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

David Bjerk*

Russell S Bock Professor of Public Economics

Claremont McKenna College

david.bjerk@cmc.edu

February 2026 Working Draft

Abstract: This article proposes that researchers employing IV methods should check whether the difference between OLS and IV coefficients has the same sign as the expected selection effect—a procedure I call the *Coefficient Difference Check*. Failures of this check imply either: the IV is invalid, the selection story is opposite of expected, or there must be very specific selection on heterogeneous treatment effects that imply the IV estimate is both fragile and a more biased estimate of the ATT than OLS. Published results often fail this check without acknowledging why or the related implications. I describe best practices given failure.

Key Words: 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.

*Special thanks to Eric Helland, Dave Abrams, Amanda Agan, Shawn Bushway, and Yotem Shem-Tov for their encouragement and comments on this project, and especially Serkan Ozbeklik for the numerous ways he helped this paper come together.

1 - Introduction

A common goal in empirical work is to determine the effect of a given treatment on some outcome of interest in settings where one expects that, in the absence of treatment, those who are treated would have had greater expected outcome values than the untreated (hereafter *positive selection-on-latent-outcomes*), or would have had lower expected outcome values than the untreated (hereafter *negative selection-on-latent-outcomes*). As is well known, in such situations, using Ordinary Least Squares (OLS) to compare actual outcomes among the treated to actual outcomes among the untreated will generally conflate any treatment effects with such selection effects even after controlling for observables. This means that in situations where one expects positive selection-on-latent-outcomes, OLS will lead to an upwardly biased estimate the average treatment effect on the treated population (i.e., the ATT). By contrast, when one expects negative selection-on-latent-outcomes, OLS will lead to a downwardly biased estimate of the ATT.

Because of this bias, in contexts where one expects there might be notable selection-on-latent-outcomes, researchers often employ an instrumental variable (IV) methodology, as this approach can provide an unbiased estimate of the average treatment effect for a subset of the treated population even when there is considerable selection regarding who is treated (Angrist, Imbens and Rubin 1996). Notably though, I show that it is fairly common, even in very prominently published papers, for the IV coefficient on the treatment to exceed the analogous OLS coefficient in contexts where an IV approach was employed due to a concern that positive selection-on-latent-outcomes would cause the OLS coefficient to be upwardly biased, or vice versa in contexts where the IV was employed due to a concern that negative selection-on-latent-outcomes would cause OLS to be downward biased. This paper considers this issue in depth, showing exactly what can lead to such seemingly contradictory results and discussing what researchers should do when it arises.

Using what I hope is a very accessible econometric treatment, I derive a simple equation showing that, given a valid IV for a binary treatment, the expected difference between the OLS coefficient and IV coefficient on the treatment will equal the *selection-on-latent-outcomes effect* plus the difference in the average treatment effects between the *always-treated* (i.e., those who get treated regardless of their value of the instrument) and

the *compliers* (i.e., those whose treatment depends on their value of the instrument), where this difference term is weighted by one minus the fraction of the treated who are *compliers*.

This equation helps to concisely synthesize much of the literature on interpreting IV results under potentially heterogenous treatment effects, and notably, also motivates a simple diagnostic related to the discussion above, which for convenience I refer to as the *Coefficient Difference Check*. Namely, I argue that when performing IV estimation, researchers should explicitly check whether the sign on the difference between the OLS and IV coefficients on the treatment is consistent with the sign of the expected selection-on-latent-outcomes effect.

Due to the relationship between the OLS coefficient and the IV coefficient inherent in the equation described above, results that fail to be consistent with this *Coefficient Difference Check* imply one of the following: (i) the required conditions for the IV to be valid do not hold, (ii) the actual selection-on-latent-outcomes effect is the opposite of what is expected, or (iii) there are notably heterogenous treatment effects such that the difference in average treatment effects between the *always-treated* and the *compliers* has the opposite sign *and* is significantly larger in absolute magnitude than the *selection-on-latent-outcomes effect*. I further show that this heterogenous treatment effects story has three additional implications. First, it implies that the IV estimate will actually provide a *more* biased estimate of the average treatment-on-the-treated (ATT) than OLS. Second, it implies that the value of the IV coefficient must be fragile to modest alterations in the group of *compliers* due to changes in the IV or setting. Third, it implies there must be very specific *selection-on-magnitude* of treatment effects that may be hard to square with theory or intuition. Notably, no such implications arise when IV results are consistent with this check.

I think it is important to note that I explicitly refer to this procedure a *check* rather than a *test*, because as stated above, results that fail to be consistent with this check do not necessarily imply that the IV is invalid. However, the argument above does suggest that when IV results fail to be consistent with the *Coefficient Difference Check*, this indicates issues with respect to the IV results that the researcher should explicitly acknowledge.[1]

[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. As I discuss in Section 4, in such

So what should researchers do when their IV results fail to be consistent with the *Coefficient Difference Check?* The first thing is to re-evaluate the validity of the candidate IV, both intuitively, and possibly explicitly via applying one or more the more complicated IV diagnostics that have been previously developed (see section 5 below for an overview of some of these diagnostics). If the researcher remains convinced the IV is valid, then the researcher should take a stand on which of the other two possible reasons accounts for the failure and attempt to provide evidence for this.

This means that, on the one hand, if the researcher argues that the results fail the *Coefficient Difference Check* due to a different selection-on-latent-outcomes story than basic theory or intuition would suggest, it would be helpful for the researcher to provide a reasonable motivation, and possibly empirical evidence consistent with this alternative story. On the other hand, if the researcher argues that the violation of the *Coefficient Difference Check* is due to heterogenous treatment effects, it would be helpful to provide evidence consistent with the existence of such large heterogeneities in treatment effects across the treated population, such as showing substantially different IV estimates across different sub-populations. Additionally, though, the researcher should also acknowledge the notable implications alluded to above that go along with this heterogenous treatment effects story. While these practices are generally not sufficient to rule out an invalid IV as the cause of the failure of the *Coefficient Difference Check*, they would provide helpful evidence in support of an alternative explanation.

For illustrative purposes, consider a simple example. Suppose a researcher uses an IV to estimate the effect of a voluntary job training treatment for disadvantaged youth on future employment. Further suppose that the researcher both expects and finds evidence consistent with *positive* selection-on-latent-outcomes regarding who is treated, but yet finds the difference between the OLS and IV coefficients is negative. Given these results fail to be consistent with the *Coefficient Difference Check*, the researcher should acknowledge either that the IV is invalid, or argue that there must be particular heterogenous treatment effects that imply (i) the IV results provide a more biased estimate of the ATT than OLS, (ii)

situations, it still seems reasonable to compare the OLS to IV coefficients as proposed here and attempt to determine what exactly accounts for the difference between the two.

the IV results would likely change with small changes in the sample or IV, and (iii) there must be negative selection-on-magnitude of the treatment effects, or in other words, those most prone to select into treatment must on be those who on average benefit the least. These are strong, but also potentially illuminating implications.

To help provide more real-world context for the issues I am discussing, I construct a couple of examples of treatments with a clear direction of selection-on-latent-outcomes bias, and then use real data to estimate the associated treatment effects using a candidate IV that is almost surely invalid due to implausibility of the required exclusion restriction. I show that the resulting IV estimates indeed fail to be consistent with the *Coefficient Difference Check*. I further confirm that there is little evidence that this failure is due to the expected selection-on-latent-outcomes story being incorrect. However, I show that some seemingly large heterogeneities in the IV estimates of treatment effects can arise across different subpopulations, even though it is almost surely the invalidity of the instrument that causes the IV results to fail to be consistent with the *Coefficient Difference Check*. This highlights another reason why researchers should be cautious to assign failures of the *Coefficient Difference Check* to heterogenous treatment effects.

I want to be clear that most researchers who employ IV methods likely already have some awareness of the basics of what I highlight in this paper, though maybe not in the detail laid out here. Regardless, as alluded to above, many papers that employ an IV approach present results inconsistent with this *Coefficient Difference Check* with little comment, or fail to include the OLS results necessary to perform this check. I highlight this issue by examining a group of roughly two dozen 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 tier journals, I show that the results in over half of these papers are inconsistent with this *Coefficient Difference Check* or do not provide the OLS results necessary to perform this check. I go on to discuss how failures of this check are also common in other IV settings, such as the return-to-education literature. Furthermore, I discuss how applying this *Coefficient Difference Check* can help us make sense of some seemingly contradictory results in the literature, as well as highlight a couple of papers that

implicitly acknowledge the issue I am analyzing in this paper, and how the *Coefficient Difference Check* and associated analyses can help clarify some potential issues therein.

In the end, standard practice when employing an IV methodology is to directly evaluate the "first-stage"—is the instrumental variable sufficiently correlated with the treatment in the expected way? I argue that applied researchers should also directly evaluate this *Coefficient Difference Check*—is the sign on the difference between the OLS and IV coefficients consistent with the expected selection-on-latent-outcomes? When IV results fail this check, researchers should provide an argument for what they think accounts for this failure, as well as delineate and consider the associated implications.

