I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18274

Understanding IV Versus OLS Estimates of Treatment Effects and the Coefficient Difference Check

David Bjerk

NOVEMBER 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18274

Understanding IV Versus OLS Estimates of Treatment Effects and the Coefficient Difference Check

David Bjerk
Claremont McKenna College and IZA

NOVEMBER 2025

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Understanding IV Versus OLS Estimates of Treatment Effects and the Coefficient Difference Check

This article derives an equation characterizing the difference between OLS and IV coefficients under potentially heterogenous treatment effects. This leads to what I call the Coefficient Difference Check, which consists of checking that the difference between the estimated OLS and IV coefficients has the same sign as the expected selection effect. I show failures of this check can arise because: IV is invalid, the expected selection story is incorrect, or there are particular heterogenous treatment effects that imply the IV estimate is both "fragile" and that it provides a more biased estimate of the ATT than OLS. Failures of this check are relatively common in the literature. I describe best practices given such failures.

Keywords: instrumental variables, selection, heterogenous treatment effects, average treatment-on-the-treated, judge fixed-effects, adjudicator propensity to treat IV, examiner tendency IV, returns to schooling

Corresponding author:
David Bjerk
Department of Economics
Claremont McKenna College
500 East Ninth Street
Claremont, CA 91711
USA

E-mail: david.bjerk@cmc.edu

1 - Introduction

A common goal in empirical work is to determine the effect of a given treatment on some outcome of interest. However, in many settings, there might be important selection effects with respect to who gets the treatment, where those who are treated might plausibly have different expected outcomes than the untreated, even in the absence of the treatment. When such selection occurs with respect to characteristics unobservable to the researcher, using Ordinary Least Squares (OLS) to compare the outcomes among the treated to the outcomes among the untreated, even after controlling for observable covariates, will give a biased estimate any treatment effects, as it will conflate treatment effects with selection effects associated with who receives treatment.

To overcome such a selection bias from OLS, researchers often employ an Instrumental Variable (IV) approach. Namely, a researcher finds a variable that is correlated with receipt of the treatment, but argues that this variable is only related to the outcome of interest through how it relates to the treatment (i.e., this variable can be "excluded" from having a direct relation with the outcome of interest). To the extent this "exclusion" restriction is plausible in a given application, the IV is used to obtain a theoretically unbiased estimate of the average treatment effect among those whose treatment status would vary depending on their value of the IV---a group generally referred to as the *compliers* (Angrist, Imbens and Rubin 1996).

Notably, though, this paper shows that given a valid IV for a binary treatment, the expected value of the OLS coefficient on the treatment will equal the sum of the selection effect plus a weighted average of the IV coefficient on the treatment (i.e., the mean treatment effect among the *compliers*) and the mean treatment effect of the *always-treated* (i.e., those who get the treatment regardless of their value of the instrument), where the weight on the IV coefficient is equal to the fraction of the treated who are *compliers*.

This result leads to a simple procedure that I call the *Coefficient Comparison Check*. Namely, I argue that when there is positive expected selection into the treatment (i.e., those treated would have greater expected outcomes than the untreated in the absence of treatment), the researcher should check whether the OLS coefficient on the treatment is greater than the IV coefficient. By contrast, when there is negative expected selection into the treatment (i.e., those treated would have lower expected outcomes than the untreated in the absence of treatment), the researcher should check whether the OLS coefficient on the treatment is less than the IV coefficient.

Due to the relationship between the OLS coefficient and the IV coefficient described above, I show that in the absence of significant measurement error with respect to who is treated, a failure of this *Coefficient Comparison Check* implies one of three things: (i) the required conditions for the IV to be valid do not hold, (ii) the actual selection effect is the opposite of what is expected, or (iii) the treatment effects among the *always-treated* must strongly differ from the treatment effects among the *compliers* in the exact opposite way of the selection effect. This last explanation becomes less plausible the closer the fraction of *compliers* among the treated is to one. Moreover, if it is this heterogenous treatment effects story that causes the *Coefficient Comparison Check* to fail, then the estimated IV coefficient on the treatment must be quite fragile in the sense that it will almost certainly change notably given even small alterations to the group of *compliers* arising from modest changes in the IV or the setting.

Using an example with real data where I employ a candidate IV that is almost surely invalid due to implausibility of the required exclusion restriction, I show that this *Coefficient Comparison Check* indeed fails to hold, highlighting the invalidity of the IV in this context. However, I explicitly call this procedure a *check* rather than a *test*, because as stated above, failures do not necessarily imply that the IV results are invalid. However, the argument above suggests a failure does constitute a red flag in that it implies that there is an issue with respect to the IV results that the researcher should explicitly acknowledge and consider which of the above reasons is the most plausible cause for the failure.[1]

I want to be clear that what I am proposing is not necessarily a wholly new concept. Indeed, most researchers who employ IV methods likely already have some awareness of the basics of my argument. However, despite being very intuitive and exceedingly easy to perform, this *Coefficient Comparison Check* is often violated and/or not discussed in papers

[1] In some situations, a researcher may posit that there could be significant selection with respect to the treatment, but it is ex-ante ambiguous whether that selection is positive or negative. In such situations, it still seems reasonable to compare the OLS to IV results as proposed here and take a stand on exactly what story is behind the difference between the two.

that employ an IV approach. I highlight this issue by examining a large swath of highly visible publications that employ one particular type of IV, namely adjudicator propensity-to-treat IVs (otherwise referred to as "judge fixed-effects" or "examiner tendency" IVs). While all of the papers I consider were published in top journals, I show that about half fail the *Coefficient Comparison Check*, or do not provide the OLS results necessary to perform this check. I discuss the ramifications of such failures with respect to a couple of these papers in more detail. I also discuss how failures of this check are also not infrequent in IV settings outside the adjudicator propensity-to-treat IV approach.

In the end, when using an IV approach, current best practices require authors to show that their chosen IV is significantly correlated with the treatment of interest in the expected way (i.e., show a valid "first-stage"). If the IV is revealed to have little correlation with the treatment, or correlated in the opposite way than expected given the motivation for the IV, then the researcher must address these issues or the IV results are generally treated with ample suspicion.[2]

I argue that there should be a similar expectation that researchers employing an IV approach should also conduct and discuss the *Coefficient Comparison Check*. This is not a large ask given this test simply consists of comparing the OLS to the IV coefficient on the treatment and determining whether the difference between them goes in the way of the expected selection or not—no new statistical code even needs to be applied. In cases where the results fail this check, researchers should then either: favor the OLS results over the IV results, credibly argue that the selection story is the opposite of what is generally expected, or discuss what the necessary heterogenous treatment effects must be and argue that such heterogeneity is indeed plausible.

Notably though, when a researcher argues that treatment effect heterogeneity is the reason why the *Coefficient Comparison Check* fails, the researcher should also then acknowledge that the estimated parameter is not very general in the sense that it would likely change substantially with modest changes in the *complier* group. Additionally, when the *Coefficient Comparison Check* fails, the researcher might consider performing some of

[2] Indeed, Angrist and Kolesar (2024) argue that one should only employ an IV if the sign of the first-stage correlation between the instrument and the treatment goes in the expected direction.

the other more involved IV validity tests and assessments that have been developed previously, several of which I discuss later in this paper.

2 – Characterizing OLS Relative to IV Estimates of a Treatment Effect

Suppose one is interested in the effect of some binary treatment $D \in \{0,1\}$ on some outcome *Y*. For simplicity, suppose the true model of outcome *Y* is given by

$$Y_i = \alpha + \delta_i D_i + \beta u_i + \varepsilon_i, \qquad (1)$$

where δ_i is how receiving the treatment will impact the outcome for individual *i*, and u_i is some characteristic of individual *i* that is *unobservable* to the researcher. Suppose this variable *u* comes from a distribution *F(u)* over the population of interest. Lastly, assume that the residual term ε_i is an individual specific independent random variable. Note that one could also include a vector X_i of other characteristics observable to the researcher that influence the outcome *Y* to equation (1), but this would just add excess notation, so I forgo doing that here. In this vein, in this context, one can think of the outcome *Y* in equation (1) as the residual from regressing the true outcome of interest, $\tilde{Y}$, on the vector X_i.

In the potential outcomes framework used by Angrist, Imbens, and Rubin (1996), let $Y_i(1)$ equal what individual *i*'s outcome would be if he receives the treatment ($D_i = 1$) and $Y_i(0)$ equal what individual *i*'s outcome would be if he stays untreated ($D_i = 0$). Given equation (1), this means

$$Y_i(1) = \alpha + \delta_i + \beta u_i + \varepsilon_i \qquad (1a)$$

and

$$Y_i(0) = \alpha + \beta u_i + \varepsilon_i, \qquad (1b)$$

implying $Y_i(1) - Y_i(0) = \delta_i$, again meaning that δ_i is the treatment effect for individual *i.*

Suppose whether or not some individual *i* receives the treatment *D* is determined by the following equation

$$D_i = \begin{cases} 0 \; if \;\; \lambda Z_i + \theta u_i - \rho < 0 \\ 1 \; if \;\; \lambda Z_i + \theta u_i - \rho \geq 0, \end{cases} \qquad (2)$$

where *Zi* is some observable individual characteristic that influences whether or not an individual is treated via $\lambda > 0$, *ui* is the same unobservable variable that was in the outcome equation (1), and ρ is just a scaling parameter. [3] Given this set-up, a person *i* whose realized value of *Zi* = *z* receives the treatment if and only if $u_i \geq u_z^*$, where

$$u_z^* = \frac{\rho - \lambda z}{\theta}. \qquad (3)$$

In words, for any given value *z*, there is a cutoff such that a person *i* with *Zi* = *z* will be treated if and only if his/her (unobserved) value of *ui* exceeds this cutoff. Given each person's *ui* is drawn from a distribution *F(u)*, this means that among individuals with a value of *Zi* = *z*, the fraction treated will be $1 - F(u_z^*)$. Moreover, from equation (3), we can see that the higher the values of *z*, the lower will be u_z^*. This then means that the likelihood an individual is treated is increasing in both the unobservable characteristics *u* and this observable characteristic *Z*.

To make explication easier, suppose *Z* can only take on two values, z1 and z0, where z1 > z0. This implies two cutoffs, $u_1^* = \frac{\rho - \lambda z_1}{\theta}$ and $u_0^* = \frac{\rho - \lambda z_0}{\theta}$, where $u_1^* < u_0^*$. This is only for simplicity though, all arguments hold if *Z* can take on more than two values.[4] These two cutoff values u_1^* and u_0^* allow us to divide the population into three distinct groups. Every person *i* with $u_i \leq u_1^*$ will not be treated regardless of their value of *Z*, a population referred to as the *never-treated*. Analogously, everyone with $u_0^* < u_i$ will be treated regardless of their value of *Z,* a population referred to as the *always-treated*.[5] Finally, everyone with $u_1^* < u_i \leq u_0^*$ will be treated if they have *Zi* = z1 and be untreated if *Zi* = z0, a population generally referred to as *compliers* and consists of those whose treatment status fluctuates depending on their value of *Z*. Note that the composition of each of these groups are specific to the particular variable *Z* in question. A variable *Z* with different possible values would mean the *compliers* would be composed of a different subgroup of *u*, a detail I will return to below.

[3] Note that selection into the treatment based on the size of the individual's specific treatment effect can be easily accommodated here. Namely, that model would correspond to defining the individual treatment effect δ_i as θu_i in equation (1). This has no impact on the argument that follows.

