Finally, I conducted a t test to determine whether there were any meaningful differences in judges' accuracy rate without versus with the PSA.

Results

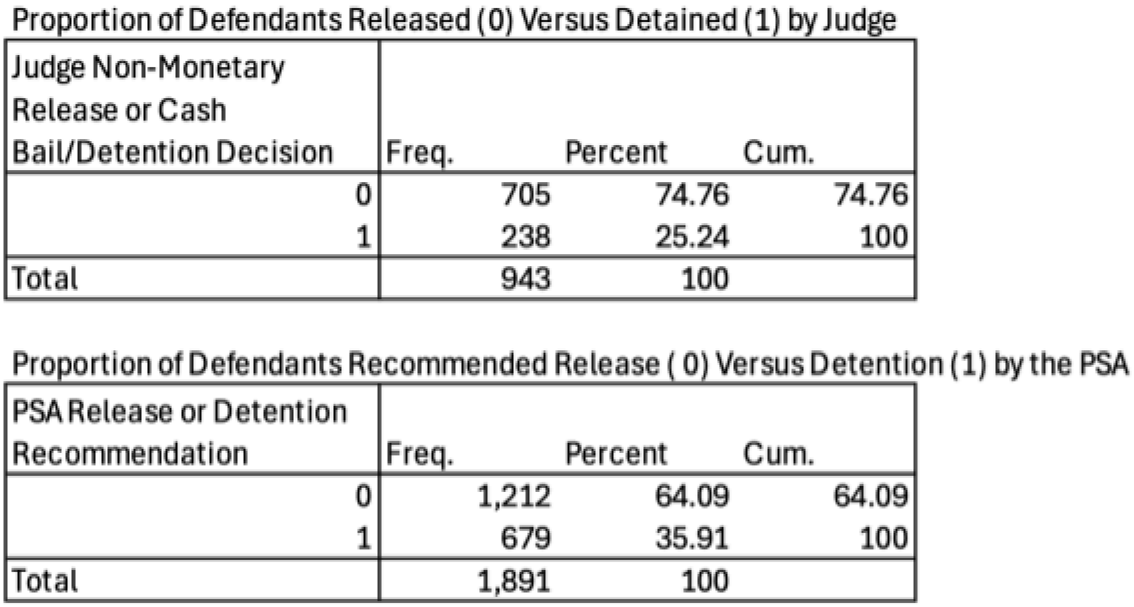
I started first by assessing the extent to which algorithm-assisted judicial decision making changed judges' behavior by examining the proportion of decisions where judges agreed with the algorithm without and with the provision of the PSA score. I find that, when judges do not see the PSA output, they agree with the PSA recommendation approximately 70% of the time, but when they are provided the PSA output their agreement increases to 75.5% of the time (Fig. 1). A t-test suggests this difference in proportions is significant, yielding a t-score of -2.76 (Fig. 1). This result suggests that judges in Dane County were meaningfully incorporating PSA outputs into their pre-trial decision making process. Prior to my regression analysis, I also assessed the relative strictness of the algorithm and the judge, comparing the percentage of defendants released by the algorithm to those released by the judge. I find that, whereas the PSA released approximately 64% of defendants, the judges released 74.76% of defendants (Fig. 2). This comparison shows that, on average, the algorithm is stricter than the judge. My analysis therefore necessitates a further examination of judges' lenience, identifying for whom judges are more or less likely to override the algorithm in favor of a defendant's release.



```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       0 |     943    .6988335    .0149474    .4590083    .6694995    .7281675
       1 |     948    .7552743    .0139707    .4301514    .7278572    .7826913
---------+--------------------------------------------------------------------
Combined |   1,891    .7271285     .010246    .4455532    .7070339    .7472231
---------+--------------------------------------------------------------------
    diff |           -.0564408    .0204563                 -.09656   -.0163215
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.7591
H0: diff = 0                                     Degrees of freedom =     1889

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0029         Pr(|T| > |t|) = 0.0059          Pr(T > t) = 0.9971
```

Figure 1: T-Test Comparing Difference in Algorithm Agreement Between Judges Provided the PSA Versus Judges Not Provided the PSA. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.



Proportion of Defendants Released (0) Versus Detained (1) by Judge

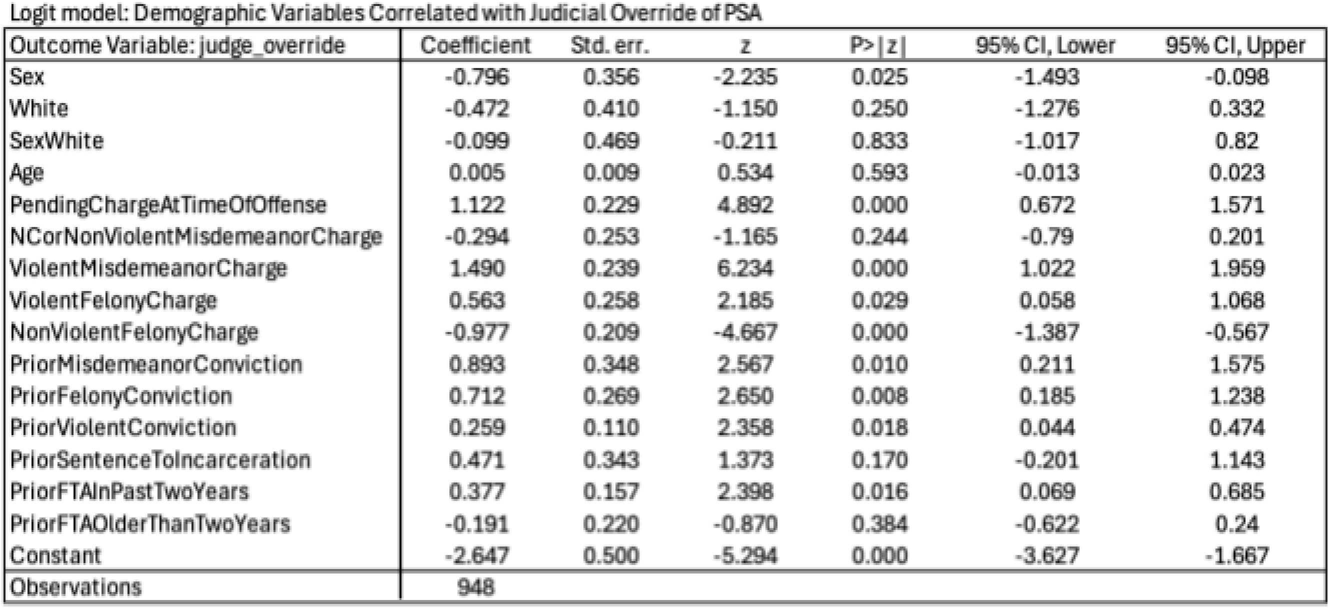
Judge Non-Monetary Release or Cash Bail/Detention Decision	Freq.	Percent	Cum.
0	705	74.76	74.76
1	238	25.24	100
Total	943	100	

