WORKING PAPER · NO. 2019-23

Incarceration, Recidivism, and Employment

Manudeep Bhuller, Gordon B. Dahl, Katrine V. Løken, Magne Mogstad

FEBRUARY 2019

Becker Friedman Institute | Ronzetti Initiative for the Study of Labor Markets

1126 E. 59th St, Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Incarceration, Recidivism, and Employment

Manudeep Bhuller* Gordon B. Dahl† Katrine V. Løken‡ Magne Mogstad§

February 11, 2019

Abstract: Understanding whether, and in what situations, time spent in prison is criminogenic or preventive has proven challenging due to data availability and correlated unobservables. This paper overcomes these challenges in the context of Norway's criminal justice system, offering new insights into how incarceration affects subsequent crime and employment. We construct a panel dataset containing the criminal behavior and labor market outcomes of the entire population, and exploit the random assignment of criminal cases to judges who differ systematically in their stringency in sentencing defendants to prison. Using judge stringency as an instrumental variable, we find that imprisonment discourages further criminal behavior, and that the reduction extends beyond incapacitation. Incarceration decreases the probability an individual will reoffend within 5 years by 29 percentage points, and reduces the number of offenses over this same period by 11 criminal charges. In comparison, OLS shows positive associations between incarceration and subsequent criminal behavior. This sharp contrast suggests the high rates of recidivism among ex-convicts is due to selection, and not a consequence of the experience of being in prison. Exploring factors that may explain the preventive effect of incarceration, we find the decline in crime is driven by individuals who were not working prior to incarceration. Among these individuals, imprisonment increases participation in programs directed at improving employability and reducing recidivism, and ultimately, raises employment and earnings while discouraging further criminal behavior. For previously employed individuals, while there is no effect on recidivism, there is a lasting negative effect on employment. Contrary to the widely embraced 'nothing works' doctrine, these findings demonstrate that time spent in prison with a focus on rehabilitation can indeed be preventive for a large segment of the criminal population.

Keywords: crime, employment, incarceration, recidivism
JEL codes: K42, J24
Acknowledgments: This paper was initially submitted in July 2016. We thank the editor, three anonymous referees, Derek Neal, Isaiah Andrews, Azeem Shaikh, Vishal Kamat and seminar participants at several universities and conferences for valuable feedback and suggestions. We are grateful to Baard Marstrand for help accessing the data and understanding institutional details, Martin E. Andresen in estimating the marginal treatment effects, and Max Kellogg in conducting the Monte Carlo simulations. The project 240653 received generous financial support from the Norwegian Research Council.

*Department of Economics, University of Oslo; Research Department, Statistics Norway; IZA; CESifo (email: manudeep.bhuller@econ.uio.no)

†Department of Economics, UC San Diego; Department of Economics, University of Bergen; NBER; IZA; CESifo (email: gdahl@ucsd.edu)

‡Department of Economics, Norwegian School of Economics; Research Department, Statistics Norway; Department of Economics, University of Bergen; IZA (email: katrine.loken@nhh.no)

§Department of Economics, University of Chicago; Research Department, Statistics Norway; Department of Economics, University of Bergen; NBER; IZA; CESifo (email: magne.mogstad@gmail.com)

1 Introduction

Over the past several decades, incarceration rates have risen dramatically in many OECD countries. In the U.S., for example, the incarceration rate has increased from 220 per 100,000 residents in 1980 to over 700 per 100,000 in 2012. In Europe, the increases (and levels) tend to be smaller but still substantial, with the average incarceration rate per 100,000 residents rising from 62 in 1980 to 112 in 2010 in Western European nations.[1] These increases raise important questions about how well ex-convicts reintegrate into society after incarceration, and in particular, whether they return to a life of crime. Prison time could convince offenders that crime does not pay, or rehabilitate them by providing vocational and life skills training. Conversely, prison time could cause human capital to depreciate, expose offenders to hardened criminals, or limit opportunities due to employment discrimination or societal stigma. Indeed, the effects of incarceration could vary in magnitude and sign depending on a prisoner's background (e.g., work history), as well as prison conditions (e.g., availability of prison programs and sentence lengths).

Understanding whether, and in what situations, time spent in prison is criminogenic or preventive has proven challenging for several reasons. One problem is data availability. The ideal dataset would be a long and representative panel with individual-level information on criminal behavior and labor market outcomes. In many countries, however, the required data sources cannot be accessed and linked together. Another major challenge is the threat to identification from correlated unobservables. While ex-convicts have relatively high rates of criminal activity and weak labor market attachment, these correlations could be driven by their unobserved characteristics as opposed to the experience of being in prison.

Due to these challenges, evidence on the causal effects of incarceration is scarce. Nagin et al. (2009), in their review article, summarize the state of the literature well: "Remarkably little is known about the effects of imprisonment on reoffending. The existing research is limited in size, in quality, [and] in its insights into why a prison term might be criminogenic or preventative." Our paper overcomes both the data and the identification challenges in the context of Norway's criminal justice system, offering new insights into how imprisonment affects subsequent criminal behavior.

Our work draws on two strengths of the Norwegian environment. First, by linking several administrative data sources we are able to construct a panel dataset containing complete records of the criminal behavior and labor market outcomes of every Norwegian. Second, we

[1] These figures come from the World Prison Brief (2016). The Western European countries used to construct the population-weighted average include Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the UK.

address threats to identification by exploiting the random assignment of criminal cases to Norwegian judges who differ systematically in their stringency. In our baseline specification, we measure judge stringency as the average incarceration rate in other cases a judge has handled. This serves as an instrument for incarceration since it is highly predictive of the judge's decision in the current case, but as we document, uncorrelated with observable case characteristics.

Our paper offers three sets of results. First, imprisonment discourages further criminal behavior. Using our measure of judge stringency as an instrument, we estimate that incarceration lowers the probability of reoffending within 5 years by 29 percentage points and reduces the corresponding number of criminal charges per individual by 11. These reductions are not simply due to an incapacitation effect. We find sizable decreases in reoffending probabilities and cumulative charged crimes even after defendants are released from prison.

Second, bias due to selection on unobservables, if ignored, leads to the erroneous conclusion that time spent in prison is criminogenic. Consistent with existing descriptive work, our OLS estimates show positive associations between incarceration and subsequent criminal behavior. This is true even when we control for a rich set of demographic and crime category controls. Using the panel structure of our data reduces the estimates somewhat, but there are noticeable changes in crime and employment in the year prior to the court case, raising concerns about the validity of offender fixed effects or lagged dependent variable models. In contrast, our IV estimates, which address the issues of selection bias and reverse causality, find that incarceration is strongly preventive for many individuals, both on the extensive and intensive margins of crime.

Third, the reduction in crime is driven by individuals who were not working prior to incarceration. Among these individuals, imprisonment increases participation in programs directed at improving employability and reducing recidivism, and ultimately, raises employment and earnings while discouraging criminal behavior.[2] The effects of incarceration for this group are large and economically important. Imprisonment causes a 35 percentage point increase in participation in job training programs for the previously nonemployed, and within 5 years, their employment rate increases by 36 percentage points. At the same time, the likelihood of reoffending within 5 years is cut in half (by 43 percentage points), and the average number of criminal charges falls by 18. A very different pattern emerges for individuals who were previously attached to the labor market. Among this group, which comprises roughly half of our sample, there is no significant effect of incarceration on either

[2]Since we observe charges and not actual crimes committed, it is in theory possible that ex-convicts do not, in fact, reduce their criminal activity, but rather learn how to avoid being caught while in prison. The fact that incarceration increases formal sector employment, which is a time substitute for criminal activity, suggests this explanation is unlikely.

the probability of reoffending or the number of charged crimes. Moreover, they experience an immediate 30 percentage point drop in employment due to incarceration and this effect continues out to 5 years. This drop is driven almost entirely by defendants losing their job with their previous employer while they are in prison. These heterogeneous effects based on prior employment status are important to keep in mind when interpreting our results.

Taken together, our findings have important implications for ongoing policy debates over the growth in incarceration rates and the nature of prison. A natural question is whether the positive effects from imprisonment found in Norway pass a cost-benefit test. While it is difficult to quantify both costs and benefits, rough calculations presented at the end of the paper suggest the high rehabilitation expenditures in Norway are more than offset by the corresponding benefits to society.

Our estimates indicate that the high rates of recidivism among ex-convicts is due to selection, and not a consequence of the experience of being in prison. Indeed, the Norwegian prison system is successful in discouraging crime and encouraging employment, largely due to changes in the behavior of individuals who were not working prior to incarceration. These individuals had no job to lose, and low levels of education and work experience. Norwegian prisons offer them access to rehabilitation programs, job training and re-entry support. Upon release, these previously unemployed individuals become more attached to the formal labor market, and find crime relatively less attractive. In contrast, for individuals with some attachment to the labor market, many of them had an actual job to loose and human capital to depreciate by going to prison. These negative effects may well offset any positive impacts of rehabilitation, and therefore help explain why incarceration does not seem to materially affect their criminal behavior or labor market outcomes.

Our paper contributes to a large literature across the social sciences on how incarceration affects both recidivism and future employment. Much of this literature focuses on incapacitation effects, finding reductions in crime while offenders are in prison.[3] There is less evidence on longer-term recidivism, and the findings are mixed. In terms of labor market outcomes, OLS studies usually find either negative or no effect on earnings and employment.[4] More sophisticated work uses panel data and offender fixed effects to minimize selection issues. For recidivism, there are fewer studies using this approach and the evidence is mixed, while for labor market outcomes a handful of studies find either no impact or a negative effect.[5]

[3]Recent studies in economics isolating incapacitation effects include Barbarino and Mastrobuoni (2014), Buonanno and Raphael (2013), and Owens (2009). We refer to Chalfin and McCrary (2017) for a recent review of the extensive literature on criminal deterrence.

[4]For example, Bernburg et al. (2006), Gottfredson (1999), and Brennan and Mednick (1994) all reach different conclusions for recidivism. For a summary of observational research on labor market outcomes, see Western et al. (2001).

[5]See Freeman (1992) and Western and Beckett (1999) for early papers using panel data. Other evidence

More closely related to our paper, some recent work has relied on the quasi-random assignment of judges to study the effects of incarceration.[6] While each of these studies uses data from the U.S, the findings are mixed. Kling (2006) presents results suggesting that time in prison improves labor market outcomes after release, although the IV estimates based on quasi-random assignment of judges are too imprecise to draw firm conclusions. Green and Winik (2010) and Loeffler (2013) report no detectable effects of incarceration on recidivism, whereas Aizer and Doyle (2015) find that juvenile incarceration results in lower high school completion rates and higher adult incarceration rates. Mueller-Smith (2015) uses data from Texas to investigate the impacts of adult incarceration and reports that incarceration increases recidivism rates, and worsens labor market outcomes.

There are several possible reasons why no consensus has emerged as to how well ex-convicts reintegrate into society. While quasi-random assignment of judges can be useful to address concerns over correlated unobservables, there remain issues that could bias the estimates. In Green and Winik (2010), for instance, the estimation sample is small and the instrument is weak, which may lead to severe bias in the IV estimates. Mueller-Smith (2015) additionally explores the importance of two other issues. He argues in his setting that standard instrumental variable estimates could be biased due to violation of the exclusion and monotonicity assumptions. To assess the relevance and validity of our instrument, we therefore perform a number of checks, all of which suggest that our instrument is strong, as good as randomly assigned, and satisfies exclusion and monotonicity.

Another possible explanation for the lack of consensus is that incarceration effects could vary depending on a prisoner's background or prison conditions. As documented later, prisoners in Norway have broadly similar observable characteristics as prisoners in many other countries. Instead, what is quite distinct, especially compared to the U.S., is the prison system. In Scandinavian countries like Norway, the prison system focuses on rehabilitation, preparing inmates for life on the outside.[7] This is done in part by investing in education and training programs, but also through extensive use of "open prisons" in which prisoners are

based on fixed effects or even study design include Grogger (1995), Kling (1999), Skardhammer and Telle (2012), and Waldfogel (1994).

[6]Similar designs in related contexts include Dobbie et al. (2018) and Stevenson (2018), which use the detention tendencies of quasi-randomly assigned bail judges to estimate the causal effects of pre-trial detention, and Di Tella and Schargrodsky (2013) which investigates the use of electronic monitoring as an alternative to prison. For studies using quasi-random assignment of examiners or judges in contexts other than crime, see e.g. Autor et al. (forthcoming), Belloni et al. (2012), Dahl et al. (2014), Dobbie and Song (2015), Doyle (2007, 2008), Doyle et al. (2012), French and Song (2014), and Maestas et al. (2013).

[7]A recent New York Times article summarizes the system's rehabilitative aims: "The goal of the Norwegian penal system is to get inmates out of it... 'Better out than in' is an unofficial motto of the Norwegian Correctional Service... It works with other government agencies to secure a home, a job and access to a supportive social network for each inmate before release."

housed in low-security surroundings and allowed frequent visits to families while electronically monitored.[8] In comparison, in many other countries rehabilitation has taken a back seat in favor of prison policies emphasizing punishment and incapacitation. In the U.S., a pivotal point was the 1974 Martinson report, concluding that "nothing works" in rehabilitating prisoners (Martinson, 1974; Lipton et al., 1975). While influential, leading criminology scholars have questioned the evidence base for this conclusion (e.g., see the review in Cullen, 2005). Our study serves as a proof-of-concept demonstrating that time spent in prison with a focus on rehabilitation can indeed be preventive.[9]

The remainder of the paper proceeds as follows. The next section provides background on the Norwegian court system, describes how criminal cases are assigned to judges, and outlines the baseline IV model. Section 3 presents our data. This section also describes similarities and differences in the criminal population and the criminal justice system of Norway versus other countries. In Section 4, we discuss our instrument and its validity. Section 5 presents our main results for recidivism, while Section 6 documents the important role of employment in reducing recidivism. Section 7 concludes.

2 Research Design

In this section, we describe our research design. We begin by reviewing key aspects of the criminal justice system in Norway, documenting how criminal court cases are randomly assigned to judges. We then describe how to use this randomization to estimate the effects of incarceration on subsequent criminal behavior and labor market outcomes.

2.1 The Norwegian Court System

The court system in Norway consists of three levels: the district court, the court of appeals, and the supreme court. The vast majority of cases are settled at the district court level. In this paper, we focus on criminal cases tried in one of the 87 district courts in existence at one time or another in Norway during the period of our study. The largest district court is located in Oslo and has around 100 judges, while the smallest courts only have a few judges.

There are two types of professional judges in district courts, regular judges and deputy

[8] Other countries are trying open prisons and finding positive results (Mastrobuoni and Terlizzese, 2015).

[9] The existing evidence base is scarce, and does not answer our research question of whether, and in what situations, imprisonment as compared to not being incarcerated is preventive or criminogenic. Kuziemko (2013) uses data on inmates in Georgia, and finds that access to parole boards increases participation in rehabilitation programs and reduces recidivism. There are also a few randomized controlled trials in the U.S. focusing primarily on post-release training and education programs for ex-convicts. These studies have estimated zero or small (and often imprecise due to small samples) effects on long-term labor market and recidivism outcomes (see Cook et al. 2015, Redcross et al. 2012, and Visher et al. 2005).

judges. Regular judges are appointed civil servants, and can only be dismissed for malfeasance. One of the regular judges is appointed as chief judge to oversee the administration of the local court. In 2010 there were 370 full-time regular judges (including chief judges); their average age was 53 and 62 percent were male. Deputy judges, like regular judges, are also law school graduates, but are appointed to a court for a limited period of time which cannot exceed three years (five years in Oslo). Deputy judges have a somewhat different caseload compared to regular judges, as discussed in the next subsection. Not all deputy judges become regular judges, and those that do typically need several of years of experience in other legal settings before applying for and being appointed as a regular judge.

Criminal cases are classified into two broad types, confession and non-confession cases. Both types are settled by trial (as opposed to the U.S. which has plea bargains). In confession cases, the accused has confessed to the police/prosecutor before his case is assigned to a judge. The confession is entered into evidence, but the prosecution is not absolved of the duty to present a full case and the judge may still decide that the defendant is innocent.[10] In practice, most confession cases are relatively straightforward. To save on time and costs, they are therefore heard by a single professional judge who decides on sentencing. Non-confession cases are heard by a panel of one professional and two lay judges, or in the case of extremely serious crimes, by two professional judges and three lay judges. The lay judges are individuals chosen from the general population to serve for a limited four year term. The professional judge presides over the case, while the lay judges participate on the questions of guilt and sentencing. As opposed to professional judges, lay judges hear only a few cases a year.[11]

One advantage of the Norwegian criminal justice system compared to some other countries is that it has no plea bargaining. For example, in the U.S. criminal defendants often know their assigned judge before deciding whether to plead guilty in exchange for a reduced sentence. The fact that these pre-trial strategies are not taking place in our setting makes the interpretation of our IV estimates easier to interpret (see Dobbie et al. (2018)). Moreover, in Norway, the judge handling the criminal court case is not necessarily the same as the pre-trial custody judge, with random re-assignment of judges for the court case.[12]

Figure 1 charts how suspected crimes are processed in Norway's criminal justice system.

[10]These rules apply to most civil law systems, in contrast to common law systems where a majority of criminal cases are settled by confession and plea bargain rather than by a trial.

[11]Lay judges must satisfy certain requirements, such as not having a criminal record and not working in certain occupations (e.g., police officer). In a municipal district the pool of lay judges is usually between 30-60 individuals. Lay judges are partially compensated for days absent from work if not covered by their employer. We do not observe the identify of the lay judges in our data, but since they are randomly assigned to judges within a court, they should not create any bias in our estimates.

[12]We verified the random re-assignment of judges by comparing the actual probability of receiving the same judge in both the court case and the custody case relative to the counterfactual probability from random assignment. The difference was close to zero and not statistically significant.

The figure reports percentages for the period 2005-2009. If the police suspect an individual of a crime, they file a formal report. A public prosecutor then decides whether the individual should be charged with a crime as well as whether the case should proceed to a court trial. As reported in the figure, about half of police reports lead to a formal criminal charge. Of these charged cases, the public prosecutor advances 43% of them to a trial. The other charged cases are either dismissed, directly assigned a fine, or sent to mediation by the public prosecutor. Around 60% of the cases that proceed to trial are non-confession cases. Once a case proceeds to trial, it is assigned to a judge. If the judge finds the accused guilty, he or she can assign a combination of possible punishments which are not necessarily mutually exclusive. In the figure, we show percentages based on the strictest penalty received, so that the percentages add up to 100%. Just over half of cases result in incarceration, with probation, community service and fines combined accounting for 44% of outcomes. In a small fraction of cases (5%), the defendant is found not guilty.

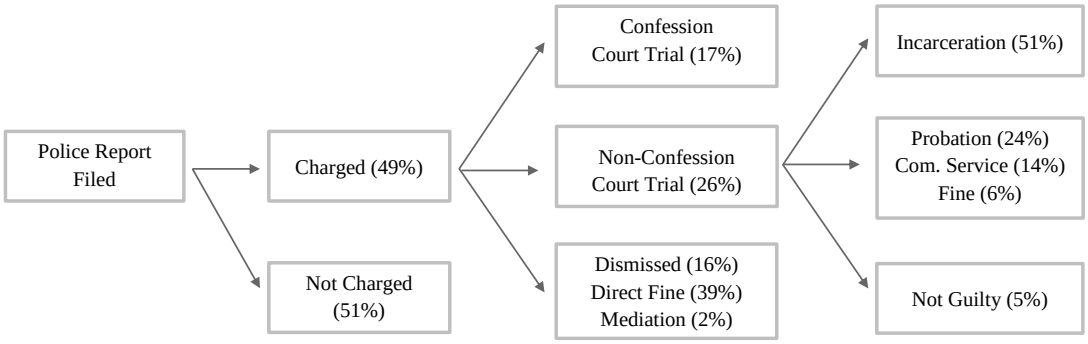


Figure 1. Processing of Suspected Crimes in Norway's Criminal Justice System.

Note: Sample consists of all criminal cases reported to the police in Norway between 2005-2009.

2.2 Assignment of Cases to Judges

In Norway, the law dictates that cases be assigned to judges according to the "principle of randomization" (Bohn, 2000; NOU, 2002). The goal is to treat all cases ex-ante equally and prevent outsiders from influencing the process of the criminal justice system. In practice, cases are assigned by the chief judge to other judges on a mechanical, rotating basis based on the date a case is received. Each time a new case arrives, it is assigned to the next judge on the list, with judges rotating between criminal and civil cases.[13]

There are some special instances where the assignment of cases does not follow the

[13]Baard Marstrand at the Norwegian Courts Administration verified that district courts are required to randomly assign cases to judges, except in a few instances which we discuss in the text. We also checked with both the Bergen District Court (the second largest court, behind Oslo) and the Nedre Telemark District Court (a medium-sized court) that they follow the principle of randomization.

principle of randomization. These include cases involving juvenile offenders, extremely serious cases which require two professional judges, and complex cases expected to take a longer time to process, all of which can be assigned to more experienced judges. The Norwegian Department of Justice provides guidelines on the types of cases that can be non-randomly assigned and the Norwegian Courts Administration has flagged such cases in our dataset. While all other cases are randomly assigned, some case types can only be assigned to regular judges, and deputy judges are assigned relatively more confession cases. This means that randomization occurs within judge type, but not necessarily across judge types. Therefore, to have a sample of *randomly assigned cases* to the *same pool of judges* we: (i) exclude the special cases described above and (ii) focus on regular judges handling non-confession cases.

A key to our design is that not only are judges randomly assigned, but they also differ in terms of their propensity to incarcerate defendants. In our baseline specification, we measure the strictness of a judge based on their incarceration rate for other randomly assigned cases they have handled, including both past and future confession and non-confession cases, and not just those cases which appear in our estimation sample. Our estimation sample has 500 judges, each of whom have presided over an average of 258 randomly assigned court cases. In our baseline specification, our measure of judge stringency is calculated as the leave-out mean judge incarceration rate. When using this measure, we always condition on fully interacted court and year fixed effects to account for the fact that randomization occurs within the pool of available judges. This controls for any differences over time or across judicial districts in the types of criminals or the strictness of judges. In a number of specification checks, we show robustness of the results to how we measure judge strictness (see Section 5.3).

Table 1 verifies that judges in our baseline sample are randomly assigned to cases. The first column regresses incarceration on a variety of variables measured before the court decision. It reveals that demographic, type of crime, and past work and criminal history variables are highly predictive of whether a defendant will be incarcerated, with most being individually significant. In column 3, we examine whether our measure of judge stringency can be predicted by this same set of characteristics. This is the same type of test that would be done to verify random assignment in a randomized controlled trial. There is no statistically significant relationship between the judge stringency variable and the various demographic, crime type and labor market variables. The estimates are all close to zero, with none of them being statistically significant at the 5% level. The variables are not jointly significant either (p-value=.920). This provides strong evidence that criminal court cases are randomly assigned to judges in our sample, conditional on fully interacted court and year fixed effects.

It is natural to ask why some judges are more likely to incarcerate than others. While we do not observe personal characteristics of judges in our data for privacy reasons, we

can measure how many cases they have handled. Using an OLS regression with the same controls as in Table 1, we find no relationship between the number of cases handled and judge stringency in our baseline sample. While there may be a variety of other reasons a judge is more or less likely to incarcerate, it is important to keep in mind that as long as judges are randomly assigned, the underlying reasons should not matter for our analysis.

Table 1. Testing for Random Assignment of Criminal Cases to Judges.

	***Dependent Variables*:**				***Explanatory Variables*:**	
	Pr(Incarcerated)		**Judge Stringency**			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Coefficient Estimate	Standard Error	Coefficient Estimate	Standard Error	Mean	Standard Deviation
Demographics and Type of Crime:						
Age	0.0036***	(0.0004)	-0.0000	(0.0000)	32.65	(11.36)
Female	-0.0520***	(0.0071)	-0.0011	(0.0007)	0.106	(0.308)
Foreign born	0.0035	(0.0062)	0.0007	(0.0007)	0.135	(0.342)
Married, year t-1	-0.0234***	(0.0117)	-0.0017	(0.0012)	0.111	(0.314)
Number of children, year t-1	-0.0011	(0.0032)	0.0002	(0.0004)	0.783	(1.244)
High school degree, year t-1	0.0109	(0.0083)	0.0004	(0.0009)	0.172	(0.377)
Some college, year t-1	-0.0532***	(0.0130)	-0.0013	(0.0015)	0.046	(0.209)
Violent crime	0.0843***	(0.0085)	0.0015	(0.0011)	0.256	(0.437)
Property crime	-0.0357***	(0.0109)	0.0011	(0.0012)	0.139	(0.346)
Economic crime	-0.0401***	(0.0116)	0.0018	(0.0015)	0.113	(0.316)
Drug related	-0.0484***	(0.0112)	-0.0000	(0.0013)	0.119	(0.324)
Drunk driving	0.0745***	(0.0128)	0.0002	(0.0014)	0.071	(0.257)
Other traffic	-0.0453***	(0.0127)	0.0003	(0.0012)	0.087	(0.281)
Missing Xs	-0.2971**	(0.1386)	-0.0088	(0.0150)	0.030	(0.170)
Past Work and Criminal History:						
Employed, year t-1	0.0284***	(0.0082)	0.0002	(0.0008)	0.352	(0.478)
Ever Employed, years t-2 to t-5	-0.0016	(0.0083)	0.0001	(0.0009)	0.470	(0.499)
Charged, year t-1	0.0498***	(0.0074)	0.0003	(0.0008)	0.459	(0.498)
Ever Charged, years t-2 to t-5	0.0447***	(0.0078)	-0.0008	(0.0010)	0.627	(0.483)
Incarcerated, year t-1	0.1423***	(0.0105)	0.0002	(0.0013)	0.139	(0.346)
Ever Incarcerated, years t-2 to t-5	0.1690***	(0.0095)	0.0009	(0.0010)	0.279	(0.448)
F-statistic for joint test	94.99		.593			
[p-value]	[.000]		[.920]			
Number of cases	33,548				33,548	

Note: Baseline sample of non-confession criminal cases processed 2005-2009. All estimations include controls for court x court entry year FEs. Reported F-statistic refers to a joint test of the null hypothesis for all variables. The omitted category for education is "Less than high school, year t-1" and the omitted category for type of crime is "Other crimes". Standard errors are two-way clustered at judge and defendant level. **p<0.1, **p<0.05, ***p<0.01.

2.3 IV Model

We are interested in the causal effects of incarceration on subsequent criminal behavior and labor market outcomes. This can be captured by the regression model

$$Y_{i,t} = \beta_t I_{i,0} + X_i^{'}\theta_t + \eta_{i,t} \tag{1}$$

where β_t is the parameter of interest, $I_{i,0}$ is an indicator variable equal to 1 if defendant i is sentenced to prison in period zero (normalized to be the time of the court decision), X_i is a vector of control variables, and $Y_{i,t}$ is the dependent variable of interest measured at some point t after individual i's court decision (e.g., cumulative criminal charges five years after the court decision). As demonstrated in Table 1, the incarcerated and non-incarcerated groups are far from comparable. This raises concerns of selection bias in OLS estimation of β_t. Our research design addresses this concern by exploiting that cases are randomly assigned to judges (conditional on year and court fixed effects) and that some judges are systematically more lenient that others. Taken together, this leads to random variation in the probability an individual will be incarcerated depending on which judge they are assigned to. We utilize this exogenous variation in $I_{i,0}$ to draw inference about the causal effects of incarceration.

Our main analysis is based on 2SLS estimation of β_t with (1) as the second stage equation and a first stage equation specified as:

$$I_{i,0} = \gamma Z_{j(i)} + X_i^{'}\delta + \nu_{i,0} \tag{2}$$

where the scalar variable $Z_{j(i)}$ denotes the stringency of judge j assigned to defendant i's case. Under the assumptions of instrument exogeneity and monotonicity, the 2SLS estimand can be interpreted as a positive weighted average of the causal effect of incarceration among the subgroup of defendants who could have received a different incarceration decision had their case been assigned to a different judge.

Given the quasi-random assignment of cases to judges, the key challenge to instrument exogeneity is that trial decisions are multidimensional, with the judge deciding on incarceration, fines, community service, probation, and guilt. In Section 5.5, we examine this threat to the exclusion restriction, showing that our estimates do not change appreciably when we augment our baseline model to either control for judge stringency in other dimensions or include and instrument for other trial sentencing decisions. In the presence of heterogenous effects, one may also be worried about the monotonicity assumption, that is, defendants who are incarcerated by a lenient judge would also need to be incarcerated by a stricter judge, and vice versa for non-incarceration. In Section 4.2, we implement two sets of tests, both of which indicate that monotonicity is likely to hold. On top of these challenges to identification, one may also be worried about exactly how to measure judge stringency $Z_{j(i)}$ and perform

statistical inference. For our main specifications, we measure $Z_{j(i)}$ as the leave-out mean incarceration rate which omits case i, that is, the average incarceration rate in other cases a judge has handled. In Section 5.3, we show robustness to alternative measures of $Z_{j(i)}$, including a split sample approach. We also make sure the conclusions do not change materially if we exclude judges with relatively few cases or if we use confidence intervals that remain valid whether or not instruments are weak. In Appendix D, we discuss potential challenges to estimation and inference in the random judge setting and perform a series of Monte Carlo simulations to assess the finite sample performance of the 2SLS estimator depending on how one measures $Z_{j(i)}$. These simulations lend support to the reliability of the statistical inference we perform when measuring $Z_{j(i)}$ as the leave-out mean incarceration rate.

In most of our analysis, we perform 2SLS estimation of equations (1) and (2) using the entire sample of all defendants in non-confession, randomly-assigned cases. However, to interpret the results and inform policy it would be useful to move beyond the resulting average causal effect and estimate the heterogeneous effect of incarceration along a variety of dimensions. One common approach to explore heterogeneity in effects would be to estimate the 2SLS model separately by subgroups. Ideally, we would want to split the sample by case characteristics (e.g. crime type, first-time versus repeated offender), demographics (e.g. age, ethnicity and prior employment status), or both. However, for reasons of sample size and power, we cannot cut the data too finely. Instead, we focus attention on how effects differ by prior employment status, as the question of whether incarceration is criminogenic or preventive is likely to depend strongly on whether a defendant has an actual job to lose and human capital to depreciate by going to prison (see Section 6). In addition to this subsample estimation we explore heterogeneity in effects according to unobservables. To do so, we first estimate the marginal treatment effects (MTE) and then use these estimates to learn about the average treatment effect (ATE), the average treatment effect on the treated (ATT) and the average treatment effect on the untreated (ATUT). The results from the sub-sample estimation and MTE analysis are reported in Section 5.4.