[4] With more than two values of Z, there are correspondingly more cutoff values of *u*. However, the only ones that really matter for the argument here are the two extreme values of Z, and their corresponding cutoff values.

[5] In the literature, these groups are generally referred to as *never-takers* and the *always-takers*, respectively. However, I think the terminology I use here is more intuitive.

Under this setup, let us suppose the researcher's goal is identify the average treatment effect δ_i over some well-defined portion of the population. To do this, first consider a simple OLS analysis. Note that given u_i is unobservable to the researcher, it can't be directly included in any regression analysis. This means that if one were to regress *Y* on the treatment *D*, one would be effectively estimating

$$Y_i = \alpha + \delta^{OLS} D_i + \eta_i,$$

where $\eta_i = \beta u_i + \varepsilon_i$. In other words, the impact of the unobserved variable *u* on the outcome of interest *Y* is included in the error term, which means this OLS error term will be correlated with the treatment variable, thus causing δ^{OLS} to be a biased estimate of the average treatment effect for any particular group of individuals. I will return to this OLS estimate in more detail below, but first let us recall how a variable like *Z* can potentially be used as an Instrumental Variable (IV) for uncovering an unbiased average treatment effect for a specific subset of the population.

The Estimated Treatment Effect Under a Valid IV

As is well known, the required aspects for a variable *Z* to be a valid IV in this context are the following (Angrist, Imbens, and Rubin 1996):

1. *SUTVA* – A person's outcome is unaffected by any other person's treatment status.[6]
2. *Random Assignment* – Asymptotically, the distribution of observable and unobservable characteristics are the same for those with $Z_i = z_0$ and $Z_i = z_1$.
3. *Relevance* – The variable *Z* is sufficiently correlated with the treatment.
4. *Monotonicity* – If person *i* would be treated with $Z_i = z_0$, he/she would also be treated if $Z_i = z_1$, and, if a person *i* is untreated with $Z_i = z_1$, he/she would also be untreated with $Z_i = z_0$.[7]
5. *Exclusion* – the variable *Z* has no direct relation with the outcome *Y*, rather it is only correlated with *Y* through its correlation with the treatment.

[6] SUTVA stands for *Stable Unit Treatment Value Assumption*.

[7] The version laid out here is equivalent to what is generally called "pairwise" monotonicity. Frandsen et al. (2023) show that this condition can be weakened to "average" monotonicity in most situations.

In our model above, *SUTVA* holds since according to equation (1), each person *i*'s outcome Y_i is unaffected by anyone else's treatment. *Random assignment* holds since the distribution of *u* does not depend on *Z*. *Relevance* holds in the sense that the fraction of those with $Z_i = z$ who are treated equals $1 - F(u_z^*)$, and since $u_1^* < u_0^*$, this fraction will be higher for those with $Z_i = z_1$ than those with $Z_i = z_0$ as long as λ is greater than zero in equation (2) and z_0 is sufficiently different than z_1, which both hold in our model above. *Monotonicity* holds in our model given the treatment selection mechanism laid out above, as $u_1^* < u_0^*$, and those with $Z_i = z_1$ are treated if and only if $u_i > u_1^*$, while those with $Z_i = z_0$ are treated if and only if $u_i > u_0^*$, This leaves the *exclusion* restriction. In this context, the *exclusion* restriction holds if and only if Z_i has not direct relation to the outcome Y_i.

Given these assumptions hold, let's consider intuitively what it means to use the variable *Z* as in IV. Figure 1 depicts the n*ever-treated, compliers*, and *always-treated* for a specific valid IV under different forms of selection and different treatment effects. For simplicity, within each of these graphs, I have depicted the treatment effects to be identical across the distribution of *u,* meaning $\delta_i = \delta$ for all *i*. However, this is only for convenience. I consider heterogenous treatment effects across the distribution of *u* in more detail below.

Each graph in Figure 1 depicts the conditional expectation functions $\mathrm{E}[Y(0)|u]$ and $\mathrm{E}[Y(1)|u]$ in different situations. In each, the dark lines capture these conditional expectation functions for individuals with $Z_i = z_1$, while the lighter lines capture these conditional expectation functions for individuals with $Z_i = z_0$. The *exclusion* restriction implies that these conditional expectation functions are identical, and hence they overlap. The solid portions of these lines depict the conditional expectation of the outcome *Y* associated with each value of *u* given the realized treatment status*,* while the dashed portions of these lines depict the counterfactual expected outcome for each value of *u* for the opposite of the realized treatment status.

This all means that, in each figure, those with u_i to the left of u_1^* are the *never-treated*, and hence have the solid portion associated with $\mathrm{E}[\mathrm{Y}(0)|u]$. Those with u_i to the right of u_0^* are the *always-treated,* and hence have the solid portion associated with $\mathrm{E}[Y(1)|u]$. Finally, those with $u_1^* < u_i \leq u_0^*$ are the *compliers*, and hence those with $Z_i = z_1$ have the solid portion associated with $\mathrm{E}[Y(1)|u]$ and the dashed portion associated with $\mathrm{E}[Y(0)|u]$, while the opposite is true for those with $Z_i = z_0$.

Figure 1a depicts *positive selection* and *positive treatment effects*. It is positive selection in that $\mathrm{E}[Y(0)|u]$ is increasing in u (equivalent to $\beta > 0$ in equation (1)), and positive treatment effect in that $\mathrm{E}[Y(1)|u] > \mathrm{E}[Y(0)|u]$ for all u (equivalent to $\delta_i = \delta > 0$ in equation (1)). Figure 1b depicts *positive selection* and *negative treatment effects*, as $\mathrm{E}[Y(0)|u]$ is increasing in u, but the treatment effect is negative in that $\mathrm{E}[Y(1)|u] < \mathrm{E}[Y(0)|u]$ for all u (equivalent to $\delta_i = \delta < 0$ in equation (1)). Figure 1c then depicts *negative selection* and *positive treatment effects*, in that $\mathrm{E}[Y(0)|u]$ is decreasing in u (equivalent to $\beta < 0$ in equation (1)), but $\mathrm{E}[Y(1)|u] > \mathrm{E}[Y(0)|u]$ for all u. Finally, Figure 1d depicts *negative selection* and *negative treatment effects*.

Now consider estimating $\mathrm{E}[Y|Z = z_1] - \mathrm{E}[Y|Z = z_0]$, or the difference in expected outcomes between those with $Z = z_1$ and those with $Z = z_0$.[8] This is often what is referred to as the *Reduced Form* estimate and is therefore denoted δ^{RF}. As can be seen in each of these graphs in Figure 1, if we first look at the *never-treated* and the *always-treated* and compare the realized expected outcomes of those with $Z = z_1$ (the solid components of the dark lines) to the realized expected outcomes for those with $Z = z_0$ (the solid components of the lighter lines), we see that their difference is zero for the *never-treated* and *always-treated*. This just leaves the difference in expected outcomes among the *compliers* with $Z = z_1$, who are all treated, and the *compliers* with $Z = z_0$, who are all untreated.

Therefore, regardless of whether there is negative or positive selection into the treatment, or whether or not the treatment effects are positive or negative, if Z is a valid IV, then comparing the average outcomes of those with $Z = z_1$ to those with $Z = z_0$ will capture a weighted average of zero and the average treatment effect among the *compliers*, where the weight on the average treatment effect among the *compliers* equals the fraction of the population that is *compliers*. Noting that the average treatment effect among the *compliers* is $E[\delta_i | u_1^* < u \leq u_0^*]$, and the fraction of the population that are *compliers* is given by $F(u_0^*) - F(u_1^*)$, we get

$$\delta^{RF} = [F(u_0^*) - F(u_1^*)]E[\delta_i | u_1^* < u \leq u_0^*] + \left[1 - \left(F(u_0^*) - F(u_1^*)\right)\right] * 0. \quad (4)$$

[8] If we assume X_i is not the null set, then these expectations would also condition on X_i.

Thus, we can obtain the actual average treatment effect among the *compliers* by upweighting this Reduced Form estimate by the inverse of the fraction of the population who are *compliers* (i.e., $F(u_0^*) - F(u_1^*)$), giving

$$\frac{\delta^{RF}}{F(u_0^*) - F(u_1^*)} = E[\delta_i | u_1^* < u \leq u_0^*].$$

If we simplify our notation a bit by defining the fraction of the population who are *compliers* as $\pi_C \equiv F(u_0^*) - F(u_1^*)$, then we can define the following:

$$\delta^{IV} \equiv \frac{\delta^{RF}}{\pi_c} = E[\delta_i | u_1^* < u \leq u_0^*]. \qquad (5)$$

This just reveals the well-known result that if Z is a valid instrumental variable, δ^{IV} is the local average treatment effect (LATE) among the compliers, or the average treatment effect among those whose treatment status is impacted by this particular instrument (Angrist, Imbens, and Rubin, 1996). See the appendix for a formal derivation of this IV estimator for this set-up.

Reconsidering the OLS Estimate of a Treatment Effect Relative to IV Estimate

Let us again consider again the OLS regression

$$Y_i = \alpha + \delta^{OLS} D_i + \varepsilon.$$

It will be asymptotically true that $\delta^{OLS} = E[Y|D = 1] - E[Y|D = 0]$, or in words, the OLS estimate will asymptotically capture the difference between the expected outcome among those actually treated and the expected outcome among those actually untreated. Calculating the expected outcome among the actually treated, which consists of the *compliers* for those with $Z_i = z_1$ and the *always-treated* regardless of their Z_i, we get

$$\begin{aligned}
E[Y|D=1] = {} & \frac{\pi_z[F(u_0^*) - F(u_1^*)]}{\pi_z(1-F(u_1^*)) + (1-\pi_z)(1-F(u_0^*))} E[Y(1)|u_1^* < u \leq u_0^*] \\
& + \frac{\pi_z[1-F(u_0^*)]}{\pi_z(1-F(u_1^*)) + (1-\pi_z)(1-F(u_0^*))} E[Y(1)|u_0^* \leq u] \\
& + \frac{(1-\pi_z)[1-F(u_0^*)]}{\pi_z(1-F(u_1^*)) + (1-\pi_z)(1-F(u_0^*))} E[Y(1)|u_0^* < u].
\end{aligned}$$

where π_z is the fraction of the population for whom $Z = z_1$. To simplify this equation a little bit, let us make the following definitions:

$\pi_C \equiv F(u_0^*) - F(u_1^*)$, or the fraction of the population who are *compliers* (as above).
$\pi_A \equiv 1 - F(u_0^*)$, or the fraction of the population who are *always-treated.*
$\pi_N \equiv F(u_1^*)$, or the fraction of the population who are *never-treated.*

Given these definitions and a little manipulation, the previous equation can be simplified to

$$E[Y|D=1] = \frac{\pi_z\pi_C}{\pi_z\pi_C + \pi_A} E[Y(1)|u_1^* < u \leq u_0^*] + \frac{\pi_A}{\pi_z\pi_C + \pi_A} E[Y(1)|u_0^* < u]. \qquad (6)$$

Not surprisingly, the above equation shows that the expected outcome among the treated is a weighted average of the expected outcome among the treated *compliers* (first term in equation above) and the *always-treated* (second term in equation above).