Proportion of Defendants Recommended Release (0) Versus Detention (1) by the PSA

PSA Release or Detention Recommendation	Freq.	Percent	Cum.
0	1,212	64.09	64.09
1	679	35.91	100
Total	1,891	100	

Figure 2: Percentage of Defendants Released By PSA Versus Judge. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

To answer the question of whether there were demographic variables systematically correlated with a judges' propensity to release a defendant, I ran a logistic regression on my judge_override variable, controlling for all other available information about defendants (e.g., demographics, prior charges, and prior failures to appear). Of the demographic variables, I find that there is a significant relationship between defendant gender and a judges' likelihood of override. Specifically, the coefficient on my Sex variable is -0.7955616, meaning that female defendants are 11.2 percentage points more likely than male defendants to have a detention recommendation from the PSA overridden, holding all other defendant demographic and criminal history factors constant (Fig. 3). These findings suggest that judges may be systematically correcting the PSA based on gender, or other factors correlated with gender not included in the dataset. Judges seem to be, on average, more lenient with female defendants in their override behavior. My findings on override behavior corroborate Imai et al.'s findings that the PSA causes judges to be more lenient with female defendants, identifying one potential channel—overrides—by which judges provide this lenience.



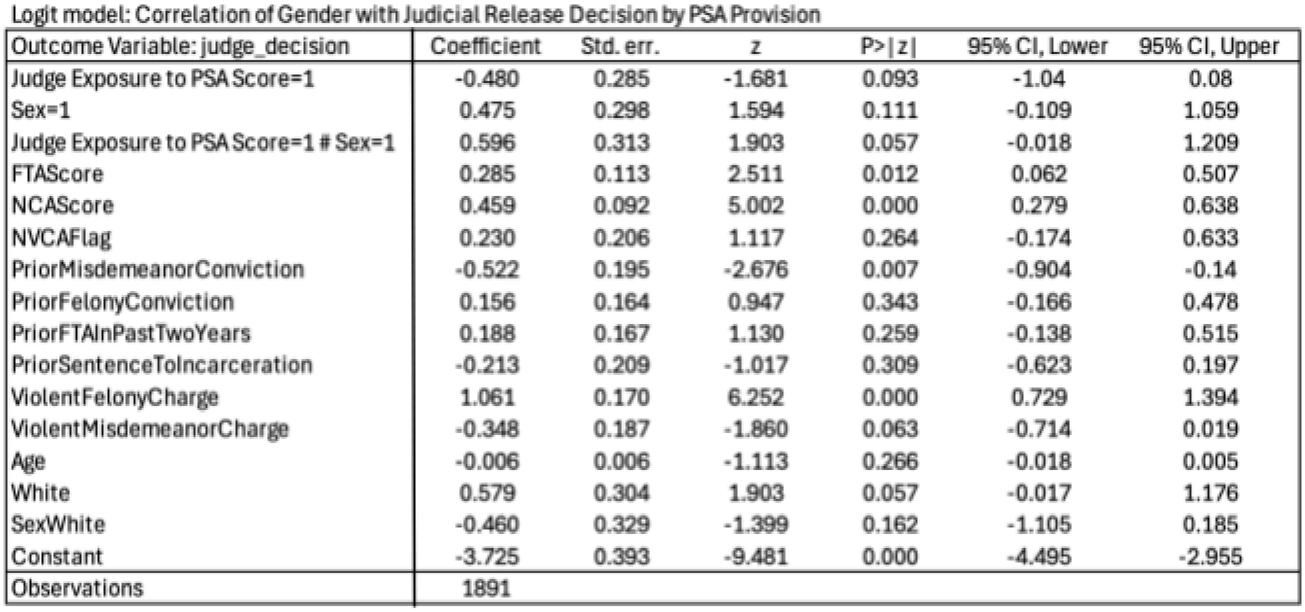
Logit model: Demographic Variables Correlated with Judicial Override of PSA

Outcome Variable: judge_override	Coefficient	Std. err.	z	P>|z|	95% CI, Lower	95% CI, Upper
Sex	-0.796	0.356	-2.235	0.025	-1.493	-0.098
White	-0.472	0.410	-1.150	0.250	-1.276	0.332
SexWhite	-0.099	0.469	-0.211	0.833	-1.017	0.82
Age	0.005	0.009	0.534	0.593	-0.013	0.023
PendingChargeAtTimeOfOffense	1.122	0.229	4.892	0.000	0.672	1.571
NCorNonViolentMisdemeanorCharge	-0.294	0.253	-1.165	0.244	-0.79	0.201
ViolentMisdemeanorCharge	1.490	0.239	6.234	0.000	1.022	1.959
ViolentFelonyCharge	0.563	0.258	2.185	0.029	0.058	1.068
NonViolentFelonyCharge	-0.977	0.209	-4.667	0.000	-1.387	-0.567
PriorMisdemeanorConviction	0.893	0.348	2.567	0.010	0.211	1.575
PriorFelonyConviction	0.712	0.269	2.650	0.008	0.185	1.238
PriorViolentConviction	0.259	0.110	2.358	0.018	0.044	0.474
PriorSentenceToIncarceration	0.471	0.343	1.373	0.170	-0.201	1.143
PriorFTAInPastTwoYears	0.377	0.157	2.398	0.016	0.069	0.685
PriorFTAOlderThanTwoYears	-0.191	0.220	-0.870	0.384	-0.622	0.24
Constant	-2.647	0.500	-5.294	0.000	-3.627	-1.667
Observations	948					

Figure 3: Logistic Regression Identifying Demographic Factors Significantly Correlated with Judge Override of PSA. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

To further understand the relationship between PSA provision and gender, I then run another logit regression along with a marginal effects analysis to identify whether there is a significant change in judges' release decision and judges' overrides of the algorithm for certain gender groups as a result of PSA provision. The regression returns a coefficient of .5958856 on my interaction term, PSA_treatment x Sex, significant at the $p < 0.1$ level (Fig. 4). Interpreting this coefficient means that after judges are provided the PSA output, male defendants are 11.2 percentage points more likely than female defendants to be detained. My marginal effects analysis corroborates this increase in gender bias in algorithm-assisted decision making. Before the PSA, the difference between female and male detention probability (22.9% and 26%, respectively), was 3.1 percentage points (Fig. 5). Once judges are provided with the PSA output, female detention probability drops to 16.9%, whereas male detention probability increases to

