Detecting Grouped Local Average Treatment Effects and Selecting True Instruments

Nicolas Apfel[1], Helmut Farbmacher[2], Rebecca Groh[2], Martin Huber[3], and Henrika Langen[3]

[1] *University of Surrey, UK*
[2] *TU Munich, Germany*
[3] *University of Fribourg, Switzerland**

July 12, 2022

Abstract

In the context of an endogenous binary treatment with heterogeneous effects and multiple instruments, we propose a two-step procedure to identify complier groups with identical local average treatment effects (LATE), despite relying on distinct instruments and even if several instruments violate the identifying assumptions. Our procedure is based on the fact that the LATE is homogeneous for any two or multiple instruments which (i) satisfy the LATE assumptions (instrument validity and treatment monotonicity in the instrument) and (ii) generate identical complier groups in terms of treatment propensities given the respective instruments. Under the (plurality) assumption that for each set of instruments with identical treatment propensities, those instruments satisfying the LATE assumptions constitute the relative majority, our procedure permits identifying these true instruments in a data driven way. We also provide a simulation study investigating the finite sample properties of our approach and an empirical application investigating the effect of incarceration on recidivism in the US with judge assignments serving as instruments.

*Corresponding address: n.apfel@surrey.ac.uk. Nicolas Apfel gratefully acknowledges funding through ESRC grant EST013567/1.

1 Introduction

In 2019, almost 1 % of the adult population in the US was incarcerated (Loeffler and Nagin, 2021). Incarceration rates have been soaring since the 1970s in the US and other countries have seen similar increases. The question of whether prisons are fulfilling their function is therefore pertinent. Specifically, does the experience of going through prison decrease the probability of inmates to reoffend? Many studies in the economics of crime have investigated this question and tend to find zero or crime-inducing effects. However, the effects might vary with geography and the institutional context (e.g. conditions under imprisonment), as Bhuller, Dahl, Løken, and Mogstad (2020) find crime-reducing effects of imprisonment in Norway. Moreover, the effect might also vary inside each country, for example by court.

Starting with the seminal paper by Kling (2006), a commonly used estimation strategy in this literature is to use an instrumental variable (IV) which is correlated with incarceration but not with the outcome directly. This IV is typically built on the stringency of each judge. Judge dummies can be used as IVs in an overidentified model to identify the effect of incarceration on reoffending. Each IV, defined by a pair of judges who differ in terms of stringency, might in fact estimate a different, so-called local average treatment effect (LATE). The LATE corresponds to the average effect of incarceration on recidivism among those offenders who comply with the IV in the sense that they are incarcerated under the more stringent judge, but not under the less stringent judge when considering a specific pair of judges as IV. As these LATEs may vary because they refer to different complier groups who are in general distinct in terms of unobservables, a conventional 2SLS based on using all IVs simultaneously might hide interesting heterogeneity in the effect of incarceration. As an empirical contribution, this study investigates effect heterogeneity in the LATE depending on whether offenders are imprisoned under more or less stringent judges and thus likely to commit relatively less or more severe offenses.

However, a general concern with IVs is that they might not be valid, i.e. by directly affecting the outcome, thus violating the so-called exclusion restriction, or through an association with unobserved characteristics affecting the outcome, thus violating IV exogeneity. In our context, judges could e.g. violate the exclusion restriction by varying in terms of the use of alternative measures to incarceration, such as electronic monitoring (Loeffler and Nagin, 2021). For the case that at least a subset of IVs is valid, an emerging literature in biostatistics has established methods which can consistently select valid and invalid IVs (Kang, Zhang, Cai, and Small, 2016; Windmeijer, Farbmacher, Davies, and Smith, 2019; Guo, Kang, Cai, and Small, 2018; Windmeijer, Liang, Hartwig, and Bowden, 2021; Apfel and Liang, 2021). A drawback of previously available IV selection approaches is that they impose homogeneous treatment effects and thus LATEs and can for this reason not distinguish violations of the exclusion restriction from effect heterogeneity. Furthermore, the methods do generally not allow for simultaneous violations of the exclusion restriction and IV exogeneity. Finally, they do not consider the possibility of heterogeneous first stage effects of the instrument on the treatment, a further issue which may jeopardize the identifiability of causal effects based on instruments. For this reason, this study proposes a method for detecting appropriate IVs in heterogeneous treatment effect models, under the conditions that (i) there exist clusters of IVs (i.e. pairs of judges) which generate identical complier groups and (ii) a plurality of IVs within such clusters satisfies the IV assumptions.

More concisely, our methodological contribution to the literature on valid IV selection under heterogeneous treatment effects consists of two steps. First, we make use of Agglomerative Hierarchical Clustering (Ward, 1963) to find clusters (which we call *clubs*) of IVs with the very same treatment probabilities, henceforth referred to as propensity scores, in order to identify complier group effects when pairing these clubs. The underlying intuition is that if two distinct pairs of judges entail the same higher and lower propensity scores (associated with the respective more and less stringent judges in each pair), then the complier groups in either pair must be identical, if the LATE assumptions outlined in Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) are satisfied. Therefore, our approach permits clustering IVs into clubs with homogeneous LATEs given that the LATE assumptions hold, namely IV validity as well as treatment monotonicity in terms of the first stage effect, implying that any offender incarcerated under a less stringent judge would also be incarcerated under a more stringent judge. We stress that the key assumption underlying this step is that pairs of judge-related propensity scores defining the first-stage effect of the instrument on the treatment are indeed clustered in the population.

Second, within a specific club, we distinguish valid from invalid IVs based on our second key assumption that the largest group of IVs inside each club indeed satisfies the LATE assumptions and thus, entails the same LATE, which is known as plurality assumption. To this end, we leverage recent advances in the selection of valid IVs in a setting with constant treatment effects (Windmeijer, Farbmacher, Davies, and Smith, 2019; Windmeijer, Liang, Hartwig, and Bowden, 2021; Apfel and Liang, 2021). Our two-step procedure broadly fits into the fast evolving literature on the integration of machine learning methods in causal inference. In our context, we fruitfully apply both unsupervised and supervised machine learning for clustering compliers based on the IVs and selecting IVs that satisfy the identifying assumptions, respectively.

In the empirical application, we apply our method to a newly collected data set on minor crimes in Minnesota for the years 2009 to 2017 to estimate the effect of incarceration on recidivism in the US. To benchmark our results against conventional approaches, in a first step we consider ordinary least squares (OLS) regression, which suggests crime-inducing effects, as well as linear IV regression based on two stage least squares (2SLS) when simultaneously relying on all judge IVs, which also yields crime-inducing effects. However, when using our method, we find that LATEs estimated by club-pairs are considerably lower and higher and statistically significant. This finding points to substantial effect heterogeneity of the effects of imprisonment on recidivism across stringency levels of judges, which is masked by 2SLS. When estimate club-pair specific LATEs and splitting into judge-specific IVs with one club as the reference category, we can test overidentifying restrictions and mostly reject the overall validity of IVs. Therefore, we use a second-step selection to weed out invalid IVs from the reduced form estimates inside each club. Doing this, we find that most club-pair specific estimates are even more heterogeneous than after the first-step selection. After the second-step selection, the tests of overidentifying restrictions do not reject anymore. In addition to the empirical application, we also consider the finite sample performance of our method in a simulation study and find that indeed the first-step selection delivers consistent classification of the propensity scores and the second-step selection consistently selects the IVs which violate the exclusion restriction. All of this also holds when the outcome equation is non-linear. We illustrate how in presence of invalid IVs without the second-step selection the club-pair specific estimates are in line with the true (oracle) LATEs we generated and without it they do not cluster at the true LATEs.

The remainder of this study is organized as follows. In Section 2, we introduce the treatment effects model, the LATE, as well as the identifying assumptions and demonstrate that two pairs of judges with identical propensity scores have the same LATE under these assumptions. Section 3 introduces our two-step procedure of (i) detecting clusters of IVs (i.e. pairs of judges) with comparable lower and higher propensity scores and (ii) selecting the IVs not violating the LATE assumptions, under the condition that the relative majority of IVs satisfies the assumptions (plurality). In Section 4, we provide a simulation study which points to a very decent finite sample performance of our method in terms of detecting clusters (or clubs) of compliers and appropriately selecting cluster-specific instruments which satisfy the LATE assumtpions. Section 5 presents our empirical application considering the effect of incarceration on recidivism in the US. Section 6 concludes.

2 Model and Assumptions

2.1 Notation and LATE assumptions

For modelling the causal effects of interest and discussing the assumptions required for their identification, we make use of the potential outcome framework, see for instance Neyman (1923) and Rubin (1974). To this end, we denote by D a binary treatment, e.g. incarceration, which can take values $d \in \{0, 1\}$. In general, we will refer to random variables by capital letters and specific values thereof by lower case letters. We are interested in the causal effect of the treatment on a discrete or continuous outcome Y, in our empirical problem on a binary indicator on recidivism. Furthermore, we denote by Z an instrumental variable (IV), which needs to satisfy certain conditions outlined below, which is assumably multivalued and discrete. That is, Z may take values $z \in \{1, ..., J\}$, with J corresponding to the total number of values of the instrument. In our empirical application, the instrument values indicate the assignment to alternative judges, which differ in terms of incarceration rates, i.e. the likelihood to issue prison sentences. Finally, $D(z)$ corresponds to the potential treatment state when setting the IV to a hypothetical value $Z = z$ in the support of Z, while $Y(d, z)$ is the potential outcome under treatment value d (either 0 or 1) and IV value z.

Following Angrist, Imbens, and Rubin (1996), individuals can be classified into compliance types as function of how someone's treatment state depends on or reacts to switching the value of the instrument. Let us to this end consider two distinct IV values z and z', which can assumably be ordered in a sensible way, e.g. two judges with a higher and lower incarceration rate, respectively. Individuals satisfying $(D(z) = 1, D(z') = 0)$ are compliers in the sense that they are treated, i.e. incarcerated under the higher value of the IV $(Z = z)$, i.e. when being assigned to a more stringent judge, while remaining non-treated under the lower value $(Z = z')$ when being assigned to a more lenient judge. Never takers neither receive the treatment under a higher, nor under a lower IV value, thus satisfying $(D(z) = D(z') = 0)$, while always takers are treated under either IV value, i.e. $(D(z) = D(z') = 1)$. Finally, defiers counteract the instrument assignment by being treated under the lower IV value implying a more lenient judge, while remaining untreated under the higher IV value implying a stricter judge, such that $(D(z) = 0, D(z') = 1)$ holds. It is important to note that different choices of values z and z' generally entail different proportions of complier types, see e.g. the discussion in Frölich (2007). For instance, the share of compliers is likely larger when considering two judges that greatly differ in terms of their stringency rather than two judges among which one is only slightly more stringent than the other.

Imbens and Angrist (1994) discuss assumptions under which an IV-based evaluation approach non-parametrically identifies a local average treatment effect (LATE) on the compliers under heterogeneous treatment effects. For two instrument values $z > z'$, this LATE is formally defined as

$$\Delta_{z,z'} = E[Y(1) - Y(0)|D(z) = 1, D(z') = 0]. \tag{1}$$

In terms of assumptions, the instrument must be as good as randomly assigned, which rules out statistical associations with background characteristics affecting the outcome, and must not have a direct effect on the outcome other than through the treatment. Furthermore, defiers must not exist, while compliers must exist in the population. We subsequently formally state these IV assumptions, which permit recovering the LATE for a specific pair of instrument values z and z'.

Assumption 1. *Exogeneity*

(i) $Z \perp (D(z), Y(z', d))$ *and (ii)* $Y(z, d) = Y(z', d) = Y(d)$.

Assumption 2. *Monotonicity*

$$\Pr(D(z) \geq D(z')) = 1.$$

Assumption 3. *First Stage*

$$E(D|Z = z) - E(D|Z = z') \neq 0.$$

Assumption 1, which is frequently referred to as IV exogeneity, consists of two conditions. The first one states that the IV values z and z' are as good as randomly assigned, i.e. independent of the potential treatments as well as the potential outcomes. This rules out confounders jointly affecting Z on the one hand and D and/or Y on the other hand. The second condition states that the instrument constructed from values z and z' does not affect the outcome conditional on the treatment, implying that Z does not have a direct effect on Y other than through D, which is known as the exclusion restriction. For this reason, we define the potential outcome as function of the treatment only (rather than also the instrument) as long as we assume no violation of the exclusion restriction, i.e. $Y(d)$. Assumption 2 requires that the potential treatment state can never decrease when shifting the instrument from z' to z and for this reason rules out the existence of defiers. Assumption 3 imposes a non-zero (first stage) effect of the instrument shift on the treatment. Together with Assumption 2, this necessarily implies the existence of compliers.

Under Assumptions 1 to 3, the LATE on compliers who are responsive to IV shifts from z' to z corresponds to a so-called Wald (1940)-estimand. The latter consists of the (reduced form) average effect of the instrument shift on the outcome scaled by the (first stage) effect of the instrument shift on the treatment:

$$\frac{E(Y|Z = z) - E(Y|Z = z')}{\Pr(D = 1|Z = z) - \Pr(D = 1|Z = z')} \tag{2}$$

For the case of a discretely distributed IV with mass points at z and z', (2) is equivalent to the probability limit of a two stage least squares regression (TSLS) in which only observations with $Z \in \{z, z'\}$ are considered. It is also worth noting that $\Pr(D = 1|Z = z) - \Pr(D = 1|Z = z')$ identifies the complier share $\Pr(D(z) = 1, D(z') = 0)$.

As an important matter of fact for our method suggested below, Vytlacil (2002) proves that Assumption 2 is equivalent to imposing a so-called threshold-crossing model on the treatment. In this model, the treatment is an additively separable function of the instrument and an unobserved term and takes the value one whenever the function on the instrument is larger than or equal to the unobserved term. Formally,

$$D = 1(\psi(Z) > V), \tag{3}$$

where $1(\cdot)$ denotes the indicator function, which is equal to one if its argument is satisfied and zero otherwise, V is a scalar (index of) unobservable(s), and $\psi(Z)$ is a nonparametric function of Z. This allows us to characterize the compliers (and other compliance types) in terms of the distribution of V. Because $D(z) = 1$ implies that $\psi(z) > V$ in (3) and $D(z') = 0$ implies that $\psi(z') \leq V$, it is easy to see that the distribution of V among compliers satisfies $\psi(z) > V \geq \psi(z')$. For this reason, it follows that

$$\Delta_{z,z'} = E[Y(1) - Y(0)|v > V \geq v'], \tag{4}$$

for values $v = \psi(z)$ and $v' = \psi(z')$ for the unobservable V.

