PROGRAM EVALUATION WITH RANDOMIZED SCREENERS: ESTIMATING HETEROGENOUS RESPONSE INSTRUMENTAL VARIABLE (HRIV) MODELS

MICHAEL MUELLER-SMITH

Abstract. In this paper, we consider a growing literature on evaluating social programs through randomized administrative screeners. The state of this discipline is reviewed, and a new methodology is proposed to allow for heterogenous response instrumental variables (HRIV). New and existing methodologies are compared on a theoretical and empirical basis. An application considering criminal justice and recidivism provides new estimates on the impacts of corrections policies in the United States.

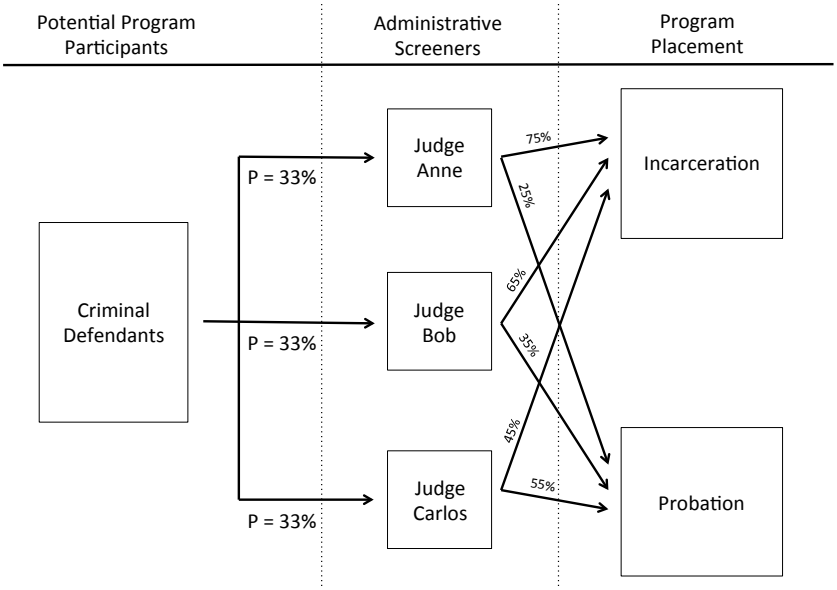
1. Introduction

The past decade has witnessed a substantial growth in the availability of digitized administrative records for empirical research in the United States. Burgeoning access to quality micro-data has encouraged many attempts at estimating the impact of numerous public programs. A variety of identification strategies to surmount the endogeneity of program participation have been developed; in particular, one small but growing subset of research designs has focused on the random (or rotational) assignment of program participants to administrative screeners. Screeners serve the role of using discretionary judgement to allocate individuals to programs, activities, or sanctions depending on the setting. Such studies leverage the quasi-random assignment of screeners and their varying program enrollment propensities as a source of exogenous variation in program take-up.

Figure 1 illustrates a generic setting with randomized screeners using criminal sentencing as a representative example. The population of potential program participants (e.g., criminal defendants) are randomly assigned to one of three screeners: Judge Anne, Judge Bob or Judge Carlos. In this simplified example, the screeners then make a single decision: whether to incarcerate or probate criminal defendants. Random assignment of the population ensures that each screener's caseload shares common observed and unobserved characteristics. If the screeners all utilize the same decision rule when assigning defendants to punishments, then the average sentencing outcomes for the three caseloads should be statistically equivalent given enough observations. In this example, the fact that Judge Anne incarcerates 30% more of her caseload compared to Judge Carlos indicates that she is a tough judge using a stricter decision rule. Thus, being randomly assigned to her courtroom increases the likelihood of incarceration for marginal defendants.

Date: April 16, 2012.

FIGURE 1. Program Participation with Administrative Screeners



Kling [2006] was the first to exploit this research design in a study of the impact of incarceration length on wages. Using a sample of 4,610 criminal offenders in California with randomly assigned judges determining their sentencing, he found a zero impact of an additional year of incarceration on short and medium-term earnings. Following up on this work, Aizer and Doyle [2011] as well as Nagin and Snodgrass [2011] have utilized judicial random assignment in juvenile and adult criminal court cases respetively to explore the impact of incarceration on recidivism.

A number of non-criminal applications have been explored with this technique. Doyle [2007] and Doyle [2008] use the random assignment of child welfare investigators that result in different child removal tendencies to estimate the causal impact of foster care placement on teenage pregnancy, criminality and employment. Autor and Houseman [2010] and Autor et al. [2012] use rotational assignment to privatized workforce development centers to estimate the impact of temporary job placements for welfare recipients. Belloni et al. [2011] use random panels of federal judges to estimate the impact of eminent domain rulings on local prices and growth. Munroe and Wilse-Samson [2012] use random assignment of civil court judges to estimate the impact of foreclosures on local property prices. One of the more creative applications of this technique comes from Doyle et al. [2012] which uses rotational assignment of ambulences and their corresponding propensities to bring patients to different hospitals to estimate the impact of various measures of hospital quality on patient outcomes.

Advancement of this identification strategy has several benefits to research and public policy. First, existing studies demonstrate the wide range of topics that can be explored

with this research design. Many social programs cannot rely on hard-and-fast rules for program assignment but instead must entrust discretionary judgement to program screeners. Future work will likely uncover new and innovative areas to apply this methodology as additional administrative records become available. Second, this research design is often applied in contexts where administering a randomized control trial would largely be considered unethical. As such, exploiting randomized screener assignment creates some of the first opportunities to explore exogenous variation in take-up for many social programs.

This study seeks to enhance this line of research through proposing a modified research design to improve estimator robustness and efficiency. The modification we propose, detailed in Section 3, is to expand current methods to allow for heterogenous response instrumental variables (HRIV). This change is motivated by economic theory and fits within the existing econometric and statistical frameworks. In the context of screener random assignment, the modified research design allows screeners to respond differently to various subsets of their overall caseload.[1] Intuitively, the difference can be seen as instrumenting with estimates of screener-specific decision rules based on participant characteristics rather than average rates of program assignment.

To complement the methodological discussion, we apply the new approach and compare it to existing estimators to study the impact of a criminal justice policies on recidivism in Section 4. Studying this important policy issue is interesting in its own right due to the broad reach and significant financial burden of criminal justice in the United States. In 2011, 1 in every 31 adults in the United States was in involved with correctional supervision, while corrections expenditures accounted for 7.2 percent of state fiscal budgets. Both of these figures reflect steady growth over the past thirty years.

The application is explored through a new dataset collected by the author, which spans 30 years of criminal court proceedings from one of the five largest metropolitan areas in the United States. In total, it includes close to 2.7 million misdemeanor and felony court proceedings and over 1 million unique defendants whose criminal activity is linked over time. More details regarding this new dataset are provided in Section 4.2.

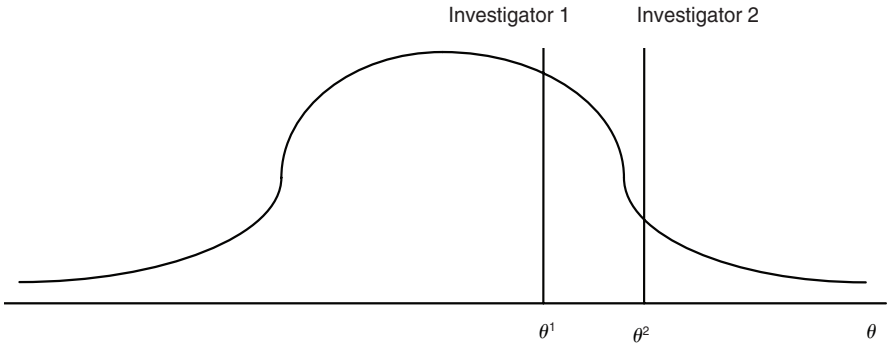
2. Existing Methods

The dominant approach to evaluating program impacts in the randomized screener context is to instrument for program participation using your assigned screener's caseload-wide rate of program enrollment. This approach, under specific assumptions, identifies the *local average treatment effect* (LATE) for marginal individuals induced into taking up the program as a result of their screener assignment.[2]

[1]While the focus of this paper is administrative screeners, the modified strategy is also applicable to research designs without administrative screeners. In particular, the modified strategy is likely to improve estimates in settings where we randomly assign incentives or some other intermediary that influences take-up programs.

[2]Several papers additionally explore *marginal treatment effects* (MTE) following Heckman and Vytlacil [2005]. The implementation of MTE estimators follows the basic assumptions of the JIVE estimates and thus will not be discussed in detail.

FIGURE 2. Program Assignment Thresholds



Source: Doyle [2007]

This framework can summarized in a standard triangular system, where:

$$D_i = 1[X_i\delta + Z_i\gamma + \nu_i > 0] \tag{1}$$

$$Y_i = X_i\beta + D_i\zeta_i + \epsilon_i \tag{2}$$

In this model, the outcome of final interest is Y_i, the program under study is D_i, observed participant characteristics are summarized in X_i and Z_i is a vector of dummy variables recording screener assignment. We allow the impact of program D_i to vary by individual, which is reflected in $\zeta_i \neq \bar{\zeta}$, to allow for potential heterogenous impacts of the program. With the standard assumptions on the covariance of Z_i, ν_i, and ϵ_i we will satisfy the necessary exclusion restrictions in the instrumental variable setting.

We can illustrate the principle behind this identification scheme in Figure 2. This figure, adapted from Doyle [2007], shows that different screeners (investigators in Doyle's context) have varying thresholds of program assignment. From Equation 1, we can see that:

$$D_i = \begin{cases} 1 & \text{if} \quad Z_i\gamma > -(X_i\delta + \nu_i) \\ 0 & \text{if} \quad Z_i\gamma \leq -(X_i\delta + \nu_i) \end{cases} \tag{3}$$

Defining $-(X_i\delta + \nu_i) = \theta$, we can view θ as a single index summarizing the relevant information a screener needs to decide program assignment. The interpretation of θ varies depending on the context: for foster care placements, Doyle interprets this as a distribution of observed abuse levels; for the incarceration versus probation decision, this could be viewed as an index collapsing the gravity the crime as well as potential risk an individual poses to the community. Different screeners exhibit different assignment thresholds (γ_z in the model, θ^z in the figure) which result in varying propensities of program allocation.

In order to avoid a small sample bias, the average propensity (p) assigned to individual i is calculated leaving out individual i's outcome and simply averaging over the rest of

screener j's caseload with N_j total individuals:

$$\hat{p}_i = \frac{1}{N_j - 1} \sum_{k=1, k \neq i}^{N_j} D_k$$

This calculation eliminates potential endogeneity when using screener rates rather than screener assignment dummies as the instrumental varaible. This strategy was originally implemented in Kling [2006], and has been utilized by most researchers seeking to leverage this general identification strategy.[3] As Kling points out, this estimation strategy is asymptotically equivalent to running a *jackknife instrumental variable estimator* (JIVE).[4]

A clear benefit of this estimation technique is that it is easy to implement. The dimensionality of screener dummy variables is reduced to one instrumental variable per program summarized in the program assignment propensity.[5] Smaller dimensionality reduces processing time as fewer computational resources are required.