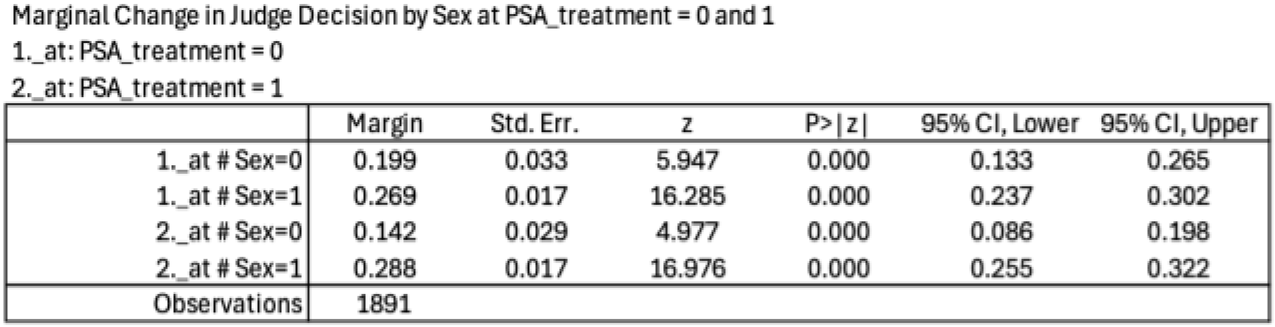
27.8%, increasing the gap between female and male detention rates to approximately 11 percentage points (Fig. 5). My results suggest that providing the PSA caused judges to respond differently for female and male defendants, inducing a more lenient judicial attitude towards female defendants and a more punitive attitude towards male defendants. My findings agree with Imai et al., who also found that PSA provision worsened gender disparities in judicial detention decisions.



Logit model: Correlation of Gender with Judicial Release Decision by PSA Provision

Outcome Variable: judge_decision	Coefficient	Std. err.	z	P>|z|	95% CI, Lower	95% CI, Upper
Judge Exposure to PSA Score=1	-0.480	0.285	-1.681	0.093	-1.04	0.08
Sex=1	0.475	0.298	1.594	0.111	-0.109	1.059
Judge Exposure to PSA Score=1 # Sex=1	0.596	0.313	1.903	0.057	-0.018	1.209
FTAScore	0.285	0.113	2.511	0.012	0.062	0.507
NCAScore	0.459	0.092	5.002	0.000	0.279	0.638
NVCAFlag	0.230	0.206	1.117	0.264	-0.174	0.633
PriorMisdemeanorConviction	-0.522	0.195	-2.676	0.007	-0.904	-0.14
PriorFelonyConviction	0.156	0.164	0.947	0.343	-0.166	0.478
PriorFTAInPastTwoYears	0.188	0.167	1.130	0.259	-0.138	0.515
PriorSentenceToIncarceration	-0.213	0.209	-1.017	0.309	-0.623	0.197
ViolentFelonyCharge	1.061	0.170	6.252	0.000	0.729	1.394
ViolentMisdemeanorCharge	-0.348	0.187	-1.860	0.063	-0.714	0.019
Age	-0.006	0.006	-1.113	0.266	-0.018	0.005
White	0.579	0.304	1.903	0.057	-0.017	1.176
SexWhite	-0.460	0.329	-1.399	0.162	-1.105	0.185
Constant	-3.725	0.393	-9.481	0.000	-4.495	-2.955
Observations	1891					

Figure 4: Logistic Regression Identifying Changes in Judicial Decisions for Different Genders With PSA Provision. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

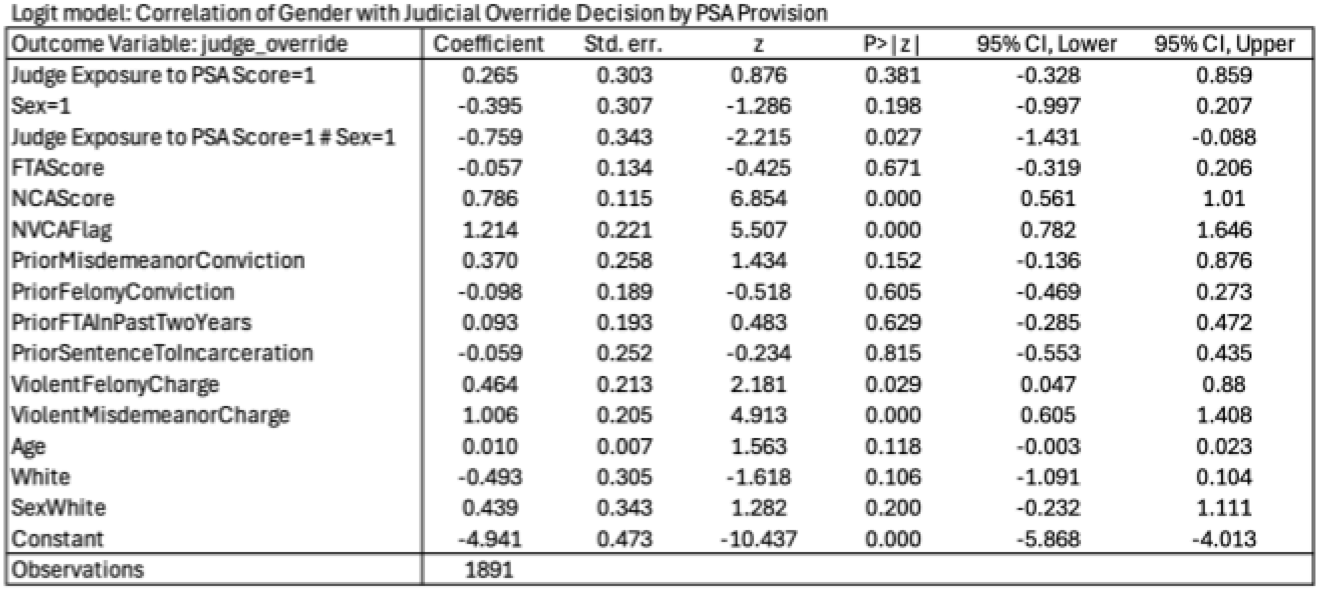


Marginal Change in Judge Decision by Sex at PSA_treatment = 0 and 1
1._at: PSA_treatment = 0
2._at: PSA_treatment = 1

