Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment[*]

Kosuke Imai[†] Zhichao Jiang[‡] D. James Greiner[§] Ryan Halen[¶] Sooahn Shin[‖]

First Draft: July 9, 2020
This Draft: December 8, 2020

Abstract

Despite an increasing reliance on fully-automated algorithmic decision making in our day-to-day lives, human beings still make highly consequential decisions. As frequently seen in business, healthcare, and public policy, recommendations produced by algorithms are provided to human decision-makers in order to guide their decisions. While there exists a fast-growing literature evaluating the bias and fairness of such algorithmic recommendations, an overlooked question is whether they help humans make better decisions. We develop a statistical methodology for experimentally evaluating the causal impacts of algorithmic recommendations on human decisions. We also show how to examine whether algorithmic recommendations improve the fairness of human decisions and derive the optimal decisions under various settings. We apply the proposed methodology to the first-ever randomized controlled trial that evaluates the pretrial Public Safety Assessment (PSA) in the criminal justice system. A goal of the PSA is to help judges decide which arrested individuals should be released. We find that the PSA provision has little overall impact on the judge's decisions and subsequent arrestee behavior. However, our analysis provides some potentially suggestive evidence that the PSA may help avoid unnecessarily harsh decisions for female arrestees regardless of their risk levels while it encourages the judge to make stricter decisions for male arrestees who are deemed to be risky. In terms of fairness, the PSA appears to increase the gender bias against males while having little effect on the existing racial biases of the judge's decisions against non-white males. Finally, we find that the PSA's recommendations might be too severe unless the cost of a new crime is sufficiently higher than the cost of a decision that may result in an unnecessary incarceration.

Keywords: algorithmic fairness, causal inference, principal stratification, randomized experiments, recommendation systems, sensitivity analysis

[*] We thank seminar participants at Harvard University, Institute for Advanced Study, Microsoft Research (New England), and Princeton University for helpful comments.

[†] Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA 02138. Email: imai@harvard.edu URL: https://imai.fas.harvard.edu

[‡] Assistant Professor, Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst MA 01003. Email: zhichaojiang@umass.edu

[§] Honorable S. William Green Professor of Public Law, Harvard Law School, 1525 Massachusetts Avenue, Griswold 504, Cambridge, MA 02138.

[¶] Data Analyst, Access to Justice Lab at Harvard Law School, 1607 Massachusetts Avenue, Third Floor, Cambridge, MA 02138.

[‖] Ph.D. student, Department of Government, Harvard University, Cambridge, MA 02138. Email: sooahnshin@g.harvard.edu URL: https://sooahnshin.com

1 Introduction

A growing body of literature has suggested the potential superiority of algorithmic decision making over purely human choices across a variety of tasks (e.g., Hansen and Hasan, 2015; He et al., 2015). Although some of this evidence is decades old (e.g., Dawes, Faust and Meehl, 1989), it has recently gained significant public attention by the spectacular defeats of humanity's best in cerebral games (e.g., Silver et al., 2018). Yet, even in contexts where research has warned of human frailties, we humans still make many important decisions to give ourselves agency and to be held accountable for highly consequential choices.

The desire for a human decision maker as well as the precision and efficiency of algorithms has led to the adoption of hybrid systems involving both. By far the most popular system uses algorithmic recommendations to inform human decision making. Such algorithm-assisted human decision making has been deployed in many aspects of our daily lives, including medicine, hiring, credit lending, investment decisions, and online shopping to name a few. And of particular interest, algorithmic recommendations are increasingly of use in the realm of evidence-based public policy making. A prominent example, studied in this paper, is the use of risk assessment instruments in the criminal justice system that are designed to improve incarceration rulings and other decisions made by judges.

While there exists a fast-growing literature in computer science that studies the bias and fairness of algorithms (see Chouldechova and Roth, 2020, for a review and many references therein), an overlooked question is whether such algorithms help human make better decisions (see e.g., Green and Chen, 2019, for an exception). In this paper, we develop a methodological framework for experimentally evaluating the impacts of algorithmic recommendations on human decision making. We conduct the first-ever real-world field experiment by providing, for randomly selected cases, a pretrial Public Safety Assessment score (PSA) to a judge who makes an initial release decision for randomly selected cases. We evaluate whether the PSA helps judges achieve their goal of preventing arrestees from committing a new crime or failing to appear in court while avoiding an unnecessarily harsh decision.

Using the concept of principal stratification from the causal inference literature (e.g., Frangakis and Rubin, 2002; Ding and Lu, 2017), we propose the evaluation quantities of interest, identification assumptions, and estimation strategies. We also develop a Bayesian sensitivity analysis to assess the robustness of empirical findings to the potential violation of a key identification assumption (see also Hirano et al., 2000; Schwartz, Li and Mealli, 2011; Mattei et al., 2013; Jiang, Ding and Geng, 2016). In addition, we also examine whether algorithmic recommendations improve the fairness of human decisions, using the concept of principal fairness that, unlike other fairness criteria, accounts for how the decision in question affects individuals (Imai and Jiang, 2020). Finally, we consider how the data from experimental evaluation can be used to inform an optimal decision rule and assess the optimality of algorithmic recommendations and human decisions. Although the proposed methodology is described and applied in the context of evaluating the PSA, it is directly applicable or at least extendable to many other settings of algorithm-assisted human decision making.

The PSA, which serves as the main application of the current paper, has played a prominent role in the literature on algorithmic fairness at least since the controversy over the potential racial bias of COMPAS risk assessment score used in the United States (US) criminal justice system (see e.g., Angwin et al., 2016; Dieterich, Mendoza and Brennan, 2016; Flores, Bechtel and Lowenkamp, 2016; Dressel and Farid, 2018). However, with few exceptions, much of this debate focused upon the accuracy and fairness properties of the PSA itself rather than how the PSA affects the decisions by judges (see e.g., Berk et al., 2018; Kleinberg et al., 2018; Rudin, Wang and Coker, 2020, and references therein). Even the studies that directly estimate the impacts of the PSA on judges' decisions are based on either observational data or hypothetical vignettes in surveys (e.g., Miller and Maloney, 2013; Berk, 2017; Stevenson, 2018; Albright, 2019; Green and Chen, 2019; Garrett and Monahan, 2020; Skeem, Scurich and Monahan, 2020).

We contribute to this literature by demonstrating how to experimentally evaluate the PSA. To the best of our knowledge, this is the first randomized controlled trial (RCT) that evaluates the impacts of modern risk assessment scores on judges' decisions in the criminal justice system (see also the 1981–82 Philadelphia Bail Experiment that evaluated the effects of a bail guideline on judges' decisions rather than those of risk assessment scores (Goldkamp and Gottfredson, 1984, 1985)). The proposed methodology allows us to evaluate the effects of the PSA on judges' decisions separately for the subgroups of arrestees with different levels of risks.

We find that the provision of the PSA has little overall impact on the judge's decisions across three outcomes we examine: failure to appear (FTA), new criminal activity (NCA), and new violent criminal activity (NVCA). However, our analysis provides some suggestive evidence that the PSA may make the judge's decisions more lenient for female arrestees regardless of their risk levels, while it encourages the judge to make stricter decisions for male arrestees who are deemed to be risky. In terms of fairness, the PSA appears to increase the gender bias against males while having no substantial impact on the existing racial bias of the judge's decisions against non-white males.

Finally, we use the experimental data to learn about the optimal decision that minimizes the negative outcomes (FTA, NCA, and NVCA) while avoiding unnecessarily harsh decisions. Our analysis suggests that the PSA's recommendations may be too severe unless the cost of negative outcomes is much higher than the cost of a decision that may result in an unnecessary incarceration.

2 Experimental Evaluation of Pretrial Public Safety Assessment

In this section, we briefly describe our field experiment after providing some background about the use of PSA in the US criminal justice system. Additional details about our experiment are given elsewhere (Greiner et al., 2020).

2.1 Background

The US criminal justice apparatus consists of thousands of diverse systems. Some are similar in the decision points they feature as an individual suspected of a crime travels from investigation to sentencing. Common decision points include whether to stop and frisk an individual in a public place, whether to arrest or issue a citation to an individual suspect of committing a crime, whether to release the arrestee while they await the disposition of any charges against them (the subject of this paper), the charge(s) to be filed against the individual, whether to find the defendant guilty of those charges, and what sentence to impose on a defendant found guilty.

At present, human judges make all of these decisions. In theory, algorithms could inform any of them, and could even make some of these decisions without human involvement. To date, algorithmic outputs have appeared most frequently in two settings: (i) at the "first appearance" hearing, during which a judge decides whether to release an arrestee pending disposition of any criminal charges, and (ii) at sentencing, in which the judge imposes a punishment on a defendant found guilty. The first of these two motivates the present paper, but the proposed methodology is applicable or extendable to other settings.

We describe a typical first appearance hearing. The key decision the judge must make at a first appearance hearing is whether to release the arrestee pending disposition of any criminal charges and, if the arrestee is to be released, what conditions to impose. Almost all jurisdictions allow the judge to release the arrestee with only a promise to reappear at subsequent court dates. In addition, most arrestee has not yet been adjudicated guilty of any charge at the time of a pretrial hearing, there exists a consensus that pretrial incarceration is to be avoided unless the cost is sufficiently high.

Judges deciding whether to release an arrestee ordinarily consider two risk factors among a variety of other concerns; the risk that the arrestee will fail to appear (FTA) at subsequent court dates, and the risk that the arrestee will engage in new criminal activity (NCA) before the case is resolved (e.g., 18 U.S.C. § 3142(e)(1)). Jurisdiction laws vary regarding how these two risks are to be weighed. Some jurisdictions direct judges to consider both simultaneously along with other factors (e.g., Ariz. Const. art. II, § 22, Iowa Code § 811.2(1)(a)), while others focus on only FTA risk (e.g., N.Y. Crim. Proc. Law § 510.30(2)(a)). Despite these variations, NCA or FTA are constant and prominent in the debate over the first appearance decisions.

The concerns about the consequential nature of the first appearance decision have led to the development of PSA, which is ordinarily offered as an input to first appearance judges. PSA can take various forms, but most focuses on classifying arrestees according to FTA and NCA risks. PSA is generally derived by fitting a statistical model to a training dataset based on the past observations from first appearance hearings and the subsequent incidences (or lack thereof) of FTA and NCA. The hope is that providing PSA will improve the assessment of FTA and NCA risks and thereby lead to better decisions. The goal of this paper is to develop a methodological framework for evaluating the impact of PSA on judges' decisions using an RCT, to which we now turn.

2.2 The Experiment

We conducted a field RCT in Dane county, Wisconsin, to evaluate the impacts of a PSA on judges' decisions. The PSA used in our RCT consists of three scores — two six-point scores separately summarizing FTA and NCA risks as well as a binary score for the risk of NVCA. These scores are based on the weighted indices of nine factors drawn from criminal history information, primarily prior convictions and FTA, and a single demographic factor, age. Notably, gender and race are not used to compute the PSA. The weights are calculated using past data. The details about the construction of the PSA and other relevant information are available at `https://advancingpretrial.org/psa/factors/`

The field operation was straightforward. In this county, a court employee assigned each matter a case number sequentially as matters entered the system. No one but this clerk was aware of the pending matter numbers, so manipulation of the number by charging assistant district attorneys was not possible. Employees of the Clerk's office scanned online record systems to calculate the PSA for all cases. If the last digit of case number was even, these employees made the PSA available to the judge at the first appearance hearing. Otherwise, no PSA was made available. Thus, the provision of PSA to judges was essentially randomized. Indeed, the comparison of observed covariate distributions suggests that this scheme produced groups comparable on background variables.

The judge presiding over the first appearance hearing by law was to consider the risk of FTA and NCA, along with other factors including ties to the community as prescribed by statute. The judge could order the arrestee released with or without bail of varying amount. The judge could also condition release on compliance with certain conditions such as certain levels of monitoring, but for the sake of simplicity, we focus on bail decisions and ignore other conditions in this paper.

When making decisions, the judge also had information other than the PSA and its inputs. In all cases, the judge had a copy of an affidavit sworn to by a police officer recounting the circumstances of the incident that led to the arrest. The defense attorney sometimes informed the judge of the following regarding the arrestee's connections to the community: length of time lived there, employment there, and family living there. When available, this information ordinarily stems from an arrestee interview conducted earlier by a paralegal. The assistant district attorney sometimes provided additional information regarding the circumstances of the arrest or criminal history. Given the lack of access to this additional information, we will develop a sensitivity analysis to address a potential unobserved confounding bias.

2.3 The Data

The field operation design called for approximately a 30-month treatment assignment period (from the middle of 2017 until the end of 2019) followed by the collection of data on FTA, NCA, NVCA, and other outcomes for a period of two years after randomization. At the time of this writing, the outcome data had been collected for a period of 24 months for the arrestees who were involved in arrest events during the first 12 months. We plan to continue our field RCT and report the results of our comprehensive analysis of a full data set in the future. Furthermore, although some arrestees had multiple cases during the study period, this paper focuses on the first arrest cases in order to avoid potential spillover effects across cases. This leads to a total of 1890 cases for our analysis, of which 40.1% (38.8%) involve white (non-white) male arrestees and 13.0% (8.1%) are white (non-white) female arrestees.

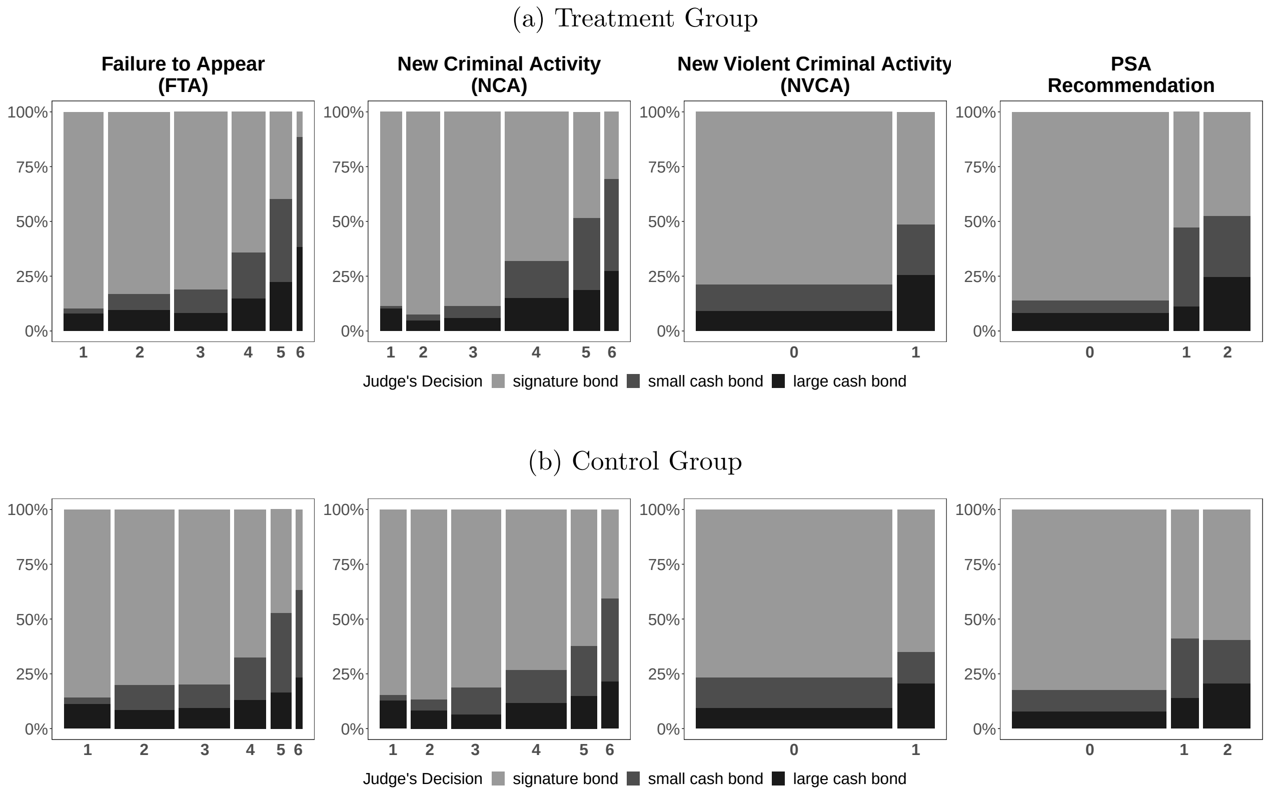
Based on the empirical distribution of bail amount and expert's opinion, we categorize the judge's decisions into three ordinal categories: signature bond, small cash bond (less than $1,000), and large cash bond (greater than or equal to $1,000). The signature bond requires an arrestee to sign a promise to return to the court for trial, but does not require any payment to be released. Table 1 summarizes the joint distribution of treatment assignment (PSA provision), the judge's decisions (three ordinal categories), and three binary outcomes. We observe that in about three quarters of cases the judge imposed signature bonds, while in the remaining cases the judge imposed bail.

	no PSA (Control Group)			PSA (Treatment Group)			
	Signature	Cash bond		Signature	Cash bond		
	bond	≤$1000	>$1000	bond	≤$1000	>$1000	Total (%)
Non-white Female	64	11	6	67	6	0	154
	(3.4)	(0.6)	(0.3)	(3.5)	(0.3)	(0.0)	(8.1)
White Female	91	17	7	104	17	10	246
	(4.8)	(0.9)	(0.4)	(5.5)	(0.9)	(0.5)	(13.0)
Non-white Male	261	56	49	258	53	57	734
	(13.8)	(3.0)	(2.6)	(13.6)	(2.8)	(3.0)	(38.8)
White Male	289	48	44	276	54	46	757
	(15.3)	(2.5)	(2.3)	(14.6)	(2.9)	(2.4)	(40.0)
FTA committed	218	42	16	221	45	16	558
	(11.5)	(2.2)	(0.8)	(11.7)	(2.4)	(0.8)	(29.4)
not committed	487	90	90	484	85	97	1333
	(25.8)	(4.8)	(4.8)	(25.6)	(4.5)	(5.1)	(70.6)
NCA committed	211	39	14	202	40	17	523
	(11.2)	(2.1)	(0.7)	(10.7)	(2.1)	(0.9)	(27.7)
not committed	494	93	92	503	90	96	1368
	(26.1)	(4.9)	(4.9)	(26.6)	(4.8)	(5.1)	(72.4)
NVCA committed	36	10	3	44	10	6	109
	(1.9)	(0.5)	(0.2)	(2.3)	(0.5)	(0.3)	(5.7)
not committed	669	122	103	661	120	107	1782
	(35.4)	(6.5)	(5.4)	(35.0)	(6.3)	(5.7)	(94.3)
Total	705	132	106	705	130	113	1891
	(37.3)	(7.0)	(5.6)	(37.3)	(6.9)	(6.0)	(100)

Table 1: The Joint Distribution of Treatment Assignment, Judge's Decisions, and Outcomes. The table shows the number of cases in each category with the corresponding percentage in parentheses. Only about 20% of all arrestees are female. Few cases result in NVCA (violent new criminal activity), while FTA (failure to appear in court) and NCA (new criminal activity) account for slightly above 25% each. A majority of decisions are signature bonds rather than cash bonds.

For the outcome variables, slightly less than 30% of arrestees commit FTA and NCA whereas the proportion of those who commit NVCA is only about 5%.

Figure 1 presents the distribution of the judge's decisions given each of the PSA scores among the cases in the treatment (top panel) and control (bottom panel) groups, to which the PSA scores are provided. The overall difference in the conditional distribution between the two groups is small though there are some differences in some subgroups (see Appendix S1). The PSA scores for FTA and NCA are ordinal, ranging from 1 (safest) to 6 (riskiest), whereas the PSA score for NVCA is binary, 0 (safe) and 1 (risky). There is also the overall PSA recommendation which is a three-category ordinal variable aggregating these three PSA scores. The PSA recommendation is 0 (a



(a) Treatment Group

(b) Control Group

Figure 1: The Distribution of the Judge's Decisions given the Pretrial Public Safety Assessment (PSA) among the Cases in the Treatment (Top Panel) and Control (Bottom Panel) Groups. There are three PSA scores, two of which are ordinal — FTA and NCA — while the other is dichotomous — NVCA. The judge's decision is coded as a three-category ordinal variable based on the type and amount of bail: a signature bond, a small cash bond (less than $1,000), and a large cash bond (greater than or equal to $1,000). The overall PSA recommendation is also coded as a three category ordinal variable: 0 (a signature bond), 1 (a small cash bond), and 2 (a large cash bond). The width of each bar is proportional to the number of cases for each value of the corresponding PSA score. There exists a positive correlation between PSA scores and the severity of the judge's decisions in both treatment and control groups.

signature bond) if the FTA and NCA scores are less than or equal to 4, and NVCA flag equals to 0, while the recommendation is 2 (a large cash bond) if the NVCA flag equals 1 or either the FTA or NCA scores is equal to 6. The remaining case is coded as the PSA recommendation of 1 (a small cash bond).

In general, we observe a positive association between the PSA scores and judge's decisions, implying that a higher PSA score is associated with a harsher decision. We also find that for FTA and NCA, the most likely scores are in the medium range, while the vast majority of NVCA cases were classified as no elevated risk. Finally, for NCA and FTA, the judge's decisions varied little when the PSA score took a value in the lower range. For the overall PSA recommendation, the judge is far more likely to give a signature bond for the cases that are actually recommended for a signature

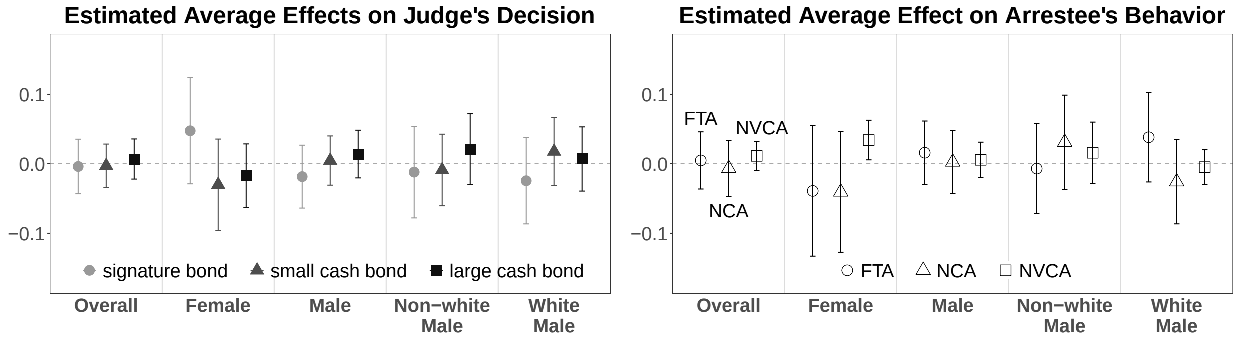


Figure 2: Estimated Average Causal Effects of PSA Provision on the Judge's Decisions and Outcome Variables. The results are based on the difference-in-means estimator. The vertical bars represent the 95% confidence intervals. In the left plot, we report the estimated effects of PSA provision on the judge's decision to charge a signature bond (solid circles), a small cash bail ($1,000 dollars or less; solid triangles), and a large cash bail (greater than $1,000; solid squares). In the right plot, we report the estimated effects of PSA provision on the three different outcome variables: FTA (open circles), NCA (open triangles), and NVCA (open squares). The PSA provision appears to have little overall effect on the judge's decision and arrestee's behavior, on average, though it may slightly increase NVCA among female arrestees.

bond.