Let us now assume that there exists another pair of instrument values z^*, z'', in our case two further judges, which also satisfies Assumptions 1 to 3 and generates exactly the same compliance behaviour as the previous pair z, z'. Formally, $v = \psi(z^*) = \psi(z)$ and $v' = \psi(z') = \psi(z'')$. It follows that

$$\Delta_{z^*,z''} = E[Y(1) - Y(0)|v > V \geq v'] = \Delta_{z,z'} \tag{5}$$

As both pairs of instruments refer to the very same complier group in terms of the distribution of unobservables V, the LATE identified by these pairs is identical. This in turn implies that the Wald estimand $\frac{E(Y|Z=z^*)-E(Y|Z=z'')}{\Pr(D=1|Z=z^*)-\Pr(D=1|Z=z'')}$ is equivalent to that in (2) and that both denominators $\Pr(D = 1|Z = z) - \Pr(D = 1|Z = z')$ and $\Pr(D = 1|Z = z^*) - \Pr(D = 1|Z = z'')$ identify the share of compliers satisfying $\Pr(v > V \geq v')$.

The identification of an identical complier effect under either pair of instrument values is driven by the fact that the satisfaction of $\psi(z) = \psi(z^*)$ for any pair z, z^* implies that $\Pr(D = 1|Z = z) = \Pr(D = 1|Z = z^*)$. This is the case because for any value z, $\Pr(D = 1|Z = z)$ is the probability of the event $(\psi(z) > V)$: $\Pr(D = 1|Z = z) = \Pr(\psi(Z) > V|Z = z) = \Pr(\psi(z) > V)$, where the first equation follows from equation (3) and the second from the independence of $D(z)$ and Z implied by Assumption 1. The converse holds as well, i.e. $\Pr(D = 1|Z = z) = \Pr(D = 1|Z = z^*)$ implies that $\psi(z) = \psi(z^*)$. To see this, let us normalize V such that it only takes values between 0 and 1. That is, rather than considering the unobservable V, we take its cumulative distribution function (cdf), denoted by F_V which is bounded between 0 and 1. Formally, $F_V \sim [0, 1]$. As noticed in the literature on marginal treatment effects, see e.g. Heckman and Vytlacil (2001) and Heckman and Vytlacil (2005), this normalization is innocuous in the sense that it does not affect the results of our treatment model postulated in Equation (3). The latter can without loss of generality be reparametrized in the following way when expressing the treatment as a function of F_V rather than V:

$$D = 1(F_V(\psi(Z)) \geq F_V)), \tag{6}$$

where $F_V(\psi(Z))$ is the cdf of V evaluated at the value $\psi(Z)$. By the definition of a cdf and treatment equation (3), $F_V(\psi(Z)) = \Pr(V \leq \psi(Z)) = \Pr(\psi(Z) > V) = \Pr(D = 1|Z)$, which is the treatment propensity score. For this reason, there exists a one-to-one correspondence between $\psi(z) = \psi(z^*)$ and $\Pr(D = 1|Z = z) = \Pr(D = 1|Z = z^*)$. This implies that distinct pairs of instruments entailing identical pairs of upper and lower propensity scores necessarily identify the LATE for the very same complier group (i.e. with the same distribution of unobserved characteristics), if Assumptions 1 to 3 hold.

2.2 Grouped propensity scores

To ease notation, we subsequently denote the treatment propensity score as a function of the instrument by $p_z = \Pr(D = 1|Z = z) = E(D|Z = z)$, where the second equality holds because D is binary. In a next step, we impose that propensity scores follow a specific cluster structure, such that the propensity scores within a cluster are homogeneous, even across different instruments (e.g. judges). This assumption is crucial in our context. To this end, we introduce some further notation. Let C_k denote some set or cluster k of instruments and p_k^0 the constant propensity score of all instruments in this cluster. K is the total number of clusters of IVs with the same propensity score. Furthermore, let $\mathbb{1}(\cdot)$ denote the indicator function, which is equal to one if its argument is satisfied and zero otherwise. Formally, we make the following cluster assumption, in analogy to Su, Shi, and Phillips (2016) (who look at panel data models more generally):

Assumption 4. *Clubs*

$$p_z = \sum_{k=1}^{K} p_k^0 \mathbb{1}(z \in C_k).$$

with cluster C_k such that $p_k^0 \neq p_{k'}^0$ for any $k \neq k'$. The clusters do not overlap and the union of all clusters is the full set of IVs, $C_k \bigcap C_k' = \emptyset$ for any $k \neq k'$ and $\bigcup\limits_{k=1}^{K} C_k = \{1, ..., J\}$.

Assumption 4 implies that there exist sets of z with the very same propensity score, which we call *clubs*, in line with the growth literature where this type of method is usually applied, where k defines the "club identity". This assumption seems restrictive at first, but as we will see soon, it is not particularly controversial as there is a way to verify it with the data. The prevalence of several clubs permits forming pairs of clubs with distinct propensity scores p_k^0 and $p_{k'}^0$ in order to construct IVs with a non-zero first stage for LATE evaluation. In fact, IV methods relying on any judge in club k and any judge in a different club k' identify the very same LATE, if the identifying assumptions outlined in Section 2.1 are satisfied, which follows from the discussion at the end of Section 2.1. The following theorem states the identification result for the LATE, which follows by equation 5.

Theorem 1. *Under Assumptions 1 to 4, IVs in the same club-pair identify the same LATE.*

2.3 Violations of the LATE Assumptions

In a next step, we will allow for violations of the LATE Assumptions. To illustrate, we consider direct effects of the judges. Our argument also holds for other violations of the LATE Assumptions.

$$Y = \eta(D, Z_j, U) \tag{7}$$

which means that a judge j has a direct effect on the outcome (or also through an unobservable, where U is a function of Z).

However, in many empirical applications, the assumption that all IVs satisfy the LATE assumptions might be challenged. Let us for instance assume that a pair of judges violates the exclusion restriction postulated in Assumption 1. In this case, one judge directly affects recidivism relative to another one, for instance by making use of distinct alternative measures to imprisonment, such as electronic monitoring. This generally implies for a pair of judges z, z' that the average direct effect of the instrument conditional on the compliance type $(D(z) = d, D(z') = d')$, denoted by $\gamma_{zz'}^{dd'}$, is non-zero:

$$\gamma_{zz'}^{dd'} := E[Y(d)|D(z) = d, D(z') = d', Z = z] - E[Y(d)|D(z) = d, D(z') = d', Z = z'] \neq 0 \text{ for } d, d' \in \{0, 1\}. \tag{8}$$

In this case, the Wald estimand defined in (2) no longer identifies the LATE but is biased, as for instance discussed in Huber (2014), who demonstrates that under a violation of the exclusion restriction, the LATE corresponds to

$$\Delta_{z,z'} = \frac{E(Y|Z = z) - E(Y|Z = z') - p_{z'} \cdot \gamma_{zz'}^{11} - (1 - p_z) \cdot \gamma_{zz'}^{00}}{p_z - p_{z'}} - \gamma_{zz'}^{10}. \tag{9}$$

Therefore,

$$\frac{E(Y|Z = z) - E(Y|Z = z')}{p_z - p_{z'}} = \Delta_{z,z'} + \frac{p_{z'} \cdot \gamma_{zz'}^{11} + (1 - p_z) \cdot \gamma_{zz'}^{00}}{p_z - p_{z'}} + \gamma_{zz'}^{10}, \tag{10}$$

where the expression to the right of the LATE $\Delta_{z,z'}$ characterizes the bias of the Wald estimand. Such biases generally also arise in the case of the violation of other LATE assumptions like monotonicity, see Angrist, Imbens, and Rubin (1996).

Given Theorem 1, the LATEs estimated from two clubs should be identical across different pairs from the same two clubs.

$$\begin{aligned} \Delta_{z,z'} = \frac{E(Y|Z=z) - E(Y|Z=z')}{p_z - p_{z'}} = c \\ c \cdot (p_z - p_{z'}) = E(Y|Z=z) - E(Y|Z=z') \end{aligned} \tag{11}$$

Hence, from judge-specific means of the outcome, $E(Y|Z=z)$, that are the same in each club and pairing these up, we will find homogeneous effects in each group-pair. A group of instruments is defined as follows:

In the discussion to follow, we will consider the scenario that some judges within a club k violate some or all of the LATE assumptions, while others satisfy them. Under specific conditions, we can in fact distinguish those judges (values of the IV) satisfying the LATE Assumptions from those violating them within a club k. To this end, we introduce a further definition, namely the IV groups. A group is a set of instrument values (in our case defined by judges) in a club with a homogeneous propensity score. We note that a group is generally a subset of a club. Formally, a group is defined as follows:

Definition 1. *Group*
$\mathcal{G}_g^k = \{z \in C_k : E(Y|Z=z) = \Gamma_g\}$

The superscript k makes explicit that groups are defined within clubs, i.e. in each club there may exist one or several groups. Summing up, a group is a subset of judges with the same reduced form coefficient, Γ_g, where g is shorthand for the presence of violations of the LATE Assumptions and therefore $g = 0$ means that there are no violations. This is in contrast with the valid IV selection literature, where groups are defined in terms of their inconsistency, which is also a function of the first-stage parameter. The difference is due to the fact that we have already pre-selected clubs in which the first-stage parameter is equal and all heterogeneity that is left is due to direct effects. Hence, we do not need to take the first stage into account as would have been the case in the IV selection literature, such as in Kang, Zhang, Cai, and Small (2016), Guo, Kang, Cai, and Small (2018), Windmeijer, Liang, Hartwig, and Bowden (2021) or Apfel and Liang (2021).

To identify those judges that satisfy the LATE Assumptions, Assumptions 1 and 2, we assume that the judges that fulfil LATE assumptions ($g = 0$) constitute the largest group, a plurality, inside of each club. More formally, let us denote by $\mathcal{G}_0^k$ the group of instruments satisfying Assumptions 1 - 2, within club k. Our so-called plurality assumption requires that the cardinality (or number) of IV values satisfying the latter assumptions exceeds the cardinality of any other group of invalid IVs:

Assumption 5. *Inside-Club Plurality*

$$|\mathcal{G}_0^k| > max(\mathcal{G}_g^k) \quad \forall k$$

This implies that within some club k, a plurality of judge-values in $\mathcal{G}_0^k$ fulfil the exclusion restriction. If we can find these maximal groups, it follows from the joint satisfaction of Assumptions 1 to 5 that club-pair specific LATEs are identified even if the LATE assumptions are fulfilled only for a subset of supposed instruments.

3 Empirical method

3.1 Overview

While Section 2 focuses on the identifying assumptions and intuition underlying LATE identification, this section will focus on estimation. We first provide an overview of the various steps and then discuss each of them in more detail. Broadly speaking, we suggest the following two-step procedure clustering procedure for (1) finding clubs and (2) detecting judges which violate exclusion:

Procedure 1. *Detecting grouped LATEs with violations of validity*

1. *Find clubs of propensity scores*
   - *(a) Use clustering to find clubs of propensity scores using estimated propensity scores*
   - *(b) Form all possible pairs of clubs*
   - *(c) Generate IV that compares two clubs (set to zero for club a and to one for club b)*
2. *Select invalid IVs*
   - *(a) Take one club as reference category and build all possible judge-IVs with another club in the pair, then detect heterogeneity via Hansen-Sargan test*
   - *(b) Inside of each club, use clustering to find the largest group of IVs in terms of judge-specific means*
   - *(c) Generate IV that compares the largest groups of judges inside each club*

The general idea of the first-step of our method is to group propensity scores into clubs, pairing these clubs, and applying IV estimation. We can directly use these estimates if we expect the LATE assumptions to hold for all instruments. In the second stage, we actually test whether LATEs are homogeneous for multiple IVs constructed with one club as the reference category and all judge-specific IVs from another club. A rejection of the overidentification test points to some heterogeneity in the judge-club IV estimates. Given that pairs of clubs should identify the same LATE, any further heterogeneity points towards a violation of LATE assumptions. If an absolute majority of IVs clustered in the first step of the procedure assumably satisfies the LATE assumptions, we can consistently estimate the LATE based on the median estimator of Han (2008). If only a relative (rather than absolute) majority of IVs assumably satisfies the LATE assumption, we suggest using machine learning methods for selecting these true instruments, see step 2 c). After this, we again generate a group-pair based IV for each club-pair.

3.2 Club-pair LATE estimation

The first step for estimating LATEs consists of detecting clusters or clubs of treatment propensity scores when invoking Assumption 4. We consider Agglomerative Hierarchical Clustering (AHC) for finding clubs. Many more classification models could be used in this step, but we choose to focus on AHC, because previous work also relies on this.[1]

We estimate club membership using the Agglomerative Hierarchical Clustering algorithm of Ward (1963). This method was first applied to the selection of valid instruments by Apfel and Liang (2021) in the case of just-identified parametric IV models with as many valid instruments as endogenous treatments. In our context, we instead apply it for clustering the treatment propensity scores as a function of the instruments. We denote by $\mathcal{C}$ the true partition into clusters with $C_k \in \mathcal{C}$ and $\hat{\mathcal{C}}$ as an estimated partition.

[1] In a previous version of the paper, we also considered C-LASSO for the estimation of club identities.

In the following, we provide a description of the algorithm, which is close to that given in Apfel and Liang (2021):

Algorithm 1. *Agglomerative Hierarchical Clustering*

1. ***Input:*** *Calculate each IV-specific propensity score and the Euclidean distance between all of these is stored in a dissimilarity matrix.*
2. ***Initialize:*** *Each propensity score is assigned to its own cluster in the beginning. Total number of clusters at initialization step: J.*
3. ***Join:*** *Two clusters closest in terms of their weighted squared Euclidean distance $\frac{|C_k||C_l|}{|C_k|+|C_l|}||\bar{C}_k - \bar{C}_l||^2$ are joined to a new cluster. $|C_k|$: number of propensity scores in cluster k. $\bar{C}_k$: mean of cluster k, which is the arithmetic mean of all propensity scores in C_k.*
4. ***Iterate:*** *Repeat joining step until all propensity scores are in one cluster.*

Each weighing scheme of the Euclidean distance corresponds to another objective function to be minimized. In the classical case of Ward (1963), which we also use here, the objective function is the sum of within-cluster variance. As discussed in Apfel and Liang (2021), the results do, however, not rely on this specific choice of the dissimilarity metric.

When running the algorithm, we obtain a path of $S = J - 1$ steps, with clusters of size $|\hat{C}_k| \in \{1, ..., J\}$ at each step. Note that each number of clusters entails a different estimated partition $\hat{C}(K)$. To demonstrate that the algorithm covers the true partition, the following high-level assumptions are introduced:

Assumption 6. *Error structure.*
Let $\mathbf{w}_i = (u_i \quad \varepsilon_i)'$. Then, $E(\mathbf{w}_i) = 0$ and $E[\mathbf{w}_i\mathbf{w}_i'] = \begin{pmatrix} \sigma_u^2 & \sigma_{u,\varepsilon} \\ \sigma_{u,\varepsilon} & \sigma_\varepsilon^2 \end{pmatrix} = \mathbf{\Sigma}$ with
$Var(u_i) = \sigma_u^2, \quad Var(\varepsilon_i) = \sigma_\varepsilon^2 \quad Cov(u_i, \varepsilon_i) = \sigma_{u,\varepsilon}$ and the elements of $\mathbf{\Sigma}$ are finite.

where $\varepsilon_i = D_i - E(D_i|Z = z)$. The next assumption invokes the satisfaction of specific laws of large numbers.