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. Using 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 ($D_i = 0$). where these potential outcomes are captured by the following equations

$$Y_i(1) = \alpha + \delta(u_i) + \beta u_i + \varepsilon_{1,i} \qquad (1a)$$

and

$$Y_i(0) = \alpha + \beta u_i + \varepsilon_{0,i}, \qquad (1b)$$

with u_i being some characteristic of individual i that is *unobservable* to the researcher, and $Y_i(1) - Y_i(0) = \delta(u_i)$, meaning that $\delta(u_i)$ is the *treatment effect* for individual *i.* The function $\delta(u)$ captures any relationship between the magnitude of the treatment effect and the unobservable characteristic u. Homogenous treatment effects correspond to $\delta(u)$ being constant over u, Heterogeneous treatment effects correspond to $\delta(u)$ being differing over different values of u. The treatment effect for individual i will be positive if $\delta(u_i) > 0$, while will be negative if $\delta(u_i) < 0$. The parameter β can be positive or negative.

Let the unobservable variable u come from a distribution $F(u)$ over the population of interest. Lastly, assume that the residual terms $\varepsilon_{0,i}$ and $\varepsilon_{1,i}$ are mean zero independent random variables. Note that one could also include some function $h(X_i)$ in equations (1a-

1b), where X_i is a vector of other characteristics observable to the researcher that influence the outcome Y. This would just add excess notation, though, so I forgo doing that here. Therefore, in this context, one can think of the outcome Y_i in equation (1) as the residual from regressing the true outcome of interest, $\tilde{Y}_i$, on the function $h(X_i)$.

Furthermore, suppose whether or not some individual i receives the treatment D is determined by the following equation

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

where Z_i is some individual characteristic observable to the researcher that influences whether or not an individual is treated, u_i is the same unobservable variable that was in the outcome equations (1a-1b), $g(u_i)$ is some increasing function of u_i, and ρ is just a scaling parameter.

Given this set-up, a person i whose realized value of $Z_i = z$ receives the treatment if and only if $u_i \geq u_z^*$, where

$$u_z^* = \rho - \lambda z. \qquad (3)$$

In words, for each value z, there is a cutoff such that a person i with $Z_i = z$ will be treated if and only if his/her value of u_i exceeds this cutoff. Given each person's u_i is drawn from a distribution $F(u)$, this means that among those with a value of $Z_i = z$, the fraction treated will be $1 - F(u_z^*)$. Without loss of generality, we can assume $\lambda > 0$, meaning this threshold u_z^* is decreasing in Z, implicitly meaning Z is defined in a way such that the likelihood an individual is treated is increasing in Z.[2] It is also without loss of generality to assume $g(u_i)$ is increasing in u_i, as it just means we have defined the unobservable component in a way such that as it increases so does the likelihood of treatment for any given Z.

To make explication easier, suppose Z can only take on two values, z_1 and z_0, where $z_1 > z_0$. This implies two cutoffs, $u_1^* = \rho - \lambda z_1$ and $u_0^* = \rho - \lambda z_0$, where $u_1^* < u_0^*$. This is only for simplicity though, as all arguments hold if Z can take on more than two values.[3] 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

[2] What is key is that λ simply cannot equal zero, as one can always define Z and u to make it true that $\lambda > 0$.
[3] With more than two values of Z, there are correspondingly more cutoff values of u. However, the only ones that really matter are the two extreme values of Z, and their corresponding cutoff values.

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*.[4] Finally, everyone with $u_1^* < u_i \leq u_0^*$ will be treated if they have $Z_i = z_1$ and be untreated if $Z_i = z_0$, a population generally referred to as *compliers* and consists of those whose treatment status would fluctuate 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, always-treated, and never-treated* would be composed of different subgroups of *u*, a detail I will return to below.

Note that this set-up allows for a variety of different selection effects. As is standard, let us define a *selection-on-latent-outcomes* effect as the difference in expected outcomes in the absence of treatment between those who are actually treated and those who aren't, or

$$
\begin{aligned}
&E[Y(0)|D_i = 1] - E[Y(0)|D_i = 0] \\
&\quad = \frac{\pi_Z \pi_C}{\pi_Z \pi_C + \pi_A} E[Y(0)|u_1^* < u \leq u_0^*] + \frac{\pi_A}{\pi_Z \pi_C + \pi_A} E[Y(0)|u_0^* < u] \qquad (4) \\
&\quad - \frac{(1-\pi_Z)\pi_C}{\pi_N + (1-\pi_Z)\pi_C} E[Y(0)|u_1^* < u \leq u_0^*] - \frac{\pi_N}{\pi_N + (1-\pi_Z)\pi_C} E[Y(0)|u \leq u_1^*].
\end{aligned}
$$

where:

$\pi_Z \equiv$ the fraction of the population with $Z = z_1$,

$\pi_C \equiv F(u_0^*) - F(u_1^*)$ which is the fraction of the population who are *compliers*,

$\pi_A \equiv 1 - F(u_0^*)$ which is the fraction of the population who are *always-treated,*

$\pi_N \equiv F(u_1^*)$ which is the fraction of the population who are *never-treated.*

This selection-on-latent-outcomes effect in equation (4) will be positive if $Y(0)$ is increasing in *u* (i.e., if $\beta > 0$ in eq. 1b), or negative if $Y(0)$ is decreasing in *u* (i.e., $\beta < 0$).

This setup also allows for different types of *selection-on-magnitude* of the treatment effect (a broadening of "selection on gains"). For example, given *g(u)* is increasing in *u* in equation (2), if $|\delta(u)|$ is increasing in *u*, this would mean increases in the magnitude of the treatment effect correspond with increasing likelihood of treatment, or positive selection-on-magnitude of treatment effects. Negative selection-on-magnitude of treatment effects

[4] In the literature, these groups are alternatively referred to as *never-takers* and the *always-takers*, respectively.

would correspond to $|\delta(u)|$ decreasing in u. Note, it is straightforward in this setup to have both for selection-on-magnitude of treatment effects and selection-on-latent-outcomes.

Given this setup, suppose the researcher's goal is identify the average treatment effect $\delta(u_i)$ over some well-defined portion of the population. The simplest approach is to simply compare the mean outcome among the treated to the mean outcome among the untreated. For simplicity, I refer to this as the OLS estimate, or δ^{OLS}. As is well known, however, when there is notable selection-on-latent-outcomes regarding who is treated, this estimate will be will be a biased estimate of the average treatment effect among the treated, as it will conflate the true average treatment effect with any selection-on-latent-outcomes effect. I return to this in more detail below, but first let us recall how a variable like *Z* can potentially be used as an Instrumental Variable (IV) for identifying an unbiased measure of the average treatment effect for a specific subset of the population.

The Estimated Treatment Effect from Using 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.[5]
2. *Random Assignment* – Asymptotically, the distribution of observable and unobservable characteristics are the same for any value of *Z*.
3. *Relevance* – Expected treatment status is a non-trivial function of Z.
4. *Monotonicity* – If we let $D_i(Z)$ *be* person *i's* treatment status given his value for Z, it must be the case that either $D_i(z_1) \geq D_i(z_0)$ for all *i,* or $D_i(z_1) \leq D_i(z_0)$ for all *i*. [6]
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.

It is straightforward to show that all these conditions hold in the model developed above. Given this, 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

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

under different forms of selection-on-latent-outcomes and different direction of treatment effects. For simplicity, within each of these graphs, I have depicted treatment effects to be homogenous, meaning $\delta(u) = \delta$ for all u. 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 latent expected outcome for each value of u for the opposite of the realized treatment status.

This all means that, in each figure, those with $u_i \leq u_1^*$ are the *never-treated*, and hence have the solid portion associated with $\mathrm{E}[Y(0)|u]$. Those with $u_i > u_0^*$ 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-on-latent-outcomes and *positive* treatment effects. Figure 1b depicts *positive* selection-on-latent-outcomes and *negative* treatment effects, Figure 1c depicts *negative* selection-on-latent-outcomes and *positive* treatment effects, and finally, Figure 1d depicts *negative* selection-on-latent-outcomes 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 $Z = z_0$. This is often what is referred to as the *Reduced Form* estimate and will therefore be denoted δ^{RF}. As can be seen in each of these graphs in Figure 1, if we 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 this 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-on-latent-outcomes with respect to treatment, or whether or not the treatment effects are positive or negative, if Z is a valid IV, then comparing the average outcomes among 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*. If we denote the average treatment effect among the *compliers* as $\delta^C \equiv E[\delta(u)|u_1^* < u \leq u_0^*]$, and recall that the fraction of the population that are *compliers* is given by $\pi_C \equiv F(u_0^*) - F(u_1^*)$, we get

$$\delta^{RF} = \pi_C \delta^C + [1 - \pi_C] * 0.$$

Thus, if we define δ^{IV} to be the above Reduced Form estimate upweighted by the inverse of the fraction of the population who are *compliers* π_C, δ^{IV} will be such that

$$\delta^{IV} \equiv \frac{\delta^{RF}}{\pi_C} = \delta^C. \qquad (5)$$

This just reveals the well-known result that if Z is a valid instrumental variable, the IV estimate of the treatment effect, δ^{IV}, captures 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 (Imbens and Angrist, 1994). In terms of Figure 1, the IV estimate will be a weighted average of the height of the area between the solid black line and solid grey line over the domain u_1^* to u_0^*. See the appendix for a formal derivation of the IV estimate of a treatment effect under this set-up.