	Margin	Std. Err.	z	P>|z|	95% CI, Lower	95% CI, Upper
1._at # Sex=0	0.199	0.033	5.947	0.000	0.133	0.265
1._at # Sex=1	0.269	0.017	16.285	0.000	0.237	0.302
2._at # Sex=0	0.142	0.029	4.977	0.000	0.086	0.198
2._at # Sex=1	0.288	0.017	16.976	0.000	0.255	0.322
Observations	1891					

Figure 5: Results of Marginal Analysis Comparing Judge Decision Probabilities for Different Genders (Sex = 1 → Male, Sex = 0 → Female) With and Without PSA Provision. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

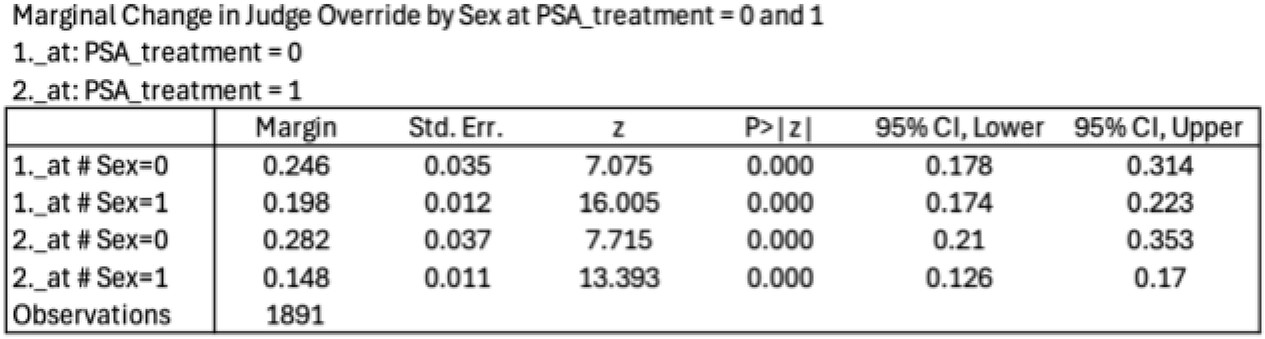
Examining judicial override behavior yielded similar results. The logit regression on the judge_override variable compared the probability of a defendant-favorable disagreement with the algorithm with and without PSA provision. The coefficient on my interaction term, PSA_treatment x Sex, is -0.759202, significant at the $p < 0.05$ level (Fig. 6). This coefficient can be interpreted as female defendants being approximately an additional 10.7 percentage points more likely than male defendants to receive a favorable override when the judge is provided the PSA outcome. The marginal analysis corroborates this finding. Without the PSA, the gap between the probability of female and male defendants receiving a favorable override was 2 percentage points, but with the PSA, this gap increases to 10.2 percentage points (Fig. 7). These findings suggest that providing judges with the PSA drastically increases the disparity between favorable judge overrulings of the algorithm for female and male defendants, with female defendants becoming more likely to receive an override and male defendants becoming less likely to receive an override.



Logit model: Correlation of Gender with Judicial Override Decision by PSA Provision

Outcome Variable: judge_override	Coefficient	Std. err.	z	P>|z|	95% CI, Lower	95% CI, Upper
Judge Exposure to PSA Score=1	0.265	0.303	0.876	0.381	-0.328	0.859
Sex=1	-0.395	0.307	-1.286	0.198	-0.997	0.207
Judge Exposure to PSA Score=1 # Sex=1	-0.759	0.343	-2.215	0.027	-1.431	-0.088
FTAScore	-0.057	0.134	-0.425	0.671	-0.319	0.206
NCAScore	0.786	0.115	6.854	0.000	0.561	1.01
NVCAFlag	1.214	0.221	5.507	0.000	0.782	1.646
PriorMisdemeanorConviction	0.370	0.258	1.434	0.152	-0.136	0.876
PriorFelonyConviction	-0.098	0.189	-0.518	0.605	-0.469	0.273
PriorFTAInPastTwoYears	0.093	0.193	0.483	0.629	-0.285	0.472
PriorSentenceToIncarceration	-0.059	0.252	-0.234	0.815	-0.553	0.435
ViolentFelonyCharge	0.464	0.213	2.181	0.029	0.047	0.88
ViolentMisdemeanorCharge	1.006	0.205	4.913	0.000	0.605	1.408
Age	0.010	0.007	1.563	0.118	-0.003	0.023
White	-0.493	0.305	-1.618	0.106	-1.091	0.104
SexWhite	0.439	0.343	1.282	0.200	-0.232	1.111
Constant	-4.941	0.473	-10.437	0.000	-5.868	-4.013
Observations	1891					

Figure 6: Logistic Regression Identifying Changes in Favorable Judicial Override for Different Genders With PSA Provision. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