3 Data and Background

3.1 Data and Sample Selection

Our analysis employs several data sources that we can link through unique identifiers for each individual. Information on the court cases comes from the Norwegian Courts Administration. The dataset contains information for all court cases over the period 2005-2014. We observe the start and end dates of every trial, various case characteristics, the verdict, and unique identifiers for both judges, defendants, and district courts. We link this information with

administrative data that contain complete records for all criminal charges, including the type of crime, when it took place, and suspected offenders. This data can be additionally linked to the prison register with information on actual time spent in prison. We merge these data sets with administrative registers provided by Statistics Norway, using a rich longitudinal database that covers every resident from 1967 to 2016. For each year, it contains individual demographic information (including sex, age, and number of children), socioeconomic data (such as years of education, earnings, employment), as well as geographical and firm identifiers.

To construct our baseline sample, we exclude the non-randomly assigned cases described in Section 2.2 and focus on regular judges handling non-confession cases.[14] This yields a sample of randomly assigned cases to the same pool of judges. Excluding the non-randomly assigned cases is straightforward, as these cases are flagged in our dataset. Our baseline sample further restricts the dataset to judges who handle at least 50 randomly assigned confession or non-confession cases between the years 2005 and 2014 (i.e., at least 50 of the cases used to construct our judge stringency instrument). Since we will be including court by year of case registration fixed effects in all our estimates, we also limit the dataset to courts which have at least two regular judges in a given year. Our main estimation sample uses cases decided between 2005 and 2009 so that each defendant can be followed for up to five years after decision, while the judge stringency instrument is based on the entire period from 2005 to 2014. Appendix Table A1 shows how the various restrictions affect the number of cases, defendants, judges and courts in our sample. After applying our restrictions, the baseline estimation sample includes 33,548 cases, 23,373 unique defendants, and 500 judges.

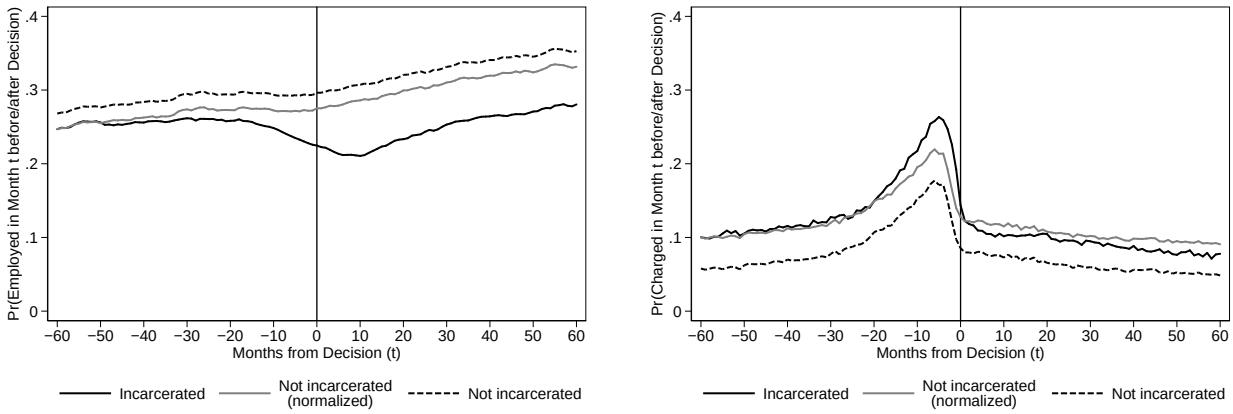
3.2 Descriptive Statistics

We now provide some summary statistics for defendants, crime types and judges. Panel A in Appendix Table A2 shows that defendants are relatively likely to be young, single men. They also have little education, low earnings and high unemployment prior to the charge, with under 40% of defendants working in the prior year. Serial offenders are common, with 38% of defendants having been charged for a different crime in the prior year. Panel B reports the fraction of cases by primary crime category. Around one fourth of cases involve violent crime, while property, economic, and drug crime each comprise a little more than 10 percent of crimes. Drunk driving, other traffic offenses, and miscellaneous crime make up the remainder.

In Figure 2, we document the typical employment and crime levels for our sample over time. Panel (a) plots the probability a defendant has any paid employment in a given month

[14] In comparison, judges are fairly similar in their incarceration rates for confession cases. Replicating the IV specification of column (3) of Table 4, but using confession cases, we estimate an effect of -0.333 (s.e. 0.311). While the magnitude of the coefficient is similar, the standard error is more than three times larger.

during the 10 year period surrounding their court decision. There are separate lines for defendants who are sentenced to incarceration versus not sentenced to incarceration. The first fact which emerges is that prior to the court decision, labor market participation is low for both groups, with less than 30% of defendants working in any month. Employment rates for the incarcerated group are a few percentage points lower; to ease comparison of changes over time, the graph also adjusts the non-incarcerated group's employment line to be the same as the incarcerated group's at the beginning of the sample period. Both groups have monthly employment rates which increase over time, reflecting the fact that employment rises as individuals get older.



(a) Pr(Employed in Month t) **(b)** Pr(Charged in Month t)

Figure 2. Employment and Criminal Charges before and after Month of Court Decision.

Note: Baseline sample consisting of 33,548 non-confession criminal cases processed 2005-2009. Defendants are categorized in two groups, either incarcerated as shown in the solid black line or not incarcerated as shown in the dashed black line. To ease the comparison of trends, in each panel we normalize the level of the not incarcerated group's outcomes to the level of the incarcerated group's outcome in month t=-60. Outcomes for this "normalized" not incarcerated group are shown by the gray solid line. In both panels, the x-axis denotes months since court decision (normalized to period 0).

The most striking pattern in the graph is the divergence in employment between the incarcerated and non-incarcerated defendants around the time of the court decision. The positively sloped pre-trends for both groups are fairly similar up until about one year before the court decision date. However, around 12 months prior to the decision, the incarcerated line trends sharply downwards. This could be the result of incarcerated individuals being more likely to lose their jobs and turn to crime prior to the court's decision, or alternatively, incarcerated individuals being more likely to commit crime and lose their jobs as a result. Either way, the divergent trends prior to treatment suggest the two groups are not comparable. The downward trend continues until about 6 months after the decision, at which point it resumes its upward trend. Comparing the two lines reveals a sizable and stubbornly persistent

drop in employment for the incarcerated group relative to the non-incarcerated.[15] Similar patterns are found for earnings and hours worked (see Appendix Figure B1).

In panel (b) of Figure 2, we plot the probability an individual is charged with at least one crime in a month over time. The figure reveals that both types of defendants have a high propensity to commit a crime. Five years before the court decision, defendants who will be incarcerated have a 10 percent chance of committing a crime in a month, compared to 7 percent for those who will not be incarcerated. Examining the pre-trends, there is a large jump around the court decision for both groups, since in order to have a court decision an individual must first be charged with a crime. While the two groups have similar trends for much of the pre period, they begin to diverge a little more than a year before the court decision, with the incarcerated group exceeding the non-incarcerated group by around 10 percent. That is, the incarcerated defendants get into more trouble with the police in the months leading up to their court decision. After the court decision, the probability of being charged with a crime returns to around 10 percent for both groups.[16]

In addition to describing our data, the graphs presented in Figure 2 highlight the hazards of using OLS or difference-in-differences to estimate the effects of incarceration. The incarcerated and non-incarcerated groups are not comparable in their pre-incarceration levels. Moreover, the trends in employment and criminal activity diverge before the court decision in ways that indicate there is an "Ashenfelter dip" prior to incarceration. These patterns motivate our quasi-experimental approach using the random assignment of judges.[17]

3.3 What Does it Mean to Be Incarcerated in Norway?

To help with interpretation, we briefly describe prison conditions in Norway (see kriminalomsorgen.no). Prisons emphasize rehabilitation and follow the "principle of normality" set forth by the Directorate of Norwegian Correctional Services. The principle dictates that "life inside will resemble life outside as much as possible" and that "offenders shall be placed in the lowest possible security regime." This means that low-level offenders go directly to open prisons,

[15] There are several reasons why employment does not drop to zero after the court decision for those sentenced to prison. First, the average waiting time after a court decision before being sent to prison is around 5 months, and many prison stays are short. Second, the receipt of employment-related payments while in prison, such as vacation pay, shows up as working for pay in our dataset. Third, a small number of individuals are allowed to work outside of prison while incarcerated.

[16] There are two reasons why both types of defendants can be charged with crimes in the months immediately following a court decision. First, we measure when an individual was charged, not when the crime was committed. Second, individuals can commit additional crimes after their court decision before they have been imprisoned (5 month waiting time on average), as well as additional crimes while in prison.

[17] While one could omit the 12 months on either side of treatment in an attempt to avoid the Ashenfelter dip, this would assume the pre-treatment changes are caused by transitory shocks rather than a trend break (see the discussion in Ashenfelter and Card, 1985).

which have minimal security, as well as more freedoms and responsibilities. Physically, these open prisons resemble dormitories rather than rows of cells with bars. More serious offenders who are at risk of violent or disruptive behaviors are sent to closed prisons, which have heightened security. The two types of prisons create a separation between minor and more hardened criminals, at least until the hardened criminals have demonstrated good behavior.[18] While more serious offenders serve the majority of their sentence in closed prisons, they are usually transferred to open prisons for resocialization and further rehabilitation before release. Overall, one third of prison beds are in open prisons and the rest are in closed prisons.

In Norway, there are a total of 61 prisons. The largest prison (in Oslo) has 392 cells, while the smallest has 13. Norway has a strict policy of one prisoner per cell and tries to place prisoners close to home so that they can maintain links with the families. This means that there is often a waiting list for non-violent individuals before they can serve their prison time. Sentenced individuals are released after their trial and receive a letter informing them when a cell opens up; in our data we calculate an average wait time of 5 months.

To help with rehabilitation, all prisons offer education, mental health and training programs. In 2014, 38% and 33% of inmates in open and closed prisons, respectively, participated in some type of educational or training program. The most common programs are for high school and work-related training although inmates can also take miscellaneous courses. All inmates are involved in some type of regular daily activity, unless they have a serious mental or physical disability. If they are not enrolled in an educational or training program, they must work within prison.[19]

All inmates have the right to daily physical exercise and access to a library and newspapers. By law, all prisoners have the same rights to health care services as the rest of the population. The Norwegian Directorate of Health is responsible for managing health programs for inmates. Most notably, 18% of inmates participate in a drug-related program while in prison. After release, there is an emphasis on helping offenders reintegrate into society, with access to programs set up to help ex-convicts find a job and access social services like housing support.[20]

3.4 Comparison to Other Countries

There are both similarities and differences in the criminal population and the criminal justice system of Norway versus the rest of the world. Along most dimensions, Norway looks broadly

[18] This separation could be important, as Bayer et al. (2009) find that inmates build "criminal capital" through interactions with other criminals.

[19] All prisoners, whether working or participating in training or education programs, receive a small stipend while in prison (around $8 per day in 2015). This stipend is not included in any of our earnings measures.

[20] It is important to realize that the initial judge assigned to a case does not determine which prison a defendant is sent to; the type of training, educational, or work program a defendant participates in; or when a defendant is eligible for parole.

similar to many other Western European countries. And while it shares some commonalities with the U.S., the U.S. is an international outlier in some respects.

Incarceration rates. Appendix Figure A1 graphs Norway's incarceration rate over time. In 1980, there were an estimated 44 incarcerated individuals per 100,000 in Norway. This rate has increased gradually over time, with a rate of 72 per 100,000 in 2012. This 64% increase is not merely due to more crime being committed over time, as there has been a more modest 25% increase in crime over the same period (Lappi-Seppälä, 2012). Norway's gradual increase is mirrored in other Western European countries as well, although Norway's rate is slightly lower. In comparison, the U.S. incarceration rate has shot up dramatically, so much so that a separate scale is needed in the figure for the U.S. Not only did the U.S. start at a higher rate of 220 in 1980, but this rate reached over 700 by 2012.[21]

Comparing Norway and the U.S. to a broader set of countries, the U.S. remains an outlier. This can be seen in Appendix Figure A2, which plots incarceration rates versus GDP for 160 countries with a population of greater than half a million. No other country comes close to the U.S. rate of roughly 700 per 100,000, and only the six countries of Rwanda, El Salvador, Turkmenistan, Thailand, Cuba and Russia have over 400 per 100,000. In contrast, the figure shows that Norway's incarceration rate is similar to the average for other Western European countries (102 per 100,000). The U.S. is particularly an outlier after controlling for GDP per capita; relative to other countries with high GDP per capita (purchasing power adjusted), the U.S. incarceration rate is several multiples higher.[22]

Inmate characteristics. Along many dimensions, the prison populations in Norway, Western Europe and the U.S. are similar.[23] Across all these countries, roughly three fourths of inmates have not completed the equivalent of high school. Five percent of prisoners in Norway are female compared to 5% in Western Europe and 7% in the U.S. In all of these countries, inmates are in their early or mid-thirties on average.

The types of offenses committed by inmates differs across countries, but perhaps less than one might expect. In terms of the fraction of prisoners who have committed a drug offense, the rates are surprisingly similar, with 24% in Norway, 22% in Western Europe and

[21] Neal and Rick (2016) show that most of the growth in incarceration rates in the U.S. can be explained by changes in sentencing policy as opposed to higher crime and arrest rates.

[22] It is more difficult to compare measures of criminal activity across countries due to differences in reporting. With this caveat in mind, the U.S. has more than double the number of reported assaults than either Norway or the rest of Western Europe according to the United Nations Survey on Crime Trends (Harrendorf et. al, 2010). Such differences cannot fully explain the large incarceration gap, however, with at least part of the difference being due to longer mandatory sentencing policies for minor crimes (see Raphael and Stoll, 2013).

[23] For details on the U.S. criminal population, see Bureau of Justice Statistics (2015) and Raphael and Stoll (2013). For Scandinavia and other European countries, see Kristoffersen (2014) and Aebi et al. (2015).

20% in the U.S. By comparison, 14% are serving a sentence for assault/battery and 4% for rape/sexual assault in Norway, respectively, compared to 11% and 7% in Western Europe and 9% and 11% in the U.S. Of course, these comparisons need to be understood in the context of a much higher incarceration rate in the U.S. But they point to a considerable overlap in the types of crimes committed by inmates across countries.[24]

Prison expenditures, sentence lengths, and post-release support. One difference across countries is the amount of money spent on prisoners. Western European countries spend an average of $66 thousand per inmate per year, which is roughly double the average of $31 thousand for the U.S. But these averages mask substantial heterogeneity, in part due to differences in labor costs, which in Norway account for two-thirds of the prison budget. For example, in Norway the yearly total cost is $118 thousand (similar to Sweden, Denmark, and the Netherlands), in Italy $61 thousand, and in Portugal $19 thousand. In the U.S., the state of New York spends $60 thousand per prisoner, Iowa $33 thousand, and Alabama $17 thousand. And in New York City, the annual cost per inmate reaches $167 thousand.[25]

Norway is able to maintain the type of prison conditions summarized in Section 3.3 in part due to its larger prison budget. In particular, more resources can be devoted to education and training programs and overcrowding is not an issue. In contrast, while most state prison systems in the U.S. aim to provide GED test preparation, adult basic education and vocational skills training, a recent RAND (2014) report finds that funding for such initiatives is scarce. The U.S. also faces serious overcrowding issues, with Federal prisons being 39% over capacity (GAO, 2012) and over half of states at or above their operational capacity (Bureau of Justice Statistics, 2014).

Another difference between Norway (and Western Europe) versus the U.S. is sentence length. The average time spent in prison using our judge stringency instrument is estimated to be 184 days, or 6 months, for our Norwegian sample. Almost 90% of spells are less than 1 year. This is considerably shorter compared to the average prison time of 2.9 years for the U.S. (Pew Center, 2011), and fairly similar to the median of 6.8 months in other Western European countries (Aebi et al., 2015). Because of this disparity in sentence lengths, the average cost per prisoner spell in Norway and Europe is smaller compared to the U.S., even though the cost per prisoner per year is generally higher.

[24]These numbers for Norway differ from our estimation sample for two reasons: we do not have illegal immigrants in our dataset, and our sample is restricted to non-confession cases which are randomly assigned. The numbers for the U.S. are the weighted average of inmates in federal and state prisons.

[25]Cost estimates are calculated by dividing total prison budgets by number of prisoners. The numbers for Western Europe (sans Belgium and Switzerland) are for the year 2013 and are purchasing power parity adjusted (Abei et al, 2015). The data for 40 U.S. states with available data are for 2010 (Vera Institute of Justice, 2012). New York City data are for 2012 (NYC Independent Budget Office, 2013) .

Norway has been a leader in reforming its penal system to help integrate inmates back into society upon release. While offenders in Norway may lose their job when going to prison, they are usually not asked or required to disclose their criminal record on most job applications. Moreover, while gaps will still appear on employment resumes, these will often span months rather than years due to shorter prison spells. Upon release all inmates have access to support from the Norwegian work and welfare services. This includes work training programs and help searching for a job, as well as access to a variety of social support programs such as unemployment benefits, disability insurance and social assistance.

4 Assessing the Instrument

4.1 Instrument Relevance

Figure 3 shows the identifying variation in our data, providing a graphical representation of the first stage. In the background of this figure is a histogram that shows the distribution of our instrument (controlling for fully interacted year and court dummies). Our instrument is the average judge incarceration rate in other cases a judge has handled, including the judge's past and future cases that may fall outside of our estimation sample. The mean of the instrument is 0.45 with a standard deviation of 0.08. The histogram reveals a wide spread in a judge's tendency to incarcerate. For example, a judge at the 90th percentile incarcerates about 54% of cases as compared to approximately 37% for a judge at the 10th percentile.

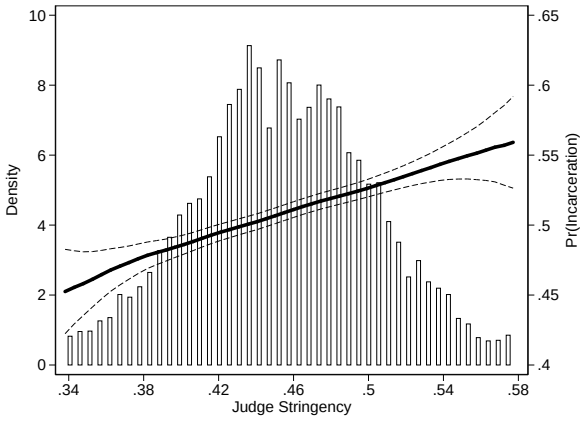


Figure 3. First Stage Graph of Incarceration on Judge Stringency.

Note: Baseline sample consisting of 33,548 non-confession criminal cases processed 2005-2009. Probability of incarceration is plotted on the right y-axis against leave-out mean judge stringency of the assigned judge shown along the x-axis. The plotted values are mean-standardized residuals from regressions on court x court entry year interacted fixed effects and all variables listed in Table 1. The solid line shows a local linear regression of incarceration on judge stringency. Dashed lines show 90% confidence intervals. The histogram shows the density of judge stringency along the left y-axis (top and bottom 2% excluded).

Figure 3 also plots the probability a defendant is sent to prison in the current case as a

function of whether he is assigned to a strict or lenient judge. The graph is a flexible analog to the first stage in equation (2), plotting estimates from a local linear regression. The likelihood of receiving a prison sentence is monotonically increasing in the judge stringency instrument, and is close to linear. Table 2 reports first stage estimates where we regress a dummy for whether a defendant is incarcerated in the current case on our stringency instrument. In panel A, we include fully interacted court and year dummies but otherwise no other controls. The first column reports the first stage estimate at the time of the court decision, whereas the other columns report first stages estimates in each of the five subsequent years. These columns are identical except for the very modest impact of sample attrition (around six percent over five years) stemming from death or emigration of defendants.[26] The point estimate of nearly

Table 2. First Stage Estimates of Incarceration on Judge Stringency.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
***Estimation Sample*:**	Time of Decision	Month 12 after Decision	Month 24 after Decision	Month 36 after Decision	Month 48 after Decision	Month 60 after Decision
***Dependent Variable*:**			**Pr(Incarcerated)**			
A. Court × Year of Court Case Registration Interacted Fixed Effects						
Judge Stringency	0.4897***	0.4922***	0.4887***	0.4818***	0.4795***	0.4699***
	(0.0665)	(0.0661)	(0.0662)	(0.0659)	(0.0661)	(0.0669)
F-stat. (Instrument)	53.56	54.67	53.69	52.79	51.89	48.61
B. Add Controls for Demographics and Type of Crime						
Judge Stringency	0.4793***	0.4811***	0.4755***	0.4694***	0.4680***	0.4587***
	(0.0666)	(0.0662)	(0.0662)	(0.0659)	(0.0661)	(0.0670)
F-stat. (Instrument)	51.11	52.07	50.82	50.09	49.41	46.20
C. Add Controls for Demographics, Type of Crime, Past Work and Criminal History						
Judge Stringency	0.4705***	0.4723***	0.4667***	0.4622***	0.4606***	0.4525***
	(0.0632)	(0.0627)	(0.0624)	(0.0622)	(0.0627)	(0.0634)
F-stat. (Instrument)	54.67	55.95	55.09	54.38	53.18	50.24
Dependent mean	0.5083	0.5077	0.5066	0.5055	0.5047	0.5045
Number of cases	33,548	33,275	32,786	32,341	31,870	31,428

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Standard errors are two-way clustered at judge and defendant level. **p<0.1, **p<0.05, ***p<0.01.

[26] Another test for selective attrition is to regress the probability of attriting on the judge stringency instrument. Performing this test, we find no evidence of a significant relationship (see Appendix Table B1).

0.5 barely moves across columns, indicating that attrition exerts a negligible impact on the first stage relationship. The estimates are highly significant, suggesting that being assigned to a judge with a 10 percentage point higher overall incarceration rate increases the probability of receiving a prison sentence by roughly 5 percentage points.[27]

4.2 Instrument Validity

Conditional Independence. For our instrument to be valid, the stringency of a judge must be uncorrelated with both defendant and case characteristics that could affect a defendant's future outcomes (controlling for fully interacted court and year dummies). As discussed in Section 2.2, Table 1 provides strong empirical support for the claim that the criminal justice system in Norway randomly assigns cases to judges within each court in a given time period.

As a second test, panels B and C of Table 2 explore what happens if a large set of control variables are added to the first stage regressions. If judges are randomly assigned, pre-determined variables should not significantly change the estimates, as they should be uncorrelated with the instrument. As expected, the coefficient does not change appreciably when demographic and crime type controls are added in panel B. As shown in panel C, this coefficient stability continues to hold when we additionally condition on lagged dependent variables capturing a defendant's prior work and criminal history.

Exclusion. Conditional random assignment of cases to judges is sufficient for a causal interpretation of the reduced form impact of being assigned to a stricter judge. However, interpreting the IV estimates as measuring the causal effect of incarceration requires an exclusion restriction: the incarceration rate of the judge should affect the defendant's outcomes only through the incarceration sentencing channel, and not directly in any other way. The key challenge here is that trial decisions are multidimensional, with the judge deciding on incarceration, fines, community service, probation, and guilt. After discussing our main results, we will present empirical evidence that the exclusion restriction holds (see Section 5.5). In particular, we will show that our estimates do not change appreciably when we augment our baseline model to either control for judge stringency in other dimensions or include an instrument for other trial sentencing decisions.

[27]Note that the number of instruments is determined by the number of moment conditions (and not the number of values the instrument takes). Even though there are many judges, our 2SLS model has one moment condition, and therefore, a single instrument. Note also that the first stage coefficient need not be one, unless the following conditions hold: (i) the sample of cases used to calculate the stringency measure is exactly the same as estimation sample, (ii) there are no covariates, and (iii) there are a large number of cases per judge. In our setting, there is no reason to expect a coefficient of one. In particular, the full set of court times year dummies breaks this mechanical relationship. In Section 5.3 we perform specification checks for the instrument, including a split-sample approach.

Monotonicity. If the causal effect of incarceration is constant across defendants, then the instrument only needs to satisfy the conditional independence and exclusion assumptions. With heterogeneous effects, however, monotonicity must also be assumed. In our setting, the monotonicity assumption requires that defendants incarcerated by a lenient judge would also be incarcerated by a stricter judge, and vice versa for non-incarceration. This assumption ensures the 2SLS estimand can be given a local average treatment effect interpretation, i.e. it is an average causal effect among the subgroup of defendants who could have received a different incarceration decision had their case been assigned to a different judge.

One testable implication of the monotonicity assumption is that the first stage estimates should be non-negative for any subsample. For this test, we continue to construct the judge stringency variable using the full sample of available cases, but estimate the first stage on the specified subsample. Results are reported in column (1) of Appendix Table B2. In panel A, we construct a composite index of all of the characteristics found in Table 1, namely predicted probability of incarceration, using the coefficients from an OLS regression of the probability of incarceration on these variables (while conditioning on fully interacted court and year dummies). We then estimate separate first stage estimates for the four quartiles of predicted incarceration. Panel B breaks the data into six crime types. Panels C and D split the data by previous labor market attachment and by whether the defendant has previously been incarcerated, respectively. Panels E, F and G split the samples by age, education, and number of children. For all of these subsamples, the first stage estimates are large, positive and statistically different from zero, consistent with the monotonicity assumption.

A second implication of monotonicity is that judges should be stricter for a specific case type (e.g., violent crimes) if they are stricter in other case types (e.g., all crimes except for violent crimes). To test this implication, we break the data into the same subsamples as we did for the first test, but redefine the instrument for each subsample to be the judge's incarceration rate for cases outside of the subsample. For example, for the violent crime subsample, we use a judge's incarceration rate constructed from all cases except violent crime cases. Column (2) of Appendix Table B2 lists the first stage estimates using this "reverse-sample instrument" which excludes own-type cases. The first stage estimates are all positive and statistically different from zero, suggesting that judges who are stricter for one type of case are also stricter for other case types.

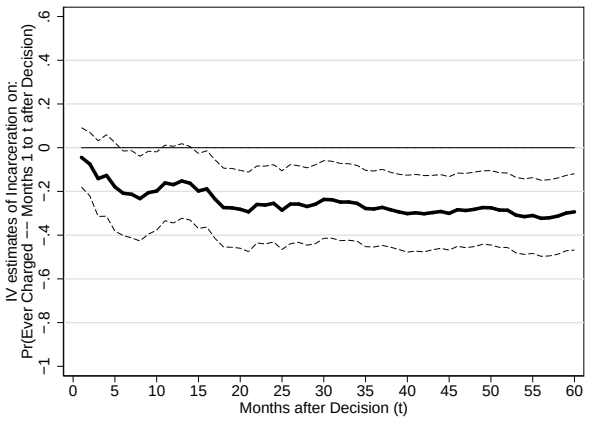
5 Effects of Incarceration on Recidivism

In this section, we present our main findings, showing that (i) incarceration causes a large reduction in the probability of reoffending, (ii) the drop is not due only to incapacitation,

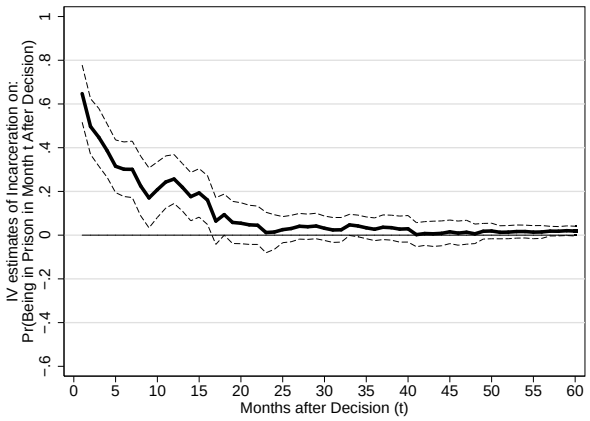
with further reductions in criminal charges after release, and (iii) the total number of charged crimes falls over time, with many individuals being diverted from a future life of crime. We then contrast these IV estimates to OLS, learning that the high rates of recidivism among ex-convicts is due to selection, and not a consequence of the experience of being in prison. Several robustness and heterogeneity checks follow.

5.1 Main Results

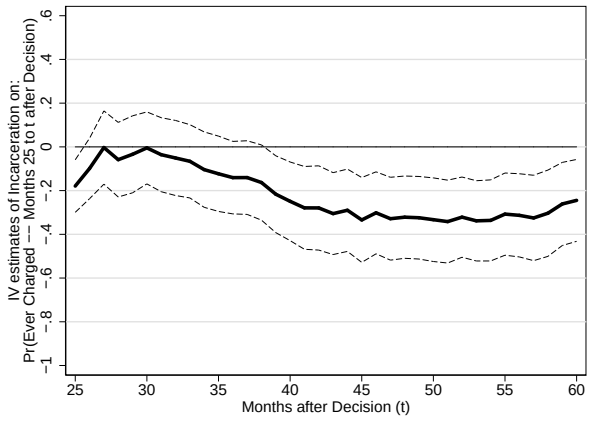
Reoffense probabilities. Panel (a) of Figure 4 graphically presents IV estimates of the effect of incarceration on the probability of reoffending. We define reoffending as the probability of being charged with at least one crime by the end of a given time period.



(a) IV Estimates: Pr(Ever Charged – Months 1 to t)



(b) IV Estimates: Pr(Being in Prison – Month t)



(c) IV Estimates: Pr(Ever Charged – Months 25 to t)

Figure 4. The Effect of Incarceration on Recidivism and Probability of Being in Prison.

Note: Baseline sample of non-confession criminal cases processed 2005-2009 (N=33,548 at time of decision and N=31,428 in month 60 after decision). Panel (b) plots prison probabilities related only to the original sentence. Dashed lines show 90% confidence intervals.

The graph presents a series of cumulative monthly estimates from 1 month to 60 months after the court decision. For example, the estimate at month 6 uses the probability an individual has been charged with at least one crime by 6 months after the decision as the dependent variable in the second stage of the IV model. As expected, there is little effect on reoffending in the first few months after the court decision, since not much time has elapsed for the committing of new crimes. But the estimate becomes more negative over time, and at around 18 months there is a large and statistically significant reduction of over 25 percentage points in recidivism for those previously sentenced to incarceration. This negative effect persists at roughly the same level all the way to 60 months.

Incapacitation versus post-release effects. The recidivism effect found in panel (a) of Figure 4 could simply be due to incapacitation, as individuals sentenced to prison time will be locked up and therefore have few criminal opportunities.[28] To better understand the role of incapacitation, Table 3 presents IV estimates of the effects of incarceration on prison time.