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

Now, let us again consider the OLS estimate

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

With no other covariates, 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. With a little manipulation, we can write down the expected outcome among the actually treated, which consists of the *compliers* among those with $Z_i = z_1$ and the *always-treated* among those with either value for Z_i, as

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

where π_Z, π_C, and $\pi_A$were defined above. 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

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

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, δ^{OLS} will simply be equation (6) minus equation (7), meaning

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

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

$$\delta^{OLS} = \pi_{D=1}^C \delta^C + (1 - \pi_{D=1}^C)\delta^A + E[Y(0)|D = 1] - E[Y(0)|D = 0] \quad (8),$$

where $\pi_{D=1}^{C}$is defined to be the fraction of the treated who are *compliers* (i.e., $\pi_{D=1}^{C} \equiv \frac{\pi_z \pi_C}{\pi_z \pi_C + \pi_A}$), δ^C is the average treatment effect among the compliers (as defined above), and $\delta^A \equiv E[\delta(u)|u_0^* \leq u]$ or 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, which we defined above to be the *selection-on-latent-outcomes effect*. In other words, equation (8) simply shows that δ^{OLS} corresponds to the average treatment-on-the-treated (ATT) plus any selection-on-latent-outcomes effect, confirming that the OLS estimate will be a biased measure of the ATT in the direction of the selection-on-latent-outcomes effe*ct*.

Moving on, if we subtract δ^{IV} from both sides, we get

$$\delta^{OLS} - \delta^{IV} = \pi_{D=1}^{C}\delta^C + (1 - \pi_{D=1}^{C})\delta^A + (E[Y(0)|D=1] - E[Y(0)|D=0]) - \delta^{IV}.$$

Finally, recalling that $\delta^{IV} = \delta^C$, the previous equation can be simplified to

$$\delta^{OLS} - \delta^{IV} = (1 - \pi_{D=1}^{C})(\delta^A - \delta^C) + \underbrace{(E[Y(0)|D=1] - E[Y(0)|D=0])}_{\textit{Selection-on-latent-outcomes effect}}, \qquad (9)$$

meaning the difference between the OLS and IV coefficients on the treatment equal the difference between the average treatment effect among the always treated and the average treatment effect among the compliers (weighted by one minus the fraction of the treated who are compliers), plus the selection-on-latent-outcomes effect. While equation (9) is quite intuitive, I have not previously seen it explicitly derived or expressed this way.

Several straightforward implications arise from equation (9). Most simply, if treatment effects are relatively homogenous across the population, implying $\delta^A \approx \delta^C$ for any valid IV, then the difference between the OLS and IV coefficients will just reflect the selection-on-latent-outcomes effect. By contrast, even if there is no selection-on-latent-outcomes, OLS and IV coefficients can still notably differ if there are large heterogeneities

in treatment effects across the population, as this could cause δ^A to differ from δ^C.[7] This implies that a Hausman test can potentially suggest a selection-on-latent-outcomes bias even if there is none. Equation (9) also shows that as the instrument approaches becoming perfect predictor of treatment, as would be the case in experimental drug RCTs for example, the fraction of the treated who are compliers ($\pi^C_{D=1}$) approaches one and the selection-on-latent-outcomes effect goes to zero, thus implying the difference between the OLS coefficient and the IV coefficient will converge to zero.

This all leads to the following theorem.

***OLS vs IV Theorem*:** If the required assumptions with respect to the proposed IV hold, then equation (9) directly implies that:

1. If there is positive selection-on-latent-outcomes, 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., $\delta^A - \delta^C < 0$) that it offsets the selection-on-latent-outcomes effect.
2. If there is negative selection-on-latent-outcomes, 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., $\delta^A - \delta^C > 0$) that it offsets the selection-on-latent-outcomes effect.

To put the above theorem in another way, under a valid IV, the sign of the difference between the OLS and IV coefficients on a treatment can differ from the direction of the selection-on-latent-outcomes effect only if there is sufficient selection-on-magnitude of treatment effects that are large enough and in the right direction such that the difference between δ^A and δ^C is sufficient to offset the selection-on-latent-outcomes effect.

So, to re-iterate the example from the introduction, consider evaluating of the effect of a voluntary job training program for disadvantaged youth on future employment. In such a context, if the IV is valid and there is *positive* selection-on-latent-outcomes, $\delta^{OLS} - \delta^{IV}$ can

[7] While expressed differently, this is similar to discussions in Ishimaru (2022) and Lochner and Moretti (2015).

only be negative if $\delta^A - \delta^C$ is sufficiently smaller than zero, which would only be true when the absolute magnitude of $\delta(u)$is falling in u—meaning those most prone to select into the program (i.e., those with high u who would be among the *always-treated*) gain the least by doing so. In other words, there would need to be substantial *negative* selection-on-magnitude of treatment effects (i.e., the opposite of "selection on gains").

3 – The Coefficient Comparison Check

The previous discussion and the *OLS vs IV Theorem* lead to the following:

***Coefficient Difference 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-on-latent-outcomes, check whether $\hat{\delta}^{OLS} - \hat{\delta}^{IV} > 0$.
2. If one expects *negative* selection-on-latent-outcomes, check whether $\hat{\delta}^{OLS} - \hat{\delta}^{IV} < 0$.

If the estimated coefficients fail to be consistent with check, then the researcher must:

(i) Argue that selection-on-latent-outcomes is opposite way of expected, or

(ii) Argue that the average treatment effect among the *always-takers* not only differs from that of the *compliers*, but also that this difference is opposite in sign *and* larger in magnitude than the selection-on-latent-outcomes effect, or

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

As the fraction of the treated who are *compliers* rises, the heterogenous treatment effects in option (ii) becomes less and less plausible.

Note that the above check is with respect to real values, not absolute values. Moreover, this check applies for both positive and negative selection-on-latent-outcomes, and positive and negative treatment effects. How this works is sufficiently nuanced that it will help to go through each of these possibilities using Figures 2a-2d, which are constructed analogously to Figure 1, but with heterogenous treatment effects. Given $\delta^{IV} = \delta^C$, we know the IV estimate in each of these figures corresponds to a weighted average of the height of the area between the black solid line and the grey solid line over the domain u_1^* to u_0^*. By

contrast, δ^A corresponds to a weighted average of the height of the area between the solid lines and the dashed lines over the domain $u > u_0^*$.

Given this, suppose one expects there is *positive selection-on-latent-outcomes* and finds a *positive* treatment effect among the *compliers* (i.e., $\hat{\delta}^{IV} > 0$), but *fails* to find that $\hat{\delta}^{OLS} - \hat{\delta}^{IV} > 0$. If the researcher wants to "keep" these IV results, then he/she must either argue that the selection-on-latent-outcomes effect is actually negative, or argue that there are notably negative selection-on-magnitude of treatment effects, in the sense that treatment effects must be smaller in magnitude for those more prone to treatment (equivalent to |δ(u)| decreasing in u) . This would be necessary for the average treatment effects among the *always-treated* (δ^A) to be sufficiently small relative to the average treatment effects among the *compliers* (δ^A) to overcome the positive selection-on-latent-outcomes effect. Such a heterogenous treatment effects story is depicted in Figure 2a.

By contrast, suppose one expects *positive* selection-on-latent-outcomes 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 $\hat{\delta}^{IV}$). If the researcher wants to "keep" these IV results, he/she must argue that selection-on-latent-outcomes is actually negative, or argue that there is notably *positive* selection-on-magnitude of treatment effects (equivalent to |δ(u)| increasing in u), so that $\delta^A - \delta^C$ is sufficiently negative to overcome the positive selection-on-latent-outcomes. This is depicted in Figure 2b.

On the other hand, suppose one expects there is *negative* selection-on-latent-outcomes and finds a *positive* treatment effect among the *compliers* (i.e., $\hat{\delta}^{IV} > 0$), but *fails* to find that $\hat{\delta}^{OLS} - \hat{\delta}^{IV} < 0$. If the researcher wants to "keep" these IV results, he/she must either argue that selection-on-latent-outcomes is actually positive, or argue that there are notably *positive* selection-on-magnitude of treatment effects (equivalent to |δ(u)| increasing in u) to ensure the average treatment effects among the *always-treated* are sufficiently more positive than they are among the *compliers* to overcome the negative selection-on-latent-outcomes. This is depicted in Figure 2c.

Finally, if one expects *negative* selection-on-latent-outcomes 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" these IV results, he/she must either argue that selection-on-latent-outcomes is actually positive, or argue that there is are notably negative selection-on-magnitude of treatment effects to ensure δ^A is sufficiently small in magnitude relative to δ^C so that their difference can overcome the negative selection-on-latent-outcomes. This is depicted in Figure 2d.