The statistical properties of this model setup have been of interest to economists for many years. Angrist et al. [1996]'s LATE framework is a standard approach for estimating program impacts with instrumental variables. This framework requires five standard assumptions to ensure unbiasedness. The first assumption, known as the *Stable Unit Treatment Value Assumption* eliminates social interaction in the model (e.g., the screener assignment of your neighbor does not affect your own outcome). The second assumption is randomization of the instrument. The third assumption, popularly referred to as the *exclusion restriction*, requires that the instrument impacts the final outcome solely through program take-up and not through any other channel. The first three assumptions are posited with support from instititutional knowledge and placebo tests.

The fourth assumption requires a nonzero average causal effect of Z on D. Relaxing this assumption creates the problem of weak instruments. Recent work in this literature has developed methods that are robust to relaxing this problem. Additionally, statistical tests have been developed to determine whether instruments are in fact weak.

The last assumption is of monotonicity[6]:

$$D_i(Z_i = z_1) \geq D_i(Z_i = z_2), \text{ or, } D_i(Z_i = z_1) \leq D_i(Z_i = z_2) \quad \forall\, z_1,\ z_2 \in Z$$

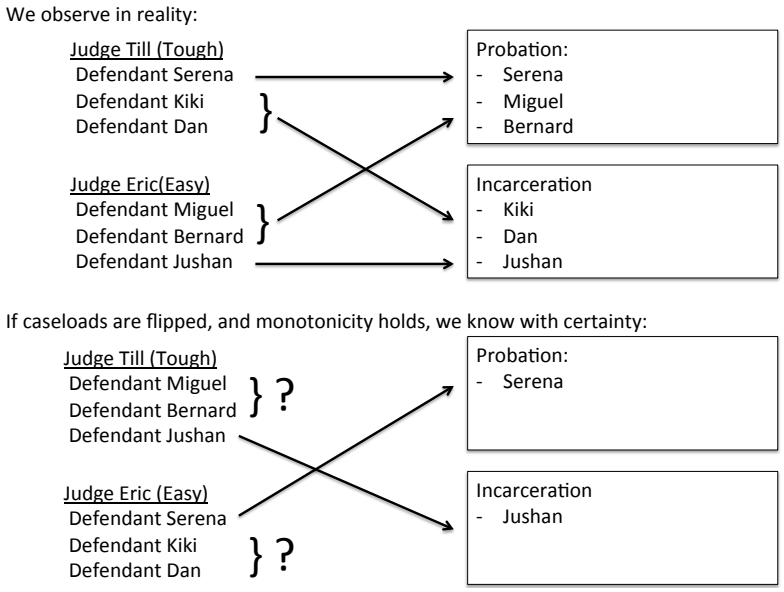
[3]One notable exception is Belloni et al. [2011].

[4]The bias from estimating allocation propensity without the leave-one-out approach is most pronounced when the researcher only observes a small number of observations per screener. This bias shrinks towards zero as the sample size per screener grows. An alternative approach to the leave-one-out methodology is to reduce the dimensionality of screeners (e.g., 100 potential caseworkers) by pooling individuals according to screener socio-economic characteristics (e.g., assignment to male or female caseworkers, etc) thereby increasing the number of observations per screener type. This latter strategy is implemented in Belloni et al. [2011].

[5]The theoretical properties of the estimator, though, are complicated by the fact that the researcher is using a predicted instrument. To account for the nature of the predicted instrument, clustering at the screener level and bootstrapping standard errors is necessary. While this complicates the theory, standard statistical packages make this easy to implement.

[6]Heckman and Vytlacil (2005) discuss at length why "monotonicity" is not an accurate moniker for this restriction and may be misleading. Instead, they believe "uniformity" to be a better descriptor.

FIGURE 3. Illustration of Monotonicity



This final assumption has received the most scrutiny in the randomized screener literature. The intuition behind montonicity is easy to illustrate with an example. In Figure 3, we observe two screeners (Judge Till and Judge Eric) and six defendants. Till is the tough judge who sentences two-thirds of his caseload to incarceration over probation, while Eric is easy and only sentences one-third of his caseload to incarceration. We never observe the counterfactual world where Till's caseload is switched with Eric's caseload, but if we did and monotonicity holds, we know with certainty that individuals incarcerated under the easy judge will continue to be incarcerated under the tough judge and individuals probated by the tough judge will also be probated by the easy judge.

Testing monotonicity is quite difficult and often must be taken as an article of faith. Failure in the monotonicity assumption, though, can have serious consequences, leading to bias in the LATE estimator. This result has been well understood for many years and is easy to illustrate. Frst, we introduce the Rubin's potential outcomes framework:

$$Y_i \quad = \quad Y_i(D_i) = D_i * Y_i(1) + (1 - D_i) * Y_i(0)$$

Here, we see that outcome Y for individual i depends on whether they enroll in program D or not. We observe Y_i, but it is truly a function of both one's realized outcome as well as their counterfactual outcome if they reversed their program participation decision. When we suspect participation in D is potentially endogenous and we cannot directly randomize D, we rely on instruments Z which affect take-up of D, but do not affect Y through any

other channel. We can formalize this relationship through writing a latent index model:

$$D_i^* = \gamma_0 + Z_i\gamma_1 + \nu_i D_i = \begin{cases} 1 & \text{if } D_i^* > 0 \\ 0 & \text{if } D_i^* \leq 0 \end{cases}$$

The bias in LATE when motonoicity fails is illustrated below. The reduced form should be proportional to the LATE, but without monotonicity it is not. To ease notation, let $\delta_i = D_i(z_1) - D_i(z_2)$.

$$\begin{aligned} E[Y_i|Z_i = z_1] - E[Y_i|Z_i = z_2] &= E[D_i(z_1) * Y_i(1) + (1 - D_i(z_1)) * Y_i(0)|Z_i = z_1] \\ &\quad -E[D_i(z_2) * Y_i(1) + (1 - D_i(z_2)) * Y_i(0)|Z_i = z_2] \\ &= E[(\delta_i) * (Y_i(1) - Y_i(0))] \\ &= Pr[\delta_i = 1] * E[Y_i(1) - Y_i(0)|\delta_i = 1] \\ &\quad -Pr[\delta_i = -1] * E[Y_i(1) - Y_i(0)|\delta_i = -1] \\ &\neq Pr[\delta_i = 1] * E[Y_i(1) - Y_i(0)|\delta_i = 1] \propto LATE \end{aligned}$$

As we can see, the reduced form is actually the net difference between the reduced form impacts for "compliers" ($\delta_i = 1$) and "defiers" ($\delta_i == -1$), and is not proportional to the LATE as long as $Pr[\delta_i = -1] * E[Y_i(1) - Y_i(0)|\delta_i = -1] \neq 0$. The bias can be summarized in the following equation:

$$\begin{aligned} &\frac{E[Y_i(z_1, D_i(z_1)) - Y_i(z_2, D_i(z_2))]}{E[\delta_i]} - E[Y_i(1) - Y_i(0)|\delta_i = 1] \\ &\qquad = \lambda * \{E[Y_i(1) - Y_i(0)|\delta_i = 1] - E[Y_i(1) - Y_i(0)|\delta_i = -1]\} \end{aligned}$$

where,

$$\lambda = \frac{Pr[\delta_i = -1]}{Pr[\delta_i = 1] - Pr[\delta_i = -1]}$$

Clearly, if motonocity holds, $Pr[\delta_i = -1] = 0$ and the bias is eliminated.[7]

To better understand how the modeling assumptions in our statistical framework constrain our estimation, we translate them into a restrictions on the behavior of agents in an economic model. Let us briefly return to the first stage in our model $D_i = 1[X_i\delta + Z_i\gamma + \nu_i > 0]$. As we observed in Equation 3 and Figure 2, the parameter vector γ can be interpreted as screener thresholds for program assignment. Identification centrally depends on variation in γ, and so it is worth considering why in fact there is variation in this key parameter in the first place. To this end, we take a short aside to model the behavior of screeners.

Screeners serve the purpose of allocating program entry to populations of potential participants. To the best of our knowledge, applications of this research design generally

[7]Two other standard assumptions besides monotonicity that eliminate this bias are assuming constant treatment effects (overall or among compliers and defiers) or assuming full or no take-up for one branch of assignment.

feature screeners who seek to maximize public welfare through program enrollments.[8] The screener j's problem is to decide whether to enroll each of N_j individuals in a program[9], can be summarized as following[10]:

$$\max_{D_i \; \forall \; i} W(D_1, ..., D_{N_j}) = \sum_{i=1}^{N_j} 1[D_i = 1] * E[\tilde{\zeta}(X_i, \xi_i)|I, I_j, X_i = x_i]$$

In this model, D_i is the binary variables indicating program assignment status for individual i. The parameter $\tilde{\zeta}$ reflects the net benefit of program entry (i.e., benefit minus cost) and is akin to the treatment effect. We use net benefit instead of treatment effect to emphasize that benefits and costs can acrue to individuals other than the participant. Because of heterogeneous effects, the net benefit depends on observed covariates X_i as well as observed individual-specific shocks ξ_i. The obvious solution for screener j is to enroll individuals in the program when the expected net benefit, given the common information shared across screeners (I) and the screener-specific information set (I_j), is positive.

We use expectations in this notation because the net benefits or treatment effects of various programs are unknown parameters. Among researchers, treatment effects are often a subject of much discussion without clear agreement on the sign or magnitude of impacts. In the context of randomized screeners, it is precisely the different perceptions, attitudes and knowledge specific to each screener and summarized in I_j that results in different enrollment rates.

The modeling assumptions made in the classic estimation framework, specifically separability between X_i and Z_i, impose a constraint on the way in which screeners form expectations regarding net benefits:

$$\begin{aligned} E[\tilde{\zeta}(X_i, \xi_i)|I, I_j, X_i = x_i] &= E[\tilde{\zeta_1}(X_i, \xi_i)|I, X_i = x_i] + E[\tilde{\zeta_2}(\xi_i)|I_j] \\ &= \psi(x_i, \xi_i) + \kappa_j \end{aligned}$$

In the above notation, $\psi(x_i, \nu_i)$ is the common component shared across screeners due the common information set I, while κ_j is a screener-specific shock. The seperability between these two components is a direct result of separability between X_i and Z_i in the first stage equation. In practice, this means that tough judges believe incarceration equally increases the net benefits for drug dealers and drunk drivers, elderly Caucasians and young African Americans, and male repeat offenders and female first-time offenders; I_j is not allowed influence the screener's assessment of the net benefits of enrollment for different subsets of the caseload.

Several empirical studies can provide us with details on the extent to which administrative screeners respond differently to covariates when determining program allocation.

[8]There also is potential that screeners also harbor nefarious motives, seeking to reward and punish different subsets of the population. Prevelance of this type of discrimination bolsters the argument but is not essential for it.

[9]Multiple programs whether mutually exclusive or not couple be accomodated.

[10]More realistic models would also include a budget constaint limitting overall program expenditures, but this is unnecessary for the argument being developed here.