Figure 2 presents the estimated average causal effect of the PSA provision on the judge's decisions (left plot) and three outcomes of interest (right plot). We use the difference-in-means estimator and display the 95% confidence intervals as well as the point estimates. We do not compute the separate estimates for white females and non-white females because we have too few female arrestees (see Table 1). The results imply that the PSA provision, on average, has little effect on the judge's decisions. In addition, the average effects of the PSA provision on the three outcomes are also largely ambiguous although there is a suggestive evidence that it may slightly increase NVCA among female arrestees. In Appendix S2.1, we also explore the average causal effects of the PSA provision across different age groups. We find some suggestive causal effects for the group of 29 – 35 years old arrestees.

Although these results show whether the PSA provision leads to a harsher or more lenient decision (and whether it increases or decreases negative outcomes), they are not informative about whether it helps judges make better decisions. In the current context, a primary goal of the judge is to make a less lenient decision on risky cases. Therefore, if the PSA is helpful, its provision should encourage the judge to impose small or no bail on safe cases and impose a greater amount of bail on risky cases. This demands the study of an important causal heterogeneity by distinguishing between cases with different risks. In addition, we may also be interested in knowing how the PSA provision affects the gender and racial fairness of judges' decision. Our goal is to develop statistical methods that directly address these and other questions.

3 The Proposed Methodology

In this section, we describe the proposed methodology for experimentally evaluating the impacts of algorithmic recommendations on human decision-making. Although we refer to our specific application throughout, the proposed methodology can be applied or extended to other settings, in which humans make decisions using algorithmic recommendations as an input. We will begin by considering a binary decision and then extend our methodology to an ordinal decision in Section 3.4.

3.1 The Setup

Let Z_i be a binary treatment variable indicating whether the PSA is presented to the judge of case $i = 1, 2, \ldots, n$. We use D_i to denote the binary detention decision made by the judge to either detain ($D_i = 1$) or release ($D_i = 0$) the arrestee prior to the trial. In addition, let Y_i represent the binary outcome. All of the outcomes in our application — NCA, NVCA, and FTA — are binary variables. For example, $Y_i = 1$ ($Y_i = 0$) implies that the arrestee of case i commits (does not commit) an NCA. Finally, we use $\mathbf{X}_i$ to denote a vector of observed pre-treatment covariates for case i. They include age, gender, race, and prior criminal history.

We adopt the potential outcomes framework of causal inference and assume the stable unit treatment value assumption (SUTVA) (Rubin, 1990). In particular, we assume no interference among cases, implying that the treatment assignment for one case does not influence the judge's decision and outcome variable in another case. Note that this assumption is reasonable in our application since we focus only on first arrests and do not analyze cases with subsequent arrests.

Let $D_i(z)$ be the potential value of the pretrial detention decision if case i is assigned to the treatment condition $z \in \{0, 1\}$. Furthermore, $Y_i(z, d)$ represents the potential outcome under the scenario, in which case i is assigned to the treatment condition z and the judge makes the decision $d \in \{0, 1\}$. Then, the observed decision is given by $D_i = D(Z_i)$ whereas the observed outcome is denoted by $Y_i = Y_i(Z_i, D_i(Z_i))$.

Throughout this paper, we maintain the following three assumptions, all of which we believe are reasonable in our application. First, because the treatment assignment is essentially randomized, the following independence assumption is automatically satisfied.

Assumption 1 (Randomization of the Treatment Assignment)

$$\{D_i(z), Y_i(z, d), \mathbf{X}_i\} \perp\!\!\!\perp Z_i \quad \text{for } z \in \{0, 1\} \text{ and all } d.$$

Second, we assume that the provision of the PSA influences the outcome only through the judge's decision. Because an arrestee would not care and, perhaps, would not even know whether the judge is presented with the PSA at their first appearance, it is reasonable to assume that their behavior, be it NCA, NVCA, or FTA, is not affected directly by the treatment assignment.

Assumption 2 (Exclusion Restriction)

$$Y_i(z, d) \ = \ Y_i(z', d) \quad \text{for } z, z' \in \{0, 1\} \text{ and all } i, d.$$

Under Assumption 2, we can simplify our notation by writing $Y_i(z, d)$ as $Y_i(d)$. A potential violation of this assumption is that the PSA may directly influence the judge's decision about release conditions, which can in turn affect the outcome. The extension of the proposed methodology to multi-dimensional decisions is left for future research.

Finally, we assume that the judge's decision monotonically affects the outcome. Thus, for NCA (NVCA), the assumption implies that each arrestee is no less likely to commit a new (violent) crime if released. If FTA is the outcome of interest, this assumption implies that an arrestee is no more likely to appear in court if released. The assumption is reasonable since being held in custody of a court makes it difficult, if not impossible, to engage in NCA, NVCA, and FTA.

Assumption 3 (Monotonicity)

$$Y_i(1) \ \leq \ Y_i(0) \quad \text{for all } i.$$

3.2 Causal Quantities of Interest

We define causal quantities of interest using principal strata that are determined by the joint values of potential outcomes, i.e., $(Y_i(1), Y_i(0)) = (y_1, y_0)$ where $y_1, y_0 \in \{0, 1\}$ (Frangakis and Rubin, 2002). Since Assumption 3 eliminates one principal stratum, $(Y_i(1), Y_i(0)) = (1, 0)$, there are three remaining principal strata. The stratum $(Y_i(1), Y_i(0)) = (0, 1)$ consists of those who would engage in NCA (NVCA or FTA) only if they are released. We call members of this stratum as "preventable cases" because keeping those arrestees in custody would prevent the negative outcome (NCA, NVCA, or FTA). The stratum $(Y_i(1), Y_i(0)) = (1, 1)$ is called "risky cases," and corresponds to those who always engage in NCA (NVCA or FTA) regardless of the judge's decision. In contrast, the stratum $(Y_i(1), Y_i(0)) = (0, 0)$ represents "safe cases," in which the arrestees would never engage in NCA (NVCA or FTA) regardless of the detention decision.

We are interested in examining how the PSA provision influences judges' detention decisions across different types of cases. We are interested in the following three average principal causal effects (APCE),

$$\begin{aligned}
\mathsf{APCEp} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 0, Y_i(0) = 1\}, \qquad (1) \\
\mathsf{APCEr} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 1, Y_i(0) = 1\}, \qquad (2) \\
\mathsf{APCEs} &= \mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(1) = 0, Y_i(0) = 0\}. \qquad (3)
\end{aligned}$$

If the PSA is helpful, its provisions should make judges more likely to detain the arrestees of the preventable cases. That is, the principal causal effect on the detention decision for the preventable cases (APCEp) should be positive. In addition, the PSA should encourage judges to release the arrestees of the safe cases, implying that the principal causal effect for the safe cases (APCEs) should be negative. The desirable direction of the principal causal effect for risky cases (APCEr) depends on various factors including the societal costs of holding the arrestees of this category in custody.

3.3 Nonparametric Identification

We consider the nonparametric identification of the principal causal effects defined above. The following theorem shows that under the aforementioned assumptions, these effects can be identified up to the marginal distributions of $Y_i(d)$ for $d = 0, 1$.

Theorem 1 (Identification) *Under Assumptions 1, 2, and 3,*

$$\begin{aligned}
\mathit{APCEp} &= \frac{\Pr(Y_i = 1 \mid Z_i = 0) - \Pr(Y_i = 1 \mid Z_i = 1)}{\Pr\{Y_i(0) = 1\} - \Pr\{Y_i(1) = 1\}}, \\
\mathit{APCEr} &= \frac{\Pr(D_i = 1, Y_i = 1 \mid Z_i = 1) - \Pr(D_i = 1, Y_i = 1 \mid Z_i = 0)}{\Pr\{Y_i(1) = 1\}}, \\
\mathit{APCEs} &= \frac{\Pr(D_i = 0, Y_i = 0 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 1)}{1 - \Pr\{Y_i(0) = 1\}}.
\end{aligned}$$

Proof is given in Appendix S3.2. Because $\Pr\{Y_i(d)\}$ is not identifiable without additional assumptions, we cannot estimate the causal effects based on Theorem 1. However, the denominators of the expressions on the right-hand side of Theorem 1 are positive under Assumption 3. As a result, the signs of the causal effects are identified from Theorem 1, which allows us to draw qualitative conclusions.

In addition, the theorem implies, for example, that the sign of APCEp is the opposite of the sign of the average causal effect on the outcome. This is intuitive because if the provision of the PSA increases the probability of NCA (NVCA or FTA), then the judge must have released more arrestees for preventable cases.

Furthermore, we can obtain the nonparametric bounds on these causal quantities by bounding

$\Pr\{Y_i(d) = y\}$ that appears in the denominators. By the law of total probability,

$$\Pr\{Y_i(d) = 1\} \quad = \quad \Pr\{Y_i = 1 \mid D_i = d\} \Pr(D_i = d) + \Pr\{Y_i(d) = 1 \mid D_i = 1 - d\} \Pr(D_i = 1 - d)$$

for $d = 0, 1$. Then, the bounds on $\Pr\{Y(d) = 1\}$ are obtained by replacing $\Pr\{Y_i(d) = 1 \mid D_i = 1-d\}$ with 0 or 1. However, these bounds may be too wide to be informative.

For point identification, we consider the following unconfoundedness assumption, which states that conditional on a set of observed pre-treatment covariates $\mathbf{X}_i$ and the PSA provision, the judge's decision is independent of the potential outcomes.

Assumption 4 (Unconfoundedness)

$$Y_i(d) \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z \quad \text{for } z \in \{0, 1\} \text{ and all } d.$$

Assumption 4 holds if $\mathbf{X}_i$ contains all the information judges have access to when making the detention decision under each treatment condition. However, as noted in Section 2.2, a judge may receive and use additional information regarding whether the arrestee has a job or a family in the jurisdiction, or perhaps regarding the length of time the arrestee has lived in the jurisdiction. If these factors have an impact on both judge's decisions and arrestee's behaviors, then the assumption is unlikely to be satisfied. Later, we address this issue by developing a sensitivity analysis for the potential violation of Assumption 4 (see Section 3.5).

To derive the identification result, consider the following principal scores (Ding and Lu, 2017), which represent in our application the population proportion (conditional on $\mathbf{X}_i$) of preventable, risky, and safe cases, respectively,

$$\begin{aligned} e_P(\mathbf{x}) &= \Pr\{Y_i(1) = 0, Y_i(0) = 1 \mid \mathbf{X}_i = \mathbf{x}\}, \\ e_R(\mathbf{x}) &= \Pr\{Y_i(1) = 1, Y_i(0) = 1 \mid \mathbf{X}_i = \mathbf{x}\}, \\ e_S(\mathbf{x}) &= \Pr\{Y_i(1) = 0, Y_i(0) = 0 \mid \mathbf{X}_i = \mathbf{x}\}. \end{aligned}$$

Under Assumptions 2, 3, and 4, we can identify the principal scores as,

$$\begin{aligned} e_P(\mathbf{x}) &= \Pr\{Y_i = 1 \mid D_i = 0, \mathbf{X}_i = \mathbf{x}\} - \Pr\{Y_i = 1 \mid D_i = 1, \mathbf{X}_i = \mathbf{x}\}, \\ e_R(\mathbf{x}) &= \Pr\{Y_i = 1 \mid D_i = 1, \mathbf{X}_i = \mathbf{x}\}, \\ e_S(\mathbf{x}) &= \Pr\{Y_i = 0 \mid D_i = 0, \mathbf{X}_i = \mathbf{x}\}. \end{aligned}$$

Then, the next theorem shows that we can identify the APCE as the difference in the weighted average of judge's decisions between the treatment and control groups.

Theorem 2 (Identification under Unconfoundedness) *Under Assumptions 1, 2, 3 and 4, APCEp, APCEr and APCEs are identified as,*

$$\begin{aligned} \textsf{\textit{APCEp}} &= \mathbb{E}\{w_P(\mathbf{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_P(\mathbf{X}_i)D_i \mid Z_i = 0\}, \\ \textsf{\textit{APCEr}} &= \mathbb{E}\{w_R(\mathbf{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_R(\mathbf{X}_i)D_i \mid Z_i = 0\}, \\ \textsf{\textit{APCEs}} &= \mathbb{E}\{w_S(\mathbf{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_S(\mathbf{X}_i)D_i \mid Z_i = 0\}, \end{aligned}$$

where

$$w_P(\mathbf{x}) = \frac{e_P(\mathbf{x})}{\mathbb{E}\{e_P(\mathbf{X}_i)\}}, \quad w_R(\mathbf{x}) = \frac{e_R(\mathbf{x})}{\mathbb{E}\{e_R(\mathbf{X}_i)\}}, \quad w_S(\mathbf{x}) = \frac{e_S(\mathbf{x})}{\mathbb{E}\{e_S(\mathbf{X}_i)\}}.$$

Proof is given in Appendix S3.2. We note that Ding and Lu (2017) also identify principal causal effects using principal scores. However, they consider principal strata based on the endogenous variable, whereas we are interested in the causal effects on the endogenous variable within each principal strata defined by the values of the potential outcome.

In some situations, we might consider the following strong monotonicity assumption instead of Assumption 3.

Assumption 5 (Strong Monotonicity)

$$Y_i(1) \ = \ 0 \quad \text{for all } i.$$

The assumption implies that the detention decision prevents FTA, NCA, or NVCA. The assumption is plausible for FTA, but may not hold for NCA/NVCA in some cases. In our data, for example, we find some NCA and NVCA among the incarcerated arrestees.

Under Assumption 5, the risky cases do not exist and hence the APCEr is not defined, and the APCEp simplifies to $\mathbb{E}\{D_i(1) - D_i(0) \mid Y_i(0) = 0\}$. This leads to the following identification result.

Theorem 3 (Identification under Strong Monotonicity) *Under Assumptions 1, 2, and 5,*

$$\begin{aligned} \textsf{\textit{APCEp}} &= \frac{\Pr(D_i = 0, Y_i = 1 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 1 \mid Z_i = 1)}{\Pr\{Y_i(0) = 1\}}, \\ \textsf{\textit{APCEs}} &= \frac{\Pr(D_i = 0, Y_i = 0 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 1)}{\Pr\{Y_i(0) = 0\}}. \end{aligned}$$

Proof is given in Appendix S3.4. As in Theorem 1, the APCEp and APCEs depend on the distribution of $Y_i(0)$, which is not identifiable. However, as before, the sign of each effect is identifiable.

For point identification, we invoke the unconfoundedness assumption. Note that under the strong monotonicity assumption, Assumption 4 is equivalent to a weaker conditional independence relation concerning only one of the two potential outcomes,

$$Y_i(0) \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$$

for $z = 0, 1$. We now present the identification result.

THEOREM 4 (IDENTIFICATION UNDER UNCONFOUNDEDNESS AND STRONG MONOTONICITY) *Under Assumptions 1, 2, 4 and 5,*

$$\begin{aligned} \mathsf{APCEp} &= \mathbb{E}\{w_P(\mathbf{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_P(\mathbf{X}_i)D_i \mid Z_i = 0\}, \\ \mathsf{APCEs} &= \mathbb{E}\{w_S(\mathbf{X}_i)D_i \mid Z_i = 1\} - \mathbb{E}\{w_S(\mathbf{X}_i)D_i \mid Z_i = 0\}, \end{aligned}$$

where

$$w_P(\mathbf{x}) = \frac{1 - e_S(\mathbf{x})}{\mathbb{E}\{1 - e_S(\mathbf{X}_i)\}}, \quad w_S(\mathbf{x}) = \frac{e_S(\mathbf{x})}{\mathbb{E}\{e_S(\mathbf{X}_i)\}}.$$

Proof is straightforward and hence omitted. The identification formulas are identical to those in Theorem 2. However, with Assumption 5, we can simply compute the principal score as $e_S(\mathbf{x}) = \Pr\{Y_i = 0 \mid D_i = 0, \mathbf{X}_i = \mathbf{x}\}$.

3.4 Ordinal Decision

We generalize the above identification results to an ordinal decision. In our application, this extension is important as the judge's release decision often is based on different amounts of cash bail or varying levels of supervision of an arrestee. We first generalize the monotonicity assumption (Assumption 3) by requiring that a decision with a greater amount of bail is no less likely to make an arrestee engage in NCA (NVCA or FTA). The assumption may be reasonable, for example, because a greater amount of bail is expected to imply a greater probability of being held in custody. The assumption could be violated if arrestees experience financial strain in an effort to post bail, causing them to commit NCA (NVCA or FTA).

Formally, let D_i be an ordinal decision variable where $D_i = 0$ is the least amount of bail, and $D_i = 1, \ldots, k$ represents a bail of increasing amount, i.e., $D_i = k$ is the largest bail amount. Then, the monotonicity assumption for an ordinal decision is given by,

ASSUMPTION 6 (MONOTONICITY WITH ORDINAL DECISION)

$$Y_i(d_1) \ \leq \ Y_i(d_2) \quad \text{for } d_1 \geq d_2.$$

To generalize the principal strata introduced in the binary decision case, we define the decision with the least amount of bail that prevents an arrestee from committing NCA (NVCA or FTA) as follows,

$$R_i \ = \ \begin{cases} \min\{d : Y_i(d) = 0\} & \text{if } Y_i(k) = 0, \\ k + 1 & \text{if } Y_i(k) = 1. \end{cases}$$

We may view R_i as an ordinal measure of risk with a greater value indicating a higher degree of risk. Note that when D_i is binary, R_i takes one of the three values, $\{0, 1, 2\}$, representing safe, preventable, and risky cases, respectively. Thus, R_i generalizes the principal strata to the ordinal case under the monotonicity assumption.

Now, we define the principal causal effects in the ordinal decision case. Specifically, for $r = 1, \ldots, k$ (excluding the cases with $r = 0$ and $r = k+1$), we define the average principal causal effect of the PSA on judge's decisions as a function of this risk measure,

$$\mathsf{APCEp}(r) \;=\; \Pr\{D_i(1) \geq r \mid R_i = r\} - \Pr\{D_i(0) \geq r \mid R_i = r\}. \tag{4}$$

Since the arrestees with $R_i = r$ would not commit NCA (NVCA or FTA) under the decision with $D_i \geq r$, $\mathsf{APCEp}(r)$ represents a reduction in the proportion of NCA (NVCA or FTA) that is attributable to the PSA provision among the cases with $R_i = r$. Thus, the expected proportion of NCA (NVCA or FTA) that would be reduced by the PSA is given by,

$$\sum_{r=1}^{k} \mathsf{APCEp}(r) \cdot \Pr(R_i = r).$$

This quantity equals the overall ITT effect of the PSA provision on the judge's decision.

Furthermore, the arrestees with $R_i = 0$ would never commit a new crime regardless of the judges' decisions. Therefore, we may be interested in estimating the increase in the proportion of the most lenient decision for these safest cases. This generalizes the APCEs to the ordinal decision case to the following quantity,

$$\mathsf{APCEs} \;=\; \Pr\{D_i(1) = 0 \mid R_i = 0\} - \Pr\{D_i(0) = 0 \mid R_i = 0\}.$$

For the cases with $R_i = k+1$ that would always result in a new criminal activity, a desirable decision may depend on a number of factors. Note that if we assume the strict monotonicity, i.e., $Y_i(k) = 0$ for all i, then such cases do not exist.

Note that like the APCEs, the $\mathsf{APCEp}(r)$ can be expressed as a function of the average principal causal effect (APCE) for each decision $d = 1, 2, \ldots, k$, i.e.,

$$\mathsf{APCE}(d, r) \;=\; \Pr\{D_i(1) = d \mid R_i = r\} - \Pr\{D_i(0) = d \mid R_i = r\}. \tag{5}$$

In our empirical analysis, we estimate this causal quantity, which has the same identification conditions.

The identification of these principal causal effects requires the knowledge of the distribution of R_i. Fortunately, under the monotonicity and unconfoundedness assumptions (Assumptions 4 and 6), this distribution is identifiable conditional on $\mathbf{X}_i$,

$$e_r(\mathbf{x}) \;=\; \Pr(R_i = r \mid \mathbf{X}_i = \mathbf{x})$$

$$
\begin{aligned}
&= \Pr(R_i \geq r \mid \mathbf{X}_i = \mathbf{x}) - \Pr(R_i \geq r+1 \mid \mathbf{X}_i = \mathbf{x}) \\
&= \Pr\{Y_i(r-1) = 1 \mid \mathbf{X}_i = \mathbf{x}\} - \Pr\{Y_i(r) = 1 \mid \mathbf{X}_i = \mathbf{x}\} \\
&= \Pr\{Y_i = 1 \mid D_i = r-1, \mathbf{X}_i = \mathbf{x}\} - \Pr\{Y_i = 1 \mid D_i = r, \mathbf{X}_i = \mathbf{x}\}, \text{ for } r = 1, \ldots, k, \quad (6) \\
e_0(\mathbf{x}) &= \Pr\{Y_i(0) = 0 \mid \mathbf{X}_i = \mathbf{x}\} = \Pr\{Y_i = 0 \mid D_i = 0, \mathbf{X}_i = \mathbf{x}\}.
\end{aligned}
$$

Since $e_r(\mathbf{x})$ cannot be negative for each r, this yields a set of testable conditions for Assumptions 4 and 6 .

Finally, we formally present the identification result for the ordinal decision case,

Theorem 5 (Identification with Ordinal Decision) *Under Assumptions 1, 2, 4 and 6,* $\mathsf{APCEp}(r)$ *is identified by*

$$
\begin{aligned}
\mathsf{APCEp}(r) &= \mathbb{E}\{w_r(\mathbf{X}_i)\mathbf{1}(D_i \geq r) \mid Z_i = 1\} - \mathbb{E}\{w_r(\mathbf{X}_i)\mathbf{1}(D_i \geq r) \mid Z_i = 0\}, \\
\mathsf{APCEs} &= \mathbb{E}\{w_0(\mathbf{X}_i)\mathbf{1}(D_i = 0) \mid Z_i = 1\} - \mathbb{E}\{w_0(\mathbf{X}_i)\mathbf{1}(D_i = 0) \mid Z_i = 0\},
\end{aligned}
$$

where $w_r(\mathbf{x}) = e_r(\mathbf{x})/\mathbb{E}\{e_r(\mathbf{X}_i)\}$ *and* $\mathbf{1}()$ *is the indicator function.*

Proof is given in Appendix S3.5.

3.5 Sensitivity Analysis

The unconfoundedness assumption, which enables the nonparametric identification of causal effects, may be violated when researchers do not observe some information used by the judges and predictive of arrestees' behavior. As noted in Section 2.2, the length of time the arrestee has lived in the community may represent an example of such unobserved confounders. Therefore, it is important to develop a sensitivity analysis for the potential violation of the unconfoundedness assumption (Assumption 4).