This all suggests that when IV results fail to be consistent with the *Coefficient Difference Check*, it should be incumbent upon the researcher to thoroughly assess the chosen IV, potentially through applying one or more of the more involved IV diagnostic methods discussed below in Section 5. If the researcher remains convinced the IV is valid, he/she should then provide a plausible argument for which other explanation he/she thinks is behind the failure. Namely, if a researcher argues that the reason IV results fail the *Coefficient Difference Check* is because selection-on-latent-outcomes is opposite of expected, then the researcher should provide a story and even evidence in support of such an argument, such as showing that the differences in observable characteristics between treated and untreated suggest a different selection-on-latent-outcomes story than simple intuition would suggest. Alternatively, if the researcher proposes that selection-on-magnitude of heterogenous treatment effects is the reason for the failure, the researcher should provide evidence in support of this, such differential IV estimates for different observable groups.[8]

Moreover, as stated in the *OLS vs IV Theorem*, for the heterogenous treatment-effects to provide a plausible argument for a *Coefficient Difference Check* failure, the difference in average treatment effects between the *always-treated* and the *compliers* must be notably greater in absolute magnitude but of opposite sign than the selection-on-latent-outcomes effect. This means that even when such a heterogenous treatment effects story is potentially plausible, the estimated $\hat{\delta}^{IV}$ must 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

[8] One might consider estimating Marginal Treatment Effects (MTEs) here (see Heckman and Vytlacil (2005), Carneiro, Heckman, Vytlacil (2011), Mogstad, Santos, Torgovitsky (2017)). Such formulation essentially try to capture how treatment effects differ by latent "resistance to treatment", which can roughly be thought of as the opposite of the unobservable *u* variable in the model developed earlier in this paper (i.e, a higher resistance to treatment is correlated with a lower likelihood of treatment). However, it should also be noted that estimated MTEs are often quite sensitive to parameterizations, method, and control variables (Andresen 2018).

large changes in the estimated $\hat{\delta}^{IV}$, as in such a scenario, the underlying δ^{IV} will be a highly variable parameter even within one particular setting depending on the exact IV employed.

This issue can be seen graphically in Figure 3. The two graphs in Figure 3 both correspond to a single setting with positive selection-on-latent-outcomes and positive treatment effects, as well as heterogenous treatment effects in which there is negative selection-on-magnitude of these treatment effects, meaning they go the opposite direction of selection-on-latent-outcomes (the same situation as depicted in Figure 2a). In other words, the magnitudes of the treatment effects are smaller for those most prone to receiving the treatment, and is one type of situation that could cause IV results to be inconsistent with the *Coefficient Difference Check* even with a valid IV. However, the two graphs in Figure 3 differ in that they characterize two different IVs that each capture a different group of *compliers* via different values for u_1^* and u_0^*.

In each graph in Figure 3, he IV estimate, which corresponds to the average treatment effect among the compliers, is essentially the height of the bold black polygons weighted by the density of *F(U)* over the domain of compliers (u_1^* to u_0^*) under each instrument. As can be seen, this IV estimate will be quite different in these two situations due to heterogeneous treatment effects and different *compliers* associated with the two IVs, despite the fact that both figures depict the *identical* underlying treatment effects scenario.

This last point relates to previous arguments made by several others, notably 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 additional siblings on the first born's education, showing different IVs can lead to dramatically different results (indeed the magnitudes and even the sign on the estimated treatment effect differs across different IVs). Importantly, IV results that fail to be consistent with the *Coefficient Difference Check* can be a simple indicator of when this might be an issue, especially in situations where a researcher only has one candidate instrumental variable.

Furthermore, a failure of the *Coefficient Difference Check* due to heterogenous treatment effects means the IV results cannot be representative of the average treatment effect over the population (ATE), nor the average treatment-on-the-treated (ATT). Indeed, given OLS should provide an upward or downward bound on the ATT depending on the direction of selection, it is straight-forward to show that the specific conditions required for heterogenous treatment effects to lead to a failure of the *Coefficient Difference Check* directly imply that the IV estimate will provide a *more biased* estimate of the ATT than the OLS estimate in such a situation (see appendix for full details). This means statements (or paper titles) suggesting the IV results provide an estimate of "the effect of treatment *D* on outcome *Y*" are particularly misleading when IV results fail the *Coefficient Difference Check*.

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

This section considers this *Coefficient Difference Check* with respect to real world data. The first part examines what happens when this *Coefficient Difference Check* is applied to IVs known to be invalid. The second part then considers the frequency of failures of the *Coefficient Difference Check* in the context of a variety of prominently published results using a particular form of IV—namely adjudicator propensity-to-treat IVs—and how this check can help us better understand the results in some of these papers. I wrap up this section by discussing findings that are inconsistent with this check more broadly.

The *Coefficient Difference Check* and "Bad" IVs

Consider trying to estimate the effect of completing college on adult earnings, or alternatively, estimating the effect of completing high school on the likelihood of arrest in young adulthood. In thinking about the first scenario, one would expect those who obtain a college degree 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, regarding the second scenario, those who complete high school almost surely would have a lower likelihood of arrest as a young adult than those who drop out of high school, even in the absence of 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 biological 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* in the situations examined here—i.e., 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 an invalid IV, as proof of concept, we can apply the *Coefficient Difference 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.[9] Column (5) then shows the results of the first-stage regression, or the results from regressing the treatment variable on the IV and any other control variables, 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 Difference 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 calculating statistical significance is all that important here).

Looking first at column (5), we see that the IV (parent college) is positively and significantly 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*.

[9] 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).

Given the expected *positive* selection-on-latent-outcomes when it comes estimating the effect of completing college on earnings, the *Coefficient Difference Check* for such specifications assesses whether the difference between the OLS coefficient and the IV coefficient is positive. 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, which means these results fail to be consistent with the *Coefficient Difference 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 results remain inconsistent with the *Coefficient Difference Check* in the sense that the difference between the OLS coefficient and the IV coefficient remains negative. However, this difference between OLS and IV is not statistically significantly negative using a one-sided test (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-on-latent-outcomes, so the *Coefficient Difference Check* assesses whether difference between the OLS coefficient and the IV coefficient is also negative. 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. Therefore, these results are also inconsistent with the *Coefficient Difference 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 no longer "statistically" so).

This exercise shows that the "bad" IVs used above will indeed cause the results to "fail" the *Coefficient Difference Check*. However, when covariates are included, one might say these failures are "marginal" in that the differences between OLS and IV are not statistically significantly different from zero in the required direction for a failure. However, the question at hand is whether the IV results provide a more informative estimate than OLS. Given the OLS coefficient should be an upper bound on the size of the ATT given positive selection-on-latent-outcomes, and a lower bound on the ATT given negative selection-on-latent-outcomes, and the standard errors will always be smaller for OLS than IV results, I'd argue that failure of the *Coefficient Difference Check* is something

that requires addressing regardless of whether or not this failure is statistically significant. To put this another way, if one motivates using an IV approach because of a concern that OLS is upwardly (downwardly) biased, it seems strange to focus on a reported IV coefficient that is even larger (smaller) than the OLS coefficient, and has larger standard errors, without any comment. Therefore, I think it reasonable to acknowledge and address results that are inconsistent with the *Coefficient Difference Check* even if the difference in coefficients is not statistically significant.

In the exercises above, the reason the IV results fail to be consistent with the *Coefficient Difference Check* was almost surely because the IV does not meet the exclusion restriction. However, it is still informative to consider what the evidence looks like for the other reasons IV results may fail to be consistent with the *Coefficient Difference Check.* In terms of selection-on-latent-outcomes when it comes to arrest as a young adult, the "treated" individuals (i.e., those who graduated from high school) were more likely to be female, less likely to have been born to a teen mom, have higher standardized test scores, and come from households with higher income. All of these are correlated with a lower likelihood of being arrested as a young adult, suggesting *positive* selection-on-latent-outcomes in this context is highly unlikely, as selection on unobservables would have be large and in the opposite direction of selection on observables. When it comes to earnings as a young adult, a similar exercise comparing the average characteristics of those who graduate from college (i.t., the "treated") to those who did not (the "untreated") suggests *negative* selection-on-latent-outcomes in this context is highly unlikely. In other words, as should be expected, at least for observable characteristics, there is little evidence to suggest that the IV results in these exercises fail to be consistent with the *Coefficient Difference Check* because the expected selection-on-latent-outcomes story is wrong.

Interestingly, however, even though we know the IV results likely fail the *Coefficient Difference Check* primarily due to an invalid IV, if one were very committed to this IV, one might be tempted to argue that the reason for the failure is heterogenous treatment effects. The reason being that, with respect to the impact of high school graduation on young adult arrest, the IV results are almost twice as large in absolute magnitude for males relative to females, and for those in the bottom half of the test score distribution relative to those in the top half. Notably, these treatment effect heterogeneities suggest those more prone to

treatment (females, those with higher test scores) incur smaller treatment effect magnitudes than those less prone (males, those with lower test scores), which would be a requirement for this explanation to account for the failure of the *Coefficient Difference Check*. On the other hand, no such heterogeneities across observable groups arise with respect to the IV results regarding the impact of college graduation on earnings.