Korn and Baumrind [1998] study clinician preferences in the United States and find that observationally equivalent patients receive different care from different clinicians and the variation in care relates to patient characteristics. Korn et al. [2001] follow up on the earlier study and find that clinicians do not even appear to show consistency in their decision making over time. Waldfogel [1998] found that differing sentencing patterns between judges in response to observed covariates could account for 9-10 percent in the variation observed in sentencing outcomes. Abrams et al. [2010] demonstrate that judges have treat race differently when sentencing criminal defendants. Price and Wolfers [2010] show that referees in the NBA call more fouls on players who are the opposite race than themselves. Each of these provides strong evidence of misspecification in the standard IV model.

Within caseload variation does not necessarily violate the monotonicity assumption, but it does create opportunities for failure. Procedures that account for differential variation within a caseload limits the potential for bias. This issue as well as the other benefits of estimating models that allow for more flexible decision rules are described in greater detail in the next section.

3. Heterogenous Response Intrumental Variables

In contrast to current methods, in this section we develop an alternative framework that allows for instruments to affect program take-up in heterogenous ways. We call this family of models *heterogenous response instrumental variables* (HRIV). We define a HRIV model in the following form:

$$\begin{aligned} D_i &= X_i\delta + Z_i\gamma(X_i) + v_i \\ &= h(Z_i, X_i) + v_i \end{aligned} \tag{4}$$

$$Y_i = X_i\beta + D_i\zeta_i + \epsilon_i \tag{5}$$

where, the coefficients on our instruments Z_i in Equation 4 to depend on the vector of observed covariates X_i. The parameterization of first stage equation can be rewritten as an unknown function $h(Z_i, X_i)$ foreshadowing the need of non-parametric methods.

The economic interpretation of this model, in the context of screener random assignment, is that each screener holds a unique vector of different thresholds for program enrollment that directly relate to observed participant characteristics. This relaxes the assumption that screeners view programs as uniformly increasing or decreasing the expected net benefits for all member of their caseload.[11]

If $h(X_i, Z_i)$ is known, we can apply standard instrumental variable methods to estimate the LATE:

$$\begin{bmatrix} \hat{\zeta}_{HRIV} \\ \hat{\beta} \end{bmatrix} = \left(\begin{bmatrix} h(Z,X) \\ X \end{bmatrix} \begin{bmatrix} D & X \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} h(Z,X) \\ X \end{bmatrix} Y \right)$$

[11] An alternative interpretation is that each screener utilizes their own unique decision making process which is a function of both their own characteristics as well as observed defendant characteristics.

In practice, $h(X_i, Z_i)$ is unknown and so we must consistently estimate it with non-parametric methods. This can be accomplished with a linear combination of series terms including polynomials, B-splines, power series, etc. When Z_i is a vector of dummy variables, as in the case of screener random assignment, we can consistently estimate $h(X_i, Z_i)$ using $Z_i \otimes f(X_i)$, where $f(X_i)$ represents a series expansion around the observed covariates. The resulting estimator is along the lines of:

$$\begin{bmatrix} \hat{\zeta}_{HRIV} \\ \hat{\beta} \end{bmatrix} = \left(\begin{bmatrix} Z_i \otimes f(X_i) \\ X \end{bmatrix} \begin{bmatrix} D & X \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} Z_i \otimes f(X_i) \\ X \end{bmatrix} Y \right)$$

This estimation is described more explicitly in Section 3.2.

3.1. **Model Comparison.** The impact of estimating local average treatment effects from a HRIV model relative to the standard models can be split into two categories: bias and efficiency. HRIV models yield weakly smaller asymptotic bias. This is the direct result of eliminating some types of monotonicity violations. Revisiting the bias derived in section two, we can rewrite it to account for covariates:

$$Bias_{IV} = \frac{\int Pr[\delta_i(x) = -1] f_X(x) dx}{\int \{Pr[\delta_i(x) = 1] - Pr[\delta_i(x) = -1]\} f_X(x) dx} \left\{ \begin{array}{c} \int \{E[Y_i(1,x) - Y_i(0,x) | \delta_i(x) = 1] - \\ E[Y_i(1,x) - Y_i(0,x) | \delta_i(x) = -1]\} f_X(x) dx \end{array} \right\}$$

$$= \int \int \int \frac{Pr[\delta_i(x_1) = -1]}{Pr[\delta_i(x_2) = 1] - Pr[\delta_i(x_2) = -1]} \left\{ \begin{array}{c} E[Y_i(1,x_3) - Y_i(0,x_3) | \delta_i(x_3) = 1] - \\ E[Y_i(1,x_3) - Y_i(0,x_3) | \delta_i(x_3) = -1] \end{array} \right\} f_X(x_1)\, f_X(x_2)\, f_X(x_3)\, dx_1\, dx_2\, dx_3$$

The bias for failure in monotonicity in a HRIV model is:

$$Bias_{HRIV} = \int \frac{Pr[\delta_i(x) = -1]}{Pr[\delta_i(x) = 1] - Pr[\delta_i(x) = -1]} \left\{ \begin{array}{c} E[Y_i(1,x) - Y_i(0,x) | \delta_i(x) = 1] - \\ E[Y_i(1,x) - Y_i(0,x) | \delta_i(x) = -1] \end{array} \right\} f_X(x)\, dx$$

Clearly, $|Bias_{IV}| - |Bias_{HRIV}| \geq 0$ as $|Bias_{IV}| = |Bias_{HRIV}| + |\kappa|$. Estimating the program impacts through a HRIV model weakly reduces the absolute magnitude of the bias.

The impact on efficiency is theoretically ambiguous. Accounting for relative within caseload variation between screeners can improve precision in the first stage; however, implementing a HRIV-based estimator introduces many instruments as a result of the non-parametric estimation. Among this large set of instruments, we confront issues with both dimensionality as well as weak correlation to program take-up. The standard way to evaluate efficiency in the context of many/weak instrumental variables is to consider whether the concentration parameter increases or decreases when we add an additional instrument to the first stage. In our context, the concentration parameter would be defined as:

$$\hat{\mu}_N^2 = \sum_{i=1}^{N} \frac{(X_i \hat{\delta} + Z_i \otimes f(X_i) \hat{\pi})^2}{E[\hat{\nu}_i^2]}$$

This statistic uses information derived from decomposing the variation in D_i into two components: observed variation and unobserved variation. The parameter takes the ratio of observed variation to the unobserved variation, which captures the degree to which

your instrumental variables explain the program participation. The distribution and corresponding accuracy of the IV estimators crucially depends on μ_N^2, with the convergence rate being $1/\mu_N$.

When omitted valid instruments are added to the first stage, $\hat{\mu}_N^2$ generally grows as the total variation of D_i explained in the right hand side variables increases while the residual will shrink. The opposite however is true when irrelevant instruments are added to the regression. The numerator of the concentration parameter increases but so does the denominator, causing the overall value of μ_N^2 to shrink. Since HRIV methods introduce many irrelevant estimators, specific procedures will need to be employed to deal with the potential loss of efficiency. These are described at length in the next subsection.

Appendices A.1, A.2, and A.3 illustrate the theoretical results on bias reduction, efficiency gain and efficiency loss through simulation exercises.

3.2. **Nonparametric Estimation.** The key challenge in estimating a HRIV model is approximating $h(X_i, Z_i) = E[D_i|X_i, Z_i]$. A significant amount of work over the past decade has advanced two key methodologies: K-class estimators and shrinkage estimators.

Among the two groups, k-class estimators are more familiar to applied economic researchers. Examples include *limited information maximum likelihood* (LIML), *Fuller's modified LIML* (FULL) and the *bias corrected two stage least squares* (BTSLS). Each of these estimators share a common esimation framework:

$$\hat{\zeta} \quad = \quad (X'P_Z X - \hat{\kappa}X'X)^{-1}(X'P_Z Y - \hat{\kappa}X'Y)$$

where, for LIML, $\hat{\kappa} = \tilde{\kappa}$ which is equal to the smallest eigen value of the matrix $(\bar{X}'\bar{X})^{-1}\bar{X}'P_Z\bar{X}$ where $\bar{X} = [y, X]$. FULL is a modification of the LIML estimator where $\hat{\kappa} = [\tilde{\kappa} - (1 - \tilde{\kappa})C/N]/[1 - (1 - \tilde{\kappa})C/N]$ for some constant C, which is approximately mean unbiased at $C = 1$. The BTSLS estimator sets $\hat{\kappa} = [1]/[(1 - (K_N - 2))/N]$, where K_N is the number of instruments.

While it is beyond the scope of this paper, a large literature has demonstrated the robustness of K-class estimators to problems with both weak and many instruments (see Bekker [1994], Staiger and Stock [1997], Donald and Newey [2001], Stock et al. [2002], Moreira [2003], Chao and Swanson [2005], and Hansen et al. [2008]). These robust properties make k-class estimators attractive options for estimating HRIV models.

An alternative framework from statistics has recently been gaining traction among econometric researchers. This framework, known as *shrinkage estimation*, focuses on optimal prediction techniques. This is achieved through selecting the strongest instruments to avoid the weak instruments problem as well as reducing dimensionality to avoid the many instruments problem. Examples of such techniques include Boosting, Principle Components, LASSO, Post-LASSO, Elastic Net. Recent econometric research in this field includes: Okui [2011], Belloni et al. [2011], Hastie et al. [2009], Ng [2011], and Bai and Ng [2010].

Many shrinkage estimators have been developed by statistical researchers. The decision of which estimator to utilize depends on the data generating process in the first stage, specifically whether the data generating process is sparse or dense in nature. A sparse data generating process refers to a setting in which only a small subset of potential instruments

have a strong correlation to the endogenous variable. Notable sparse procedures include LASSO and Post-LASSO. In contrast, a dense data generating process refers to a setting where among the set of many correlated potential instruments each contributes a weak correlation to the endogenous variable. The prefered approach in this context would be using a principle components-type estimator. Hybrid procedures that integrate aspects of both dense and sparse models have also been developed; elastic net is one such example.

Work comparing the performance of k-class estimators and shrinkage estimators do not give clear conclusions on which procedures are most optimal. Some studies have found instances where shrinkage estimators outperform k-class estimators, but these are often dependent on the specific context of the data generating process.

4. Application: U.S. Corrections Policy and Criminal Recidivism

We implement the methodologies considered in Sections 2 and 3 to an empirical application studying criminal justice and recidivism in the United States. The application aids in illustrating the deficiency of current empirical methods; standard IV-based estimators yield results that are statistically and economically inconsistent with findings from the more robust HRIV-based methods.