Assumption 7.

$$\frac{1}{\sum_{z=1}^{J} N_z}\mathbf{Z}'\mathbf{Z} \xrightarrow{P} E(\mathbf{z}_i\mathbf{z}_i') = \mathbf{Q} \text{ with } \mathbf{Q} \text{ a finite and full rank matrix,} \qquad \frac{1}{N_z}\sum_{i=1}^{N_z} D_i \xrightarrow{P} E(D_i|Z_i = z)\forall z$$

Denote $\mathbf{Z}$ a matrix including all judge dummies and $\mathbf{z}_i$ is the $J \times 1$ observation vector. By Assumption 7, the IV-specific average converges in probability to the conditional expectation of the treatment. In our empirical setup, the IVs are mutually exclusive dummies and hence the matrix product $\frac{1}{\sum_{z=1}^{J} N_z}\mathbf{Z}'\mathbf{Z}$ is a diagonal matrix with entries $\frac{N_z}{\sum_{z=1}^{J} N_z}$, the number of unit observations for each IV. As the inverse of this is the diagonal matrix with entries $\frac{\sum_{z=1}^{J} N_z}{N_z}$, assuming that it is invertible is equivalent to assuming that for each judge the number of cases goes to infinity, $N_z \to \infty$, and the fraction of overall cases and judge-specific cases converges to a constant, $\frac{N_z}{\sum_{z=1}^{J} N_z} \to c$. Furthermore, we assume that a central limit theorem (CLT) holds such that

Assumption 8. $\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{w} \xrightarrow{d} N(0, \mathbf{\Sigma} \otimes \mathbf{Q})$.

Next, we show that the true partition $\mathcal{C}^0$ is on the selection path of the Agglomerative Hierarchical Clustering algorithm. The proof is very close to the one in Apfel and Liang (2021), but we reproduce it here to suit the setting and notation of the first-stage propensity scores.

Lemma 1. *Clusters with propensity scores from the same club C_k are merged first*
Let q and q' denote two clusters such that all propensity scores in the two clusters both satisfy $p_z \in C_k$ (i.e. the propensity scores in the two clusters belong to the same single club). Let q'' be a different cluster s.t. $\exists p_z : p_z \in \hat{C}_{q''}$ and $p_z \in C_{k'}$ but $k' \neq k$ (i.e. there is at least one propensity score in this cluster which does not belong to the club from which the previous two clusters come). Under Assumptions 7 and 4, as $N \to \infty$, clusters q and q' are merged with probability converging to 1.

We subsequently demonstrate on a high level that the probability that an estimated cluster $\hat{C}_q$ with elements from the true club C_k is merged with another estimated cluster with elements from the same club $\hat{C}_k$ goes to 1. The proof closely follows the one in Apfel and Liang (2021) and is divided into three parts for proving the subsequent statements:

1. The means of clusters which include propensity scores from the same club also converge to the same value as each propensity score in the club.
2. The algorithm merges the two clusters that have minimal distance.
3. Clusters with propensity scores from the same club have a distance of zero and clusters with propensity scores from more than one club have a non-zero distance, with the probability going to one.

Proof. Part 1: Consider

$$z, z' \in C_k \quad z'' \in C_{k'} \quad k \neq k'.$$

Under Assumption 7 and 4, it follows that

$$\begin{aligned} plim(\hat{p}_z) = plim(\hat{p}_{z'}) = p_k \\ plim(\hat{p}_{z''}) = p_{k'} \end{aligned} \tag{12}$$

Let $\hat{C}_q$ and $\hat{C}_{q'}$ be estimated clusters with propensity scores from the same club: $\hat{C}_q$, $\hat{C}_{q'} \subset C_k$ and $\hat{C}_{q''} \subset C_{k'}$.

$$plim\ \bar{\hat{C}}_q = \frac{\sum\limits_{\hat{p}_z \in C_k^0} \hat{p}_z}{|\hat{C}_q|} = \frac{|\hat{C}_q| p_k}{|\hat{C}_q|} \text{ where } \hat{C}_q \subset C_k^0 \tag{13}$$

and hence

$$plim\ \bar{\hat{C}}_q = p_k^0.$$

Part 2: Consider the case that the algorithm decides whether to merge two estimated clusters, $\hat{C}_q$ and $\hat{C}_{q'}$, containing propensity scores from the same club, or to merge two estimated clusters containing propensity scores from more than one underlying club, C_q and $C_{q''}$. The two clusters which are closest in terms of their weighted Euclidean distance are merged first. Hence, we need to consider the distances between C_q and $C_{q'}$, C_q and $C_{q''}$, as well as $C_{q'}$ and $C_{q''}$.

C_q is merged with a cluster with elements of its own club C_k^0 iff $\frac{|C_q||C_{q'}|}{|C_q|+|C_{q'}|}||\bar{\hat{C}}_q - \bar{\hat{C}}_{q'}||^2 < \frac{|C_q||C_{q''}|}{|C_q|+|C_{q''}|}||\bar{\hat{C}}_q - \bar{\hat{C}}_{q''}||^2$. Therefore, the following two are equivalent

$$\begin{aligned} & lim\, P(C_q \cup C_{q'} = C_{q,q'} \subseteq C_k^0) = 1 \\ \Leftrightarrow \quad & lim\, P(\frac{|C_q||C_{q'}|}{|C_q| + |C_{q'}|}||\bar{\hat{C}}_q - \bar{\hat{C}}_{q'}||^2 < \frac{|C_q||C_{q''}|}{|C_q| + |C_{q''}|}||\bar{\hat{C}}_q - \bar{\hat{C}}_{q''}||^2) = 1 \end{aligned} \tag{14}$$

where $C_{q,q'}$ is the merged cluster.

Part 3: Our objective is to prove (14). We can then prove $lim\, P(\frac{|C_q||C_{q'}|}{|C_q|+|C_{q'}|}||\bar{\hat{C}}_q - \bar{\hat{C}}_{q'}||^2 < \frac{|C_{q'}||C_{q''}|}{|C_{q'}|+|C_{q''}|}||\bar{\hat{C}}_{q'} - \bar{\hat{C}}_{q''}||^2) = 1$ by changing the subscripts.

First, define $a := \frac{|C_q||C_{q'}|}{|C_q|+|C_{q'}|}||\bar{\hat{C}}_q - \bar{\hat{C}}_{q'}||^2$, $b := \frac{|C_q||C_{q''}|}{|C_q|+|C_{q''}|}||\bar{\hat{C}}_q - \bar{\hat{C}}_{q''}||^2$ and $c := \frac{|C_q||C_{q''}|}{|C_q|+|C_{q''}|}(p_k^0 - p_{k'}^0)^2$.

Under (13)

$$plim(a) = \mathbf{0},$$
$$plim(b) = c$$

The remainder of this proof is identical to the one in Apfel and Liang (2021). It needs to be shown that $\lim_{n\to\infty} P(a < b) = 1$, which is obtained by a proof by contradiction, by showing that $\lim_{n\to\infty} P(b < a) \neq 0$ entails a contradiction. To ease notation, lim stands for $\lim_{n\to\infty}$ in the subsequent discussion. By the definitions of convergence in probability, it follows that

$$\lim P(a < \varepsilon) = 1 \tag{15}$$

and

$$\lim P(|b - c| < \varepsilon) = 1. \tag{16}$$

for any ε. Therefore, $\lim P(b < a) \neq 0$ and $\lim P(a < \varepsilon) = 1$ imply $\lim P(b < \varepsilon) \neq 0$.

Now, consider $\varepsilon < \frac{1}{2}c$. Then,

$$\lim P(b < \frac{1}{2}c) \neq 0 \tag{17}$$

Because of the absolute value of $b - c$, consider two cases, $b < c$ and $b > c$. If $b < c$, then $\lim P(c - b < \frac{1}{2}c) = 1 \Leftrightarrow \lim P(c - b > \frac{1}{2}c) = 0. \Rightarrow \lim P(b < \frac{1}{2}c) = 0$, which contradicts (17). If $b \geq c$, then $a < \varepsilon < \frac{1}{2}c < c \leq b$ and hence $\lim P(a < b) = 1 \Leftrightarrow \lim P(b \leq a) = 0$, which is a contradiction, too. □

Corollary 1. *The true partition is on the selection path*
As $N_z \to \infty\ \forall z$ and $K = K^0$: $lim\, P(\hat{\mathcal{C}}(K) = \mathcal{C}^0) = 1$.

$K = K^0$ means that the number of clusters is equal to the true number of clubs. This follows directly from Lemma 1, because we start with $K = J$ and then merge only clusters which contain propensity scores from the same club. Only when all p_z from each club are in a cluster respectively, the algorithm starts to merge clusters with members of different clubs. For details, refer to Apfel and Liang (2021).

Choice of K via F-test It is unclear at which step of the Algorithm to stop. The penalty parameter to choose in this context is the number of clusters K. Note that we reduced the number of penalty parameters to choose from two to only one. We propose the following procedure to select the tuning parameter.

A method to choose the number of clusters K is to use first-stage F-tests for equality of the propensity scores. Starting with $K = 1$, we test whether all of the propensity scores are equal. If the test is rejected, we proceed to the step of the AHC with $K = 2$ and test for equality of all propensity scores *inside of each cluster.*

Algorithm 2. *Selecting number of clusters via F-test*

1. *Select a significance level α*
2. *Set $K = 1$*
3. *Testing: Perform F-test at significance level α, with $H0:$ Propensity scores inside each cluster are equal*
4. *Updating: If rejected, increase K by 1*
5. *Iterate testing and updating until F-test does not lead to rejection anymore*

The F-test statistic is defined as

$$F = \frac{(\mathbf{R}\hat{p} - \mathbf{r})^\intercal \left[\mathbf{R}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{R}^\intercal\right]^{-1} (\mathbf{R}\hat{p} - \mathbf{r})/q}{s^2} \tag{18}$$

where R is a hypothesis matrix with dimension $(J - q) \times J$. For the IV of the first club to be tested for equality, the matrix entry is 1 and for the remaining IVs it is zero, except for another IV in the club, for which it is -1. In this way, we test the equality of all coefficients. $\mathbf{r}$ is a $(J \times 1)$ vector of zeros. Under the $H0$, the test statistic follows an F-distribution with $J - 1$ and $N - J$ degrees of freedom.

After identifying clubs of IVs with comparable propensity scores, such clubs with distinct propensity score values may be paired, to obtain the first stage effect of the instrument on the treatment as the difference across club-specific propensity scores. In this context, one club of judge IVs in the pair may be considered as reference category and the other as focal category. To compute the LATE estimate, we use the subset of the data for which the judges from the reference category in the pair and those of the focal category have ruled. When a defendant faces a judge from the focal category, then a dummy variable for that club takes the value one and thus serves a club-specific (rather than judge-specific) instrumental variable. We hence have a just-identified model. We may then test the strength of these club pair-specific aggregations of judge-based IVs, e.g. by 2SLS regression. When running the regressions, whether we control for the other judges or clubs does not make a difference, because we are using a subset of the data where there is no variation in the judge dummies other than the for the reference and focal categories.

Furthermore, we may create multiple IVs from the same club pair based on the different judges in the respective clubs and run a 2SLS regression based on these instruments. We call a set of IVs which originates from the very same club pair a *union* of IVs. We can take one club as reference category and create $|\hat{C}_k|$ judge-club specific dummies, generating dummies that take the value 1 for one specific judge in the focal club. Again, we can test the strength of these unions in each first-stage regression. The estimates associated with club pairs that pass the first-stage test are then reported, along with their confidence intervals, the estimates of club pairs that do not pass the first-stage test can be discarded. In this way, we get a map of heterogeneous, club pair-specific first-stage coefficients and estimates.

To summarize, Stage 1 of Procedure 1 works as follows: We use the IVs, the treatment, outcome and controls as input. Then, we find the clubs of judges for which the propensity score is the same. We compute all possible pairs of clubs. We construct single IVs or unions of IVs from each club-pair. We compute the first-stage F-statistic and check if it exceeds a conventional threshold. Next, we can discard club-pairs which do not pass the first-stage test. Finally, report IV or 2SLS estimates for the remaining unions of IVs.

3.3 Allowing for violations of the LATE assumptions

We now move to the second step of our procedure 1. Once we have identified clubs of IVs with the same propensity scores, we have a setting with constant LATEs across judges within pairs of clubs with specific higher and lower propensity scores, under the condition that the LATE assumptions hold. This fact may be used to address the problem that not all judge IV - pairs might satisfy these identifying assumptions. Given that our classification procedures imply that we should observe homogeneous effects, any remaining heterogeneity should now be due to violations of the LATE assumptions.

Concretely, we can proceed in three further ways to deal with violations of the LATE assumptions. Our first approach is based on tests of homogeneity of the estimates, such as the Hansen-Sargan test. When relying on parametric IV models with homogeneous treatment effects, the issue of such tests is that they may not only have power against a violation of the LATE assumptions, but also against a violation of effect homogeneity. In our framework with club pair-specific propensity scores, the true LATEs are, however, homogeneous such that a rejection of the Hansen-Sargan test or another test of overidentifying restrictions implies a violation of the LATE assumptions. If the test rejects, the estimate in a union is hence deemed uninformative.

Secondly, if a majority of IVs in a union is valid, then the median estimator will be consistent, as outlined in Han (2008). Therefore, we can pair up all judges in two clubs and take the median of the resulting estimates. However, inference is not simple for the median estimator. We could bootstrap the standard errors, but we would need to do this for many unions, adding to the computational cost of the method.

Third, we can leverage approaches that select valid and invalid IVs. To achieve this, an emerging literature in statistics and econometrics has been selecting valid and invalid instruments. Building on a seminal paper by Andrews (1999), Kang, Zhang, Cai, and Small (2016) proposed to select valid IVs via the LASSO. We note that IV validity in their study refers to the satisfaction of the exclusion restriction, which does not include other potential violations of the IV assumptions like monotonicity as we explicitly consider in our approach. Their procedure consistently selects valid IVs when the majority of IVs is valid. Windmeijer, Farbmacher, Davies, and Smith (2019) show that the adaptive Lasso, using the median estimator as initial estimate, consistently selects valid IVs under an absolute majority rule, without further assumptions on the strength and correlation of IVs. Guo, Kang, Cai, and Small (2018) show that a procedure based on pairwise testing can select valid IVs under a plurality rule - a potential relaxation of the majority rule. Windmeijer, Liang, Hartwig, and Bowden (2021) simplify this procedure by reducing the number of pairwise tests, using an algorithm that uses the confidence intervals of each IV-specific just-identified estimate. Apfel and Liang (2021) further simplify the selection procedure by using agglomerative hierarchical clustering which outperforms the other methods in presence of weak IVs.