Doing a similar calculation for the expected outcome among the untreated we get

$$\begin{aligned}
E[Y|D=0] = {} & \frac{(1-\pi_z)[F(u_0^*) - F(u_1^*)]}{\pi_z F(u_1^*) + (1-\pi_z)F(u_0^*)} E[Y(0)|u_1^* < u \leq u_0^*] \\
& + \frac{(1-\pi_z)F(u_1^*)}{\pi_z F(u_1^*) + (1-\pi_z)F(u_0^*)} E[Y(0)|u < u_1^*] \\
& + \frac{\pi_z F(u_1^*)}{\pi_z F(u_1^*) + (1-\pi_z)F(u_0^*)} E[Y(0)|u \leq u_1^*],
\end{aligned}$$

which, given the above definitions, simplifies to

$$E[Y|D=0] = \frac{(1-\pi_z)\pi_C}{\pi_N + (1-\pi_z)\pi_C} E[Y(0)|u_1^* < u \leq u_0^*] + \frac{\pi_N}{\pi_N + (1-\pi_z)\pi_C} E[Y(0)|u \leq u_1^*]. \quad (7)$$

This equation shows that the expected outcome among the untreated is a weighted average of the expected outcome among the untreated *compliers* (first term in equation above) and the *never-treated* (second term in equation above).

Therefore, given $\delta^{OLS} = E[Y|D=1] - E[Y|D=0]$, it will be true that δ^{OLS} will simply be equation (6) minus equation (7), or

$$\begin{aligned}\delta^{OLS} = {} & \frac{\pi_z\pi_C}{\pi_z\pi_C + \pi_A} E[Y(1)|u_1^* < u \leq u_0^*] + \frac{\pi_A}{\pi_z\pi_C + \pi_A} E[Y(1)|u_0^* < u] \\ & - \frac{(1-\pi_z)\pi_C}{\pi_N + (1-\pi_z)\pi_C} E[Y(0)|u_1^* < u \leq u_0^*] - \frac{\pi_N}{\pi_N + (1-\pi_z)\pi_C} E[Y(0)|u \leq u_1^*].\end{aligned}$$

While I leave the subsequent derivation details to the appendix, given the data generation process in equation (1), it can be shown that the above equation can be represented as

$$\begin{aligned}\delta^{OLS} = {} & \pi_{D=1}^C \delta^{IV} + (1-\pi_{D=1}^C)\delta^A \\ & + E[Y(0)|D=1] - E[Y(0)|D=0] \qquad (8),\end{aligned}$$

where $\pi_{D=1}^C$is defined to be the fraction of the treated who are *compliers* (i.e., $\pi_{D=1}^C = \frac{\pi_z\pi_C}{\pi_z\pi_C+\pi_A}$) and δ^A is the average treatment effect among the *always-treated* (i.e., average treatment effect for those with $u_i > u_0^*$). The lower line in equation (8) is simply the expected (latent) outcome among the treated *if they hadn't received the treatment* minus the expected outcome for the untreated in the absence of treatment. In other words, this lower term is simply the *selection effect* associated with the treatment.

Therefore, equation (8) shows that δ^{OLS} is a weighted average of the average treatment effect among the *compliers* (δ^{IV}) and the average treatment effect among the *always-treated* (δ^A), plus any selection effect. This just confirms that even with heterogenous treatment effects, if there are any selection effects, then δ^{OLS} will be a biased measure of the average treatment effect for any definable group.

Moving on, if we subtract δ^{IV} from both sides of equation (8) and call the second row of equation (8) the *selection effect*, we get

$$\delta^{OLS} - \delta^{IV} = (1 - \pi^{C}_{D=1})(\delta^{A} - \delta^{IV}) + selection\ effect \qquad (9).$$

Note that given the *selection effect* is defined to be $E[Y(0)|D = 1] - E[Y(0)|D = 0]$, there will be *positive* selection into the treatment if β > 0 in equation (1) and *negative* selection if β < 0 in equation (1). Equation (9) directly leads to the following theorem.

OLS vs IV Theorem: If the required assumptions with respect to the proposed IV hold, and there is negligible measurement error with respect to the treatment, then regardless of the sign of δ^{IV}:

1. When there is positive selection, then $\delta^{OLS} - \delta^{IV} < 0$ only if (i) the fraction of the treated who are *compliers* ($\pi^{C}_{D=1}$) is relatively small, and (ii) the average treatment effect among the *always-treated* is so much lower than it is among the *compliers* (i.e., $\delta^{A} - \delta^{IV} < 0$) that it offsets the selection effect.
2. When there is negative selection, then $\delta^{OLS} - \delta^{IV} > 0$ only if (i) the fraction of the treated who are *compliers* is relatively small, and (ii) the average treatment effect among the *always-treated* is so much higher than it is among the *compliers* (i.e., $\delta^{A} - \delta^{IV} > 0$) that it offsets the selection effect.

Note that one implication of equation (9) and the theorem above is that even with no selection effect, the OLS and IV coefficients might differ substantially given large heterogeneities in treatment effects. While derived and expressed differently, this is similar to what is discussed in Ishimaru (2022), which in turn builds on Lochner and Moretti (2015).[9]

[9] Also, note that if $\pi^{C}_{D=1}$ goes to one, meaning everyone is a *complier*, then equation (9) also shows the OLS coefficient converges to the IV coefficient since there are no selection effect when everyone is a *complier*.

3 – The Coefficient Comparison Check

The *OLS vs IV Theorem* leads to the following:

Coefficient Comparison Check: Given a plausible IV, estimate $\hat{\delta}^{OLS}$ and $\hat{\delta}^{IV}$. If there is little reason to believe there is measurement error with respect to who is deemed treated, then:

1. If one expects *positive* selection, check whether $\hat{\delta}^{OLS} - \hat{\delta}^{IV} > 0$.
2. If one expects *negative* selection, check whether $\hat{\delta}^{OLS} - \hat{\delta}^{IV} < 0$.

If the estimated coefficients fail this check, then the researcher must either:

(i) Argue that selection goes the opposite way than expected, or

(ii) Argue that the average treatment effect among the *always-takers* not only differs substantially from that of the *compliers*, but also that this difference goes in the opposite direction of the selection-effects, or

(iii) Conclude the IV is invalid due to failure of monotonicity or exclusion restrictions.

As the fraction of the treated who are *compliers* rises, the heterogenous treatment effects option (ii) becomes less and less plausible. Moreover, if one argues that option (ii) is the reason for the failure, one must acknowledge that the estimated $\hat{\delta}^{IV}$ is fragile in the sense it can change dramatically if there are expansions or contractions with respect to who is included among the *compliers*.

Note that the above check is with respect to real values, not absolute values. Moreover, this check works for both positive and negative selection and positive and negative treatment effects. For example, if one expects there is *positive selection* and finds a *positive treatment effect* among the *compliers* (i.e., $\hat{\delta}^{IV} > 0$), but *fails* to find that $\hat{\delta}^{OLS} - \hat{\delta}^{IV} > 0$, then if the researcher wants to "keep" the IV results, he/she must either argue selection is actually negative, or argue that the treatment effects among the *always-treated* are much less positive than they are among the *compliers*.

By contrast, if one expects there is *positive selection* and finds a *negative treatment effect* among the *compliers* (i.e., $\hat{\delta}^{IV} < 0$), but *fails* to find that $\hat{\delta}^{OLS} - \hat{\delta}^{IV} > 0$ (i.e., fails to find that $\hat{\delta}^{OLS}$ is less negative than $\hat{\delta}$), then if the researcher wants to "keep" the IV results, he/she must either argue that selection is actually negative, or argue that the treatment effects among the *always-treated* are much more negative than they are among the *compliers*.

On the other hand, if one expects there is *negative selection* and finds a *positive treatment effect* among the *compliers* (i.e., $\hat{\delta}^{IV} > 0$), but *fails* to find that $\hat{\delta}^{OLS} - \hat{\delta}^{IV} < 0$, then if the researcher wants to "keep" the IV results, he/she must either argue that selection is actually positive, or argue that the treatment effects among the *always-treated* are much more positive than they are among the *compliers*.

Finally, if one expects there is *negative selection* and finds a *negative treatment effect* among the *compliers* (i.e., $\hat{\delta}^{IV} < 0$), but *fails* to find that $\hat{\delta}^{OLS} - \hat{\delta}^{IV} < 0$ (i.e., fails to find that $\hat{\delta}^{OLS}$ is more negative than $\hat{\delta}^{IV}$), then if the researcher wants to "keep" the IV results, he/she must either argue that selection is actually positive, or argue that the treatment effects among the *always-treated* are much less negative than they are among the *compliers*.

As stated above though, the heterogenous treatment-effects argument for why the *Coefficient Comparison Check* fails becomes less plausible as the fraction of *compliers* among the treated approaches one. Moreover, even when such a heterogenous treatment effects story is potentially plausible, it still must be acknowledged that it inherently means that $\hat{\delta}^{IV}$ is the estimated LATE for the specific set of *compliers* associated with the particular instrument used in the context at hand, not representative of the average causal effect of the treatment on the population (the ATE), the average effect of the treatment on the treated (the ATT), or even the average treatment effect among a different the set of *compliers* associated with a slightly different instrument or population. Indeed, given this story requires substantial heterogeneity in treatment effects, the estimated $\hat{\delta}^{IV}$ will be fragile in the sense that small changes in who is impacted by the instrument, i.e., small changes in the group of *compliers*, can cause large changes in the estimated $\hat{\delta}^{IV}$, as the underlying δ^{IV} is not a unique parameter even within one particular setting.

This last issue can be seen graphically in Figure 2. The two graphs in Figure 2 both correspond a single setting with positive selection and positive treatment effects, such as in Figure 1a. However, unlike Figure 1a, Figure 2 depicts a situation in which there are heterogenous treatment effects such that the size of these treatment effects go in the opposite direction of the selection. This would be the type of heterogeneous treatment effects story that could cause a violation of the *Coefficient Comparison Check* even if the IV is valid. However, the two graphs in Figure 2 differ in that they characterize two IVs that capture a slightly different group of *compliers* via different values for u_1^* and u_0^*.

The average treatment effect among the compliers in each graph in Figure 2, δ^{IV}, is essentially the height of the black trapezoids weighted by the density of *F(U)* over the range of compliers in each. As can be seen, δ^{IV} will be quite different in these two situations due to heterogeneous treatment effects and the changes in the composition of *compliers* associated with slightly different IVs*,* despite the fact that both depict the exact same underlying treatment effects model. In other words, if one argues that the *Coefficient Comparison Check* fails due to a heterogenous treatment effects that are in opposition to the selection effect, then the estimated parameter is almost certainly quite sensitive to even relatively small changes in the instrument. This means one should be very careful before applying such an estimate to any policy counterfactuals that would alter who gets treated.

This last point relates to the arguments previously made by some other papers, including Mogstad and Torgovitsky (2018) and Heckman and Vytlacil (2005), the latter stating that, given heterogenous treatment effects, "(t)wo economists analyzing the same dataset but using different valid instruments will estimate different parameters that have different interpretations." Brinch, Mogstad, and Wiswall (2017) provide an example of this in the context of estimating the impact of having additional siblings on one's own education, showing different IVs can lead to dramatically different results (indeed the sign on the estimated IV coefficient differs across IVs).[10]

Moreover, given a failure of the *Coefficient Comparison Check* can also arise due to an invalid IV, this suggests that given such a failure, researchers might also consider of employing one of the IV validity checks discussed below in Section 5.