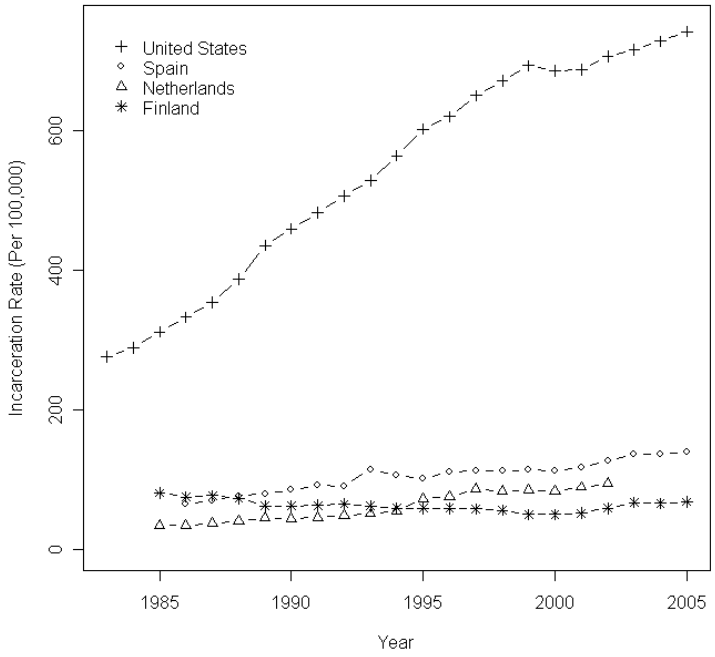
Using our modified estimators, we provide new evidence on the causal impacts of several apparatuses of criminal justice in the United States. We focus on penalty types (and intensities) that are most prevelant in the United States, specifically: probation, incarceration, and fines. We also able to explore the impact of having a criminal record (independent of penalty) through a unique sentencing outcome available in the data.

Aside from the methodological contribution of this paper, studying criminal justice in the United States is incredibly important in and of itself. Over the past 30 years, the incarcerated population in the United States has grown by close to 300 percent, vastly outpacing population growth over the same time period (about 50 percent). This trend is unique to the United States; other OECD countries have remained stable in their per capita incarceration rates over time (see Figure 4).

The result of this dramatic growth is that the United States now has the highest per capita levels of incarceration in the entire world. Table 1 lists the thirty countries with the highest incarceration rates among those with at least 1 million inhabitants. The United States is the clear leader in incarceration, more than doubling the tenth highest country, Ukraine. Few other OECD countries, besides the United States, are present on the table.

Criminal justice in the United States, however, extends well beyond just prisons and jails. The Bureau of Justice Statistics reports that in 2009, 7.2 million adults in the United States were under "correctional supervision", a status that includes probation and parole in addition to incarceration. On a per capita basis, this would translate into 1 out of every 31 adults. Subgroups of the population exhibit even higher rates: 1 in every 18 adult males and 1 in every 11 adult African Americans were under correctional supervision in 2009. These are static measure, demonstrating the extent of criminal justice at a given point in time; if we were to consider ever being under correctional supervision, these figures would be much higher.

FIGURE 4. Incarceration Trends in Four Wealthy Countries



Source: Wildeman [2011]

The wide reach of corrections activity translates into substantial portions of state budgets. The National Association for State Budget Officers reports that states spent $51.7 billion dollars in fiscal year 2011 on corrections programs, reflecting 7.4 percent of total state budgets. These figures would be substantially larger if expenditures on other forms of public safety like police enforcement or other crime prevention efforts were included.

Finally, there has also been renewed focus on reducing recidivism in the United States as the number of children with incarcerated parents has grown. According to the Bureau for Justice Statistics (2008), in 2007, 1.7 million children in America had an incarcerated parent, up 80 percent since 1991. Parental incarceration has been associated with a host of negative outcomes for children, increasing the odds of poverty and criminality among our most at-risk children. To stabilize these marginal families, we need a better understanding of how to prevent reoffending.

4.1. **Related Literature.** A substantial literature of varying quality has developed in response to growing concern regarding incarceration in the United States. A sizeable portion of this work examines the cross-sectional relationship between various corrections policies and outcomes for criminal offenders. This literature is fundamentally flawed due to classic omitted variables bias; those who receive worse punishments are more likely to exhibit unobserved characteristics that would also lead to poor outcomes regardless

TABLE 1. Highest incarceration rates among countries with 1 million or more inhabitants

Rank	County	Prisoners per 100,000 of national population			
1.	United States of America	756	16.	Trinidad and Tobago	270
2.	Russian Federation	629	17.	Singapore	267
3.	Rwanda	604	18.	Tunisia	263
4.	Cuba	531	19.	Estonia	259
5.	Belarus	468	20.	Thailand	257
6.	Georgia	415	21.	Mongolia	244
7.	Kazakhstan	378	22.	United Arab Emirates	238
8.	Puerto Rico	330	23.	Lithuania	234
9.	Israel	326	24.	Azerbaijan	229
10.	Ukraine	323	25.	Brazil	227
11.	South Africa	318	26.	Moldova	227
12.	Chile	305	27.	Turkmenistan	224
13.	Panama	295	28.	Iran	222
14.	Latvia	288	29.	Poland	221
15.	Taiwan	276	30.	El Salvador	207

Source: Walmsley [2009]

of sentence. Because of the sensitve nature of crime and punishment, randomized control trials have not been extensively explored as a method to evaluate criminal justice programs.

The past decade, however, has observed a series of creative approaches to exploring the dynamics of criminal justice in the United States. Several authors have sought to evalaute the determinants of criminal behavior. Establishing whether harsh criminal sentencing deters crime has been a major focus in this research agenda. Katz et al. [2003] find that bad prison conditions (as measured by prisoner death rates) are significantly correlated with lower levels of crime providing some evidence of deterrence. However, Lee and McCrary [2009], using Florida administrative data and a discontunity in sentencing guidelines for criminal offenders around age 18, find a very small negative deterrant effect of harsher sentencing. McCrary and Sanga [2012], expand this paper with data from six states and find no evidence of a deterrant effect, observing a smooth function of criminal across the threshold of turning 18 years old despite a large increase in the severity of punishments.

Another branch of research in the determinants of crime literature seeks to answer whether there is a contagious element to criminality. Bayer et al. [2009] finds that the criminal background of cellmates (who are effectively randomly assigned) influence the future criminal behavior of juvenile offenders providing credence to the contagion hypothesis. Ludwig and Kling [2007], however, provides evidence against a contagion effect in the Moving To Opportunity experiment; in this randomized control trial, which randomly relocated

families to new neighborhoods, there was no correlation between the ambient levels of crime in the new neighbhoods and the criminality of the study participants. In considering the macro-relationships between different types of offenses, Levitt and Kuziemko [2004] find that shift towards incarcerating drug offenders during the late 1980's and early 1990's led to a small reducation in violent and property crimes. They interpret the drop in violent and property crimes to be the result of decreased drug consumption in the population as increased incarceration for drug offenses pushed up drug prices.

A growing literature focused on clean identification has developed around studying of the impacts of criminal sentencing. Kling [2006] studied the relationship between incarceration length and earnings using random assignment of federal judges in California as an instrument to determine if human capital erodes while incapacitated. He found no impact of incarceration length on earnings, which he interprets as evidence against human capital erosion. Using similar identification techniques, Nagin and Snodgrass [2011] using data from Pennsylvania find no significant impact of incarceration compared to probation on recidivism, while Aizer and Doyle [2011] using data from Chicago find juvenile incarceration increases the likelihood of adult reoffending. Using a different identification technique, Owens [2009] finds that individuals who were released on early parole as an unexpected policy change committed more crimes of a more serious nature compared to individuals compared to those who served the full duration of their original sentence. Kuziemko [2011] though suggests the need to consider equilibrium behavior as she finds that when early release from incarceration is made at the discretion of parole boards, prisoners invest in signals that lower their recidivism risk such education and work programs during imprisonment.

As we observe in the literature, several distinct theoretical channels could be active when determining the impact of sentencing on future outcomes. If there is uncertainty regarding the consequences of criminal behavior, personally experiencing a criminal sentence could have a deterrant effect on future criminality. Understanding the consequences of breaking the law or the austerities of prison life, may deter individuals from committing future offenses. Additionally, incarceration and probation both feature incapacitation effects, where by being placed behind bars or meeting with a probation officer on a weekly basis, you are physically prevented from committing criminal acts. While both of these channels suggest that harsh sentencing should decrease the risk of recidivism, the deterrance effect should operate over a long horizon while the incapcitation effect will only last as long as the individual is under correctional supervision.

On the other side, it can be argued that harsh criminal sentencing leads to the atrophy of human and social capital. Incarceration can disrupt education and employment, and lead to long spells of detachment from the labor force. This can make it difficult to secure employment after the conclusion of one's sentence, leading individuals to seek illegal sources of income. In addition, incarceration may destabilize relationships and families, eroding the social ties defendants have to the community. This would lower the benefits of remaining out of jail for defendants, increasing the likelihood of recidivism.

Lastly, there is potential for there to be a stigmatizing or scarring effect of criminal sentencing. Many employers and educational institutions require that individuals report

if they have ever been convicted of a felony offense. This can legally be used a screening device making reintegration in the labor market difficult or impossible.[12] This same stigmatization process may also occur informally among social networks regardless of whether the conviction was for a felony or a misdemeanor charge.

4.2. **Data and Research Setting.** As discussed extensively throughout this paper, we leverage randomized administrative screeners as a source of exogenous variation in criminal court outcomes. The analysis is based on data from Harris County, Texas, which includes the city of Houston as well as several surrounding municipalities. The Houston metropolitan statistical area is the fifth largest in the United States and encompases a geographical area slightly larger than the state of New Jersey. The population is economically and demographically diverse; this is reflected in the observed population of criminal defendants.

Criminal court cases in Harris County are randomly assigned to judges at the time of filing through a computerized system. The purpose of random assignment is to maintain a balanced caseload across all of the judges. The majority of cases fall into one of two tracks: misdemeanor or felony cases.[13] Misdemeanor cases, which account for lesser crimes and are only eligible for a maximum of one year of incarceration, are administered by the Harris County Criminal Courts at Law, of which there are fifteen. Felony cases, which are more serious in nature, are handled by the twenty-two Harris County District Courts.[14] Judges in both court systems are elected officials and serve exclusively their own caseloads (i.e., only felony cases or only misdemeanor cases).[15]

Criminal judges in Texas hold siginificant discretion over court proceedings. Texas is one of the few states that does not adhere to federal sentencing guidelines. Guidelines have been established at the state level (see Table 2), but the guidelines are quite broad and are still not required by law. In most cases but murder, judges have the option to probate jail time so the defendant serves their sentence under community supervision, commonly known as probation. The duration of probation is also at the discretion of the court judge.

In addition, judges influence the outcomes of trials through several other channels. First, they have the discretion to determine what evidence is admissible in court; this potentially affects the final verdict as well as bargaining power should the defendant to elect to negotiate for a plea bargain. Another feature, unique to Texas, is that the Harris County courts did not have a public defenders office until 2011; in the event a defendant could not afford legal representation, the trial judge was responsible for appointing a lawyer for the

[12]Sex offenders are extreme example of stigmatization as they must register with local authorities and are restricted in terms of where they can live and work.

[13]The lowest level misdemeanor cases are handled by Justices of the Peace Courts and are not randomly assigned.

[14]Randomization occurs separately depending on whether a case is slotted for misdemeanor or felony status.