Table 3. The Effect of Incarceration on Prison Time.

	(1)	**(2)**	**(3)**
***Dependent Variable*:**	**Days of Prison Sentence (Potential Prison Time)**	**Days Spent Outside Prison Before Serving Sentence (Waiting Time)**	**Days of Prison Sentence Served (Actual Prison Time)**
RF: Judge Stringency	104.57**	67.89***	83.19***
	(49.03)	(19.47)	(25.15)
2SLS: Incarcerated	231.088**	150.02***	183.83***
	(91.72)	(38.12)	(49.78)
Dependent mean	153.75	69.92	69.20
Number of cases		31,428	

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1. Days spent in pre-trial custody are included in actual prison time in column (3). Standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

We find that, on average, being incarcerated leads to a sentence of 231 days in prison. But this is sentencing time (i.e., potential prison time), not actual time served. Using the IV model, we estimate that being incarcerated leads to 184 days, or approximately 6 months, in actual prison time served. This smaller number makes sense, as Norway allows individuals to be released on parole after serving about two-thirds of their prison sentence for good behavior.[29] In column (2) of the Table 3, we also estimate the average wait time between the

[28]Individuals may be charged with a crime while serving prison time, as they can commit crimes while in prison. They may also have other cases working their way through the system while in prison.

[29]The IV estimate suggests that a majority of individuals receive parole in our dataset. If an inmate

court decision and when individuals start serving their prison sentence. The average wait time is estimated to be around 5 months.

In panel (b) of Figure 4, we plot a series of IV estimates for the probability of being in prison, 1 to 60 months after the court decision. The figure is similar to a survival function, in that if all treated individuals (i.e., those sentenced to prison) started out in prison in month 1, the estimates would map out 1 minus the probability of exit from prison. It is not exactly a survival function though, because not all individuals sentenced to prison begin serving their sentences immediately due to waiting times for an open space. As expected, the probability of being in prison if an individual is sentenced to prison starts out high. This probability falls rapidly, with fewer than 30 percent of incarcerated individuals being in prison for the original criminal charge six months after the court decision. By month 18, only around 5 percent of these individuals are still in prison, and by month 24 almost none are still in prison.

The main point to take away from panel (b) of Figure 4 is that any incapacitation effect from being incarcerated at time zero can only operate in the first two years. Using this insight, we now graph the probability of ever being charged with a crime between months 25 and 60 in panel (c) of Figure 4. By ignoring crimes committed within the first two years after the decision, we are estimating incarceration effects which cannot be attributed to the original incapacitation spell. As in panel (a) of Figure 4, it takes a few months for individuals to start being charged with a crime in this window. But by 15 months after the start of this new window (i.e., 39 months after the court decision), there is a strong and statistically significant reduction in crimes for individuals previously sentenced to prison. The effect is a sizable 25 percentage point reduction in reoffending at least once between months 25 and 60.

In Appendix Table B3, we provide further granularity by running year-by-year models for crimes committed in a particular year. The table documents a negative recidivism effect in the first year (when most individuals are in prison for the current case), the second year (after the majority are already released from prison), and in each of years three through five (when virtually all are out of prison). The individual point estimates are somewhat noisy, but the estimates all go in the same direction. In Table 4 we group the first two years together and years 3 to 5 together for increased precision. That table reveals sizable reductions in recidivism, both in years 1 and 2, as well as in years 3 to 5, consistent with a reduction in crime which is separate from an incapacitation effect.

In theory, it is possible that future pre-trial detentions or prison spells could induce an incapacitation effect even after the original prison spell is completed. This could happen if a prior prison sentence flags an individual as higher risk so that in a future case they are either remanded to custody while awaiting trial or have an increased chance of being sent to prison.

commits a new offense while on parole, this counts as a new charge in our dataset.

To explore this possibility, in Appendix Table B4 we examine whether judge stringency in the current case affects time spent in prison for new charges unrelated to the current case. We first estimate how an incarceration sentence in the current case affects the probability of being sent to prison in the future, either due to pre-trial detention or a new incarceration sentence. We find only a small, insignificant effect (a 1 percentage point increase relative to the mean of 42 percent). This small impact likely reflects two opposing forces. Incarceration reduces the likelihood of recidivism, thus lowering the chances of being charged with a future crime. At the same time, we find evidence suggesting the probability of being incarcerated in the future *conditional on* an individual being charged with a future crime is higher, consistent with the notion that judges are tougher on repeat offenders (see Appendix Table B4).

Table 4. The Effects of Incarceration on Recidivism.

***Dependent Variable*:**	**Pr(Ever Charged)**			**Number of Charges**
	Months 1-24 after Decision	*Months 25-60 after Decision*	*Months 1-60 after Decision*	*Months 1-60 after Decision*
	(1)	**(2)**	**(3)**	**(4)**
OLS: Incarcerated	0.130***	0.115***	0.113***	5.275***
No controls	(0.007)	(0.007)	(0.006)	(0.321)
OLS: Incarcerated	0.126***	0.109***	0.105***	5.369***
Demographics & Type of Crime	(0.007)	(0.007)	(0.006)	(0.310)
OLS: Incarcerated	0.068***	0.050***	0.052***	2.917***
All controls	(0.006)	(0.007)	(0.006)	(0.278)
OLS: Incarcerated	0.057***	0.042***	0.049***	1.595***
Complier Re-weighted	(0.007)	(0.007)	(0.006)	(0.251)
RF: Judge Stringency	-0.108**	-0.111**	-0.133***	-5.196**
All controls	(0.047)	(0.048)	(0.045)	(2.452)
IV: Incarcerated	-0.239**	-0.245**	-0.293***	-11.482**
All controls	(0.113)	(0.113)	(0.106)	(5.705)
Dependent mean	0.57	0.57	0.70	10.21
Complier mean if not incarcerated	0.56	0.57	0.73	13.62
Number of cases		31,428		

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1. RF and IV in addition also control for court x court entry year FEs. OLS standard errors are clustered at the defendant level, while RF and IV standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

The small effect on future incarceration helps interpret the mechanisms behind our main estimates. In particular, they suggest that incapacitation effects due to future prison spells do not explain the large and persistent reduction in recidivism. For example, estimates for

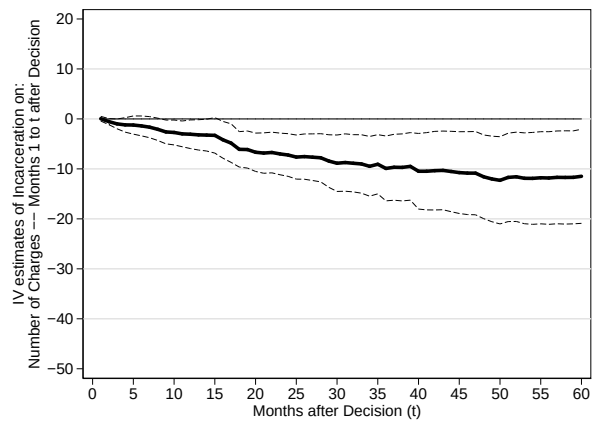
the cumulative number of days spent in prison for new cases is just 4.5 days. This increase is small compared to the direct increase of 184 days of prison time served reported in Table 3.

Number of crimes. A comparison of panels (a) and (c) in Figure 4 suggests that incarceration not only prevents an individual from ever committing a crime (the extensive margin), but it also prevents individuals from committing a series of future crimes (the intensive margin). In panel (a), after month 18, the probability of ever being charged with a crime is flat, suggesting that additional individuals are not being prevented from committing a crime after that time. But in panel (c), we see that the probability an individual will commit a crime between 25 and 60 months is affected by an incarceration decision at time zero. This means that many of the individuals who were prevented from committing a crime in panel (a) are also being prevented from committing another crime in panel (c).

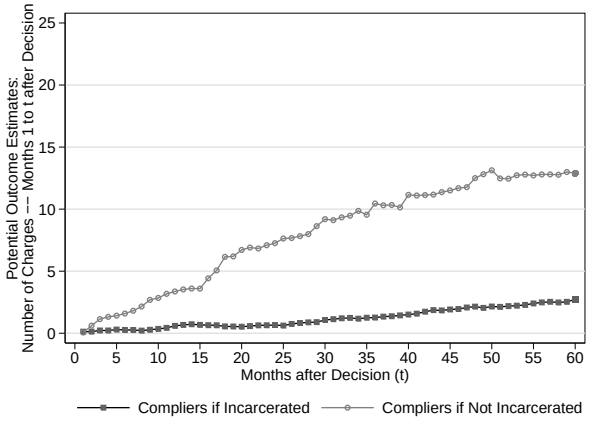
To further explore the intensive margin, panel (a) of Figure 5 plots IV estimates for the cumulative number of charges in the months after the court decision. The estimated effects become more negative over time. After one year, the estimated effect of an incarceration decision is around 3 fewer crimes per individual, whereas after two years, the effect is 7 fewer crimes. By four years, the effect is 11 fewer crimes per individual (see also Table 4).

Potential crimes. Our IV estimates represent the average causal effects for compliers who could have received a different court decision had their case been assigned to a different judge. To better understand this LATE, we follow Imbens and Rubin (1997) and Dahl et al. (2014) in decomposing the IV estimates into the average potential outcomes if the compliers would have been incarcerated and if they would not have been incarcerated. The top line in panel (b) of Figure 5 is the number of potential charges if the compliers would not have been incarcerated. The line trends upward in close to a linear fashion, with approximately 2 to 3 extra criminal charges per year and around 13 crimes on average after five years. In sharp contrast, the compliers would have been charged with far fewer crimes if incarcerated; even by month 60, they would only have been charged with 2 crimes on average.

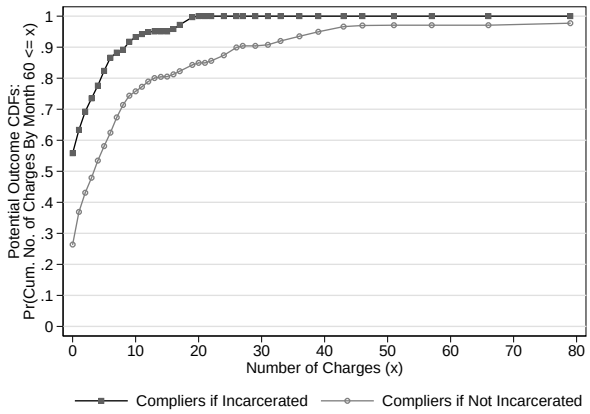
Panel (c) plots the distribution functions for cumulative potential charges as of year 5, for compliers if they would have been incarcerated and if they would not have been incarcerated. The difference between the two CDF's when the number of charges is one is around 30 percentage points, which mirrors the IV effect graphed in Figure 4, panel (a), at 5 years out. Comparing the CDF's further to the right (i.e., for a larger number of charges) makes clear that incarceration is not simply preventing low-crime individuals from committing future crime. To see this, suppose that incarceration caused individuals who would have been charged with 5 crimes or less (or some similarly small number of crimes) from being



(a) IV Estimates: No. of Charges – Months 1 to t



(b) Potential Outcomes: No. of Charges in Months 1 to t



(c) Potential Outcome CDFs: No. of Charges by Month 60

Figure 5. The Effect of Incarceration on Number of Charges.

Note: Baseline sample of non-confession criminal cases processed 2005-2009 (N=33,548 at time of decision and N=31,428 in month 60 after decision). Dashed lines show 90% confidence intervals.

charged with any crimes, but that more hardened criminals (those charged with more than 5 crimes) were unaffected. In this case, the two lines in panel (c) would lie on top of each other starting at 5 charges. But, in fact, the two lines diverge at one charge, remain fairly parallel until around 18 charges, and do not get close to each other until around 45 charges. For instance, 12% of compliers would have been charged with more than 18 crimes if they were not incarcerated, whereas few, if any, compliers would have been charged with this many crimes if incarcerated. Taken together, the results suggest that incarceration must be preventing some individuals from being charged with a large number of crimes, and stopping some individuals from a life of crime entirely.[30]

[30]From the graph, one cannot infer whether an individual charged with 35 crimes reduces their charges to 0 versus whether an individual charged with 35 crimes reduces their crime to 15 while the individual charged

5.2 Comparison to OLS

With few exceptions, the bulk of the research on recidivism is based on OLS regressions with controls for observable confounding factors. In Table 4, we present OLS estimates of equation (1) with and without a rich set of controls. The first OLS specification in Table 4 regresses whether an individual has reoffended (i.e., been charged with a new crime after the court decision) on whether the defendant was sentenced to prison, but includes no other control variables. The OLS estimates 0-2 years after the decision, 2-5 years after the decision, and 0-5 years after the decision are all positive and significant; for example, individuals sent to prison are 11 percentage points more likely to reoffend at least once over the next 5 years.

In the next specification of Table 4, we add a host of defendant characteristics, including demographic variables and the type of crime they are being charged with. These controls affect the estimates only slightly. In the third specification, we additionally add lagged variables for whether defendants have previously been charged with a crime, whether they have previously been incarcerated for a previous crime, and whether they have worked in the prior year (i.e., including all of the variables listed in Table 2 as controls). This brings the coefficient down to 5 percentage points.

The divergence between the OLS estimates and the IV estimates in Table 4 is stark. The OLS estimates always remain positive, while the IV estimates are negative and large. One possible explanation is that the OLS estimates suffer from selection bias due to correlated unobservables. If this is the case, we can conclude that the high rates of recidivism among ex-convicts is due to selection, and not a consequence of the experience of being in prison.

Another possible explanation for the differences between the IV and OLS estimates is effect heterogeneity, so that the average causal effects for the compliers differ in sign compared to the mean impacts for the entire population. To explore this possibility, it is useful to characterize compliers by their observable characteristics. We begin by splitting our sample into eight mutually exclusive and collectively exhaustive subgroups based on prior labor market attachment and the predicted probability of incarceration (see Appendix Table B5). The predicted probability of incarceration is a composite index of all of the observable characteristics while prior employment is key source of heterogeneity in effects, as discussed in the next section. Next, we estimate the first stage equation (2) separately for each subsample, allowing us to calculate the proportion of compliers by subgroup. We then reweight the estimation sample so that the proportion of compliers in a given subgroup matches the share of the estimation sample for that subgroup. The fourth row of Table 4 presents OLS estimates based on this reweighted sample. The results suggest the differences between the IV and OLS

with 15 reduces their crime to 0. But the shapes of the CDF's do imply that high volume criminals must reduce their number of charged crimes.

estimates cannot be accounted for by heterogeneity in effects, at least due to observables.

5.3 Specification Checks

Before exploring the results further, we present specification checks related to the construction of the instrument and the procedure for inference (see also Appendix D). The first column of Appendix Table B6 presents our baseline results for comparison. In this specification, we include any defendant whose judge handled at least 50 cases. In the next three specifications, we instead require judges to handle at least 25 cases, at least 75 cases or at least 100 cases, respectively. These changes do not materially affect the estimated effects. This is reassuring as one might be worried the statistical inference becomes unreliable if the number of cases per judge is too small.

The next two specification checks examine sensitivity to changing how the instrument is constructed. In column (5), we randomly split our sample in half and use one half of the sample to calculate the average incarceration rate of each judge. We next use these measures of judge leniency as an instrument for incarceration in the other half of the sample. The resulting estimates (and standard errors) do not materially change. The last column shows that our findings are not sensitive to whether we calculate judge stringency based on non-confession cases only or if we include all randomly assigned cases (both confession and non-confession cases) in these calculations.

As a final robustness check, panel D in Appendix Table B6 reports Anderson-Rubin (AR) confidence intervals. The confidence intervals remain valid whether or not the instrument is weak, in the sense that their probability of incorrectly rejecting the null hypothesis and covering the true parameter value, respectively, remains well-controlled. Since IV estimates are non-normally distributed when an instrument is weak, the AR procedure does not rely on point estimates and standard errors but instead uses test inversion.[31] We find that the confidence intervals do not materially change. Consistent with this finding, we can strongly reject the null hypothesis of weak instrument using the test proposed by Montiel Olea and Pflueger (2013).

5.4 Heterogeneous Effects

First time offenders. We now examine whether there are heterogeneous effects in the recidivism result. We first limit the sample to first time offenders, defined as defendants who have not previously served time in prison and for whom this is the first court case observed in our sample. Appendix Table C2 reports results analogous to Table 4 for this subsample.

[31] For details on the procedure, see the review article by Andrews et al. (2019).

The 5 year cumulative estimates in column (3) are somewhat larger for first time offenders, with the probability of recidivism dropping by 43 percentage points. Interestingly, the effect is concentrated in the months 25-60 after their court decision, with less evidence for a drop in crime during the period which includes their imprisonment (months 1-24).

Looking at first time offenders is useful not only for exploring heterogeneous effects, but also for ease of interpretation. In our baseline sample, individuals can appear more than once in our dataset if they are brought to trial for multiple crimes over time. Individuals appearing multiple times could be in the incarcerated group in one year and the non-incarcerated group in another year. While judges are randomly assigned for each case, and hence the baseline estimate is still causal, the interpretation is more nuanced. With first time offenders, each individual appears only once in the sample. The cost of looking only at an individual's first criminal case is that the sample drops in half, from over 30,000 observations to less than 15,000. Given the results are qualitatively similar, but with less precision for the smaller sample, we focus on results using the more comprehensive dataset which contains all cases with random assignment. A more complete set of results for first time offenders, which mirror those found for the full sample in what follows, can be found in Appendix Tables C1-C3.

Open versus closed prisons. A second type of heterogeneity is the type of prison an individual is sent to. As a reminder, there are two types of prisons in Norway: open and closed. As described in Section 3.3, open prisons have minimal security, as well as more freedoms and responsibilities compared to closed prisons. The two types of prisons not only result in different day-to-day activities, but also create a separation between minor and more hardened criminals. Whether a convicted defendant is initially sent to an open or closed prison depends both on the severity of the crime, as well as geographical proximity and available space at open versus closed prisons. Judges do not directly determine whether individuals are sent to open versus closed prisons. Moreover, when we run a multinomial regression with three outcomes (incarcerated in open prison, incarcerated in closed prison, not incarcerated), we find that a judge's stringency does not differentially affect whether an individual is sent to an open versus a closed prison.[32]

To explore whether the types of individuals sent to open versus closed prisons experience different outcomes, we first predict whether an individual will be sent to an open versus closed prison based on the pre-determined characteristics in Table 2. We then create dummy variables for whether an individual's probability of being sent to an open prison is above or below the median. Finally, we interact these dummy variables with our judge stringency

[32]In a multinomial logit regression which includes the same controls as in panel C of Table 2, judge stringency has an average marginal effect of 0.24 (s.e. 0.04) for open prison versus 0.22 (s.e. .04) for closed prison, with not incarcerated being the omitted category.

measure, and create two analogous instruments. In Appendix Table B7 we re-estimate our main IV specification, but with two separate endogenous variables and instruments based on the interactions. We find remarkably similar effects of incarceration on recidivism for those individuals with below and above median probabilities of being sent to an open versus a closed prison (see column (4)). However, it is important not to overinterpret these results, since the two groups could experience heterogeneous effects from incarceration if prison type were held fixed.

Marginal treatment effects. Finally, we explore heterogeneity by examining marginal treatment effects (MTEs). Ignoring subscripts for simplicity, we model the observed outcome as $Y = I \times Y(1) + (1 - I) \times Y(0)$, where I is an indicator for treatment (being incarcerated) and $Y(1)$ and $Y(0)$ are the associated potential outcomes which are a linear function of both observable (X) and unobservable factors. The choice of treatment by a judge is given by $I = 1[\upsilon(X, Z) - U]$, where υ is an unknown function, U is an unobserved continuous random variable, and Z is our judge stringency instrument.[33] One can normalize the distribution of $U|X = x$ to be uniformly distributed over [0,1] for every value of X. Under this normalization, it is straightforward to show that $\upsilon(X, Z)$ is equal to the propensity score $p(X, Z) \equiv P[D = 1|X = x, Z = z]$.

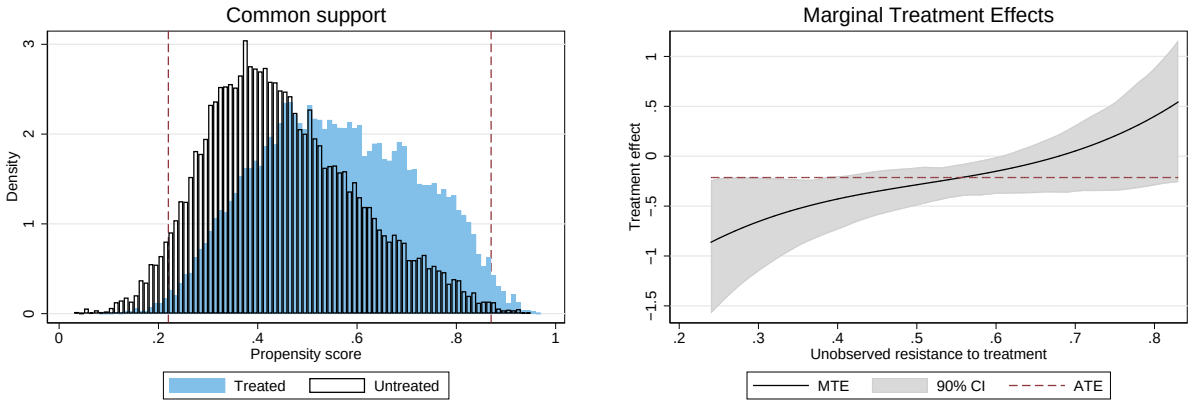
The MTE is defined as $E[Y(1) - Y(0)|U = u, X = x]$. The dependence of the MTE on U for a fixed X reflects unobserved heterogeneity in treatment effects, as indexed by a judge's latent propensity to choose incarceration for a defendant (where U captures unobserved characteristics of the defendant which influence the judge's choice). The choice equation implies that, given X, defendants with lower values of U are more likely to take treatment regardless of their realization of Z. Following Brinch et al. (2017), we assume separability between observed and unobserved heterogeneity in the treatment effects. Together with the assumption of an exogenous instrument that satisfies monotonicity, this restriction on the potential outcomes is sufficient to allow point identification of MTE over the unconditional support of the propensity score $p(X, Z)$.[34] We probe the stability of MTE estimates to various specifications of the empirical MTE model. Reassuringly, the estimates based on a linear, quadratic, cubic, or quartic specification all yield similar estimates, as does a semiparametric specification based on local linear regressions.

Panel (a) in Figure 6 graphs the propensity score distributions for the treated and

[33]The weakly separable choice equation is equivalent to assuming monotonicity (Imbens and Angrist, 1994).

[34]Separability between observed and unobserved heterogeneity in the treatment effect is weaker than additive separability between I and X, which is a standard auxiliary assumption in applied work using IV. Furthermore, it is implied by (but does not imply) full independence $(Z, X \perp Y(1), Y(0), U)$, a common assumption in applied work estimating MTEs. See Mogstad and Torgovitsky (2018).

untreated samples. The dashed red lines indicate the upper and the lower points of the propensity score with common support (after trimming 1% of the sample with overlap in the distributions of propensity scores). Panel (b) of Figure 6 plots MTE estimates by the unobserved resistance to treatment (i.e., the latent variable U) based on a local instrumental variables approach using a global cubic polynomial specification. The MTE estimates are most negative for those with a low unobserved resistance to treatment, and rise as unobserved resistance to treatment increases. This implies that incarceration reduces recidivism the most for defendants whose unobservables would make them very likely to go to prison regardless of the stringency of their judge. In contrast, defendants whose unobservables would make them very unlikely to go to prison experience, if anything, an increase in recidivism due to treatment, with the caveat that the estimates are noisy.



(a) Common Support

(b) Marginal Treatment Effect Estimates: Pr(Ever Charged – Months 1 to 60)

Figure 6. Common Support and Marginal Treatment Effects.

Note: Baseline sample of non-confession criminal cases processed 2005-2009 (N=31,428 in month 60 after decision). The dashed red lines in figure (a) indicate the upper and the lower points of the propensity score with common support (based on 1% trimming). The MTE estimates plotted in figure (b) are based on a local instrumental variables (IV) approach using a global cubic polynomial specification for the trimmed sample with common support (N=28,275). Standard errors are constructed based on 100 bootstrap replications.

As shown by Heckman and Vytlacil (1999, 2005, 2007), all conventional treatment parameters can be expressed as different weighted averages of the MTE. Recovering these treatment parameters for the entire population, however, requires full support of the propensity score $p(X, Z)$ on the unit interval. Since we do not have full support, we follow Carneiro et al. (2011) in rescaling the weights so that they integrate to one over the region of common support. Appendix Table B8 uses the MTE estimates to construct such rescaled estimates of the average treatment effect on the treated (ATT), the average treatment effect (ATE), and the average treatment effect on the untreated (ATUT). These weighted averages are obtained

by integrating the MTE over the propensity score for the relevant sample. The ATT estimates reveal the recidivism effects of imprisonment are especially large for the treated; for example, the linear specification yields an estimate of -0.42, which is more negative than either the LATE or the ATE. By comparison, the estimated ATE (also plotted as the horizontal line in Figure 6) is similar to the LATE. The ATUT, in contrast, is closer to zero and not statistically significant.

5.5 Threats to Exclusion Restriction

As discussed in Section 4.2, interpreting the IV estimates as average causal effects of incarceration requires the judge stringency instrument to affect the defendant's outcomes only through the prison sentencing channel. A potential issue is that trial decisions are multidimensional, with judges deciding on incarceration, fines, community service, probation and guilt (where the penalties are not mutually exclusive).

To make this issue precise, it is useful to extend the baseline IV model given by equations (1) and (2), distinguishing between the incarceration decision and other trial decisions:

$$I_{i,0}^{Incar} = \alpha Z_{j(i)}^{Incar} + \gamma Z_{j(i)}^{Other} + X_i^{'}\delta + v_{i,0} \tag{3}$$

$$I_{i,0}^{Other} = \zeta Z_{j(i)}^{Incar} + \lambda Z_{j(i)}^{Other} + X_i^{'}\psi + u_{i,0} \tag{4}$$

$$Y_{i,t} = \beta_t I_{i,0}^{Other} + \theta_t I_{i,0}^{Incar} + X_i^{'}\omega_t + \eta_{i,t} \tag{5}$$

where j denotes the judge that handles defendant i's case, $I_{i,0}^{Incar}$ is an indicator variable equal to 1 if defendant i is sentenced to prison in period zero, $I_{i,0}^{Other}$ is an indicator variable equal to 1 if defendant i is sentenced to fines, community service or probation, $Z_{j(i)}^{Incar}$ denotes the judge stringency instrument for the incarceration decision, $Z_{j(i)}^{Other}$ denotes the judge stringency instrument for trial decisions other than incarceration, and X_i is a vector of control variables that includes a full set of case year by court dummy variables. The omitted reference category is not guilty. As in the baseline model, we measure $Z_{j(i)}^{Incar}$and $Z_{j(i)}^{Other}$ as leave-out means.

There are two cases in which the baseline IV estimates based on (1)-(2) are biased because they abstract from trial decisions other than incarceration. The first case is if $Z_{j(i)}^{Incar}$correlates with $Z_{j(i)}^{Other}$, and $Z_{j(i)}^{Other}$ directly affects $Y_{i,t}$ (conditional on X_i). This would violate the exclusion restriction in the baseline IV model because $Z_{j(i)}^{Incar}$ not only affects $Y_{i,t}$ through $I_{i,0}^{Incar}$ but also through its correlation with $Z_{j(i)}^{Other}$. However, controlling for $Z_{j(i)}^{Other}$ in both (1) and (2) will eliminate this source of bias. The second case is if $Z_{j(i)}^{Incar}$ correlates with $I_{i,0}^{Other}$ conditional on $Z_{j(i)}^{Other}$, and $I_{i,0}^{Other}$ affects $Y_{i,t}$ holding $I_{i,0}^{Incar}$ fixed (conditional on X_i). In the baseline IV model, this would violate the exclusion restriction because $Z_{j(i)}^{Incar}$ not only

affects $Y_{i,t}$ through $I_{i,0}^{Incar}$ but also through to its influence on $I_{i,0}^{Other}$. The augmented IV model given by (3)-(5) addresses this issue by including $I_{i,0}^{Other}$ as an additional endogenous regressor and $Z_{j(i)}^{Other}$ as an extra instrument.[35]

In Appendix Tables B9 and B10, we examine these two cases, finding support for the exclusion restriction. To start, we first calculate a judge's tendencies on trial decisions other than incarceration.[36] For example, we measure a judge's probation stringency as the average probation rate in the other cases a judge has handled. The top panel of Table B9 repeats our baseline specification for comparison. In panel B, we add a judge's probation stringency, community service stringency, and fine stringency as three additional controls in both the first and second stages. A decision of not guilty is the omitted category. The IV estimates for both recidivism outcomes are similar to our baseline, albeit with standard errors which are larger. To increase precision, panel C combines these three control variables into a single "probation, community service or fine" stringency variable. Again, the IV estimates for recidivism are similar to the baseline in panel A, but the standard errors are considerably larger.

We next estimate the augmented IV model given by (3)-(5). Appendix Table B10 presents the first stage, reduced form and IV estimates. To make sure we have enough precision and avoid problems associated with weak instruments, we use a specification with three decision margins: "incarceration," "probation, community service or fine," and "not guilty." For the incarceration first stage, the judge stringency instrument for the incarceration decision has a similar coefficient as before. For the other first stage, the judge stringency instrument for the incarceration decision matters little if anything, but the other instrument is strongly significant. To formally evaluate the overall strength of the instruments we report the Sanderson-Windmeijer F-statistics, indicating that weak instruments are not an issue. Looking at the reduced form estimates, the coefficients on the judge stringency instrument for the incarceration decision are virtually unchanged compared to the baseline IV model. In contrast, we find that the judge stringency instrument for the other decisions has almost no effect on recidivism in the reduced form. Likewise, the IV estimates for incarceration in the final columns of Appendix Table B10 are similar to those from the baseline IV model which does not include and instrument for the other decision margins.

A useful byproduct of examining the threats to exclusion from trial decisions other than incarceration is that it helps with interpretation. The baseline IV model compares the

[35]Note that a causal interpretation of the IV estimates based on (3)-(5) requires assumptions in addition to the instrument exogeneity and monotonicity conditions discussed in Section 4.2. As shown by Kirkeboen et al. (2016), one may either assume the effects of each treatment are the same across individuals or invoke additional restrictions on individuals' choice behavior.