We begin by proposing a nonparametric sensitivity analysis for the ordinal decision under the monotonicity assumption (Assumption 6). We introduce the following sensitivity parameters, $\xi_d(\mathbf{x})$ for $d = 0, \ldots, k$, to characterize the deviation from the unconfoundedness assumption,

$$
\xi_d(\mathbf{x}) = \frac{\Pr\{D_i(1) = d \mid Y_i(d) = 1, \mathbf{X}_i = \mathbf{x}\}}{\Pr\{D_i(1) = d \mid Y_i(d) = 0, \mathbf{X}_i = \mathbf{x}\}},
$$

which is equal to 1 for all d and $\mathbf{x}$ when the unconfoundedness assumption holds. The randomization of treatment assignment implies,

$$
\begin{aligned}
\Pr\{D_i(1) = d \mid Y_i(d) = 0, \mathbf{X}_i = \mathbf{x}\} &= \frac{\Pr\{Y_i(d) = 0, D_i(1) = 1 \mid \mathbf{X}_i = \mathbf{x}\}}{\Pr\{Y_i(d) = 0 \mid \mathbf{X}_i = \mathbf{x}\}} \\
&= \frac{\Pr\{Y_i = 0, D_i = d \mid Z_i = 1, \mathbf{X}_i = \mathbf{x}\}}{\Pr\{Y_i(d) = 0 \mid \mathbf{X}_i = \mathbf{x}\}},
\end{aligned}
$$

$$
\begin{aligned}
\Pr\{D_i(1) = d \mid Y_i(d) = 1, \mathbf{X}_i = \mathbf{x}\} &= \frac{\Pr\{Y_i(d) = 1, D_i(1) = d \mid \mathbf{X}_i = \mathbf{x}\}}{\Pr\{Y_i(d) = 1 \mid \mathbf{X}_i = \mathbf{x}\}} \\
&= \frac{\Pr\{Y_i = 1, D_i = d \mid Z_i = 1, \mathbf{X}_i = \mathbf{x}\}}{\Pr\{Y_i(d) = 1 \mid \mathbf{X}_i = \mathbf{x}\}}.
\end{aligned}
$$

Therefore, for a given value of $\xi_d(\mathbf{x})$, we have,

$$
\frac{\Pr\{Y_i = 1, D_i = d \mid Z_i = 1, \mathbf{X}_i = \mathbf{x}\}}{\Pr\{Y_i(d) = 1 \mid \mathbf{X}_i = \mathbf{x}\}} = \xi_d(\mathbf{x}) \cdot \frac{\Pr\{Y_i = 0, D_i = d \mid Z_i = 1, \mathbf{X}_i = \mathbf{x}\}}{\Pr\{Y_i(d) = 0 \mid \mathbf{X}_i = \mathbf{x}\}}.
$$

Solving this equation yields,

$$
\Pr\{Y_i(d) = 1 \mid \mathbf{X}_i = \mathbf{x}\} = \frac{\Pr(Y_i = 1, D_i = d \mid Z_i = 1, \mathbf{X}_i = \mathbf{x})}{\xi_1(\mathbf{x}) \cdot \Pr(Y_i = 0, D_i = d \mid Z_i = 1, \mathbf{X}_i = \mathbf{x}) + \Pr(Y_i = 1, D_i = d \mid Z_i = 1, \mathbf{X}_i = \mathbf{x})}.
$$

We can then calculate $e_r(\mathbf{x})$ for $r = 0, \ldots, k+1$. By using these results and Theorem 5, we can identify the $\mathsf{APCEp}(r)$ and APCEs with given values of the sensitivity parameters.

Since the above nonparametric sensitivity analysis requires too many sensitivity parameters, we propose an alternative parametric sensitivity analysis. We consider the following bivariate ordinal probit model for the observed judge's decision D and the latent risk measure R_i,

$$
\begin{aligned}
D_i^*(z) &= \beta_Z z + \mathbf{X}_i^\top \beta_X + z \mathbf{X}_i^\top \beta_{ZX} + \epsilon_{i1}, && (7) \\
R_i^* &= \mathbf{X}_i^\top \alpha_X + \epsilon_{i2}, && (8)
\end{aligned}
$$

where

$$
\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),
$$

and

$$
D_i(z) = \begin{cases} 0 & D^*(z) \leq \theta_{z1} \\ 1 & \theta_{z1} < D_i^*(z) \leq \theta_{z2} \\ \vdots & \vdots \\ k-1 & \theta_{z,k-1} < D_i^*(z) \leq \theta_{zk} \\ k & \theta_{zk} < D_i^*(z) \end{cases}, \quad R_i = \begin{cases} 0 & R_i^* \leq \delta_0 \\ 1 & \delta_0 < R_i^* \leq \delta_1 \\ \vdots & \vdots \\ k & \delta_{k-1} < R_i^* \leq \delta_k \\ k+1 & \delta_k < R_i^* \end{cases}.
$$

The error terms $(\epsilon_{i1}, \epsilon_{i2})$ are assumed to follow a bivariate normal distribution, implying a bivariate ordinal probit model for the two ordinal variables (D_i, R_i). In the literature, Frangakis, Rubin and Zhou (2002), Barnard et al. (2003), and Forastiere, Mealli and VanderWeele (2016) also model the distribution of principal strata using the ordinal probit model.

Under this model, ρ represents a sensitivity parameter since $\rho = 0$ implies Assumption 4. If the value of ρ is known, then the other coefficients, i.e., β_X, α_X and β_Z, can be estimated, which in turn leads to the estimation of the $\mathsf{APCEp}(r)$ and APCEs. Because R_i is a latent variable, the estimation of this model is not straightforward. In our empirical application, we conduct a Bayesian analysis to estimate the causal effects (see e.g., Hirano et al., 2000; Schwartz, Li and Mealli, 2011; Mattei et al., 2013; Jiang, Ding and Geng, 2016, for other applications of Bayesian sensitivity analysis). Appendix S4 presents the details of the Bayesian estimation. We also perform a frequentist analysis, based on Theorem 2, that does not require an outcome model, assessing the robustness of the results to the outcome model.

3.6 Fairness

Next, we discuss how the above causal effects relate to the fairness of the judge's decision. In particular, Imai and Jiang (2020) introduce the concept of "principal fairness." The basic idea is that within each principal strata a fair decision should not depend on protected attributes (race, gender, etc.). Imai and Jiang (2020) provide a detailed discussion about how principal fairness is related to the existing definitions of fairness, which are based on the predictive accuracy (see also Corbett-Davies et al., 2017; Chouldechova and Roth, 2020, and references therein). Although Coston et al. (2020) consider the potential outcome, they only focus on one potential outcome $Y_i(0)$ rather than the joint potential outcomes $(Y_i(0), Y_i(1))$.

Formally, let $A_i \in \mathcal{A}$ be a protected attribute such as race and gender. We first consider a binary decision. We say that decisions are fair on average with respect to A_i if it does not depend on the attribute within each principal stratum, i.e.,

$$\Pr\{D_i = 1 \mid A_i, Y_i(1) = y_1, Y_i(0) = y_0\} \;=\; \Pr\{D_i = 1 \mid Y_i(1) = y_1, Y_i(0) = y_0\} \tag{9}$$

for all $y_1, y_0 \in \{0, 1\}$. We can generalize this definition to the ordinal case as,

$$\Pr(D_i \geq d \mid A_i, R_i = r) \;=\; \Pr(D_i \geq d \mid R_i = r)$$

for $1 \leq d \leq k$ and $0 \leq r \leq k + 1$.

The degree of fairness for principal stratum $R_i = r$ can be measured using the maximal deviation among the distributions for different groups,

$$\Delta_r(z) \;=\; \max_{a,a',d} \left|\Pr\{D_i(z) \geq d \mid A_i = a, R_i = r\} \;-\; \Pr\{D_i(z) \geq d \mid A_i = a', R_i = r\}\right| \tag{10}$$

for $z = 0, 1$. By estimating $\Delta_r(z)$, we can use the experimental data to examine whether or not the provision of the PSA improves the fairness of judge's decisions. Specifically, the PSA provision improves the fairness of judge's decisions for the principal stratum r if $\Delta_r(1) \leq \Delta_r(0)$.

3.7 Optimal Decision

The discussion so far has focused on estimating the impacts of algorithmic recommendations on human decisions. We now show that the experimental data can also be used to derive optimal decision rules given a certain objective. By comparing human decisions and algorithmic recommendations with optimal decision rules, we can evaluate their efficacy. In our application, one goal is to prevent as many NCAs (NVCAs or FTAs) as possible while avoiding unnecessarily harsh initial release decisions. To achieve this, we must carefully weigh the cost of negative outcomes and that of unnecessarily harsh decisions. We show how to empirically assess this tradeoff using the experimental data.

Formally, let δ be the judge's decision based on $\mathbf{X}_i$, which may include the PSA. We consider a deterministic decision rule, i.e., $\delta(\mathbf{x}) = d$ if $\mathbf{x} \in \mathcal{X}_d$ where $\mathcal{X}_d$ is a non-overlapping partition of the covariate space $\mathcal{X}$ with $\mathcal{X} = \bigcup_{r=0}^{k} \mathcal{X}_r$ and $\mathcal{X}_r \cap \mathcal{X}_{r'} = \emptyset$. We consider the utility function of the following general form,

$$U_i(\delta) \;=\; \begin{cases} -c_0 & \delta(\mathbf{X}_i) < R_i \\ 1 & \delta(\mathbf{X}_i) = R_i \\ 1 - c_1 & \delta(\mathbf{X}_i) > R_i \end{cases},$$

where c_0 and c_1 represent the cost of an NCA (NVCA or FTA) and that of an unnecessarily harsh decision, respectively. Under this setting, preventing an NCA (NVCA or FTA) with the most lenient decision ($\delta(\mathbf{X}_i) = R_i$) yields the utility of one, i.e., $U_i(\delta) = 1$, while we incur the cost c_1 for an unnecessarily harsh decision ($\delta(\mathbf{X}_i) > R_i$), leading to the net utility of $1 - c_1$.

The relative magnitude of these two cost parameters, c_0 and c_1, may depend on the consideration of various factors including the potential harm to the public and arrestees caused by the negative outcomes and unnecessarily harsh decisions, respectively. When $c_0 = c_1 = 0$, for example, $U_i(\delta)$ reduces to $\mathbf{1}\{\delta(\mathbf{X}_i) \geq R_i\}$, which is non-zero only if the decision is sufficiently harsh so that it prevents the negative outcome. The optimal decision under this utility is the most stringent decision, i.e., $\delta(\mathbf{X}_i) = k$, for all cases. If $c_0 = 2$ and $c_1 = 1$, the resulting utility function implies that the cost of NCA (NVCA or FTA) is twice as large as that of an unnecessarily harsh decision.

We derive the optimal decision rule δ^* that maximizes the expected utility,

$$\delta^* \;=\; \operatorname*{argmax}_{\delta} \mathbb{E}\{U_i(\delta)\}.$$

For $r = 0, \ldots, k+1$ and $d = 0, \ldots, k$, we can write

$$\mathbb{E}[\mathbf{1}\{\delta(\mathbf{X}_i) = d, R_i = r\}] \;\; = \;\; \mathbb{E}\{\mathbf{1}(\mathbf{X}_i \in \mathcal{X}_d, R_i = r)\} \; = \; \mathbb{E}\left\{\mathbf{1}(\mathbf{X}_i \in \mathcal{X}_d) \cdot e_r(\mathbf{X}_i)\right\}.$$

Therefore, we can express $\mathbb{E}\{U_i(\delta)\}$ as

$$\begin{aligned}
& \sum_{r=0}^{k+1}\sum_{d=0}^{k} \mathbb{E}[\mathbf{1}\{\delta(\mathbf{X}_i) = d, R_i = r\}] \\
= \; & \sum_{r=0}^{k+1} \left[\sum_{d \geq r} \mathbb{E}\left\{\mathbf{1}(\mathbf{X}_i \in \mathcal{X}_d) \cdot e_r(\mathbf{X}_i)\right\} - c_0 \sum_{d<r} \mathbb{E}\left\{\mathbf{1}(\mathbf{X}_i \in \mathcal{X}_d) \cdot e_r(\mathbf{X}_i)\right\} - c_1 \sum_{d>r} \mathbb{E}\left\{\mathbf{1}(\mathbf{X}_i \in \mathcal{X}_d) \cdot e_r(\mathbf{X}_i)\right\}\right] \\
= \; & \sum_{d=0}^{k} \mathbb{E}\left[\mathbf{1}(\mathbf{X}_i \in \mathcal{X}_d)\left\{\sum_{r \leq d} e_r(\mathbf{X}_i) - c_0 \cdot \sum_{r>d} e_r(\mathbf{X}_i) - c_1 \cdot \sum_{r<d} e_r(\mathbf{X}_i)\right\}\right]. \qquad (11)
\end{aligned}$$

This yields the following optimal decision,

$$\delta^*(\mathbf{x}) \; = \; \operatorname*{argmax}_{d \in \{0,\ldots,k\}} g_d(\mathbf{x}) \quad \text{where} \quad g_d(\mathbf{x}) \; = \; \sum_{r \leq d} e_r(\mathbf{x}) - c_0 \cdot \sum_{r>d} e_r(\mathbf{x}) - c_1 \cdot \sum_{r<d} e_r(\mathbf{x}). \qquad (12)$$

We can use the experimental estimate $e_r(\mathbf{x})$ to learn about the optimal decision.

Policy makers could derive the optimal decision rule by using the above result and then use it as the PSA recommendation. However, this may not be useful if the judge decides to follow the algorithmic recommendation selectively for some cases or ignore it altogether. We may wish to construct PSA scores that maximize the optimality of the judge's decision. Unfortunately, the derivation of such an optimal PSA score is difficult since the PSA scores were not directly randomized in our experiment. In Appendix S5, we instead consider the optimal provision of the PSA given the same goal considered above (i.e., prevent as many NCAs (NVCAs or FTAs) as possible with the minimal amount of bail).

4 Empirical Analysis

In this section, we apply the proposed methodology to the data from the field RCT described in Section 2.

4.1 Preliminaries

As explained in Section 2.3, we use the ordinal decision variable with three categories — the signature bond ($D_i = 0$), the bail amount of \$1,000 or less ($D_i = 1$), and the bail amount of greater than \$1,000 ($D_i = 2$). Given this ordinal decision, we call the principal strata as safe ($R_i = 0$), easily preventable ($R_i = 1$), preventable ($R_i = 2$), and risky cases ($R_i = 3$).

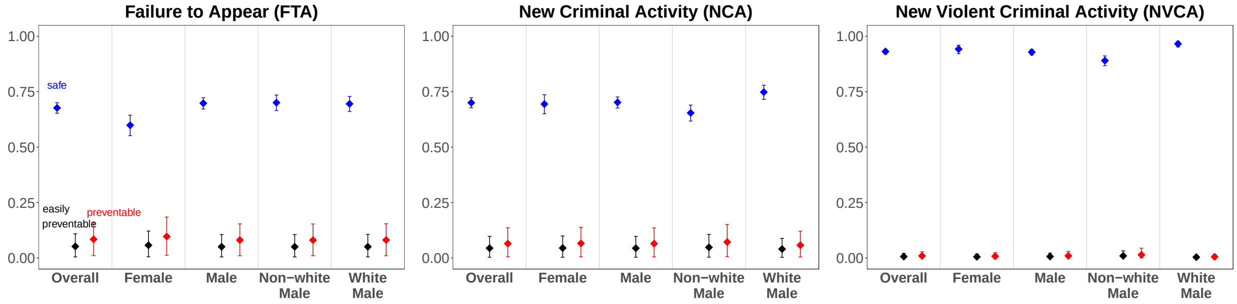


Figure 3: Estimated Population Proportion of Each Principal Stratum. Each plot represents the result using one of the three outcome variables (FTA, NCA, and NVCA), where the blue, black, and red diamonds represent the estimates for safe, easily preventable, and preventable cases, respectively. These three estimates do not necessarily sum to one because there is an additional, very small, principal stratum of risky case. The solid vertical lines represent the 95% Bayesian credible intervals. The results show that a vast majority of cases are safe across subgroups and across different outcomes. The proportion of safe cases is estimated to be especially high for NVCA.

We fit the Bayesian model defined in equations (7) and (8) with a diffuse prior distribution as specified in Appendix S4, separately for each of three binary outcome variables — FTA, NCA, and NVCA. The model incorporates following pre-treatment covariates: gender (male or female), race (white or non-white), the interaction between gender and race, age, and several indicator variables regarding the current and past charges. They include the presence of current violent offense, pending charge (either felony, misdemeanor, or both) at time of offense, felony charge, misdemeanor charge, prior misdemeanor conviction, prior violent conviction, prior felony conviction, prior sentence to incarceration, and prior FTA.

We use the Gibbs sampling and run five Markov chains of 100,000 iterations each with random starting values independently drawn from the prior distribution. Based on the Gelman-Rubin statistic for convergence diagnostics, we retain the second half of each chain and combine them to be used for our analysis. Appendix S4 presents the computational details including the Gibbs sampling algorithm we use.

We begin by computing the estimated population proportion of each principal stratum based on equation (6). Figure 3 presents the results. We find that the overall proportion of safe cases (blue circles) is estimated to be 68%, whereas those of easily preventable (black triangles) and preventable (red squares) cases are 5% and 8%, respectively. A similar pattern is observed for FTA and NCA across different racial and gender groups, while the estimated overall proportion of safe cases is even higher for NVCA, exceeding 93%.

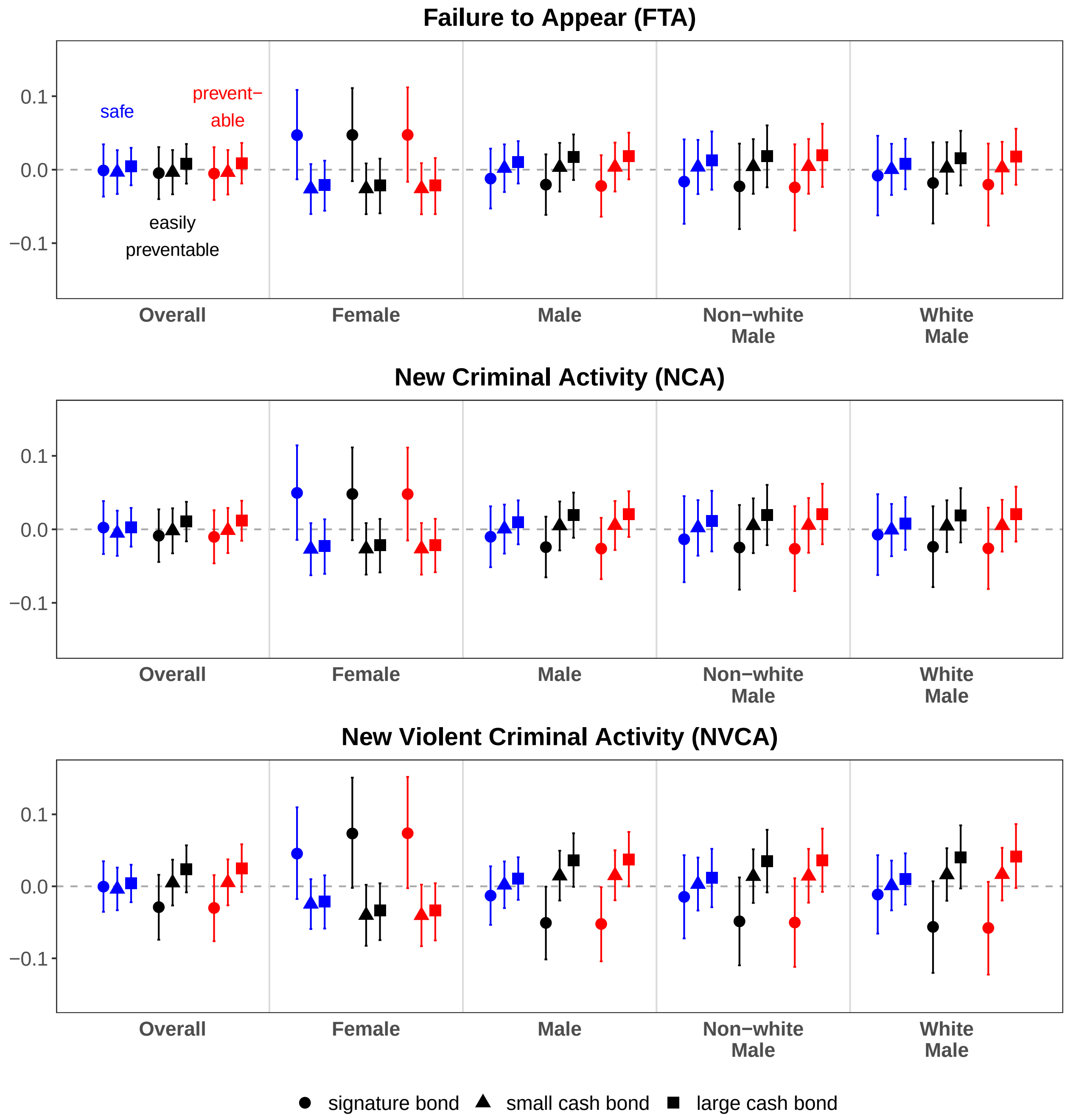


Figure 4: Estimated Average Principal Causal Effects (APCE) of PSA Provision on the Judge's Decision. Each panel presents the overall and subgroup-specific results for a different outcome variable. Each column within a panel shows the estimated APCE of PSA provision for safe (blue), easily preventable (black), and preventable (red) cases. For each of these principal strata, we report the estimated APCE on the judge's decision to impose a signature bond (circles), a small cash bail amount of 1,000 dollars or less (triangles), and a large cash bail amount of greater than 1,000 (squares). The vertical line for each estimate represents the Bayesian 95% credible interval. The results show that the PSA provision may make the judge's decision more lenient for female arrestees regardless of their risk levels. The PSA provision may also encourage the judge to make a harsher decision for male arrestees with a greater risk level.

4.2 Average Principal Causal Effects

Figure 4 presents the estimated APCE of the PSA provision on the three ordinal decision categories, separately for each of the three outcomes and each principal stratum (see equation (5)). The overall and subgroup-specific results are given for each of the three principal strata — safe (blue), easily preventable (black), and preventable (red) cases. For a given principal stratum, we present the estimated APCE on each decision category — signature bond (circle), small cash bond (triangle), and large cash bond (square). The left column of each panel shows that the PSA provision has little overall impact on the judge's decision across three principal strata for FTA and NCA. There is a suggestive, but inconclusive, evidence that the PSA provision leads to an overall harsher decision for NVCA among preventable and easily preventable cases.

We also present the estimated APCE for different gender and racial groups in the remaining columns of each panel. We find potentially suggestive evidence that the PSA provision may make it more likely for the judge to impose signature bonds (circles) on female arrestees instead of cash bonds (triangles and squares) across three outcomes. Interestingly, for all outcomes, this pattern appears to hold for any of the three principal strata, implying that the PSA provision might not help the judge distinguish different risk types of female arrestees. Our analysis also finds that for NVCA the PSA provision may lead to a harsher decision for easily preventable and preventable cases among male arrestees while it has little effect on the safe cases. This suggests that the PSA provision may help distinguish different risk types among male arrestees, resulting in better decisions. Interestingly, there is no discernible racial difference in these effects.