In principle, we can use any of these methods to select valid IVs. We opt to use the Agglomerative Hierarchical Clustering procedure (AHC), because it operates mainly using point-estimates and therefore lends itself to the changes presented next. Moreover, it is easily applicable to settings with multiple endogenous regressors.

Following the reasoning introduced in equation 11, we now search for clusters among the reduced form estimates. The reduced form parameters are

$$E(Y|Z=z) = \Gamma_z \tag{19}$$

and can be obtained from a regression of Y on Z. The clustering is done based on these parameters. In short, the procedure works as follows:

Algorithm 3. *Agglomerative Hierarchical Clustering - Second stage*

1. ***Input:*** *Compute all Γ_z from a regression of Y on Z. The Euclidean distance between all of these is stored in a dissimilarity matrix.*

2. ***Initialize:*** *Each reduced form estimate from club k is assigned to its own cluster in the beginning. Total number of clusters at initialization step:* $|\hat{C}_k|$.
3. ***Join:*** *Two clusters closest in terms of their weighted squared Euclidean distance* $\frac{|C_k||C_l|}{|C_k|+|C_l|}||\bar{C}_k - \bar{C}_l||^2$ *are joined to a new cluster.* $|C_k|$*: number of estimates in cluster k.* $\bar{C}_k$*: mean of cluster k, which is the arithmetic mean of all propensity scores in* C_k.
4. ***Iterate:*** *Repeat joining step until all propensity scores are in one cluster.*

Note that the clustering procedure is performed for the reduced form coefficients inside of each club. This is exactly the procedure in Apfel and Liang (2021), with the difference that we're taking judge-specific reduced form estimates here. Note that this is very similar to the clustering algorithm for the propensity scores, Algorithm 1. To determine the number of clusters to choose, we use the downward testing procedure, as proposed Andrews (1999) and applied in Windmeijer, Liang, Hartwig, and Bowden (2021) again using the F-test, just as in Algorithm 2:

Algorithm 4. *Selecting number of clusters via F-test*

1. *Select a significance level* α
2. *Set* $K = 1$
3. *Testing: Perform F-test at significance level* α*, with* $H0$: *Reduced form coefficients inside each cluster are equal*
4. *Updating: If rejected, increase* K *by 1*
5. *Iterate testing and updating until F-test does not lead to rejection anymore*

When a plurality rule holds the method has oracle properties.[2]

It is worth noting that usually, IV selection methods select valid IVs, but in our case, we want to select valid IV values which take part in creating valid IVs. As an alternative to selecting inside clubs, we could also compute all judge-pairs inside a union and IV estimates for each of them. There are $J_a \cdot J_b$ such judge pairs for two clubs a and b where J is the number of judges in a club. The AHC then clusters the IV estimates that are closest to each other in terms of the weighted Euclidean distances. After each step, we extract the IVs that are involved in the estimation of the largest cluster of estimates and verify homogeneity of the estimates via the Sargan test. In an earlier version of this paper we were using this procedure instead of the one based on the reduced form parameters. The simulation results were very similar and we opted for clustering of the reduced form estimates for simplicity.

4 Simulation

In the following exercise we simulate data to mimic real-world applications but with the difference that we know the underlying DGP. We first show a setting with no invalid, then with a few invalid and then with many invalid judges. By invalid judges we denote any judge that violates any of the LATE Assumptions. For each setting we show simulations with a small and a large sample to verify classification consistency.

[2] For further details on the respective IV selection methods, please refer to the original studies.

4.1 Setup

We choose a setting like the one in the application with an unbalanced panel, but with fewer judges. We set the number of judges to $J = 30$. The number of cases per judge is selected as follows: draw from a uniform distribution $U(0.3, 5)$, multiply this value by 50 for a small sample and by 100 for a large sample and round to the next integer. Repeat this for all J judges. We construct judge-specific dummies: $Z_{ji} = 1$ if $j = k$, otherwise 0 for $k \in \{1, 2, ..., 30\}$.

The first-stage error is uniformly distributed: $V \sim Unif(0, 1)$. The reduced-form error is: $U = 0.5 \cdot V + W_i$ where $W_i = Unif(0, 1)$. The RF-error is correlated with the FS-error, and this correlation creates the endogeneity of the treatment. The first stage coefficients (or propensity scores) of the judge dummies are set to: $\Pi = (0.97, 0.5, 0.2, 0.05)'$ and the corresponding size of the clubs are 12, 12, 3 and 3, respectively. Hence we have four clubs.

The first stage is defined as:

$$D_i = I(\mathbf{Z} \times \Pi > V_i)$$

where $\mathbf{Z}$ is the $(N \cdot J) \times J$ matrix of judge dummies. The outcome is created as

$$Y_i = (D_i \cdot 0.5 + D_i \cdot U_i + \mathbf{Z}\gamma + U_i)^4$$

where γ is the vector of direct effects of the judges on the outcome. Keep in mind that violations of the exclusion restriction are just one way in which the IV can be invalid. We choose this violation for simplicity in this simulation but one could also consider violations of monotonicity or random assignment. We can calculate the (oracle) Local Average Treatment Effects for complier populations that get treatment (go to jail) with judge j, but don't go to jail with judge k, directly from the generated data, as the conditional expectation

$$LATE(j, k) = E\left(Y(D = 1) - Y(D = 0) \mid D(Z_j = 1) = 1, D(Z_k = 1) = 0\right).$$

Analogously we can calculate the oracle LATE for complier populations that go to jail with a valid judge from club a but don't go to jail with a valid judge from club b as the conditional expectation

$$LATE(a, b) = E\left(Y(D = 1) - Y(D = 0) \mid D(C_b = 1) = 1, D(C_a = 1) = 0\right).$$

4.2 Simulation Results

We start with a setting where there are no invalid judges. We ran 1000 Monte Carlo repetitions. In Table A.1 we show the 2SLS coefficient estimate and standard error which is 28.27 and 0.84. These results are for the small sample setting. All judges are classified into the correct club with probability at 0.52 when the significance level in the F-test to select the number of clusters is set to 0.05, and can be lower when the significance level is lower. The mean number of clusters is close to four but can be slightly lower in the settings with a lower significance level than 0.05. The fraction of repetitions for which the number of clusters selected is four varies between 0.53 and 0.73.

In the first line of Table A.2 we compute the oracle LATE for the pairs of clubs, taking only valid judges into account. In this table we summarize 2SLS, Median and AHC-2SLS estimates as discussed in the preceding sections with the restriction that we take outcomes when the number of clusters selected is correct (four). Therefore, there are LATEs for six club pairs (unions). The mean of the estimates is always close to the oracle LATE, which is not surprising as there are no invalid IVs. Standard errors are rather low, the CI coverage is mostly close to 0.95, but sometimes there is 100 % coverage. The CI never includes the 0 as illustrated by the power being 1. The mean of Hansen-Sargan test p-values are close to 0.5 on average.

In Figure A.1 we compare the standard 2SLS, which does not account for invalidity and the AHC-2SLS estimates (i.e. AHC is used in the second stage), which do select invalid IVs. The red histogram illustrates the standard 2SLS estimates and AHC illustrates the AHC-2SLS estimates. The red vertical lines illustrate the oracle LATEs. Unlike in the preceding table, we also compare estimates for when the first stage selection yielded more than four clubs. Knowing that in this setting there is no invalidity yet and having seen the comparison in the preceding table, it does not come as a surprise that the histograms that visualize the results of the two methods show a large overlap.

Increasing the sample size for this setting with no invalid judges leads to correct classification in 79-84% of the cases and the mean number of clusters approaches four. The fraction of repetitions when $K = 4$ ranges from 0.86 to 0.92. This can be seen in Table 1. The results of the classification tables are very similar in the following settings and hence we omit their discussion. When increasing the sample size, the standard errors decrease, but bias, power and overidentification test results are fairly similar (Table A.3). Coverage increases in some settings and approaches one. The improved classification when increasing sample size is an indication that in fact classification is consistent.

Next, we select a setting with a few invalid judges. To achieve this, we set $\gamma = (0.1, 0.2, 0.3, 0i_9, -1.0, -1.5, -2.0, 0i_{15})$, where i_n is a vector of ones with n entries. This implies that the first three judges in club 1 and 2 are invalid because they have a direct effect and the two remaining clubs are valid. As can be seen in Tables A.5 and A.7, now the 2SLS and median estimates are biased away from the oracle LATEs as soon as a club including invalid IVs is involved in the estimation. Hansen-Sargan p-values for specifications that do not account for invalid IVs are close to 0. CI coverage of the standard 2SLS is too low and can for some unions never cover the true value. For AHC-2SLS which selects valid IVs, the average of estimates is still close to the oracle LATE, keeping the SE low, coverage close to 0.95, power at 1 and Hansen-Sargan p-values close to 0.5 on average suggest no rejection. The illustrations in Figures A.3 and A.4 suggest that now some of the point estimates lie above and below of oracle LATEs, when estimating club-pair specific IV without accounting for invalid IVs and that is not the case for AHC-2SLS as indicated by the blue histogram that indicates that most AHC-2SLS estimates lie within the range of oracle LATEs.

Finally, we choose a setting with many invalid judges to put our method to the test. We set $\gamma = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0i_6, -1.0, -1.5 - 2.0 - 2.5, 0i_8, 0.7, 0, 0, 0.8, 0, 0)$, meaning that the first six IVs in the first club, the first four in the second club and the first in the last two clubs are invalid, leading to a setting with many invalid IVs. In a small sample setting, the 2SLS and median results are much farther off from the oracle LATE than with few invalid IVs. As can be seen in Table A.9, for the first union, the mean estimates are at 59, instead of the oracle LATE of 34, for example. CI coverage is too low for all unions, power is not one for all unions and Hansen-Sargan p-values close to zero suggest the presence of invalid IVs. Similarly, the median estimates are also far away from the oracle LATEs. The AHC-2SLS results are closer to the oracle LATEs, mostly have lower standard errors, higher coverage, power and higher mean p-values of the HS-test. Figure A.5 suggests that a large fraction of invalidity-unadjusted club-pair IV estimates is outside of the range of oracle LATEs. However, the performance of AHC-2SLS is still far from optimal, with biased estimates and coverage between 0.5 and 0.85. When increasing the sample size, as seen in Table 2, AHC-2SLS increases as measured by all performance statistics, now has low bias, coverage closer to the nominal level and power one, while club-pair IV without correction for invalid IVs still performs badly. Figure 1 confirms that AHC-2SLS is concentrated around the oracle values while uncorrected club-pair IV is outside of the interval spanned by the oracle LATE estimands.

Overall, the simulations produce two main findings: First, the first-stage classification is indeed consistent. Second, with an increasing number of invalid IVs the 2SLS and median estimators quickly decline in their performance, while the corrected 2SLS estimator which selects valid IVs in the second stage performs well and approaches oracle performance as the sample size increases.

Figure 1: Illustration of results, Many invalid, Large sample

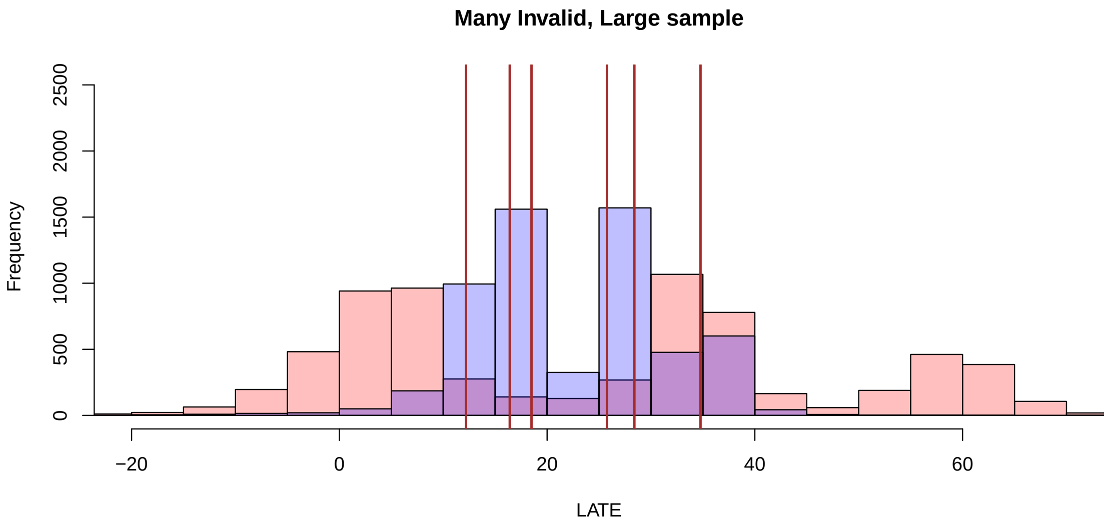


Table 1: Simulation: No invalid, Large Sample - Classification

α	2SLS	SE	Cons Class	# clubs	Rep K=4
0.001	28.4	0.77	0.84	3.97	0.91
0.01	28.4	0.77	0.84	4.04	0.92
0.05	28.4	0.77	0.79	4.13	0.86

Note: ConsClass indicates the fraction of times that all judges have been assigned to the correct club. # Clubs indicates the mean number of clusters selected by the algorithm. Rep $K = 4$ shows the fraction of times that the algorithm selected $K = 4$. In the first line, we show the results for the oracle estimator, i.e. when the club membership is known. The following LATE estimates are reported only when the number of clubs selected is 4.