[36]While not a trial decision per se, judges could also differ in how quickly they process cases. Creating a second instrument based on a judge's average processing time in other cases they have handled, and redoing the empirical tests reported below with processing time as an additional covariate yields similar conclusions.

potential outcomes if incarcerated to the outcomes that would have been realized if not incarcerated. The augmented IV model helps to clarify what is meant by not incarcerated, distinguishing between not guilty as opposed to alternative sentences to imprisonment. The IV estimates in Appendix Table B10 suggest significant effects of being sentenced to prison compared to being found not guilty, whereas the probation, community service or fine category does not have a statistically different effect compared to not guilty.

5.6 Sentence Length

It is possible that strict judges (as measured by our judge stringency instrument) are both more likely to incarcerate defendants and to give them longer sentences. If this is the case, our baseline estimates capture a linear combination of the extensive margin effect of being incarcerated and the intensive margin of longer sentences. However, most of the sentences observed in the data are short, so there is limited variation along the intensive dimension. As shown in Appendix Figure B2, the median sentence length is 3 months in our sample, with roughly 80% of sentences being less than one year. Empirically, there is little difference in sentence lengths across judges holding incarceration rates fixed. This is consistent with judges having discretion on the incarceration decision, but using mandatory rules or guidelines for sentence lengths.

Keeping these caveats in mind, we now explore various models which use sentence length. To provide context, panel (a) of Appendix Figure B3 graphs sentence length in days (including zeros) as a function of our judge incarceration stringency instrument. The upward slope largely reflects the fact that stricter judges send more defendants to prison. Panel (b) illustrates how sentence length is affected by our instrument. It plots estimates of the probability a sentence length will exceed a given number of days (including zeros) as a function of the judge stringency instrument, and reveals that most of the action is for relatively short sentences.[37] As shown in Table 3, using our judge stringency instrument results in an increase of roughly 7.5 months spent in prison, which helps in interpreting our main estimates in Table 4.

A complementary analysis is to replace the endogenous variable of incarceration with sentence length, but still use our judge incarceration stringency variable as the instrument. As shown by Imbens and Angrist (1994), 2SLS applied to an IV model with variable treatment intensity (such as days in prison) captures a weighted average of causal responses to a unit change in treatment, for those whose treatment status is affected by the instrument. The weight attached to the jth unit of treatment is proportional to the number of people who, because of the instrument, change their treatment from less than j to j or more. In our setting,

[37]To calculate these estimates, we use a similar specification as equation 2, but replace the dependent variable for incarceration with indicators for a sentence length exceeding a given number of days.

this means that coding the endogenous regressor as days in prison (instead of incarceration) permits identification of a weighted average of the effect of another day in prison. Thus, this parameter captures a convex combination of the extensive margin effect of going to prison and the intensive margin effects of longer sentencing. When estimating this model with days in prison as the endogenous regressor, the results are consistent with those using the binary incarceration measure. The effect of increasing sentence length by 250 days (roughly the average sentence length), yields estimates which are similar in size to our estimates based on the binary endogenous variable of incarceration, but with standard errors which are 75% larger (see Appendix Table B11).

Finally, we consider models which include both incarceration and sentence length simultaneously. Our first exploration is what happens if we control for a judge's sentence length stringency, defined as the average sentence length in the other cases a judge has handled. In Table B9, Panel D, when we add in controls for sentence length stringency, it has little effect on our IV estimates. When we try to go a step further, treating both incarceration and sentence length as endogenous variables and using two instruments, the standard errors for IV blow up due to the multicolinearity of incarceration stringency and sentence length stringency (see Appendix Table B12).[38] This means we cannot separately identify the intensive and extensive margin effects. But it is worth noting that the reduced form regression which includes both instruments finds a similar estimate for incarceration stringency compared to baseline, and no significant effect for sentence length stringency.

6 Employment and Recidivism

This section explores factors that may explain the preventive effect of incarceration, showing that the decline in crime is driven by individuals who were not working prior to incarceration. Among these individuals, imprisonment increases participation in programs directed at improving employability and reducing recidivism, and ultimately, raises employment and earnings while discouraging further criminal behavior.

6.1 Recidivism as a Function of Prior Employment

To examine heterogeneity in effects by labor market attachment, we assign defendants to two similarly sized groups based on whether they were employed before the crime for which they are in court occurred. We classify people as 'previously employed' if they were working in at

[38]These patterns are similar whether or not we also adds in controls for probation, community service or fine stringency. The first stage graph of sentence length stringency on sentence length can be found in panel (c) of Appendix Figure B3.

least one of the past five years; the other individuals are defined as 'previously non-employed'.[39] We then re-estimate the IV model separately for each subgroup.

Figure 7 presents the IV estimates for the two subsamples of the effect of incarceration on the probability of reoffending. The results show the effect is concentrated among the previously non-employed. The effects of incarceration for this group are large and economically important. In particular, the likelihood of reoffending within 5 years is cut in half due to incarceration, from 96 percent to 50 percent. Examining the results in Appendix Figure B4 reveals that incarceration not only reduces the probability of re-offending among the previously non-employed, but also the number of crimes they commit. Five years out, this group is estimated to commit 22 fewer crimes per individual if incarcerated. By comparison, previously employed individuals experience no significant change in recidivism due to incarceration.

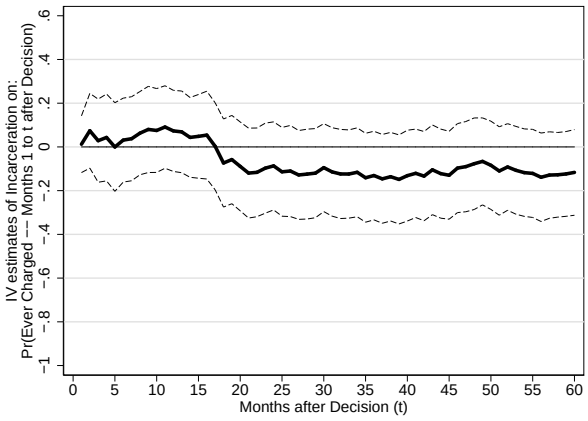
A natural question is whether the heterogeneous effects are due to labor market attachment per se or to variables correlated with prior employment. To explore this, we first compare the characteristics of the previously employed and non-employed subsamples. As seen in Appendix Table A2, the two subsamples differ in characteristics other than prior employment.[40] The non-employed group is about two years younger, less likely to be married, and have lower education. They are more likely to commit property and drug-related crimes instead of economic and traffic-related offenses, including drunk driving. Both groups are charged with an equal number of violent crimes. The non-employed individuals are also 50 percent more likely to have been charged with a crime in the year before their court case.

These comparisons make clear that the previously employed and non-employed have different characteristics. To find out whether these differences can explain the contrasting recidivism effects, we reweight the subsamples so that they are similar based on observables. To do this, we estimate the probability of being in the previously employed group using all of the control variables listed in Table 1 (excluding the variables on past work history). Appendix Figure B5 plots the estimated propensity scores for both the previously employed and non-employed groups. There is substantial overlap for the entire range of employment probabilities. Using these propensity scores, we weight each subsample so that they have the same distribution as the opposite subsample.

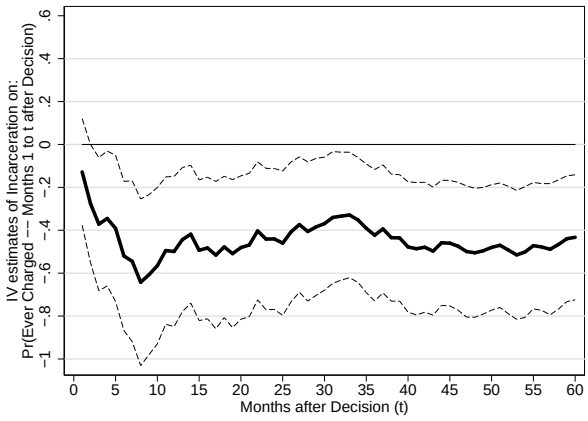
[39] As in Kostøl and Mogstad (2014), an individual is defined as employed in a given year if his annual earnings exceed the yearly substantial gainful activity threshold (used to determine eligibility to government programs like unemployment insurance). In 2010, this amount was approximately NOK 72,900 ($12,500). Our results are not sensitive to exactly how we define employment.

[40] A few of the non-employed will have earned more than the minimum threshold in the year before their court case, even though by definition the five years before their crime they were non-employed. This is because the date of a court case does not line up precisely with the date of a crime.

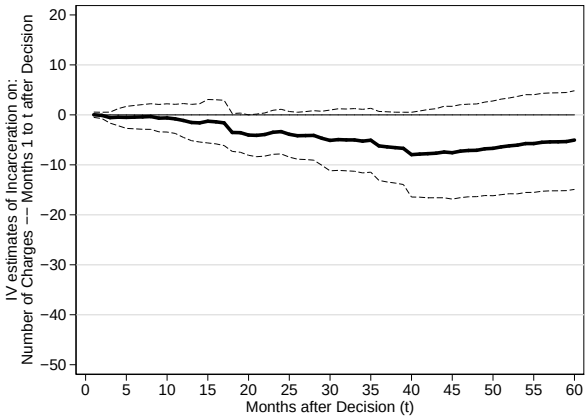
Column A: Previously Employed | **Column B: Previously Non-employed**



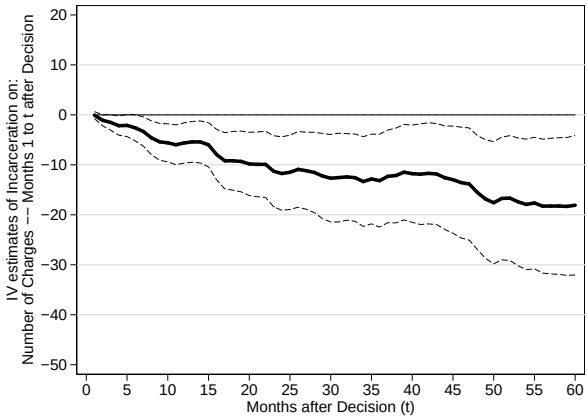
(a) IV Estimates: Pr(Ever Charged – Months 1 to t)



(b) IV Estimates: Pr(Ever Charged – Months 1 to t)



(c) IV Estimates: No. of Charges – Months 1 to t



(d) IV Estimates: No. of Charges – Months 1 to t

Figure 7. The Effect of Incarceration on Recidivism by Previous Labor Market Attachment.

Note: Baseline sample consisting of non-confession criminal cases processed 2005-2009 (N=33,548 at time of decision and N=31,428 in month 60 after decision). Dashed lines show 90% confidence intervals.

Table 5 reports estimates in columns (1) and (3) for the baseline balanced sample five years after the court decision, without any reweighting. Consistent with the figures discussed above, prison time dramatically reduces the extensive and intensive recidivism margins for the previously non-employed defendants. The same is not true for the previously employed, where the effects are much smaller in absolute value, and not statistically significant. The table then reports the weighted results in columns (2) and (4). This has little effect on the estimates, indicating that differences in observable characteristics are not driving the contrasting results. Instead, it appears the differential effects are driven by labor market attachment per se or correlated unobservable characteristics.

Table 5. The Effect of Incarceration on Recidivism by Previous Labor Market Attachment.

	Sub-sample:		*Sub-sample:*	
Dependent Variable:	**Previously Employed**		**Previously Non-employed**	
A. Pr(Ever Charged)	**(1)**	**(2)**	**(3)**	**(4)**
Months 1-60 after Decision	*Baseline*	*Re-weighted*	*Baseline*	*Re-weighted*
RF: Judge Stringency	-0.062	-0.079	-0.183***	-0.157***
All controls	(0.063)	(0.068)	(0.060)	(0.069)
IV: Incarcerated	-0.117	-0.146	-0.433**	-0.365*
All controls	(0.119)	(0.126)	(0.177)	(0.192)
Dependent mean	0.62	0.58	0.79	0.76
Complier mean if not incarcerated	0.55	0.60	0.96	0.86
Dependent Variable:	**Previously Employed**		**Previously Non-employed**	
B. Number of Charges	**(1)**	**(2)**	**(3)**	**(4)**
Months 1-60 after Decision	*Baseline*	*Re-weighted*	*Baseline*	*Re-weighted*
RF: Judge Stringency	-2.686	-2.304	-7.637**	-8.448***
All controls	(3.134)	(2.953)	(3.167)	(3.046)
IV: Incarcerated	-5.042	-4.280	-18.085**	-19.688**
All controls	(5.983)	(5.584)	(8.452)	(8.672)
Dependent mean	7.29	6.10	13.45	11.92
Complier mean if not incarcerated	3.61	5.16	24.01	21.97
Number of cases	16,547		14,881	

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. In columns (2) and (4), we use propensity score re-weighting to adjust for differences in observable characteristics across sub-samples; see discussion of the re-weighting procedure in Section 6.1. *p<0.1, **p<0.05, ***p<0.01.

6.2 The Effect of Incarceration on Future Employment

Why are the reductions in recidivism concentrated among the group of defendants with no prior employment? To shed light on this question, we turn to an examination of the labor market consequences of incarceration depending on prior employment.

Using the previously non-employed sample, panel (b) of Figure 8 plots the IV estimates for the probability of being ever employed by a given time period. Two years after the court decision, previously non-employed defendants experience a 30 percentage point increase in employment if incarcerated. This employment boost grows further to a nearly 40 percentage point increase within 5 years. Panel (b) of Appendix Figure B6 decomposes the IV estimates into the potential employment rates of the compliers. This decomposition reveals that only 12 % of the previously non-employed compliers would have been employed if not incarcerated. By comparison, these compliers would experience a steady increase in employment if incarcerated, with over 50% being employed by month 60. In column A of Figure 8, a different story emerges for the previously employed. They experience an immediate 25 percentage point

drop in employment due to incarceration and this effect continues out to 5 years.

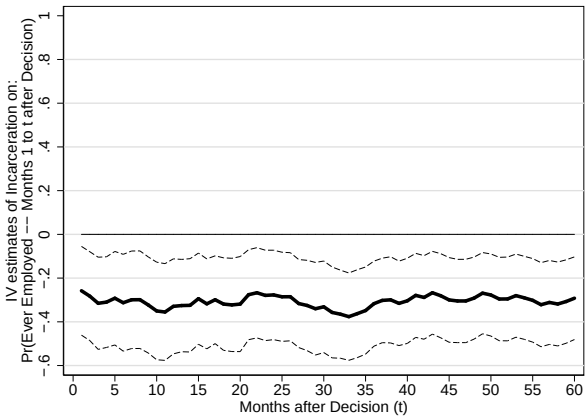
We complement the employment results by examining the effects of incarceration on cumulative hours of work and earnings. Starting with the previously non-employed defendants, in panel (d) of Figure 8 we see a steady increase in the number of hours worked due to incarceration. The IV estimate increases modestly for the first two years, and then starts to increase at a faster rate. By month 60, incarceration increases labor supply by 2,700 hours per individual, translating into more than 550 additional hours per year on average. The decomposition by potential outcomes in panels (d) and (f) of Appendix Figure B6 helps explain what is happening. If not incarcerated, few of the previously non-employed compliers would have gotten a job. As a result of incarceration, they get a job and continue to accumulate hours over time.

Looking at previously employed individuals in column A of Figure 8, we see a different pattern. Incarceration has a negative effect on hours worked, consistent with the drop in employment observed for this group. Interestingly, the potential employment rate of the previously employed compliers is fairly similar to that of the previously non-employed compliers if they are incarcerated (compare panels (c) and (d) in Appendix Figure B6). This suggests that incarceration can take an individual who previously had almost no attachment to the labor market, and make them look like someone who also served prison time, but was previously employed.

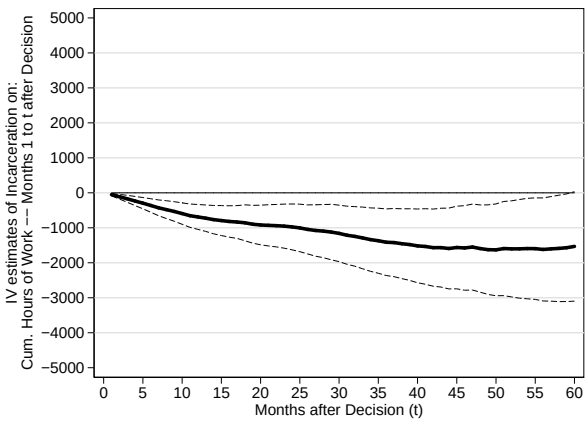
Panels (e) and (f) of Figure 8 repeats the same exercise, but this time for cumulative earnings. The general patterns found for employment and hours of work are mirrored in these figures, as are the decompositions based on potential outcomes in Appendix Figure B7. Lastly, Table 6 shows that differences in observable characteristics other than prior employment are not driving the contrasting labor market effects for previously employed versus previously non-employed defendants.

One question is whether the positive cumulative effects found for the previously non-employed reflects initial boosts in employment and hours associated with participation in a re-entry program after release from prison, or whether the employment effects are more long lasting. To assess this, in the top panel of Appendix Table B13 we estimate year-by-year effects for employment outcomes in a given year. We find sizable and lasting effects of incarceration on future employment for the previously non-employed. The table documents statistically significant increases in hours of work in years 2, 3, 4, and 5. If anything, the effect grows larger with each passing year, with estimates increasing from 115 more hours of work in year 1 to 734 more hours in year 5. Similar results are found for both ever employed and earnings outcomes, although the ever employed estimates are more imprecise. This suggests that we are not simply estimating the cumulative impact of a short term effect, but

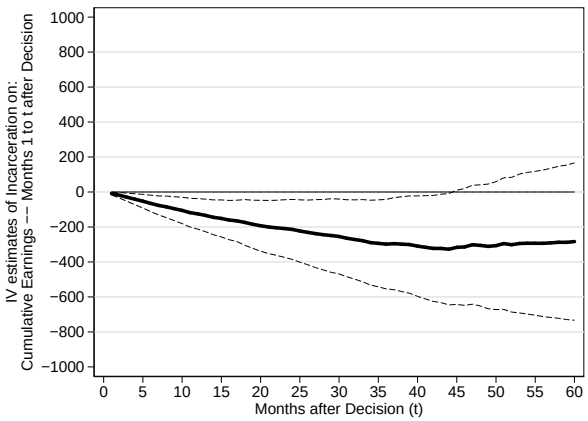
Column A: Previously Employed



(a) IV Estimates: Pr(Ever Employed – Months 1 to t)

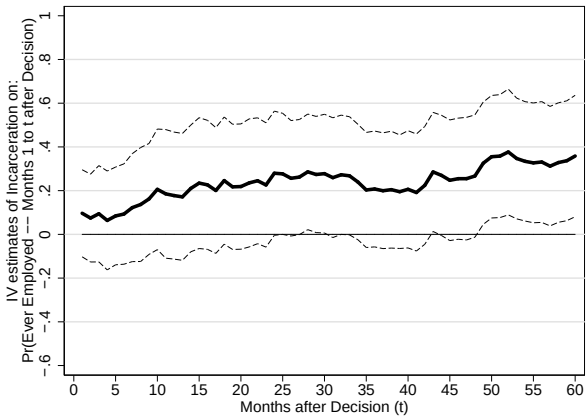


(c) IV Estimates: Hours of Work – Months 1 to t

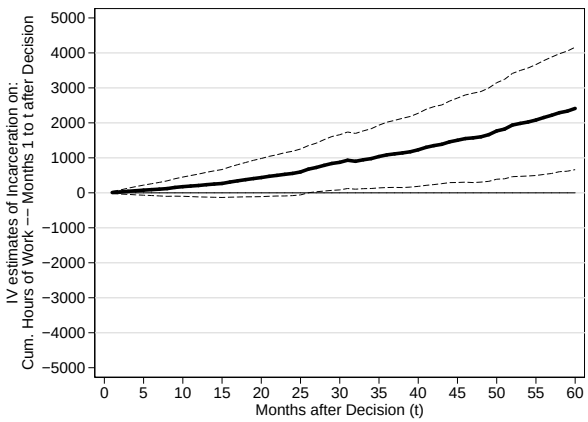


(e) IV Estimates: Cumulative Earnings – Month 1 to t

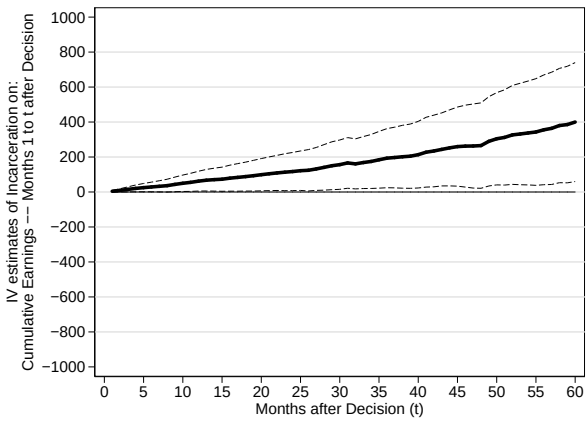
Column B: Previously Non-employed



(b) IV Estimates: Pr(Ever Employed – Months 1 to t)



(d) IV Estimates: Hours of Work – Months 1 to t



(f) IV Estimates: Cumulative Earnings – Month 1 to t

Figure 8. The Effect of Incarceration on Employment by Previous Labor Market Attachment.

Note: Baseline sample consisting of non-confession criminal cases processed 2005-2009 (N=33,548 at time of decision and N=31,428 in month 60 after decision). Dashed lines show 90% confidence intervals.

Table 6. The Effect of Incarceration on Future Employment by Previous Labor Market Attachment.

	Sub-sample:		*Sub-sample:*	
Dependent Variable:	**Previously Employed**		**Previously Non-employed**	
A. Pr(Ever Employed)	**(1)**	**(2)**	**(3)**	**(4)**
Months 1-60 after Decision	*Baseline*	*Re-weighted*	*Baseline*	*Re-weighted*
RF: Judge Stringency	-0.155***	-0.176***	0.151**	0.210***
All controls	(0.058)	(0.063)	(0.064)	(0.070)
IV: Incarcerated	-0.292**	-0.327**	0.358**	0.490**
All controls	(0.115)	(0.127)	(0.168)	(0.199)
Dependent mean	0.70	0.72	0.43	0.44
Complier mean if not incarcerated	0.82	0.75	0.13	0.13
Dependent Variable:	**Previously Employed**		**Previously Non-employed**	
B. Cumulative Hours of work	**(1)**	**(2)**	**(3)**	**(4)**
Months 1-60 after Decision	*Baseline*	*Re-weighted*	*Baseline*	*Re-weighted*
RF: Judge Stringency	-815.9	-1075.8*	1019.1***	1385.9***
All controls	(507.3)	(554.1)	(364.9)	(415.6)
IV: Incarcerated	-1531.7	-1998.2*	2413.1**	3230.0**
All controls	(948.1)	(1048.1)	(1060.7)	(1289.9)
Dependent mean	3804.2	4063.7	1514.3	1613.8
Complier mean if not incarcerated	4410.7	3838.3	51.4	47.8
Dependent Variable:	**Previously Employed**		**Previously Non-employed**	
C. Cumulative Earnings	**(1)**	**(2)**	**(3)**	**(4)**
Months 1-60 after Decision	*Baseline*	*Re-weighted*	*Baseline*	*Re-weighted*
RF: Judge Stringency	-151.0	-206.1	168.9**	254.2**
All controls	(146.0)	(172.0)	(73.7)	(92.6)
IV: Incarcerated	-283.5	-382.9	399.9*	592.4**
All controls	(272.6)	(319.9)	(206.1)	(272.7)
Dependent mean	834.3	920.4	255.7	279.0
Complier mean if not incarcerated	914.2	788.3	9.95	9.70
Number of cases	16,547		14,881	

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. In columns (2) and (4), we use propensity score re-weighting to adjust for differences in observable characteristics across sub-samples; see discussion of re-weighting in Section 6.1. Cumulative earnings are reported in 1000 Norwegian Kroner. *p<0.1, **p<0.05, ***p<0.01.

rather a persistent employment effect. These findings mirror what we see when looking at the intensive margins of the cumulative number of hours and earnings over time in panels (d) and (f) of Figure 8.

A similar analysis is done for the previously employed in the bottom panel of Appendix Table B13. The year-by-year estimates for this group are generally negative, with the largest effect in the first year (consistent with them losing their job while in prison) with smaller

effects by year 5 (suggesting they start to recover by the end). The cumulative effects on hours and earnings in panels (c) and (e) of Figure 7 are another way to illustrate these negative effects.

6.3 The Role of Job Loss and Job Training Programs

The differences in labor market effects depending on prior employment are striking. For the previously employed group, the negative effects are perhaps not unexpected, as these individuals had an actual job to lose by going to prison. To test whether job loss is the explanation, we take advantage of the fact that we can link firms to workers in our data. In particular, we follow the previously employed defendants from years 2 to 5 after their court case (after virtually all incarcerated individuals should be out of prison), and track whether their first employment, if any, during this period was with the same firm as they worked in before the court case. We then run two new IV regressions, quantifying the effect of incarceration on i) the probability of being employed at a new firm and ii) the chance of being employed at the previous firm. The first column in Appendix Table B14 shows the overall employment effect in any firm, which is a 29 percentage point drop due to incarceration. As shown in the next two columns of this table, the drop in employment is almost entirely due to a reduction in the likelihood of employment at the previous firm, whereas there is only a small, and statistically insignificant, effect of incarceration on the probability of employment at a new firm.

Individuals who were not working prior to incarceration had no job to lose. However, serving time in prison could give access to educational and job training programs, both while in prison and immediately after. We collected individual-level data on participation in a variety of job training and classroom training programs. The most common job training program is on-the-job training in a regular or sheltered workplace, where the employer receives a temporary subsidy (normally up to one year) to train the individual and expose them to different jobs. Job training is specifically targeted to those who need work experience in order to find employment. It is often paired with job finding assistance, where a personal counselor helps the individual find a suitable workplace and negotiate wages and employment conditions. The classroom training programs include short skill-focused courses, vocational training and ordinary education. Classroom training is limited to 10 months for skill courses, 2 years for vocational training and 3 years for ordinary education. Thirty-three percent of the previously non-employed sample participates in job training and 25% participates in classroom training. In comparison, among the previously employed, 25% participate in job training and 25% in classroom training.

Table 7 reports IV estimates for both types of training using our judge stringency

instrument. We focus on the first two years after the court decision, so as to capture the training while in prison and immediately after. For the previously employed group, there are hints that participation in both job and classroom training programs increases due to incarceration, but nothing which is statistically significant. For the previously non-employed group, there is likewise no statistically significant evidence for an increase in classroom training, although the estimate is positive. Instead, what changes significantly due to incarceration is the probability that previously non-employed defendants participate in job training programs. We estimate that being incarcerated makes these individuals 35 percentage points more likely to attend a job training program. By comparison, few if any of the previously non-employed compliers would have participated in job training programs if not incarcerated.

Table 7. The Effect of Incarceration on Participation in Job Training Programs (JTP) and Classroom Training Programs (CTP).

	Sub-sample: **Previously Employed**		***Sub-sample:*** **Previously Non-employed**	
Dependent Variable:	**(1)** *Pr(Participated in Job Training Programs)*	**(2)** *Pr(Participated in Classrom Training Programs)*	**(3)** *Pr(Participated in Job Training Programs)*	**(4)** *Pr(Participated in Classrom Training Programs)*
	Months 1-24 after Decision		*Months 1-24 after Decision*	
RF: Judge Stringency	0.056	0.073	0.147**	0.054
All controls	(0.063)	(0.065)	(0.063)	(0.067)
IV: Incarcerated	0.106	0.138	0.348**	0.127
All controls	(0.118)	(0.122)	(0.168)	(0.164)
Dependent mean	0.17	0.19	0.22	0.17
Complier mean if not incarcerated	0.16	0.18	0.00	0.04
Number of cases	16,547		14,881	

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Control variables include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

6.4 Putting the Pieces Together

So far, we have demonstrated the decline in crime from incarceration is driven by individuals who were not working prior to incarceration. Among these individuals, imprisonment increases participation in programs directed at improving employability and reducing recidivism, and ultimately, raises employment and earnings while discouraging further criminal behavior. A

natural question is whether the people who, due to incarceration, commit fewer crimes are the same individuals as those who become more likely to participate in job training programs and work more. Or does the decline in crime occur independently of the increase in program participation and employment? We investigate this question in Appendix Table B15. In columns (2) and (3), we first break up the probability of re-offending into the probability of re-offending and employed plus the probability of re-offending and not employed. Using the IV model, we report estimates for how each of these joint probabilities are affected by incarceration. As shown in column (2), there is little change in the joint probability of re-offending and employment due to incarceration. Instead, the entire drop in recidivism appears to be driven by a reduction in the joint probability of re-offending and not employed. The only conclusion consistent with all of our estimates is that individuals who are induced to start working are the same individuals who stop committing crimes.[41]

Going a step further, in columns (4) and (5), we estimate the joint probability of re-offending, employment and job training. We find the entire drop in recidivism reported in column (1) is due to a reduction in the joint probability of being charged, not employed and not participating in a job training program. We therefore conclude that the drop in crime we find for the previously non-employed is driven by the same individuals who, due to incarceration, participate in job training and become gainfully employed.

7 Implications for Cost-Benefit Calculations

A natural question is whether the positive effects from imprisonment found in Norway pass a cost-benefit test. It is difficult to estimate the benefits of crime reduction and the costs of imprisonment, with researchers making strong assumptions and extrapolations to do so (see McCollister et al., 2010 and Garcia et al., 2017). With this caveat in mind, we attempt a simple cost-benefit comparison. Our rough calculations suggest the high rehabilitation expenditures in Norway are more than offset by the corresponding benefits to society.

To calculate the costs of incarceration reported in Appendix Table B16, we first compute the direct daily cost per prisoner of incarceration. To do this, we take the total prison spending reported by the Norwegian Correctional Services divided by the total number of prison days served across all prisoners in 2013. This gives a direct prison cost of $323.50 per

[41] To see this, let C denote crime, E denote employment, and I denote incarceration. By definition, P(C) = P(C $\bigcap$ E) + P(C $\bigcap$ not E). We estimate that dP(C)/dI $<$ 0 is driven by dP(C $\bigcap$ not E)/dI $<$ 0, since dP(C $\bigcap$ E)/dI $\approx$ 0. Notice that dP(C $\bigcap$ not E)/dI $<$ 0 means that some individuals with C=1, E = 0 if I = 0 change behavior if I = 1. There are three possibilities for change: (i) C = 0, E = 0, (ii) C=1, E = 1 and (iii) C = 0, E = 1. However, (i) is inconsistent with dP(E)/dI $>$ 0 and (ii) is inconsistent with dP(C)/dI $<$ 0. Only (iii) is consistent with with dP(E)/dI $>$ 0, dP(C)/dI $<$ 0, and dP(C $\bigcap$ not E)/dI $<$ 0. Note that this is an argument about net effects; while there may be some of types (i) and (ii), they would have to be offset by even more of type (iii).

day. We then create an outcome variable which multiplies the number of days spent in prison for an individual's current court case by $323.50. The IV estimate which uses this outcome measure yields a cost of $60,515 per incarceration sentence. We note this measure captures the average cost of incarceration, even though ideally one would like to use the marginal cost of incarceration.