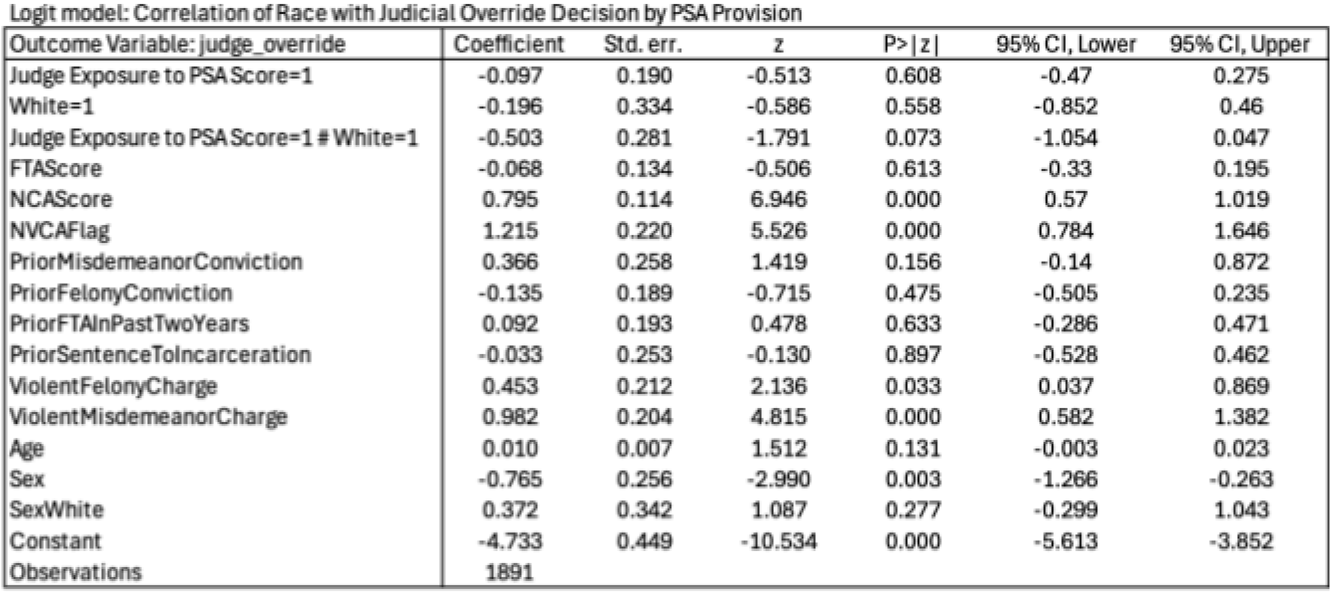


Marginal Change in Judge Override by Sex at PSA_treatment = 0 and 1
1._at: PSA_treatment = 0
2._at: PSA_treatment = 1

	Margin	Std. Err.	z	P>|z|	95% CI, Lower	95% CI, Upper
1._at # Sex=0	0.246	0.035	7.075	0.000	0.178	0.314
1._at # Sex=1	0.198	0.012	16.005	0.000	0.174	0.223
2._at # Sex=0	0.282	0.037	7.715	0.000	0.21	0.353
2._at # Sex=1	0.148	0.011	13.393	0.000	0.126	0.17
Observations	1891					

Figure 7: Results of Marginal Analysis Comparing Judge Favorable Override Probabilities for Different Genders With and Without PSA Provision. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

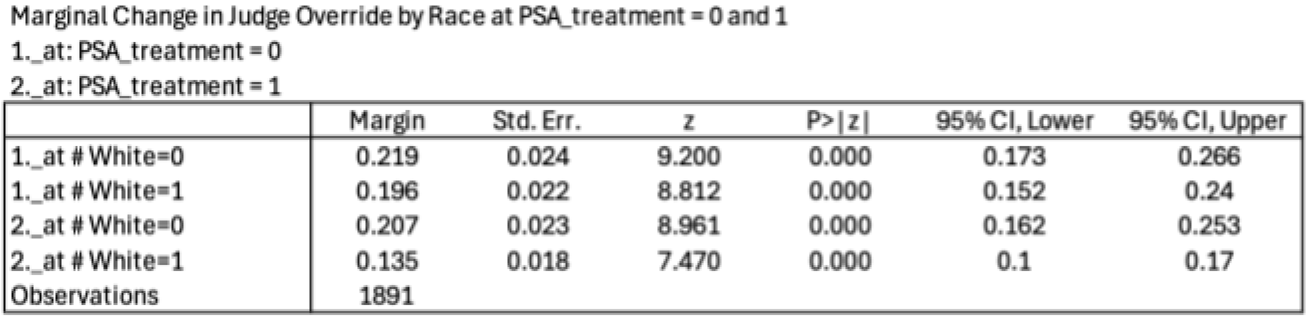
Beyond gender, I also analyzed whether there were meaningful differences in judges' decisions and judges' overrides for white versus non white defendants. While I find no significant difference in judges' release versus detention decisions for white versus nonwhite defendants when provided the PSA, there is a statistically significant difference in favorable judicial overrides for white versus non-white defendants. The coefficient in my logit regression is -0.5031729, significant at the $p < 0.1$ level (Fig. 8). Interpreting this coefficient means that when the PSA was provided, non-white defendants were an additional 7.1 percentage points more likely to receive a judicial override. The marginal effect analysis confirms this discrepancy: the gap in the likelihood of a favorable disagreement in a judges' final decision between white and non white defendants is 2.3 percentage points without the PSA and increases to 7.3 percentage points with the PSA (Fig. 9). The increase in the gap is primarily driven by a reduction in the probability of override for white defendants. These findings suggest, similar to the gender outcomes, that judges may be systematically correcting for perceived racial bias, shown by the decrease in probability of override for white defendants. My results challenge Imai et al.'s findings that algorithm-assisted judicial decision making had no significant correlation with race. The analysis shows that judicial override behavior became relatively harsher towards white defendants when the PSA was provided.



Logit model: Correlation of Race with Judicial Override Decision by PSA Provision

Outcome Variable: judge_override	Coefficient	Std. err.	z	P>|z|	95% CI, Lower	95% CI, Upper
Judge Exposure to PSA Score=1	-0.097	0.190	-0.513	0.608	-0.47	0.275
White=1	-0.196	0.334	-0.586	0.558	-0.852	0.46
Judge Exposure to PSA Score=1 # White=1	-0.503	0.281	-1.791	0.073	-1.054	0.047
FTAScore	-0.068	0.134	-0.506	0.613	-0.33	0.195
NCAScore	0.795	0.114	6.946	0.000	0.57	1.019
NVCAFlag	1.215	0.220	5.526	0.000	0.784	1.646
PriorMisdemeanorConviction	0.366	0.258	1.419	0.156	-0.14	0.872
PriorFelonyConviction	-0.135	0.189	-0.715	0.475	-0.505	0.235
PriorFTAInPastTwoYears	0.092	0.193	0.478	0.633	-0.286	0.471
PriorSentenceToIncarceration	-0.033	0.253	-0.130	0.897	-0.528	0.462
ViolentFelonyCharge	0.453	0.212	2.136	0.033	0.037	0.869
ViolentMisdemeanorCharge	0.982	0.204	4.815	0.000	0.582	1.382
Age	0.010	0.007	1.512	0.131	-0.003	0.023
Sex	-0.765	0.256	-2.990	0.003	-1.266	-0.263
SexWhite	0.372	0.342	1.087	0.277	-0.299	1.043
Constant	-4.733	0.449	-10.534	0.000	-5.613	-3.852
Observations	1891					

Figure 8: Logistic Regression Identifying Changes in Favorable Judicial Override for Different Racial Groups With PSA Provision. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.