In Appendix S2.2, we explore the estimated APCE for different age groups. We find that the PSA provision may lead to a harsher decision for arrestees of the 26–35 years old group across three outcomes. This pattern seems to generally hold across all principal strata though for NVCA the effects are more pronounced for preventable and easily preventable cases. Finally, our analysis yields suggestive evidence that across all outcomes, the PSA provision may make the judge's decision more lenient for the oldest (46 years old or above) group. This appears to be true across all three outcomes except that for NVCA this effect may exist only for safe cases.

Finally, we conduct two robustness analyses. First, we perform a frequentist analysis that is based on Theorem 2 and does not assume an outcome model. The results are shown in Appendix S6, and are largely consistent with those shown here though the estimation uncertainty of the frequentist analysis, which makes less stringent assumptions than Bayesian analysis, is greater as expected. Second, we conduct a sensitivity analysis using the methodology described in Section 3.5. In particular, we set the value of correlation parameter ρ to 0.05, 0.1, and 0.3, and examine how the estimated APCE changes. The results in Appendix S7 show that the results appear to be largely consistent across different values of ρ although the effects for females tend to exhibit a large degree of estimation uncertainty especially when the correlation is high and particularly for NVCA. This is not surprising given the small sample size of female arrestees and only a handful of them commit NVCA.

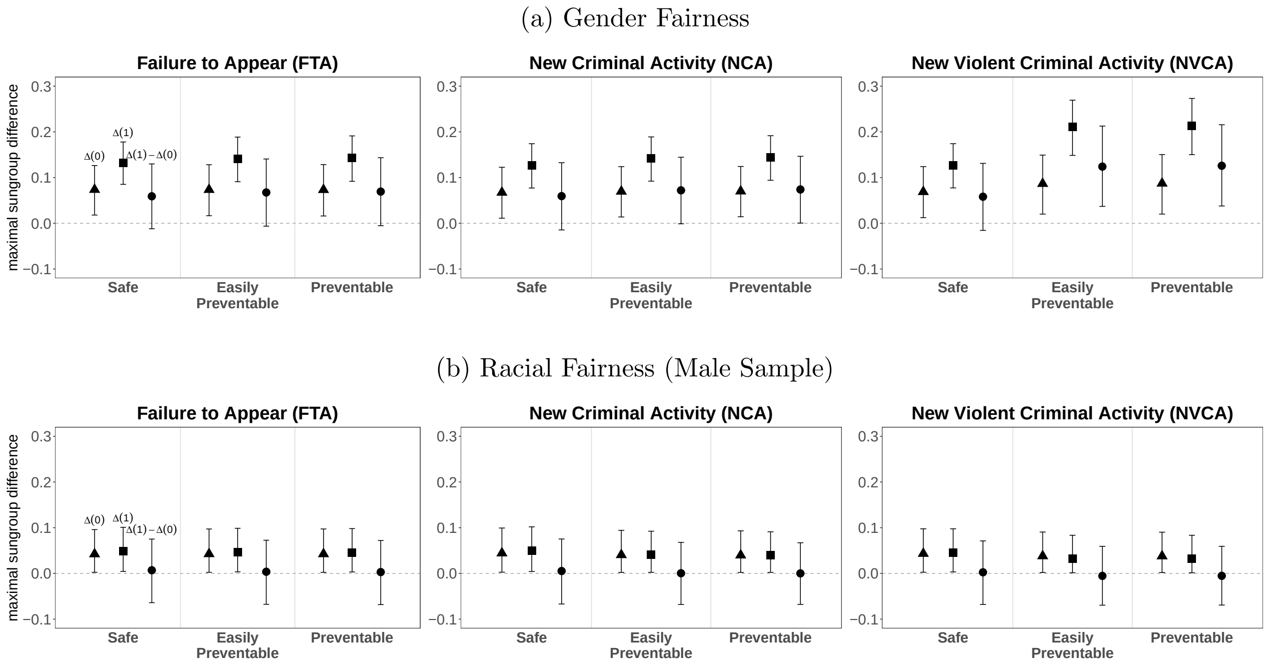
4.3 Gender and Racial Fairness

We now examine the impacts of the PSA provision on gender and racial fairness. Specifically, we evaluate the principal fairness of the PSA provision as discussed in Section 3.6. We use gender (female vs. male) and race (white male vs. non-white male) separately as a protected attribute, and analyze the two subgroups defined by each of the two variables. While the gender analysis is based on the entire sample, the racial analysis is based on the male sample due to the limited sample size for females.

Figure 5 presents the results for gender (top panel) and racial (bottom panel) fairness across the principal strata and separately for each of the three outcomes. Each column within a given plot presents $\Delta_r(z)$ defined in equation (10), which represents the maximal subgroup difference in the judge's decision probability distribution within the same principal stratum $R_i = r$ under the provision of PSA $z = 1$ (no provision $z = 0$). In this application, the maximal difference always occurs at $d = 1$, allowing us to interpret $\Delta_r(z)$ as the difference in probability of imposing a cash bond rather than a signature bond. We also present the estimated difference caused by the PSA provision in the two maximal subgroup differences, i.e., $\Delta_r(1) - \Delta_r(0)$. If this difference is estimated to be positive, then the PSA provision reduces the fairness of judge's decisions by increasing the maximal subgroup difference.

When the PSA is not provided, the maximal subgroup differences in the judge's decision probability $\Delta_r(0)$ are relatively small but significantly greater than zero in terms of both gender and race. For example, within each risk category, the judge is more likely to impose a cash bond on male arrestees than on female arrestees. In addition, it appears that among male arrestees, the judge is more likely to impose a cash bond on non-whites than on whites even though they belong to the same risk category. This suggests that when the PSA is not provided, the judges' decision may be biased against males and non-whites according to the principal fairness criterion.

We find that the PSA provision might worsen the gender fairness of judge's decisions. When the PSA is provided, the maximal gender difference in the judge's decision probability is on average greater than that when it is not provided. The effect is particularly large and statistically significant



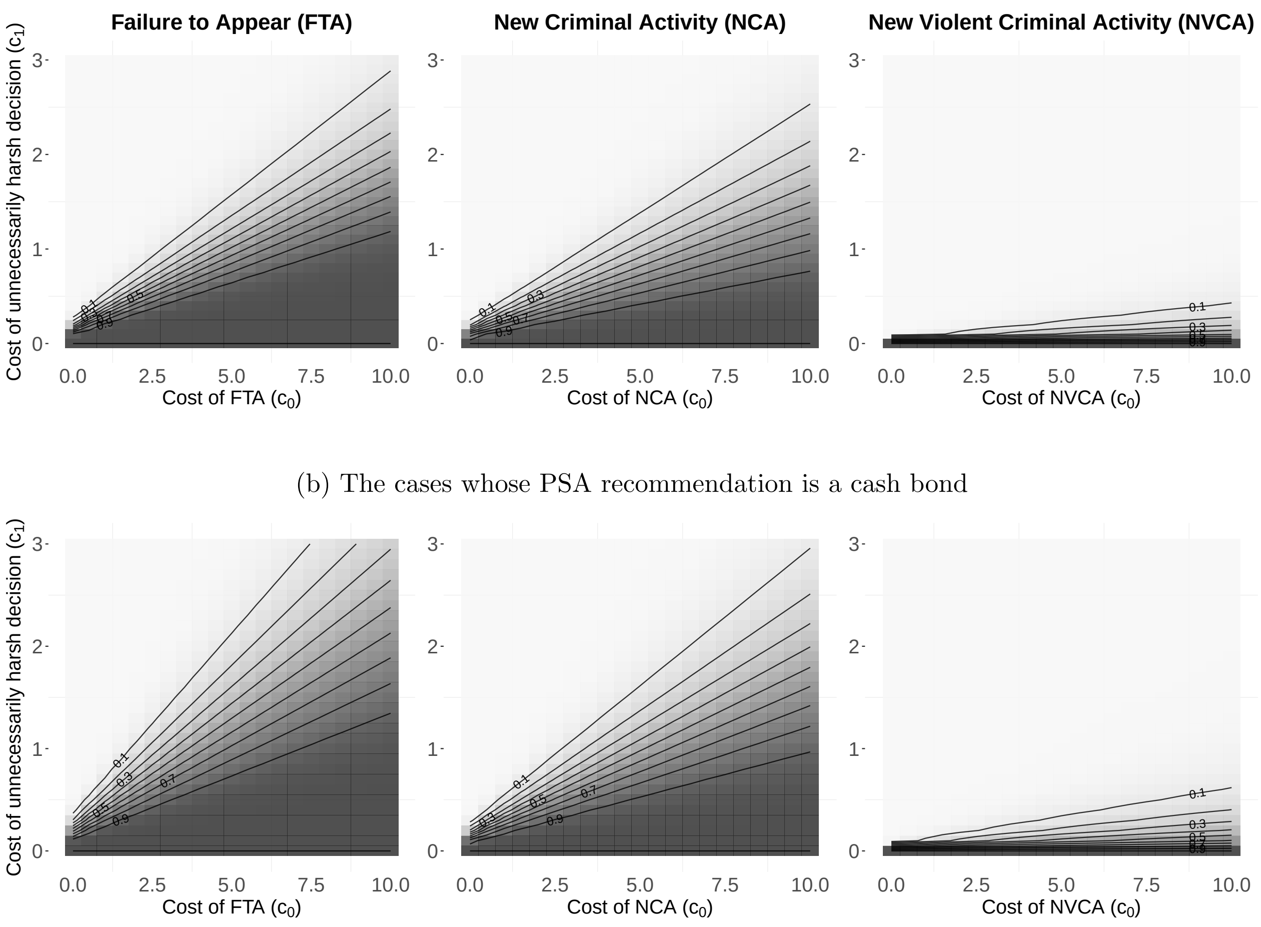
(a) Gender Fairness

(b) Racial Fairness (Male Sample)

Figure 5: Gender and Racial Fairness of the PSA Recommendation and Judge's Decisions. Within each plot, we show three estimates separately for each principal stratum — the maximal subgroup difference in the judge's decision probability of imposing a cash bond with the PSA provision (squares; $\Delta(1)$) and without it (triangles; $\Delta(0)$) as well as the difference between them (circles; $\Delta(1) - \Delta(0)$). The vertical solid lines represent the 95% Bayesian credible intervals. A positive value of the difference would imply that the PSA reduces the fairness of the judge's decisions. For the gender analysis (top panel), even without the PSA, the judge seems to be more likely to impose a cash bond on male arrestees when compared to female arrestees with the same risk levels. The PSA provision appears to worsen this existing bias. For the race analysis (bottom panel), the PSA provision has little impact on the existing bias of the judge's decision against non-whites across all outcomes and risk levels.

for NVCA and for preventable and easily preventable cases. This is consistent with our finding that especially for NVCA, the PSA provision might make the judge's decision more lenient for female arrestees while it leads to a harsher decision for male arrestees among preventable and easily preventable cases. Thus, the PSA provision appears to reduce gender fairness.

However, the PSA provision does not have a statistically significant impact on the racial fairness of judges' decisions among male arrestees. For instance, in the principal stratum of safe cases, we find that when the PSA is provided, the maximal difference in the judge's decision probability (between non-white males and white males) is essentially identical to that when it is not provided. This suggests that the PSA may not alter the racial bias of judge's decisions against non-whites.


(a) The cases whose PSA recommendation is a signature bond

(b) The cases whose PSA recommendation is a cash bond

Figure 6: Estimated Proportion of Cases for Which Cash Bond is Optimal. Each column represents the results based on one of the three outcomes (FTA, NCA, and NVCA). The top (bottom) panel shows the results for the cases whose PSA recommendation is a signature (cash) bond. In each plot, the contour lines represent the estimated proportions of cases for which a cash bond is optimal, given the cost of an unnecessarily harsh decision (c_1; y-axis) and that of a negative outcome (c_0; x-axis). A dark grey area represents a greater proportion of such cases. The results show that regardless of the PSA recommendation, a signature bond is optimal unless the cost of a negative outcome is much greater than the cost of an unnecessarily harsh decision.

4.4 Optimal Decision

Finally, we empirically investigate the optimal decision as discussed in Section 3.7 by comparing it with the PSA recommendation under different values of the costs. Given a specific pair of cost parameters (c_0, c_1) and the experimental estimate of $e_r(\mathbf{x})$ for $r = 0, \ldots, k$, we can compute the optimal decision for each case according to equation (12) and then obtain the estimated proportion of cases, for which a cash bond (either small or large amount) is optimal. We repeat this process for a grid of different values for the cost of a negative outcome (c_0; FTA, NCA, and NVCA) and that of an unnecessarily harsh decision (c_1).

The top panel of Figure 6 presents the results for the cases whose PSA recommendation is a signature bond (FTA score less than or equal to 4, NCA score less than or equal to 4, and NVCA flag equals to 0). In contrast, the bottom panel of the figure shows the results for the other cases (i.e., the PSA recommendation is a cash bond). In each plot, a darker grey region represents a greater proportion. The results suggest that unless the cost of a negative outcome is much larger than the cost of an unnecessarily harsh decision, imposing a signature bond is the optimal decision for a vast majority of cases.

We also find that for all three outcomes, a cash bond is optimal for a greater proportion of cases when the PSA recommendation is indeed a cash bond. However, this difference is small, suggesting that the PSA recommendation is only mildly informative. Similar results are found even if we separately examine three PSA scores (see Figure S13 in Appendix S8).

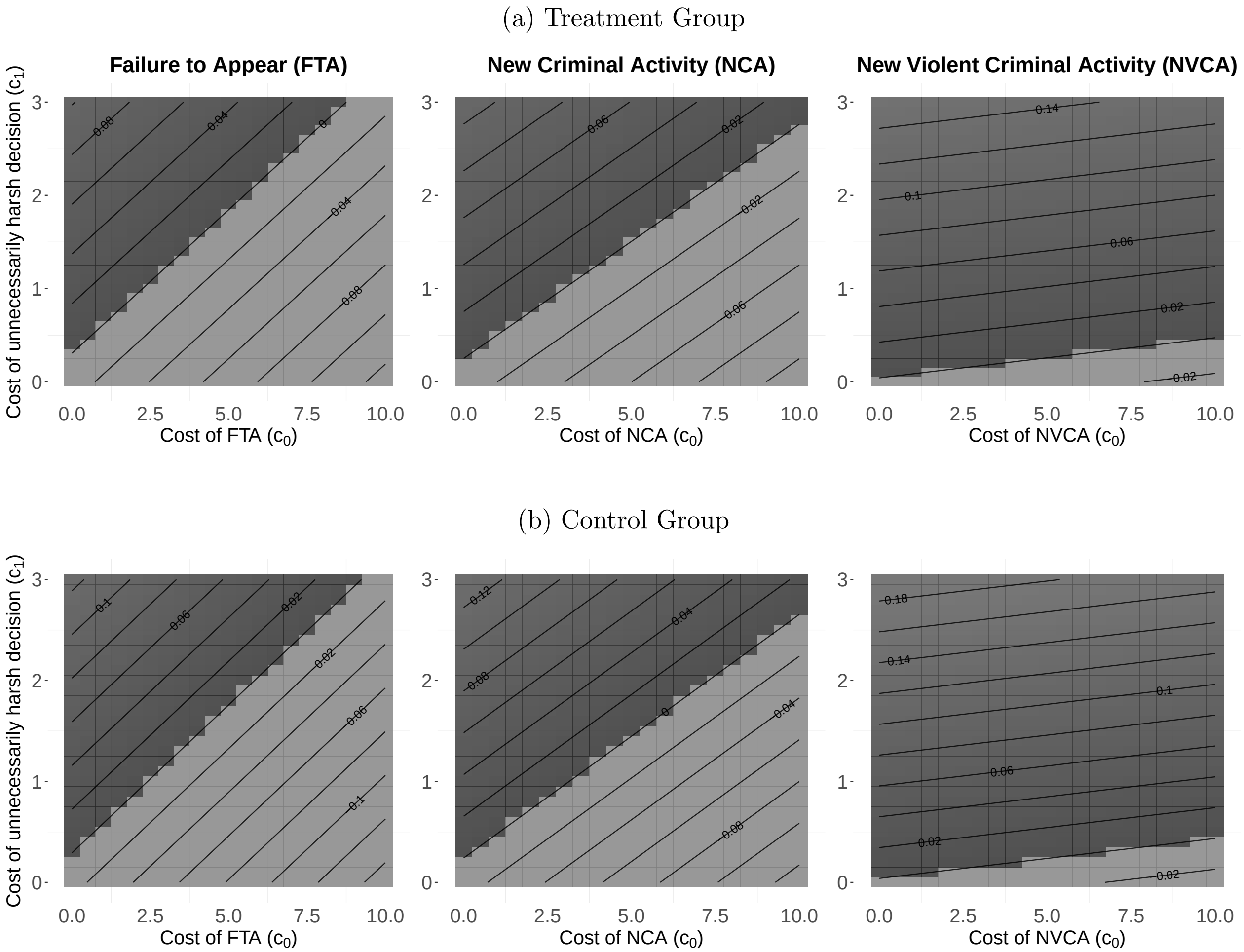
4.5 Comparison between the Judge's Decisions and PSA Recommendations

Lastly, we compare the judge's actual decision with the PSA recommendation in terms of the expected utility given in equation (11). The top panel of Figure 7 represents the results for the treatment group (i.e., the judge's decision with PSA), whereas the bottom panel represents those for the control group (i.e., the judge's decision without PSA). A darker grey area indicates that the expected utility for the judge's decision is estimated to be greater than the PSA recommendation. Most of these estimates are statistically significant (see Figure S14 for more details). We find that unless the cost of a negative outcome is much greater than the cost of an unnecessarily harsh decision, the judge's decision (with or without PSA scores) yields a greater expected utility than the PSA recommendation. This is especially true for NVCA. In other words, the PSA recommendations may be unnecessarily more stringent than the judge's decisions.

5 Concluding Remarks

In today's data-rich society, many human decisions are guided by algorithmic-recommendations. While some of these algorithmic-assisted human decisions may be trivial and routine (e.g., online shopping and movie suggestions), others that are much more consequential include judicial and medical decision-making. As algorithmic recommendation systems play increasingly important roles in our lives, we believe that a policy-relevant question is how such systems influence human decisions and how the biases of algorithmic recommendations interact with those of human decisions. Thus, it is essential to empirically evaluate the impacts of algorithmic recommendations on human decisions.

In this paper, we present a set of general statistical methods that can be used for the experimental


(a) Treatment Group

(b) Control Group

Figure 7: Estimated Difference in the Expected Utility between the Judge's Decisions and PSA Recommendations for the Treatment (top panel) and Control (bottom panel) Group. Each column represents the results base on one of the three outcomes with a darker region indicating the values of the costs (the cost of a negative outcome and the cost of an unnecessarily harsh decision) for which the Judge's decision yields a higher expected utility (i.e., more optimal) than the corresponding PSA recommendation. The results show that the judge's decision is more optimal than the PSA recommendation unless the cost of a negative outcome is much higher than the cost of an unnecessarily harsh decision. This pattern holds for all outcomes and is unchanged by the provision of PSA.

evaluation of algorithm-assisted human decision making. We applied these methods to the first-ever randomized controlled trial for assessing the impacts of the PSA provision on judges' pretrial decisions. There are several findings that emerge from our analysis. First, we find that the PSA provision has little overall impacts on the judge's decisions. Second, we find potentially suggestive evidence the PSA provision may encourage the judge to make more lenient decisions for female arrestees regardless of their risk levels while leading to more stringent decisions for males who are classified as risky. Third, the PSA provision appears to widen the existing gender bias of the judge's decisions against male arrestees whereas it does not seem to alter the existing racial bias against non-whites among male arrestees. Finally, we find that for a vast majority of cases, the optimal decision is to impose a signature bond rather than a cash bond unless the cost of a new crime is much higher than that of a decision that may result in unnecessary incarceration. This suggests that the PSA recommendations may be harsher than necessary. These results might bring into question the utilities of using PSA in judicial decision-making.

References

Albright, Alex. 2019. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. Technical Report. Department of Economics, Harvard University.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barnard, John, Constantine E Frangakis, Jennifer L Hill and Donald B Rubin. 2003. "Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City." *Journal of the American Statistical Association* 98:299–323.

Berk, Richard. 2017. "An impact assessment of machine learning risk forecasts on parole board decisions and recidivism." *Journal of Experimental Criminology* 13:193–216.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research.* DOI:10.1177/0049124118782533.

Chouldechova, Alexandra and Aaron Roth. 2020. "A Snapshot of the Frontiers of Fairness in Machine Learning." *Communications of the ACM* 63:82–89.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *KDD'17.* August 13–17, 2017 Halifax, NS, Canada: .

Coston, Amanda, Alan Mishler, Edward H. Kennedy and Alexandra Chouldechova. 2020. Counterfactual Risk Assessments, Evaluation, and Fairness. In *FAT* '20.* January 27—30, 2020 Barcelona, Spain: .

Dawes, Robyn M., David Faust and Paul E. Meehl. 1989. "Clinical Versus Actuarial Judgment." *Science* 243:1668–1674.

Dieterich, William, Christina Mendoza and Tim Brennan. 2016. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity." `http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf`. Northpointe Inc. Research Department.

Ding, Peng and Jiannan Lu. 2017. "Principal stratification analysis using principal scores." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 79:757–777.

Dressel, Julia and Hany Farid. 2018. "The accuracy, fairness, and limits of predicting recidivism." *Science Advances* 4:eaao5580.

Flores, Anthony W., Kristin Bechtel and Christopher Lowenkamp. 2016. "False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."." *Federal Probation Journal* 80:28–46.

Forastiere, Laura, Fabrizia Mealli and Tyler J VanderWeele. 2016. "Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification." *Journal of the American Statistical Association* 111:510–525.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58:21–29.

Frangakis, Constantine E, Donald B Rubin and Xiao-Hua Zhou. 2002. "Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms." *Biostatistics* 3:147–164.

Garrett, Brandon L. and John Monahan. 2020. "Judging Risk." *California Law Review*.

Goldkamp, John S. and Michael R. Gottfredson. 1984. *Judicial Guidelines for Bail: The Philadelphia Experiment*. Washington D.C.: U.S. Department of Justice, National Institute of Justice.

Goldkamp, John S. and Michael R. Gottfredson. 1985. *Policy Guidelines for Bail: An Experiment in Court Reform*. Temple University Press.

Green, Ben and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19).* January 29–31, 2019 Atlanta, GA, USA: pp. 90–99.

Greiner, D. James, Ryan Halen, Matthew Stubenberg and Christopher L. Griffin, Jr. 2020. Randomized Control Trial Evaluation of the Implementation of the PSA-DMF System in Dane County, WI. Technical Report. Access to Justice Lab, Harvard Law School.

Hansen, John H. L. and Taufiq Hasan. 2015. "Speaker Recognition by Machines and Humans: A Tutorial Review." *IEEE Signal Processing Magazine* 32:74–99.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *The IEEE International Conference on Computer Vision (ICCV).* pp. 1026–1034.

Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin and Xiao-Hua Zhou. 2000. "Assessing the Effect of an Influenza Vaccine in an Encouragement Design." *Biostatistics* 1:69–88.