On the benefit side, there are three broad categories. First, there is a reduction in criminal justice system expenditures due to fewer crimes being committed. Following the approach that McCollister et al. (2010) used for the U.S., we calculate police savings per crime avoided as total operating costs reported by the Norwegian Police Service divided by the total number of reported crimes. Police savings are computed to be $3,670 per reported crime. Likewise, we calculate court savings as total operating costs reported by the Norwegian Courts, scaled by the fraction of criminal cases in the courts, and then divided by the total number of criminal cases processed in 2013. Court savings are computed to be $2,533 per court case. We then create an outcome variable which takes the total number of crimes committed by an individual multiplied by $3,670 plus the total number of future court cases for an individual multiplied by $2,533. Using this combined criminal justice cost as the outcome variable and our IV setup, we estimate a savings of $71,225 per incarceration sentence.

The second category of benefits is due to increased employment, which results in higher taxes paid and lower transfer payments. We estimate the increase in taxes minus transfers to be $67,086 per incarceration sentence using IV, although we note this estimate is noisy. Net transfers include all cash transfers received minus all income taxes paid over the five year period following the court decision. Either of these first two benefit categories would justify the direct costs of prisons. Note that our calculations only cover the 5 years after the court decision; any benefits in the future would further add to the benefits (see Garcia et al., 2017).

The third benefit category is the reduction in victimization costs due to fewer crimes being committed in the future. Victimization costs are notoriously difficult to estimate, so we instead simply note that this category would make the comparison of benefits versus costs even more favorable. Of course, the importance of this category depends on whether the avoided crimes are serious from a welfare perspective. We lack the power to precisely estimate the decrease in the number of crimes for all crime types. With this caveat in mind, we find that roughly 40% of the overall reduction is due to drops in property crime, with 4.3 fewer property crimes (s.e. = 2.1), and roughly 20% is due to fewer traffic violations (estimate = -2.4, s.e. = 1.2). The remaining decrease is spread across other crime types, such as violent crimes, drug crimes, and drunk driving, but these estimates are not statistically significant.

8 Concluding Remarks

A pivotal point for prison policy was the 1974 Martinson report, which concluded that "nothing works" in rehabilitating prisoners. Around this time, incarceration rates started to rise dramatically, especially in the U.S. where they more than tripled, as an increasing emphasis was placed on punishment and incapacitation. In recent years, researchers and policymakers have questioned whether incarceration is necessarily criminogenic or whether it can instead be preventive. Our study serves as a proof-of-concept demonstrating that time spent in prison with a focus on rehabilitation can indeed be preventive. The Norwegian prison system is successful in increasing participation in job training programs, encouraging employment, and discouraging crime, largely due to changes in the behavior of individuals who were not working prior to incarceration.

While this paper establishes an important proof of concept, several important questions remain for future research. Our results do not imply that prison is necessarily cost effective or preventative in all settings. Evidence from other countries and populations would be useful to assess the generalizability of our findings. Moreover, while we provide some evidence that job training and employment are part of the story, it would be interesting to quantify their effects more precisely, as well as to analyze other possible mechanisms such as sentence lengths, prison conditions, drug treatment programs and post-release support. Additional research along these lines will aid policymakers as they tackle the challenging task of prison reform.

References

Aebi, M., M. Tiago, and C. Burkhardt (2015). *Survey on Prison Populations (SPACE I – Prison Populations Survey 2014) Survey 2014*. Council of Europe Annual Penal Statistics.

Aizer, A. and J. J. Doyle (2015). Juvenile Incarceration, Human Capital and Future Crime: Evidence from Randomly-Assigned Judges. *Quarterly Journal of Economics 130*(2), 759–803.

Andrews, I., J. Stock, and L. Sun (2019). Weak Instruments in IV Regression: Theory and Practice. *Annual Review of Economics 11*.

Ashenfelter, O. and D. Card (1985). Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *Review of Economics and Statistics 67*(4), 648–60.

Autor, D., A. R. Kostøl, M. Mogstad, and B. Setzler (forthcoming). Disability Benefits, Consumption Insurance, and Household Labor Supply. *American Economic Review*.

Barbarino, A. and G. Mastrobuoni (2014). The Incapacitation Effect of Incarceration: Evidence from Several Italian Collective Pardons. *American Economic Journal: Economic Policy 6*(1), 1–37.

Bayer, P., R. Hjalmarsson, and D. Pozen (2009). Building Criminal Capital Behind Bars: Peer Effect In Juvenile Corrections. *Quarterly Journal of Economics 124*(1).

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica 80*(6), 2369–2429.

Bernburg, J. G., M. D. Krohn, and C. J. Rivera (2006). Official Labeling, Criminal Embeddedness, and Subsequent Delinquency: a Longitudinal Test of Labeling Theory. *Journal of Research in Crime and Delinquency 43*(1), 67–88.

Bohn, A. (2000). *Domsstolloven, Kommentarutgave [Law of Courts, Annotated Edition]*. Universitetsopplaget, Oslo (in Norwegian).

Brennan, P. A. and S. A. Mednick (1994). Learning Theory Approach to the Deterrence of Criminal Recidivism. *Journal of Abnormal Psychology 103*(3), 430–440.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017). Beyond LATE with a Discrete Instrument. *Journal of Political Economy 125*(4), 985–1039.

Buonanno, P. and S. Raphael (2013). Incarceration and Incapacitation: Evidence from the 2006 Italian Collective Pardon. *American Economic Review 103*(6), 2437–2465.

Bureau of Justice Statistics (2014). *Prisoners in 2013*. U.S. Department of Justice.

Bureau of Justice Statistics (2015). *Prisoners in 2014*. U.S. Department of Justice.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating Marginal Returns to Education. *American Economic Review 101*(6), 2754–81.

Chalfin, A. and J. McCrary (2017). Criminal Deterrence: A Review of the Literature. *Journal of Economic Literature 55*(1), 5–48.

Cook, P. J., S. Kang, A. A. Braga, J. Ludwig, and M. E. OBrien (2015). An Experimental Evaluation of a Comprehensive Employment-oriented Prisoner Re-entry Program. *Journal of Quantitative Criminology 31*(3), 355–382.

Cullen, F. T. (2005). The Twelve People Who Saved Rehabilitation: How the Science of Criminology Made a Difference-The American Society of Criminology 2004 Presidential Address. *Criminology 43*(1), 1–42.

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014). Family Welfare Cultures. *Quarterly Journal of Economics 129*(4), 1711–1752.

Di Tella, R. and E. Schargrodsky (2013). Criminal Recidivism after Prison and Electronic Monitoring. *Journal of Political Economy 121*(1), 28–73.

Dobbie, W., J. Goldin, and C. S. Yang (2018). The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. *American Economic Review 108*(2), 201–40.

Dobbie, W. and J. Song (2015). Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection. *American Economic Review 105*(3), 1272–1311.

Doyle, J. J. (2007). Child Protection and Child Outcomes: Measuring the Effects of Foster Care. *American Economic Review 97*(5), 1583–1610.

Doyle, J. J. (2008). Child Protection and Adult Crime: Using Investigator Assignment to Estimate Causal Effects of Foster Care. *Journal of Political Economy 116*(4), 746–770.

Doyle, J. J., J. A. Graves, J. Gruber, and S. Kleiner (2012). *Do High-Cost Hospitals Deliver better Care? Evidence from Ambulance Referral Patterns.* NBER Working Paper No. 17936.

Freeman, R. B. (1992). *Crime and the Economic Status of Disadvantaged Young Men.* in: George E. Peterson and Wayne Vroman, eds., Urban Labor Markets and Job Opportunities (Urban Institute Press Washington, DC), 112–152.

French, E. and J. Song (2014). The Effect of Disability Insurance Receipt on Labor Supply. *American Economic Journal: Economic Policy 6*(2), 291–337.

GAO-12-743 (2012). *Growing Inmate Crowding Negatively Affects Inmates, Staff, and Infrastructure.* United States Government Accountability Office.

Garcia, J. L., J. J. Heckman, D. E. Leaf, and M. J. Prados (2017). *Quantifying the Life-Cycle Benefits of a Prototypical Early Childhood Program.* IZA DP No. 10811.

Gottfredson, D. M. (1999). *Effects of Judges' Sentencing Decisions on Criminal Careers*. US Department of Justice, Office of Justice Programs, National Institute of Justice.

Green, D. P. and D. Winik (2010). Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism among Drug Offenders. *Criminology 48*(2), 357–387.

Grogger, J. (1995). The Effect of Arrests on the Employment and Earnings of Young Men. *Quarterly Journal of Economics 110*(1), 51–71.

Harrendorf, S., M. Heiskanen, and S. Malby (2010). *International Statistics on Crime and Justice*. European Institute for Crime Prevention and Control, affiliated with the United Nations (HEUNI).

Heckman, J. J. and E. J. Vytlacil (1999). Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences 96*(8), 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation. *Handbook of Econometrics 6*, 4779–4874.

Imbens, G. W. and J. D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and D. B. Rubin (1997). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *Review of Economic Studies 64*(4), 555–574.

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016). Field of Study, Earnings, and Self-Selection. *Quarterly Journal of Economics 131*(3), 1057–1111.

Kling, J. (1999). *The Effect of Prison Sentence Length on the Subsequent Employment and Earnings of Criminal Defendants*. Princeton University, Discussion Papers in Economics, Discussion Paper No. 208.

Kling, J. R. (2006). Incarceration Length, Employment, and Earnings. *American Economic Review 96*(3), 863–876.

Kostøl, A. R. and M. Mogstad (2014). How Financial Incentives Induce Disability Insurance Recipients to Return to Work. *American Economic Review 104*(2), 624–655.

Kristoffersen, R. (2014). *Correctional Statistics of Denmark, Finland, Iceland, Norway and Sweden 2009 - 2013*. Correctional Service of Norway Staff Academy.

Kuziemko, I. (2013). How Should Inmates be Released from Prison? An Assessment of Parole versus Fixed-sentence Regimes. *Quarterly Journal of Economics 128*(1), 371–424.

Lappi-Seppälä, T. (2012). Penal Policies in the Nordic Countries 1960–2010. *Journal of Scandinavian Studies in Criminology and Crime Prevention 13*(1), 85–111.

Lipton, D., R. Martinson, and J. Wilks (1975). *The Effectiveness of Correctional Treatment: A Survey of Treatment Evaluation Studies*. New York Office of Crime Control Planning.

Loeffler, C. E. (2013). Does Imprisonment Alter the Life Course? Evidence on Crime and Employment from a Natural Experiment. *Criminology 51*(1), 137–166.

Maestas, N., K. J. Mullen, and A. Strand (2013). Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt. *American Economic Review 103*(5), 1797–1829.

Martinson, R. (1974). What Works? - Questions and Answers About Prison Reform. *The Public Interest 35*, 22–54.

Mastrobuoni, G. and D. Terlizzese (2014). *Rehabilitation and Recidivism: Evidence from an Open Prison*. Working Paper.

McCollister, K. E., M. T. French, and H. Fang (2010). The Cost of Crime to Society: New Crime-Specific Estimates for Policy and Program Evaluation. *Drug and Alcohol Dependence 108*(1-2), 98–109.

Mogstad, M. and A. Torgovitsky (2018). Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics* (0).

Montiel Olea, J. L. and C. Pflueger (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics 31*(3), 358–369.

Mueller-Smith, M. (2015). *The Criminal and Labor Market Impacts of Incarceration*. University of Michigan Working Paper.

Nagin, D. S., F. T. Cullen, and C. L. Jonson (2009). Imprisonment and Reoffending. *Crime and Justice 38*(1), 115–200.

Neal, D. and A. Rick (2016). The Prison Boom and Sentencing Policy. *Journal of Legal Studies 45*(1), 1–41.

NOU (2002). *Dømmes av Likemenn [Judged by Peers]*. Ministry of Justice and Public Security, Norway (In Norwegian).

NYC Independent Budget Office (2013). *NYC Independent Budget Office Annual Data*. NYC Independent Budget Office (IBO).

Owens, E. G. (2009). More Time Less Crime? Estimating the Incapacitative Effect of Sentence Enhancements. *Journal of Law and Economics 52*(3), 551–579.

Pew Center (2011). *State of Recidivism. The Revolving Door of America's Prisons*. The Pew Center on the States, Washington, DC.

RAND (2014). *How Effective Is Correctional Education, and Where Do We Go from Here?* Rand Corporation.

Raphael, S. and M. A. Stoll (2013). *Why Are So Many Americans in Prison?* Russell Sage Foundation.

Redcross, C., M. Millenky, T. Rudd, and V. Levshin (2012). More than a Job: Final Results from the Evaluation of the Center for Employment Opportunities (CEO) Transitional Jobs Program. *OPRE Report 2011-18*.

Skardhamar, T. and K. Telle (2012). Post-Release Employment and Recidivism in Norway. *Journal of Quantitative Criminology 28*(4), 629–649.

Stevenson, M. T. (2018). Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes. *Journal of Law, Economics, and Organization 34*(4), 511–542.

Vera Institute of Justice (2012). *The Price of Prisons: What Incarceration Costs Taxpayers*. Technical Report, Center on Sentencing and Corrections.

Visher, C. A., L. Winterfield, and M. B. Coggeshall (2005). Ex-offender Employment Programs and Recidivism: A Meta-analysis. *Journal of Experimental Criminology 1*(3), 295–316.

Waldfogel, J. (1994). The Effect of Criminal Conviction on Income and the Trust "Reposed in the Workmen". *Journal of Human Resources 29*(1), 62–81.

Western, B. and K. Beckett (1999). How Unregulated is the US Labor Market? The Penal System as a Labor Market Institution. *American Journal of Sociology 104*(4), 1030–60.

Western, B., J. R. Kling, and D. F. Weiman (2001). The Labor Market Consequences of Incarceration. *Crime and Delinquency 47*(3), 410–427.

World Prison Brief (2016). *World Prison Population List (11th edition)*. Institute for Criminal Policy Research (Author: Roy Walmsley).

Appendix Figures and Tables

Appendix A. Additional Descriptives

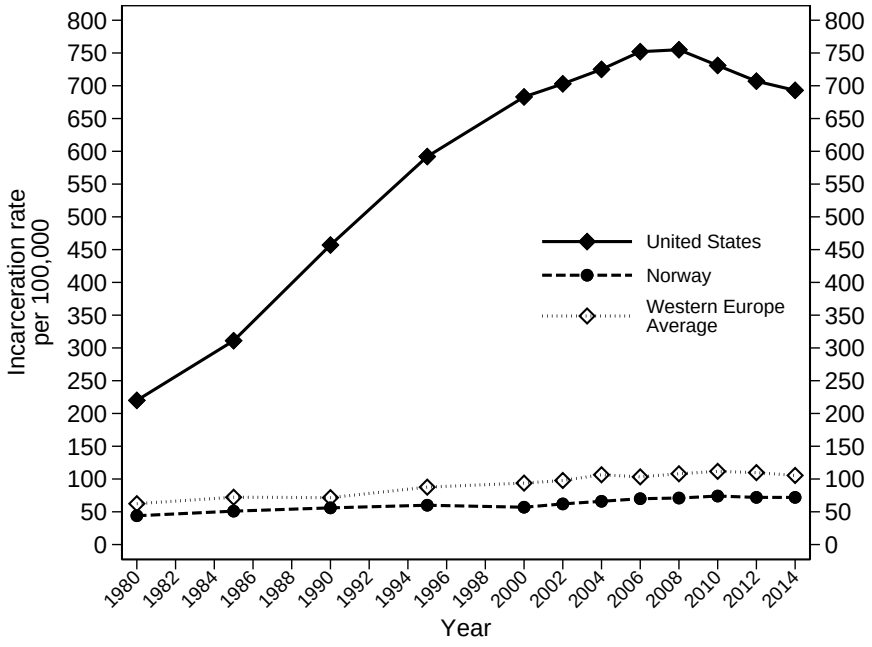


Figure A1. Incarceration Trends in Norway, Western Europe and the U.S.

Note: The Western European countries used to construct the population-weighted average include Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the UK. Source: Institute for Criminal Policy Research.

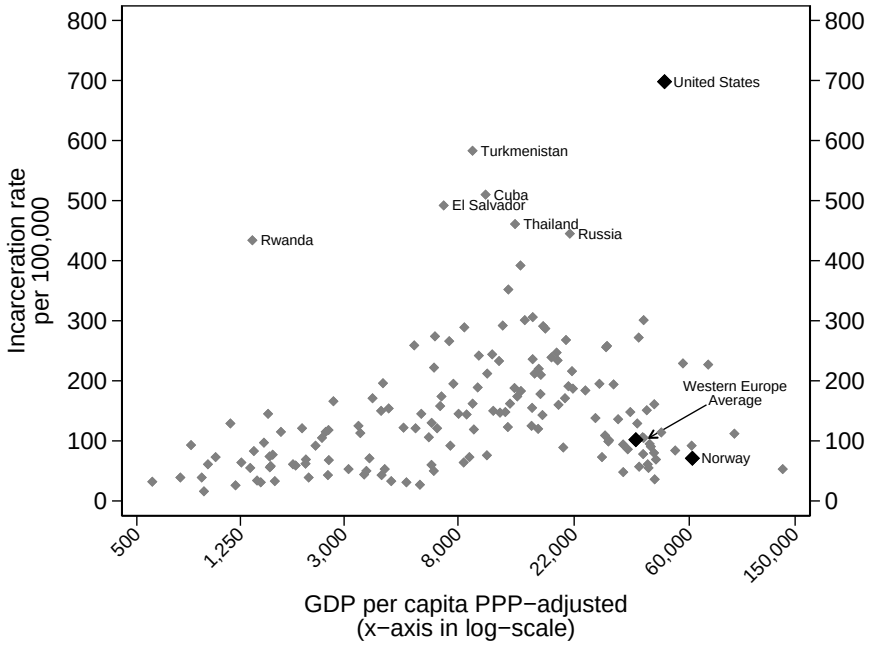


Figure A2. Incarceration Rates versus GDP per Capita.

Note: Sample consists of 160 countries with population greater than 0.5 million and with available data on incarceration and GDP. Incarceration rates and GDP are for the latest available year. GDP per capita is adjusted for purchasing power parity (PPP) and reported in 2010 US dollars. The Western European countries used to construct the population-weighted average include Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the UK. Sources: Institute for Criminal Policy Research, International Monetary Fund and the World Bank.

Table A1. Sample Restrictions.

	Sample Sizes (Remaining After Each Restriction):			
	No. of Cases (1)	**No. of Defendants (2)**	**No. of Judges (3)**	**No. of Courts (4)**
A. All Cases:	**128,804**	**83,143**	**1,059**	**87**
• drop cases with time limits (juveniles; defendant custody)	126,453	82,568	1,059	87
• drop cases with statutory sentence above 6 years	119,860	79,593	1,059	87
• drop cases where the defendant doesn't have right to a lawyer	101,930	68,042	1,050	87
B. Non-Confession Cases:	**76,609**	**49,989**	**1,053**	**87**
• drop cases with time limits (juveniles; defendant custody)	74,258	49,247	1,052	87
• drop cases with statutory sentence above 6 years	68,601	46,358	1,051	87
• drop cases where the defendant doesn't have right to a lawyer	50,671	33,182	1,041	87
C. Non-Confession Cases	**55,098**	**37,934**	**562**	**87**
Assigned to Regular Judges:				
• drop cases with time limits (juveniles; defendant custody)	53,294	37,238	562	87
• drop cases with statutory sentence above 6 years	47,776	34,167	561	87
• drop cases where the defendant doesn't have right to a lawyer	35,129	24,299	558	87
• drop courts with less than 2 regular judges stationed	34,554	23,936	557	84
• drop judges who have handled less than 50 criminal cases (baseline)	33,548	23,373	500	83

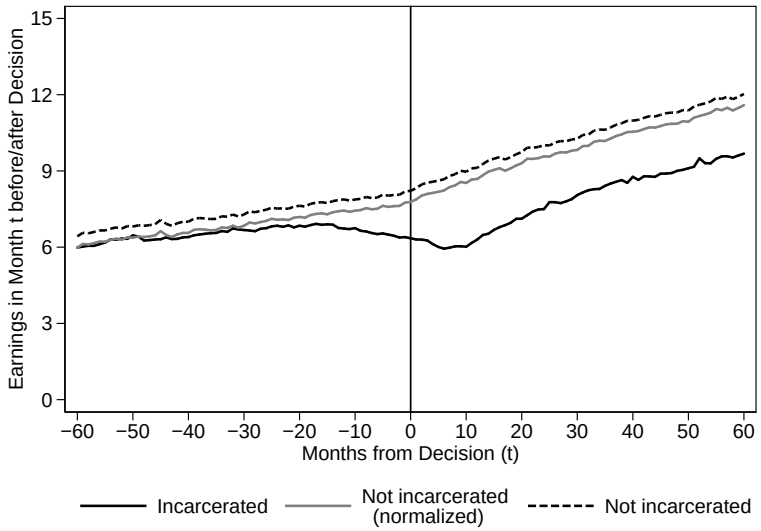
Note: The initial sample consists of all criminal cases processed in Norwegian district courts between 2005-2009.

Table A2. Descriptive Statistics by Previous Labor Market Attachment.

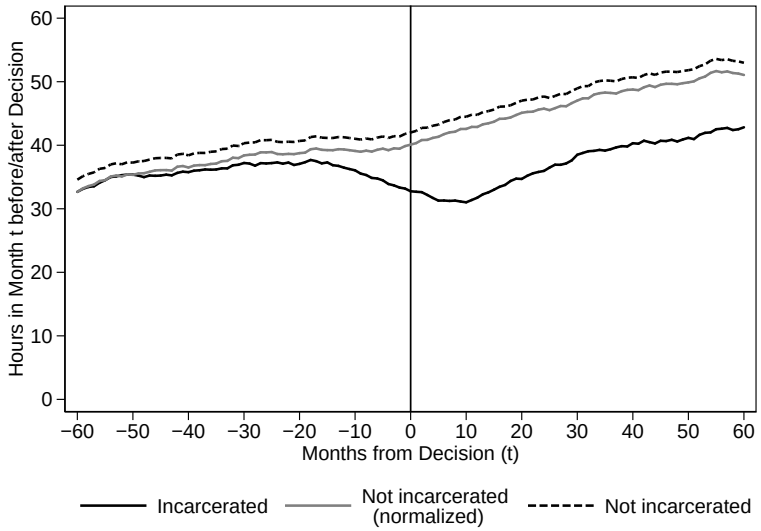
	Baseline Sample		*Sub-sample:* Previously Employed		*Sub-sample:* Previously Non-employed	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
	(1)	(2)	(3)	(4)	(5)	(6)
A. Defendant Characteristics:						
Demographics:						
Age	33.07	(11.78)	34.54	(10.60)	31.23	(12.88)
Female	0.119	(0.324)	0.102	(0.302)	0.141	(0.348)
Foreign born	0.148	(0.355)	0.143	(0.350)	0.153	(0.360)
Married, year t-1	0.128	(0.334)	0.168	(0.374)	0.078	(0.269)
Number of children, year t-1	0.822	(1.284)	0.956	(1.322)	0.656	(1.215)
Some college, year t-1	0.056	(0.229)	0.079	(0.270)	0.025	(0.157)
High school degree, year t-1	0.186	(0.389)	0.267	(0.442)	0.084	(0.278)
Less than high school, year t-1	0.758	(0.428)	0.654	(0.474)	0.890	(0.312)
Missing Xs	0.034	(0.181)	0.024	(0.153)	0.046	(0.210)
Past Work and criminal history:						
Employed, year t-1	0.393	(0.488)	0.680	(0.467)	0.037	(0.188)
Ever Employed, years t-2 to t-5	0.505	(0.499)	0.887	(0.317)	0.031	(0.174)
Charged, year t-1	0.378	(0.485)	0.314	(0.464)	0.458	(0.498)
Ever Charged, years t-2 to t-5	0.572	(0.495)	0.536	(0.499)	0.618	(0.486)
Incarcerated, year t-1	0.087	(0.282)	0.067	(0.249)	0.113	(0.316)
Ever Incarcerated, years t-2 to t-5	0.204	(0.403)	0.176	(0.380)	0.240	(0.427)
Number of defendants	23,373		12,938		10,435	
B. Type of Crime:						
Violent crime	0.256	(0.437)	0.246	(0.431)	0.267	(0.442)
Property crime	0.139	(0.346)	0.105	(0.306)	0.176	(0.381)
Economic crime	0.113	(0.316)	0.157	(0.364)	0.064	(0.246)
Drug related	0.119	(0.324)	0.104	(0.305)	0.136	(0.343)
Drunk driving	0.071	(0.257)	0.082	(0.274)	0.059	(0.236)
Other traffic	0.087	(0.281)	0.101	(0.301)	0.071	(0.256)
Other crimes	0.215	(0.419)	0.205	(0.404)	0.225	(0.418)
Number of cases	33,548		17,421		16,127	

Note: Baseline sample of non-confession criminal cases processed 2005-2009.

Appendix B. Additional Results for the Baseline Sample



(a) Earnings in Month t (in 1000 Norwegian Kroner)



(b) Hours Worked in Month t

Figure B1. Earnings and Hours Worked before and after Month of Court Decision.

Note: Baseline sample consisting of 33,548 non-confession criminal cases processed 2005-2009. Defendants are categorized in two groups, either incarcerated as shown in the solid black line or not incarcerated as shown in the dashed black line. To ease the comparison of trends, in each panel we normalize the level of the not incarcerated group's outcomes to the level of the incarcerated group's outcome in month t=-60. Outcomes for this "normalized" not incarcerated group are shown by the gray solid line. In both panels, the x-axis denotes months since court decision (normalized to period 0).

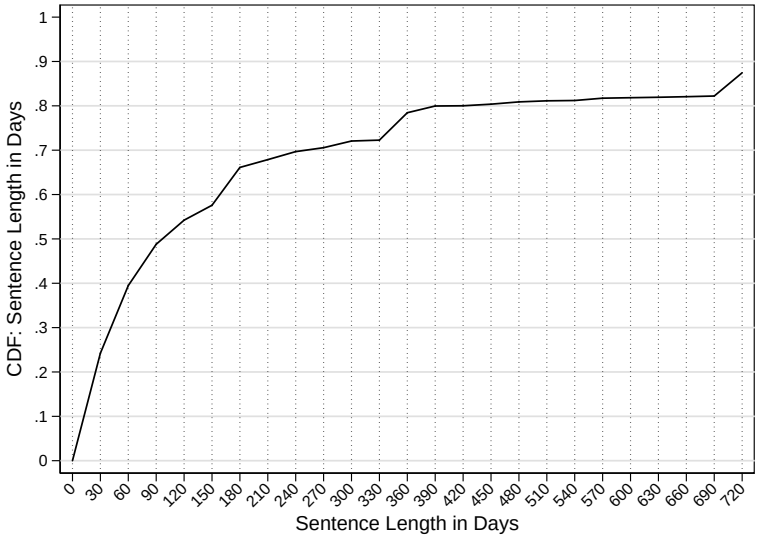
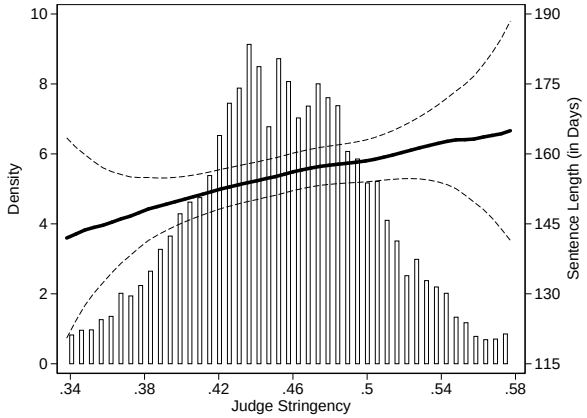
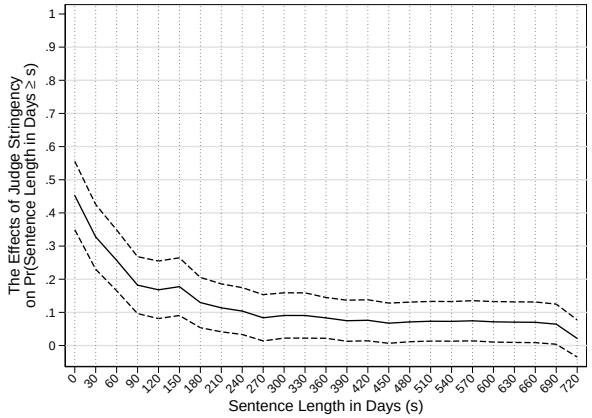


Figure B2. The CDF of Sentence Length in Days Conditional on Incarceration.

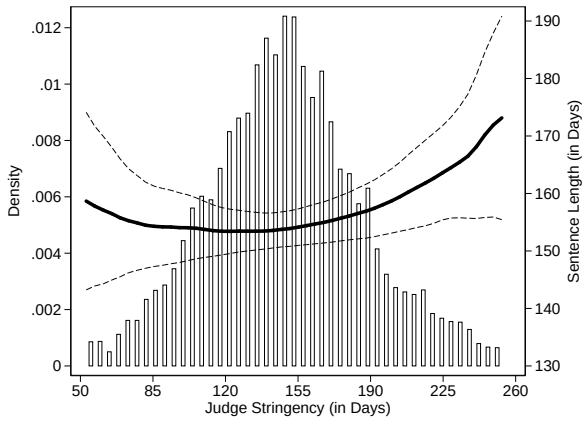
Note: Sample consisting of 17,052 non-confession criminal cases processed 2005-2009 which resulted in an incarceration decision.