Table 2: Simulation results - Many Invalid, Large sample

Est	α	L1	SE	CI	P	HS	L2	SE	CI	P	HS	L3	SE	CI	P	HS	L4	SE	CI	P	HS	L5	SE	CI	P	HS	L6	SE	CI	P	HS
				cov.					cov.					cov.					cov.					cov.					cov.		
Oracle		34.77					28.4					25.76					18.5					16.41					12.19				
2SLS	0.001	59.19	1.76	0	1	0	36.33	1.77	0.09	1	0	33.05	1.48	0.04	1	0	0.71	1.68	0	0.57	0	5.67	0.98	0	0.9	0	15.63	3.52	0.32	0.78	0
	0.01	59.22	1.76	0	1	0	36.42	1.8	0.09	1	0	33.07	1.49	0.04	1	0	0.72	1.7	0	0.58	0	5.7	0.99	0	0.89	0	15.58	3.55	0.33	0.77	0
	0.05	59.25	1.76	0	1	0	36.36	1.81	0.1	1	0	33.1	1.5	0.04	1	0	0.66	1.7	0	0.58	0	5.75	0.99	0	0.9	0	15.93	3.58	0.33	0.77	0
Med.	0.001	50.95					30.74					27.87					1.2					5.65					12.92				
	0.01	50.94					30.82					27.87					1.19					5.62					12.89				
	0.05	50.92					30.76					27.88					1.08					5.66					12.9				
AHC	0.001	35.87	1.51	0.86	1	0.33	28.51	1.6	0.94	1	0.45	25.81	1.32	0.95	1	0.38	16.91	1.93	0.89	0.99	0.45	15.24	1.22	0.85	1	0.38	11.88	2.19	0.81	0.94	0.39
	0.01	35.87	1.51	0.85	1	0.33	28.49	1.6	0.94	1	0.45	25.78	1.33	0.95	1	0.38	16.89	1.94	0.88	0.99	0.45	15.21	1.23	0.84	1	0.38	11.82	2.22	0.8	0.94	0.38
	0.05	35.87	1.51	0.86	1	0.34	28.42	1.61	0.93	1	0.45	25.78	1.33	0.95	1	0.38	16.76	1.94	0.88	0.99	0.45	15.24	1.23	0.85	1	0.38	11.95	2.26	0.8	0.94	0.38

Note: 1000 iterations. L1-L6 indicate Local Average Treatment Effects calculated using one group as basis and the other one as focus group for a union of instruments. CI-cover indicates whether the 95% includes the true LATE. 2SLS is the mean estimate when using all IVs. HS indicates p-value of Hansen-Sargan test. SE is the corresponding mean standard error of the estimates. CI cov. stands for coverage rate of 95% CI. P stands for power, probability with which 0 is outside CI.

5 Application

5.1 Assessing the effect of incarceration on recidivism

The threat of incarceration is thought to affect crime by acting as a deterrent while incarceration itself could either potentially increase or decrease the probability of future criminal behavior (i.e. recidivism). Incarceration can reduce the likelihood of recidivism if prisons help rehabilitate offenders and prepare them for reintegration into society. If, however, the stay in prison fosters integration into criminal networks or leads to a loss of human capital, incarceration might increase the likelihood of recidivism.

In recent years, several studies from the fields of economics and criminology have addressed the question whether incarceration affects the likelihood of recidivism or, in other words, whether sentencing offenders to prison affects the likelihood of them relapsing into criminal behavior. In order to answer this question, authors typically aim at estimating the following empirical model:

$$Y = \phi(D, X, U) \tag{20}$$

where Y denotes the outcome, an indicator for recidivism, i.e. re-indictment or re-incarceration in a certain time window after conviction. D is the treatment variable, indicating whether a judge has sentenced the convict to a term of imprisonment or has decided for a non-prison sentence, such as probation. X are observed covariates potentially affecting the probability of recidivism, which might serve as control variables. Nagin, Cullen, and Jonson (2009) suggest to use prior record, offence type, age, race and sex as key control variables. U reflects unobserved characteristics affecting the outcome.

Estimating equation (20) by OLS is generally inconsistent if the association between the outcome and the right-hand side variables is nonlinear and/or unobserved confounders jointly affect D and Y even after controlling for X. For instance, the judge might have information (not available to the researcher) about the defendant's previous offenses or possible addictions, which may simultaneously influence the judge's sentencing and the likelihood of recidivism. Likewise, the sentence may be influenced by the defendant's behavior in court, which in turn may shed light on the defendant's potential to recidivate. In order to address this issue of unobserved confounders, several studies have leveraged the fact that in many courts cases are randomly assigned to judges whose use of prison sentences varies systematically. They have considered dummies for being assigned to a particular judge or the incarceration rate by judge as instrumental variables.

Studies attempting to estimate the effect of incarceration on recidivism by means of IVs mostly consider data from the U.S., see for instance the seminal study by Kling (2006), but there is also some research on data from other countries, such as Chile (Cortés, Grau Veloso, and Rivera Cayupi, 2019) and Norway (Bhuller, Dahl, Løken, and Mogstad, 2020). Most studies examine the incarceration effect only among convicts, while some, such as Cortés, Grau Veloso, and Rivera Cayupi (2019) and Leslie and Pope (2017), assess the impact of pre-trial detention among all individuals accused of a crime. Loeffler and Nagin (2021) provide a review of 13 published IV-based studies on the effect of incarceration on recidivism concluding that those based on data from U.S. courts mainly find either an insignificant or a significant recidivism-increasing effect. It appears that U.S. prisons are ineffective in terms of rehabilitation and resocialization, and thus do not serve to fight crime beyond general deterrence effects. In contrast, a study by Bhuller, Dahl, Løken, and Mogstad (2020) finds that incarceration in Norwegian prisons has a significant recidivism-reducing effect and a positive impact on employment, pointing to considerable heterogeneity in the effect of incarceration across regions, possibly due to differences in prison infrastructure and support services, such as labor market training opportunities.

Effect heterogeneity can also arise within the same institutional context, in that defendants sentenced to prison by different judges generally differ in their background characteristics (such as personality traits) and these in turn can influence the effect of incarceration. For this reason, the LATE, i.e. the effect of incarceration on recidivism among individuals who would be incarcerated under a stricter but not a more lenient judge might generally differ across distinct pairs of judges (or relatedly, distinct propensities of incarceration). Such effect heterogeneity would be masked when applying 2SLS simultaneously to all judge IVs, which is one motivation for using our approach.

Further, in any of the IV approaches to the instruments might fail to satisfy the identifying assumptions outlined in Section 2.1. For instance, judges could differ in terms of their use of alternative measures to imprisonment, such as electronic monitoring (Loeffler and Nagin, 2021), which would violate the exclusion restriction postulated in Assumption 1. Further, the IVs might violate Assumption 2, requiring monotonicity of the treatment in the instrument, see for instance the discussion in (Frandsen, Lefgren, and Leslie, 2019). Depending on how different judges weigh the individual aspects of a case, a rather stringent judge with a relatively high rate of prison sentences could refrain from incarcerating a particular defendant, who would have been imprisoned by a judge with a lower incarceration rate. i.e. their (potential) sentences could contradict the judges' order of (average) severity. Such a situation entails the existence of defiers and thus a violation of weak monotonicity of the treatment in the instrument.

To nevertheless consistently estimate the LATEs of interest, we invoke Assumption 4, i.e. the existence of clusters (or clubs) of judges with the same propensity score to incarcerate, as well as Assumption 5, i.e. plurality, such that a relative majority of judges satisfies the LATE assumptions. The cluster assumption appears realistic as long as judges can be plausibly categorized into a limited number of judge types, each with a distinct rate of prison sentences.

5.2 Data

The model developed in Section 2 is applied to a data set on offenders in the U.S. state of Minnesota, which is composed of data from two primary sources: the first one is an extract from the Minnesota Judicial Branch case database containing information on the offender, including his/her full name, date of birth and place of residence, for all 2009 to 2020 criminal cases; the second source is a data set from the Minnesota Sentencing Guidelines Commission on all adult offender cases in Minnesota between 2001 and 2017, with information on the type and severity of the offense as well as the judge hearing the case. We link these two datasets using the case number as identifier in order to obtain a data set with years ranging from 2009 to 2017 on all criminal offense cases that includes information on offender, judge, offense and conviction.

For each offender in our data set, we identify all criminal cases in Minnesota in which he/she was involved between 2009 and 2017, using the offenders' full names and dates of birth as identifiers. Despite also having information on the offenders' place of residence, we do not include this information for identifying recidivism in order to account for changes of residence, which occur frequently, especially after returning from jail or prison. This way, we accept the low risk of incorrectly linking the cases of two individuals with the same name and date of birth, while reducing the risk of not detecting recidivism. In addition, we cannot detect recidivism if an offender commited a crime in another state, has moved to another state/country or has changed his or her name.

For estimating the effect of incarceration on offender recidivism, we consider recidivism within 3 years of sentencing as the outcome. Therefore, in order to observe the 3-year post-conviction period of every offender, we must reduce our dataset to the years 2009 to 2014. According to the Minnesota Order for Assignment of Cases, all criminal cases in Minnesota are randomly assigned to a judge having jurisdiction in the county in which the crime is tried. After being assigned to a case, a judge in active service must preside over that case until its resolution, i.e., random judge assignment, as required for our IV approach, is guaranteed by Minnesota state case assignment rules. Senior judges, however, are permitted to opt out from hearing a case. We therefore remove all cases heard by a senior judge in order to ensure random judge assignment. Then, although the vast majority of judges in Minnesota remain in one and the same court throughout their tenure, there are some judges in the sample that have changed court during our observation period. These judges are assigned a different ID for each court such that the terms at different courts are treated as if belonging to different judges, in order to account for differences in county crime profiles, which in turn are reflected in the judges' sentencing practices.

To ensure that the vast majority of offenders have been able to re-offend in the data within the three years following sentencing, i.e., are not in prison for the entire three years during which we observe potential recidivism, we need to reduce the data set to minor crimes. An analysis of the recidivism effect based on the entire dataset would require strict control for crime and offender profiles in order to avoid bias in the estimated effect caused by the inclusion of observations for which recidivism is highly unlikely due to long-term incarceration. At the same time, including offenses that result in a long prison sentence would not improve the quality of the estimator for the recidivism effect. Recent studies on the effect of incarceration have reduced their datasets based on different rules: Loeffler (2013) have concentrated on cases of the three lowest charge classes, Green and Winik (2010) on drug offenses. We follow Bhuller et al. (2020) by reducing the dataset based on the sentence lengths that an independent institution - in our case the Minnesota Sentencing Guidelines Commission - recommends for each observed offense given the type of crime, the severity of the offense and the offender's criminal history. We reduce our dataset to cases with presumptive sentences of up to three years (a detailed list of the included offenses can be found in Appendix B.1.2). The resulting sample contains 48,849 cases involving 38,874 unique offenders. Some 82% of the offenses in the final dataset have resulted in a sentence of up to one year in county jails, while in about 14% of the observed cases, offenders were sentenced to one to two years in state prison, meaning in 96% of the cases offenders had at least one year after their official release date to recividate. In only some 0.9% of the cases the offender was convicted to three or more years in state prison. Given that usually only two-thirds of a sentence are served in prison and the rest is on probation, the share of offenders having at least one year to recidivate is even higher than 96%.

5.3 Descriptive Statistics

Table B.11 in Appendix B.1.1 provides some descriptive statistics for our data, namely the mean of outcome and covariates in the total sample, as well as among those offenses resulting in a prison or jail sentence ($D = 1$) and those that are not sanctioned with a prison or jail sentence ($D = 0$). It shows that only some 10% of the 48,849 offenses in our sample did not result in incarceration. The descriptive statistics suggest that the distribution of prison/jail sentences differs not only in terms of crime type but also in terms of the offender's race and gender. The proportion of cases that resulted in a prison or jail sentence is lower among white or Hispanic offenders than among offenders of other races and higher among men than among women. The share of property crimes sanctioned with incarceration is smaller than that of crimes against persons, drug crimes, weapon offenses and sex offenses. The table also shows that the severity of crimes and the likelihood of incarceration are positively correlated, where the variable "Severity" is an indicator for the seriousness of a crime as defined by the Minnesota Sentencing Guidelines Commission, ranging from low (Severity = 1) to high (Severity = 11)[3].

Following Bhuller et al. (2020), we perform a robustness check on our data in order to confirm that

[3] The Minnesota Sentencing Guidelines Commission defines a different severity grid for sex offenders. The severity levels of the sex offenders in our sample are translated into the standard severity level according to the sentence lengths the Minnesota Sentencing Guidelines Commission's suggests for each severity level.

there is no evidence of judges not being randomly assigned to cases. For each case, we construct a judge stringency measure as the average incarceration rate for all other cases heard by the same judge as the one assigned to the respective case, i.e. the judge stringency measure is the leave-current-case-out average of the incarceration dummy. This judge stringency measure will also be used as instrumental variable in one of or baseline IV approaches. In line with Bhuller et al. (2020), we restrict the sample to cases heard by judges who decided at least 50 randomly assigned cases between 2009 and 2017 and, in addition, to cases tried in counties where no fewer than two of these judges are stationed in any given year. The resulting sample contains 42,437 cases.

For the robustness check, the stringency measure is regressed on all covariates considered in our analysis, meaning the offender's age, gender and race, the type and severity of the offense as well as interacted court-year fixed effects, with the standard errors clustered by offender and judge. This way, we can assess whether the covariates can to some extent predict judge stringency, which would indicate that the principle of random judge assignment may be violated in our dataset. We find no statistically significant relationship between the measure of judge stringency and the covariates; all estimated coefficients on the covariates are close to zero and not statistically significant at the 10% level. The covariates are not jointly significant either. The results can be found in Table B.13 of Appendix B.1.1.

5.4 Baseline IV Approach

In a first baseline IV estimation, we follow the approach proposed by Bhuller et al. (2020). Using the judge stringency defined in Section 5.3 as instrumental variable, we estimate the effect of incarceration on recidivism within three years of conviction. In a second baseline IV estimation, we use judge dummies as instrumental variables. For doing so, we reduce the sample further so that all judges who have heard fewer than 50 minor crime cases between 2009 and 2014 are removed from the sample. Then, the sample is again reduced to cases tried in counties where at least two of these judges are stationed in any given year, leaving us with a sample including 40,097 cases. This second IV approach is close to the approach by Green and Winik (2010) who use dummies for judicial calendars as instruments and is furthermore closer to the econometric procedure developed in this paper.

Variable	Coef.	(p-Value)	Sign. Level	F-Statistic	(p-Value)	Sign. Level
A. No Control						
Judge Stringency	0.9635	(0)	***	1981.388	(0)	***
Judge Dummies				15.1904	(0)	***
B. Court x Year FE						
Judge Stringency	0.6854	(1.11e-09)	***	37.1342	(1.11e-09)	***
Judge Dummies				4.553	(1.16e-93)	***
C. Only Covariates						
Judge Stringency	0.9492	(0)	***	3318.6578	(0)	***
Judge Dummies				15.0684	(0)	***
D. Covariates & Court ×x Year FE						
Judge Stringency	0.6835	(1.03e-08)	***	32.7912	(1.03e-08)	***
Judge Dummies				4.6013	(2.43e-95)	***

Table 3: F-statistics on the first-stage regression estimates of incarceration on judge stringency and judge dummies. Panel A includes controls for fully interacted county - year fixed effects. Panel B additionally includes all covariates. The standard errors are two-way clustered at judge and offender level. Significance levels: . p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

As is evident from the F-statistic on the instruments in the first-stage regression (see Table 3), both, the judge stringency and the judge dummies, are strong instruments for incarceration. We have estimated the first stage when controlling for fully interacted county - year fixed effects (Panel A) and when additionally controlling for all covariates on offender and offense characteristics (Panel B). The coefficients on the judge stringency measure suggest that assignment to a judge with a 10 percentage point higher stringency, i.e., a 10 percentage point higher overall incarceration rate, increases the probability of receiving a prison sentence by some 7 percentage points.