[15]Felony judges have the discretion to lower the judged offense for convicted offenders from a low felony to a high misdemeanor, but such judges would never hear a case that only included misdemeanor charges in the first place.

TABLE 2. Charges, Crimes and Texas Sentencing Guidelines

Charge	Typical Crimes	Eligible Penalty
Capital Felony	Murder of a public safety officer, Multiple Murders, Murder of a child	Death or Life without Parole
First-degree Felony	Murder, Possession of a controlled substance (CS) with intent to distribute, Theft over $200,000	5 to 99 years in a state prison and/or a fine of not more than $10,000
Second-degree Felony	Possession of a CS $>$ 4 grams and $\leq$ 200 grams, Aggravated Assault with a deadly weapon, Indecency with a child (by contact), Intoxicated Manslaughter	2 to 20 years in a state prison and/or a fine of not more than $10,000
Third-degree Felony	Possession of CS $>$ 1 gram and $\leq$ 4 grams, Aggravated Assault, DWI (3rd Offense), Solicitation of a minor	2 to 10 years in a state prison and/or a fine of not more than $10,000
State jail Felony	Possession of CS $\leq$ 1 gram, DWI with a minor under the age of 15 in the vehicle, Third theft conviction of any amount	180 days to 2 years in a state jail and/or a fine of not more than $10,000
Class A Misdemeanor	DWI (2nd offense), Assault causing bodily injury, Possession of marijuana (between 2 oz. and 4 oz.), Illegal possession of prescription drugs	Not more than 1 year in a county jail and/or a fine of not more than $4,000
Class B Misdemeanor	DWI (1st offense), Possession of Marijuana (less than 2 oz.), Prostitution	Not more than 180 days in a county jail and/or a fine of not more than $2,000
Class C Misdemeanor	Assault by contact, Drug paraphernalia, Disorderly conduct, Theft under $50	A fine of not more than $500

representative. Bright [2000] discusses the variety of ways Texan judges manipulated this system at the expense of consistent application of the law in the context of capital crimes:

> Texas trial judges-some treating the appointment of counsel to defend poor defendants as political patronage and some assigning lawyers not to provide zealous advocacy but to help move their dockets-have frequently appointed incompetent lawyers to defend those accused of capital crimes.

Popular press in the early 2000's documented cases where appointed counsel were underqualified, intoxicated, and/or asleep at the time of trial.[16]

As is common in many states in the U.S., Texan criminal court records are treated as a matter of public record; in Texas, this is legally established through the Public Information Act (Texas Government Code Chapter 552)[17]. The Harris County District Clerk has primary responsibility for maintaining and permitting access to the records for both the felony and misdemeanor courts. The criminal records can be accessed onsite in the Harris

[16]See for example New York Times, 11 June 2000, "Texas Lawyer's Death Row Record a Concern"
[17]http://www.statutes.legis.state.tx.us/docs/GV/htm/GV.552.htm

County District Clerks office or alternatively via an online database hosted by the office.[18] The online database served as the primary source of data collection for this project.[19]

A significant effort was required to translate the text records from the online database into an empirical dataset, particularly when coding schemes changed over time. The end result is a dataset containing information regarding the charges of a crime, judicial assignment, court trial outcomes as well as detailed penalty information, defendant demographics and defendant personal identifying information.[20] Specific elements are displayed in Table 3.

In total, we leverage thirty years worth of criminal court records from Harris County, Texas. These represent close to 2.7 million court proceedings, of which there are over 1 million unique defendants.[21] The records are the universe of non-federal criminal proceeding in Harris County between 1980 and 2009 excluding two key groups. First, cases charged as crimes at the Misdemeanor C level were not collected as they are administered by Justice of the Peace Courts which are not randomly assigned. Second, cases that have been sealed by the court to the public are not included; these reflect roughly 15,000 total cases.

Trends of the aggregate caseload based on the collected micro-data are summarized in Figure 5. The size of both misdemeanor and felony caseloads have sustained continuing growth over this time period. In a given year, about 80 percent of cases are found guilty, of which 10 to 15 percent receive a deferred adjudication of guilty ruling. This status indicates that if the defendant successfully completes his punishment without any issues, the conviction will be erased from his record as if it never happened. Lastly, slightly less than half of misdemeanor trials and slightly more than half of felony trials end in incarceration for the accused.

Table 3 shows summary statistics for misdemeanor and felony defendants. Both the misdemeanor and felony caseloads are predominantly male with mean age around 30 years old. Misdemeanor cases split roughly into even splits between non-Hispanic Caucasians, African American and Hispanics, while felonies have a larger proportion of African Americans. Detailed physical descriptions are available for the majority of the sample. The preponderance of cases with missing information for these fields were charged during the early 1980's when detailed records were not maintained.

Individuals facing misdemeanor charges have on average been charged with and convicted of fewer previous crimes compared to felony defendants; close to half of the misdemeanor cases are first-time offenders while only one third of the felony caseload are first-time offenders. The most common crime types for misdemeanor cases are driving while intoxicated (DWI), theft and drugs; for felony cases, the most common are drugs, theft and assault.

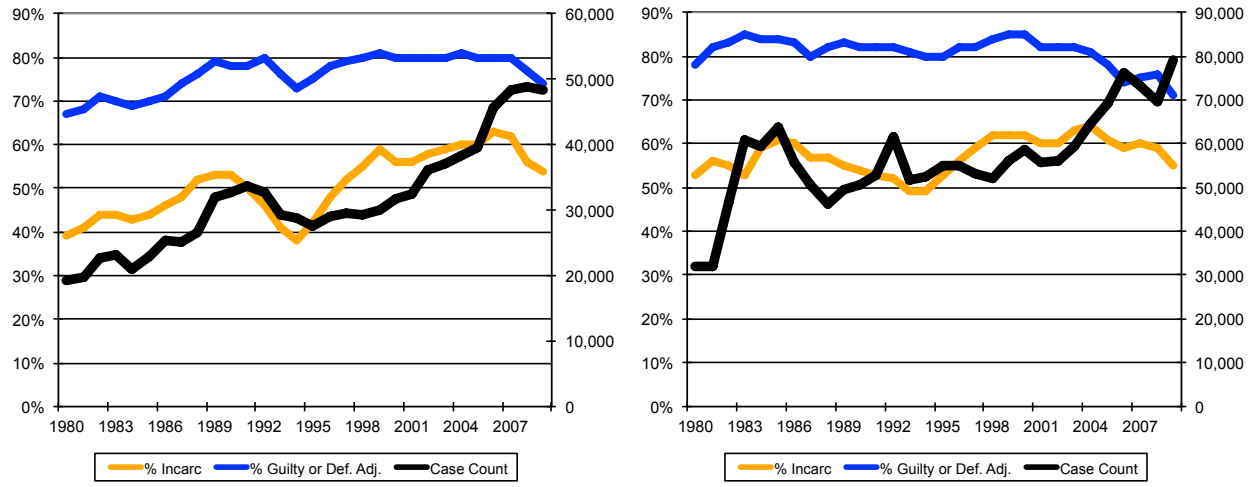
[18]http://www.hcdistrictclerk.com/eDocs/Public/Search.aspx

[19]Data collection was approved under Columbia IRB protocol number: IRB-AAAI1323

[20]Future work is seeking to leverage the personal identifying information (full name, date-of-birth and home address) to merge this dataset with other sources of administrative data, specifically UI Wage Records and Vital Statistics Birth Records.

[21]Repeat offenders in Harris County system can be observed through a unique identifier linking the records over time.

FIGURE 5. Trends in the Misdemeanor and Felony Caseloads, Harris County, TX (1980-2009)



(A) Misdemeanors (B) Felonies

Evidence that randomization occurred and was abided by in practice can be tested by regressing pre-existing case characteristics on the vector of judge dummy variables and computing the F statistic testing the joint hypothesis that all coefficients are equal to 0 (when including a constant in the equation). Tables 4 and 5 show a subset of these balance test results, selecting a random subset of judges working between 1995 and 1999. While there is some degree a minute variation between judges, the characteristics are generally commonly distributed across judges.

In contrast, Tables 6 and 7 demonstrate a large amount of variation between judges in average trial outcomes. This test does not even make use of the type of within caseload variation discussed in Section 3, indicating that we should have substantial power when we use judicial assignment as instruments for trial outcomes.

TABLE 3. Characteristics of Defendants, 1980-2009

Characteristics	Misdemeanors	Felonies	Characteristics	Misdemeanors	Felonies
Sex = Female	21%	18%	Weight (lbs)	170	173
Race = White	39%	30%	Height (Inches)	68	68
Race = Black	31%	46%	Build = Heavy, Obese	8%	9%
Race = Hispanic	28%	22%	Build = Medium	61%	66%
Age	29.6	30.0	Build = Skinny, Light	11%	12%
			Build = Missing	20%	12%
Skin = Black	4%	7%			
Skin = Dark	4%	7%	Tattoos = 1	11%	16%
Skin = Dark Brown	9%	14%	Visible Scars	6%	9%
Skin = Fair	15%	13%			
Skin = Light	6%	5%	Cumulative Felony Charges	0.5	1.3
Skin = Light Brown	7%	6%	Cumulative Misd. Charges	1.2	1.5
Skin = Medium	22%	22%	Cumulative Felony Convictions	0.4	1.1
Skin = Medium Brown	9%	10%	Cumulative Misd. Convictions	1.0	1.3
Skin = Olive	3%	3%	First Time Offender	47%	33%
Skin = Missing	21%	13%			
			Crime Type = DWI	23%	-
Eyes = Green, Blue	18%	15%	Crime Type = Theft	17%	26%
Eyes = Brown, Black	66%	75%	Crime Type = Drug	12%	35%
Eyes = Miss	16%	10%	Crime Type = Traffic	11%	-
			Crime Type = Assault	8%	14%
Hair = Blonde, Red	7%	6%	Crime Type = Fugitive	7%	-
Hair = Black, Brown	75%	83%	Crime Type = Weapon	5%	8%
Hair = Missing	16%	9%	Crime Type = Fraud	-	8%
			Crime Type = Deadly Conduct	-	7%
N	1,699,734	946,524	N	1,699,734	946,524

TABLE 4. Balance in Pre-Existing Characteristics by Misdemeanor Judge, 1995-1999

VARIABLES	Judge A	Judge B	Judge C	Judge D	Judge E	Judge F	Judge G	Judge H	Judge I	Judge J	F Stat	N
Sex = Female	21%	20%	21%	20%	21%	20%	20%	21%	21%	20%	1.4	271,379
Race = White	37%	35%	37%	39%	37%	38%	36%	36%	37%	37%	3.5	271,379
Race = African American	31%	33%	31%	32%	31%	30%	31%	31%	31%	31%	1.4	271,379
Race = Hispanic	31%	30%	30%	28%	30%	31%	30%	31%	31%	31%	5.2	271,379
Age	29.8	29.6	29.8	30.1	29.8	29.7	29.6	29.7	29.9	29.6	3.1	269,103
Weight (lbs)	171	172	171	172	172	171	171	171	172	171	1.5	255,438
Height (Inches)	68.1	68.2	68.2	68.3	68.2	68.2	68.2	68.2	68.2	68.2	1.8	255,820
First Time Offender	44%	43%	44%	46%	44%	43%	43%	43%	44%	45%	4.1	271,379
Crime Type = Drug	10%	12%	10%	9%	10%	10%	10%	10%	11%	11%	3.5	271,379
Crime Type = DUI	19%	16%	19%	16%	18%	19%	18%	18%	18%	18%	6.3	271,379
Crime Type = Theft	16%	16%	17%	14%	17%	16%	16%	16%	16%	16%	5.7	271,379