(a) Judge Incarceration Stringency on Sentence Length



(b) Judge Incarceration Stringency on Pr(Sentence Length $\geq s$)



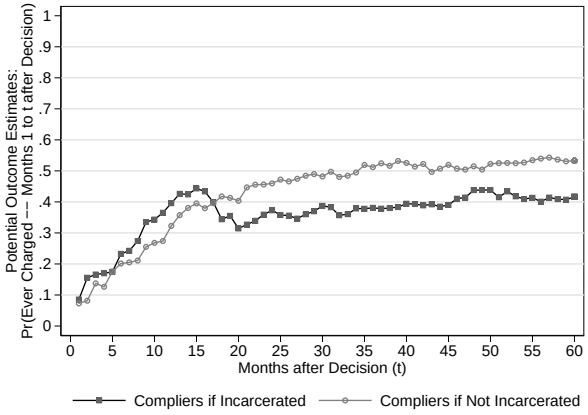
(c) Judge Sentence Length Stringency on Sentence Length

Figure B3. First Stage Graphs of Sentence Length on Judge Stringency.

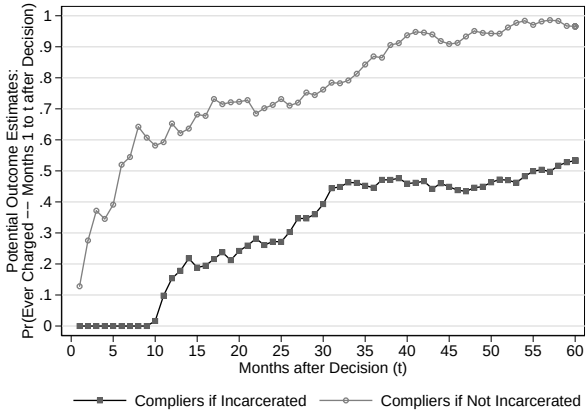
Note: Baseline sample consisting of 33,548 non-confession criminal cases processed 2005-2009. Sentence length is plotted on the right y-axis against leave-out mean judge incarceration stringency (plot a) and leave-out mean judge sentence length stringency (plot c) of the assigned judge shown along the x-axis. The plotted values are mean-standardized residuals from regressions on court x court entry year interacted fixed effects and all variables listed in Table 1. The solid line shows a local linear regression of sentence length on judge stringency. The histograms in plots a and c shows the density of judge stringency along the left y-axis (top and bottom 2% excluded). Plot b shows the estimates of judge incarceration stringency on Pr(Sentence Length $\leq$ s). Dashed lines show 90% confidence intervals.

Column A: Previously Employed

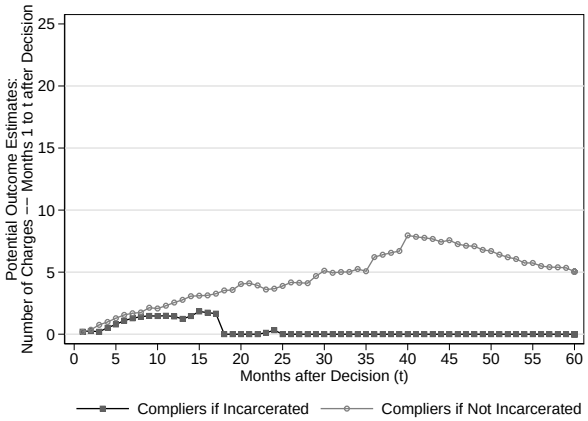
Column B: Previously Non-employed



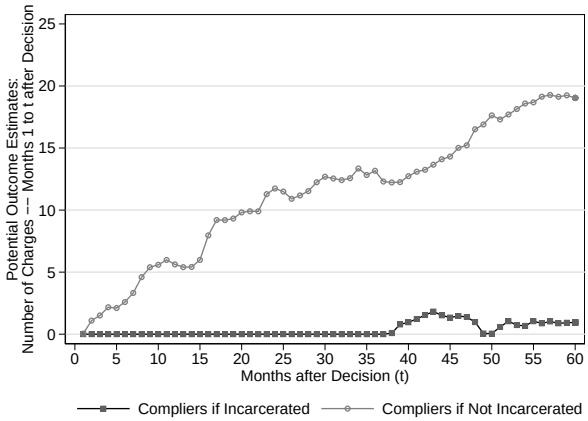
(a) Potential Outcomes: Pr(Ever Charged – Months 1 to t)



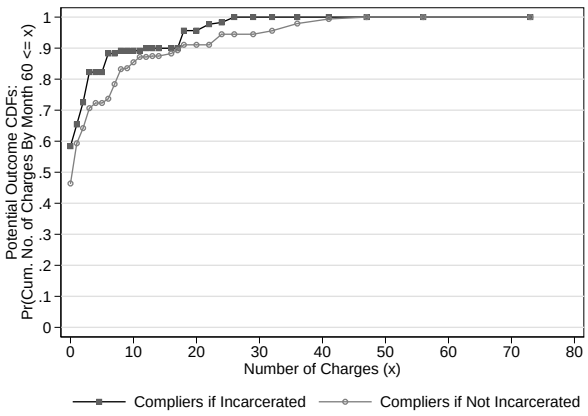
(b) Potential Outcomes: Pr(Ever Charged – Months 1 to t)



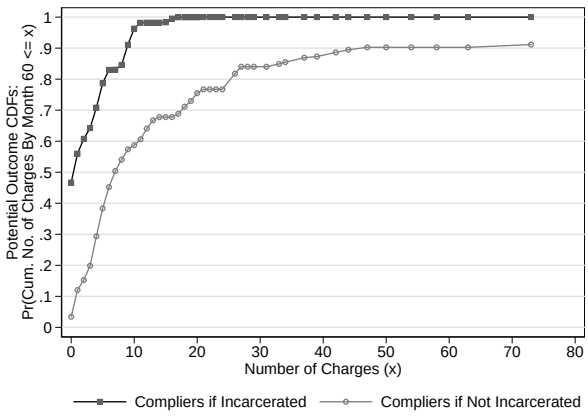
(c) Potential Outcomes: No. of Charges – Months 1 to t



(d) Potential Outcomes: No. of Charges – Months 1 to t



(e) Potential Outcome CDFs: No. of Charges by Month 60



(f) Potential Outcome CDFs: No. of Charges by Month 60

Figure B4. Potential Outcomes for Recidivism by Previous Labor Market Attachment.

Note: Baseline sample consisting of non-confession criminal cases processed 2005-2009 (N=33,548 at time of decision and N=31,428 in month 60 after decision). Dashed lines show 90% confidence intervals.

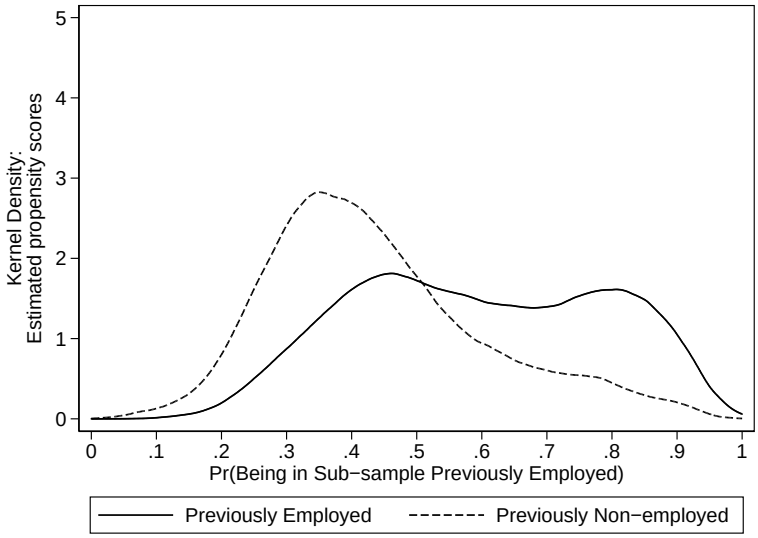
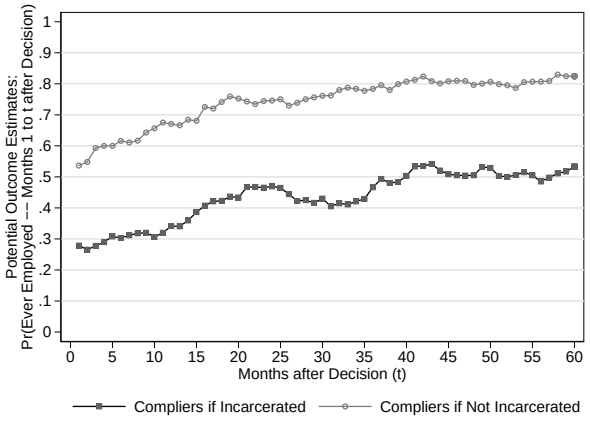


Figure B5. Propensity Score Overlap by Previous Employment Status.

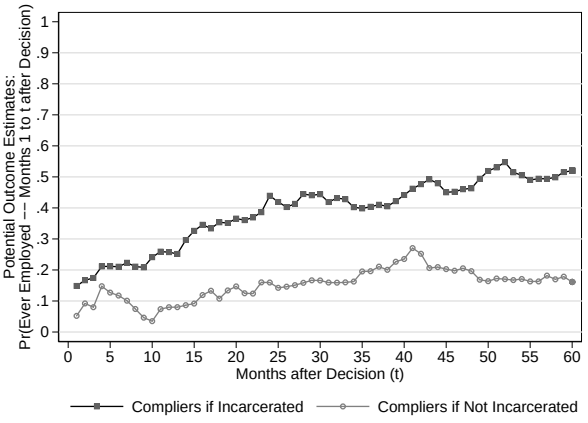
Note: Baseline balanced sample consisting of 33,548 non-confession criminal cases processed 2005-2009. The estimated propensity score is a composite index of all variables listed in Table 1, excluding the variables directly capturing past work history which would also fully predict the probability of being in either sub-sample with no overlap. In Tables 5-6, columns (2) and (4), we use the estimated propensity scores to adjust for differences in observable characteristics across sub-samples; see discussion of the re-weighting procedure in Section 6.1.

Column A: Previously Employed

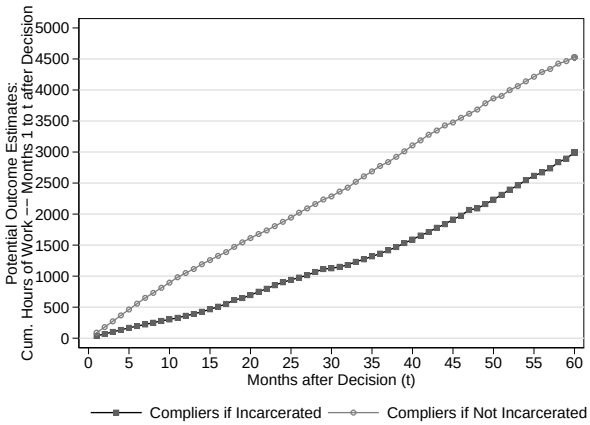
Column B: Previously Non-employed



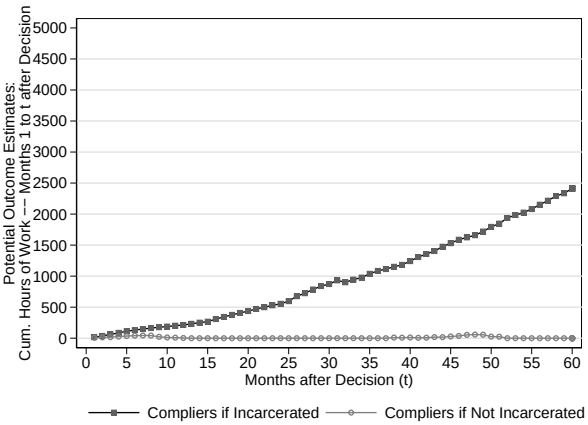
(a) Potential Outcomes: Pr(Ever Employed – Months 1 to t)



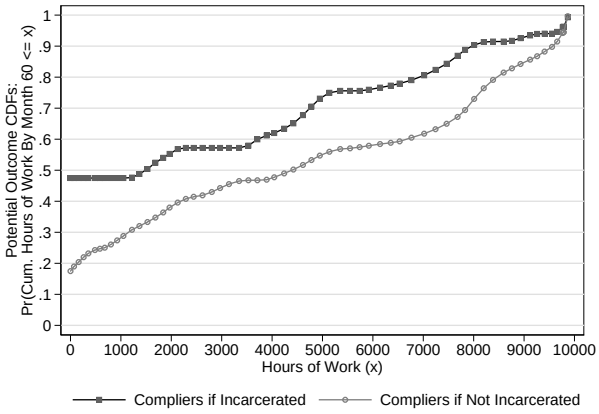
(b) Potential Outcomes: Pr(Ever Employed – Months 1 to t)



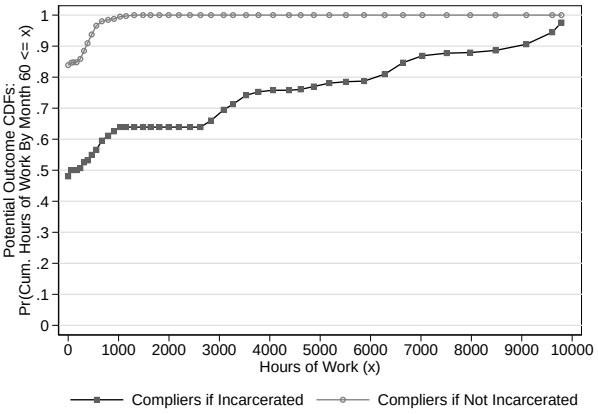
(c) Potential Outcomes: Hours of Work – Months 1 to t



(d) Potential Outcomes: Hours of Work – Months 1 to t



(e) Potential Outcome CDFs: Hours of Work by Month 60

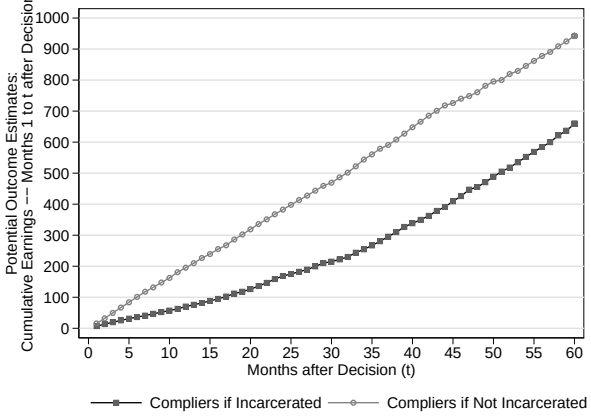


(f) Potential Outcome CDFs: Hours of Work by Month 60

Figure B6. Potential Outcomes for Employment and Hours of Work by Previous Labor Market Attachment.

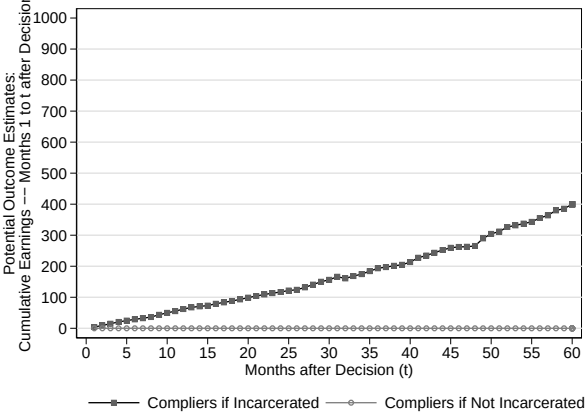
Note: Baseline sample of non-confession criminal cases processed 2005-2009 (N=33,548 at time of decision and N=31,428 in month 60 after decision). Dashed lines show 90% confidence intervals.

Column A: Previously Employed

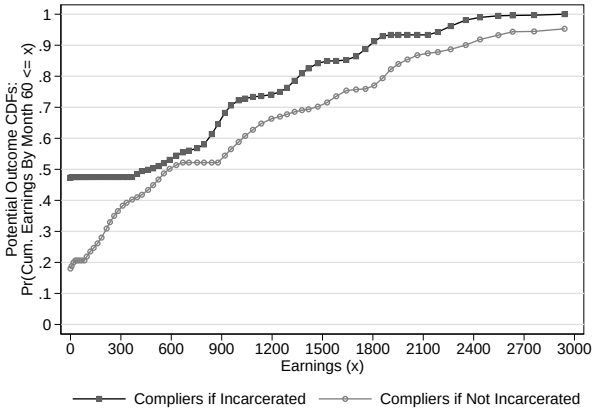


(a) Potential Outcomes: Cum. Earnings – Month 1 to t

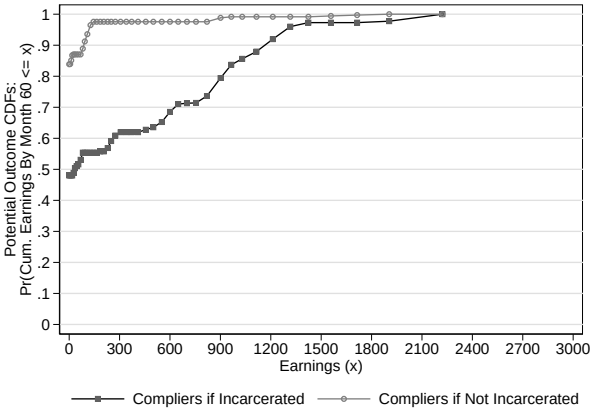
Column B: Previously Non-employed



(b) Potential Outcomes: Cum. Earnings – Month 1 to t



(c) Potential Outcome CDFs: Cum. Earnings by Month 60



(d) Potential Outcome CDFs: Cum. Earnings by Month 60

Figure B7. Potential Outcomes for Earnings by Previous Labor Market Attachment.

Note: Baseline sample of non-confession criminal cases processed 2005-2009 (N=33,548 at time of decision and N=31,428 in month 60 after decision). Earnings are reported in 1,000 Norwegian Kroner. Dashed lines show 90% confidence intervals.

Table B1. Test for Selective Sample Attrition.

	(1)	(2)	(3)	(4)	(5)
***Dependent Variable*:**	Pr(Attrited by Month 12 after Decision)	Pr(Attrited by Month 24 after Decision)	Pr(Attrited by Month 36 after Decision)	Pr(Attrited by Month 48 after Decision)	Pr(Attrited by Month 60 after Decision)
RF: Judge Stringency	-0.009	-0.012	-0.011	-0.005	-0.042
	(0.009)	(0.020)	(0.023)	(0.026)	(0.026)
Dependent mean	0.008	0.023	0.036	0.050	0.063
Number of cases			33,548		

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1. Standard errors are two-way clustered at judge and defendant level. Sample attrition can only occur due to either death or emigration and without such 'natural' attrition our baseline sample would be fully balanced over months 1-60 after decision. **p<0.1, **p<0.05, ***p<0.01.

Table B2. Tests for the Monotonicity Assumption.

	Baseline Instrument:	Reverse-sample Instrument:
	(1)	(2)
	First Stage	First Stage
Dependent Variable:	*Pr(Incarcerated)*	*Pr(Incarcerated)*
A. INCARCERATION PROPENSITY (ALL COVARIATES):		
1. Sub-sample: Incarceration propensity – 1st quartile (lowest)		
Estimate	0.5047***	0.5331***
(SE)	(0.1130)	(0.1144)
Dependent mean	0.3737	0.3741
Number of cases	7,856	7,693
2. Sub-sample: Incarceration propensity – 2nd quartile		
Estimate	0.5691***	0.6004***
(SE)	(0.0951)	(0.0936)
Dependent mean	0.4702	0.4695
Number of cases	7,858	7,708
3. Sub-sample: Incarceration propensity – 3rd quartile		
Estimate	0.4058***	0.4495***
(SE)	(0.1143)	(0.1071)
Dependent mean	0.5437	0.5415
Number of cases	7,857	7,714
4. Sub-sample: Incarceration propensity – 4th quartile (highest)		
Estimate	0.3975***	0.3793***
(SE)	(0.0956)	(0.0979)
Dependent mean	0.6303	0.6300
Number of cases	7,857	7,676
B. TYPE OF CRIME:		
1. Sub-sample: Type of crime – Violent crimes		
Estimate	0.6303***	0.5648***
(SE)	(0.1117)	(0.1146)
Dependent mean	0.5530	0.5520
Number of cases	8,118	8,016
2. Sub-sample: Type of crime – Drug-related crimes		
Estimate	0.4088***	0.3653***
(SE)	(0.1423)	(0.1448)
Dependent mean	0.5227	0.5221
Number of cases	4,354	4,292
3. Sub-sample: Type of crime – Property crimes		
Estimate	0.4820***	0.5402***
(SE)	(0.1551)	(0.1445)
Dependent mean	0.4281	0.4268
Number of cases	3,586	3,554
4. Sub-sample: Type of crime – Economic crimes		
Estimate	0.5552***	0.5352***
(SE)	(0.1615)	(0.1587)
Dependent mean	0.4774	0.4767
Number of cases	3,697	3,671
5. Sub-sample: Type of crime – Drunk-driving & other traffic offenses		
Estimate	0.2873**	0.3319***
(SE)	(0.1367)	(0.1253)
Dependent mean	0.5009	0.4998
Number of cases	4,965	4,782
6. Sub-sample: Type of crime – Other crimes		
Estimate	0.4275***	0.4673***
(SE)	(0.1164)	(0.1195)
Dependent mean	0.4924	0.4911
Number of cases	6,708	6,5697
C. PREVIOUS LABOR MARKET ATTACHMENT:		
1. Sub-sample: Previously Employed		
Estimate	0.5327***	0.3607***
(SE)	(0.0810)	(0.0765)
Dependent mean	0.5060	0.5059
Number of cases	16,547	14,694

(continued on the next page)

Table B2. Tests for the Monotonicity Assumption.

	Baseline Instrument:	Reverse-sample Instrument:
	(1)	(2)
	First Stage	First Stage
Dependent Variable:	*Pr(Incarcerated)*	*Pr(Incarcerated)*
(continued from the previous page)		
C. PREVIOUS LABOR MARKET ATTACHMENT:		
2. Sub-sample: Previously Non-employed		
Estimate	0.4223***	0.3618***
(SE)	(0.0905)	(0.0787)
Dependent mean	0.5029	0.5059
Number of cases	14,881	14,101
D. PREVIOUS INCARCERATION STATUS:		
1. Sub-sample: Previously Non-incarcerated		
Estimate	0.41400***	0.3291***
(SE)	(0.0806)	(0.0732)
Dependent mean	0.3948	0.3973
Number of cases	15,386	13,833
2. Sub-sample: Previously Incarcerated		
Estimate	0.4784***	0.4784***
(SE)	(0.0810)	(0.0810)
Dependent mean	0.6096	0.6085
Number of cases	16,042	15,311
E. AGE:		
1. Sub-sample: Age at time of first offense > 30		
Estimate	0.4007***	0.3160***
(SE)	(0.0871)	(0.0851)
Dependent mean	0.5238	0.5275
Number of cases	15,472	14,314
2. Sub-sample: Age at time of first offense < 30		
Estimate	0.5586***	0.4664***
(SE)	(0.0774)	(0.0850)
Dependent mean	0.4858	0.4881
Number of cases	15,956	14,760
F. LEVEL OF EDUCATION:		
1. Sub-sample: Less than high school at time of first offense		
Estimate	0.4584***	0.3064***
(SE)	(0.0698)	(0.0789)
Dependent mean	0.5110	0.5098
Number of cases	22,651	16,718
2. Sub-sample: High school or above at time of first offense		
Estimate	0.4638***	0.4622***
(SE)	(0.1055)	(0.1054)
Dependent mean	0.4876	0.4866
Number of cases	8,777	8,5245
G. NUMBER OF CHILDREN:		
1. Sub-sample: Had no children at time of first offense		
Estimate	0.4657***	0.3667***
(SE)	(0.0746)	(0.0691)
Dependent mean	0.4956	0.4995
Number of cases	19,711	15,743
2. Sub-sample: Had children at time of first offense		
Estimate	0.4506***	0.4573***
(SE)	(0.1000)	(0.0932)
Dependent mean	0.5194	0.5181
Number of cases	11,717	11,305

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. To reduce noise in judge stringency instruments, judges with less than 50 handled cases are dropped from each sample. **p<0.1, **p<0.05, ***p<0.01.

Table B3. Year-By-Year Estimates of the Effects of Incarceration on Recidivism.

	Months 1-12 after Decision	*Months 13-24 after Decision*	*Months 25-36 after Decision*	*Months 37-48 after Decision*	*Months 49-60 after Decision*
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
***Dependent Variable*:**			**A. Pr(Ever Charged)**		
OLS: Incarcerated	0.062***	0.053***	0.060***	0.041***	0.049***
All controls	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)
RF: Judge Stringency	-0.079	-0.090*	-0.048	-0.096*	-0.027
All controls	(0.050)	(0.052)	(0.046)	(0.054)	(0.053)
IV: Incarcerated	-0.175	-0.199*	-0.106	-0.213*	-0.059
All controls	(0.114)	(0.119)	(0.105)	(0.122)	(0.117)
Dependent mean	0.44	0.40	0.38	0.36	0.35
***Dependent Variable*:**			**B. Number of Charges**		
OLS: Incarcerated	0.543***	0.594***	0.705***	0.485***	0.478***
All controls	(0.090)	(0.081)	(0.081)	(0.073)	(0.070)
RF: Judge Stringency	-1.734**	-1.440**	-1.179	-0.592	-0.251
All controls	(0.727)	(0.667)	(0.868)	(0.852)	(0.584)
IV: Incarcerated	-3.832**	-3.182**	-2.604	-1.309	-0.554
All controls	(1.722)	(1.587)	(1.979)	(1.874)	(1.291)
Dependent mean	2.42	2.22	2.03	1.84	1.71
Number of cases			31,428		

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1. RF and IV in addition also control for court x court entry year FEs. OLS standard errors are clustered at the defendant level, while RF and IV standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Table B4. The Effects of Incarceration on Re-Incarceration and Days Spent in Prison in New Cases.

Dependent Variable:	**Pr(Ever Charged)**	**Pr(Ever Incarcerated in New Cases)**	**Pr(Ever Incarcerated in News Cases | Ever Charged)**	**Total Days Spent in Prison in New Cases**
	(1)	**(2)**	**(3)**	**(4)**
RF: Judge Stringency	-0.133***	0.005	0.123*	2.06
All controls	(0.044)	(0.048)	(0.065)	(25.08)
IV: Incarcerated	-0.293***	0.011	0.310*	4.55
All controls	(0.106)	(0.107)	(0.161)	(55.37)
Dependent mean	0.70	0.42	0.59	93.10
Number of cases	31,428	31,428	22,031	31,428

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1. RF and IV in addition also control for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Table B5. Characterization of Compliers.

		(1) Previously Employed	(2) Previously Non-Employed
1. Sub-sample: Incarceration propensity – 1st quartile (lowest)			
Population share:	$Pr[X_i = x]$	0.139	0.113
Complier share:	$Pr[Complier \mid X_i = x]$	0.152	0.155
Complier cond. pop. share:	$Pr[X_i = x \mid Complier]$	0.159	0.132
Complier relative likelihood:	$\frac{Pr[X_i=x \mid Complier]}{Pr[X_i=x]}$	1.144	1.168
2. Sub-sample: Incarceration propensity – 2nd quartile			
Population share:	$Pr[X_i = x]$	0.129	0.116
Complier share:	$Pr[Complier \mid X_i = x]$	0.220	0.114
Complier cond. pop. share:	$Pr[X_i = x \mid Complier]$	0.213	0.099
Complier relative likelihood:	$\frac{Pr[X_i=x \mid Complier]}{Pr[X_i=x]}$	1.651	0.853
3. Sub-sample: Incarceration propensity – 3rd quartile			
Population share:	$Pr[X_i = x]$	0.121	0.126
Complier share:	$Pr[Complier \mid X_i = x]$	0.116	0.104
Complier cond. pop. share:	$Pr[X_i = x \mid Complier]$	0.106	0.099
Complier relative likelihood:	$\frac{Pr[X_i=x \mid Complier]}{Pr[X_i=x]}$	0.876	0.786
4. Sub-sample: Incarceration propensity – 4th quartile (highest)			
Population share:	$Pr[X_i = x]$	0.134	0.119
Complier share:	$Pr[Complier \mid X_i = x]$	0.143	0.093
Complier cond. pop. share:	$Pr[X_i = x \mid Complier]$	0.148	0.083
Complier relative likelihood:	$\frac{Pr[X_i=x \mid Complier]}{Pr[X_i=x]}$	1.072	0.697
Number of cases		17,421	16,127

Note: Baseline sample of non-confession criminal cases processed 2005-2009. We split the sample into eight mutually exclusive and collectively exhaustive subgroups based on previous labor market attachment and quartiles of the predicted probability of incarceration which is estimated based on all variables listed in Table 1. We estimate the first stage equation (2) separately for each subgroup, which allows us to calculate the proportion of compliers by subgroup. For each subgroup, we report the population share (row 1), the complier share (row 2), and the probability of being in a subgroup conditional on being a complier (row 3). Finally, we also report the complier relative likelihood (row 4), which is the ratio of group-specific complier share to the overall complier share estimated to be 0.133 for the full baseline sample.

Table B6. Specification Checks.

		Sample Selection			*Definition of Instrument*	
	Baseline	$\geq$ 25 cases handled by each judge	$\geq$ 75 cases handled by each judge	$\geq$100 cases handled by each judge	Split-sample instrument	Non-confession sample instrument
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Dependent Variable:			**A. Pr(Incarcerated)**			
First stage	0.4525***	0.4516***	0.4631***	0.4305***	0.4278***	0.3622***
	(0.0634)	(0.0591)	(0.0664)	(0.0677)	(0.0734)	(0.0507)
Dep. mean	0.5045	0.5045	0.5035	0.5072	0.5058	0.5054
No. of cases	31,428	32,152	30,582	29,339	15,476	28,860
Dependent Variable:			**B. Pr(Ever Charged)**			
			Months 1-60 after Decision			
RF	-0.133***	-0.113***	-0.127***	-0.111**	-0.110**	-0.087**
	(0.045)	(0.044)	(0.045)	(0.048)	(0.055)	(0.038)
IV	-0.293***	-0.249**	-0.275**	-0.259**	-0.258*	-0.240**
	(0.106)	(0.102)	(0.105)	(0.119)	(0.111)	(0.111)
Dep. mean	0.70	0.70	0.70	0.70	0.70	0.70
No. of cases	31,428	32,011	30,582	29,339	15,476	28,860
Dependent Variable:			**C. Number of Charges**			
			Months 1-60 after Decision			
RF	-5.196**	-4.074*	-5.426**	-4.593*	-5.661**	-4.230*
	(2.452)	(2.208)	(2.566)	(2.775)	(2.485)	(2.083)
IV	-11.482**	-9.021*	-11.739**	-10.668	-13.235**	-11.680*
	(5.705)	(5.050)	(5.866)	(6.759)	(6.285)	(6.246)
Dep. mean	10.21	10.23	10.21	10.21	10.28	10.20
No. of cases	31,428	32,011	30,582	29,339	15,476	28,860

	D. Weak Instrument Robust Inference	
	Pr(Ever Charged)	**Number of Charges**
Weak Instrument Robust Confidence Intervals:		
Standard Wald CIs (at 95% conf. level)	[-0.501, -0.086]	[-22.664, -0.301]
Anderson-Rubin CIs (at 95% conf. level)	[-0.532, -0.105]	[-23.906, -0.866]
Montiel Olea-Pflueger Weak Instrument Test:		
Effective F-statistic (at 5% confidence level)	50.60	
Critical Value 2SLS (τ=10% of worse case bias)	23.11	

Note: Controls include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. Column (1) shows baseline estimates using leave-case-out mean judge stringency as an instrument for incarceration decision. Baseline estimates in column (1) include cases assigned to judges who have handled at least 50 cases. In columns (2)-(4), we instead require judges to handle at least 25, 75 and 100 cases, respectively. In column (5), we first i) randomly split the baseline estimation sample in two equal-sized and mutually exclusive sub-samples, ii) retain one of these subsample and construct the instrument using each judge's case decisions in the other subsample, and finally, iii) estimate the IV model given by equations (2)-(1) using the retained subsample. In column (6), we only use non-confession cases to construct measures of leave-out mean judge stringency. Panel D reports weak instrument robust tests and confidence intervals for the baseline sample and specification (see details in Section 5.3 and Appendix D). *p<0.1, **p<0.05, ***p<0.01.