[10] A failure of the *Coefficient Comparison Check* may also suggest researchers investigate the Marginal Treatment Effects (MTEs) associated with their results (see for example Heckman and Vytlacil (2005) and Mogstad, Santos, and Torgovitsky (2017)), though such procedures may not be completely informative in cases like what is discussed here, as MTEs are often not identified for desired policy relevant questions, and therefore may not necessarily speak to the issues I am highlighting here associated with slight changes to the group of *compliers* due to a slightly different instrument or slight change in context.

4 – Applying the *Coefficient Comparison Check* to Real Data

This section considers this *Coefficient Comparison Check* with respect to real world data. The first part examines what happens when this *Coefficient Comparison Check* is applied to IVs known to be invalid. The second part then considers the frequency of failures of the *Coefficient Comparison Check* in the context of a variety of prominently published results using a particular form of IV---namely adjudicator propensity-to-treat IVs. I wrap up this section by discussing failures of this check in the literature more broadly.

The *Coefficient Comparison Check* and "Bad" IVs

Consider trying to estimate the effect of completing college on adult earnings, or estimating the effect of completing high school on the likelihood of arrest in young adulthood. In thinking about such exercises, one would expect those who obtain a college degree or more would, on average, have had higher earnings than those who don't obtain a college degree, even in the absence obtaining the college degree. In other words, OLS estimates of the effect of a college degree on earnings are almost surely *positively* biased. On the other hand, those who complete high school or more almost surely would have a lower expected likelihood of arrest as a young adult than those who drop out of high school, even in the absence of those in the first group completing high school. Therefore, OLS estimates of the effect of a high school degree on early adulthood arrest are almost surely *negatively* biased.

Now, suppose we "attempt" to overcome this OLS bias by instrumenting for an individual's attainment of a college degree or a high school degree using whether or not one of his/her biological parents obtained a college degree as an IV. One could certainly provide a plausible argument that a person's educational attainment is positively related to whether or not one or more of his/her parents obtained a college degree, which is required for the *relevance condition* to hold (and is directly testable). However, I put "attempt" in quotes above, as this IV almost certainly violates the required *exclusion restriction* for a valid instrument in the situations examined here---parents' education almost surely impacts their child's expected income and expected arrest likelihood in young adulthood in ways other than just through how it impacts the child's own educational attainment.

Given this is almost certainly a faulty IV, as proof of concept, we can apply the *Coefficient Comparison Check* and see if the estimation results fail this check. To implement this exercise, I use data from the National Longitudinal Survey of Youth 1997. Table 1 shows the results of this exercise. Column (1) shows the treatment, column (2) shows the "bad" IV that is being employed, and column (3) shows the outcome of interest. Column (4) reports whether or not covariates are included in the regressions.[11] Column (5) then shows the results of the first-stage regression, or the results from regressing the treatment variable on the IV, which is used to assess the "relevance" of the instrument. Column (6) shows the expected bias in the OLS coefficient, columns (7) and (8) report the OLS and IV results, and column (9) reports the results of the *Coefficient Comparison Check*. If this check fails, I report the one-sided p-value associated with the z-statistic related to whether the OLS coefficient and IV coefficient differ in the expected direction (although, as I discuss below, I don't think statistical significance is all that important here).

Looking first at column (5), we see that the IV (parent college) is positively and significantly (p-val < 0.01) related to the treatments of interest (college degree and high school degree respectively). In other words, even though we know this IV is almost surely invalid for these situations, it passes the *relevance condition*.

Given the expected *positive* selection when it comes assessing the effect of completing college on earnings, the *Coefficient Comparison Check* for such specifications assesses whether the OLS coefficient is greater than the IV coefficient. Looking at top row of Table 1, we see that the OLS coefficient in column (7) is actually much smaller than the IV coefficient in column (8), with column (9) showing this difference is negative (along with the p-value on a one-sided z-test of this difference). This constitutes a failure of the *Coefficient Comparison Check*. Looking at the second row, we can see that the difference between the OLS and IV coefficients closes quite a bit when I include covariates, but the *Coefficient Comparison Check* still fails in the sense that the difference between the OLS coefficient and the IV coefficient remains negative. However, we might call this a "marginal"

[11] Covariates include sex, race, whether the mother gave birth to the respondent while she was a teenager, the respondent's AFQT test score, and household income in 1997 (when the respondent was between 12 and 16).

failure in the sense that the (one-sided) difference between OLS and IV is not statistically significant (more on this below).

We can look next at the third and fourth rows of Table 1, where the treatment is a high school degree and the outcome of interest is arrest in early adulthood. In this case, there is expected to be negative selection, so the *Coefficient Comparison Check* assesses whether the OLS coefficient is less (in real terms) than the IV coefficient. Looking at the third row, the OLS coefficient in column (7) is actually greater (in real terms) than the IV coefficient in column (8), with column (9) showing this positive difference and the p-value on a one-sided z-test equaling 0.048. Therefore, this also constitutes a failure of the *Coefficient Comparison Check*. The fourth row shows that when we include covariates, the difference between the OLS and IV coefficients again closes, but the difference between the OLS coefficient and the IV coefficient is still positive, though not "statistically" so. This again could be called a "marginal" failure in this context.

This exercise shows that the "bad" IVs used above will fail this *Coefficient Comparison Check*. However, when covariates are included, these failures are "marginal" in the sense that the differences between OLS and IV are not statistically significantly different from the required direction. However, given the question is whether the IV results provide a more informative estimate than OLS, and OLS should be an upper bound on the size of the treatment effect under relatively homogenous treatment effects given positive selection, and a lower bound given negative selection, and the standard errors will always be smaller for OLS than IV results, I'd argue that failure of the *Coefficient Comparison Check* is something that requires addressing regardless of whether or not it is statistically significant.

In these exercises, failure of the *Coefficient Comparison Check* is almost surely due to a failure of the IV exclusion restriction. However, as discussed above, there are two other reasons why the *Coefficient Comparison Check* can fail. Below, I discuss this in more detail with respect to actual published results with respect to one particular type of IVs.

The *Coefficient Comparison Check* Applied to Adjudicator Propensity-to-Treat IVs

This subsection applies the *Coefficient Comparison Check* to a variety of published IV results where the IVs are all of a particular class. Namely, I focus on applications where treatments are issued by some sort of adjudicator. For example, one particular type of adjudicator is a criminal court judge who must decide whether or not to incarcerate the guilty defendants in his/her court. In such cases, any given judge is arguably more prone to impose incarceration on the defendants whom he or she thinks are more likely to quickly re-offend. So, if a researcher is trying to assess the treatment effect of incarceration on the likelihood to re-offend, most researchers would likely expect that there is positive selection into this treatment in the sense that those treated with incarceration would have a higher expected re-offending rate than the untreated even in the absence of the incarceration treatment. Therefore, if one compares the average re-offending rates among those treated to those untreated (essentially the OLS estimate), one expects this will be an upwardly biased estimate of the impact of incarceration on re-offending, as it conflates the true treatment effect (which could be positive or negative) and the selection effect (which in this case one expects to be positive).

To overcome such a bias in these situations, researchers have often exploited the fact that individuals are often as-good-as randomly allocated to adjudicators within some particular domain (e.g., same courthouse, office, etc.), and different adjudicators have different propensities to assign treatment. Researchers then effectively use the average propensity-to-treat of the adjudicator each subject is assigned to as an IV for treatment. The argument is that if subjects are randomly assigned to adjudicators, and adjudicators only impact each subject's outcome through whether or not they impose the treatment (the exclusion restriction), then an adjudicator's average propensity-to-treat is a valid IV. Because this approach can be applied to situations with different types of adjudicators (e.g., judges, case workers, prosecutors, etc.), I refer to these as adjudicator propensity-to-treat IVs. However, in the literature, they have also been referred to as "judge fixed-effects" or "examiner tendency" IVs (see Chyn, Frandsen, and Leslie (2025) for a thorough discussion of this approach).

Table 2 below shows the results from 39 different estimates from 24 different published studies of the effect of some treatment on some outcome using an adjudicator propensity-to-treat IV. Each of these papers was published in one of the most prestigious peer reviewed journals in economics, sociology, or criminology, and are all excellent (and often innovative) papers. Indeed, the overall strength and influence of these papers is exactly why they make good case studies for the issue I am raising in this paper.

The first column of Table 2 simply lists the citation. Columns (2) and (3) show the treatment of interest and the outcome of interest. Many papers have more than one outcome of interest. When these outcomes are similar in nature, I have only included one to save space. When the outcomes seem to be relatively distinct, I include these different outcomes in separate rows. In cases in which there were multiple estimates with respect to the same outcome coming from specifications with different covariates, I tried to take the one that the authors seem to prefer or the one from the specification with the most covariates. None of this has any notable impact on any of my findings though.

In column (4) of Table 2 I've done my best to capture the expected OLS bias given the context. Columns (5) and (6) show the OLS and IV coefficients on the treatment respectively (along with their standard errors in parentheses). The last column in Table 2 shows the difference between columns (5) and (6) and states whether this *Coefficient Comparison Check* passes or fails, or cannot be determined due to lack of OLS results reported in the paper.

As can be seen, while a many of these results pass the *Coefficient Comparison Check*, there are also a substantial number that either fail or do not allow one to conduct the *Coefficient Comparison Check* since they don't report the OLS results. Indeed, of the 39 separate estimates shown, 17 fail this check, and in 6 cases the check cannot be performed due to OLS results not being reported. In these papers where the results fail this check, only a few explicitly acknowledge this issue (more on this below).

Interestingly, some subsets of papers in Table 2 examine essentially the same research question using essentially the same method, but just differ in contexts, and find opposing results. For example, both the Aizer and Doyle (2015) paper (see rows 10a and 10b) and the Eren and Mocan (2021) paper (see rows 19a and 19b) use juvenile judge propensity-to-incarcerate as an IV to estimate the impact of juvenile incarceration on both high-school graduation and recidivism (in Chicago and Louisiana respectively). In each case, the authors employ the IV approach because they expect a particular form of selection. As stated by Aizer and Doyle, "(t)he main challenge inherent in estimating the causal impact of incarceration is to control or otherwise account for the influence of individual characteristics that may jointly influence incarceration and future human capital accumulation, criminal activity, and labor market outcomes. These characteristics include greater socioeconomic disadvantage, lower levels of cognitive achievement, and less self-control" (Aizer and Doyle 2015: pp 761). This suggests that, in the absence of incarceration, juveniles who end up incarcerated would have had higher expected recidivism rates (positive selection) and lower high-school graduation rates (negative selection) than the juveniles who weren't incarcerated. Notably, Aizer and Doyle's IV results suggest juvenile incarceration increases recidivism and decreases high-school graduation, while Eren and Mocan's IV results suggest juvenile incarceration has little impact on either.

The selection stories discussed above imply that the *Coefficient Comparison Check* consists of checking whether the difference between the OLS coefficient and the IV coefficient is positive when it comes to recidivism and negative when it comes to high-school graduation. As can be seen in Table 2, Aizer and Doyle's results (rows 10a and 10b) fail the *Coefficient Comparison Check*, but Eren and Mocan's (rows 19a and 19b) results do not.

This makes one wonder why Aizer and Doyle's results fail the *Coefficient Comparison Check*. One reason for failure of this check could be that the exclusion restriction is invalid for this IV (option *iii* in the *Coefficient Comparison Check*). For example, in Aizer and Doyle's context of Chicago, more lenient judges may help the juveniles they adjudicate in ways other than just through being less likely to incarcerate, such as through what they say to these youths in court or via other treatments they can apply.