In summary, this section used two examples to show that when an IV was highly likely to be invalid, the IV results indeed fail to be consistent with the *Coefficient Difference Check*. Moreover, comparing the characteristics of the treated versus the untreated would provide little evidence consistent with the selection-on-latent-outcomes story being opposite of what was expected. However, there are a couple of notable heterogeneities with respect to the IV results which suggests that finding the presence of heterogeneities in treatment effects after a failure of the *Coefficient Difference Check* is likely not sufficient to rule out that an invalid IV as the cause of the failure.

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

This subsection applies the *Coefficient Difference 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, a criminal court judge (adjudicator) who must decide whether or not to incarcerate (treat) the guilty defendants in his/her court. In such cases, any given judge is arguably more prone to impose incarceration on the defendants whose characteristics suggest he or she is likely to quickly re-offend, thought all such characteristics are not likely to be observed by researchers. 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-on-latent-outcomes into this treatment in the sense that those treated with incarceration would have a higher expected re-offending rate than the untreated 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, negative, or negligible) and the selection-on-latent-outcomes effect (which one expects to be positive here).

To overcome such a bias in these situations, researchers have often exploited the fact that potentially treated subjects 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" and "examiner tendency" IVs (see Chyn, Frandsen, and Leslie (2025) for a thorough discussion).

Table 2 below shows the results from 37 different estimates from 21 different published studies of the effect of some treatment on some outcome using an adjudicator propensity-to-treat IV.[10] 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.[11] Indeed, the overall strength and influence of these papers is exactly why they make good case studies for the issues I am raising in this article.

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

[10] Note, I exclude four prominent papers that use this IV methodology to look at the impact of disability payments on different outcomes (Maestas, Mullen, and Strand (2013), Dahl, Kostol, Mogstad (2014), French and Song (2014), Autor et al. (2019)), as a notable debate on the expected direction of the OLS bias in this literature. Though I return to discuss the results of one of these papers later.

[11] The included papers appeared in American Economic Review, Quarterly Journal of Economics, Journal of Political Economy, Review of Economics and Statistics, American Economic Journal: Economic Policy, American Economic Journal: Applied Economics, Journal of Law & Economics, Criminology, and American Journal of Sociology.

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 direction of the expected selection-on-latent-outcomes bias given the context. Columns (5) and (6) show the OLS and IV coefficients on the treatment respectively. The last column in Table 2 shows the difference between columns (5) and (6) and states whether the IV results pass or fail the *Coefficient Difference Check*, or cannot be determined due to lack of OLS results reported in the paper.

As can be seen, while a many of these results are consistent the *Coefficient Difference Check*, there are also a substantial number that either fail or do not allow one to conduct the *Coefficient Difference Check* since they don't report the OLS results. Indeed, of the 37 separate estimates shown, 23 either fail to be consistent with this check or do not present the OLS results necessary to perform this check. In these papers where the results fail to be consistent with this check, only a few explicitly acknowledge this (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 8a and 8b) and the Eren and Mocan (2021) paper (see rows 15a and 15b) 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-on-latent-outcomes. 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-on-latent-outcomes) and lower high-school graduation rates (negative selection-on-latent-outcomes) than the juveniles who weren't incarcerated. Notably, Aizer and Doyle's IV results suggest juvenile incarceration

substantially 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 in the previous paragraph imply that the *Coefficient Difference Check* in this context consists of checking whether the difference between the OLS coefficient and the IV coefficient is positive with respect to recidivism, and negative when it comes to high-school graduation. As can be seen in Table 2, Aizer and Doyle's results (rows 8a and 8b) fail to be consistent with the *Coefficient Difference Check*, while Eren and Mocan's results (rows 15a and 15b) are consistent with this check.

This makes one wonder why Aizer and Doyle's results are not consistent with the *Coefficient Difference Check*. One reason for this inconsistency could be that the exclusion restriction is invalid for this IV (option *iii* in the *Coefficient Difference Check*). For example, in Aizer and Doyle's context of Chicago, more lenient judges may have helped the juveniles they adjudicated in ways other than just being less likely to incarcerate, such as through what they said to these youths in court or via other treatments they may have applied.

However, if one doesn't believe this invalid IV story, then one must acknowledge that either the selection-on-latent-outcomes story is opposite of expected (option *i*) or that there is selection-on-magnitude of heterogeneous treatment effects sufficient to offset the selection-on-latent-outcomes effect (option *ii*). Given Aizer and Doyle's argument for using the IV approach, it seems unlikely that the selection-on-latent-outcomes story goes in the opposite direction than expected here. This leaves option *ii*, or that the treatment effects among the *always-treated* are sufficiently different than the treatment effects among the *compliers* to overcome the *selection effect*.

The type of selection-on-magnitude of treatment effects that could account for why Aizer and Doyle's recidivism results fail to be consistent with the *Coefficient Difference Check* corresponds to what was depicted in Figure 2a and, equivalently, Figure 3, where the treatment effects are smaller for those most prone to receiving the treatment. As Figure 3 showed, IV results under such a scenario are likely quite fragile, as small changes in the IV, such as what would occur with the retiring of a couple of particularly harsh judges, or some

judges becoming more lenient over time, would shift the bolded "polygon of *compliers*" rightward in Figure 3, thereby notably shrinking the IV results. [12]

Given Eren and Mocan's results are consistent with the *Coefficient Difference Check*, one might argue that Eren and Mocan's results are the more robust of the two. However, the fact that the results in one of these studies fail the *Coefficient Difference 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 different groups of *compliers*. However, regardless of whether the failure is due to an invalid IV or heterogeneous treatment effects, the argument here implies that Aizer and Doyle's results cannot be providing any sort of *general* estimate of the effect of juvenile incarceration on future offending or high school graduation.

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 (the "treatment") for those arrestees who they think will be less prone to commit future crimes. Hence, one would suspect there is negative selection-on-latent-outcomes in this context. Given this, and the fact that Agan, Doleac, and Harvey's IV estimates are negative in sign, the *Coefficient Difference Check* would consist of examining whether the OLS results are less than (i.e., more negative) than the IV results. Looking at Table 2 (row 19), we see that their OLS results are higher (i.e., less negative), than the IV results, meaning their results fail to be consistent with this check.

In their discussion of these results, Agan, Doleac, and Harvey (2023) essentially acknowledge this issue, arguing 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,

[12] Aizer and Doyle directly acknowledge that their results only apply to the *compliers* given their instrument and context, though they don't discuss what might account for why their results implicitly fail the *Coefficient Difference Check*, nor what the implications of this are. This is not meant as criticism. Rather, the point of this paper is to clarify this issue going forward.

while counterintuitive, "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 Difference Check* is because the "expected" selection-on-latent-outcomes story is wrong—there is actually *positive* selection-on-latent-outcomes rather than *negative* in their context. While such positive selection-on-latent-outcomes is certainly possible, for it to be true, it means that there must be sizeable enough positive selection with respect to unobservable characteristics to overcome the seemingly negative selection occurring with respect to observable characteristics, as the treated, (i.e., those not prosecuted) on average had fewer previous convictions and "serious" counts against them, both of which tend to be correlated with lower recidivism.[13]

Another interesting paper to consider is Dahl, Kostol, and Mogstad (2014). They study whether a parent's receipt of disability impacts the likelihood a child receives disability as an adult, using an adjudicator propensity-to-treat IV methodology. As discussed in a footnote above, this paper was not included in Table II, as it is a point of contention in this literature which way selection-on-latent-outcomes might go.[14] However, these authors still consider what may be driving the difference between their OLS and IV estimates. Their analysis suggests negative selection-on-latent-outcomes. Given negative selection-on-latent-outcomes and the fact that their OLS results are close to zero and the IV results are significantly positive, their findings are consistent with the *Coefficient Difference Check*. Even so, the authors also do a further analysis suggesting negative selection-on-magnitude of treatment effects. Performing the types of analyses developed in this paper would be especially useful when the selection-on-latent-outcomes story is seemingly more clear-cut, yet the IV results are not consistent with the *Coefficient Difference Check*.

Failures of the *Coefficient Difference Check* More Broadly

[13] The authors do note, however, that the treated group does tend to younger (i.e., < age 23), which would generally be associated with more future criminality. However, their results show a pretty modest difference here, with the fraction < age 23 being 0.23 among the treated compared to 0.22 among the untreated.

[14] This is especially true here since the examined population is only children whose parents appealed a denial.

Thus far I've focused on papers using the adjudicator propensity-to-treat IV methodology, but failures of the *Coefficient Difference Check* aren't just an issue with this class of IV papers. Indeed, 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 (Card 1995), and actually includes in his paper 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 almost all of these IV estimates fail to be consistent with the *Coefficient Difference Check*.

Obviously, Card does not describe what he is talking about as failures of the *Coefficient Difference Check*, but he does acknowledge the issue at hand, referring to it is as a "puzzle." He goes on to highlight a few possible explanations for this puzzle, but primarily focuses on heterogenous treatment effects. In the context of the argument made in his paper, Card's argument suggests that even though there is likely positive selection-on-latent-outcomes with respect to who goes to college, there must be notable negative selection-on-magnitude of treatment effects, wherein those with the highest return-to-college are actually substantially less likely to go to college. Card suggests the reason for such negative selection-on-magnitude of treatment effects is notable credit constraints among those individuals with the highest returns to college.