Imai, Kosuke and Zhichao Jiang. 2020. "Principal Fairness for Human and Algorithmic Decision-Making." *Working paper available at https://arxiv.org/pdf/2005.10400.pdf.*

Jiang, Zhichao, Peng Ding and Zhi Geng. 2016. "Principal causal effect identification and surrogate end point evaluation by multiple trials." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78:829–848.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133:237–293.

Mattei, Alessandra, Fan Li, Fabrizia Mealli et al. 2013. "Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program." *The Annals of Applied Statistics* 7:2336–2360.

Miller, Joel and Carrie Maloney. 2013. "Practitioner Compliance With Risk/Needs Assessment Tools: A Theoretical and Empirical Assessment." *Criminal Justice and Behavior* 40:716–736.

Rubin, Donald B. 1990. "Comments on "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed." *Statistical Science* 5:472–480.

Rudin, Cynthia, Caroline Wang and Beau Coker. 2020. "The Age of Secrecy and Unfairness in Recidivism Prediction." *Harvard Data Science Review* 2.
URL: *https://hdsr.mitpress.mit.edu/pub/7z10o269*

Schwartz, Scott L, Fan Li and Fabrizia Mealli. 2011. "A Bayesian semiparametric approach to intermediate variables in causal inference." *Journal of the American Statistical Association* 106:1331–1344.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan and Demis Hassabi. 2018. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science* 362:1140–1144.

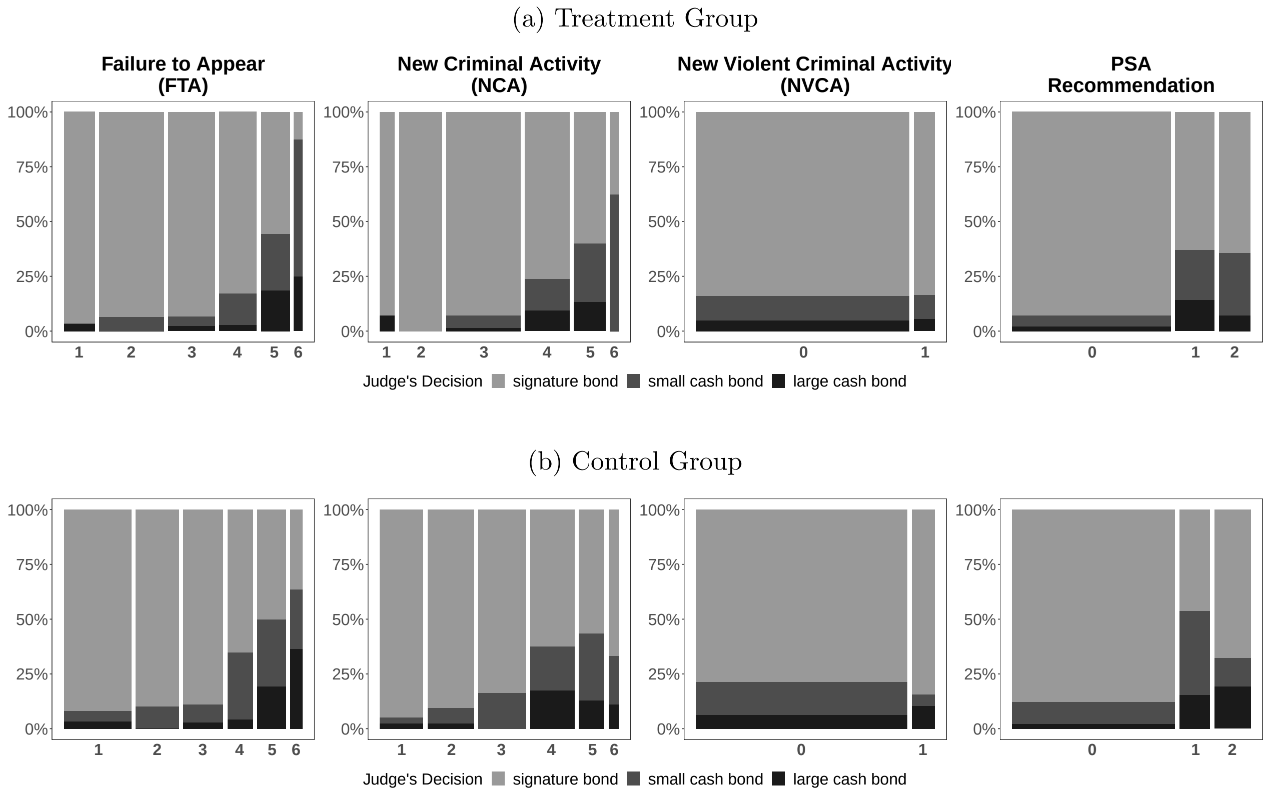
Skeem, Jennifer, Nicholas Scurich and John Monahan. 2020. "Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendant." *Law and Human Behavior* 44:51–59.

Stevenson, Megan. 2018. "Assessing Risk Assessment in Action." *Minnesota Law Review.*

Supplementary Appendix

S1 Distribution of Judge's Decisions given the PSA for Subgroups

S1.1 Female Arrestees

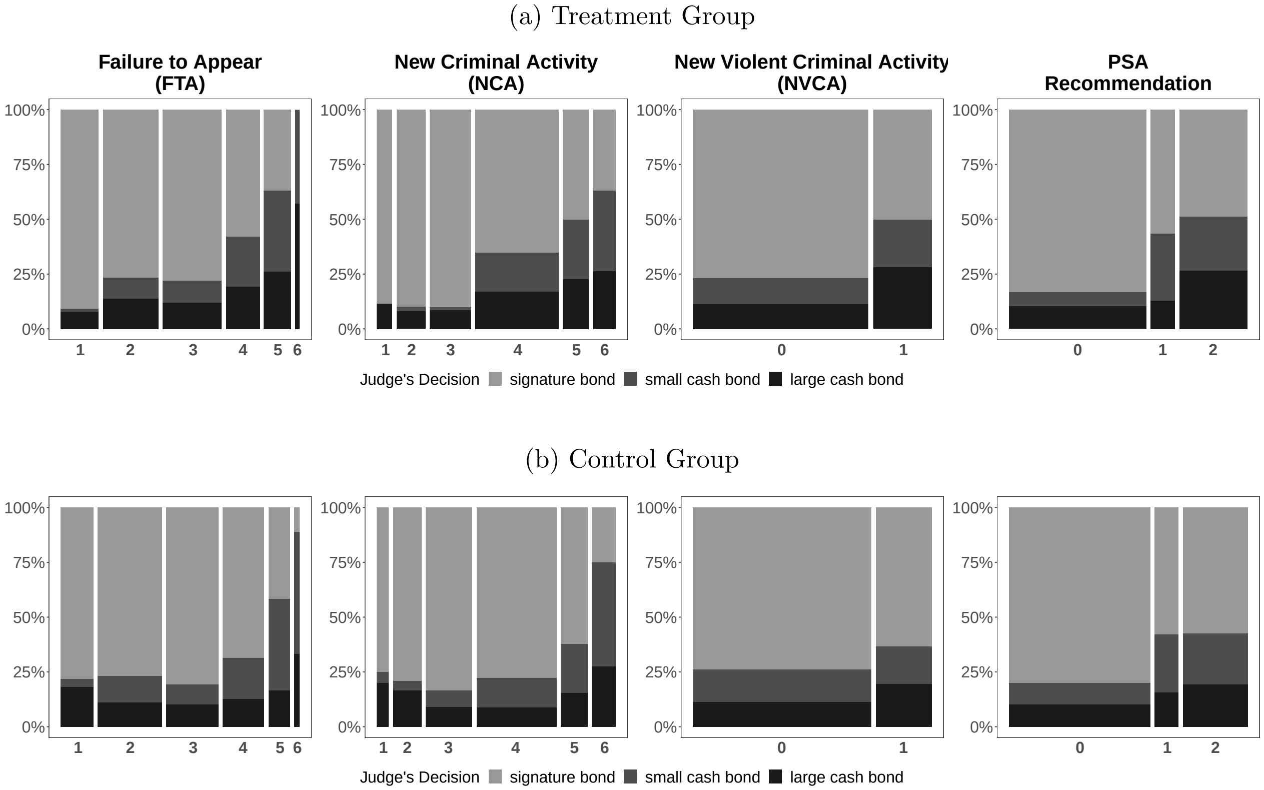


(a) Treatment Group

(b) Control Group

Figure S1: The Distribution of Judge's Decisions given the Pretrial Public Safety Assessment (PSA) among the Cases in the Treatment (Top Panel) and Control (Bottom Panel) Groups Among Female Arrestees.

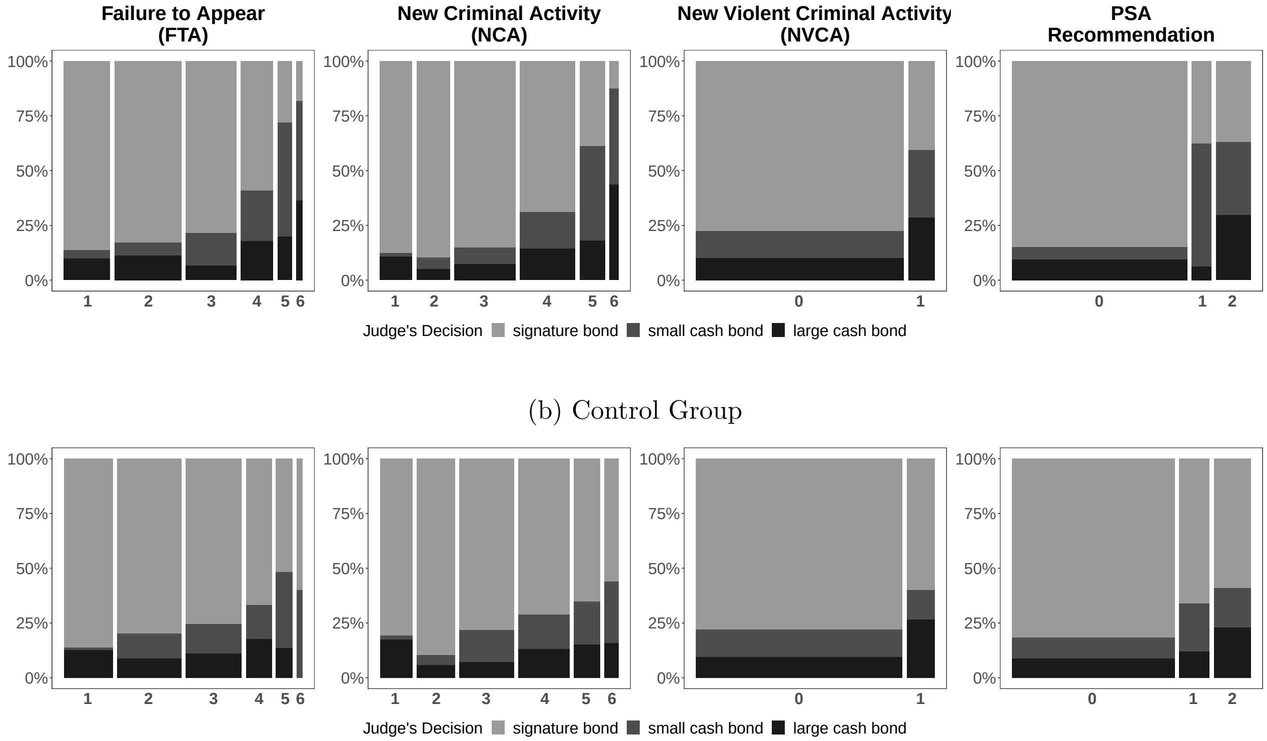
S1.2 Non-white Male Arrestees


(a) Treatment Group

(b) Control Group

Figure S2: The Distribution of Judge's Decisions given the Pretrial Public Safety Assessment (PSA) among the Cases in the Treatment (Top Panel) and Control (Bottom Panel) Groups Among Non-white Male Arrestees.

S1.3 White Male Arrestees


(a) Treatment Group

(b) Control Group

Figure S3: The Distribution of Judge's Decisions given the Pretrial Public Safety Assessment (PSA) among the Cases in the Treatment (Top Panel) and Control (Bottom Panel) Groups Among White Male Arrestees.

S2 Subgroup Analysis for Age Groups

In this appendix, we conduct the subgroup analysis for different age groups.

S2.1 Age Distribution, Descriptive Statistics, and Average Causal Effects

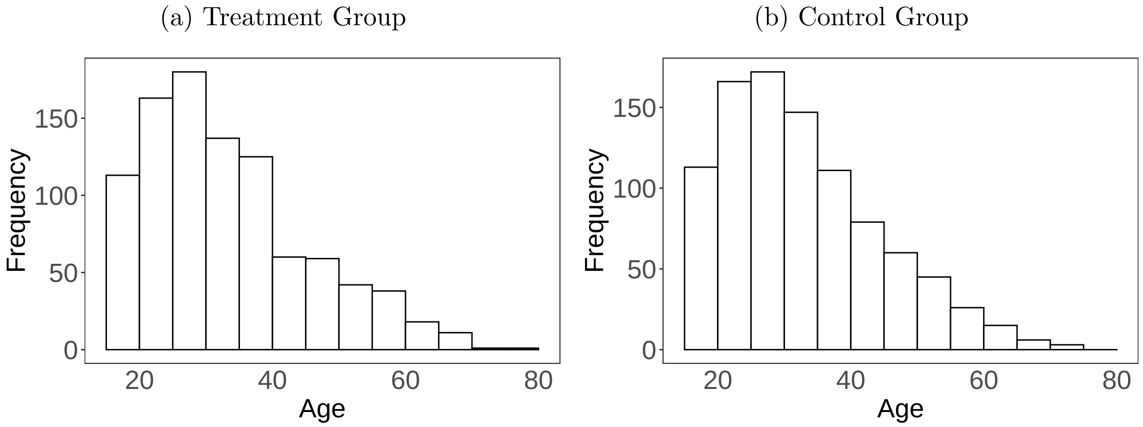


Figure S4: The Distribution of Age in the Treatment (Left Panel) and Control (Right Panel) Groups Among Arrestees.

	no PSA			PSA			
	Signature bond	Cash bond ≤$1000	Cash bond >$1000	Signature bond	Cash bond ≤$1000	Cash bond >$1000	Total (%)
22 or below	135	24	22	136	24	16	357
	(7.1)	(1.3)	(1.2)	(7.2)	(1.3)	(0.8)	(18.9)
23 – 28	158	25	23	148	29	28	411
	(8.4)	(1.3)	(1.2)	(7.8)	(1.5)	(1.5)	(21.7)
29 – 35	157	40	14	151	33	28	423
	(8.3)	(2.1)	(0.7)	(8.0)	(1.7)	(1.5)	(22.3)
36 – 45	142	22	26	133	30	22	375
	(7.5)	(1.2)	(1.4)	(7.0)	(1.6)	(1.2)	(19.9)
46 or above	113	21	21	137	14	19	325
	(6.0)	(1.1)	(1.1)	(7.2)	(0.7)	(1.0)	(17.1)

Table 2: The Joint Distribution of Treatment Assignment, Decisions, and Age. The table shows the number of cases in each category with the corresponding percentage in parentheses.

Figure S4 presents the distribution of age for the treatment and control groups. As expected, the two distribution is similar. We observe that the age distribution is right skewed with many more young arrestees. Table 2 presents the descriptive statistics for different age groups examined here. We divide the arrestees into five subgroups with different ranges of age (aged 22 or below, between 23 to 28, between 29 to 35, between 36 to 45, 46 or above). Within each age group, the signature bond appears to be the dominant decision.

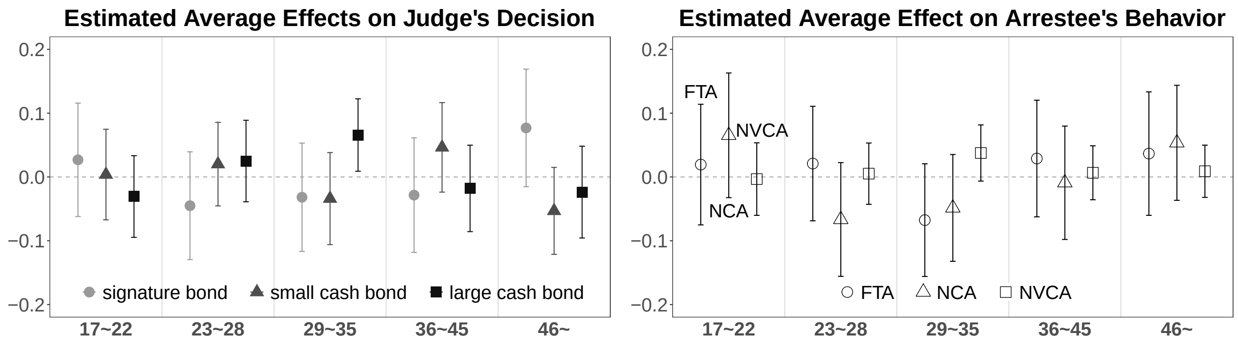


Figure S5: Estimated Average Causal Effects of PSA Provision on Judge's Decisions and Outcome Variables for First Arrest Cases (FTA, NCA, and NVCA). The results are based on the difference-in-means estimator. The vertical bars represent the 95% confidence intervals. In the left figure, we report the estimated average causal effect of the PSA provision on the decision to charge a signature bond (circles), a small cash bail ($1,000 dollars or less; triangles), and a large cash bail (greater than $1,000; squares). In the right figure, we report the estimated average causal effect of the PSA provision on the three different outcome variables: FTA (open circles), NCA (open triangles), and NVCA (open squares).

Figure S5 presents the estimated ITT effects of PSA provision on judge's decisions (top panel) and arrestee's behaviors (bottom panel). We find that the PSA provision has little effect on the judge's decisions with the exception of the 29 – 35 years old group and the oldest group. For the 29 – 35 years old group, the PSA appears to lead to a harsher decision while for the 46 or older group the effect is opposite. As for the effects on arrestee's behavior, our analysis suggests that the PSA provision may increase NVCA among female arrestees and the 29 – 35 years old group.

S2.2 Principal Stratum Proportion and Average Principal Strata Effects

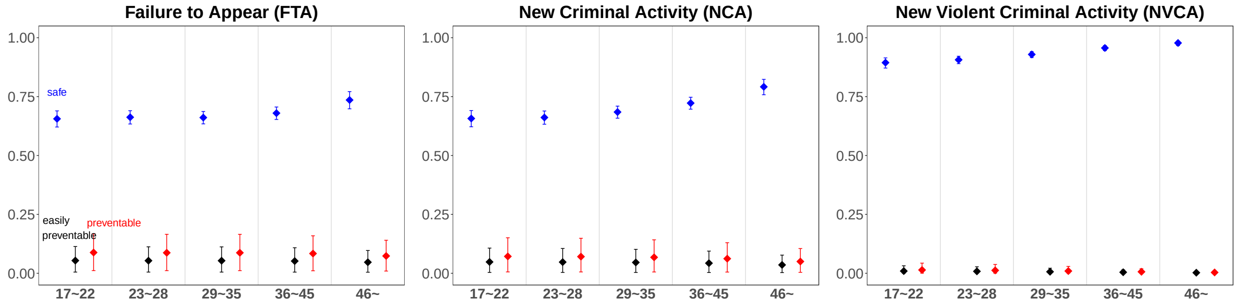


Figure S6: Estimated Population Proportion of Each Principal Stratum. Each panel represents the result using three different outcome variables (FTA, NCA, and NVCA). In each column, the blue, black, and red diamonds represent the estimates for safe, easily preventable, and preventable cases, respectively. These three estimates do not sum to one because there is an additional principal stratum of risky cases that represents a group of arrestees who will commit a new FTA (or NCA/NVCA) regardless of judges' decisions. The solid vertical lines represent 95% Bayesian credible intervals.

Figure S6 presents the estimated proportion of each principal stratum for different age groups. We observe that the principal stratum size is similar across age groups with the safe cases being the most dominant. The proportion of safe cases appears to be greater for older age groups though the rate of increase is modest. The interpretation of Figure S7 is given in the last paragraph of Section 4.2.

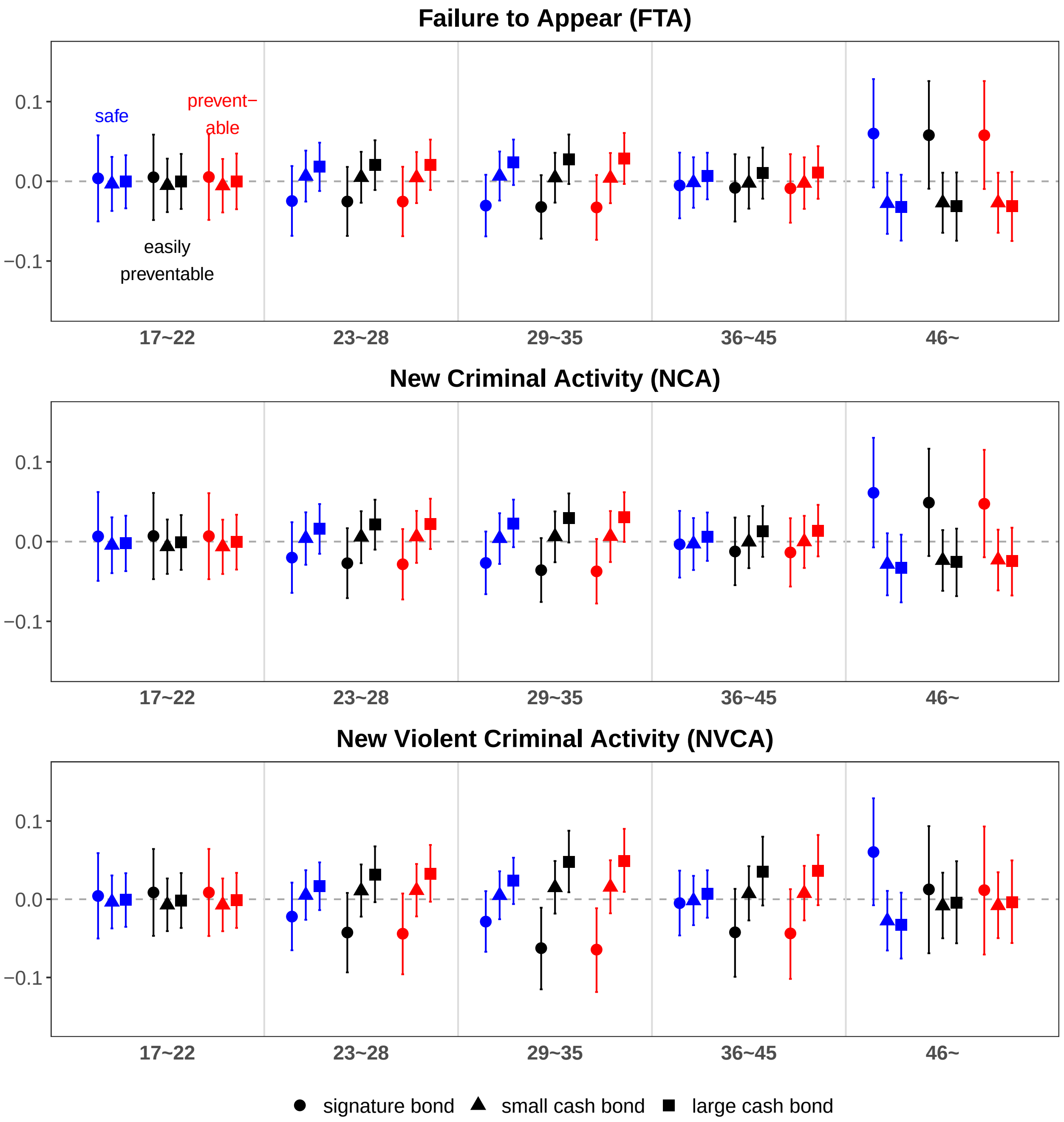


Figure S7: Estimated Average Principal Causal Effects (APCE) of the PSA Provision on Judges' Decision. Each plot presents the age group-specific results for a different outcome variable. Each column of a plot shows the estimated APCE of the PSA provision for safe (blue), easily preventable (black), and preventable (red) cases. For each of these principal strata, we report the estimated APCE on the decision to charge a signature bond (circles), a small cash bail amount of 1,000 dollars or less (triangles), and a large cash bail amount of greater than 1,000 (squares). The vertical line for each estimate represents the Bayesian 95% credible interval.