Marginal Change in Judge Override by Race at PSA_treatment = 0 and 1

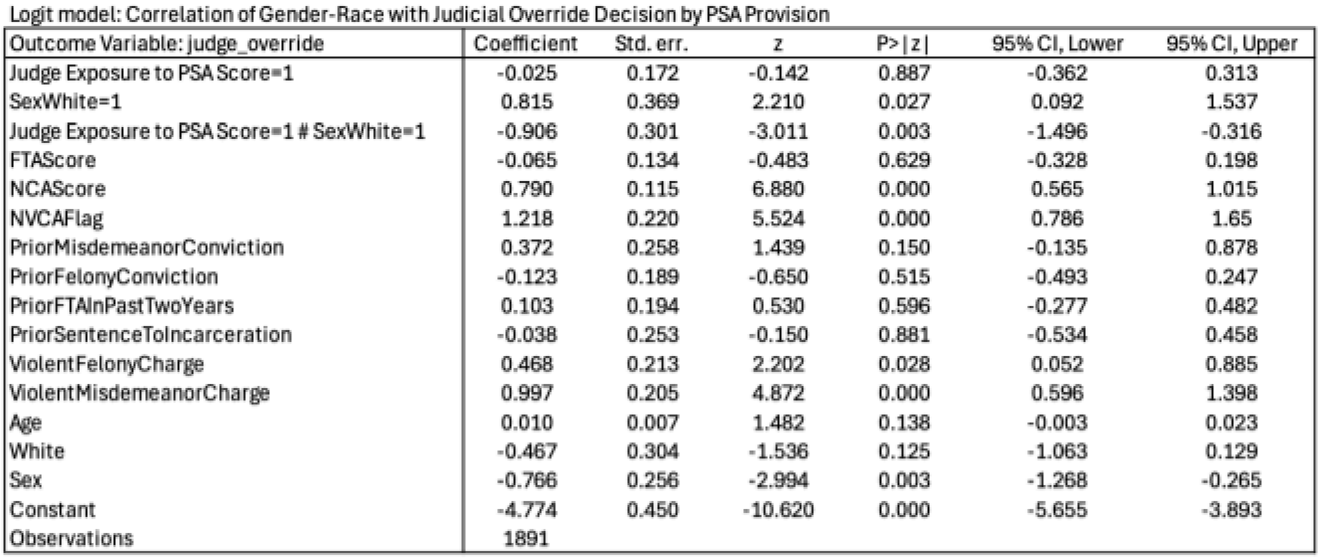
1._at: PSA_treatment = 0

2._at: PSA_treatment = 1

	Margin	Std. Err.	z	P>|z|	95% CI, Lower	95% CI, Upper
1._at # White=0	0.219	0.024	9.200	0.000	0.173	0.266
1._at # White=1	0.196	0.022	8.812	0.000	0.152	0.24
2._at # White=0	0.207	0.023	8.961	0.000	0.162	0.253
2._at # White=1	0.135	0.018	7.470	0.000	0.1	0.17
Observations	1891					

Figure 9: Results of Marginal Analysis Comparing Judge Favorable Override Probabilities for Different Racial Groups With and Without PSA Provision. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

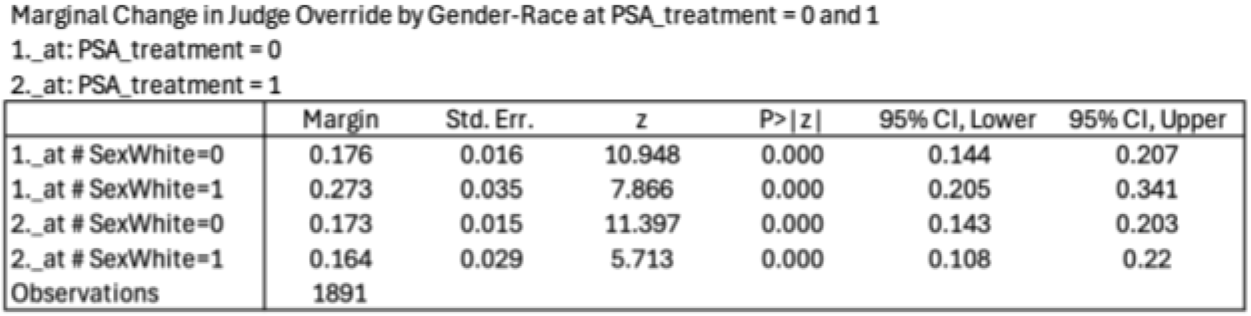
Finally, I repeated the analysis for gender and race on the gender-race interaction term, SexWhite. Similar to the outcome for racial groups, I found no significant difference in judicial detention versus release decisions for white males compared to other defendants when a judge was provided the PSA versus when they were not. However, I do find a significant change in the probability of judicial override across gender-racial groups when a judge was provided the PSA versus when they were not. The coefficient on the interaction term of PSA_treatment x SexWhite is -0.9064556, significant at the p < 0.05 level (Fig. 10). After interpretation, this coefficient indicates that the probability of a favorable override for white male defendants dropped 12.8 percentage points once the PSA was provided. The marginal effects analysis shows that, without the PSA, white male defendants had a higher probability of a favorable judicial override, with a 27.3% chance of getting an override compared to 17.6% chance of non white male defendants (Fig. 11). However, with the PSA, the probability of an override for white male defendants dropped to 16.39%, while the probability for non white male defendants stayed roughly constant at 17.3%. These findings suggest that the PSA might have helped reduce racial bias in judges' decisions, reducing the leniency judges provided to white male defendants when they were deciding without the PSA.



Logit model: Correlation of Gender-Race with Judicial Override Decision by PSA Provision

Outcome Variable: judge_override	Coefficient	Std. err.	z	P>|z|	95% CI, Lower	95% CI, Upper
Judge Exposure to PSA Score=1	-0.025	0.172	-0.142	0.887	-0.362	0.313
SexWhite=1	0.815	0.369	2.210	0.027	0.092	1.537
Judge Exposure to PSA Score=1 # SexWhite=1	-0.906	0.301	-3.011	0.003	-1.496	-0.316
FTAScore	-0.065	0.134	-0.483	0.629	-0.328	0.198
NCAScore	0.790	0.115	6.880	0.000	0.565	1.015
NVCAFlag	1.218	0.220	5.524	0.000	0.786	1.65
PriorMisdemeanorConviction	0.372	0.258	1.439	0.150	-0.135	0.878
PriorFelonyConviction	-0.123	0.189	-0.650	0.515	-0.493	0.247
PriorFTAInPastTwoYears	0.103	0.194	0.530	0.596	-0.277	0.482
PriorSentenceToIncarceration	-0.038	0.253	-0.150	0.881	-0.534	0.458
ViolentFelonyCharge	0.468	0.213	2.202	0.028	0.052	0.885
ViolentMisdemeanorCharge	0.997	0.205	4.872	0.000	0.596	1.398
Age	0.010	0.007	1.482	0.138	-0.003	0.023
White	-0.467	0.304	-1.536	0.125	-1.063	0.129
Sex	-0.766	0.256	-2.994	0.003	-1.268	-0.265
Constant	-4.774	0.450	-10.620	0.000	-5.655	-3.893
Observations	1891					

Figure 10: Logistic Regression Identifying Changes in Favorable Judicial Override for Different Gender-Racial Groups With PSA Provision. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.