Table B7. IV Model Interacted with Sub-Sample Indicators for High or Low Open Prison Propensity.

	First Stages		Reduced Form		IV
	(1)	(2)	(3)		(4)
	Decision:	*Decision:*	*Months 1-60 after Decision:*		*Months 1-60 after Decision:*
	Pr(Incarcerated and Below-Median Pr(Open Prison))	**Pr(Incarcerated and Above-Median Pr(Open Prison))**	**Pr(Ever Charged)**		**Pr(Ever Charged)**
A. Baseline Specification					
Instrument:				*Decision:*	
Judge Stringency	0.453***		-0.134***	*Incarcerated*	-0.293***
	(0.063)		(0.045)		(0.106)
F-stat (Instrument)	50.24				
Dependent Mean	0.5045		0.7010	Dependent Mean	0.7010
B. Interacted Specification					
Instruments:				*Decisions:*	
Judge Stringency X	0.782***	-0.297***	-0.116***	*Incarcerated X*	-0.269***
Below-Median Pr(Open Prison)	(0.042)	(0.043)	(0.046)	*Below-Median Pr(Open Prison)*	(0.103)
Judge Stringency X	-0.277***	0.700***	-0.147***	*Incarcerated X*	-0.317***
Above-Median Pr(Open Prison)	(0.041)	(0.045)	(0.045)	*Above-Median Pr(Open Prison)*	(0.110)
SW F-stat (Instrument)	47.51	46.56			
Dependent Mean	0.2836	0.2209	0.7010	Dependent Mean	0.7010
Number of Cases	31,428	31,428	31,428		31,428

Note: Baseline sample consisting of non-confession criminal cases processed 2005-2009 (N=31,428 in month 60 after decision). All estimations include controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. **p<0.1, **p<0.05, ***p<0.01.

Table B8. Summary Measures of Treatment Effects Based on the 2SLS and the MTE.

A. Treatment Parameters Based on the 2SLS

	Local Average Treatment Effect (LATE) for the Baseline Sample	**Local Average Treatment Effect (LATE) for the Common Support Sample**
	(1)	(2)
Pr(Ever Charged)	-0.293***	-0.241***
	(0.106)	(0.105)
Number of cases	31,428	28,275

B. Treatment Parameters Based on the MTE for the Common Support Sample

	Average Treatment Effect on the Treated (ATT)	**Average Treatment Effect (ATE)**	**Average Treatment Effect on the Untreated (ATUT)**
	(1)	(2)	(3)
1. Linear Specification			
Pr(Ever Charged)	-0.424**	-0.252***	-0.100
	(0.143)	(0.094)	(0.121)
Number of cases	28,275	28,275	28,275
2. Global Quadratic Polynomial			
Pr(Ever Charged)	-0.423***	-0.247**	-0.091
	(0.158)	(0.124)	(0.160)
Number of cases	28,275	28,275	28,275
3. Global Cubic Polynomial			
Pr(Ever Charged)	-0.508**	-0.213*	0.047
	(0.200)	(0.122)	(0.165)
Number of cases	28,275	28,275	28,275
4. Global Quartic Polynomial			
Pr(Ever Charged)	-0.510***	-0.213**	0.048
	(0.173)	(0.104)	(0.161)
Number of cases	28,275	28,275	28,275
5. Semiparametric Specification			
Pr(Ever Charged)	-0.332**	-0.220*	-0.119
	(0.146)	(0.123)	(0.169)
Number of cases	28,275	28,275	28,275

Note: Baseline sample of non-confession criminal cases processed 2005-2009 and trimmed sample with common support. The rescaled treatment parameters are weighted averages (for the treated (ATT), for all (ATE), and for the untreated (ATUT)) over the MTE curves over the area with common support (weights sum to 1). The semiparametric specification is a local linear regression with 100 gridpoints. Standard errors are constructed based on 100 nonparametric bootstrap replications.

Table B9. Controlling for Judge Stringency in Decision Margins Other Than Incarceration.

	First Stage	Reduced Form		IV	
	(1)	(2)	(3)	(4)	(5)
	Decision:	*Months 1-60 after Decision:*		*Months 1-60 after Decision:*	
	Pr(Incarcerated)	**Pr(Ever Charged)**	**No. of Charges**	**Pr(Ever Charged)**	**No. of Charges**
A. Baseline Specification					
	0.4525***	-0.133***	-5.196**	-0.293***	-11.482**
	(0.0634)	(0.045)	(2.452)	(0.106)	(5.704)
F-stat. (Instrument)	50.23 [0.000]				
B. Controls for 'Probation Stringency', 'Community Service Stringency' and 'Fine Stringency'					
	0.3912***	-0.111*	-5.621*	-0.283*	-14.369*
	(0.0742)	(0.060)	(2.869)	(0.165)	(8.244)
F-stat. (Instrument)	27.83 [0.000]				
C. Control for 'Probation, Community Service or Fine Stringency'					
	0.4337***	-0.130***	-5.000**	-0.299***	-11.530*
	(0.0626)	(0.046)	(2.413)	(0.114)	(5.942)
F-stat. (Instrument)	47.98 [0.000]				
D. Controls for 'Probation, Community Service or Fine Stringency' and 'Sentence Length Stringency'					
	0.4151***	-0.137**	-2.795	-0.329**	-6.734
	(0.0804)	(0.057)	(2.948)	(0.154)	(7.254)
F-stat. (Instrument)	26.67 [0.000]				
Number of cases			31,428		

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Table B10. IV Model with Three Decision Margins: 'Incarceration', 'Probation, Community Service or Fine', and 'Not Guilty'.

	First Stages		Reduced Form			IV	
	(1)	(2)	(3)	(4)		(5)	(6)
	Decision:		*Months 1-60 after Decision:*			*Months 1-60 after Decision:*	
	Pr(Incarcerated)	**Pr(Probation, CS or Fine)**	**Pr(Ever Charged)**	**No. of Charges**		**Pr(Ever Charged)**	**No. of Charges**
A. Baseline Specification							
Instrument:					*Decision:*		
Incarceration stringency	0.4525***		-0.133***	-5.196**	Incarcerated	-0.293***	-11.482**
	(0.0634)		(0.045)	(2.452)		(0.106)	(5.705)
F-stat. (Instrument)	50.23 [0.000]						
Dependent mean	0.50		0.70	10.21		0.70	10.21
B. Specification with Three Decision Margins							
Instruments:					*Decisions:*		
Incarceration stringency	0.4337***	-0.0160	-0.130***	-5.000**	Incarcerated	-0.297***	-11.512**
	(0.063)	(0.039)	(0.046)	(2.413)		(0.111)	(5.839)
Probation, CS, or Fine	0.2202**	0.4970***	-0.003	-2.287	Probation, CS,	0.063	0.498
stringency	(0.102)	(0.081)	(0.079)	(3.910)	or Fine	(0.194)	(8.675)
Sanderson-Windmeijer F-stat.	43.76 [0.000]	30.10 [0.000]					
Dependent mean	0.50	0.89	0.70	10.21		0.70	10.21
Number of cases				31,428			

Note: Baseline sample of non-confession criminal cases processed 2005-2009 . Control variables include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01. CS = Community Service. The omitted decision category in Panel A is "Probation, Community Service, Fine, or Not Guilty", while in Panel B the omitted category is "Not Guilty". *p<0.1, **p<0.05, ***p<0.01.

Table B11. The Effects of Sentence Length on Recidivism.

***Dependent Variable*:**	**Pr(Ever Charged)**			**Number of Charges**
	Months 1-24 after Decision	*Months 25-60 after Decision*	*Months 1-60 after Decision*	*Months 1-60 after Decision*
	(1)	**(2)**	**(3)**	**(4)**
OLS: Sentence length	0.042***	0.040***	0.034***	2.14***
No controls	(0.002)	(0.003)	(0.002)	(0.14)
OLS: Sentence length	0.035***	0.034***	0.028***	1.81***
Demographics & Type of Crime	(0.002)	(0.002)	(0.002)	(0.13)
OLS: Sentence length	0.023***	0.022***	0.018***	1.24***
All controls	(0.002)	(0.002)	(0.002)	(0.12)
RF: Judge Stringency	-0.108**	-0.111**	-0.133***	-5.20**
All controls	(0.047)	(0.048)	(0.045)	(2.45)
IV: Sentence length	-0.258	-0.265	-0.317*	-12.42
All controls	(0.173)	(0.181)	(0.181)	(8.31)
Dependent mean	0.57	0.57	0.70	10.21
Number of cases		31,428		

Note: Baseline sample of non-confession criminal cases processed 2005-2009. The estimates show the effects of an increase in sentence length by 250 days. Controls include all variables listed in Table 1. RF and IV in addition also control for court x court entry year FEs. OLS standard errors are clustered at the defendant level, while RF and IV standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Table B12. IV Model with Three Decision Margins: 'Incarceration', 'Sentence Length', and 'Probation, Community Service, Fine, or Not Guilty'.

	First Stages		Reduced Form			IV	
	(1)	(2)	(3)	(4)		(5)	(6)
	Decision:		*Months 1-60 after Decision:*			*Months 1-60 after Decision:*	
	Pr(Incarcerated)	**Sentence Length**	**Pr(Ever Charged)**	**No. of Charges**		**Pr(Ever Charged)**	**No. of Charges**
A. Baseline Specification							
Instrument:					*Decision:*		
Incarceration stringency	0.4525***		-0.133***	-5.196**	Incarcerated	-0.293***	-11.482**
	(0.0634)		(0.045)	(2.452)		(0.106)	(5.705)
F-stat. (Instrument)	50.23 [0.000]						
Dependent mean	0.50		0.70	10.21		0.70	10.21
B. Specification with Three Decision Margins							
Instruments:					*Decisions:*		
Incarceration stringency	0.411***	0.352	-0.136**	-2.797	Incarcerated	-0.793	52.192
	(0.081)	(0.223)	(0.057)	(2.948)		(2.839)	(341.00)
Sentence length stringency	0.019	0.031	0.001	-1.131	Sentence length	0.540	-68.885
	(0.025)	(0.079)	(0.016)	(0.808)		(3.076)	(372.16)
Sanderson-Windmeijer F-stat.	0.04 [0.8409]	0.04 [0.8441]					
Dependent mean	0.50	0.62	0.70	10.21		0.70	10.21
Number of cases	31,428						

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Control variables include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01. The omitted decision category is "Probation, Community Service, Fine, or Not Guilty". *p<0.1, **p<0.05, ***p<0.01.

Table B13. Year-By-Year Estimates of the Effects of Incarceration on Future Employment by Previous Labor Market Attachment.

	Months 1-12 after Decision	*Months 13-24 after Decision*	*Months 25-36 after Decision*	*Months 37-48 after Decision*	*Months 49-60 after Decision*
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
A. Sub-sample: Previously Non-employed					
***Dependent Variable*:**			**a. Pr(Ever Employed)**		
RF: Judge Stringency	0.057	0.111*	0.080	0.109*	0.132*
All controls	(0.069)	(0.067)	(0.065)	(0.063)	(0.070)
IV: Incarcerated	0.136	0.262	0.190	0.259	0.313*
All controls	(0.168)	(0.169)	(0.159)	(0.164)	(0.181)
Dependent mean	0.24	0.26	0.27	0.27	0.28
***Dependent Variable*:**			**b. Cumulative Hours of Work**		
RF: Judge Stringency	48.5	170.9*	229.3**	260.5***	309.9***
All controls	(78.8)	(87.7)	(92.8)	(97.0)	(104.1)
IV: Incarcerated	114.8	404.7*	542.9**	616.8**	733.8**
All controls	(192.1)	(234.9)	(249.7)	(277.2)	(304.6)
Dependent mean	238.7	282.8	313.0	327.9	352.0
***Dependent Variable*:**			**c. Cumulative Earnings (in 1000 NOK)**		
RF: Judge Stringency	19.9	26.2*	41.6**	37.9*	43.3**
All controls	(12.5)	(15.4)	(17.2)	(20.2)	(22.0)
IV: Incarcerated	47.1	62.0	98.6**	89.8*	102.4*
All controls	(32.9)	(40.3)	(47.2)	(54.3)	(59.2)
Dependent mean	33.4	44.0	52.4	59.4	66.4
Number of cases			14,881		
B. Sub-sample: Previously Employed					
***Dependent Variable*:**			**a. Pr(Ever Employed)**		
RF: Judge Stringency	-0.161**	-0.093	-0.111*	-0.080	-0.069
All controls	(0.075)	(0.067)	(0.063)	(0.066)	(0.069)
IV: Incarcerated	-0.302**	-0.174	-0.208*	-0.151	-0.129
All controls	(0.141)	(0.127)	(0.120)	(0.123)	(0.128)
Dependent mean	0.53	0.52	0.51	0.51	0.51
***Dependent Variable*:**			**b. Cumulative Hours of Work**		
RF: Judge Stringency	-376.0***	-125.3	-208.2*	-84.5	-21.9
All controls	(122.9)	(115.1)	(114.3)	(130.7)	(129.6)
IV: Incarcerated	-706.0***	-235.1	-390.9*	-158.6	-41.1
All controls	(237.3)	(216.1)	(220.4)	(242.6)	(242.6)
Dependent mean	738.2	747.1	763.2	771.1	784.5
***Dependent Variable*:**			**c. Cumulative Earnings (in 1000 NOK)**		
RF: Judge Stringency	-68.0**	-44.3	-37.3	-3.2	1.9
All controls	(31.4)	(30.9)	(31.1)	(33.8)	(36.2)
IV: Incarcerated	-127.7**	-83.2	-70.1	-6.1	3.5
All controls	(59.7)	(58.2)	(58.7)	(63.3)	(68.0)
Dependent mean	148.3	158.1	168.2	175.0	184.7
Number of cases			16,547		

Note: Baseline sample of non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1. RF and IV in addition also control for court x court entry year FEs. OLS standard errors are clustered at the defendant level, while RF and IV standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Table B14. Decomposing the Employment Effect of Incarceration into Employment at the Previous Firm versus a New Firm for the 'Previously Employed' Subsample.

		Change in Employer:	
		Months 1-60 after Decision	
	(1)	(2)	(3)
Dependent Variable:	*Pr(Ever Employed)*	*Pr(Employed at Previous Firm)*	*Pr(Employed at New Firm)*
RF: Judge Stringency	-0.156***	-0.133*	-0.023
All controls	(0.058)	(0.068)	(0.063)
IV: Incarcerated	-0.292***	-0.249*	-0.043
All controls	(0.114)	(0.129)	(0.119)
Dependent mean	0.70	0.39	0.31
Complier mean if not incarcerated	0.82	0.52	0.30
Number of cases		16,547	

Note: Sub-sample of previously employed defendants in non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1. RF and IV in addition also control for court x court entry year FEs. RF and IV standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Table B15. Decomposing the Effect of Incarceration on Reoffending, Future Employment and Job Training Program (JTP) Participation for the 'Previously Non-employed' Subsample.

		Future Employment: *Months 1-60 after Decision*		**Future Employment & JTP Participation:** *Months 1-60 after Decision*	
	(1)	(2)	(3)	(4)	(5)
Dependent Variable:	*Pr[Ever Charged]*	*Pr[Ever Charged ∩ Ever Employed]*	*Pr[Ever Charged ∩ Not Employed]*	*Pr[Ever Charged ∩ (Ever Employed ∪ On JTP)]*	*Pr[Ever Charged ∩ Not Employed ∩ Not on JTP]*
RF: Judge Stringency	-0.183***	0.024	-0.207***	0.021	-0.204***
All controls	(0.060)	(0.067)	(0.068)	(0.072)	(0.067)
IV: Incarcerated	-0.433**	0.056	-0.489**	0.050	-0.483**
All controls	(0.177)	(0.158)	(0.196)	(0.171)	(0.194)
Dependent mean	0.79	0.32	0.48	0.42	0.38
Complier mean if not incarcerated	0.96	0.13	0.83	0.13	0.83
Number of cases			14,881		

Note: Sub-sample of previously non-employed defendants in non-confession criminal cases processed 2005-2009. Controls include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Table B16. The Costs and Benefits of Incarceration (in 2012 USD).

	Direct Prison Costs	***Legal Costs***	***Net Transfers***
	Time Served in Prison for Current Sentence	***From Future Crimes***	***Future Cash Transfers Received - Taxes Paid***
	(1)	(2)	(3)
IV: Incarcerated	60 514.6***	-71 224.8**	-67 086.0
	(16,917.1)	(35,387.0)	(80,753.9)
Dependent Mean	23,919	63,318	31,024
Number of Observations		31,428	

Note: Baseline sample and specification with all controls. Standard errors are two-way clustered. Direct prison costs are set to USD 323.5 per day of incarceration, corresponding to the total prison spending reported by the Norwegian Correctional Services divided by the total number of prison days served across all prisoners in 2013. Legal costs include police and court processing costs, set to USD 3,670 per reported crime and USD 2,533 per court case, respectively, which correspond to the total operating costs reported by the Norwegian Police Service divided by the total number of crimes reported and the total operating costs reported by the Norwegian Courts divided by the total number of criminal court cases (and scaled by the fraction of criminal cases) processed in 2013. Net transfers include all cash transfers received subtracted all income taxes paid over the five year period after court case decision. All numbers are PPP adjusted and reported in 2012 USD. *p<0.1, **p<0.05, ***p<0.01.

Appendix C. Additional Results for the Sample of First-Time Offenders

Table C1. First Stage Estimates of Incarceration on Judge Stringency for First-Time Offenders.

	(1)	(2)	(3)	(4)	(5)	(6)
***Estimation Sample*:**	Time of Decision	Month 12 after Decision	Month 24 after Decision	Month 36 after Decision	Month 48 after Decision	Month 60 after Decision
***Dependent Variable*:**			**Pr(Incarcerated)**			
A. Court × Year of Court Case Registration Interacted Fixed Effects						
Judge Stringency	0.4399***	0.4383***	0.4466***	0.4346***	0.4242***	0.4191***
	(0.0819)	(0.0819)	(0.0828)	(0.0823)	(0.0825)	(0.0834)
F-stat. (Instrument)	28.08	27.90	28.34	27.12	25.73	24.60
B. Add Controls for Demographics and Type of Crime						
Judge Stringency	0.4416***	0.4403***	0.4459***	0.4363***	0.4267***	0.4182***
	(0.0792)	(0.0793)	(0.0802)	(0.0802)	(0.0802)	(0.0814)
F-stat. (Instrument)	30.22	30.02	30.08	28.82	27.50	25.68
C. Add Controls for Demographics, Type of Crime, Past Work and Criminal History						
Judge Stringency	0.4436***	0.4436***	0.4494***	0.4400***	0.4304***	0.4227***
	(0.0784)	(0.0784)	(0.0792)	(0.0793)	(0.0794)	(0.0804)
F-stat. (Instrument)	31.12	31.14	31.27	29.95	28.54	26.83
Dependent mean	0.3851	0.3851	0.3850	0.3838	0.3837	0.3835
Number of cases	15,603	15,505	15,323	15,143	14,982	14,797

Note: Sample of first time non-confession criminal cases with previously unincarcerated defendants processed 2005-2009. Standard errors are two-way clustered at judge and defendant level. **p<0.1, **p<0.05, ***p<0.01.

Table C2. The Effects of Incarceration on Recidivism for First-Time Offenders.

***Dependent Variable*:**	**Pr(Ever Charged)**			**Number of Charges**
	Months 1-24 after Decision	*Months 25-60 after Decision*	*Months 1-60 after Decision*	*Months 1-60 after Decision*
	(1)	**(2)**	**(3)**	**(4)**
OLS: Incarcerated	0.030***	0.023**	0.036***	0.131
No controls	(0.010)	(0.010)	(0.010)	(0.242)
OLS: Incarcerated	0.053***	0.0425***	0.055***	0.697***
Demographics & Type of Crime	(0.010)	(0.019)	(0.009)	(0.229)
OLS: Incarcerated	0.054***	0.043***	0.055***	0.761***
All controls	(0.009)	(0.009)	(0.009)	(0.227)
RF: Judge Stringency	-0.030	-0.169**	-0.180**	-2.780
All controls	(0.069)	(0.077)	(0.071)	(1.823)
IV: Incarcerated	-0.071	-0.399**	-0.427**	-6.577
All controls	(0.164)	(0.191)	(0.182)	(4.477)
Dependent mean	0.40	0.40	0.54	4.36
Number of cases			14,797	

Note: Sample of first time non-confession criminal cases with previously unincarcerated defendants processed 2005-2009. Controls include all variables listed in Table 1. RF and IV in addition also control for court x court entry year FEs. OLS standard errors are clustered at the defendant level, while RF and IV standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Table C3. The Effect of Incarceration on Recidivism and Future Employment by Previous Labor Market Attachment for First-Time Offenders.

	Sub-sample:	***Sub-sample:***
Dependent Variable:	**Previously Employed**	**Previously Non-employed**
A. Pr(Ever Charged)	**(1)**	**(2)**
Months 1-60 after Decision		
RF: Judge Stringency	-0.085	-0.257**
All controls	(0.085)	(0.111)
IV: Incarcerated	-0.258	-0.412**
All controls	(0.266)	(0.200)
Dependent mean	0.47	0.65
Dependent Variable:	**Previously Employed**	**Previously Non-employed**
B. Number of Charges	**(1)**	**(2)**
Months 1-60 after Decision		
RF: Judge Stringency	0.449	-6.695**
All controls	(1.731)	(3.156)
IV: Incarcerated	1.370	-10.719**
All controls	(5.303)	(5.463)
Dependent mean	2.88	6.54
Dependent Variable:	**Previously Employed**	**Previously Non-employed**
C. Pr(Ever Employed)	**(1)**	**(2)**
Months 1-60 after Decision		
RF: Judge Stringency	-0.182**	0.302**
All controls	(0.077)	(0.115)
IV: Incarcerated	-0.555*	0.483**
All controls	(0.297)	(0.208)
Dependent mean	0.79	0.55
Dependent Variable:	**Previously Employed**	**Previously Non-employed**
D. Cumulative Hours of work	**(1)**	**(2)**
Months 1-60 after Decision		
RF: Judge Stringency	-1025.2	2490.4***
All controls	(748.3)	(688.9)
IV: Incarcerated	-3129.4	3987.2***
All controls	(2414.5)	(1323.6)
Dependent mean	4821.7	2151.7
Dependent Variable:	**Previously Employed**	**Previously Non-employed**
E. Cumulative Earnings	**(1)**	**(2)**
Months 1-60 after Decision		
RF: Judge Stringency	-60.8	402.6***
All controls	(234.3)	(142.7)
IV: Incarcerated	-185.5	644.5**
All controls	(713.3)	(259.6)
Dependent mean	1090.1	364.6
Number of cases	8,828	5,969

Note: Sample of first time non-confession criminal cases with previously unincarcerated defendants processed 2005-2009. Controls include all variables listed in Table 1, plus controls for court x court entry year FEs. Standard errors are two-way clustered at judge and defendant level. *p<0.1, **p<0.05, ***p<0.01.

Appendix D. Monte Carlo Simulations for the Random Judge Design

Notation, Assumptions, and Parameters

Consider the problem of using random assignment of cases to judges to draw inference about the causal effects of incarceration. For simplicity, we abstract from covariates and assume that each of the J judges are unconditionally randomly assigned to cases. Unless necessary, we also suppress the subscript for defendants. Let the observed outcome be denoted Y and let $Y(1)$ and $Y(0)$ represent the realization of Y that would have been experienced by an individual had their incarceration decision I been exogenously set to 0 or 1. For any given defendant, the relationship between observed outcome Y and potential outcomes is given by:

$$Y = I \times Y(1) + (1 - I) \times Y(0)$$

while the relationship between the observed incarceration decision and the potential incarceration decision is given by:

$$I = \sum_{j=1}^{J} \mathbf{1}\{Z = j\} \times I(j)$$

where judge assignment is represented by the discrete variable $Z \in \mathbb{Z} = \{1, ..., J\}$ and $I(j)$ denotes the potential incarceration value for each judge j.

Without loss of generality,

$$I(j) = \mathbf{1}\{V(j) \leq \tau_j\}, \quad \text{for } j \in \mathbb{Z}, \text{ where } V(j) \sim U[0, 1] \tag{D1}$$

Here, $V(j)$ can be viewed as the rank under assignment to judge $Z = j$. Assume instrument exogeneity, implying that $Y(1), Y(0), V(1), ..., V(J)$ are independent of Z. For each $j \in \mathbb{Z}$, the propensity score τ_j can then be identified by:

$$\tau_j = Prob(I = 1|Z = j)$$

Without loss of generality, suppose Z is ordered such that $\tau_1 \geq ... \geq \tau_J$. Assume rank invariance, i.e. $V(1) = V(2) = ... = V(J) = V$, or, equivalently, instrument monotonicity, i.e. $I(1) \leq I(2) \leq ... \leq I(J)$. Given instrument exogeneity and monotonicity, we get the usual selection equation,

$$I = \mathbf{1}\{V \geq g(Z)\}$$

where $g : \mathbb{Z} \to [0, 1]$ is such that $g(Z) = \tau_j$.

In the paper, we consider several parameters of interest. One is the 2SLS estimand:

$$\beta_{2SLS} \quad = \quad cov(Y, g(Z))/cov(I, g(Z)) = \sum_{j=2}^{J} w_j E[Y(1) - Y(0)|\tau_j \leq V < \tau_{j+1}]$$

which is a positively weighted average of the LATEs among defendants who would be incarcerated by judge $j+1$ but not by judge j (Imbens and Angrist, 1994). More generally, we are interested in the MTE, $E[Y(1) - Y(0)|V = v]$, and functions thereof, such as the ATE, ATT, ATUT, and LATE.

Issues of Estimation and Statistical Inference

Given the data $(Y, I, Z) : 1 \leq i \leq n$, identification of the parameters of interest follows from standard arguments. Using this data, we can first estimate τ_j and order the judges by their estimated propensity score. Next, we can estimate β_{2SLS} (or other parameters of interest) by the sample analogue. However, the need to estimate τ_j introduces two problems related to inference.

The *first problem* is that the number of cases per judge could be small. As a result, the estimated propensity score $\hat{\tau}_j$ may be a poor estimate for the true, latent propensity score τ_j. This will affect inference as we may get a poor estimate of the ranking of the instrument by the propensity score and even if we get the rankings correct, we may get a poor estimate of β_{2SLS} since the number of effective observations used in its estimation is small. In our simulations, we will capture this problem by examining how the estimators of β_{2SLS} perform with few versus many cases per judge.

The *second problem* is that the instrument may be weak. For example, suppose $J = 2$. Then, having a weak instrument means the difference between τ_1 and τ_2 is sufficiently small (where "small" depends on the underlying data generating process and on the sample size). This will affect inference as the estimated denominator of β_{2SLS} is close to zero. In particular, with a weak instrument 2SLS is biased towards OLS, and 2SLS tests have the wrong size. In our simulations, we will capture this problem by examining how the estimators of β_{2SLS} perform if we make the propensity scores more similar.

Of course, while these two problems are conceptually distinct, they could arise together in practice, which would further complicates the inference problem. Our simulations will therefore consider the combination of few cases per judge and a weak instrument.

Before proceeding, it is necessary to decide on how to estimate the propensity scores. We consider three estimators, all of which consistently estimate τ_j as the number of cases per judge tends to infinity. However, their performance may differ in finite sample. The

first estimator of τ_j that we consider is the sample average incarceration rate among cases assigned to judge j:

$$\hat{\tau}_j^{mean} = \frac{\sum_{i=1}^n \mathbf{1}\{Z_i = j\} I_i}{\sum_{i=1}^n \mathbf{1}\{Z_i = j\}}$$

The 2SLS estimator using $\hat{\tau}_j^{mean}$ as the instrument for I_i is numerically equivalent to the 2SLS estimator using a full of set judge dummy variables as instruments for I_i.

One drawback with these estimators is the mechanical relationship between the defendant's own case outcome and the sample average incarceration rate of the defendant's judge. As a result, any correlation between the defendant's own case outcome and unobservables in the second stage will (with a finite number of cases per judge) lead to *some* correlation between the sample average incarceration rate of the defendant's judge and those same unobservables. To address this issue, many researchers have opted to estimate the propensity scores as the leave out mean incarceration rate:

$$\hat{\tau}_{j,-i}^{leaveout} = \frac{\sum_{k \neq i} \mathbf{1}\{Z_k = j\} I_k}{\sum_{k \neq i} \mathbf{1}\{Z_k = j\}}$$

The 2SLS estimator using $\hat{\tau}_{j,-i}^{leaveout}$as an instrument for I_i is numerically equivalent to a jackknife IV estimator (JIVE) using a full set of judge dummy variables as instruments for I_i.