	Coef.	Standard Error	Sign. Level	Sargan-J	p-val.
OLS: No Controls	0.0815	0.008	***		
OLS: Only Court x Year FE	0.077	0.009	***		
OLS: Only Covariates	0.0674	0.008	***		
OLS: Covariates & Court x Year FE	0.0701	0.009	***		
IV Stringency: No Controls	0.092	0.022	***		
IV Stringency: Only Court x Year FE	0.0744	0.083			
IV Stringency: Only Covariates	0.0541	0.022	*		
IV Stringency: Covariates & Court x Year FE	0.0995	0.095			
IV Judge Dummies: No Controls	0.0578	0.041		409.16	0
IV Judge Dummies: Only Court x Year FE	0.02	0.048		233.28	0.001
IV Judge Dummies: Only Covariates	0.0431	0.038		358.96	0
IV Judge Dummies: Covariates & Court x Year FE	0.0538	0.055		216.55	0.012

Table 4: The estimations include controls for fully interacted county - year fixed effects. The omitted category for race is “White", the one for crime type is “Property Crime". The standard errors are two-way clustered at judge and offender level. Significance levels: . p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

Table 4 shows the results by means of a simple OLS approach, the IV approach as proposed by Bhuller et al. (2020), as well as the approach with judge dummies as instrumental variables. We estimate each approach without covariates and when controlling only for fully interacted county - year fixed effects, only for the set of covariates as well as for the full set of covariates and fixed effects. A comparison of the results from the OLS and the two IV approaches shows substantial differences. While all approaches suggest a recidivism-increasing effect of incarceration, the OLS results indicate a highly statistically significant effect while the effect estimated with the IV approaches is similar in magnitude but not statistically significant. The results of the two baseline IV estimations are in line with the IV literature on the effect of incarceration on recidivism in the U.S. that mostly find an increasing effect or a null effect of incarceration on recidivism (see e.g. Loeffler and Nagin, 2021).

5.5 Subset analysis

We tackle the problem of accounting for observable controls in the following two ways: first we reduce the data to a subset with similar observables and perform the clustering of clubs inside this reduced dataset. We limit the data to male, white defendants of ages up to 100 with crimes against persons with a severity index equal to four. Moreover, we only look at judges with a minimum case count of 20, so that the calculation of propensity scores remains reliable.

Table 5: Effect of imprisonment on recidivism

	OLS	2SLS	AHC-21	AHC-31	AHC-32
Prison	0.0545	0.130	0.289*	0.0203	0.0607
	(0.0279)	(0.0707)	(0.142)	(0.0843)	(0.0943)
Nr IVs		103	62	62	39
F		4.856	6.674	5.062	2.958

Standard errors in parentheses, hetereoskedasticity robust.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In this case, the OLS estimate is statistically insignificant with a coefficient of 0.05, the 2SLS estimate is more than double the OLS estimate but it is still statistically indistinguishable from zero. When using algorithms 1 and 2 we find three clubs. The judge clubs have size 62, 39 and a small cluster with only two judges. When pairing these clubs, we find three unions of IVs. Pairing clubs two and one, we get a 2SLS estimate of almost 0.3, which is more precisely estimated. For the other two unions, the coefficient estimates are much lower and statistically insignificant. Hence, it seems that the initial 2SLS analysis was masking some heterogeneity in the Local Average Treatment Effects. We don't proceed to the selection of invalid IVs at this step, because the Hansen-Sargan test is not rejected in the original 2SLS estimate. One main drawback of this analysis is that the first-stage F-statistics are worryingly low.

Table 6: First step clustering - Subset analysis

Clubs	Mean	Count
1	0.98	62
2	0.87	39
3	0.40	2

5.6 Including controls

If we want to preserve statistical power, instead of looking at a subset, we may also take the entire dataset and include a rich set of controls. In this application, we include race-, gender-, offense-type-, year-dummies, severity, age, the squares of the latter two and race-gender interaction dummies.

Results are presented in table 8. The OLS estimate is now at 0.068 and it is estimated with precision. The 2SLS estimate is at 0.054 and it is also statistically significant. Its first-stage F-statistic is moderate, at 18 and the Hansen-Sargan test of overidentifying restrictions rejects clearly. This might well be because if using all judge-specific dummies, some of the focal judges are in the same club as the reference judge, leading to first stage relationships that are zero, and hence to a weak IV problem.

We start with our first-step AHC and choose the significance level for the F-test as 0.01. In table 7, we show the results of the first stage clustering. We first partial out the controls from the outcome, treatment and the IVs. We then run the first-stage regression of the treatment (imprisonment) on all judge dummies. With that we get five clubs, of sizes 1, 17, 20, 98 and 104. The singleton club has a negative mean and we hence exclude it. We also exclude it because the second stage selection would be pointless with a singleton club and we cannot be sure the judge is valid. The other propensity score means lie between 0.35 and 0.63, with most of them lying between 0.56 and 0.63.

Once we pair the four remaining clubs up and build an IV which compares two clubs, we get a wide range of coefficients, from -0.237 to 0.305. These estimates are denoted by Modus: "Single". The number of judges involved here lies between 37 and 202. F-statistics are high and range between 178 and 993. A red flag is that even though we found different estimates with relevant first stages, the tests of overidentifying restrictions still reject clearly, with p-values close to zero.

If we look for the largest group of reduced form coefficients and reduce the number of IVs further via our second-step AHC, we get lower first-stage F-statistics, which are however still high, between 88 and 982. The Hansen-Sargan tests now do not reject anymore at any conventional significance level. Now,

Table 7: Result of first-stage AHC - Including controls

we find estimates that are even more spread out than before: coefficient estimates now range from -0.483 to 0.415, and all of them are significantly different from zero.

The key takeaway of this application is that our method can discover different LATEs from a large set of IVs and can provide a list of LATEs that would otherwise be collapsed into the 2SLS estimate. Using the first-step AHC, we can find different clubs of propensity scores which yield very high first-stage F-statistics. Whether there is a monotonously increasing effect of the disaggregation of the 2SLS estimate into several LATEs on the first-stage F-statistic is unclear and is left for future research. Using the second-step AHC we find subsets of IVs in the club-pairs that seem more likely to fulfil the exclusion restriction.

One might think about even more flexible ways to control for observables, such as through random forests or other machine learning methods.

Table 8: Effect of Imprisonment on Recidivism

	OLS	2SLS	1-2	1-2	1-3	1-3	1-4	1-4	2-3	2-3	2-4	2-4	3-4	3-4
Prison	0.0680	0.0540	-0.00640	-0.0859	-0.101	-0.255	-0.237	-0.483	0.305	0.415	0.173	0.243	0.0876	0.0894
	(0.00675)	(0.0207)	(0.0338)	(0.0354)	(0.0453)	(0.0508)	(0.104)	(0.118)	(0.0723)	(0.0868)	(0.0538)	(0.0700)	(0.0897)	(0.125)
Modus			Single	Double	Single	Double	Single	Double	Single	Double	Single	Double	Single	Double
J			115	85	121	71	37	28	202	122	118	79	124	65
N	44520	44520	22866	16798	20495	12200	6170	4779	37603	23240	23278	15819	20907	11221
P(HS)		0	0	.1773	.0004	.9614	.0006	.0645	.0002	.2626	0	.269	.00024	.935
F		18.53	992.2	981.5	545.5	497.7	115.0	101.2	687.4	407.0	541.1	303.6	178.0	88.39

Cluster-robust standard errors in parentheses. Significance level in testing procedure: 0.01.

6 Conclusion

In this paper, we have shown how to use retrieve grouped LATEs from data under violations of the LATE assumptions. We first group IV values by their propensity scores and use these to estimate Local Average Treatment Effects. In a second step, we can search for violations of the LATE assumptions among the judges in each club. This procedure can be useful in the economics of crime, when using judge instruments, but could be extended to settings with non-binary IVs. Applications in epidemiology, in Mendelian Randomization for example, might especially benefit from these new developments. Also, settings where the researcher already knows that there are clubs of propensity scores that are identical, for example through a policy, would be of particular interest.

In future work, it would be interesting to allow for propensity scores that deviate from a common cluster mean with a local-to-zero deviation.

References

Andrews, D. W. (1999). Consistent Moment Selection Procedures for Generalized Method of Moments Estimation. *Econometrica 67*(3), 543–563.

Angrist, J., G. Imbens, and D. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of American Statistical Association 91*(434), 444–472 (with discussion).

Apfel, N. and X. Liang (2021). Agglomerative Hierarchical Clustering for Selecting Valid Instrumental Variables. *arXiv preprint arXiv:2101.05774*.

Bhuller, M., G. B. Dahl, K. V. Løken, and M. Mogstad (2020). Incarceration, recidivism, and employment. *Journal of Political Economy 128*(4), 1269–1324.

Cortés, T., N. Grau Veloso, and J. Rivera Cayupi (2019). Juvenile incarceration and adult recidivism.

Frandsen, B. R., L. J. Lefgren, and E. C. Leslie (2019). Judging judge fixed effects. *NBER working paper 25528, National Bureau of Economic Research.*

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics 139*(5), 35–75.

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.

Guo, Z., H. Kang, T. T. Cai, and D. S. Small (2018). Confidence Intervals for Causal Effects with Invalid Instruments by Using Two-Stage Hard Thresholding with Voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 793–815.

Han, C. (2008). Detecting Invalid Instruments Using L1-GMM. *Economics Letters 101*(3), 285–287.

Heckman, J. J. and E. Vytlacil (2001). Local instrumental variables. In C. Hsiao, K. Morimune, and J. Powell (Eds.), *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya.* Cambridge: Cambridge University Press.

Heckman, J. J. and E. Vytlacil (2005, May). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica 73*(3), 669–738.

Huber, M. (2014). Sensitivity Checks for the Local Average Treatment Effect. *Economics Letters 123*(2), 220–223.

Imbens, G. W. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Kang, H., A. Zhang, T. T. Cai, and D. S. Small (2016). Instrumental Variables Estimation with Some Invalid Instruments and Its Application to Mendelian Randomization. *Journal of the American Statistical Association 111*(513), 132–144.

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

Leslie, E. and N. G. Pope (2017). The unintended impact of pretrial detention on case outcomes: Evidence from new york city arraignments. *The Journal of Law and Economics 60*(3), 529–557.

Loeffler, C. E. (2013). Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology 51*(1), 137–166.

Loeffler, C. E. and D. S. Nagin (2021). The impact of incarceration on recidivism. *Annual Review of Criminology 5.*

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

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science Reprint, 5*, 463–480.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688–701.

Su, L., Z. Shi, and P. C. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*(6), 2215–2264.

Vytlacil, E. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica 70*(1), 331–341.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics 11*(3), 284–300.

Ward, J. H. J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association 58*(301), 236–244.

Windmeijer, F., H. Farbmacher, N. Davies, and G. D. Smith (2019). On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments. *Journal of the American Statistical Association 114*(527), 1339–1350.

Windmeijer, F., X. Liang, F. Hartwig, and J. Bowden (2021). The Confidence Interval Method for Selecting Valid Instrumental Variables. *Journal of the Royal Statistical Society: Series B Forthcoming.*

Appendices

A Simulations

A.1 Tables

Table A.1: Simulation: No invalid, Large Sample - Classification

α	2SLS	SE	Cons Class	# clubs	Rep K=4
0.001	28.27	0.84	0.34	3.62	0.53
0.01	28.27	0.84	0.47	3.84	0.68
0.05	28.27	0.84	0.52	4.03	0.73

Note: ConsClass indicates the fraction of times that all judges have been assigned to the correct club. # Clubs indicates the mean number of clusters selected by the algorithm. Rep $K = 4$ shows the fraction of times that the algorithm selected $K = 4$. In the first line, we show the results for the oracle estimator, i.e. when the club membership is known. The following LATE estimates are reported only when the number of clubs selected is 4.

Table A.2: Simulation results - No Invalidity, Small sample

Est	α	L1	SE	CI	P	HS	L2	SE	CI	P	HS	L3	SE	CI	P	HS	L4	SE	CI	P	HS	L5	SE	CI	P	HS	L6	SE	CI	P	HS
				cov.					cov.					cov.					cov.					cov.					cov.		
Oracle		34.68					28.38					25.74					18.47					16.38					12.21				
2SLS	0.001	34.48	1.75	0.98	1	0.48	29.72	2.03	0.88	1	0.48	25.92	1.41	1	1	0.48	17.88	2.77	0.98	0.96	0.5	16.48	1.34	0.98	1	0.48	13.05	1.83	0.86	1	0.49
	0.01	34.62	1.75	0.98	1	0.49	29.46	2.03	0.9	1	0.49	25.9	1.45	1	1	0.48	17.92	2.66	0.98	0.96	0.5	16.43	1.37	0.98	1	0.48	12.78	1.93	0.89	1	0.5
	0.05	34.68	1.74	0.98	1	0.48	29.15	1.99	0.92	1	0.48	25.89	1.49	1	1	0.48	18.06	2.53	0.99	0.97	0.49	16.39	1.4	0.99	1	0.48	12.65	2	0.91	1	0.49
Med.	0.001	34.41					29.71					25.85					18.03					16.46					13.04				
	0.01	34.53					29.44					25.82					18.16					16.42					12.82				
	0.05	34.59					29.15					25.81					18.21					16.36					12.65				
AHC	0.001	34.39	1.77	0.9	1	0.51	29.56	1.98	0.86	1	0.53	25.77	1.4	0.97	1	0.39	18.25	3.01	0.98	0.94	0.52	16.57	1.36	0.98	1	0.44	13.06	1.88	0.87	1	0.46
	0.01	34.44	1.75	0.9	1	0.52	29.27	1.97	0.88	1	0.53	25.69	1.44	0.97	1	0.41	18.46	2.87	0.99	0.95	0.52	16.56	1.39	0.99	1	0.45	12.81	1.97	0.89	1	0.46
	0.05	34.58	1.74	0.89	1	0.53	29	1.91	0.9	1	0.53	25.72	1.48	0.96	1	0.43	18.56	2.68	0.99	0.96	0.52	16.51	1.42	0.99	1	0.45	12.69	2.06	0.91	1	0.47

Note: 1000 iterations. L1-L6 indicate Local Average Treatment Effects calculated using one group as basis and the other one as focus group for a union of instruments. CI-cover indicates whether the 95% includes the true LATE. 2SLS is the mean estimate when using all IVs. HS indicates p-value of Hansen-Sargan test. SE is the corresponding mean standard error of the estimates. CI cov. stands for coverage rate of 95% CI. P stands for power, probability with which 0 is outside CI.