S3 Proofs of the Theorems

S3.1 Lemmas

To prove the theorems, we need some lemmas.

Lemma S1 *Consider two random variables X and Y with finite moments. Let $f_1(x)$ and $f_2(y)$ be their density functions. Then, any function $g(\cdot)$*

$$\mathbb{E}\{g(X)\} = \mathbb{E}\left\{\frac{f_1(Y)}{f_2(Y)}g(Y)\right\}.$$

Proof is straightforward and hence omitted.

Lemma S2 *For a binary decision, Assumption 4 implies $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$ under Assumption 3. For a discrete decision, Assumption 4 implies $R_i \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$ under Assumption 6.*

Proof of Lemma S2. For a binary decision, we have

$$\begin{aligned}
\Pr\{Y_i(1) = 1, Y_i(0) = 1 \mid D_i, \mathbf{X}_i, Z_i = z\} &= \Pr\{Y_i(1) = 1 \mid D_i, \mathbf{X}_i, Z_i = z\} \\
&= \Pr\{Y_i(1) = 1 \mid \mathbf{X}_i, Z_i = z\} \\
&= \Pr\{Y_i(1) = 1, Y_i(0) = 1 \mid \mathbf{X}_i, Z_i = z\},
\end{aligned}$$

where the first and third equality follow from Assumption 3 and the second equality follows from Assumption 4. Similarly, we have

$$\begin{aligned}
\Pr\{Y_i(1) = 0, Y_i(0) = 0 \mid D_i, \mathbf{X}_i, Z_i = z\} &= \Pr\{Y_i(0) = 0 \mid D_i, \mathbf{X}_i, Z_i = z\} \\
&= \Pr\{Y_i(0) = 0 \mid \mathbf{X}_i, Z_i = z\} \\
&= \Pr\{Y_i(1) = 0, Y_i(0) = 0 \mid \mathbf{X}_i, Z_i = z\},
\end{aligned}$$

where the first and third equality follow from Assumption 3 and the second equality follows from Assumption 4. As a result, $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$ because $\{Y_i(1), Y_i(0)\}$ takes only three values.

For a discrete decision D_i taking values in $\{0, \ldots, k\}$, we have

$$\begin{aligned}
\Pr(R_i = r \mid D_i, \mathbf{X}_i, Z_i = z) &= \Pr(R_i \geq r \mid D_i, \mathbf{X}_i, Z_i = z) - \Pr(R_i \geq r + 1 \mid D_i, \mathbf{X}_i, Z_i = z) \\
&= \Pr(Y_i(r-1) = 1 \mid D_i, \mathbf{X}_i, Z_i = z) - \Pr(Y_i(r) = 1 \mid D_i, \mathbf{X}_i, Z_i = z) \\
&= \Pr(Y_i(r-1) = 1 \mid \mathbf{X}_i, Z_i = z) - \Pr(Y_i(r) = 1 \mid \mathbf{X}_i, Z_i = z) \\
&= \Pr(R_i \geq r \mid \mathbf{X}_i, Z_i = z) - \Pr(R_i \geq r + 1 \mid \mathbf{X}_i, Z_i = z) \\
&= \Pr(R_i = r \mid D_i, \mathbf{X}_i, Z_i = z),
\end{aligned}$$

where the second and the fourth equality follow from the definition of R_i and the third equality follows from Assumption 4. As a result, $R_i \perp\!\!\!\perp D_i \mid \mathbf{X}_i, Z_i = z$. □

S3.2 Proof of Theorem 1

First, Assumption 3 implies,

$$\begin{aligned}
\Pr\{Y_i(0) = 0, Y_i(1) = 0\} &= \Pr\{Y_i(0) = 0\}, \quad \Pr\{Y_i(0) = 1, Y_i(1) = 1\} = \Pr\{Y_i(1) = 1\}, \\
\Pr\{Y_i(0) = 1, Y_i(1) = 0\} &= 1 - \Pr\{Y_i(0) = 0\} - \Pr\{Y_i(1) = 1\}.
\end{aligned}$$

Second, we have

$$\begin{aligned}
& \Pr\{D_i(z) = 1, Y_i(0) = 0, Y_i(1) = 0\} \\
= \; & \Pr\{Y_i(0) = 0, Y_i(1) = 0\} - \Pr\{D_i(z) = 0, Y_i(0) = 0, Y_i(1) = 0\} \\
= \; & \Pr\{Y_i(0) = 0\} - \Pr\{D_i(z) = 0, Y_i(0) = 0\} \\
= \; & \Pr\{Y_i(0) = 0\} - \Pr\{D_i(z) = 0, Y_i(D_i(z)) = 0 \mid Z_i = z\} \\
= \; & \Pr\{Y_i(0) = 0\} - \Pr(D_i = 0, Y_i = 0 \mid Z_i = z),
\end{aligned}$$

where the second equality follows from Assumption 3 and the third equality follows from Assumption 1. Similarly, we can obtain

$$\begin{aligned}
\Pr\{D_i(z) = 1, Y_i(0) = 1, Y_i(1) = 1\} &= \Pr\{D_i(z) = 1, Y_i(1) = 1\} \\
&= \Pr\{D_i(z) = 1, Y_i(D_i(z)) = 1 \mid Z_i = z\} \\
&= \Pr(D_i = 1, Y_i = 1 \mid Z_i = z).
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \Pr\{D_i(z) = 1, Y_i(0) = 1, Y_i(1) = 0\} \\
= \; & \mathrm{pr}\{D_i(z) = 1\} - \Pr\{D_i(z) = 1, Y_i(0) = 0, Y_i(1) = 0\} - \Pr\{D_i(z) = 1, Y_i(0) = 1, Y_i(1) = 1\} \\
= \; & \mathrm{pr}\{D_i = 1 \mid Z_i = z\} - \Pr\{Y_i(0) = 0\} + \Pr(D_i = 0, Y_i = 0 \mid Z_i = z) - \Pr(D_i = 1, Y_i = 1 \mid Z_i = z) \\
= \; & \Pr(Y_i = 0 \mid Z_i = z) - \Pr\{Y_i(0) = 0\}.
\end{aligned}$$

Finally, we have,

$$\begin{aligned}
\mathsf{APCEp} &= \frac{\Pr\{D_i(1) = 1, Y_i(0) = 1, Y_i(1) = 0\} - \Pr\{D_i(0) = 1, Y_i(0) = 1, Y_i(1) = 0\}}{\Pr\{Y_i(0) = 1, Y_i(1) = 0\}} \\
&= \frac{\Pr(Y_i = 1 \mid Z_i = 0) - \Pr(Y_i = 1 \mid Z_i = 1)}{\Pr\{Y_i(0) = 1\} - \Pr\{Y_i(1) = 1\}},
\end{aligned}$$

$$\begin{aligned}
\mathsf{APCEr} &= \frac{\Pr\{D_i(1) = 1, Y_i(0) = 1, Y_i(1) = 1\} - \Pr\{D_i(0) = 1, Y_i(0) = 1, Y_i(1) = 1\}}{\Pr\{Y_i(0) = 1, Y_i(1) = 1\}} \\
&= \frac{\Pr(D_i = 1, Y_i = 1 \mid Z_i = 1) - \Pr(D_i = 1, Y_i = 1 \mid Z_i = 0)}{\Pr\{Y_i(1) = 1\}},
\end{aligned}$$

and

$$\begin{aligned}
\mathsf{APCEs} &= \frac{\Pr\{D_i(1) = 1, Y_i(0) = 0, Y_i(1) = 0\} - \Pr\{D_i(0) = 1, Y_i(0) = 0, Y_i(1) = 0\}}{\Pr\{Y_i(0) = 0\}} \\
&= \frac{\Pr(D_i = 0, Y_i = 0 \mid Z_i = 0) - \Pr(D_i = 0, Y_i = 0 \mid Z_i = 1)}{\Pr\{Y_i(0) = 0\}}.
\end{aligned}$$

□

S3.3 Proof of Theorem 2

Assumption 4 and Lemma S2 imply,

$$\begin{aligned}
\mathbb{E}\{D_i(z) \mid Y_i(1) = y_1, Y_i(0) = y_0\} &= \mathbb{E}\left[\mathbb{E}\{D_i(z) \mid \mathbf{X}_i, Y_i(1) = y_1, Y_i(0) = y_0\} \mid Y_i(1) = y_1, Y_i(0) = y_0\right] \\
&= \mathbb{E}\left[\mathbb{E}\{D_i(z) \mid \mathbf{X}_i\} \mid Y_i(1) = y_1, Y_i(0) = y_0\right].
\end{aligned}$$

Based on Lemma S1,

$$\begin{aligned}
& \mathbb{E}\left[\mathbb{E}\{D_i(z) \mid \mathbf{X}_i\} \mid Y_i(1)=y_1, Y_i(0)=y_0\right] \\
= & \mathbb{E}\left[\frac{\Pr\{\mathbf{X}_i \mid Y_i(1)=y_1, Y_i(0)=y_0\}}{\Pr(\mathbf{X}_i)}\mathbb{E}\{D_i(z) \mid \mathbf{X}_i\}\right] \\
= & \mathbb{E}\left(\mathbb{E}\left[\frac{\Pr\{\mathbf{X}_i \mid Y_i(1)=y_1, Y_i(0)=y_0\}}{\Pr(\mathbf{X}_i)}D_i(z)\middle|\mathbf{X}_i\right]\right) \\
= & \mathbb{E}\left(\mathbb{E}\left[\frac{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0 \mid \mathbf{X}_i\}}{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0\}}D_i(z)\middle|\mathbf{X}_i\right]\right) \\
= & \mathbb{E}\left[\frac{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0 \mid \mathbf{X}_i\}}{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0\}}D_i(z)\right] \\
= & \mathbb{E}\left[\frac{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0 \mid \mathbf{X}_i\}}{\Pr\{Y_i(1)=y_1, Y_i(0)=y_0\}}D_i\middle|Z_i=z\right],
\end{aligned} \tag{S1}$$

where the last equality follows from Assumption 1. We can then obtain the expressions for APCEp, APCEr and APCEs by choosing different values of y_1 and y_0 in (S1). □

S3.4 Proof of Theorem 3

Assumption 1 implies,

$$\Pr\{D_i(z)=d, Y_i(d)=y\} = \Pr\{D_i(z)=d, Y_i(D_i(z))=y \mid Z_i=z\} = \Pr(D_i=d, Y_i=y \mid Z_i=z).$$

Therefore,

$$\begin{aligned}
\Pr\{D_i(z)=1 \mid Y_i(0)=y\} &= \frac{\Pr\{D_i(z)=1, Y_i(0)=y\}}{\Pr\{Y_i(0)=y\}} \\
&= \frac{\Pr\{Y_i(0)=y\} - \Pr\{D_i(z)=0, Y_i(0)=y\}}{\Pr\{Y_i(0)=y\}} \\
&= \frac{\Pr\{Y_i(0)=y\} - \Pr(D_i=0, Y_i=y \mid Z_i=z)}{\Pr\{Y_i(0)=y\}}
\end{aligned}$$

As a result, we have

$$\begin{aligned}
\mathsf{APCEp} &= \frac{\Pr(D_i=0, Y_i=1 \mid Z_i=0) - \Pr(D_i=0, Y_i=1 \mid Z_i=1)}{\Pr\{Y_i(0)=1\}}, \\
\mathsf{APCEs} &= \frac{\Pr(D_i=0, Y_i=0 \mid Z_i=0) - \Pr(D_i=0, Y_i=0 \mid Z_i=1)}{\Pr\{Y_i(0)=0\}}.
\end{aligned}$$

□

S3.5 Proof of Theorem 5

Using the law of total expectation, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid R_i=r] &= \mathbb{E}(\mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid \mathbf{X}_i, R_i=r] \mid R_i=r) \\
&= \mathbb{E}(\mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid \mathbf{X}_i] \mid R_i=r) \\
&= \mathbb{E}\left(\frac{\Pr(\mathbf{X}_i \mid R_i=r)}{\Pr(\mathbf{X}_i)}\mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid \mathbf{X}_i]\right) \\
&= \mathbb{E}\left(\frac{\Pr(R_i=r \mid \mathbf{X}_i)}{\Pr(R_i=r)}\mathbb{E}[\mathbf{1}\{D_i(z) \geq r\} \mid \mathbf{X}_i]\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{\Pr(R_i = r \mid \mathbf{X}_i)}{\Pr(R_i = r)} \mathbf{1}\{D_i(z) \geq r\}\right] \\
&= \mathbb{E}\left[\frac{\Pr(R_i = r \mid \mathbf{X}_i)}{\Pr(R_i = r)} \mathbf{1}\{D_i \geq r\} \mid Z_i = z\right],
\end{aligned}
$$

where the second equality follows from Assumption 4 and Lemma S2, and the last equality follows from Assumption 1. Thus,

$$
\mathsf{APCEp}(r) \;=\; \mathbb{E}\{w_r(\mathbf{X}_i)\mathbf{1}(D_i \geq r) \mid Z_i = 1\} - \mathbb{E}\{w_r(\mathbf{X}_i)\mathbf{1}(D_i \geq r) \mid Z_i = 0\}.
$$

We can prove the expression for APCEs similarly. □

S4 Details of the Bayesian Estimation

We only consider the algorithm for sensitivity analysis with ordinal decision since the computation of the original analysis is straightforward by setting the sensitivity parameters to zero. Consider the model given in equations (7) and (8). We can write equation (7) in terms of the observed data as,

$$
D_i^* \;=\; \beta_Z Z_i + \mathbf{X}_i^\top \beta_X + Z_i \mathbf{X}_i^\top \beta_{ZX} + \epsilon_{i1}, \tag{S2}
$$

where

$$
D_i = \begin{cases} 0 & D^* \leq \theta_{Z_i,1} \\ 1 & \theta_{Z_i,1} < D_i^* \leq \theta_{Z_i,2} \\ \vdots & \vdots \\ k-1 & \theta_{Z_i,k-1} < D_i^* \leq \theta_{Z_i,k} \\ k & \theta_{Z_i,k} < D_i^* \end{cases}.
$$

We then consider equation (8). For $r = 0, \dots, k$, because $R_i \geq r+1$ is equivalent to $Y_i(r) = 1$, we have

$$
\Pr\{Y(r) = 1\} \;=\; \Pr(R_i^* > \delta_r) = \Pr(\boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2} > \delta_r) = \Pr(-\delta_r + \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2} > 0).
$$

Therefore, we can introduce a latent variable $Y^*(r)$, and write

$$
Y_i^*(r) = -\delta_r + \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2}, \tag{S3}
$$

where $Y_i(r) = 1$ if $Y_i^*(r) > 0$ and $Y_i(r) = 0$ if $Y_i^*(r) \leq 0$. We can further write (S3) in terms of the observed data as

$$
Y_i^* = -\sum_{r=0}^{k} \delta_r \mathbf{1}(D_i = r) + \boldsymbol{X}_i^\top \alpha_X + \epsilon_{i2}, \tag{S4}
$$

where $Y_i = 1$ if $Y_i^* > 0$ and $Y_i = 0$ if $Y_i^* \leq 0$.

Combining (S2) and (S4), we have

$$
\begin{aligned}
D_i^* &= \beta_Z Z_i + \mathbf{X}_i^\top \beta_X + Z_i \mathbf{X}_i^\top \beta_{ZX} + \epsilon_{i1}, \\
Y_i^* &= -\sum_{d=0}^{k} \delta_d \mathbf{1}(D_i = d) + \mathbf{X}_i^\top \alpha_X + \epsilon_{i2},
\end{aligned}
$$

where

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

and

$$D_i = \begin{cases} 0 & D^* \le \theta_{Z_i,1} \\ 1 & \theta_{Z_i,1} < D_i^* \le \theta_{Z_i,2} \\ \vdots & \vdots \\ k-1 & \theta_{Z_i,k-1} < D_i^* \le \theta_{Z_i,k} \\ k & \theta_{Z_i,k} < D_i^* \end{cases}, \quad Y_i = \begin{cases} 0 & Y_i^* \le 0 \\ 1 & Y_i^* > 0 \end{cases}$$

with $\delta_d \le \delta_{d'}$ for $d \le d'$.

We choose multivariate normal priors for the regression coefficients, $(\beta_Z, \beta_X^\top, \beta_{ZX}^\top) \sim \boldsymbol{N}_{2p+1}(\mathbf{0}, \mathbf{\Sigma}_D)$ and $\alpha_X \sim \boldsymbol{N}_p(\mathbf{0}, \mathbf{\Sigma}_R)$. We choose the priors for θ and δ in the following manner. We first choose a normal prior for θ_{z1} and δ_0, $\theta_{z1} \sim N(0, \sigma_0^2)$ and $\delta_0 \sim N(0, \sigma_0^2)$ for $z = 0, 1$. We then choose truncated normal priors for other parameters, $\theta_{zj} \sim N(0, \sigma_0^2)\mathbf{1}(\theta_{zj} \ge \theta_{z,j-1})$ for $j = 2, \ldots, k$ and $\delta_l \sim N(0, \sigma_0^2)\mathbf{1}(\delta_l \ge \delta_{l-1})$ for $l = 1, \ldots, k$. In this way, we guarantee that θ's and δ's are increasing. In our empirical analysis, we choose $\mathbf{\Sigma}_D = 0.01 \cdot \mathbf{I}_{2p+1}$, $\mathbf{\Sigma}_D = 0.01 \cdot \mathbf{I}_p$, and $\sigma_0 = 10$

Treating Y_i^* and D_i^* as missing data, we can write the complete-data likelihood as

$$\begin{aligned} & L(\theta, \beta, \delta, \alpha) \\ = & \prod_{i=1}^n L_i(\theta, \beta, \delta, \alpha) \\ \propto & \prod_{i=1}^n \exp\left(-\frac{1}{2(1-\rho^2)} \left[(D^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX})^2 + \left\{ Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X \right\}^2 \right.\right. \\ & \qquad \left.\left. -2\rho (D^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i) \left\{ Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X \right\} \right] \right). \end{aligned}$$

Imputation Step. We first impute the missing data given the observed data and parameters. Using R package *tmvtnorm*, we can jointly sample Y_i^* and D_i^*. Given $(D_i, Y_i, Z_i, \mathbf{X}_i^\top, \theta, \beta, \alpha, \delta)$, (D_i^*, Y_i^*) follows a truncated bivariate normal distribution whose means are given by $\mathbf{X}_i^\top \beta_X + \beta_Z Z_i + Z_i \mathbf{X}_i^\top \beta_{ZX}$ and $-\sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) + \mathbf{X}_i^\top \alpha_X$, and whose covariance matrix has unit variances and correlation ρ where D^* is truncated within interval $[\theta_{zd}, \theta_{z,d+1}]$ if $Z_i = z$ and $D_i = d$ (we define $\theta_0 = -\infty$ and $\theta_{k+1} = \infty$) and Y_i^* is truncated within $(-\infty, 0)$ if $Y_i = 0$ and $[1, \infty)$ if $Y_i = 1$.