However, when Carneiro and Heckman (2002) consider this issue, they argue that that credit constraints at the time of the college going decision are not the reason why the IV coefficient on college attendance often exceeds the OLS coefficient. Rather, they argue that selection-on-latent-outcomes could actually be negative in that, in the absence of college, those who go to college would have lower wages than non-college goers, as college goers might be less skilled than non-college goers in the types of jobs available to non-college grads. Moreover, they provide evidence suggesting that many of the instruments used in this literature are likely invalid due to the exclusion restriction not holding.

5 - Other IV Diagnostics

This is certainly not the first paper to propose a diagnostic check on IV results. Below I discuss other methods proposed for highlighting potentially problematic IVs,

though all are more complicated to implement than the *Coefficient Difference Check*. While a failure in the tests below can indicate when proposed IV violates the necessary assumptions for a valid IV, not failing these tests does not guarantee that all the assumptions hold. However, they do provide another way of assessing the credibility of IV results, which might be especially useful for results that fail the *Coefficient Difference 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-on-latent-outcomes 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 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 Difference 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, Lefgren and Leslie (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-on-latent-outcomes 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 are arguably more suitable for identifying treatment effects due to smaller standard errors.

6 - Conclusion

This paper derives a simple equation that explicitly characterizes what drives the difference between OLS and IV estimates of a treatment effect, and motivates what I refer to as the *Coefficient Difference Check*. Namely, this check amounts to assessing whether the difference between the OLS coefficient and the IV coefficient has the same sign as the expected selection-on-latent-outcomes effect.

Results that fail to be consistent with this check do not necessarily imply that the IV is invalid, but I argue that such failures do require attention. Given results inconsistent with the *Coefficient Difference Check*, this paper argues that a researcher who still believes his/her IV is valid must convince the reader either that the selection-on-latent-outcomes effect is actually the opposite of expected, or that the difference in average treatment

effects between the *always-treated* and the *compliers* is large and goes in the exact opposite direction of the selection-on-latent-outcomes effect. However, even in cases where such heterogenous treatment effects are a plausible, the researcher should acknowledge that this means the IV coefficient actually provides a more biased estimate of the ATT than the OLS, the IV estimate will likely will be quite sensitive to small changes in who is impacted by the instrument, and be clear about the particular form of selection-on-magnitude treatment effects that this story must imply.

The issue at the heart of this paper is by no means unknown. Indeed, Card (2001) considered this type of issue over twenty years ago. Yet, this issue still seems to often go unacknowledged, even in papers published in the most prestigious journals. 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 check and responses that should be included with IV results, researchers using IV methods will apply this simple check, and when results fail to be consistent with this check, 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:

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

and

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

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

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

and

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

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

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

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

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

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

where δ^C is the average treatment effect among the compliers, which is the equation stated in the text above.

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 $Z_i = z_1$, 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 $\delta^C \equiv \frac{\int_{u_1^*}^{u_0^*} \delta_i f(u) du}{\pi_C}$, 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

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

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

Furthermore, noting 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

$$\delta^{OLS} = \pi_{D=1}^{C} \delta^{C} + (1 - \pi_{D=1}^{C}) \delta^{A}$$
$$+ \left[\frac{\pi_Z \pi_C}{\pi_Z \pi_C + \pi_A} E[Y(0)|u_1^* < u \leq u_0^*] + \frac{\pi_A}{\pi_Z \pi_C + \pi_A} E[Y(0)|u_0^* < u] \right]$$
$$- \left[\frac{(1 - \pi_Z)\pi_C}{\pi_N + (1 - \pi_Z)\pi_C} E[Y(0)|u_1^* < u \leq u_0^*] + \frac{\pi_N}{\pi_N + (1 - \pi_Z)\pi_C} E[Y(0)|u \leq u_1^*] \right].$$

Now, referring to equation (4), we can see that the second and third lines in the equation above simply correspond to the selection-on-latent-outcomes-effect, or $E[Y(0)|D = 1] - E[Y(0)|D = 0]$, meaning

$$\delta^{OLS} = \pi_{D=1}^{C} \delta^{C} + (1 - \pi_{D=1}^{C}) \delta^{A}$$
$$+ E[Y(0)|D = 1] - E[Y(0)|D = 0].$$

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^{C} and the average treatment effects among the always-treated δ^{A}, plus the selection-on-latent-outcomes effect, which is the difference in expected outcomes between those actually treated and those actually untreated in the counterfactual world where no one is treated.

OLS vs IV Results As Estimates of the ATT Under a Failure of the Coefficient Difference Check Due to Heterogeneous Treatment Effects

First note that the ATT will equal a weighted average of the average treatment effect of the *compliers* and the *always-treated*, where the weights depend on the fraction of the treated who are compliers, or in the notation used in the paper,

$$ATT = \pi_{D=1}^{C} \delta^{C} + (1 - \pi_{D=1}^{C}) \delta^{A}.$$

From equation (8) and our definition of the selection-on-latent-outcomes effect, we know

$$\delta^{OLS} = \pi^{C}_{D=1}\delta^{C} + (1 - \pi^{C}_{D=1})\delta^{A} + \text{selection-on-latent-outcomes.}$$

From earlier we also know

$$\delta^{IV} = \delta^{C}.$$

Calculating the absolute difference between the equation for the ATT and δ^{OLS}, we get

$$\left|ATT - \delta^{OLS}\right| = |\text{selection-on-latent-outcomes}|. \qquad (A1)$$

The analogous difference for the IV coefficient will be

$$\left|ATT - \delta^{IV}\right| = \left|(1 - \pi^{C}_{D=1})\left(\delta^{A} - \delta^{C}\right)\right|. \qquad (A2)$$

A direct implication of a failure of the *Coefficient Difference Check* due to heterogenous treatment effects is that

$$\left|(1 - \pi^{C}_{D=1})\left(\delta^{A} - \delta^{C}\right)\right| - |\text{selection-on-latent-outcomes}| > 0.$$

Given equations (A1) and (A2), the above equation then implies that when IV results fail the *Coefficient Difference Check* due to heterogenous treatment effects, it must be the case that $|ATT - \delta^{IV}| > |ATT - \delta^{OLS}|$. This in turn means that, in expectation, δ^{IV} is a more biased estimate of the ATT than is δ^{OLS} when heterogenous treatment effects are the reason for the failure of the *Coefficient Difference Check*.

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.

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.

Bureau of Labor Statistics, U.S. Department of Labor. National Longitudinal Survey of Youth 1997 cohort, 1997-2024. Produced and Distributed by the Center for Human Resource Research (CHR), The Ohio State University: Columbus OH.

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.

Carneiro, Pedro and James J. Heckman. (2002). "The Evidence on Credit Constraints in Post-Secondary Schooling." The Economic Journal 112: 705-734.

Carneiro, Pedro, James J. Heckman, Edward Vytlacil. (2011). "Estimating Marginal Returns to Education." American Economic Review 101: 2754-2781.

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.

Cohen, Elinor. (2024). "Housing the Homeless: The Effect of Placing Single Adults Experiencing Homelessness in Housing Programs on Future Homelessness and Socioeconomic Outcomes." American Economic Journal: Applied Economics 162: 130-175.

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.

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.

Imbens, Guido and Joshua Angrist. (1994). “Identification and Estimation of Local Average Treatment Effects.” Econometrica 62: 467-475.

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.

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.

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.

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

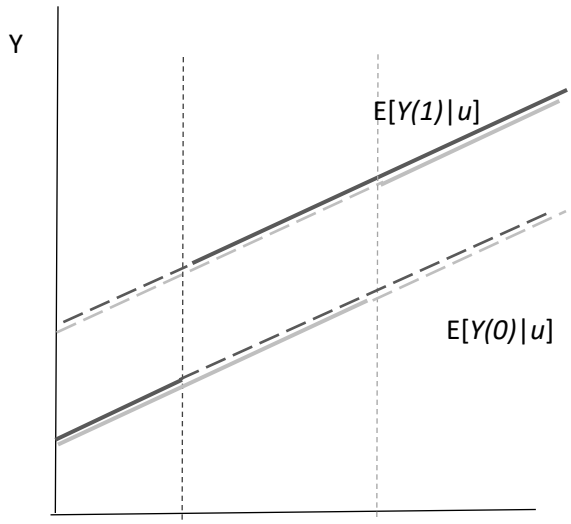


Fig 1a: Positive Treatment Effect (TE), Positive Selection-on-latent-outcomes, Uniform magnitude of TEs.

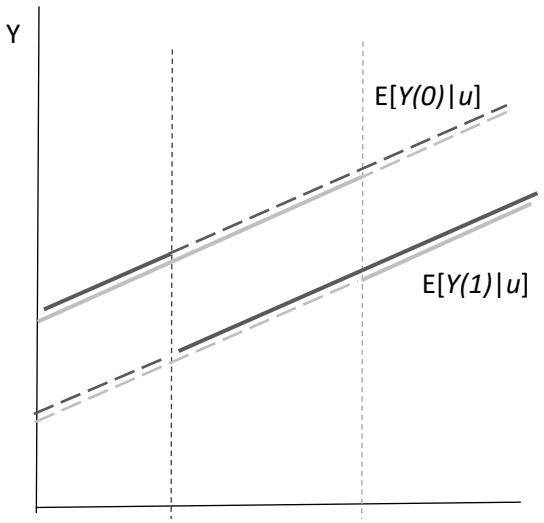


Fig 1b: Negative Treatment Effect (TE), Positive Selection-on-latent-outcomes, Uniform magnitude of TEs.