As an alternative to using $\hat{\tau}_{j,-i}^{leaveout}$ as the estimator of τ_j, one may use a split sample approach. This can be done by randomly splitting the sample in half and using the first half of the sample to estimate the average incarceration rate among cases assigned to a given judge within this sample,

$$\hat{\tau}_j^{split} = \frac{\sum_{i=1}^n \mathbf{1}\{S_i = 1\}\mathbf{1}\{Z_i = j\} I_i}{\sum_{i=1}^n \mathbf{1}\{S_i = 1\}\mathbf{1}\{Z_i = j\}}$$

where S_i is an indicator variable for being assigned to the first half of the sample. Next, the estimated propensity scores $\hat{\tau}_j^{split}$ can be used an instrument for I_i in 2SLS estimation based on the other half of the sample. While sample splitting avoids any dependence between the defendant's own case and the estimated propensity scores of her judge, it comes at the cost of smaller sample size and less power.

Data Generating Process

In the tables below, we consider how various estimators of β_{2SLS} depend on the number of judges and cases, the differences across the propensity scores, and the estimators of the propensity scores. Our primary goal is to examine second stage performance, which we do by comparing the bias, test size, and precision of the estimated values for β_{2SLS} in finite samples. A secondary goal is to examine whether a test for a weak instrument can provide an indicator for when the various estimators of β_{2SLS} perform well.

The simulations are based on the following data generating process. The propensity score associated with judge j is generated according to a truncated normal distribution centered at 0.5 and in the interval $[0, 1]$:

$$\tau_j \sim N(0.5, \sigma_z^2), \quad \text{if } \tau_j \notin [0,1], \text{ redraw } \tau_j$$

We randomly assign each judge to n/J defendants. Consistent with the assumptions of instrument exogeneity and monotonicity, the defendant incarceration outcome is defined as:

$$I = \mathbf{1}\{V \leq \tau_j\} \quad \text{where } V \sim U[0,1]$$

The outcome depends on the incarceration decision and a composite error term:

$$Y = \beta I + .5V + \eta \quad \text{where } \eta \sim N(0,1)$$

We set the true treatment effect to be $\beta = 1$. The appearance of V in the outcome equation creates a correlation between the composite error term and treatment, which biases the OLS estimator of the treatment effect β.

With this structure, we implement simulations that vary the number of cases per judge (20, 50, 100, 250, or 500), the number of judges (20, 50, 100, 250, or 500), the variance across judges' propensity scores (none: σ_z^2=0, small: σ_z^2=.0025, and large: σ_z^2=.01), and estimators of the propensity scores ($\hat{\tau}_j^{mean}$, $\hat{\tau}_{j,-i}^{leaveout}$, and $\hat{\tau}_j^{split}$). As a benchmark, we compare the peformance of these estimators to the case where the latent τ_j is known. We take 2,000 draws for each data generating process, and estimate both the first and second stages.

Simulation Results

We begin by examining the performance of the various approaches for the second stage estimate of β. Appendix Tables D1-D3 report the bias, the standard deviation, and the test size. To understand the general structure of the tables, begin with the upper left panel of Appendix Table D1. That panel considers the case where there are 20 judges and the variance in judge's propensity scores is large (σ_z^2=.01), with the rows varying the number of cases per judge from 20 to 500. In each column of the panel, a different estimator for the propensity score is used. The first column uses the latent propensity scores, and columns 2 through 4 use the mean, leave-out, and split-sample estimators of the propensity scores. The other panels in the table have a similar structure, but vary the number of judges and the variance in the propensity scores. In the tables, entries are bolded if the median first stage robust F-statistic across simulations (see Appendix Table D5) exceeds the critical value proposed by Montiel Olea and Pflueger (2013) as a test for a weak instrument.[42] We return

[42]Note that the number of instruments is determined by the number of moment conditions (and not the number of values the instrument takes). Even though there are many judges, our 2SLS model has one moment

to the usefulness of this test statistic after our discussion of second stage performance in Appendix Tables D1-D3.

Appendix Table D1 reports the estimated bias based on 2,000 Monte Carlo draws. Consider first the simulations where the variation in judge propensity scores is large (σ_z^2=.01), which corresponds to the top panels of the table. As expected, there is almost no bias when the latent propensity scores are known, regardless of the number of judges or cases per judge. The mean estimator exhibits a moderate amount of bias when there are few cases per judge, but this bias shrinks as the number of cases increases. The leave-out and split-sample estimators show even less bias, with trivial amounts of bias even with a moderate number of cases per judge. Turning to the middle panels, we explore what happens when the variation in propensity scores is small (σ_z^2=.025). In this case, the bias using the mean is generally larger, and while the bias shrinks as the number of cases rises, it is not completely eliminated even when there are 500 cases per judge. The leave-out and split-sample estimators generally perform better than the mean estimator in terms of bias, but compared to the top panel, more cases per judge are needed to eliminate bias. It is also interesting to note that for the leave-out and split-sample estimators, holding the number of cases per judge fixed, the bias decreases as the number of judges increases. The bottom panels consider the case when there is no variation across judges in their propensity scores. In this case, the mechanical bias from the mean estimator dominates, with a bias of around 0.25 regardless of the number of cases per judge or the number of judges. The leave-out and split-sample estimators do not perform well in the bottom panels.

Of course, the mean bias needs to be considered in concert with the variation in the second stage estimators and test size. In Appendix Table D2, we examine the standard deviation of the second stage estimates across the 2,000 Monte Carlo draws. The standard deviations for the case when the latent judge propensity scores are known serve as a useful benchmark for the amount of variability in the DGP. Relative to this benchmark, the mean estimator understates statistical uncertainty, particularly when the number of cases per judge is small. In contrast, the leave-out estimator has a larger standard deviation compared to the latent model when the number of cases per judge is small, as it should since estimating the instrument introduces sampling uncertainty. The split-sample estimator has an even larger standard deviation, reflecting the fact that it uses a fraction of the data to estimate the propensity scores. Comparing results with different amounts of variation in judge propensity scores, the standard deviation goes up when there is less variation across judges. As the number of judges increases, the standard deviation generally falls. When there is no variation

condition, and therefore, a single instrument. Thus, the appropriate critical value should reflect that there is one and not multiple instruments.

across judges (the bottom panels), both the leave-out and split-sample estimators have extremely large standard deviations.

Appendix Tables D3 and D4 report second stage test size, calculated as the fraction of simulations in which the true null of $\beta = 1$ is rejected at the 95% confidence level. If the estimators are performing well, this fraction should be close to 0.05. In Appendix Table D3, we report test size based on the t-statistic using the estimated β and its associated standard error. Appendix Table D3 documents that for the mean estimator, size gets worse as the number of judges increases. While size is better controlled as the number of cases per judge increases in the top and middle panels, even with 500 cases per judge the mean estimator is distorted. In contrast, the leave-out estimator does a good job in terms of size when the variation across judges is large, as well as when the variation across judges is small but the number of cases per judges is large. In the other cases in the table, the leave-out estimator understates size, which results in overly conservative hypothesis tests. The split-sample estimator shows similar patterns as the leave-out estimator, but requires more cases per judge before size is close to the target of 0.05. In Appendix Table D4, we report test size based on inverting the Anderson-Rubin (1949) test statistic, an approach which is robust to the presence of a weak instrument. The mean estimator continues to do poorly, with little change in size compared to Appendix Table D3. In contrast, size for both the leave-out and split-sample estimators is close to the target, even when the variation across judges is weak or non-existent. In particular, the size is no longer overly conservative when there are few cases per judge or little to no variation in judge propensity scores.

We next assess whether first stage F tests can provide a useful diagnostic for second stage performance. Staiger and Stock (1997) and Stock and Yogo (2005) ran simulation studies based on conditional homoskedasticity, and recommend using a homoskedastic F-statistic of roughly 10 or smaller as an indicator for problems due to a weak instrument. Building on this work, Montiel Olea and Pflueger (MOP, 2013) propose a conservative test which is robust to the worst type of heteroscedasticity, serial correlation, and clustering in the second stage. In the case of a single instrument, they suggest using the robust F-statistic (which allows for heteroskedasticity) with an adjusted critical value of 23.11.[43] In Appendix Table D5 we report the median of the robust first stage F-statistics across simulations. The results in this table are used to bold the entries in the other tables, based on whether the median robust F-statistic exceeds 23.11. Looking back at the bolded entries in Appendix Tables D1 and D3, we conclude the MOP test does a reasonably good job of identifying cases where the

[43]Their recommendation is based on testing the null that the Nagar bias exceeds 10% of a "worst-case" bias with a size of 5%. As they recognize, their test is overly conservative if heteroscedasticity, serial correlation, and clustering are not present, particularly when the number of instruments is small.

second stage will perform well in terms of bias and have the correct size for the leave-out and split-sample estimators. It correctly flags cases which exhibit bias and have the wrong size, which occurs more often when the variation in propensity scores is small or nonexistent and when the number of cases per judge is small. In contrast, for the mean estimator the MOP diagnostic does a poor job, especially as the number of judges increases. With the mean estimator, the bias can be large and the size grossly overstated, even when the robust F-statistic is much larger than the MOP rule of thumb.

We conclude from our simulations that the mean estimator is a poor choice for a random judge design. Finite sample bias is substantial, size is distorted, and MOP tests do not do a good job at identifying when problems arise. In contrast, the leave-out and split-sample estimators both perform well when the number of cases per judge is large enough, and MOP tests are reasonable indicators for problems in the second stage estimates. The cost of the split-sample estimator is the reduction in sample size, and as a consequence is more often flagged as having problems in the second stage, especially when the variation across judges is small. We also conclude that if one is worried about a weak instrument creating an inference problem, the use of AR confidence intervals are a good idea as a robustness check. Of course, this is just one set of simulations, and it is important to recognize that different data generating processes could yield different insights.

Table D1: 2nd Stage Point Estimate Bias

σ^2_τ	No. Cases Per Judge	20 Judges Latent	20 Judges Mean	20 Judges Leave-out	20 Judges Split	50 Judges Latent	50 Judges Mean	50 Judges Leave-out	50 Judges Split	100 Judges Latent	100 Judges Mean	100 Judges Leave-out	100 Judges Split	250 Judges Latent	250 Judges Mean	250 Judges Leave-out	250 Judges Split	500 Judges Latent	500 Judges Mean	500 Judges Leave-out	500 Judges Split
.01	20	-.03	**.131**	-.43	1174	**-.009**	**.137**	-.011	-.014	**.002**	**.135**	**-.029**	-.037	**.005**	**.138**	**-.004**	**.043**	**.001**	**.135**	**-.007**	**.047**
	50	**-.005**	**.078**	**.015**	.058	**-.011**	**.075**	**-.022**	-.042	**-.005**	**.079**	**-.008**	**-.006**	**-.003**	**.081**	**-.003**	**.011**	**.001**	**.082**	**0**	**.018**
	100	**-.001**	**.055**	**-.005**	.057	**.003**	**.05**	**-.003**	**.006**	**-.001**	**.049**	**-.002**	**.002**	**-.001**	**.049**	**0**	**0**	**-.001**	**.047**	**-.002**	**.001**
	250	**-.002**	**.021**	**-.006**	**-.013**	**.001**	**.023**	**0**	**-.009**	**-.003**	**.019**	**-.004**	**-.001**	**0**	**.022**	**0**	**0**	**-.001**	**.022**	**-.001**	**.002**
	500	**.001**	**.012**	**-.001**	**-.003**	**-.002**	**.009**	**-.003**	**.001**	**0**	**.011**	**-.001**	**-.001**	**-.001**	**.011**	**-.001**	**0**	**-.001**	**.011**	**-.001**	**0**
.0025	20	.1	**.201**	-5827	-65	-.051	**.221**	.195	122	-.01	**.214**	-.265	-.382	**-.014**	**.211**	-5.75	-1.6	**-.008**	**.207**	-.38	-.047
	50	.014	**.155**	.046	1.27	**.037**	**.178**	-.217	2.38	**-.004**	**.164**	-.43	-.273	**.001**	**.169**	**-.006**	.068	**-.002**	**.164**	**-.016**	.023
	100	.003	**.139**	-.066	2.52	**.005**	**.132**	.177	-.042	**.005**	**.124**	**-.016**	.042	**-.004**	**.122**	**-.01**	.01	**-.003**	**.124**	**-.004**	**.005**
	250	**-.01**	**.075**	**.082**	-.587	**-.004**	**.073**	**-.006**	-1.63	**-.001**	**.071**	**-.004**	.004	**.002**	**.072**	**0**	**.001**	**-.002**	**.07**	**-.002**	**-.001**
	500	**-.011**	**.034**	**-.019**	.517	**-.002**	**.04**	**-.005**	**-.004**	**-.004**	**.039**	**-.005**	**-.011**	**0**	**.041**	**-.002**	**-.001**	**.001**	**.042**	**.001**	**.006**
0	20	-	.26	10979	-149	-	**.252**	3058	1.62	-	**.247**	-.286	-.019	-	**.249**	1.5	1	-	**.25**	-.157	1.07
	50	-	.254	-.566	-.323	-	**.262**	.084	-.993	-	**.251**	3.29	4.21	-	**.248**	-.132	2.49	-	**.252**	-.608	-9.27
	100	-	.26	-.948	-.436	-	**.255**	-1.76	.231	-	**.251**	3.74	-3.12	-	**.251**	.956	-1.61	-	**.252**	4.6	-1.64
	250	-	.241	.453	7.33	-	**.252**	-.549	-.342	-	**.249**	3.81	-2.14	-	**.256**	-1.71	-4.81	-	**.247**	1.35	-2
	500	-	.242	.379	2.46	-	**.253**	3.01	1.03	-	**.253**	2.52	5.83	-	**.247**	-5.26	-1.81	-	**.248**	-1.53	.335

Note: Cells report the second stage bias, defined as the average of $\hat{\beta} - \beta$ across simulations, when using the latent or the mean, leave-out, or split-sample estimators of the propensity scores. Entries are bold when the median heteroskedasticity robust F-statistic from the first stage exceeds 23.11 (see Table D5 and its notes). All cells are based on results from 2,000 Monte Carlo draws.

Table D2: 2nd Stage Point Estimate Standard Deviation

σ^2_τ	No. Cases Per Judge	20 Judges Latent	20 Judges Mean	20 Judges Leave-out	20 Judges Split	50 Judges Latent	50 Judges Mean	50 Judges Leave-out	50 Judges Split	100 Judges Latent	100 Judges Mean	100 Judges Leave-out	100 Judges Split	250 Judges Latent	250 Judges Mean	250 Judges Leave-out	250 Judges Split	500 Judges Latent	500 Judges Mean	500 Judges Leave-out	500 Judges Split
.01	20	.712	**.36**	13.9	49172	**.338**	**.218**	1.28	3.91	**.239**	**.155**	**.451**	1.4	**.15**	**.097**	**.228**	**.445**	**.102**	**.068**	**.159**	**.296**
	50	**.36**	**.277**	**3.53**	11.9	**.214**	**.171**	**.28**	.786	**.145**	**.118**	**.182**	**.416**	**.092**	**.075**	**.113**	**.247**	**.064**	**.053**	**.079**	**.178**
	100	**.249**	**.22**	**.328**	10.3	**.146**	**.132**	**.168**	**.396**	**.102**	**.092**	**.116**	**.256**	**.064**	**.057**	**.071**	**.156**	**.046**	**.041**	**.051**	**.112**
	250	**.156**	**.147**	**.167**	**.383**	**.093**	**.089**	**.099**	**.21**	**.066**	**.063**	**.07**	**.141**	**.042**	**.04**	**.043**	**.09**	**.028**	**.027**	**.03**	**.061**
	500	**.108**	**.105**	**.112**	**.26**	**.066**	**.064**	**.068**	**.13**	**.046**	**.044**	**.047**	**.093**	**.028**	**.028**	**.029**	**.055**	**.02**	**.019**	**.02**	**.039**
.0025	20	38.1	**.455**	260633	54743	1.17	**.265**	28.2	5478	.53	**.19**	30.7	31	**.299**	**.117**	262	70	**.205**	**.082**	19.4	5.73
	50	2.7	**.392**	15.1	50.8	**.743**	**.247**	12.2	117	**.303**	**.166**	12.2	13.3	**.178**	**.103**	**.35**	1.81	**.127**	**.074**	**.237**	.609
	100	.568	**.338**	10.4	107	**.308**	**.215**	8.79	5.67	**.21**	**.145**	**.327**	3.98	**.128**	**.089**	**.186**	.749	**.091**	**.065**	**.133**	**.381**
	250	**.334**	**.26**	**5.09**	22.1	**.185**	**.155**	**.23**	77.8	**.128**	**.109**	**.158**	.558	**.081**	**.067**	**.095**	**.292**	**.057**	**.048**	**.068**	**.205**
	500	**.225**	**.203**	**.322**	27.5	**.133**	**.12**	**.148**	**.69**	**.093**	**.086**	**.105**	**.3**	**.057**	**.052**	**.063**	**.174**	**.04**	**.037**	**.045**	**.123**
0	20	-	.489	865313	69842	-	**.294**	136815	89	-	**.207**	28.2	39.5	-	**.129**	80.4	52.4	-	**.089**	83.1	34.4
	50	-	.474	71.7	42.2	-	**.289**	34.9	38.3	-	**.207**	143	114	-	**.125**	75.3	471	-	**.088**	57.9	603
	100	-	.491	82.3	44.5	-	**.292**	78.3	63.3	-	**.208**	128	95.7	-	**.129**	62.5	117	-	**.087**	168	59.1
	250	-	.481	84.5	367	-	**.29**	27.3	167	-	**.204**	106	180	-	**.128**	140	186	-	**.09**	138	82.3
	500	-	.475	41.1	116	-	**.287**	92.2	105	-	**.203**	153	182	-	**.127**	179	73.4	-	**.09**	70.4	57.5

Note: Cells report the standard deviation across simulations in the point estimate of β, when using the latent or the mean, leave-out, or split-sample estimators of the propensity scores. Entries are bold when the median heteroskedasticity robust F-statistic from the first stage exceeds 23.11 (see Table D5 and its notes). All cells are based on results from 2,000 Monte Carlo draws.

Table D3: 2nd Stage Test Size

σ_τ^2	No. Cases Per Judge	20 Judges Latent	20 Judges Mean	20 Judges Leave-out	20 Judges Split	50 Judges Latent	50 Judges Mean	50 Judges Leave-out	50 Judges Split	100 Judges Latent	100 Judges Mean	100 Judges Leave-out	100 Judges Split	250 Judges Latent	250 Judges Mean	250 Judges Leave-out	250 Judges Split	500 Judges Latent	500 Judges Mean	500 Judges Leave-out	500 Judges Split
.01	20	.023	**.059**	.009	.004	**.04**	**.089**	.02	.005	**.047**	**.146**	**.037**	.015	**.058**	**.308**	**.051**	**.034**	**.049**	**.509**	**.049**	**.047**
	50	**.043**	**.057**	**.031**	.009	**.051**	**.075**	**.04**	.015	**.039**	**.101**	**.045**	**.028**	**.056**	**.194**	**.047**	**.043**	**.051**	**.347**	**.052**	**.051**
	100	**.043**	**.054**	**.043**	.02	**.046**	**.068**	**.048**	**.037**	**.041**	**.091**	**.048**	**.037**	**.045**	**.135**	**.044**	**.046**	**.061**	**.219**	**.058**	**.051**
	250	**.051**	**.058**	**.05**	**.04**	**.052**	**.054**	**.045**	**.047**	**.051**	**.068**	**.054**	**.047**	**.054**	**.086**	**.056**	**.058**	**.045**	**.13**	**.049**	**.043**
	500	**.048**	**.05**	**.047**	**.051**	**.051**	**.051**	**.051**	**.047**	**.047**	**.052**	**.05**	**.051**	**.046**	**.064**	**.049**	**.043**	**.041**	**.084**	**.043**	**.05**
.0025	20	.003	**.06**	.005	.001	.019	**.122**	.004	.001	.026	**.221**	.006	.002	**.041**	**.437**	.013	.006	**.043**	**.71**	.026	.018
	50	.016	**.051**	.004	.001	**.031**	**.128**	.014	.002	**.04**	**.163**	.02	.002	**.038**	**.369**	**.039**	.012	**.047**	**.6**	**.043**	.021
	100	.03	**.063**	.014	.003	**.041**	**.105**	.031	.002	**.047**	**.14**	**.033**	.006	**.045**	**.275**	**.044**	.02	**.051**	**.491**	**.056**	**.034**
	250	**.045**	**.063**	**.035**	.006	**.038**	**.072**	**.045**	.021	**.047**	**.105**	**.051**	.027	**.051**	**.18**	**.043**	**.038**	**.047**	**.306**	**.051**	**.054**
	500	**.038**	**.06**	**.049**	.026	**.052**	**.065**	**.049**	**.027**	**.053**	**.085**	**.056**	**.036**	**.049**	**.115**	**.053**	**.047**	**.045**	**.216**	**.049**	**.053**
0	20	-	.073	.007	.001	-	**.141**	.002	0	-	**.227**	.004	.002	-	**.493**	.003	.001	-	**.791**	.001	.002
	50	-	.052	.004	.001	-	**.147**	.004	.001	-	**.238**	.003	0	-	**.499**	.003	.001	-	**.797**	.002	0
	100	-	.075	.003	.001	-	**.133**	.002	0	-	**.24**	.002	0	-	**.506**	.004	0	-	**.817**	.002	.001
	250	-	.061	.003	0	-	**.123**	.002	0	-	**.227**	.004	0	-	**.521**	.002	0	-	**.78**	.004	0
	500	-	.06	.002	0	-	**.136**	.001	0	-	**.241**	.002	0	-	**.489**	.002	0	-	**.794**	.003	0

Note: Cells report the share of simulations in which the null ($\beta = 1$) is rejected at the 95% confidence level based on the t-statistic, when using the latent or the mean, leave-out, or split-sample estimators of the propensity scores. Entries are bold when the median heteroskedasticity robust F-statistic from the first stage exceeds 23.11 (see Table D5 and its notes). All cells are based on results from 2,000 Monte Carlo draws.

Table D4: 2nd Stage Weak Instrument Robust Test Size

σ_τ^2	No. Cases Per Judge	20 Judges Latent	20 Judges Mean	20 Judges Leave-out	20 Judges Split	50 Judges Latent	50 Judges Mean	50 Judges Leave-out	50 Judges Split	100 Judges Latent	100 Judges Mean	100 Judges Leave-out	100 Judges Split	250 Judges Latent	250 Judges Mean	250 Judges Leave-out	250 Judges Split	500 Judges Latent	500 Judges Mean	500 Judges Leave-out	500 Judges Split
.01	20	.0495	**.0645**	.051	.064	**.0475**	**.088**	.0435	.042	**.052**	**.144**	**.0495**	.0425	**.0585**	**.303**	**.0575**	**.0485**	**.0505**	**.504**	**.053**	**.0535**
	50	**.05**	**.062**	**.0515**	.0515	**.0545**	**.0755**	**.0535**	.0415	**.0395**	**.099**	**.0495**	**.047**	**.0565**	**.191**	**.049**	**.0475**	**.0505**	**.344**	**.0535**	**.052**
	100	**.049**	**.053**	**.053**	.0465	**.049**	**.0635**	**.0515**	**.052**	**.0425**	**.0895**	**.047**	**.044**	**.0445**	**.132**	**.0445**	**.0495**	**.06**	**.217**	**.0575**	**.053**
	250	**.0545**	**.06**	**.052**	**.0485**	**.053**	**.055**	**.0465**	**.0525**	**.0515**	**.067**	**.0545**	**.0505**	**.0545**	**.0855**	**.057**	**.0595**	**.045**	**.13**	**.0515**	**.0425**
	500	**.05**	**.0505**	**.0485**	**.057**	**.052**	**.0515**	**.051**	**.0505**	**.047**	**.051**	**.05**	**.052**	**.045**	**.064**	**.049**	**.043**	**.041**	**.0835**	**.0435**	**.0505**
.0025	20	.047	**.0745**	.055	.052	.0565	**.123**	.053	.0575	.045	**.221**	.053	.06	**.0535**	**.433**	.0415	.048	**.046**	**.706**	.0515	.0555
	50	.0465	**.0645**	.052	.056	**.045**	**.128**	.059	.052	**.0465**	**.159**	.0525	.0525	**.042**	**.364**	**.047**	.0485	**.05**	**.596**	**.052**	.05
	100	.0515	**.066**	.044	.0535	**.0535**	**.105**	.059	.0585	**.0485**	**.138**	**.048**	.0505	**.045**	**.271**	**.0465**	.0405	**.0525**	**.486**	**.0615**	**.047**
	250	**.054**	**.065**	**.0485**	.0465	**.0445**	**.072**	**.0525**	.052	**.048**	**.102**	**.053**	.0475	**.051**	**.178**	**.0455**	**.0475**	**.047**	**.302**	**.0515**	**.0575**
	500	**.0435**	**.0625**	**.058**	.0605	**.0535**	**.0645**	**.051**	**.051**	**.0545**	**.0835**	**.058**	**.046**	**.05**	**.113**	**.054**	**.05**	**.045**	**.214**	**.0485**	**.056**
0	20	-	.091	.0535	.0605	-	**.144**	.0485	.0595	-	**.225**	.0545	.0485	-	**.488**	.0525	.0485	-	**.789**	.0495	.0565
	50	-	.0745	.0475	.0555	-	**.152**	.0465	.049	-	**.236**	.052	.055	-	**.493**	.0475	.056	-	**.791**	.043	.052
	100	-	.088	.051	.0525	-	**.138**	.051	.0505	-	**.236**	.0585	.0505	-	**.501**	.057	.0515	-	**.813**	.046	.0545
	250	-	.0795	.053	.054	-	**.13**	.052	.0555	-	**.224**	.054	.043	-	**.513**	.0495	.0535	-	**.777**	.051	.0485
	500	-	.078	.048	.0455	-	**.14**	.0465	.0565	-	**.235**	.055	.041	-	**.481**	.052	.0515	-	**.789**	.053	.0545

Note: Cells report the share of simulations in which the null ($\beta = 1$) is rejected at the 95% confidence level based on inverting the Anderson Rubin confidence interval, when using the latent or the mean, leave-out, or split-sample estimators of the propensity scores. Entries are bold when the median heteroskedasticity robust F-statistic from the first stage exceeds 23.11 (see Table D5 and its notes). All cells are based on results from 2,000 Monte Carlo draws.

Table D5: 1st Stage Median Test Statistics

σ_τ^2	No. Cases Per Judge	20 Judges Latent	20 Judges Mean	20 Judges Leave-out	20 Judges Split	50 Judges Latent	50 Judges Mean	50 Judges Leave-out	50 Judges Split	100 Judges Latent	100 Judges Mean	100 Judges Leave-out	100 Judges Split	250 Judges Latent	250 Judges Mean	250 Judges Leave-out	250 Judges Split	500 Judges Latent	500 Judges Mean	500 Judges Leave-out	500 Judges Split
.01	20	17.1	**41**	5.41	1.83	**43.4**	**110**	16.1	4.98	**86.5**	**225**	**34.3**	9.97	**227**	**576**	**91.3**	**26.4**	**453**	**1156**	**183**	**52.7**
	50	**40.8**	**63.6**	**23.7**	5.41	**109**	**170**	**67**	13.6	**223**	**350**	**142**	**29.3**	**565**	**883**	**360**	**72.9**	**1128**	**1771**	**724**	**145**
	100	**82.2**	**105**	**62.2**	13.2	**217**	**276**	**167**	**33.6**	**447**	**568**	**346**	**68.1**	**1129**	**1439**	**882**	**174**	**2255**	**2873**	**1760**	**343**
	250	**205**	**226**	**181**	**46.4**	**546**	**605**	**489**	**109**	**1116**	**1238**	**1004**	**216**	**2817**	**3122**	**2536**	**537**	**5656**	**6272**	**5097**	**1074**
	500	**410**	**432**	**386**	**117**	**1095**	**1153**	**1035**	**262**	**2233**	**2349**	**2111**	**516**	**5636**	**5942**	**5345**	**1254**	**11293**	**11896**	**10700**	**2494**
.0025	20	3.76	**25.4**	1.08	.539	9.8	**69**	1.75	.907	20.6	**140**	2.82	1.54	**51.4**	**356**	7.48	3.98	**102**	**713**	15.4	7.95
	50	9.75	**29.5**	2.4	.672	**25.1**	**78.9**	7.16	1.33	**50.6**	**160**	15.3	2.74	**127**	**405**	**40.2**	7.03	**257**	**811**	**81.4**	14
	100	19.3	**39.1**	8.11	1.34	**49.5**	**102**	22.8	2.88	**102**	**207**	**48.1**	6.34	**257**	**525**	**125**	15.6	**516**	**1056**	**254**	**31.9**
	250	**47.1**	**66.3**	**30.9**	4.92	**124**	**175**	**85.3**	10.7	**253**	**356**	**176**	21.4	**643**	**908**	**456**	**53.6**	**1289**	**1818**	**912**	**106**
	500	**93.4**	**113**	**75.1**	16.3	**249**	**300**	**204**	**29.5**	**509**	**610**	**417**	**56.6**	**1272**	**1535**	**1053**	**138**	**2580**	**3105**	**2139**	**273**
0	20	-	20.6	1.08	.39	-	**56.2**	1.02	.454	-	**115**	1.01	.514	-	**290**	.899	.93	-	**582**	.873	1.57
	50	-	19.4	1.07	.362	-	**51.1**	1.05	.347	-	**104**	.879	.363	-	**264**	.85	.443	-	**529**	.918	.564
	100	-	18.6	.987	.327	-	**49.9**	.875	.294	-	**102**	1.02	.296	-	**255**	.947	.381	-	**513**	.93	.347
	250	-	18.5	.966	.299	-	**48.4**	.988	.287	-	**98.8**	.966	.277	-	**251**	.851	.282	-	**506**	.93	.322
	500	-	18.3	1.09	.349	-	**48.6**	.964	.293	-	**99.2**	.906	.267	-	**250**	.925	.26	-	**501**	.935	.274

Note: Cells report the median heteroskedasticity robust F-statistic across simulations for the first stage, when using the latent or the mean, leave-out, or split-sample estimators of the propensity scores. Entries are bold when the median heteroskedasticity robust F-statistic from the first stage exceeds 23.11. This is the conservative critical value recommended in Montiel Olea and Pflueger (2013) as a test for weak instruments. All cells are based on results from 2,000 Monte Carlo draws.

References

Anderson, T. W. and H. Rubin (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics 20*(1), 46–63.

Imbens, G. W. and J. D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–475.

Montiel Olea, J. L. and C. Pflueger (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics 31*(3), 358–369.

Staiger, D. and J. H. Stock (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica 65*(3), 557–586.

Stock, J. and M. Yogo (2005). *Testing for Weak Instruments in Linear IV Regression*, pp. 80–108. in: Identification and Inference for Econometric Models (edited by: Donald W.K. Andrews), Cambridge University Press.