Table A.3: Simulation results - No Invalidity, Large sample

Est	α	L1	SE	CI	P	HS	L2	SE	CI	P	HS	L3	SE	CI	P	HS	L4	SE	CI	P	HS	L5	SE	CI	P	HS	L6	SE	CI	P	HS
				cov.					cov.					cov.					cov.					cov.					cov.		
Oracle		34.77					28.42					25.79					18.47					16.39					12.27				
2SLS	0.001	34.77	1.18	0.99	1	0.5	28.44	1.28	0.98	1	0.5	25.83	1.07	0.99	1	0.5	18.27	1.5	0.99	1	0.48	16.37	0.99	1	1	0.48	12.31	1.54	0.99	1	0.49
	0.01	34.78	1.18	0.99	1	0.5	28.42	1.28	0.98	1	0.5	25.83	1.09	0.99	1	0.5	18.25	1.5	0.99	1	0.48	16.37	1	1	1	0.48	12.31	1.56	0.99	1	0.49
	0.05	34.78	1.18	1	1	0.5	28.42	1.29	0.98	1	0.5	25.84	1.09	0.99	1	0.5	18.26	1.51	0.99	1	0.49	16.38	1.01	1	1	0.49	12.3	1.58	0.99	1	0.49
Med.	0.001	34.71					28.41					25.77					18.51					16.41					12.26				
	0.01	34.74					28.4					25.78					18.44					16.39					12.25				
	0.05	34.74					28.4					25.79					18.43					16.4					12.26				
AHC	0.001	34.66	1.2	0.87	1	0.48	28.36	1.27	0.96	1	0.51	25.76	1.07	0.97	1	0.49	18.5	1.54	0.99	1	0.51	16.43	1	1	1	0.49	12.31	1.57	0.99	1	0.49
	0.01	34.7	1.2	0.87	1	0.49	28.37	1.28	0.96	1	0.52	25.79	1.08	0.97	1	0.49	18.44	1.54	0.99	1	0.51	16.41	1.02	1	1	0.49	12.31	1.59	0.99	1	0.49
	0.05	34.72	1.2	0.87	1	0.51	28.38	1.29	0.96	1	0.52	25.8	1.09	0.97	1	0.5	18.44	1.55	0.99	1	0.51	16.42	1.02	1	1	0.5	12.3	1.61	1	1	0.49

Note: 1000 iterations. L1-L6 indicate Local Average Treatment Effects calculated using one group as basis and the other one as focus group for a union of instruments. CI-cover indicates whether the 95% includes the true LATE. 2SLS is the mean estimate when using all IVs. HS indicates p-value of Hansen-Sargan test. SE is the corresponding mean standard error of the estimates. CI cov. stands for coverage rate of 95% CI. P stands for power, probability with which 0 is outside CI.

Table A.4: Simulation: Few invalid, Small Sample - Classification

α	2SLS	SE	Cons Class	# clubs	Rep K=4
0.001	32.23	1.09	0.35	3.58	0.53
0.01	32.23	1.09	0.46	3.83	0.71
0.05	32.23	1.09	0.51	4.03	0.76

Note: ConsClass indicates the fraction of times that all judges have been assigned to the correct club. # Clubs indicates the mean number of clusters selected by the algorithm. Rep $K = 4$ shows the fraction of times that the algorithm selected $K = 4$. In the first line, we show the results for the oracle estimator, i.e. when the club membership is known. The following LATE estimates are reported only when the number of clubs selected is 4.

Table A.5: Simulation results - Few Invalid, Small sample

Est	α	L1	SE	CI	P	HS	L2	SE	CI	P	HS	L3	SE	CI	P	HS	L4	SE	CI	P	HS	L5	SE	CI	P	HS	L6	SE	CI	P	HS
				cov.					cov.					cov.					cov.					cov.					cov.		
Oracle		34.78					28.42					25.77					18.5					16.41					12.22				
2SLS	0.001	43.72	2.01	0.03	1	0.02	33.48	2.23	0.59	1	0.02	28.46	1.52	0.63	1	0.02	12.2	2.82	0.26	0.93	0	11.81	1.24	0.12	0.99	0	12.58	1.81	0.87	1	0.4
	0.01	43.79	2.02	0.02	1	0.02	33.41	2.31	0.61	1	0.02	28.53	1.56	0.63	1	0.02	11.95	2.9	0.25	0.92	0	11.83	1.28	0.14	0.99	0	12.39	1.9	0.88	0.99	0.4
	0.05	43.66	1.97	0.02	1	0.02	32.81	2.16	0.64	1	0.02	28.47	1.6	0.67	1	0.02	11.94	2.67	0.24	0.95	0	11.91	1.31	0.15	1	0	12.38	1.97	0.91	0.99	0.41
Med.	0.001	40.86					31.74					26.92					16.53					14.98					12.78				
	0.01	41.02					31.7					26.98					16					15.1					12.55				
	0.05	40.82					31.04					26.92					16.21					15.14					12.44				
AHC	0.001	35.38	2.09	0.84	1	0.47	29.91	2.03	0.84	1	0.49	25.99	1.4	0.93	1	0.41	35.38	2.09	0.84	1	0.47	29.91	2.03	0.84	1	0.49	25.99	1.4	0.93	1	0.41
	0.01	35.35	2.08	0.84	1	0.48	29.89	2.02	0.84	1	0.5	26.01	1.44	0.92	1	0.42	18.67	3.41	0.96	0.94	0.47	16.25	1.37	0.94	1	0.48	12.69	1.95	0.87	0.99	0.48
	0.05	35.18	2.04	0.85	1	0.5	29.52	1.97	0.86	1	0.52	25.99	1.47	0.93	1	0.44	18.65	3.14	0.96	0.96	0.48	16.23	1.4	0.95	1	0.49	12.62	2.03	0.91	1	0.49

Note: 1000 iterations. L1-L6 indicate Local Average Treatment Effects calculated using one group as basis and the other one as focus group for a union of instruments. CI-cover indicates whether the 95% includes the true LATE. 2SLS is the mean estimate when using all IVs. HS indicates p-value of Hansen-Sargan test. SE is the corresponding mean standard error of the estimates. CI cov. stands for coverage rate of 95% CI. P stands for power, probability with which 0 is outside CI.

Table A.6: Simulation: Few invalid, Large Sample - Classification

α	2SLS	SE	Cons Class	# clubs	Rep K=4
0.001	32.36	0.66	0.88	3.97	0.93
0.01	32.36	0.66	0.88	4.02	0.94
0.05	32.36	0.66	0.85	4.09	0.9

Note: ConsClass indicates the fraction of times that all judges have been assigned to the correct club. # Clubs indicates the mean number of clusters selected by the algorithm. Rep $K = 4$ shows the fraction of times that the algorithm selected $K = 4$. In the first line, we show the results for the oracle estimator, i.e. when the club membership is known. The following LATE estimates are reported only when the number of clubs selected is 4.

Table A.7: Simulation results - Few Invalid, Large sample

Est	α	L1	SE	CI	P	HS	L2	SE	CI	P	HS	L3	SE	CI	P	HS	L4	SE	CI	P	HS	L5	SE	CI	P	HS	L6	SE	CI	P	HS
				cov.					cov.					cov.					cov.					cov.					cov.		
Oracle		34.75					28.4					25.76					18.47					16.38					12.17				
2SLS	0.001	43.78	1.35	0	1	0	31.48	1.41	0.41	1	0	28.32	1.17	0.38	1	0	11.87	1.44	0.08	1	0	12.04	0.93	0.05	1	0	12.14	1.54	0.99	1	0.5
	0.01	43.79	1.35	0	1	0	31.43	1.42	0.43	1	0	28.33	1.18	0.38	1	0	11.91	1.42	0.08	0.99	0	12.01	0.94	0.06	1	0	12.16	1.56	0.99	1	0.5
	0.05	43.8	1.35	0	1	0	31.39	1.42	0.44	1	0	28.32	1.18	0.39	1	0	11.91	1.43	0.08	0.99	0	12.01	0.95	0.06	1	0	12.18	1.57	0.99	1	0.5
Med.	0.001	39.72					29.44					26.55					17.17					15.57					12.13				
	0.01	39.74					29.4					26.56					17.16					15.57					12.14				
	0.05	39.75					29.37					26.57					17.15					15.58					12.16				
AHC	0.001	34.96	1.39	0.88	1	0.51	28.56	1.31	0.95	1	0.51	25.88	1.07	0.96	1	0.46	18.52	1.57	0.99	1	0.5	16.39	1	1	1	0.48	12.2	1.57	0.99	1	0.48
	0.01	34.97	1.39	0.88	1	0.51	28.53	1.31	0.95	1	0.51	25.88	1.08	0.96	1	0.46	18.52	1.57	0.99	1	0.51	16.4	1.01	1	1	0.48	12.22	1.59	0.99	1	0.49
	0.05	34.97	1.39	0.88	1	0.52	28.51	1.32	0.95	1	0.52	25.88	1.09	0.96	1	0.47	18.5	1.58	0.99	1	0.51	16.4	1.01	1	1	0.48	12.22	1.6	0.99	1	0.49

Note: 1000 iterations. L1-L6 indicate Local Average Treatment Effects calculated using one group as basis and the other one as focus group for a union of instruments. CI-cover indicates whether the 95% includes the true LATE. 2SLS is the mean estimate when using all IVs. HS indicates p-value of Hansen-Sargan test. SE is the corresponding mean standard error of the estimates. CI cov. stands for coverage rate of 95% CI. P stands for power, probability with which 0 is outside CI.

Table A.8: Simulation: Many invalid, Small Sample - Classification

α	2SLS	SE	Cons Class	# clubs	Rep K=4
0.001	39.08	1.41	0.33	3.61	0.53
0.01	39.08	1.41	0.46	3.85	0.7
0.05	39.08	1.41	0.52	4.05	0.73

Note: ConsClass indicates the fraction of times that all judges have been assigned to the correct club. # Clubs indicates the mean number of clusters selected by the algorithm. Rep $K = 4$ shows the fraction of times that the algorithm selected $K = 4$. In the first line, we show the results for the oracle estimator, i.e. when the club membership is known. The following LATE estimates are reported only when the number of clubs selected is 4.

Table A.9: Simulation results - Many Invalid, Small sample

Est	α	L1	SE	CI	P	HS	L2	SE	CI	P	HS	L3	SE	CI	P	HS	L4	SE	CI	P	HS	L5	SE	CI	P	HS	L6	SE	CI	P	HS
				cov.					cov.					cov.					cov.					cov.					cov.		
Oracle		34.76					28.41					25.77					18.47					16.39					12.21				
2SLS	0.001	59.02	2.59	0.01	1	0	39.42	2.83	0.14	1	0	33.42	1.96	0.06	1	0	1.55	3.03	0.06	0.5	0	5.63	1.4	0.02	0.83	0	14.25	3.96	0.4	0.72	0.02
	0.01	58.94	2.6	0.01	1	0	39.28	2.9	0.16	1	0	33.37	2	0.08	1	0	1.35	3.08	0.06	0.51	0	5.62	1.42	0.02	0.84	0	14.64	4.2	0.41	0.71	0.02
	0.05	59.07	2.55	0.01	1	0	38.22	2.64	0.17	1	0	33.37	2.05	0.09	1	0	0.89	2.77	0.05	0.51	0	5.71	1.42	0.02	0.83	0	14.95	4.48	0.42	0.7	0.01
Med.	0.001	50.87					33.93					28.44					2.93					6.07					11.87				
	0.01	50.88					33.78					28.45					2.76					6.03					11.87				
	0.05	50.84					32.78					28.41					2.05					6.23					12.21				
AHC	0.001	37.34	2.22	0.73	1	0.32	29.7	2.32	0.79	1	0.37	25.46	1.63	0.85	1	0.23	14.18	3.71	0.68	0.9	0.37	12.79	1.56	0.54	0.98	0.27	10.72	3.04	0.55	0.78	0.27
	0.01	37.25	2.2	0.73	1	0.32	29.61	2.36	0.8	1	0.36	25.35	1.66	0.84	1	0.23	14.44	3.65	0.69	0.9	0.36	12.88	1.58	0.54	0.98	0.27	10.35	3.27	0.55	0.76	0.27
	0.05	37.08	2.18	0.74	1	0.34	29.08	2.27	0.82	1	0.36	25.24	1.69	0.83	1	0.24	14.05	3.27	0.67	0.91	0.35	12.69	1.61	0.52	0.97	0.26	9.88	3.56	0.54	0.74	0.26

Note: 1000 iterations. L1-L6 indicate Local Average Treatment Effects calculated using one group as basis and the other one as focus group for a union of instruments. CI-cover indicates whether the 95% includes the true LATE. 2SLS is the mean estimate when using all IVs. HS indicates p-value of Hansen-Sargan test. SE is the corresponding mean standard error of the estimates. CI cov. stands for coverage rate of 95% CI. P stands for power, probability with which 0 is outside CI.

Table A.10: Simulation: Many invalid, Large Sample - Classification

α	2SLS	SE	Cons Class	# clubs	Rep K=4
0.001	39.26	1.1	0.88	3.97	0.94
0.01	39.26	1.1	0.87	4.03	0.94
0.05	39.26	1.1	0.83	4.1	0.89

Note: ConsClass indicates the fraction of times that all judges have been assigned to the correct club. # Clubs indicates the mean number of clusters selected by the algorithm. Rep $K = 4$ shows the fraction of times that the algorithm selected $K = 4$. In the first line, we show the results for the oracle estimator, i.e. when the club membership is known. The following LATE estimates are reported only when the number of clubs selected is 4.