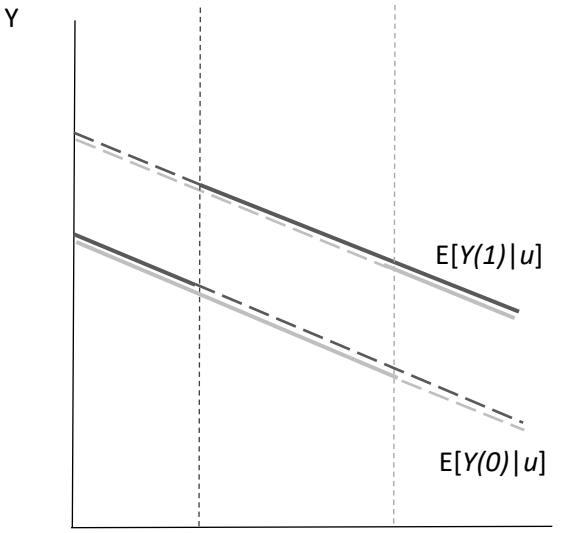


Fig 1c: Positive TEs, Negative Selection-on-latent-outcomes, Uniform magnitude

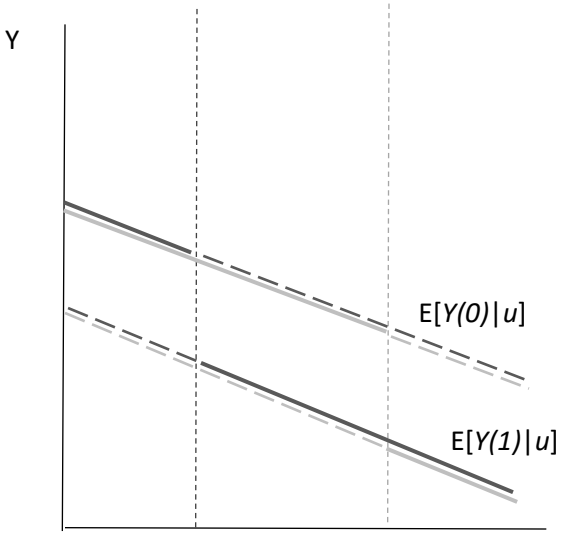


Fig 1d: Negative TEs, Negative Selection-on-latent-outcomes, Uniform magnitude

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. The dark lines and light lines overlapping is a byproduct of the exclusion restriction associated with a valid IV (i.e., the instrument has no direct relation to expected outcomes).

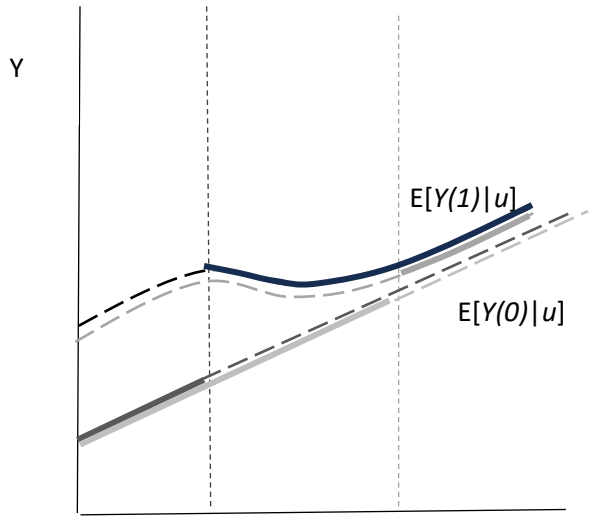


Fig 2a: Positive Treatment Effect (TE), Positive Selection-on-latent-outcomes, Negative Selection-on-Magnitude of TEs.

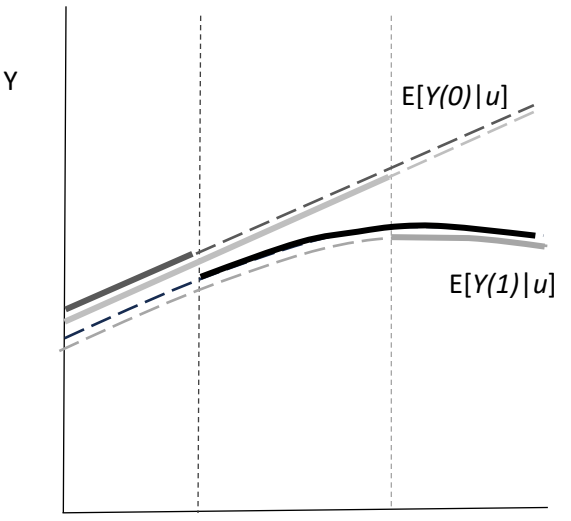


Fig 2b: Negative Treatment Effect (TE), Positive Selection-on-latent-outcomes, Positive Selection-on-Magnitude of TEs.

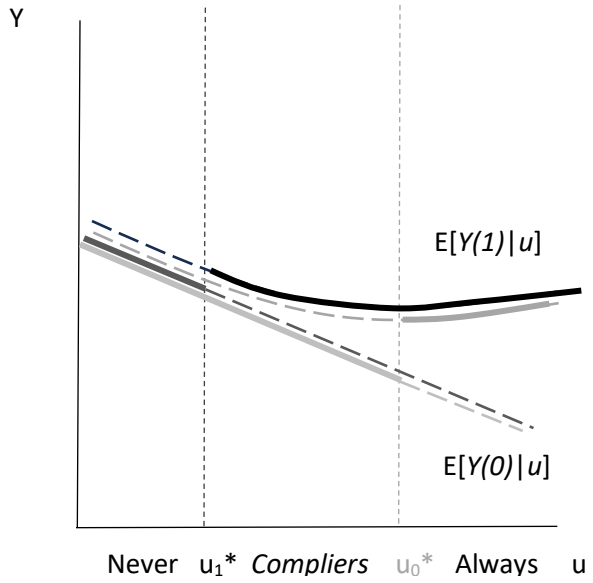


Fig 2c: Positive TE, Negative Selection-on-latent-outcomes, Positive Selection-on- Magnitude of TEs.

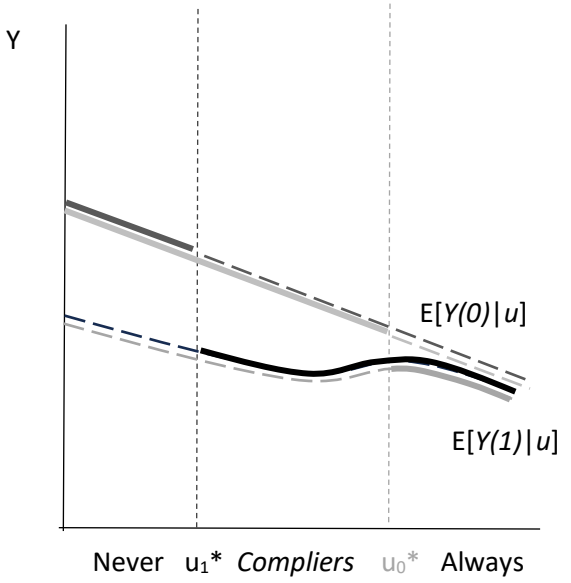


Fig 2d: Negative TE, Negative Selection-on-latent-outcomes, Negative Selection-on- Magnitude of TEs.

Figure 2 – Graphical depictions of the types of selection and treatment effects that could cause IV results to fail to be consistent with *Coefficient Difference Check*. 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. The dark lines and light lines overlapping is a byproduct of the exclusion restriction associated with a valid IV.