Marginal Change in Judge Override by Gender-Race at PSA_treatment = 0 and 1

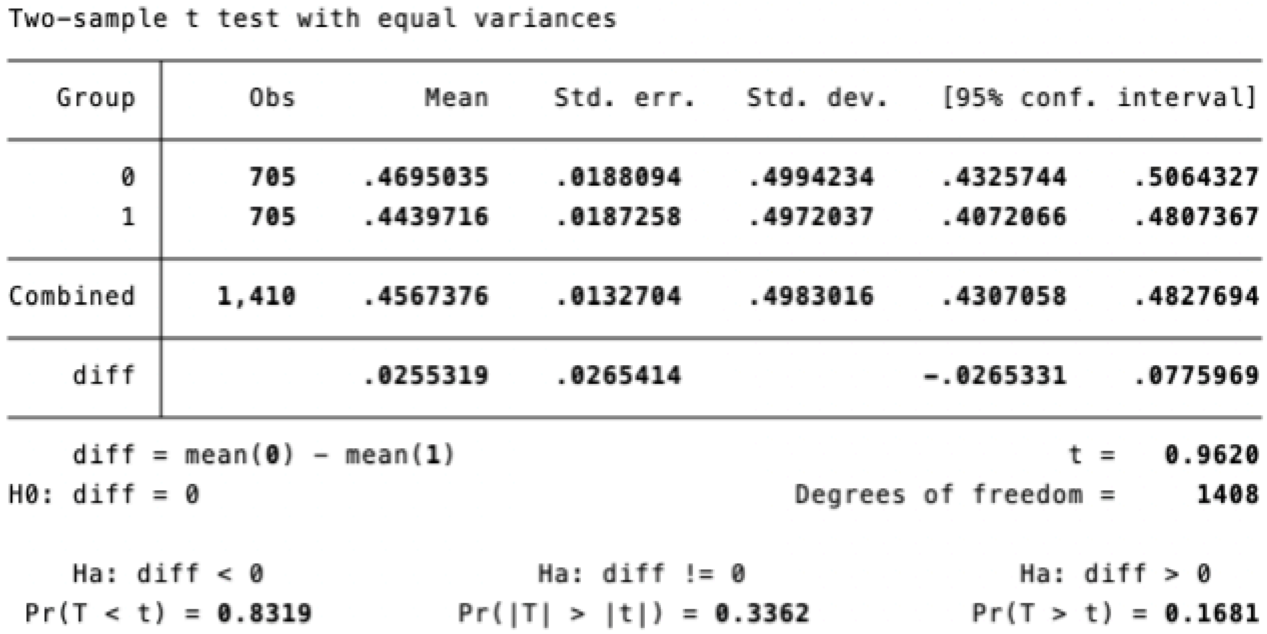
1._at: PSA_treatment = 0

2._at: PSA_treatment = 1

	Margin	Std. Err.	z	P>|z|	95% CI, Lower	95% CI, Upper
1._at # SexWhite=0	0.176	0.016	10.948	0.000	0.144	0.207
1._at # SexWhite=1	0.273	0.035	7.866	0.000	0.205	0.341
2._at # SexWhite=0	0.173	0.015	11.397	0.000	0.143	0.203
2._at # SexWhite=1	0.164	0.029	5.713	0.000	0.108	0.22
Observations	1891					

Figure 11: Results of Marginal Analysis Comparing Judge Favorable Override Probabilities for Different Gender-Racial Groups With and Without PSA Provision. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

To compare accuracy in judicial decision making with and without the PSA, I run a t-test on the difference in proportion of defendants released who have any FTA/NCA/NVCA flag. While there is a difference in proportion of defendants with any flags, where the proportion is 0.47 for defendants released without the PSA and 0.44 for defendants released with the PSA, implying a gain in accuracy, the t-test does not return a statistically significant difference in mean value of the anyFlag variable (Fig. 12). My findings agree with those of Ben-Michael et al. in that algorithm-assisted judicial decision making did not increase accuracy significantly.



```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       0 |     705    .4695035    .0188094    .4994234    .4325744    .5064327
       1 |     705    .4439716    .0187258    .4972037    .4072066    .4807367
---------+--------------------------------------------------------------------
Combined |   1,410    .4567376    .0132704    .4983016    .4307058    .4827694
---------+--------------------------------------------------------------------
    diff |            .0255319    .0265414               -.0265331    .0775969
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.9620
H0: diff = 0                                     Degrees of freedom =     1408

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.8319         Pr(|T| > |t|) = 0.3362          Pr(T > t) = 0.1681
```

Figure 12: T-Test Results Comparing Proportion of Defendants Released with FTA/NCA/NVCA for Judges without PSA versus Judges with PSA. Data Source: Imai, Kosuke, Zhichao Jiang, D. James Greiner, Ryan Halen, and Sooahn Shin. 2023. "Replication Data for: Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/L0NHQU.

Conclusions

Overall, my findings have focused on judicial override behavior and changes in accuracy from PSA-assisted judicial decision making. While I did not find significant changes in judicial decisions of detention versus release for different gender, racial, or gender-racial groups following the provision of the PSA, I did find that the provision of the PSA caused systematic patterns of judicial override behavior in favor of female defendants and less in favor of white defendants and white male defendants. I also found no significant accuracy gains when judges used the PSA, though the data do suggest that judges are considering PSA outputs when making pretrial release or detention decisions.

The lack of accuracy gain suggests that adopting the PSA does not improve economic efficiency of pre-trial decision making, with efficiency defined by more accurate detention outcomes and thus lower crime rates. My findings did suggest that judges' decision making was influenced by seeing PSA results, with judge-algorithm agreement increasing once the judges were given the risk score. Given the flaws discussed with the PSA, such as its relatively harsher decision making compared to judges and vulnerability of outputs to zombie predictions, the use of the PSA may be merely adding randomness to the harshness of judicial decisions without improving accuracy. Moore et al.'s findings that most high risk scores do not correspond to more violent crimes or severe criminal activity also suggest that reliance on the PSA may cause overdeterrence where the costs of incarceration outweigh the societal costs saved in crime prevention.