A.2 Figures

Figure A.1: Illustration of results, No invalid, Small sample

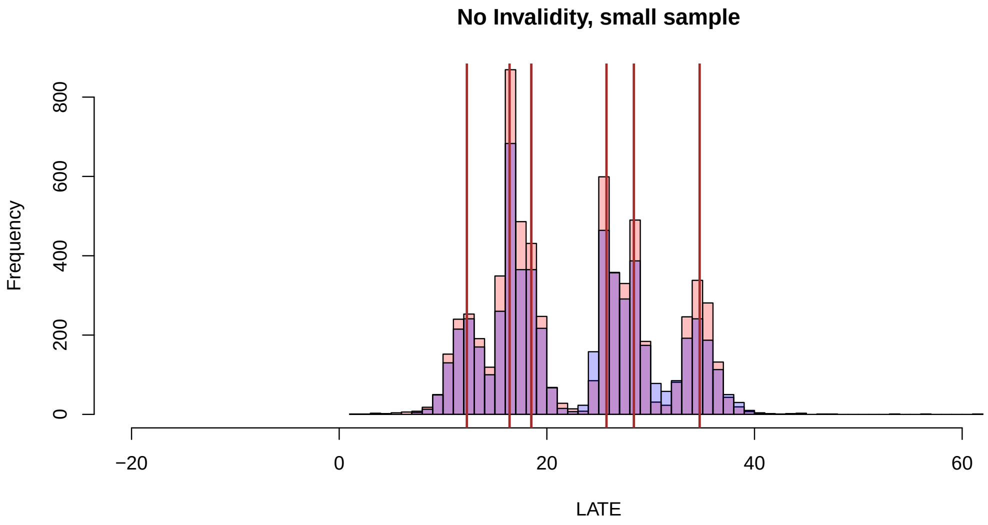


Figure A.2: Illustration of results, No invalid, Large sample RF

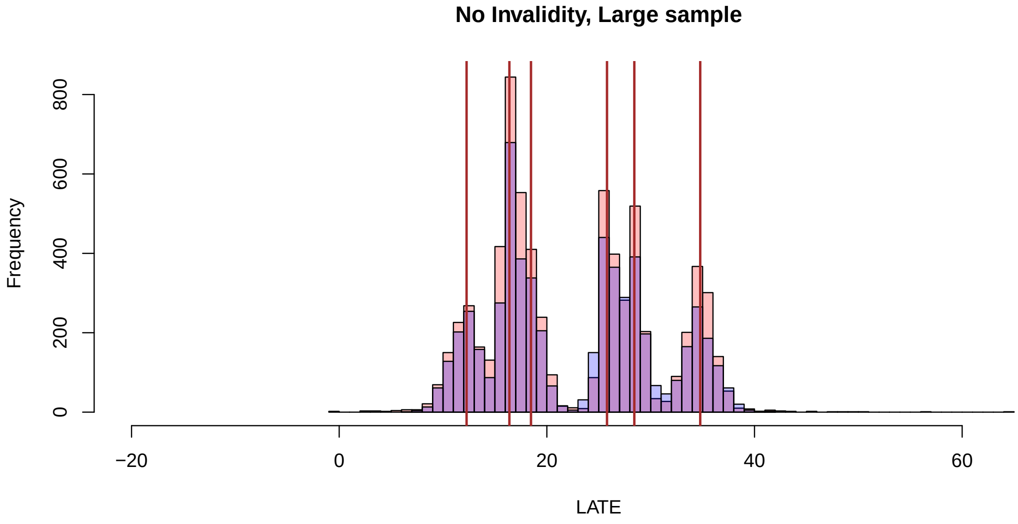


Figure A.3: Illustration of results, Few invalid, Small sample

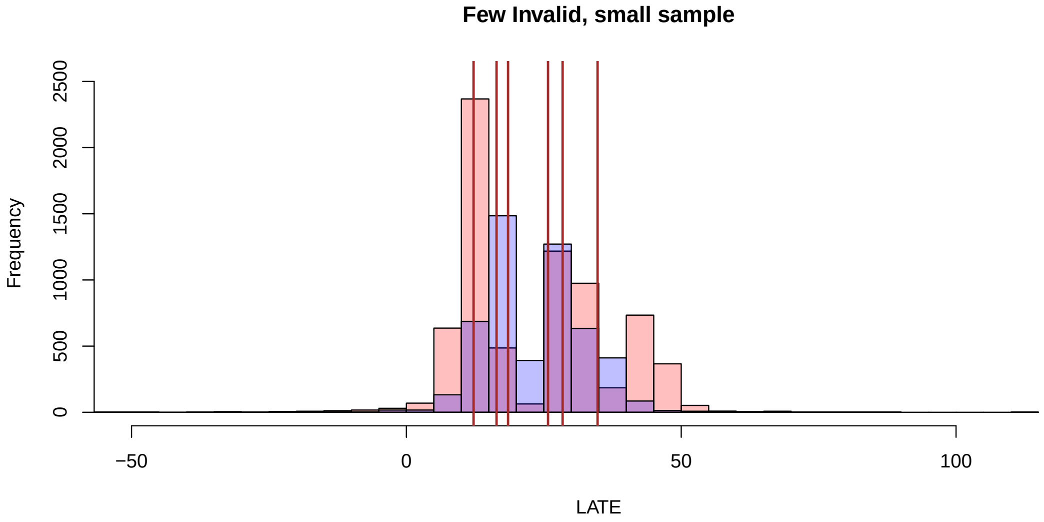


Figure A.4: Illustration of results, Few invalid, Large sample

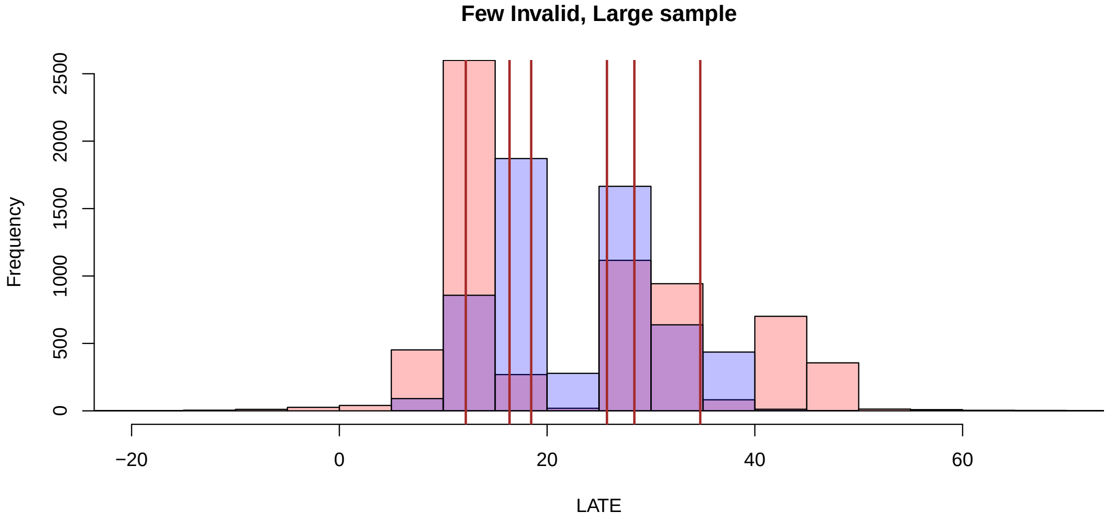
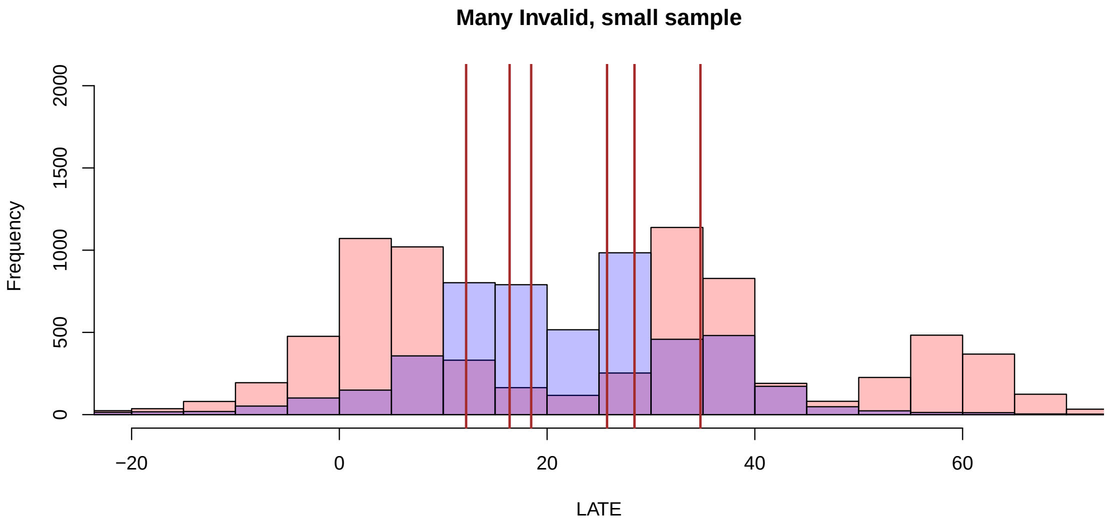


Figure A.5: Illustration of results, Many invalid, Small sample



B Application

B.1 Data

B.1.1 Descriptive Statistics

Variable	Overall	$D = 1$	$D = 0$	Diff	pval
Recidivism	0.292	0.3	0.219	0.08	4.4e-37
Female	0.183	0.173	0.276	-0.1	4.3e-53
Age at Sentence	33.04	33.02	33.2	-0.18	0.28
Race					
White	0.59	0.585	0.638	-0.05	6.9e-13
Black	0.261	0.267	0.204	0.06	6.3e-24
Amerindian	0.073	0.074	0.065	0.01	0.021
Hispanic	0.05	0.047	0.074	-0.03	5.6e-12
Asian	0.026	0.027	0.018	0.01	5e-05
Unknown	0	0	0	0	0.73
Crime Type					
Property Crime	0.332	0.324	0.412	-0.09	9.7e-32
Crime against a Person	0.299	0.306	0.235	0.07	8e-28
Drug Crime	0.238	0.242	0.2	0.04	7.5e-12
Sex Offenses	0.056	0.056	0.054	0	0.49
Weapons Offense	0.007	0.007	0.009	0	0.19
Other	0.068	0.065	0.091	-0.03	1.3e-09
Severity	3.44	3.46	3.25	0.21	6.1e-09

Table B.11: ‘Overall’, ‘D = 1’ and ‘D = 0’, report the mean of the respective variable in the total sample, among the treated and the non-treated in the baseline sample of 2009-14 criminal cases. ‘Diff’ and ‘p-val’ provide the mean difference (across treatment states) and the p-value of a two-sample t-test.

B.1.2 Crimes Considered in Baseline Sample

B.1.3 Robustness

Crime Type	Crime Category	N
Theft	Property	3416
Theft of Firearm	Property	89
Theft Over 35K	Property	144
Possession of Shoplifting Gear	Property	85
Theft From Pers	Property	301
Theft of Motor Vehicle	Property	159
Motor Vehicle Use Without Consent	Property	1341
Receiving Stolen Property	Property	1330
Arson, 1st degree	Property	10
Arson, 2nd degree	Property	64
Arson, 3rd degree	Property	25
Burglary, 1st degree (Occupied Dwelling)	Property	483
Burglary, 2nd degree (Dwelling/Bank/Government Building/ Religious Est./ Historic Property/School Building)	Property	1330
Burglary, 2nd degree (Pharmacy/Tool)	Property	212
Burglary 3rd degree (Non Residential)	Property	1954
Possesion of Burglary Tools	Property	552
Criminal Damage	Property	617
Check Forgery (Over $35,000)	Property	8
Check Forgery (Over $2,500)	Property	276
Check Forgery ($251-$2,500)	Property	1144
Check Forgery ($250 or Less)	Property	81
Other Forgery	Property	306
Issuance of Dishonored Checks	Property	422
Financial Transaction Card Fraud	Property	989
Welfare/Unemployment Benefits/Food Stamp Fraud	Property	303
Identity Theft	Property	177
Possession or Sale of Counterfeit Check	Property	133
Mail Theft	Property	48
Other Property Crimes	Property	238
Criminal Vehicular Homicide or Injury, severity=5	Person	92
Criminal Vehicular Homicide or Injury, severity=3	Person	270
Assault, 2nd degree	Person	970
Assault, 3rd degree	Person	1637
Assault, 4th degree	Person	705
Assault, 5th degree	Person	253
Domestic Assault	Person	2280
Domestic Assault by Strangulation	Person	1147
Simple Robbery	Person	420
Aggravated Robbery, 1st degree	Person	88
Aggravated Robbery, 2nd degree	Person	102
Kidnapping (Safe Release/No Great Bodily Harm)	Person	17
Kidnapping (Unsafe Release/Great Bodily Harm/Victim Under 16)	Person	1
False Imprisonment	Person	65
Depriving Another of Cust. or Parental Rights	Person	37
Coercion	Person	13
Accidents	Person	9
Malicious Punishment of Child	Person	101
Threats of Violence (Terror/Evacuation)	Person	2900
Threats of Violence (Replica Firearm/Bomb Threat)	Person	87
Stalking, severity=4	Person	121
Stalking, severity=5	Person	160
Violation of Harassment Restraining Order	Person	2941
Tampering with a Witness	Person	24
Burglary 1st Degree (w/Weapon or Assault)	Person	8
Prostitution	Person	31
Other Person	Person	133
Controlled Substance Crime, 2nd degree	Drug	18
Controlled Substance Crime, 3rd degree	Drug	1454
Controlled Substance Crime, 4th degree	Drug	584
Controlled Substance Crime, 5th degree	Drug	9322
Possession of Substances with Intent to Manufacture Methamphetamine	Drug	46
Other Drug Offense	Drug	206
Drive-by Shooting (Unoccupied Motor Vehicle/Building)	Weapon	9
Discharge of Firearm	Weapon	139
Other Weapon Related Crimes	Weapon	204
Criminal Sexual Conduct, 2nd degree	Sex	240
Criminal Sexual Conduct, 3rd degree	Sex	373
Criminal Sexual Conduct, 4th degree	Sex	283
Criminal Sexual Conduct, 5th degree	Sex	4
Solicitation of Minors to Engage in Sexual Conduct	Sex	94
Failure to Register as Predator	Sex	1380
Possession of Child Pornography	Sex	334
Other Sex Crimes	Sex	12
Bribery	Other	13
Perjury	Other	35
Escape, severity=3	Other	240
Fleeing a Police Officer	Other	1660
Aiding Offender to Avoid Arrest	Other	124
Accomplice After the Fact	Other	54
Obstruction of Legal Procedure	Other	26
Lottery Fraud	Other	38
Felony Driving while Intoxicated	Other	530
Failure to Appear in Court	Other	65
Tax Offenses	Other	72
Other	Other	441

Table B.12: Overview of the offenses considered in the application, the broader crime category to which each offense belongs, and the number of observations per offense in the baseline sample of 2009-14 criminal cases.

Variable	Coefficient Estimate	Standard Error	Sign. Level
Female	-3e-04	7e-04	
Age at Sentence	0	0	
Black	3e-04	5e-04	
Amerindian	9e-04	6e-04	
Hispanic	-3e-04	9e-04	
Asian	0.001	8e-04	
Unknown	0.0065	0.0087	
Crime against a Person	0.0019	0.0028	
Drug Crime	-5e-04	8e-04	
Sex Offense	0.0024	0.0026	
Weapons Offense	0.0042	0.0035	
Other	0.0014	0.0023	
Severity	0	1e-04	
(Intercept)	0.9378	0.0102	***
F-Statistic for Joint Test		0.9436	
(p-value)		(0.5059)	

Table B.13: The estimation includes controls for fully interacted county - year fixed effects. The omitted category for race is “White", the one for crime type is “Property Crime". The standard errors are two-way clustered at judge and offender level.