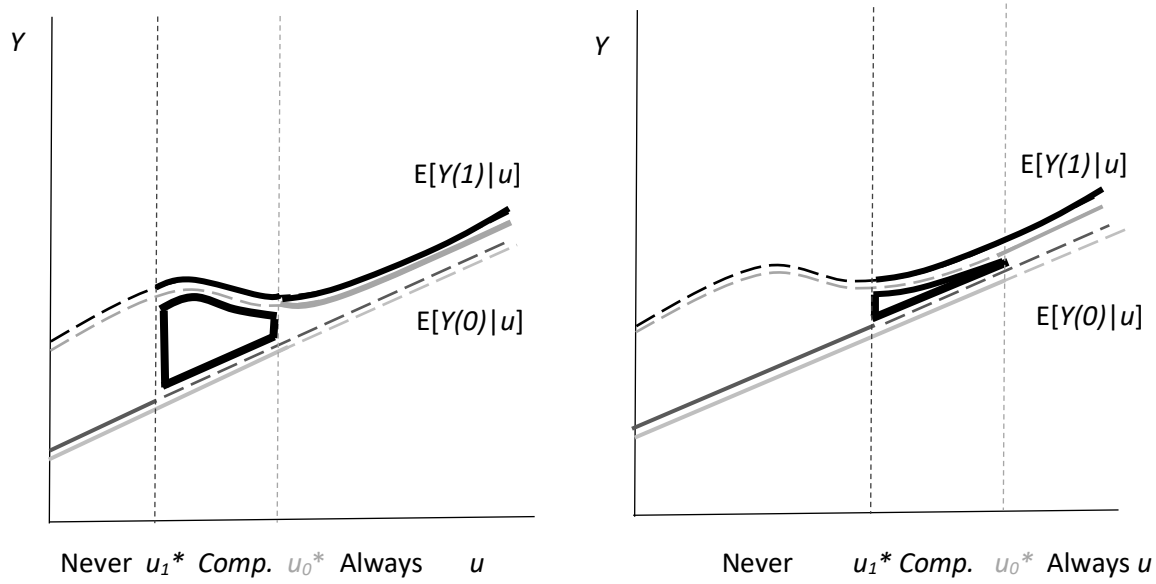


Figure 3 – Both graphs above depict the same underlying situation with Positive Selection-on-latent-outcomes along with Positive Treatment Effects, but where there is Negative Selection-on - Magnitude of these Treatment Effects. This is a situation that could cause a failure of *Coefficient Difference Check*. Lines depict the 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 outcomes, dashed portions indicate unobserved counterfactual outcomes. The dark lines and light lines overlapping is a byproduct of the exclusion restriction associated with a valid IV. Graphs reveal that δ^{IV} (the average treatment effect among compliers), which is weighted average of the height of the black polygons, will change substantially with shifts in group of *compliers* arising from alterations to instrument.

Table 1 - Applying *Coefficient Difference Check* to "Bad" IV Estimates

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

(1) Paper	(2) Treatment	(3) Outcome	(4) "Expected" OLS Bias	(5) OLS Coeff	(6) IV Coeff	(7) Coefficient Diff. Check
1 - Kling (2006)	Incarceration	Earnings	Negative	-44 (32)	248 (294)	-292 Pass
2a - Doyle (2007)	Foster Care	Juvenile Delinquency	Positive	0.00 (0.01)	0.35 (0.14)	-0.35 Fail
2b - Doyle (2007)	Foster Care	Teen Motherhood	Positive	0.09 (0.01)	0.29 (0.17)	-0.20 Fail
2c - Doyle (2007)	Foster Care	Earnings	Negative	-50 (30.6)	-1296 (626.0)	1246 Fail
3a - Doyle (2008)	Foster Care	Adult Arrest	Positive	0.060 (0.008)	0.391 (0.182)	-0.331 Fail
3b - Doyle (2008)	Foster Care	Adult Incarceration	Positive	0.031 (0.005)	0.225 (0.102)	-0.194 Fail
4 - Green and Winik (2010)	Incarceration	Recidivism	Positive	-0.006 (0.001)	0.009 (0.008)	-0.015 Fail
5 - Di Tella, Schargrodsky (2013)	Elec. Monitor (vs. Prison)	Recidivism	Negative	-0.09 t-stat rep	-0.13 t-stat rep	0.04 Fail
6a - Loeffler (2013)*	Incarceration	Recidivism	Positive	0.031 (0.008)	0.075 (0.104)	-0.044 Fail
6b - Loeffler (2013)*	Incarceration	Employment	Negative	-0.360 (0.006)	0.031 (0.083)	-0.391 Pass
7 - Dobbie, Song (2015)	Chapter 13 Bankruptcy	Home Foreclosure	Positive	Not shown	-0.19 (0.034)	Can't det.
8a - Aizer, Doyle (2015)	Juvenile Incarceration	Incarceration	Positive	0.155 (0.007)	0.234 (0.076)	-0.079 Fail
8b - Aizer, Doyle (2015)	Juvenile Incarceration	High School Grad	Negative	-0.073 (0.004)	-0.125 (0.043)	0.052 Fail
9 - Gupta, Hansman, Frenchman (2016)	Bail (vs Release)	Recidivism	Positive	Not Shown	0.07 (0.008)	Can't det.

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

(1) Paper	(2) Treatment	(3) Outcome	(4) "Expected" OLS Bias	(5) OLS Coeff	(6) IV Coeff	(7) Coefficient Diff. Check
10 - Leslie and Pope (2017)	Pre-trial Detention	Felony Recidivism	Positive	-0.100 (0.003)	-0.188 (0.020)	0.088 Pass
11 - Dobbie et al. (2017)	Chapter 13 Bankruptcy	Financial Strain Index	Positive	Not shown	-0.323 (0.071)	Can't det.
12a - Dobbie, Goldin, Yang (2018)	Pre-trial Release	Failure to Appear	Negative	0.01 (0.008)	0.156 (0.046)	-0.146 Pass
12b - Dobbie, Goldin, Yang (2018)	Pre-trial Release	Recidivism	Negative	-0.015 (0.006)	0.015 (0.063)	-0.03 Pass
13a - Harding et al. (2018)**	Incarceration	Employment 3 yrs Postrelease	Negative	0.02 (0.00)	-0.04 (0.03)	0.06 Fail
13b - Harding et al. (2018)**	Incarceration	Re-incarceration 3 yrs Postrelease	Positive	0.12 (0.00)	0.29 (0.02)	-0.17 Fail
14 - Bhuller, Dahl, Loken, Mogstad (2020)	Incarceration	Recidivism	Positive	0.052 (0.006)	-0.293 (0.106)	0.345 Pass
15a - Eren, Mocan (2021)	Juvenile Incarceration	Adult Convictions	Positive	0.119 (0.014)	0.013 (0.160)	0.106 Pass
15b - Eren, Mocan (2021)	Juvenile Incarceration	High School Graduation	Negative	-0.052 (0.016)	-0.002 (0.091)	-0.05 Pass
16 - Norris, Pecenco, Weaver (2021)	Parental Incarceration	Incarceration by Age 25	Positive	0.015 (0.004)	-0.049 (0.020)	0.064 Pass
17a - Gross, Baron (2022)	Foster Care	Child Well-Being	Negative	0.026 (0.110)	0.392 (0.164)	-0.366 Pass
17b - Gross, Baron (2022)	Foster Care	Subsequent Maltreatment	Positive	-0.032 (0.004)	-0.132 (0.058)	0.1 Pass
17c - Gross, Baron (2022)	Foster Care	Math Score	Negative	0.057 (0.013)	0.356 (0.203)	-0.299 Pass
17d - Gross, Baron (2022)	Foster Care	Juvenile Delinquency	Positive	0.041 (0.004)	-0.028 (0.040)	0.069 Pass

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

(1) Paper	(2) Treatment	(3) Outcome	(4) "Expected" OLS Bias	(5) OLS Coeff	(6) IV Coeff	(7) Coefficient Diff. Check
18a - Bald, Chyn, Hastings, Machelette (2022)	Foster Care	Test Scores (Girls)	Negative	Not Shown	1.367 (0.567)	Can't det.
18b - Bald, Chyn, Hastings, Machelette (2022)	Foster Care	Test Scores (Boys)	Negative	Not Shown	0.044 (0.562)	Can't det.
19 - Agan, Doleac, Harvey (2023)***	No Misd. Prosecution	Recidivism	Negative	-0.10 (0.010)	-0.29 (0.100)	0.19 Fail
20a - Collinson et. al (2024)	Eviction	Emergency Shelter 2-years post filing	Positive	0.014 (0.001)	-0.001 (0.013)	0.015 Pass
20b - Collinson et. al (2024)	Eviction	Homeless Serv. 2-years post filing	Positive	0.02 (0.001)	0.036 (0.036)	-0.017 Fail
20c - Collinson et. al (2024)	Eviction	Earnings 5-8 quarters post filing	Negative	-269 (13)	-613 (248)	344 Fail
20d - Collinson et. al (2024)	Eviction	Finacial Health 5-8 q. post filing	Negative	-0.103 (0.001)	-0.141 (0.036)	0.038 Fail
21a - Cohen (2024)****	Housing Assistance	Return to Homelessness	Negative	-0.088 (0.016)	-0.213 (0.060)	0.125 Fail
21b - Cohen (2024)****	Housing Assistance	Crime	Negative	-0.011 (0.006)	-0.066 (0.021)	0.055 Fail

Positive OLS bias means $\beta > 0$ in equation (1), negative OLS bias means $\beta < 0$ in equation (1). Reported Standard errors shown in parentheses. *Loeffler (2013) also uses a subsample of only judges with very high or low propensities to incarcerate. Under this sample, IV passes the Coefficient Difference Check.**Harding et al. (2018) use a slightly different IV methodology in their preferred specification than the standard judge-propensity for treatment IV. Under this preferred specification, their results still constitute a very marginal fail of the Coefficient Comparison Check. ***With respect to Agan, Doleac, Harvey (2023), see text for further discussion. ****I say negative expected bias here, as Cohen (2024) shows that those who receive housing assistance were less likely to have previously had emergency/crisis services and less likely to have had prior jail time than those who didn't receive housing assistance.