I do find that providing the PSA systematically changes judicial override behavior in favor of marginalized groups. However, whether the override patterns are a positive or negative outcome depends on the normative values of how society perceives the goals of the criminal justice system. My findings disagree with those of Copp et al., who found that judges in their dataset were overriding the algorithm to impose harsher penalties on Black and Brown defendants. In my analysis, judges became less likely to override the algorithm in favor of white and white male defendants, while override rates remained largely unchanged for non-white defendants. The disparities between my conclusions and those of Copp et al. can potentially be explained by cultural differences between the counties studied. Whereas Copp et al. study a large, southeastern U.S. county, my dataset covers Dane County, Wisconsin, a county in the midwestern U.S. known for their progressive attitudes towards criminal justice system reform. As evidence of the difference in attitudes affecting judicial decision making, Copp et al. found that the vast majority of judicial overrides in their dataset were in the direction of more restrictive decisions, whereas my analysis found that judges in Dane County were more lenient on defendants than the PSA (Copp et al. 346, 2022). This disparity in baseline lenience of judges reflects a potential cultural difference in judges' attitudes toward pretrial detention in Dane County versus the southeastern county Copp et al. study; Dane County judges may adopt a more reform-minded approach to interpreting PSA outcomes, being skeptical of how these tools may disproportionately impact minority populations, whereas the judges studied by Copp et al. may be less progressive, viewing the algorithmic recommendation as an excessively-lenient baseline recommendation. Ultimately, it is a normative question of whether the equalization in override rates between racial and racial-gender groups is a positive outcome. Similarly, judges' increased lenience towards female defendants leads to more gender disparity in pre-trial outcomes, but whether this inequality is inherently problematic once again remains a normative question.

Further research in the field should continue to explore judicial override behavior through qualitative and quantitative studies of how judges weigh and consider algorithmic predictions in their decisions and why judges may choose to override the algorithm. Focus groups and conversations with judges may lend credence to my data analysis which suggests some systematic behavior when judges choose to override the algorithm. Another interesting area for further exploration is how judicial usage of the PSA changes over time. The data from Dane County was collected early on in the county's implementation of the PSA. Seven years later, Dane County judges could be using PSA outputs differently, now that they have become more familiarized with the tool. It would also be insightful to consider under what circumstances judicial override of the algorithm is more likely to occur. For example, perhaps during busier periods of court schedules, judges may be more likely to defer to the algorithm due to well-documented judicial mental depletion (Danziger et al. 6889, 2011). Similarly, judges who have used the PSA for several years may be more comfortable with and inclined to override its recommendations. All of these areas for further research expand upon the ever-growing body of research around algorithm-assisted judicial decision making.

My conclusions yield several policy implications. While use of the PSA may not affect accuracy of judicial decision making—defined as correctly predicting failures to appear or new criminal activity—, the PSA does reshape judicial override behavior, reducing racial and gender-racial disparities while exacerbating gender disparities. Policymakers should therefore recognize that judges interpret the PSA through their own cognitive and normative biases, systematically overriding the PSA in deliberate ways. The PSA may also moderate certain pre-existing forms of preferential treatment. Without PSA exposure, white and white male defendants especially enjoyed relatively more lenience from judges in their release decisions. If a policy goal is to reduce historic discrepancies in judge leniency based on identity factors, judges' use of the could help standardize treatment of different demographic groups. However, it is also important to note that judicial override of the PSA increased gender disparities, depicting the need for continuous auditing of how judges are systematically responding to and overriding PSA outputs to ensure the downstream effects of algorithm-assisted judicial decision making are in line with normative goals of justice. Finally, the lack of accuracy improvement from the PSA suggests that policymakers must reconsider the stated purpose of adopting the PSA. The way the PSA is currently used by judges does not seem to reduce crime rates in a meaningful way, even if the algorithm could help correct for disparities in judicial leniency.

Bibliography

Angelova, Victoria, Will S Dobbie, and Crystal Yang. Algorithmic Recommendations and Human Discretion. *NBER WORKING PAPER SERIES*. n.d.

Ben-Michael, Eli, D. James Greiner, Melody Huang, Kosuke Imai, Zhichao Jiang, and Sooahn Shin. "Does AI Help Humans Make Better Decisions? A Statistical Evaluation Framework for Experimental and Observational Studies." *Proceedings of the National Academy of Sciences* 122, no. 38 (2025): e2505106122. https://doi.org/10.1073/pnas.2505106122.

Copp, Jennifer E., William Casey, Thomas G. Blomberg, and George Pesta. "Pretrial Risk Assessment Instruments in Practice: The Role of Judicial Discretion in Pretrial Reform." *Criminology & Public Policy* 21, no. 2 (2022): 329–58. https://doi.org/10.1111/1745-9133.12575.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences* 108, no. 17 (2011): 6889–92. https://doi.org/10.1073/pnas.1018033108.

Federal Register. "Annual Determination of Average Cost of Incarceration Fee (COIF)." December 6, 2024. https://www.federalregister.gov/documents/2024/12/06/2024-28743/annual-determination-of-average-cost-of-incarceration-fee-coif.

Frank-Loron, Rhonda. *Pretrial Services Updates*. n.d.

Green, Ben, and Yiling Chen. "Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (2021): 1–33. https://doi.org/10.1145/3479562.

Hamilton, Melissa. "Evaluating Algorithmic Risk Assessment." *New Criminal Law Review* 24, no. 2 (2021): 156–211. https://doi.org/10.1525/nclr.2021.24.2.156.

Imai, Kosuke, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. *Experimental Evaluation of Algorithm- Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment*. n.d.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 133, no. 1 (2018): 237–93. https://doi.org/10.1093/qje/qjx032.

Koepke, John Logan, and David G Robinson. *Danger Ahead: Risk Assessment and the Future of Bail Reform*. n.d.

Mattu, Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya. “Machine Bias.” *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Moore, Cristopher, Elise Ferguson, and Paul Guerin. “Pretrial Risk Assessment on the Ground: Algorithms, Judgments, Meaning, and Policy.” *MIT Case Studies in Social and Ethical Responsibilities of Computing*, no. Summer 2023 (August 2023). https://doi.org/10.21428/2c646de5.b016a7b3.

Sloan, Carly Will, George S Naufal, and Heather Caspers. *The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes*. 2018.

Solow-Niederman, Alicia; Choi, YooJung; Van Den Broeck, Guy. *The Institutional Life of Algorithmic Risk Assessment*. Berkeley Technology Law Journal, 2019. https://doi.org/10.15779/Z38WD3Q226.

Stevenson, Megan T. “Assessing Risk Assessment in Action.” *SSRN Electronic Journal*, ahead of print, 2017. https://doi.org/10.2139/ssrn.3016088.

Young, Delia, and Shar-Ron Buie. *Office of Justice Reform Adds Two Staff*. n.d.