Note: F Statistics are conditional on week of charge fixed effects

TABLE 5. Balance in Pre-Existing Characteristics by Felony Judge, 1995-1999

VARIABLES	Judge A	Judge B	Judge C	Judge D	Judge E	Judge F	Judge G	Judge H	Judge I	Judge J	F Stat	N
Sex = Female	20%	20%	19%	21%	20%	22%	21%	22%	22%	22%	1.4	145,018
Race = White	29%	28%	27%	26%	27%	27%	27%	28%	27%	31%	1.8	145,018
Race = African American	43%	45%	46%	48%	46%	46%	47%	45%	45%	47%	2.9	145,018
Race = Hispanic	27%	25%	25%	24%	25%	25%	25%	26%	27%	20%	1.7	145,018
Age	29.5	30.0	30.0	29.5	29.7	29.7	29.9	29.8	30.2	30.9	2.0	143,313
Weight (lbs)	174	174	173	173	175	174	174	174	172	175	1.7	139,837
Height (Inches)	68.3	68.3	68.3	68.4	68.4	68.3	68.3	68.3	68.3	68.4	1.1	139,983
First Time Offender	33%	32%	32%	31%	33%	31%	33%	32%	31%	27%	3.0	145,018
Crime Type = Assault	16%	16%	14%	16%	15%	15%	15%	15%	16%	15%	2.2	145,018
Crime Type = Drug	32%	35%	38%	38%	36%	36%	37%	37%	40%	42%	1.7	145,018
Crime Type = Theft	22%	21%	20%	21%	20%	20%	20%	20%	19%	20%	1.0	145,018

Note: F Statistics are conditional on week of charge fixed effects

TABLE 6. Variation in Trial Outcomes by Misdemeanor Judge, 1995-1999

VARIABLES	Judge A	Judge B	Judge C	Judge D	Judge E	Judge F	Judge G	Judge H	Judge I	Judge J	F Stat	N
Verdict = Innocent	16%	17%	15%	17%	16%	29%	16%	16%	15%	18%	135.7	271,333
Verdict = Guilty	68%	71%	75%	74%	71%	60%	73%	72%	74%	70%	85.4	271,333
Verdict = Def. Adj. of Guilt	16%	11%	10%	10%	13%	10%	11%	12%	11%	13%	42.1	271,333
Probation Length > 0	29%	23%	19%	20%	24%	21%	22%	24%	21%	24%	48.4	271,379
Probation Length (Years)	0.38	0.27	0.23	0.26	0.29	0.25	0.27	0.29	0.26	0.31	60.3	271,333
Incarceration Length > 0	54%	59%	65%	62%	59%	49%	62%	59%	62%	57%	74.0	271,379
Incarceration Length (Years)	0.06	0.05	0.06	0.06	0.05	0.04	0.06	0.06	0.06	0.05	45.4	271,333
Fine Amount > 0	44%	40%	38%	40%	40%	25%	41%	41%	38%	41%	116.3	271,379
Fine Amount ($1,000)	0.12	0.10	0.12	0.11	0.10	0.23	0.11	0.12	0.11	0.11	112.2	271,333

Note: F Statistics are conditional on week of charge fixed effects

TABLE 7. Variation in Trial Outcomes by Felony Judge, 1995-1999

VARIABLES	Judge A	Judge B	Judge C	Judge D	Judge E	Judge F	Judge G	Judge H	Judge I	Judge J	F Stat	N
Verdict = Innocent	20%	23%	24%	21%	21%	23%	21%	19%	23%	25%	8.5	144,534
Verdict = Guilty	66%	52%	60%	56%	63%	57%	59%	63%	55%	59%	27.8	144,534
Verdict = Def. Adj. of Guilt	14%	26%	16%	23%	17%	20%	20%	18%	22%	16%	29.1	144,534
Probation Length > 0	17%	29%	32%	36%	20%	31%	31%	29%	34%	37%	34.9	145,018
Probation Length (Years)	0.8	1.2	2.0	1.9	1.0	1.5	1.7	1.9	2.1	2.3	48.8	144,534
Incarceration Length > 0	63%	49%	48%	46%	59%	49%	53%	53%	46%	42%	25.4	145,018
Incarceration Length (Years)	2.1	1.3	1.9	1.9	2.0	2.0	1.8	2.5	1.7	1.7	5.7	144,534
Fine Amount > 0	14%	16%	11%	14%	13%	18%	11%	15%	14%	13%	26.6	145,018
Fine Amount ($1,000)	0.08	0.12	0.06	0.08	0.08	0.11	0.07	0.14	0.12	0.08	4.2	144,534

Note: F Statistics are conditional on week of charge fixed effects

4.3. **Results.** In this section, we present the results of our empirical analysis. We study the probability of being charged with a new crime within 2 and 10 years of one's original filing date. We also consider the probability of being convicted of another crime conditional on being charged again in the future.[22]

For each outcome, we compute three estimators: $\hat{\zeta}_{OLS}$, $\hat{\zeta}_{IV}$ and $\hat{\zeta}_{HRIV}$.

$$\begin{aligned}
\hat{\zeta}_{OLS} &= \left(\begin{bmatrix} D \\ X \end{bmatrix} \begin{bmatrix} D & X \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} D \\ X \end{bmatrix} Y \right) \\
\hat{\zeta}_{IV} &= \left(\begin{bmatrix} Z \\ X \end{bmatrix} \begin{bmatrix} D & X \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} Z \\ X \end{bmatrix} Y \right) \\
\hat{\zeta}_{HRIV} &= \left(\begin{bmatrix} Z \otimes f(X)' \\ X \end{bmatrix} \begin{bmatrix} D & X \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} Z \otimes f(X)' \\ X \end{bmatrix} Y \right)
\end{aligned}$$

To implement the HRIV-based estimator, we simplify computation by setting $f(x) = [1, Race_i = White, Sex_i = Female, Age_i, FirstOffender_i = 1]$. Future work will seek to explore the more sophisticated methods described in Section 3.2 to estimate $\hat{\zeta}_{HRIV}$.

Tables 8 and 9 show the results for misdemeanor cases at 2 year and 10 year intervals; tables 10 and 11 show the corresponding results for felony cases. From a methodological standpoint, it is clear that different models yield divergent results. Not surprisingly, the OLS results are generally inconsistent with the instrumental variables regressions (either IV or HRIV) most likely reflecting an omitted variables bias. The IV versus HRIV comparison is quite interesting; one-third of the coefficients report statistically or economically conflicting results. There are fifteen instances where coefficients go from being statistically insignificant to statistically significant when changing from the IV to the HRIV estimators. In six instances, the HRIV model kills the significant result observed in the IV model. Most alarming are two cases in which both the IV and HRIV models report statistically significant results but each with opposite signs. Such results are highly disturbing as these different causal treatment effect estimates would lead to very different conclusions for public policy, despite clear satisfaction of the exclusion restriction.

Because of the robust properties of HRIV-based estimators, we focus on the magnitudes and significance found in these models when interpreting our results for public policy. We do not observe systematic evidence in favor of a stigmatization effect. Most results find a common impact of having a guilty or deferred adjudication of guilt verdict.[23] Having any duration of incarceration does increase the likelihood of being charged with an additional crime, which might indicate that stigma operates mainly through incarceration rather than other punishments.

With regard to intensity of sentence, there is an interesting dynamic in Table 8 when looking at probation length. Defendants are less likely to commit crimes for each additional

[22]In future work, we hope to also consider escalation and diversification in crime behavior as outcomes.

[23]Recidivism may not be an appropriate environment to evaluate this channel, however, as the court can always view the results for these cases, including those with a deferred adjudication of guilt ruling. As such, defendants are not shielded from the stigma of their previous crimes in the court system.

year of probation, but more likely to be convicted if charged with a crime. This may indicate that the heightened monitoring associated with probation deters criminals from reoffending in the future. In the event, they do reoffend, more evidence appears to be available to convict them.

The duration of incarceration for misdemeanor offenses increases the likelihood of being charged with additional crimes, while we observe the opposite effect for felony cases. The difference in results is likely due to the average duration of incarceration for misdemeanor cases being substantially shorter compared to felony cases. This would indicate that the incapacitation effect has a limited role for misdemeanor cases during our follow-up window, and the impacts we observe are likely more associated with atrophy of human and/or social capital channels. For felony offenses, which result in substantially longer prison spells, the incapacitation effect appears to dominate. The other potential reason for disagreement in these results is that misdemeanor and felony convictions are incarcerated in different facilities, of which felony facilities (prisons) generally have much harsher conditions (e.g., lack of air conditioning, less personal freedom, higher rates of violence). As such, there is likely a stronger deterrence effect for felony offenses. Further work exploring heterogeneity in response for first-time compared to repeat offenders is needed to help determine the extent of the deterrence effect.[24]

5. Conclusion

In this paper, we have reviewed a growing literature focused on evaluating social programs with randomized administrative screeners. We have considered current methodologies and provided new intuition (based on economic modeling) on the source of identification and the restrictive impliciations of current methods. We have suggested an alternative framework for estimating program impacts that produce estimators which are theoretically less biased and more precise. The theoretical results are illustratedd with simulation exercises.

The methodologies are applied to study the impact of criminal justice polices on recidivism in the United States. The empirical analysis is performed on previously unstudied dataset, which was the result of an original data collection effort by the authors. Empirical estimates show divergent trends between standard IV-based procedures and the modified HRIV-based estimators.

Based on the HRIV estimates, we find that there is little evidence of stigmatization for non-incarcerated defendants. Further work is required to distinguish between the magnitude of incapcitation, deterence and atrophy of human/social capital channels.

[24]To explore this issue in greater detail, we could also apply to merge our court records with the Harris County prison records to observe exactly which institutions prisoners end up at.