However, if one doesn't believe this invalid IV story, then one must argue either that the selection story is opposite of expected (option *i*) or that there are heterogeneous treatment effects to go in opposition to the selection (option *ii*). Given it seems unlikely that the selection story goes in the opposite direction than expected here, this leaves option *ii*, or that the treatment effects for the *always-incarcerated* are sufficiently smaller than the treatment effects among the *compliers* to overcome the selection effects. While this is arguably possible, it would mean that in Aizer and Doyle's context, the impact of incarceration on recidivism among the *always-treated* must be extremely small or even negative for this to account for the failure in the *Coefficient Comparison Check*.

Indeed, the type of heterogeneity that could account for why Aizer and Doyle's recidivism results fail the *Coefficient Comparison Check* is what was depicted in Figure 2 above. As that figure showed, IV results under such a scenario are likely quite fragile, as small changes in the IV, such as what would occur with the hiring or retiring of a couple of particularly lenient or harsh judges, or some judges changing their behavior over time, would shift the "trapezoid of *compliers"* in Figure 2, thereby notably altering the IV results.

Given that Eren and Mocan's (2021) results do not fail the *Coefficient Comparison Check*, this suggests that Eren and Mocan's results are arguably the more robust of the two. However, the fact that the results in one of these studies fails the *Coefficient Comparison Check* means the conflicting findings might just reflect that there are very large heterogeneities in the effects of juvenile incarceration on future outcomes, and the two studies simply have slightly different groups of *compliers*. Either way though, this suggests that Aizer and Doyle's results unlikely provide a general result about the effects of juvenile incarceration on future offending or high school graduation. This same dynamic is at play in any pairs of studies where the research question and IV method are similar, but results are conflicting and one fails the *Coefficient Comparison Check* while the other does not.

The Agan, Doleac, and Harvey (2023) paper presents an interesting case to consider for other reasons. Their paper looks at the impact of avoiding a misdemeanor conviction (conditional on a misdemeanor arrest) on future arrests. In this context, one would arguably expect prosecutors to be more likely to drop the misdemeanor charge for those arrestees that they think will be less prone to commit future crimes. Hence, one would suspect there is negative selection in this context. Given this, and the fact that their IV results are negative in sign, the *Coefficient Comparison Check* consists of examining whether the OLS results are less than (i.e., more negative) than the IV results. Looking at Agan, Doleac, and Harvey's (2023) results in Table 2 (row 23), we see that the OLS results are higher (i.e., less negative), than the IV results, meaning their results fail this check.

In their discussion of these results, Agan, Doleac, and Harvey (2023) essentially acknowledge this issue, pointing out that for their IV results to be valid, the Assistant District Attorneys (ADAs) who are in charge of prosecution, must be "on average, choosing not to prosecute defendants who have a higher risk of subsequent criminal justice contact than marginal defendants. This may at first glance be counterintuitive." They go to say that, while counterintuitive, such positive selection may occur because "there are a variety of characteristics that ADAs might interpret as mitigating circumstances, making defendants less culpable for their crimes or more worthy of a second chance, but that also increase the risk of subsequent criminal justice contact, for example mental health issues, drug addiction, or age." (Agan, Doleac, and Harvey, 2023: pp. 1479) In other words, the authors are implicitly suggesting that the reason their results fail the *Coefficient Comparison Check* is because the "expected" selection story is wrong---there is actually *positive* selection rather than *negative* selection in their context. They then further apply the Frandsen et al. (2023) test to explicitly consider potential failures in the required IV restrictions (more on this test below) and estimate marginal treatment effects (MTEs) among the compliers to consider the scope of heterogenous treatment effects.

A related example arises in Dahl, Kostol, and Mogstad (2014), who also implicitly acknowledge that their results with respect to the impact of parent receipt of disability on child's receipt of disability (row 8 in Table 2) go in the opposite way of expected selection. However, they thoroughly discuss and assess the extent to which heterogenous treatment effects could be behind this puzzle, as well as emphasize the "local" and therefore not necessarily generalizable nature of their results.

In the end, these types of discussions by Agan, Doleac, and Harvey (2023) and Dahl, Kostol, and Mogstad (2014) align with what I am arguing should be "best practices" when IV results fail the *Coefficient Comparison Check*.

Failures of the *Coefficient Comparison Check* More Broadly

Thus far I've focused on papers using the adjudicator propensity-to-treat IV methodology, but failures of the *Coefficient Comparison Check* aren't just an issue with this category of IV papers. Indeed, the elephant in the room is that this issue also arises with respect to some of the most classic papers in the IV literature, namely those examining the impact of additional schooling on earnings. Card (2001) discusses several of these papers including one of his own (i.e, Card 1995), and actually includes a table analogous to Table 2 in this paper, but for IV papers estimating the impact of additional schooling on earnings (see Table II, Card 2001: pp 1146). What is notable is that Card shows that almost all of these IV estimates implicitly violate the *Coefficient Comparison Check*, several notably so.

Obviously, Card does not describe what he is talking about as failures of the *Coefficient Comparison Check*, but he does acknowledge the issue at hand, referring to is as a "puzzle." He goes on to highlight several possible explanations for this puzzle, including measurement error in education (though he deems this unlikely), "specification searching" wherein researchers tend to favor reporting specifications that have higher t-statistics (Ashenfelter, Harmon, and Oosterbeek 1999), as well as heterogenous treatment effects in which those with higher costs to obtaining more education also have much higher returns. More recently, in his in his terrific book *Causal Inference-The Mixtape*, Scott Cunningham comments on this exact issue with respect to Card (1995) by saying, "(a)ll we are left saying is that for some reason, the higher marginal cost of attending college is causing these people to underinvest in schooling; that in fact their returns are much higher." (Cunningham 2021: pp 356)

5 - Other Tests of IVs

This is certainly not the first paper to propose a check on IV results. The tests I discuss below provide other methods for highlighting potentially problematic IVs, though all are more complicated to implement than the *Coefficient Comparison Check* introduced here. While these tests can sometimes detect when a proposed IV violates the necessary assumptions for a valid IV, they cannot guarantee that all these assumptions hold for a proposed IV. However, they do provide another way of assessing the credibility of IV results, which might be especially useful for results that fail the *Coefficient Comparison Check*.

Kitagawa (2015) builds on the work of Balke and Pearl (1997) and Heckman and Vytlacil (2005), by proposing a test for a binary proposed instrument Z based on the argument that after dividing the sample into subsamples of those treated and those not treated, the distribution of observed outcomes for those with $Z = 0$ must be embedded in the distribution of the outcomes for those with $Z = 1$ in the treated subsample, while the distribution of observed outcomes for those with $Z = 1$ must be embedded in the distribution of the outcomes for those with $Z = 0$ in the untreated subsample. Intuitively, if there is significant selection and the exclusion restriction holds, then the group that includes the compliers in each subsample should have more variation in realized outcomes.

To implement this test, one must estimate differences in distributions, and use a particular bootstrapping procedure to obtain p-values for hypothesis testing. Interestingly, Kitagawa's test rejects the proximity to college IV for college attendance as proposed by Card (1995), which was one of the cases discussed above that also fails the *Coefficient Comparison Check*. However, when Kitagawa includes a variety of other conditioning variables, Card's (1995) proposed IV no longer fails his test. Mourifie and Wan (2017) build on Kitagawa's (2015) approach, but instead of testing for differences in distributions, their test looks at a series of conditional moment inequalities that can be evaluated using the intersection bounds framework of Chernozhukov, Lee, and Rosen (2013).

Huber and Mellace (2015) offer an alternative test for whether the validity of the necessary IV conditions hold. Namely, they argue that if the IV conditions hold for some binary instrument Z, then one could use the average outcome of the treated individuals with $Z = 0$ (who consist of the always-treated) to derive bounds on the expected outcome among those treated with $Z = 1$ (which includes always-treated and compliers). An analogous version of this test can be derived using the average outcome of the untreated individuals with $Z = 1$ (the never-treated) to derive bounds on the expected outcome for untreated individuals with $Z = 0$ (which includes never-treated and compliers). Like Kitagawa (2015), when Huber and Mellace (2015) apply their test to Card's (1995) study, they reject the validity of the IV when no covariates are included, but do not reject its validity when covariates are included.

Fransden et al. (2023) develop a test for evaluating whether the monotonicity and exclusion restrictions hold for the adjudicator propensity-to-treat (i.e., "judge fixed-effects") IVs discussed above. This test exploits the fact that the conditional expectation of outcomes across adjudicators has a bounded slope between any two adjudicators. Finding that the slope of this estimated conditional expectation function lies outside of these bounds suggests the necessary IV restrictions do not hold. However, like the tests above, a non-failure of this test does not necessarily mean the IV restrictions hold with certainty.

Finally, Black et al. (2022) develop a test to highlight whether there is significant enough selection to warrant an IV, rather than the validity of the IV itself. Specifically, they propose a relatively simple set of tests based on dividing observations into the treated and untreated subsamples to determine whether the Conditional Mean Independence

Assumption (CMIA) holds. If one cannot reject that the CMIA holds, non-IV methods can be used to identify treatment effects.

6 - Conclusion

This paper proposes that a standard practice when estimating IV regressions should be to conduct what I refer to as the *Coefficient Comparison Check*. Namely, if the expected selection effect that motivates the IV is positive, then this check amounts to assessing whether the OLS coefficient is greater than (in real terms) the IV coefficient. By contrast, when the expected selection effect that motivates the IV is negative, then this check amounts to assessing whether the OLS coefficient is less than (in real terms) than the IV coefficient.

Failures of this check do not necessarily imply that the IV is invalid, but I argue that such failures do require attention. Given a failure of the *Coefficient Comparison Check*, a researcher who still believes his/her IV is valid must convince the reader either that selection is actually the opposite of expected, or that there are very large differences in treatment effects between the *compliers* and the *always-treated,* and these differences go in the exact opposite direction of the selection effect. This heterogenous treatment effects explanation is less plausible the higher the fraction of the treated who are *compliers*. Even in cases where such heterogenous treatment effects are plausible, the researcher should acknowledge that the failure of the *Coefficient Comparison Check* inherently means that the estimated parameter is not very general, as it is likely quite sensitive to small changes in who is impacted by the instrument.

The issue at the heart of this paper is by no means unknown. Indeed, Card (2001) considered this type of issue over 20 years ago. My point here is that, at this point in the literature, researchers using IV methods should take this issue seriously and acknowledge it when it arises. My hope is that by working out this issue in full detail, and giving a name to the type of checks and responses that should be included with IV results, researchers using IV methods will apply this check, and when it fails, discuss what explanation most likely accounts for this failure and what the exact implications of this explanation are.

7 – Appendix

Deriving What a Valid IV Estimator Identifies with Potentially Heterogeneous Treatment Effects

Consider the IV estimator defined as $\delta^{IV} \equiv \frac{\delta^{RF}}{\pi_c}$, where π_c is the fraction of the population who are compliers or $F(u_0^*) - F(u_1^*)$, and $\delta^{RF} = E[Y|Z = z_1] - E[Y|Z = z_0]$ is the "reduced-form" estimate, or simply the difference in expected outcomes between those with Z = z_1 and those with Z = z_0.