Posterior Sampling Step. The posterior distribution is proportional to

$$\begin{aligned} & \prod_{i=1}^n \exp\left(-\frac{1}{2(1-\rho^2)} \left[(D^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX})^2 + \left\{ Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X \right\}^2 \right.\right. \\ & \qquad \left.\left. -2\rho (D^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \left\{ Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) - \mathbf{X}_i^\top \alpha_X \right\} \right] \right) \\ & \cdot \exp\left\{ -\frac{(\beta_Z, \beta_X^\top, \beta_{ZX}^\top) \mathbf{\Sigma}_D^{-1} (\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top}{2} \right\} \cdot \exp\left(-\frac{\alpha_X^\top \mathbf{\Sigma}_R^{-1} \alpha_X}{2} \right) \end{aligned}$$

$$\cdot \exp\left(-\frac{\theta_{11}^2}{2\sigma_0^2}\right)\exp\left(-\frac{\delta_0^2}{2\sigma_0^2}\right)\prod_{j=2}^{k}\left\{\exp\left(-\frac{\theta_{1j}^2}{2\sigma_0^2}\right)\mathbf{1}(\theta_{1j}\geq\theta_{1,j-1})\right\}\prod_{l=1}^{k}\left\{\exp\left(-\frac{\delta_l^2}{2\sigma_0^2}\right)\mathbf{1}(\delta_l\geq\delta_{l-1})\right\}$$
$$\cdot \exp\left(-\frac{\theta_{01}^2}{2\sigma_0^2}\right)\prod_{j=2}^{k}\left\{\exp\left(-\frac{\theta_{0j}^2}{2\sigma_0^2}\right)\mathbf{1}(\theta_{0j}\geq\theta_{0,j-1})\right\}.$$

We first sample $(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)$. From the posterior distribution, we have

$$\begin{aligned}
& f(\beta_Z, \beta_X^\top, \beta_{ZX}^\top \mid \cdot) \\
\propto\ & \prod_{i=1}^{n}\exp\left(-\frac{1}{2(1-\rho^2)}\left[(D_i^* - \mathbf{X}_i^\top\beta_X - \beta_Z Z_i - Z_i\mathbf{X}_i^\top\beta_{ZX})^2\right.\right. \\
& \left.\left.-2\rho(D_i^* - \mathbf{X}_i^\top\beta_X - \beta_Z Z_i - Z_i\mathbf{X}_i^\top\beta_{ZX})\left\{Y_i^* + \sum_{d=0}^{k}\delta_d\mathbf{1}(D_i=d) - \mathbf{X}_i^\top\alpha_X\right\}\right]\right)\cdot\exp\left\{-\frac{(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top\mathbf{\Sigma}_D^{-1}(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)}{2}\right\} \\
\propto\ & \prod_{i=1}^{n}\exp\left(-\frac{1}{2(1-\rho^2)}\left[(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)(Z_i, \mathbf{X}_i^\top, Z_i\mathbf{X}_i^\top)^\top(Z_i, \mathbf{X}_i^\top, Z_i\mathbf{X}_i^\top)(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top - 2D_i^*(Z_i, \mathbf{X}_i^\top, Z_i\mathbf{X}_i^\top)(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top\right.\right. \\
& \left.\left.+2\rho\left\{Y_i^* + \sum_{d=0}^{k}\delta_d\mathbf{1}(D_i=d) - \mathbf{X}_i^\top\alpha_X\right\}(Z_i, \mathbf{X}_i^\top, Z_i\mathbf{X}_i^\top)(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top\right]\right)\cdot\exp\left\{-\frac{(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top\mathbf{\Sigma}_D^{-1}(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)}{2}\right\}.
\end{aligned}$$

Therefore, we can sample

$$(\beta_Z, \beta_X^\top, \beta_{ZX}^\top)^\top \mid \cdot \sim \boldsymbol{N}_{p+1}(\widehat{\mu}_D, \widehat{\mathbf{\Sigma}}_D),$$

where

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_D &= \left\{\frac{1}{1-\rho^2}\sum_{i=1}^{n}(Z_i, \boldsymbol{X}_i^\top, Z_i\mathbf{X}_i^\top)^\top(Z_i, \boldsymbol{X}_i^\top, Z_i\mathbf{X}_i^\top) + \mathbf{\Sigma}_D^{-1}\right\}^{-1}, \\
\widehat{\mu}_D &= \widehat{\mathbf{\Sigma}}_D\left(\frac{1}{1-\rho^2}\sum_{i=1}^{n}(Z_i, \boldsymbol{X}_i^\top, Z_i\mathbf{X}_i^\top)^\top\left[D_i^* - \rho\left\{Y_i^* + \sum_{d=0}^{k}\delta_d\mathbf{1}(D_i=d) - \mathbf{X}_i^\top\alpha_X\right\}\right]\right).
\end{aligned}$$

We then consider sampling α_X. We have

$$\begin{aligned}
& f(\alpha_X \mid \cdot) \\
\propto\ & \prod_{i=1}^{n}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left\{Y_i^* + \sum_{d=0}^{k}\delta_d\mathbf{1}(D_i=d) - \mathbf{X}_i^\top\alpha_X\right\}^2\right.\right. \\
& \left.\left.-2\rho(D_i^* - \mathbf{X}_i^\top\beta_X - \beta_Z Z_i - Z_i\mathbf{X}_i^\top\beta_{ZX})\left\{Y_i^* + \sum_{d=0}^{k}\delta_d\mathbf{1}(D_i=d) - \mathbf{X}_i^\top\alpha_X\right\}\right]\right)\cdot\exp\left(-\frac{\alpha_X^\top\mathbf{\Sigma}_R^{-1}\alpha_X}{2}\right) \\
\propto\ & \prod_{i=1}^{n}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\alpha_X^\top\mathbf{X}_i^\top\mathbf{X}_i\alpha_X - 2\left\{Y_i^* + \sum_{d=0}^{k}\delta_d\mathbf{1}(D_i=d)\right\}\mathbf{X}_i\alpha_X + 2\rho(D_i^* - \mathbf{X}_i^\top\beta_X - \beta_Z Z_i - Z_i\mathbf{X}_i^\top\beta_{ZX})\mathbf{X}_i\alpha_X\right]\right) \\
& \cdot\exp\left(-\frac{\alpha_X^\top\mathbf{\Sigma}_R^{-1}\alpha_X}{2}\right).
\end{aligned}$$

Therefore, we can sample

$$\alpha_X \mid \cdot \sim \boldsymbol{N}_p(\widehat{\mu}_R, \widehat{\mathbf{\Sigma}}_R),$$

where

$$\widehat{\mathbf{\Sigma}}_R = \left\{\frac{1}{1-\rho^2}\sum_{i=1}^{n}\boldsymbol{X}_i^\top\boldsymbol{X}_i + \mathbf{\Sigma}_R^{-1}\right\}^{-1},$$

$$\widehat{\mu}_R \;=\; \widehat{\mathbf{\Sigma}}_D \left(\frac{1}{1-\rho^2} \sum_{i=1}^n \boldsymbol{X}_i \left[\left\{ Y_i^* + \sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) \right\} - \rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \right] \right).$$

To sample δ's, we write $\sum_{d=0}^k \delta_d \mathbf{1}(D_i = d) = \delta_0 + \sum_{d=1}^k (\delta_d - \delta_{d-1}) \mathbf{1}(D_i \geq d)$ and denote $\boldsymbol{W}_i = (1, \mathbf{1}(D_i \geq 1), \ldots, \mathbf{1}(D_i \geq k))$ and $\delta = (\delta_0, \delta_1 - \delta_0, \ldots, \delta_k - \delta_{k-1})$. Thus, we have

$$\begin{aligned} & f(\delta \mid \cdot) \\ \propto \; & \prod_{i=1}^n \exp\left(-\frac{1}{2(1-\rho^2)} \left[\left\{ Y_i^* + \boldsymbol{W}_i \delta - \mathbf{X}_i^\top \alpha_X \right\}^2 - 2\rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \left\{ Y_i^* + \boldsymbol{W}_i \delta - \mathbf{X}_i^\top \alpha_X \right\} \right] \right) \\ & \quad \cdot \exp\left(-\frac{\delta_0^2}{2\sigma_0^2} \right) \prod_{d=1}^k \left\{ \exp\left(-\frac{\delta_l^2}{2\sigma_0^2} \right) \mathbf{1}(\delta_d - \delta_{d-1} \geq 0) \right\} \\ \propto \; & \prod_{i=1}^n \exp\left(-\frac{1}{2(1-\rho^2)} \left[\delta^\top \boldsymbol{W}_i^\top \boldsymbol{W}_i \delta + 2\left(Y_i^* - \mathbf{X}_i^\top \alpha_X \right) \boldsymbol{W}_i \delta - 2\rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) \boldsymbol{W}_i \delta \right] \right) \\ & \quad \cdot \exp\left(-\frac{\delta_0^2}{2\sigma_0^2} \right) \prod_{d=1}^k \left\{ \exp\left(-\frac{\delta_l^2}{2\sigma_0^2} \right) \mathbf{1}(\delta_d - \delta_{d-1} \geq 0) \right\}. \end{aligned}$$

Therefore, we can draw from a truncated normal distribution with mean and covariance matrix

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_\delta &= \left\{ \frac{1}{1-\rho^2} \sum_{i=1}^n \boldsymbol{W}_i^\top \boldsymbol{W}_i + \sigma_0^{-2} \right\}^{-1}, \\ \widehat{\mu}_\delta &= \widehat{\mathbf{\Sigma}}_D \left[\frac{1}{1-\rho^2} \sum_{i=1}^n \boldsymbol{W}_i^\top \left\{ \rho(D_i^* - \mathbf{X}_i^\top \beta_X - \beta_Z Z_i - Z_i \mathbf{X}_i^\top \beta_{ZX}) - \left(Y_i^* - \mathbf{X}_i^\top \alpha_X \right) \right\} \right], \end{aligned}$$

where the 2-th to $(k+1)$-th element is truncated within interval $[0, \infty)$. We can then transform δ to obtain $(\delta_0, \delta_1, \ldots, \delta_k)$.

Finally, we sample

$$\theta_{z1} \mid \cdot \sim TN(0, \sigma_0^2; \max_{i: Z_i = z, D_i = 0} D_i^*, \min_{i: Z_i = z, D_i = 1} (D_i^*, \theta_2)).$$

We then sample

$$\theta_{zj} \mid \cdot \sim TN(0, \sigma_0^2; \max_{i: Z_i = z, D_i = j-1} (D_i^*, \theta_{j-1}), \min_{i: Z_i = z, D_i = j} (D_i^*, \theta_{j+1}))$$

for $j = 2, \ldots, k-1$, and

$$\theta_{zk} \mid \cdot \sim TN(0, \sigma_0^2; \max_{i: Z_i = z, D_i = k-1} (D_i^*, \theta_{k-1}), \min_{i: Z_i = z, D_i = k} D_i^*).$$

The MCMC gives the posterior distributions of the parameters and therefore we can obtain the posterior distributions of $\Pr(D_i \mid R_i, \mathbf{X}_i = \mathbf{x}, Z_i = z)$ and $\Pr(R_i \mid \mathbf{X}_i = \mathbf{x})$. As a result, for $r = 1, \ldots, k$, we have

$$\begin{aligned} \mathsf{APCEp}(r) &= \Pr\{D_i(1) \geq r \mid R_i = r\} - \Pr\{D_i(0) \geq r \mid R_i = r\} \\ &= \frac{\mathbb{E}\left\{\Pr(D_i(1) \geq r, R_i = r \mid \mathbf{X}_i)\right\}}{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i)\}} - \frac{\mathbb{E}\left\{\Pr(D_i(0) \geq r, R_i = r \mid \mathbf{X}_i)\right\}}{\mathbb{E}\{\Pr(R_i = r \mid \mathbf{X}_i)\}}, \\ \mathsf{APCEs} &= \Pr\{D_i(1) = 0 \mid R_i = 0\} - \Pr\{D_i(0) = 0 \mid R_i = 0\} \\ &= \frac{\mathbb{E}\left\{\Pr(D_i(1) = 0, R_i = 0 \mid \mathbf{X}_i)\right\}}{\mathbb{E}\{\Pr(R_i = 0 \mid \mathbf{X}_i)\}} - \frac{\mathbb{E}\left\{\Pr(D_i(0) = 0, R_i = 0 \mid \mathbf{X}_i)\right\}}{\mathbb{E}\{\Pr(R_i = 0 \mid \mathbf{X}_i)\}}. \end{aligned}$$

We can calculate the conditional probabilities $\Pr\{D_i(z), R_i \mid \mathbf{X}_i\}$ and $\Pr(R_i \mid \mathbf{X}_i)$ based on the posterior sample of the coefficients, and then replace the expectation with the empirical average to obtain the estimates.

S5 Optimal PSA Provision

In this appendix, we consider the optimal PSA provision rule and conduct an empirical analysis. Let ξ be a PSA provision rule, i.e., $\xi(\mathbf{x}) = 1$ (the PSA is provided) if $\mathbf{x} \in \mathcal{B}_1$ and $\xi(\mathbf{x}) = 0$ (the PSA is not provided) if $\mathbf{x} \in \mathcal{B}_0$, where $\mathcal{X} = \mathcal{B}_0 \bigcup \mathcal{B}_1$ and $\mathcal{B}_0 \cap \mathcal{B}_1 = \emptyset$. The judges will make their decisions based on the PSA and other available information included in $\mathbf{X}_i = \mathbf{x}$. To consider the influence of the PSA on judges' decision, we define δ_{i1} the potential decision rule of case i if the judge received the PSA and δ_{i0} if not. Thus, $\delta_{iz}(\mathbf{x}) = d$ if $\mathbf{x} \in \mathcal{X}_{i,zd}$ where $\mathcal{X}_{i,zd}$ is a partition of the covariate space with $\mathcal{X} = \bigcup_{d=0}^{k} \mathcal{X}_{i,zd}$ and $\mathcal{X}_{i,zd} \cap \mathcal{X}_{i,zd'} = \emptyset$ for $z = 0, 1$. Although we allow the judge to make a different decision even if the observed case characteristics $\mathbf{X}_i$ are identical, we assume that the judges' decisions are identically distributed given the observed case characteristics and the PSA provision. That is, we assume $\Pr\{\delta_{iz}(\mathbf{x}) = d\} = \Pr\{\delta_{i'z}(\mathbf{x}) = d\}$ for fixed $\mathbf{x}$, z and $i \neq i'$, where the probability is taken with respect to the super population of all cases.

Given this setup, we derive the optimal PSA provision rule. As before, we consider the 0–1 utility $U_i(\xi) = \mathbf{1}\{\delta_{i,\xi(\mathbf{X}_i)}(\mathbf{X}_i) = R_i\}$. This utility equals one, if the judge makes the most lenient decision to prevent an arrestee from engaging in NCA (NVCA or FTA), and equals zero otherwise. As before, we begin by rewriting the expected utility in the following manner,

$$\begin{aligned}
\mathbb{E}\{U_i(\xi)\} &= \mathbb{E}\left[\mathbf{1}\{R_i = \delta_{i,\xi(\mathbf{X}_i)}(\mathbf{X}_i)\}\right] \\
&= \sum_{r=0}^{k} \mathbb{E}\left[\mathbf{1}\{R_i = r, \delta_{i,\xi(\mathbf{X}_i)}(\mathbf{X}_i) = r\}\right] \\
&= \sum_{r=0}^{k} \sum_{z=0}^{1} \mathbb{E}[\mathbf{1}\{R_i = r, \delta_{iz}(\mathbf{X}_i) = r, \mathbf{X}_i \in \mathcal{B}_z\}].
\end{aligned}$$

Under the unconfoundedness assumption, we can write,

$$\begin{aligned}
\mathbb{E}[\mathbf{1}\{R_i = r, \delta_{iz}(\mathbf{X}_i) = r, \mathbf{X}_i \in \mathcal{B}_z\}] &= \mathbb{E}[\Pr(R_i = r \mid \mathbf{X}_i) \cdot \Pr\{\delta_{iz}(\mathbf{X}_i) = r \mid \mathbf{X}_i\} \cdot \mathbf{1}\{\mathbf{X}_i \in \mathcal{B}_z\}] \\
&= \mathbb{E}[e_r(\mathbf{X}_i) \cdot \Pr\{\delta_{iz}(\mathbf{X}_i) = r\} \cdot \mathbf{1}\{\mathbf{X}_i \in \mathcal{B}_z\}].
\end{aligned}$$

Because in the experiment, the provision of PSA is randomized, we can estimate $\Pr\{\delta_{iz}(\mathbf{X}_i) = r\} = \Pr(D_i = r \mid Z_i = z, \mathbf{X}_i)$ from the data. Therefore, we obtain

$$\mathbb{E}\{U_i(\xi)\} = \sum_{z=0,1} \mathbb{E}\left(\left[\sum_{r=0}^{k} e_r(\mathbf{X}_i) \cdot \Pr(D_i = r \mid Z_i = z, \mathbf{X}_i)\right] \cdot \mathbf{1}\{\mathbf{X}_i \in \mathcal{B}_z\}\right).$$

Then, the optimal PSA provision rule is,

$$\xi(\mathbf{x}) = \operatorname*{argmax}_{z=0,1} h_z(\mathbf{x}) \quad \text{where} \quad h_z(\mathbf{x}) = \sum_{r=0}^{k} e_r(\mathbf{x}) \cdot \Pr(D_i = r \mid Z_i = z, \mathbf{X}_i). \tag{S5}$$

Thus, we can use the experimental data to derive the optimal PSA provision rule.

S6 Frequentist Analysis

In this appendix, we implement frequentist analysis and present the results. We fit the model defined in equation (S4) with probit regression. Recall that for $r = 0, \ldots, k$, $R_i \geq r + 1$ is equivalent to $Y_i(r) = 1$. Hence, we can estimate the conditional probabilities $e_r(\mathbf{X}_i)$ for each $r = 0, \ldots, k$ based on the estimates of the regression coefficients. We estimate $\mathsf{APCEp}(r)$ and APCEs using Hajek estimator as follows,

$$\begin{aligned}
\widehat{\mathsf{APCEp}}(r) &= \frac{\sum_i \hat{w}_r(\mathbf{X}_i)\mathbf{1}(D_i \geq 1)\mathbf{1}(Z_i = 1)}{\sum_i \hat{w}_r(\mathbf{X}_i)\mathbf{1}(Z_i = 1)} - \frac{\sum_i \hat{w}_r(\mathbf{X}_i)\mathbf{1}(D_i \geq 1)\mathbf{1}(Z_i = 0)}{\sum_i \hat{w}_r(\mathbf{X}_i)\mathbf{1}(Z_i = 0)}, \\
\widehat{\mathsf{APCEs}} &= \frac{\sum_i \hat{w}_0(\mathbf{X}_i)\mathbf{1}(D_i = 0)\mathbf{1}(Z_i = 1)}{\sum_i \hat{w}_0(\mathbf{X}_i)\mathbf{1}(Z_i = 1)} - \frac{\sum_i \hat{w}_0(\mathbf{X}_i)\mathbf{1}(D_i = 0)\mathbf{1}(Z_i = 0)}{\sum_i \hat{w}_0(\mathbf{X}_i)\mathbf{1}(Z_i = 0)},
\end{aligned}$$

where $\hat{w}_r(\mathbf{x}) = \hat{e}_r(\mathbf{x})/\{\frac{1}{n}\sum_i \hat{e}_r(\mathbf{X}_i)\}$. We use bootstrap to compute the 95% confidence interval.

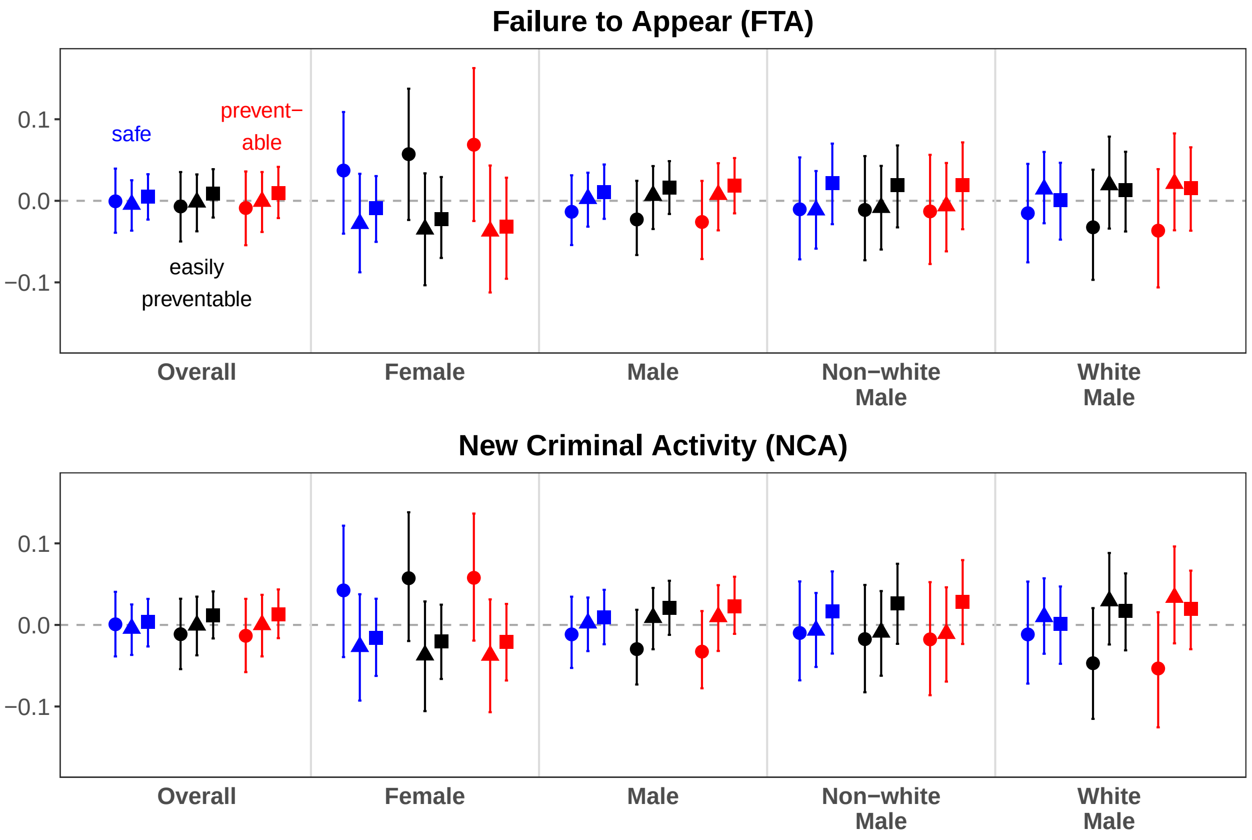


Figure S8: Estimated Average Principal Causal Effects (APCE) of PSA Provision on Judge's Decision based on Frequentist Analysis. Each plot presents the overall and subgroup-specific results for a different outcome variable. Each column of a plot shows the estimated APCE of the PSA provision for safe (blue), easily preventable (black), and preventable (red) cases. For each of these principal strata, we report the estimated APCE on the decision to charge a signature bond (circles), a small cash bail amount of 1,000 dollars or less (triangles), and a large cash bail amount of greater than 1,000 (squares). The vertical line for each estimate represents the 95% confidence interval.

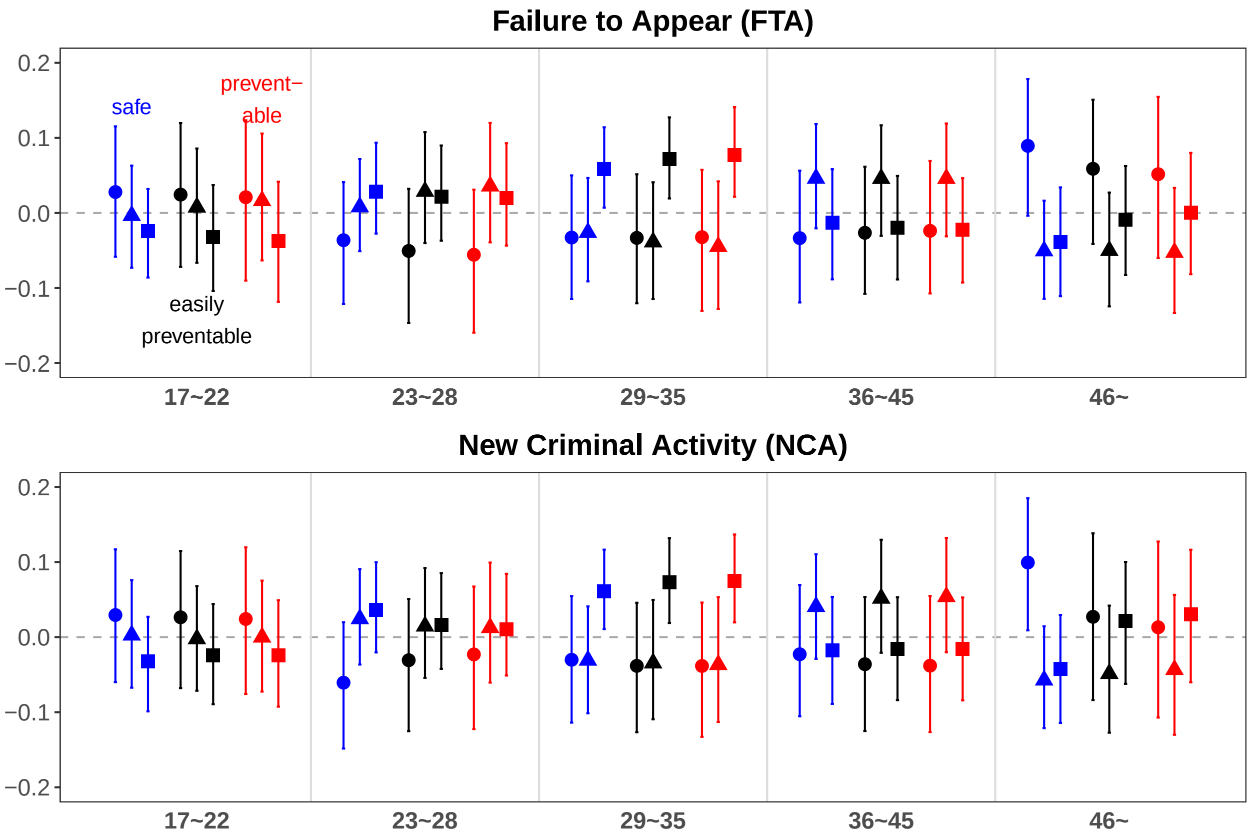


Figure S9: Estimated Average Principal Causal Effects (APCE) of PSA Provision on Judge's Decision based on Frequentist Analysis. Each plot presents the age group-specific results for a different outcome variable. Each column of a plot shows the estimated APCE of the PSA provision for safe (blue), easily preventable (black), and preventable (red) cases. For each of these principal strata, we report the estimated APCE on the decision to charge a signature bond (circles), a small cash bail amount of 1,000 dollars or less (triangles), and a large cash bail amount of greater than 1,000 (squares). The vertical line for each estimate represents the 95% confidence interval.