TABLE 8. Recidivism Results at 2 Years, Misdemeanor Cases

	Charged with New Offense			Found Guilty | Charge = 1		
Model	OLS	IV	HRIV	OLS	IV	HRIV
Verdict = Guilty	-0.0361***	-0.0156	-0.0457	0.0680***	0.801***	0.323***
	(0.00294)	(0.135)	(0.0693)	(0.00699)	(0.213)	(0.0926)
Verdict = Def. Adj. of Guilt	-0.0537***	-0.0232	0.00274	0.0427***	0.694***	0.278***
	(0.00344)	(0.146)	(0.0696)	(0.00968)	(0.231)	(0.0982)
Probation Length > 0	0.00325	0.0195	0.152**	-0.0187	-0.107	-0.000903
	(0.00342)	(0.151)	(0.0673)	(0.0141)	(0.241)	(0.101)
Probation Length (Years)	0.00400***	-0.00502	-0.0533***	0.0107	-0.115*	0.0778***
	(0.00106)	(0.0335)	(0.0101)	(0.00659)	(0.0629)	(0.0255)
Incarceration Length > 0	0.0466***	0.101	0.118*	0.0646***	-0.294	0.120
	(0.00276)	(0.134)	(0.0661)	(0.00681)	(0.208)	(0.0869)
Incarceration Length (Years)	0.153***	-0.296***	-0.0334	-0.00735*	-0.0551	-0.103*
	(0.00548)	(0.104)	(0.0593)	(0.00418)	(0.0993)	(0.0583)
Fine Amount > 0	-0.0400***	0.0803***	0.104***	-0.0360***	-0.0314	0.0139
	(0.00108)	(0.0263)	(0.0123)	(0.00178)	(0.0424)	(0.0205)
Fine Amount ($1,000)	-0.00242**	-0.0683***	0.00750	-0.0119***	0.0351	0.00812
	(0.00112)	(0.0130)	(0.00788)	(0.00399)	(0.0260)	(0.0140)
Observations	1,550,354	1,550,354	1,550,354	475,245	475,245	475,245

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

TABLE 9. Recidivism Results at 10 Years, Misdemeanor Cases

	Charged with New Offense			Found Guilty | Charge = 1		
Model	OLS	IV	HRIV	OLS	IV	HRIV
Verdict = Guilty	0.0190***	0.173	0.00995	0.0423***	0.386*	0.367***
	(0.00515)	(0.190)	(0.120)	(0.00726)	(0.209)	(0.104)
Verdict = Def. Adj. of Guilt	1.80e-05	0.128	0.171	0.0296***	0.296	0.355***
	(0.00561)	(0.205)	(0.121)	(0.00781)	(0.225)	(0.107)
Probation Length > 0	-0.0264***	-0.0708	-0.206*	0.0283***	0.0447	-0.138
	(0.00546)	(0.207)	(0.120)	(0.00785)	(0.224)	(0.107)
Probation Length (Years)	0.00278**	-0.00213	0.116***	0.00467***	-0.0634	0.0356**
	(0.00109)	(0.0363)	(0.0134)	(0.00165)	(0.0403)	(0.0162)
Incarceration Length > 0	0.0241***	-0.0290	-0.00588	0.0643***	-0.0801	-0.106
	(0.00498)	(0.190)	(0.120)	(0.00711)	(0.205)	(0.103)
Incarceration Length (Years)	0.0597***	-0.215*	0.400***	-0.0151***	-0.0688	0.0482
	(0.00580)	(0.124)	(0.0786)	(0.00383)	(0.0852)	(0.0549)
Fine Amount > 0	-0.0302***	0.0319	0.122***	-0.0212***	-0.0416	-0.0502***
	(0.00145)	(0.0356)	(0.0166)	(0.00143)	(0.0355)	(0.0165)
Fine Amount ($1,000)	-0.0126***	-0.102***	-0.0745***	-0.0158***	0.0366	0.00576
	(0.00246)	(0.0275)	(0.0162)	(0.00362)	(0.0331)	(0.0169)
Observations	1,037,241	1,037,241	1,037,241	490,657	490,657	490,657

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

TABLE 10. Recidivism Results at 2 Years, Felony Cases

	Charged with New Offense			Found Guilty | Charge = 1		
Model	OLS	IV	HRIV	OLS	IV	HRIV
Verdict = Guilty	-0.0789***	-0.155***	-0.184***	0.0391***	0.361***	0.276***
	(0.00234)	(0.0367)	(0.0245)	(0.00317)	(0.0407)	(0.0237)
Verdict = Def. Adj. of Guilt	-0.0369***	-0.135**	-0.120***	0.0631***	0.314***	0.350***
	(0.00317)	(0.0536)	(0.0354)	(0.00595)	(0.0736)	(0.0462)
Probation Length > 0	-0.159***	0.00621	-0.0381	-0.113***	0.0146	-0.0814
	(0.00303)	(0.0502)	(0.0369)	(0.00653)	(0.0798)	(0.0500)
Probation Length (Years)	0.00837***	-0.00974**	0.00268	0.00375***	0.00354	6.84e-05
	(0.000351)	(0.00447)	(0.00324)	(0.000724)	(0.00829)	(0.00521)
Incarceration Length > 0	-0.0282***	0.00276	0.0770***	0.00761**	-0.0732**	-0.0381*
	(0.00218)	(0.0283)	(0.0225)	(0.00297)	(0.0327)	(0.0225)
Incarceration Length (Years)	-0.00389***	-0.0114***	-0.0186***	-0.00302***	0.00767**	0.000601
	(9.06e-05)	(0.00276)	(0.00104)	(0.000147)	(0.00298)	(0.00133)
Fine Amount > 0	-0.0383***	0.0290	-0.0215	-0.0469***	-0.118**	-0.0868**
	(0.00183)	(0.0227)	(0.0156)	(0.00470)	(0.0590)	(0.0365)
Fine Amount ($1,000)	-0.000608***	-0.0249	0.0492***	-0.00293	0.104	0.0298
	(0.000213)	(0.0222)	(0.0124)	(0.00351)	(0.0664)	(0.0354)
Observations	827,100	827,100	827,100	268,426	268,426	268,426

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

TABLE 11. Recidivism Results at 10 Years, Felony Cases

	Charged with New Offense			Found Guilty | Charge = 1		
Model	OLS	IV	HRIV	OLS	IV	HRIV
Verdict = Guilty	-0.0126***	-0.0887**	-0.121***	0.0499***	0.284***	0.193***
	(0.00257)	(0.0435)	(0.0293)	(0.00229)	(0.0344)	(0.0211)
Verdict = Def. Adj. of Guilt	0.0471***	-0.0823	0.0495	0.0632***	0.288***	0.267***
	(0.00390)	(0.0618)	(0.0419)	(0.00428)	(0.0538)	(0.0349)
Probation Length > 0	-0.181***	0.0989	-0.120***	-0.0438***	0.0331	-0.0174
	(0.00379)	(0.0625)	(0.0455)	(0.00451)	(0.0570)	(0.0376)
Probation Length (Years)	0.00302***	-0.00960	0.00281	-0.00120**	-0.00744	-0.00503
	(0.000443)	(0.00590)	(0.00452)	(0.000572)	(0.00618)	(0.00428)
Incarceration Length > 0	0.00600**	0.0902***	0.0840***	0.0131***	-0.0458*	-0.0229
	(0.00240)	(0.0324)	(0.0248)	(0.00192)	(0.0237)	(0.0172)
Incarceration Length (Years)	-0.00751***	-0.0165***	-0.0213***	-0.00310***	0.00723**	-0.000386
	(0.000113)	(0.00339)	(0.00148)	(0.000134)	(0.00286)	(0.00124)
Fine Amount > 0	-0.0397***	-0.0277	-0.0754***	-0.0167***	-0.0318	-0.0629**
	(0.00258)	(0.0316)	(0.0229)	(0.00327)	(0.0446)	(0.0280)
Fine Amount ($1,000)	-0.00115***	-0.0282	0.00389	-0.00661***	-0.0351	0.0182
	(0.000365)	(0.0243)	(0.0122)	(0.00193)	(0.0433)	(0.0230)
Observations	518,346	518,346	518,346	263,352	263,352	263,352

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

References

D. S. Abrams, M. Bertrand, and S. Mullainathan. Do Judges Vary in Their Treatment of Race? *Journal of Legal Studies*, 2010.

A. Aizer and J. Doyle. Juvenile incarceration and adult outcomes: Evidence from randomly-assigned judges. *NBER Working Paper*, 2011.

J. Angrist, G. Imbens, and D. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

D. Autor, S. N. Houseman, and S. P. Kerr. The Effect of Work First Job Placements on the Distribution of Earnings: An Instrumental Variable Quantile Regression Approach. *NBER Working Paper*, 2012.

D. H. Autor and S. N. Houseman. Do temporary-help jobs improve labor market outcomes for low-skilled workers: Evidence from "work first". *American Economic Journal: Applied Economics*, 2(3):96–128, 2010.

J. Bai and S. Ng. Instrumental variable estimation in a data rich environment. *Econometric Theory*, 26(6):1577–1606, 2010.

P. Bayer, R. Hjalmarsson, and D. Pozen. Building criminal capital behind bars: Peer effects in juvenile corrections. *Quarterly Journal of Economics*, 124(1):105–147, 2009.

P. Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, pages 657–681, 1994.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *unpublished manuscript*, 2011.

S. Bright. Elected judges and the death penalty in texas: Why full habeas corpus review by independent federal judges is indispensable to protecting constitutional rights. *Tex. L. Rev.*, 78:1805, 2000.

J. Chao and N. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005.

S. Donald and W. Newey. Choosing the number of instruments. *Econometrica*, 69(5): 1161–1191, 2001.

J. Doyle. Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, 97(5):1583–1610, 2007.

J. Doyle. Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *Journal of Political Economy*, 116(4):746–770, 2008.

J. Doyle, J. Graves, J. Gruber, and S. Kleiner. Do high-cost hospitals deliver better care? evidence from ambulance referral patterns. *NBER Working Paper*, 2012.

C. Hansen, J. Hausman, and W. Newey. Estimation with many instrumental variables. *Journal of Business and Economic Statistics*, 26(4):398–422, 2008.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2009.

J. Heckman and E. Vytlacil. Structural equations, treatment effects and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005.

L. Katz, S. Levitt, and E. Shustorovich. Prison conditions, capital punishment, and deterrence. *American Law and Economics Review*, 5(2):318–343, 2003.

J. Kling. Incarceration length, employment and earnings. *American Economic Review*, 96 (3):863–876, 2006.

E. Korn and S. Baumrind. Clinician preferences and the estimation of causal treatment differences. *Statistical Science*, 13(3):209–235, 1998.

E. Korn, D. Teeter, and S. Baumrind. Using explicit clinician preferences in nonrandomized study designs. *Journal of statistical planning and inference*, 96(1):67–82, 2001.

I. Kuziemko. Should prisoners be released via rules or discretion? *NBER Working Paper*, 2011.

D. Lee and J. McCrary. The deterrence effect of prison: Dynamic theory and evidence. *unpublished manuscript*, 2009.

S. Levitt and I. Kuziemko. An empirical analysis of imprisoning drug offenders. *Journal of Public Economics*, 88(9-10):2043–2066, 2004.

J. Ludwig and J. Kling. Is crime contagious? *Journal of Law and Economics*, 50(3): 491–518, August 2007.

J. McCrary and S. Sanga. Youth offenders and the deterrence effect of prison. *unpublished manuscript*, 2012.

M. Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71 (4):1027–1048, 2003.

D. Munroe and L. Wilse-Samson. Measuring local externalities of home foreclosures. *unpublished manuscript*, 2012.

D. Nagin and M. G. Snodgrass. The effect of incarceration on offending: Evidence from a natural experiment in pennsylvania. *unpublished manuscript*, 2011.