Recalling that those with *Z = z_1* will be treated if $u_1^* < u$ and those with *Z = z_0* will be treated if $u_0^* < u$, we can write out $E[Y|Z = z_1]$ and $E[Y|Z = z_0]$ in equation form as follows:

$$E[Y|Z = z_1] = F(u_1^*)E[Y(0)|Z = z_1, u \leq u_1^*] + [1 - F(u_1^*)]E[Y(1)|Z = z_1, u_1^* < u]$$

and

$$E[Y|Z = z_0] = F(u_0^*)E[Y(0)|Z = z_0, u \leq u_0^*] + [1 - F(u_0^*)]E[Y(1)|Z = z_0, u_0^* < u].$$

Recalling that $u_1^* < u_0^*$, the above equations in turn can be re-written as

$$E[Y|Z = z_1] = \pi_N E[Y(0)|Z = z_1, u \leq u_1^*] +$$
$$[1 - F(u_1^*)]\left(\frac{\pi_C}{1 - F(u_1^*)}E[Y(1)|Z = z_1, u_1^* < u \leq u_0^*] + \frac{\pi_A}{1 - F(u_1^*)}E[Y(1)|Z = z_1, u_0^* < u]\right)$$

and

$$E[Y|Z = z_0] = F(u_0^*)\left(\frac{\pi_N}{F(u_0^*)}E[Y(0)|Z = z_0, u \leq u_1^*] + \frac{\pi_C}{F(u_0^*)}E[Y(0)|Z = z_0, u_1^* < u \leq u_0^*]\right)$$
$$+ \pi_A E[Y(1)|Z = z_0, u_0^* < u].$$

Where π_N is the fraction of the population that are never-treated or $F(u_1^*)$, π_C is the fraction of the population who are compliers or $F(u_0^*) - F(u_1^*)$, and π_A is the fraction of the population who are always-treated or $1 - F(u_0^*)$.

Subtracting $E[Y|Z = z_0]$ from $E[Y|Z = z_1]$ we get

$$\delta^{RF} = \pi_N(E[Y(0)|Z = z_1, u \leq u_1^*] - E[Y(0)|Z = z_0, u \leq u_1^*])$$

$$+\pi_c(E[Y(1)|Z = z_1, u_1^* < u \leq u_0^*] - E[Y(0)|Z = z_0, u_1^* < u \leq u_0^*])$$
$$+\pi_A(E[Y(1)|Z = z_1, u_0^* < u] - E[Y(1)|Z = z_0, u_0^* < u])$$

Now, noting from equations (1a) and (1b) that $E[Y(0)|Z = z_1, u] = E[Y(0)|Z = z_0, u] = E[Y(0)|u]$ and $E[Y(1)|Z = z_1, u] = E[Y(1)|Z = z_0, u] = E[Y(1)|u]$ for all u, we know the top and bottom lines in the above equation net out to zero, leaving

$$\delta^{RF} = \pi_c(E[Y(1)|\, u_1^* < u \leq u_0^*] - E[Y(0)|u_1^* < u \leq u_0^*]).$$

Given the outcome generation process in equation (1), the above equation can be re-written

$$\delta^{RF} = \pi_c\left(\frac{\int_{u_1^*}^{u_0^*}(\delta_i + \beta u)f(u)du}{\pi_C} - \frac{\int_{u_1^*}^{u_0^*}(\beta u)f(u)du}{\pi_C}\right),$$

where *f(u)* is the pdf corresponding to the cdf *F(u)*, which can be simplified to

$$\delta^{RF} = \int_{u_1^*}^{u_0^*}\delta_i\, f(u)du.$$

This means that since $\delta^{IV} \equiv \frac{\delta^{RF}}{\pi_c}$, the above equation implies

$$\delta^{IV} = \frac{\int_{u_1^*}^{u_0^*}\delta_i\, f(u)du}{\pi_c} = E[\delta_i|u_1^* < u_i \leq u_0^*],$$

which is what was stated in the text.

Deriving what OLS Estimator Identifies with Potentially Heterogeneous Treatment Effects

Asymptotically, the basic OLS estimate of the treatment effect, just captures the difference between the expected outcomes among the treated versus the expected outcomes among the untreated, or $\delta^{OLS} = E[Y|D = 1] - E[Y|D = 0]$. As shown in the text, this can be captured by the following equation:

$$\delta^{OLS} = \frac{\pi_z\pi_C}{\pi_z\pi_C + \pi_A}E[Y(1)|u_1^* < u \leq u_0^*] + \frac{\pi_A}{\pi_z\pi_C + \pi_A}E[Y(1)|u_0^* < u]$$
$$- \frac{(1-\pi_z)\pi_C}{\pi_N + (1-\pi_z)\pi_C}E[Y(0)|u_1^* < u \leq u_0^*] - \frac{\pi_N}{\pi_N + (1-\pi_z)\pi_C}E[Y(0)|u \leq u_1^*].$$

where π_z is the fraction of the population with *Zi = z1*, π_C is the fraction of the population who are compliers or $F(u_0^*) - F(u_1^*)$, π_A is the fraction of the population who are always-treated or $1 - F(u_0^*)$, and π_N is the fraction of the population that are never-treated or $F(u_1^*)$.

Given equations (1a) and (1b) characterizing $Y(1)$ and $Y(0)$, and noting that *u* is distributed over $(\underline{u}, \overline{u})$ according the cdf *F(u)*, where *f(u)* is the corresponding pdf, we can re-write the above equation as

$$\delta^{OLS} = \frac{\pi_z\pi_C}{\pi_z\pi_C + \pi_A}\frac{\int_{u_1^*}^{u_0^*}(\delta_i + \beta u)f(u)ds}{\pi_C} + \frac{\pi_A}{\pi_z\pi_C + \pi_A}\frac{\int_{u_0^*}^{\overline{u}}(\delta_i + \beta u)f(u)du}{\pi_A}$$
$$- \frac{(1-\pi_z)\pi_C}{\pi_N + (1-\pi_z)\pi_C}E[Y(0)|u_1^* < u \leq u_0^*] - \frac{\pi_N}{\pi_N + (1-\pi_z)\pi_C}E[Y(0)|u < u_1^*].$$

This equation can in turn be re-written to be

$$\delta^{OLS} = \frac{\pi_z\pi_C}{\pi_z\pi_C + \pi_A}\left[\frac{\int_{u_1^*}^{u_0^*}\delta_i\, f(u)du}{\pi_C} + \frac{\int_{u_1^*}^{u_0^*}\beta u\, f(u)du}{\pi_C}\right]$$
$$+ \frac{\pi_A}{\pi_z\pi_C + \pi_A}\left[\frac{\int_{u_0^*}^{\overline{u}}\delta_i\, f(u)du}{\pi_A} + \frac{\int_{u_0^*}^{\overline{u}}\beta u\, f(u)du}{\pi_A}\right]$$
$$- \frac{(1-\pi_z)\pi_C}{\pi_N + (1-\pi_z)\pi_C}E[Y(0)|u_1^* < u \leq u_0^*] - \frac{\pi_N}{\pi_N + (1-\pi_z)\pi_C}E[Y(0)|u \leq u_1^*].$$

From above we know $\delta^{IV} = \frac{\int_{u_1^*}^{u_0^*}\delta_i f(u)du}{\pi_C}$, and defining $\delta^A \equiv \frac{\int_{u_0^*}^{\overline{u}}\delta_i f(u)du}{\pi_A}$ (i.e., the average treatment effect among the *always-treated*), and doing some re-arranging we get

$$\delta^{OLS} = \frac{\pi_z\pi_C}{\pi_z\pi_C + \pi_A}\delta^{IV} + \frac{\pi_A}{\pi_z\pi_C + \pi_A}\delta^A$$
$$+ \frac{\pi_z\pi_C}{\pi_z\pi_C + \pi_A}\frac{\int_{u_1^*}^{u_0^*}(\beta u)f(u)du}{\pi_C} + \frac{\pi_A}{\pi_z\pi_C + \pi_A}\frac{\int_{u_0^*}^{\overline{u}}(\beta u)f(s)du}{\pi_A}$$
$$- \frac{(1-\pi_z)\pi_C}{\pi_N + (1-\pi_z)\pi_C}E[Y(0)|u_1^* < u \leq u_0^*] - \frac{\pi_N}{\pi_N + (1-\pi_z)\pi_C}E[Y(0)|u \leq u_1^*].$$

Now, let us denote the fraction of the treated who are compliers as $\pi^C_{D=1} \equiv \frac{\pi_z \pi_C}{\pi_z \pi_C + \pi_A}$.

Furthermore, note that $\frac{\pi_A}{\pi_z \pi_C + \pi_A} = 1 - \frac{\pi_z \pi_C}{\pi_z \pi_C + \pi_A}$, and that $\frac{\int_{u_1^*}^{u_0^*} (\beta u) f(u) du}{\pi_C} =$ $E[Y(0)|u_1^* < u \leq u_0^*]$ and $\frac{\int_{u_0^*}^{\bar{u}} (\beta u) f(u) du}{\pi_A} = E[Y(0)|u_0^* < u]$, the above equation becomes

$$
\begin{aligned}
\delta^{OLS} = {} & \pi^C_{D=1} \delta^{IV} + (1 - \pi^C_{D=1}) \delta^A \\
& + \left[\frac{\pi_z \pi_C}{\pi_z \pi_C + \pi_A} E[Y(0)|u_1^* < u \leq u_0^*] + \frac{\pi_A}{\pi_z \pi_C + \pi_A} E[Y(0)|u_0^* < u] \right] \\
& - \left[\frac{(1 - \pi_z) \pi_C}{\pi_N + (1 - \pi_z) \pi_C} E[Y(0)|u_1^* < u \leq u_0^*] + \frac{\pi_N}{\pi_N + (1 - \pi_z) \pi_C} E[Y(0)|u \leq u_1^*] \right].
\end{aligned}
$$

Now, recognize that the second line is simply the expected outcome of those who are treated in the absence of treatment, while the third line is simply the expected outcome of those who are untreated in the absence of treatment. This means

$$
\begin{aligned}
\delta^{OLS} = {} & \pi^C_{D1} \delta^{IV} + (1 - \pi^C_{D1}) \delta^A \\
& + E[Y(0)|D = 1] - E[Y(0)|D = 0].
\end{aligned}
$$

This is what was shown above in the text, and reveals that δ^{OLS} is a weighted average of the average treatment effects among the compliers δ^{IV} and the average treatment effects among the always-treated δ^A, plus the selection-effect, which is the difference in expected outcomes between those treated and those untreated in the absence of anyone being treated.

8– References

Agan, Amanda, Jennifer Doleac, and Anna Harvey. (2023). “Misdemeanor Prosecution.” Quarterly Journal of Economics 138: 1453-1505.

Aizer, Anna and Joseph Doyle. (2015). “Juvenile Incarceration, Human Capital, and Future Crime.” Quarterly Journal of Economics 130: 759-804.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. (1996). “Identification of Causal Effects Using Instrumental Variables.” Journal of the American Statistical Association 91: 444-472

Angrist, Joshua and Michael Kolesar. (2024). “One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV.” Journal of Econometrics 240(2): 105398.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. (1999). “A Review of Estimates of the Schooling/Earnings Relationship, with Tests of Publication Bias.” Labour Economics 6: 453-470.

Autor, David, Andreas Kostol, Magne Mogstad, and Bradley Setzler. (2019). “Disability Benefits, Consumption Insurance, and Household Labor Supply.” American Economic Review 109: 2613-2654.

Bald, Anthony, Eric Chyn, Justine Hastings, and Margarita Machelett. (2022). “The Causal Impact of Removing Children from Abusive and Neglectful Homes.” Journal of Political Economy 130: 1919-1962.