Figures S8 presents the estimated APCE of the PSA provision on the three ordinal decision categories, separately for FTA and NCA within each principal stratum. The results for NVCA are not presented due to the fact that the number of events is too small for an informative subgroup analysis. The results are largely consistent with those of the Bayesian analysis presented in the main text. Figure S9 presents the results for each age group similar to the one in Appendix S2.

S7 Sensitivity Analysis

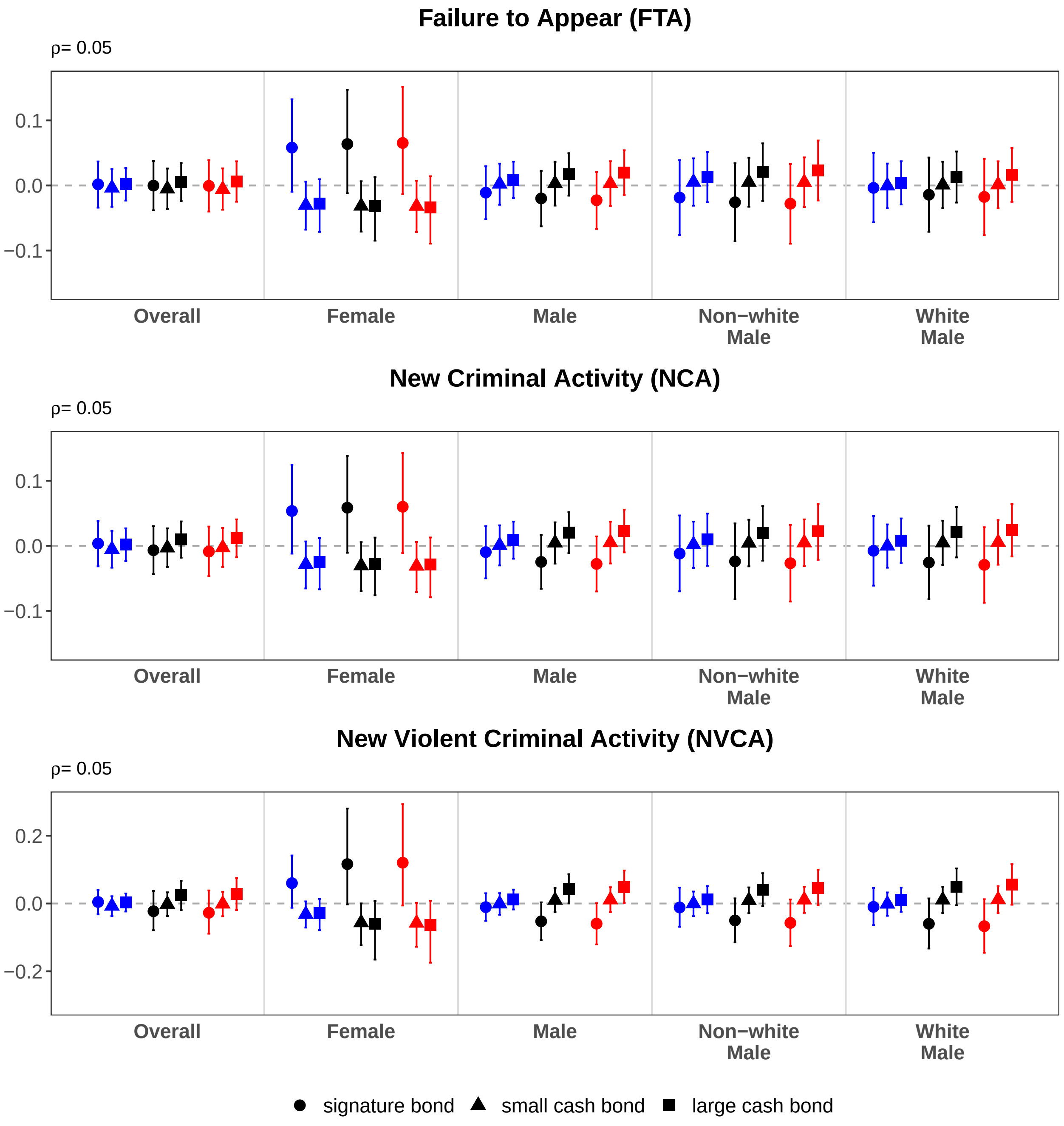


Figure S10: Estimated Average Principal Causal Effects (APCE) of PSA Provision on Judge's Decision with $\rho = 0.05$. Each plot presents the overall and subgroup-specific results for a different outcome variable. Each column of a plot shows the estimated APCE of the PSA provision for safe (blue), easily preventable (black), and preventable (red) cases. For each of these principal strata, we report the estimated APCE on the decision to charge a signature bond (circles), a small cash bail amount of 1,000 dollars or less (triangles), and a large cash bail amount of greater than 1,000 (squares). The vertical line for each estimate represents the Bayesian 95% credible interval.

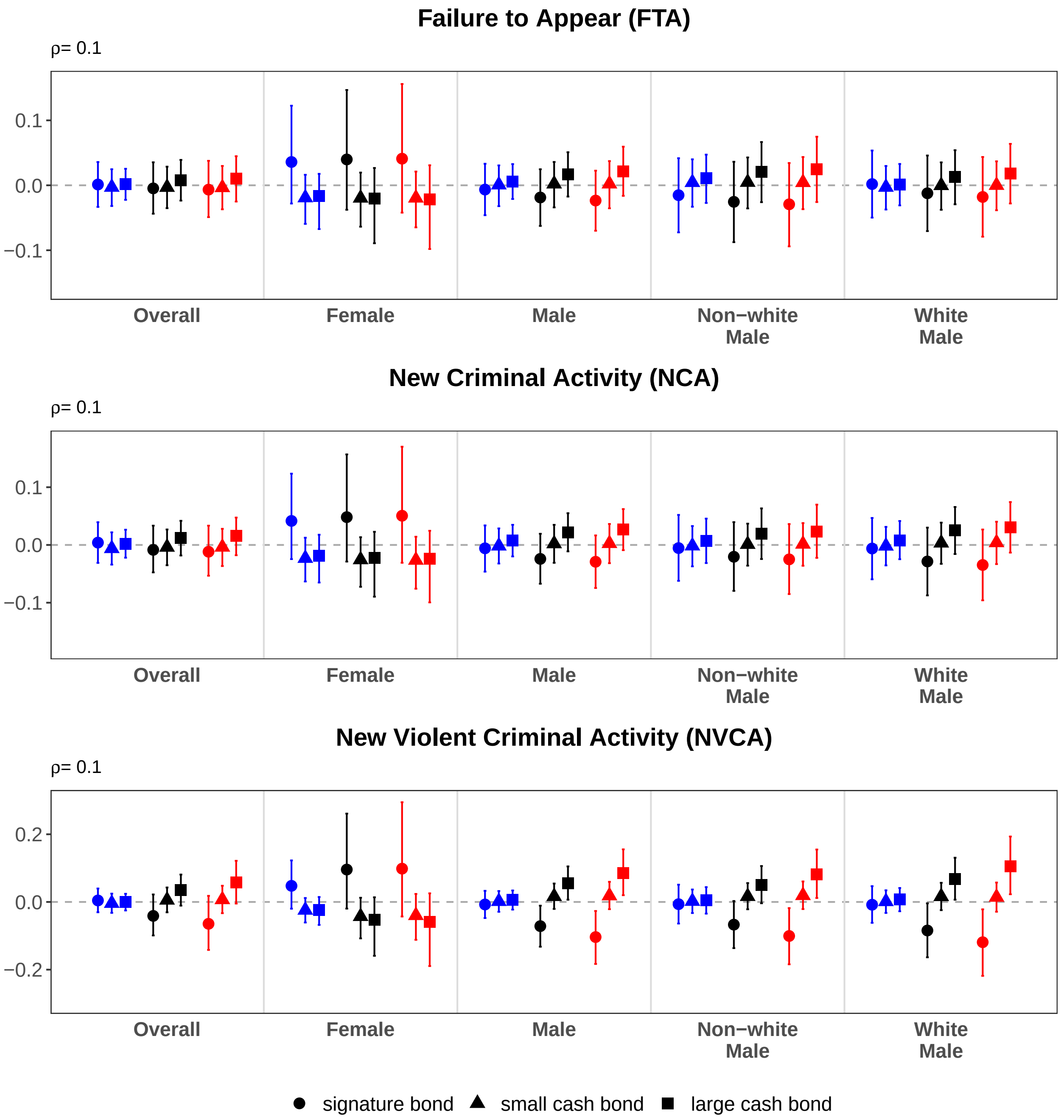
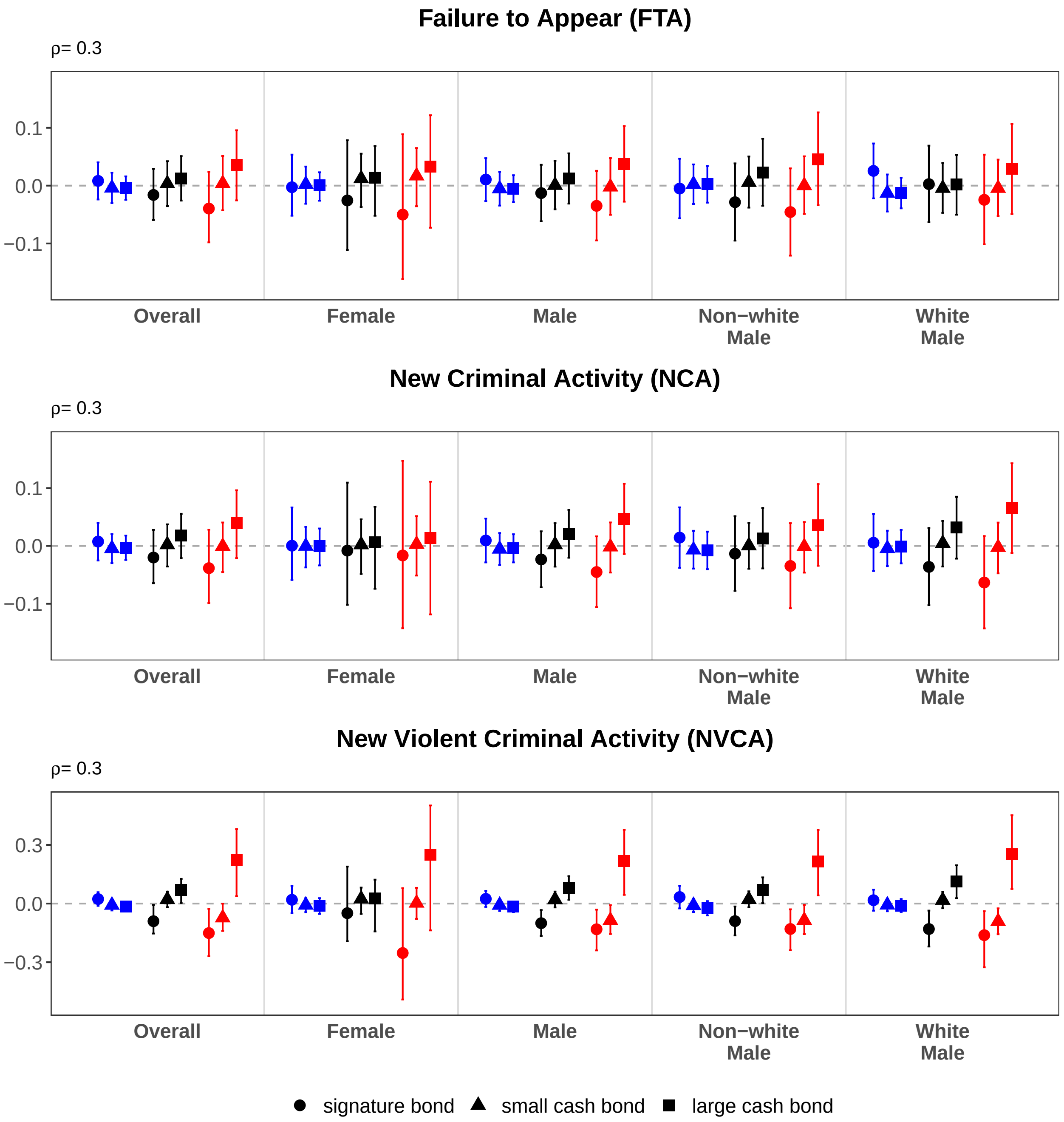


Figure S11: Estimated Average Principal Causal Effects (APCE) of PSA Provision on Judge's Decision with $\rho = 0.1$. Each plot presents the overall and subgroup-specific results for a different outcome variable. Each column of a plot shows the estimated APCE of the PSA provision for safe (blue), easily preventable (black), and preventable (red) cases. For each of these principal strata, we report the estimated APCE on the decision to charge a signature bond (circles), a small cash bail amount of 1,000 dollars or less (triangles), and a large cash bail amount of greater than 1,000 (squares). The vertical line for each estimate represents the Bayesian 95% credible interval.

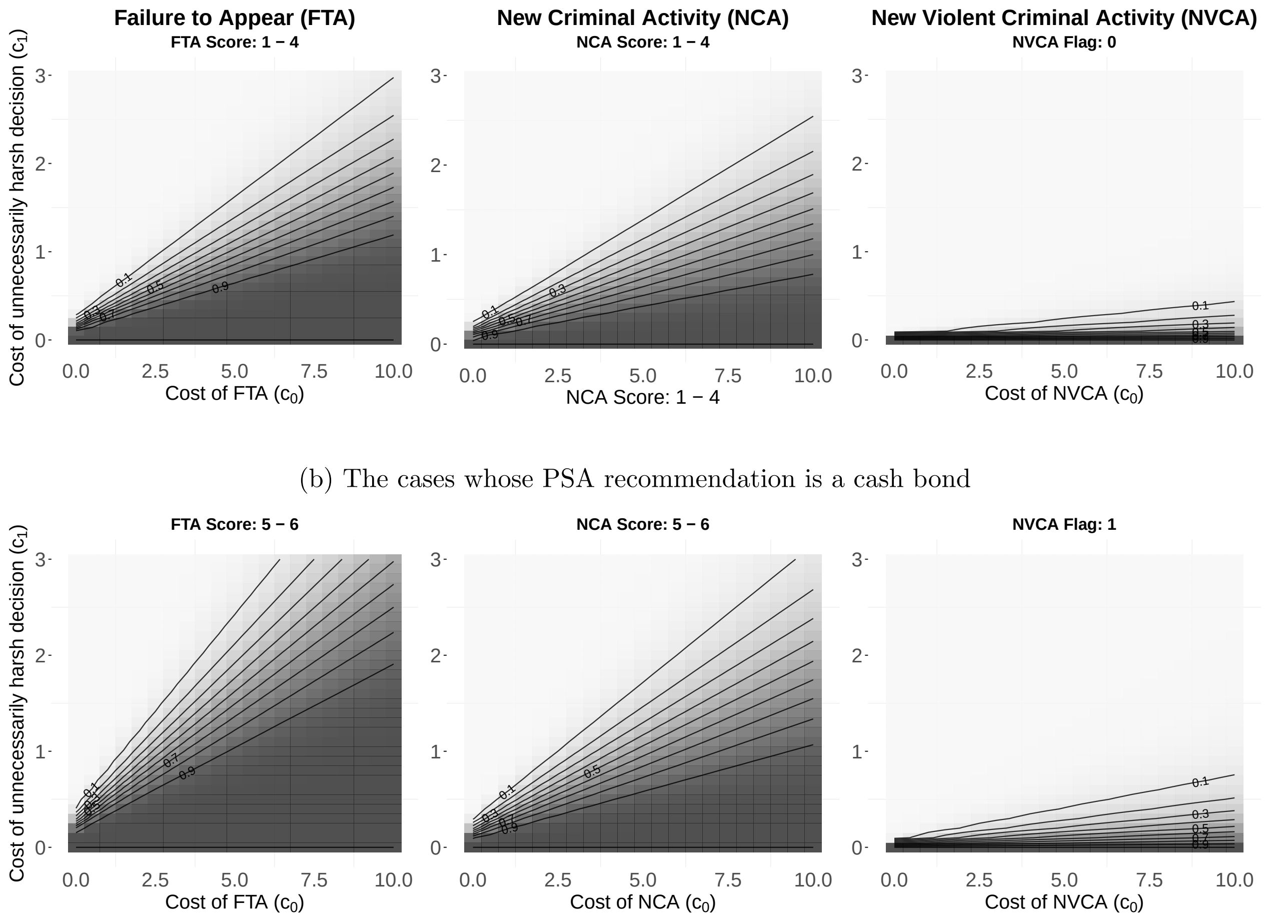

Failure to Appear (FTA)

New Criminal Activity (NCA)

New Violent Criminal Activity (NVCA)

Figure S12: Estimated Average Principal Causal Effects (APCE) of PSA Provision on Judge's Decision with $\rho = 0.3$. Each plot presents the overall and subgroup-specific results for a different outcome variable. Each column of a plot shows the estimated APCE of the PSA provision for safe (blue), easily preventable (black), and preventable (red) cases. For each of these principal strata, we report the estimated APCE on the decision to charge a signature bond (circles), a small cash bail amount of 1,000 dollars or less (triangles), and a large cash bail amount of greater than 1,000 (squares). The vertical line for each estimate represents the Bayesian 95% credible interval.

S8 Additional Results for Optimal Decision

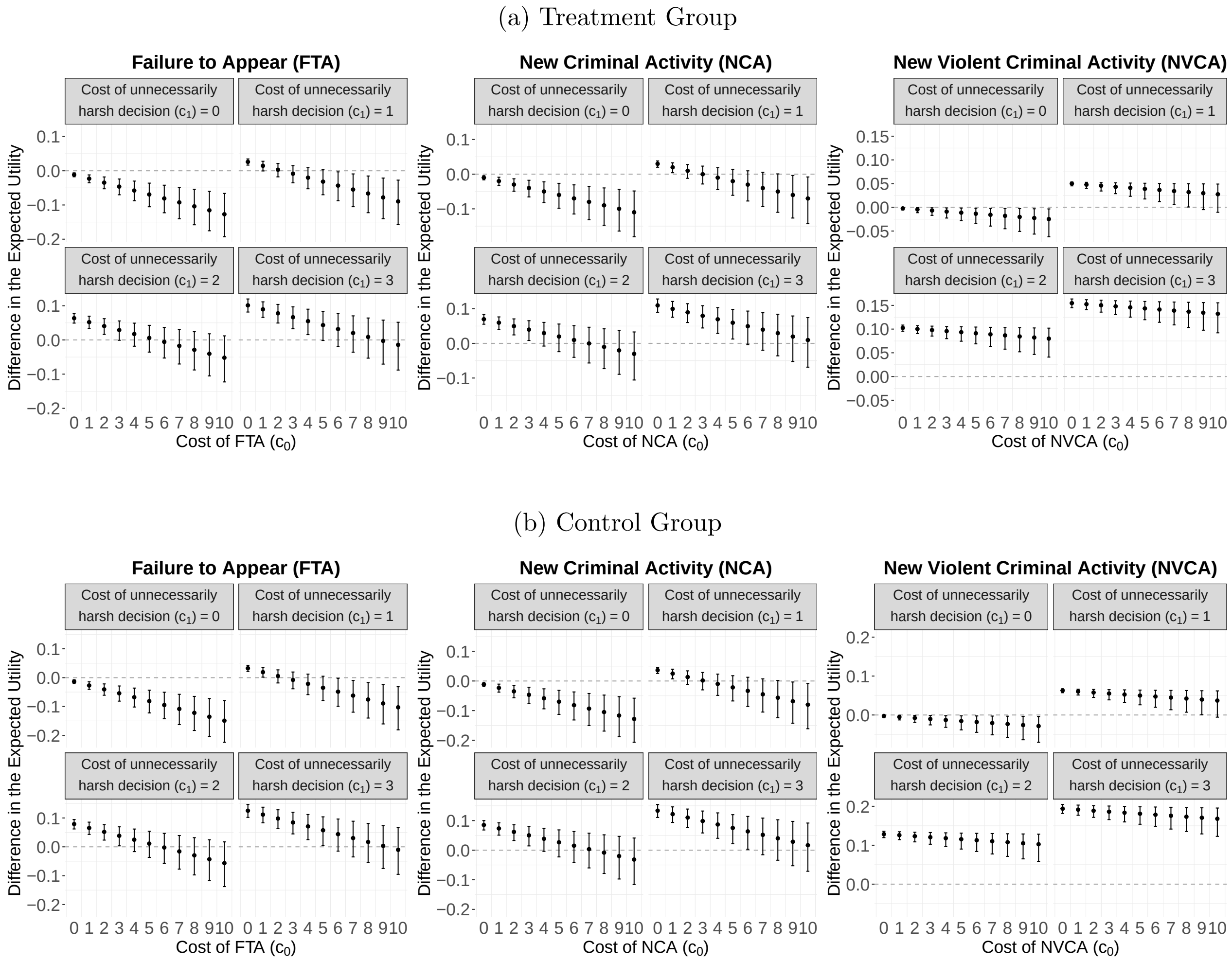


(a) The cases whose PSA recommendation is a signature bond

(b) The cases whose PSA recommendation is a cash bond

Figure S13: Estimated Proportion of Cases for Which Cash Bond is Optimal. Each column represents the results based on one of the three outcomes (FTA, NCA, and NVCA). The top (bottom) panel shows the results for the cases whose PSA recommendation is a signature (cash) bond. Unlike Figure 6, which uses the combined PSA recommendation, the results are based on the separate PSA recommendation for each outcome. In each plot, the contour lines represents the estimated proportion of cases, for which a cash bond is optimal, given the cost of an unnecessarily harsh decision (y-axis) and that of a negative outcome (x-axis). A grey area represents a greater proportion of such cases.

S9 Additional Results for Comparison between the Judge's Decisions and PSA Recommendation


(a) Treatment Group

(b) Control Group

Figure S14: Estimated Difference in the Expected Utility under Selected Values of Cost Parameters between the Judge's Decisions and PSA Recommendations for the Treatment (top row) and Control (bottom row) Group. Each column represents the results base on one of the three outcomes, given the cost of an unnecessarily harsh decision (c_1; each panel) and that of a negative outcome (c_0; x-axis). A positive value implies that the Judge's decision yields a higher expected utility (i.e., more optimal) than the corresponding PSA recommendation. The vertical line for each estimate represents the Bayesian 95% credible interval.