S. Ng. Variable selection in predictive regressions. *unpublished manuscript*, 2011.

R. Okui. Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics*, 2011.

E. Owens. More time, less crime? estimating the incapacitative effect of sentence enhancements. *Journal of Law and Economics*, 52(3):551–579, 2009.

J. Price and J. Wolfers. Racial Discrimination Among NBA Referees. *Quarterly Journal of Economics*, 125(4):1859–1887, 2010.

D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, May 1997.

J. Stock, J. Wright, and M. Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of business and economic statistics*, 20(4): 518–529, 2002.

J. Waldfogel. Does inter-judge disparity justify empirically based sentencing guidelines? *International Review of Law and Economics*, 18(3):293–304, 1998.

R. Walmsley. *World Prison Population List (8th edition)*. International Centre for Prison Studies, 2009.

C. Wildeman. Incarceration and population health in wealthy democracies. *unpublished manuscript*, 2011.

Appendix A. Simulation Exercises

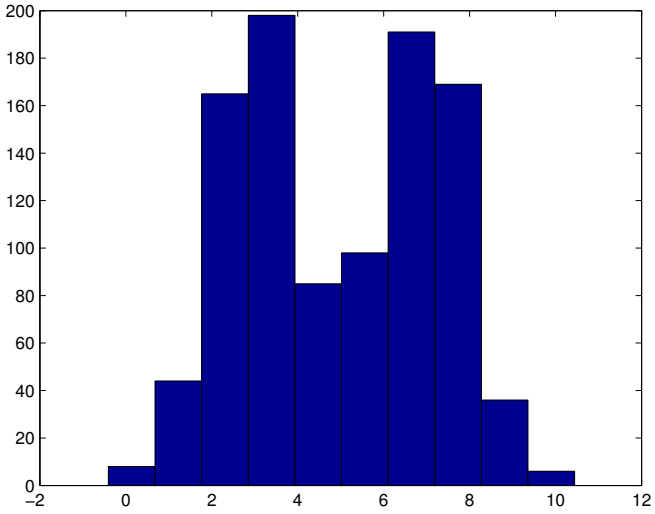
A.1. **Bias Reduction.** This simulation exercise will demonstrate robustness of $LATE_{HRIV}$ compared to $LATE_{IV}$ when the monotonicity fails and bias occurs. Consider the following functional form:

$$\begin{aligned} Y_i &= X_i * B + D_i * \zeta_i + e_i \\ D_i &= \left\{ X_i * b + \sum_{z=\{0,1\}} 1[Z_i = z] * \gamma_z(X_i) + \nu_i > 0 \right\} \end{aligned}$$

$$Z_i \in \{0,1\} \; ; \; X_i = [1, X_{i,1}] \; ; \; X_{i,1} \in \{0,1\} \; ; \; \nu_i \sim N(0,1) \; ; \; e_i \sim N(\nu_i, 1)$$

We draw a distribution of treatment effects for the same using $\zeta_i = 3 + 4 * X_{i,1} + \epsilon_i$ to reflect heterogenous treatment effects:

Figure 6. Empirical Distribution of Treatment Effects



We create a failure in monotonicity through imposing:

$$\begin{aligned} \gamma_0(X_i) &= 1 - 1X_{i,1} \\ \gamma_1(X_i) &= -0.25 + 0.75X_{i,1} \end{aligned}$$

This formulation results in the following program enrollment rates:

% [D=1]	$X_1 = 0$	$X_1 = 1$	Total
Z = 0	94%	60%	76%
Z = 1	42%	74%	58%
Difference	-52%	14%	-19%

If we use overall assignment propensities, Screener 0 has a higher enrollment rate than Screener 1. But, when we delve into subsets of the caseload, we observe that this relationship reverses for individuals with $X_i = 1$.

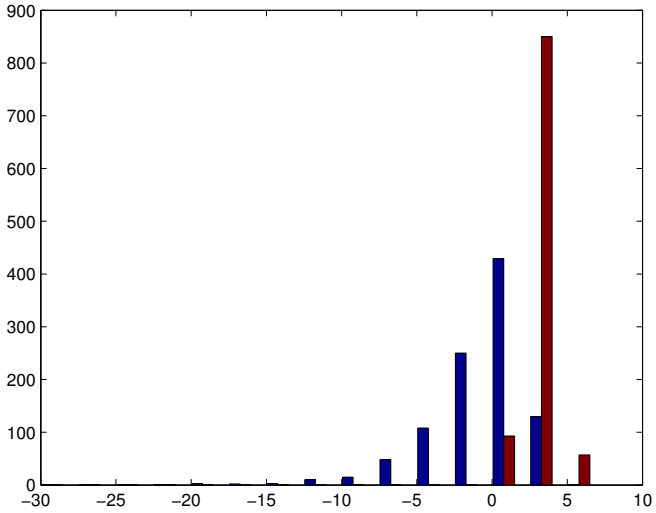
Using this sample, I will estimate the following two estimators:

$$\begin{aligned}\hat{\zeta}_{IV} &= \left(\begin{bmatrix} Z & X \end{bmatrix}' \begin{bmatrix} D & X \end{bmatrix}\right)^{-1} \left(\begin{bmatrix} Z & X \end{bmatrix}' Y\right) \\ \hat{\zeta}_{HRIV} &= \left(\begin{bmatrix} Z \otimes f(X)' & X \end{bmatrix}' \begin{bmatrix} D & X \end{bmatrix}\right)^{-1} \left(\begin{bmatrix} Z \otimes f(X)' & X \end{bmatrix}' Y\right)\end{aligned}$$

Where $f(x)'$ is simply just $[1\ X]$ since X only contains a constant and a binary dummy variable.

Figure 7 plots the estimated coefficents from 1,000 replications. The blue bars show $\hat{\zeta}_{IV}$ and the red bars reflect $\hat{\zeta}_{HRIV}$. Note that despite the true distribution of ζ_i being strictly

FIGURE 7. Bias Reduction: Distribution of $\hat{\zeta}_{IV}$ (Blue) and $\hat{\zeta}_{HRIV}$ (Red)



positive, the majority of $\hat{\zeta}_{IV}$ estimates are very close to zero or negative. The HRIV-based estimator however avoids this bias.

A.2. **Efficiency Gain.** In the second example, we maintain the same distribution of ζ_i as shown in Figure 6. However, we ensure that monotonicity is maintained through defining γ_z as follows:

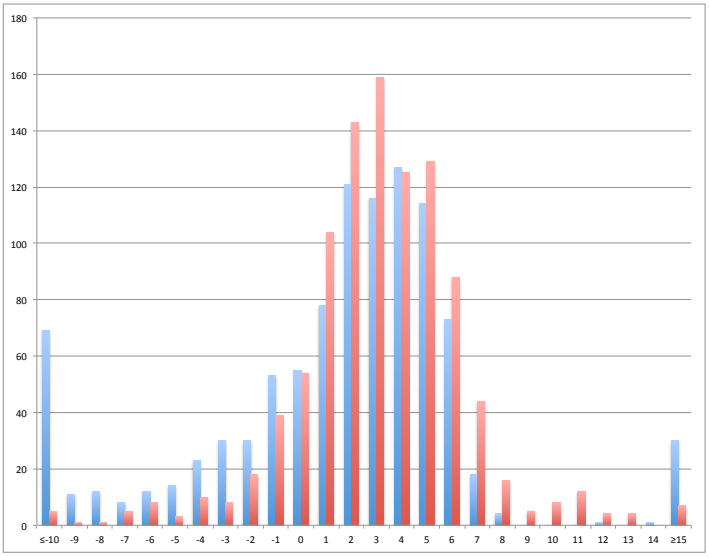
$$\begin{aligned}\gamma_0(X_i) &= 0.2 - 0.2X_{i,1} \\ \gamma_1(X_i) &= -0.1 + 0.1X_{i,1}\end{aligned}$$

As a result of this specification, we still observe an overall higher enrollment rate for Screener 0 compared to Screener 1. What is different from the first example is that Screener

0 and Screener 1 treat individuals with $X_{i,1} = 1$ equally. Instead, all of the variation stems from differential treatment of the subpopulation with $X_{i,1} = 0$. Adding the interaction term to the first stage will help zero in on this source of variation and improve our precision in the $\hat{\zeta}_{HRIV}$ estimator.

Figure 8 shows the results of 1,000 replications. Here, $\hat{\zeta}_{IV}$ in blue has wider dispersion and particularly large tails compared to $\hat{\zeta}_{HRIV}$ in red.

FIGURE 8. Efficiency Gain: Distribution of $\hat{\zeta}_{IV}$ (Blue) and $\hat{\zeta}_{HRIV}$ (Red)



A.3. **Efficiency Loss.** In the third and final example, we expand the dimension of covariates such that $dim(X_i) = 8$, where each element of X_i is an orthogonal binary variable. In contrast to examples 1 and 2, the true data generating process is a standard IV model where $\gamma_z(X_i) = \gamma_z$. Specifically, $\gamma_0 = 0$ and $\gamma_1 = 1$. We estimate $\hat{\zeta}_{IV}$ and $\hat{\zeta}_{HRIV}$ in 1,000 replications where $\hat{\zeta}_{HRIV}$ utilizes interactions between Z and every possible combination of X_i as instruments. This introduces hundreds of instruments that have zero correlation to program assignment. To account for the additional covariates, a new distribution of ζ_i is drawn for the population based on all 8 covariates (X_i) and their interactions:

This simulation exercise demonstrate the inefficiency of estimating LATE's from a HRIV model with a moderate number of covariates, when a standard IV model is accurate. Figure 10 shows wider dispension in the $\hat{\zeta}_{HRIV}$ estimator (in red) compared to $\hat{\zeta}_{IV}$ (in blue).

FIGURE 9. Empirical Distribution of Treatment Effects

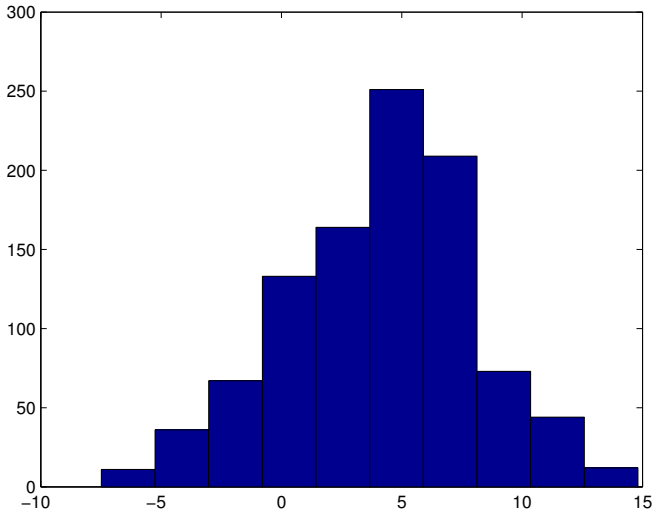


FIGURE 10. Efficiency Loss: Distribution of $\hat{\zeta}_{IV}$ (Blue) and $\hat{\zeta}_{HRIV}$ (Red)