Balke, A., and J. Pearl. (1997). “Bounds on Treatment Effects from Studies with Imperfect Compliance.” Journal of the American Statistical Association 92: 1171-1176.

Bhuller, Maudeep, Gordon Dahl, Katrine Loken, and Magne Mogstad. (2020). “Incarceration, Recidivism, and Employment.” Journal of Political Economy 128: 1269-1324.

Black, Dan, Joonhwi Joo, Robert LaLonde, Jeffrey Smith, and Evan Taylor (2022). “Simple Tests for Selection: Learning More from Instrumental Variables.” Labour Economics 79: 1-14.

Brinch, Christian N., Magne Mostad, and Matthew Wiswall. (2017). “Beyond LATE with a Discrete Instrument.” Journal of Political Economy 125: 985-1039.

Card, David. (1995). “Using Geographic Variation in College Proximity to Estimate the Return to Schooling.” In Aspects of Labour Market Behavior: Essays in Honour of John Vandercamp, ed. By Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky. University of Toronto Press: Toronto ON.

Card, David. (2001). “Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems.” Econometrica 69: 1127-1160.

Chernozhukov, V., S. Lee, and A.M. Rosen. (2013). “Intersection Bounds: Estimation and Inference.” Econometrica 81: 667-737.

Chyn, Eric, Brigham Frandsen, and Emily Leslie. (2025). “Examiner and Judge Designs in Economics: A Practitioner’s Guide.” Journal of Economic Literature 63: 401-39.

Collinson, Robert, John Eric Humphries, Nicholas Mader, Davin Reed, Daniel Tannenbaum, Winnie Van Dijk. (2024). “Eviction and Poverty in American Cities: Evidence from Chicago and New York.” Quarterly Journal of Economics 139: 57-120.

Cunningham, Scott. (2021). Causal Inference-The Mixtape. Yale University Press: New Haven, NJ.

Dahl, Gordon, Andreas Ravndal KostØl, and Magne Mogstad. (2014). “Family Welfare Cultures.” Quarterly Journal of Economics 129: 1711-1752.

Di Tella, Rafael and Ernesto Schargodsky. (2013). “Criminal Recidivism after Prison and Electronic Monitoring.” Journal of Political Economy 121: 28-73.

Dobbie, Will, Paul Goldsmith-Pinkham, and Crystal S. Yang. (2017). “Consumer Bankruptcy and Financial Health.” Review of Economics and Statistics 99: 853-869.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. (2018). “The Effects of Pretrial Detention on Conviction, Future, Crime, and Employment: Evidence from Randomly Assigned Judges.” American Economic Review 108: 201-240.

Doyle, Joseph J. (2007). “Child Protection and Child Outcomes: Measuring the Effects of Foster Care.” American Economic Review 97: 1583-1610.

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

Eren, Ozkan and Naci Mocan. (2021). “Juvenile Punishment, High School Graduation, and Adult Crime: Evidence from Idiosyncratic Judge Harshness.” Review of Economics and Statistics 10: 34-47.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. (2023). “Judging Judge Fixed Effects.” American Economic Review 113: 253-277.

French, Eric and Jae Song. 2014. “The Effect of Disability Receipt on Labor Supply.” American Economic Journal: Economic Policy 6: 291-337.

Green, Donald P. and Daniel Winik. (2010). “Using Random Judge Assignments to Estimate the Effects of Incarceration and Probation on Recidivism Among Drug Offenders.” Criminology 48: 357-387.

Gross, Max and E. Jason Baron. (2022). “Temporary Stays and Persistent Gains: The Causal Effects of Foster Care.” American Economic Journal: Applied Economics 14: 170-199.

Harding, David, Shawn Bushway, Jeffrey D. Morenoff, and Anh P. Nguyen. (2018). "Imprisonment and Labor Market Outcomes: Evidence from a Natural Experiment." American Journal of Sociology 124: 49-110.

Heckman, James J. and Edward Vytlacil. (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." Econometrica 73: 669-738.

Huber, Martin and Giovanni Mellace. (2015). "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints." Review of Economics and Statistics 97: 398-411.

Ishimaru, Shoya. (2022). "Empirical Decomposition of the IV-OLS Gap With Heterogenous and Nonlinear Effects." Review of Economics and Statistics 106: 505-520.

Kitagawa, Toru. (2015). "A Test For Instrument Validity." Econometrica 83: 2043-2063.

Kling, Jeffrey R. (2006). "Incarceration Length, Employment, and Earnings." American Economic Review 96: 863-876.

Leslie, Emily and Nolan Pope. (2017). "The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments." Journal of Law and Economics 60: 529-557.

Lochner, Lance and Enrico Moretti. (2015). "Estimating and Testing Models with Many Instrument Levels and Limited Instruments." Review of Economics and Statistics 97: 387-397.

Loeffler, Charles E. (2013). "Does Imprisonment Alter the Life Course? Evidence on Crime and Employment from a Natural Experiment." Criminology 51: 137-166.

Mogstad, Magne, Andreas Santos, and Alexander Torgovitsky. (2018). "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters." Econometrica 86: 1589-1619.

Mogstad, Magne and Alexander Torgovitsky. (2018). "Identification and Extrapolation of Causal Effects with Instrumental Variables." Annual Review of Economics 10: 577-613.

Mourifie, Ismael and Yuanyuan Wan. (2017). "Testing Local Average Treatment Effect Assumptions." Review of Economics and Statistics 99: 305-313.

Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. (2013). "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." American Economic Review 103: 1797-1829.

Norris, Samuel. Matthew Pecenco, and Jeffrey Weaver. (2021). "The Effects of Parental and Sibling Incarceration: Evidence from Ohio." American Economic Review 111: 2926-2963.

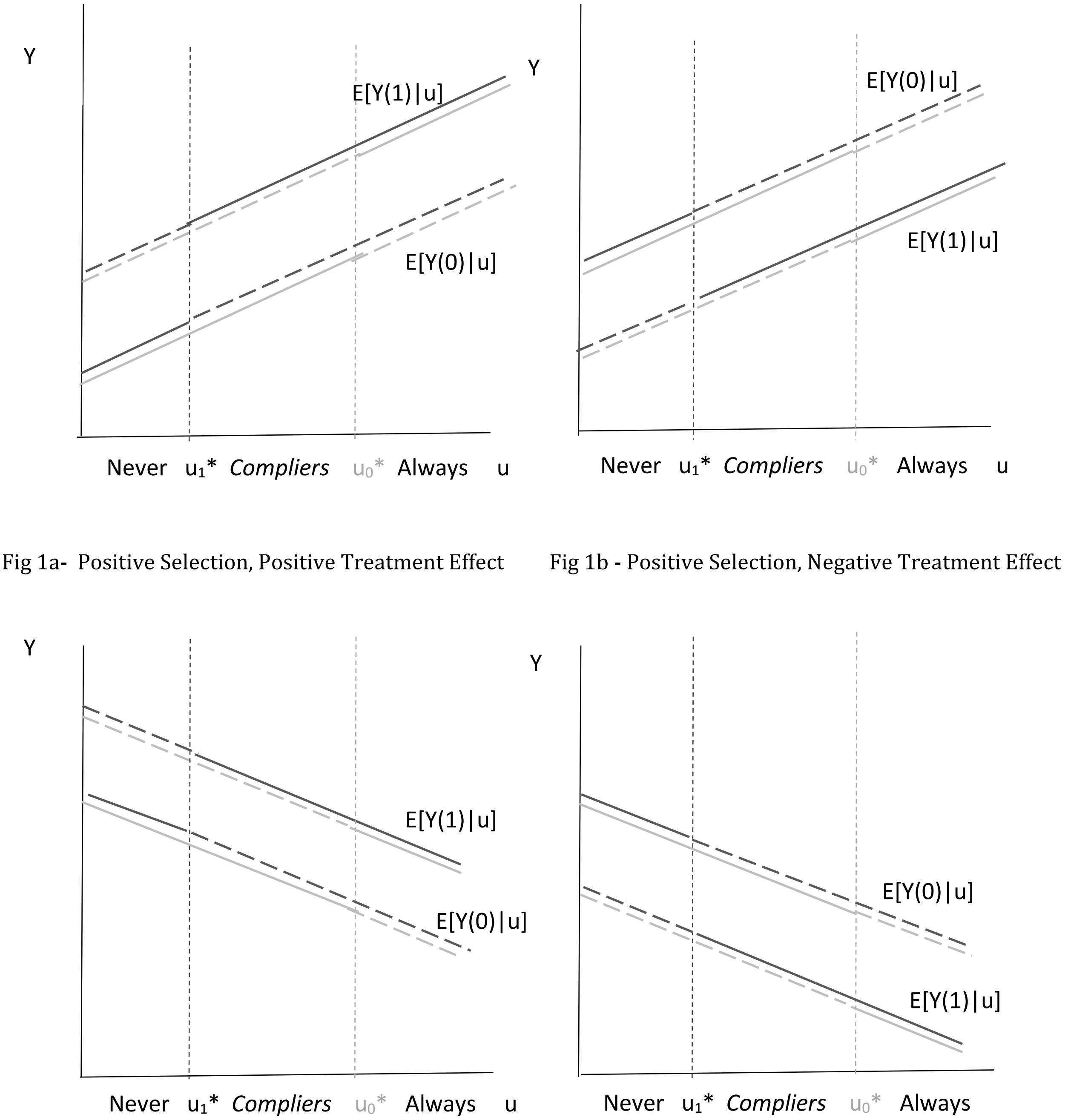

Fig 1a- Positive Selection, Positive Treatment Effect

Fig 1b - Positive Selection, Negative Treatment Effect

Fig 1c - Negative Selection, Positive Treatment Effect

Fig 1d - Negative Selection, Negative Treatment Effect

Figure 1 – Graphical Depictions of Selection and (homogeneous) Treatment Effects. Lines depict expected outcome given treatment and no treatment, conditional on u. The dark lines indicate these for those with $Z_i = z_1$, while lighter lines indicate these for those with $Z_i = z_0$. Solid portions indicate observed realizations, dashed portions indicate unobserved counterfactual realizations.

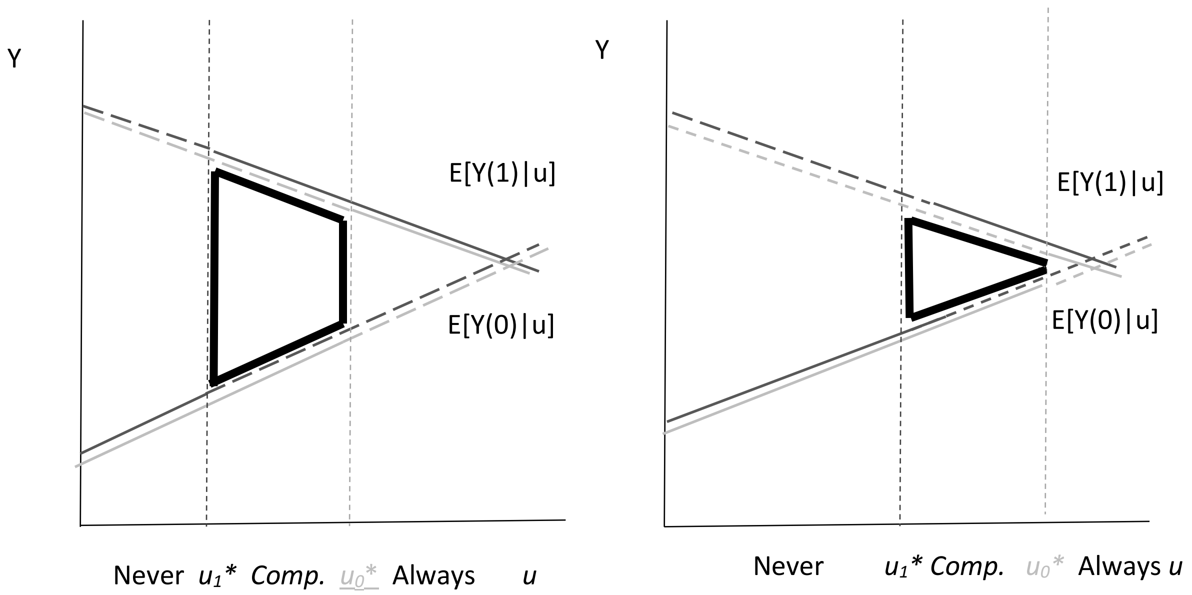
Y
E[Y(1)|u]
E[Y(0)|u]
Never u_1* Comp. u_0* Always u
Y
E[Y(1)|u]
E[Y(0)|u]
Never u_1* Comp. u_0* Always u

Figure 2 – Graphs above depict a situation with Positive Selection along with Positive but heterogenous Treatment Effects that are in opposition to selection. This is a situation that could cause a failure of *Coefficient Comparison Check*. Lines depict expected outcome given treatment and no treatment, conditional on *u*. Dark lines indicate those with $Z_i = z_1$, while lighter lines indicate those with $Z_i = z_0$. Solid portions indicate observed realizations, dashed portions indicate unobserved counterfactual realizations. Graphs reveal that δ^{IV} (the average treatment effect among compliers), which is weighted average of the height of the black trapezoids, will change substantially with shifts in group of *compliers* arising from alterations to instrument.

Table 1 - Applying IV Selection Test to "Bad" IV Estimates

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					Expected			**Coefficient**
Treatment	**"Bad" IV**	**Outcome**	**Covariates?**	**First-Stage**	**OLS Bias**	**OLS Coeff**	**IV Coeff**	**Comp. Check**
Coll. Grad.	Parent	Earnings	No	0.33	Positive	31,845	59,934	-28,089
	College	in 2019		(0.02)		(1774)	(5476)	Fail (p < 0.01)
Coll. Grad.	Parent	Earnings	Yes	0.16	Positive	23,012	35,857	-12,845
	College	in 2019		(0.02)		(1958)	(11825)	Fail (p=0.14)
H.S. Grad.	Parent	Arrested	No	0.14	Negative	-0.30	-0.47	0.17
	College	Ages 22-27		(0.01)		(0.02)	(0.10)	Fail (p=0.048)
H.S. Grad.	Parent	Arrested	Yes	0.05	Negative	-0.25	-0.39	0.14
	College	Ages 22-27		(0.01)		(0.02)	(0.29)	Fail (p=0.32)

Positive OLS bias means $\beta^{OLS} > \beta^{IV}$. Negative OLS bias means $\beta^{OLS} < \beta^{IV}$. Standard errors shown in parentheses. Covariates include sex, race, mother gave birth to respondent while a teen, AFQT test score, household income when respondent was a youth (1997). P in Column (9) refers to p-value on one-sided z-test regarding differences in OLS and IV coefficients.

Table 2 - Applying *Coefficient Comparison Check* to Adjudicator Propensity-to-Treat IV Estimates

(1)	(2)	(3)	(4)	(5)	(6)	(7)
			"Expected"			Coefficient
Paper	Treatment	Outcome	OLS Bias	OLS Coeff	IV Coeff	Comp. Check
1 - Kling (2006)	Incarceration	Earnings	Negative	-44	248	-292
				(32)	(294)	Pass
2a - Doyle (2007)*	Foster Care	Juvenile	Positive	0.00	0.35	-0.35
		Delinquency		(0.01)	(0.14)	Fail (p=0.01)
2b - Doyle (2007)*	Foster Care	Teen	Positive	0.09	0.29	-0.20
		Motherhood		(0.01)	(0.17)	Fail (p=0.12)
2c - Doyle (2007)*	Foster Care	Earnings	Negative	-50	-1296	1246
				(30.6)	(626.0)	Fail (p=0.00)
3a - Doyle (2008)	Foster Care	Adult Arrest	Positive	0.060	0.391	-0.331
				(0.008)	(0.182)	Fail (p=0.03)
3b - Doyle (2008)	Foster Care	Adult	Positive	0.031	0.225	-0.194
		Incarceration		(0.005)	(0.102)	Fail (p=0.029)
4 - Green and Winik (2010)	Incarceration	Recidivism	Positive	-0.006	0.009	-0.015
	for drugs			(0.001)	(0.008)	Fail (p =0.03)
5 - Di Tella, Schargrodsky	Elec. Monitor	Recidivism	Negative	-0.09	-0.13	0.04
(2013)	(vs. Prison)			t-stat rep	t-stat rep	Fail (can't det. p)
6a - Loeffler (2013)**	Incarceration	Recidivism	Positive	0.031	0.075	-0.044
				(0.008)	(0.104)	Fail (p=0.33)
6b - Loeffler (2013)**	Incarceration	Employment	Negative	-0.360	0.031	-0.391
				(0.006)	(0.083)	Pass
7 - Maestas, Mullen,	Disability	Earnings	Negative	-7715	-3007	-4708
Strand (2013)	Insurance			t stat rep	t stat rep	Pass
8 - Dahl, Kostol,	Disability	Child on	Positive	0.01	0.06	-0.050
Mogstad (2014)	Insurance	Disability Ins.		(0.010)	(0.023)	Fail (p=0.02)
9 - French, Song (2014)	Disability	Labor Supply	Negative	-0.265	-0.256	-0.009
	Insurance			(0.002)	(0.006)	Pass

Table 2 - Applying *Coefficient Comparison Check* to Adjudicator Propensity-to-Treat IV Estimates (cont.)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
			"Expected"			Coefficient
Paper	Treatment	Outcome	OLS Bias	OLS Coeff	IV Coeff	Comp. Check
10a - Aizer, Doyle (2015)	Juvenile Incarceration	Incarceration	Positive	0.155 (0.007)	0.234 (0.076)	-0.079 Fail (p = 0.15)
10b - Aizer, Doyle (2015)	Juvenile Incarceration	High School Grad	Negative	-0.073 (0.004)	-0.125 (0.043)	0.052 Fail (p=0.11)
11 - Dobbie, Song (2015)	Chapter 13 Bankruptcy	Home Foreclosure	Positive	Not shown	-0.19 (0.034)	Can't det.
12 - Gupta, Hansman, Frenchman (2016)	Bail (vs Release)	Recidivism	Positive	Not Shown	0.07 (0.008)	Can't det.
13 - Leslie and Pope (2017)	Pre-trial Detention	Felony Recidivism	Positive	-0.100 (0.003)	-0.188 (0.020)	0.088 Pass
14 - Dobbie et al. (2017)	Chapter 13 Bankruptcy	Financial Strain Index	Positive	Not shown	-0.323 (0.071)	Can't det.
15a - Dobbie, Goldin, Yang (2018)	Pre-trial Release	Failure to Appear	Negative	0.01 (0.008)	0.156 (0.046)	-0.146 Pass
15b - Dobbie, Goldin, Yang (2018)	Pre-trial Release	Recidivism	Negative	-0.015 (0.006)	0.015 (0.063)	-0.03 Pass
16a - Harding et al. (2018)***	Incarceration	Employment 3 yrs Postrelease	Negative	0.02 (0.00)	-0.04 (0.03)	0.06 Fail (p < 0.01)
16b - Harding et al. (2018)***	Incarceration	Re-incarceration 3 yrs Postrelease	Positive	0.12 (0.00)	0.29 (0.02)	-0.17 Fail (p < 0.01)
17 - Autor et al. (2019)	Disability Insurance	Earnings	Negative	Not shown	-5660 (2706)	Can't det.
18 - Bhuller, Dahl, Loken, Mogstad (2020)	Incarceration	Recidivism	Positive	0.052 (0.006)	-0.293 (0.106)	0.345 Pass
19a - Eren, Mocan (2021)	Juvenile Incarceration	Adult Convictions	Positive	0.119 (0.014)	0.013 (0.160)	0.106 Pass
19b - Eren, Mocan (2021)	Juvenile Incarceration	High School Graduation	Negative	-0.052 (0.016)	-0.002 (0.091)	-0.05 Pass

Table 2 - Applying *Coefficient Comparison Check* to Adjudicator Propensity-to-Treat IV Estimates (cont.)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
			"Expected"			Coefficient
Paper	Treatment	Outcome	OLS Bias	OLS Coeff	IV Coeff	Comp. Check
20 - Norris, Pecenco, Weaver (2021)	Parental Incarceration	Incarceration by Age 25	Positive	0.015 (0.004)	-0.049 (0.020)	0.064 Pass
21a - Gross, Baron (2022)	Foster Care	Child Well-Being	Negative	0.026 (0.110)	0.392 (0.164)	-0.366 Pass
21b - Gross, Baron (2022)	Foster Care	Subsequent Maltreatment	Positive	-0.032 (0.004)	-0.132 (0.058)	0.1 Pass
21c - Gross, Baron (2022)	Foster Care	Math Score	Negative	0.057 (0.013)	0.356 (0.203)	-0.299 Pass
21d - Gross, Baron (2022)	Foster Care	Juvenile Delinquency	Positive	0.041 (0.004)	-0.028 (0.040)	0.069 Pass
22a - Bald, Chyn, Hastings, Machelette (2022)	Foster Care	Test Scores (Girls)	Negative	Not Shown	1.367 (0.567)	Can't det.
22b - Bald, Chyn, Hastings, Machelette (2022)	Foster Care	Test Scores (Boys)	Negative	Not Shown	0.044 (0.562)	Can't det.
23 - Agan, Doleac, Harvey (2023)****	No Misd. Prosecution	Recidivism	Negative	-0.10 (0.010)	-0.29 (0.100)	0.19 Fail (p=0.03)
24a - Collinson et. al (2024)	Eviction	Emergency Shelter 2-years post filing	Positive	0.014 (0.001)	-0.001 (0.013)	0.015 Pass
24b - Collinson et. al (2024)	Eviction	Homeless Serv. 2-years post filing	Positive	0.02 (0.001)	0.036 (0.036)	-0.017 Fail (p=0.32)
24c - Collinson et. al (2024)	Eviction	Earnings 5-8 quarters post filing	Negative	-269 (13)	-613 (248)	344 Fail (p=0.08)
24d - Collinson et. al (2024)	Eviction	Finacial Health 5-8 q. post filing	Negative	-0.103 (0.001)	-0.141 (0.036)	0.038 Fail (p=0.15)

Positive OLS bias means β > 0 in equation (1), negative OLS bias means β < 0 in equation (1). Reported Standard errors shown in parentheses. The p shows the p-value of a one-sided test z-test of the difference in coefficients. *Doyle (2007) reports coefficients from Probit and Probit IV specificatons. **Loeffler (2013) also uses a subsample of only judges with very high or low propensities to incarcerate. Under this sample, IV passes the Coefficient Comparison Check.***Harding et al. (2018) use a slightly different IV in their preferred specification than the standard judge-propensity for treatment IV. Under this preferred specification, their results constitute a very marginal fail of the Coefficient Comparison Check. ****With respect to Agan, Doleac, Harvey (2023), see text for discussion of about expected OLS Selection bias.