When Should We (Not) Interpret Linear IV Estimands as LATE?[*]

Tymon Słoczyński[†]

Abstract

In this paper I revisit the interpretation of the linear instrumental variables (IV) estimand as a weighted average of conditional local average treatment effects (LATEs). I focus on a situation in which additional covariates are required for identification while the reduced-form and first-stage regressions may be misspecified due to an implicit homogeneity restriction on the effects of the instrument. I show that the weights on some conditional LATEs are negative and the IV estimand is no longer interpretable as a causal effect under a weaker version of monotonicity, *i.e.* when there are compliers but no defiers at some covariate values and defiers but no compliers elsewhere. The problem of negative weights disappears in the interacted specification of Angrist and Imbens (1995), which avoids misspecification and seems to be underused in applied work. I illustrate my findings in an application to the causal effects of pretrial detention on case outcomes. In this setting, I reject the stronger version of monotonicity, demonstrate that the interacted instruments are sufficiently strong for consistent estimation using the jackknife methodology, and present several estimates that are economically and statistically different, depending on whether the interacted instruments are used.

Keywords: instrumental variables, local average treatment effects, model misspecification, monotonicity, negative weights, two-stage least squares

JEL classification: C21, C26, C52, K42

[*] First arXiv draft: November 13, 2020. This version: October 21, 2025. I am very grateful to the Editor, three anonymous referees, Martin Andresen, Isaiah Andrews, Carolina Caetano, Brant Callaway, Todd Elder, Brigham Frandsen, Xavier D'Haultfœuille, Molly Hawkins, Phillip Heiler, Stefan Hoderlein, Felix Holub, Martin Huber, Peter Hull, Shoya Ishimaru, Krzysztof Karbownik, Toru Kitagawa, Patrick Kline, Michael Knaus, Michal Kolesár, Giovanni Mellace, Anna Mikusheva, Áureo de Paula, Alex Poirier, Jonathan Roth, Pedro Sant'Anna, Vira Semenova, Jesse Shapiro, Gary Solon, Liyang Sun, Alex Torgovitsky, Edward Vytlacil, Chris Walters, Jeff Wooldridge, Kaspar Wüthrich, Yinchu Zhu, seminar participants at Binghamton University, Brandeis University, Duke University, Emory University, Goethe University Frankfurt, University of Chicago, and University of Warsaw, and participants at many conferences for helpful comments and discussions. I thank Megan Stevenson for sharing data and code, Tiemen Woutersen for sharing code, Anne Laski and Qihui Lei for excellent research assistance, as well as Danny Hellstein for assistance with the Brandeis HPC cluster. Qihui Lei has also coauthored the companion R and Stata packages `fejiv`. Finally, I acknowledge financial support from the Theodore and Jane Norman Fund.

[†] Brandeis University. Correspondence: Department of Economics, Brandeis University, MS 021, 415 South Street, Waltham, MA 02453. E-mail: tslocz@brandeis.edu.

1 Introduction

Many instrumental variables are only valid after conditioning on additional covariates. The draft eligibility instrument in Angrist (1990) requires controlling for the year of birth. The college proximity instrument in Card (1995) is invalid without conditioning on several individual characteristics of workers (Kitagawa, 2015). Even in the case of randomized experiments with noncompliance, it is often necessary to control for covariates correlated with treatment probability, such as household size and survey wave in Finkelstein *et al.* (2012).

When conditioning on additional covariates is necessary for instrument validity, interpreting the linear instrumental variables (IV) and two-stage least squares (2SLS) estimands becomes complicated. Angrist and Imbens (1995) (hereafter, AI) provide an influential interpretation of the 2SLS estimand in this context as a convex combination of conditional local average treatment effects (LATEs), *i.e.* average effects of treatment for individuals whose treatment status is affected by the instrument. However, this result is restricted to saturated models with discrete covariates and first-stage regressions that include a complete set of interactions between these covariates and the instrument. Such specifications are rare in empirical work, as is evident from several recent surveys of applications of IV methods.[1] This makes AI's result inappropriate for interpreting the vast majority of IV estimates encountered in economic applications (cf. Abadie, 2003).

In this paper I revisit the question of the causal interpretability of standard instrumental variables estimands. In particular, I focus on whether these estimands can be written as weighted averages of conditional LATEs with positive weights and, if so, whether these weights have an intuitive interpretation. To do so, I consider two variants of the usual monotonicity assumption: "weak monotonicity," which postulates that at every covariate value, the instrument either does not discourage or does not encourage anyone to take treatment, and "strong monotonicity," which additionally requires that the direction of this effect is uniform across covariate values.

My first contribution is to demonstrate that under weak monotonicity, the weights on some conditional LATEs may be negative in the usual application of IV, which restricts the first-stage effects of the instrument to be homogeneous. This finding implies that the resulting estimand is not a useful summary measure of average treatment effects; this parameter could be negative (positive) even if treatment effects are positive (negative) for everyone in the population. Under the same assumptions, all weights are necessarily positive in AI's interacted specification.

My second contribution is to explicitly compare the weights in both specifications with the "desired" weights, which recover the unconditional LATE parameter. Under strong monotonicity, when the weights in the usual application of IV and AI's specification are positive, both specifica-

[1] Blandhol, Bonney, Mogstad, and Torgovitsky (2022, 2025) consider a sample of 99 papers and find a single application of AI's specification. Mogstad, Torgovitsky, and Walters (2021) consider a sample of 122 papers and identify seven with specifications that include *some* covariate interactions with a single instrument.

tions overweight the effects in groups with large variances of the instrument, while the latter also overweights the effects in groups with strong first stages. It follows that the usual application of IV might be preferable when violations of strong monotonicity are not an issue.

However, if weak monotonicity is plausible but strong monotonicity is not, my theoretical results suggest that AI's interacted specification is preferable to the usual application of IV. Unfortunately, AI's specification is also difficult to estimate without bias; when the researcher divides the sample into many groups and subsequently creates an interacted instrument for each, 2SLS will be subject to the "many instrument" bias (see, *e.g.*, Bekker, 1994). An alternative approach to estimating AI's specification, such as the fixed effect jackknife IV (FEJIV) estimator of Chao, Swanson, and Woutersen (2023), should be used instead. Another concern about specifications with many instruments is whether they are jointly strong enough to enable consistent estimation. In this context, I consider a recent pretest for weak identification developed by Mikusheva and Sun (2022). As an illustration, I perform an extensive simulation study. In these simulations, Mikusheva and Sun (2022)'s pretest does a great job differentiating between cases where the best estimators of AI's specification, such as FEJIV, perform well and cases where all estimators perform badly.

To corroborate the concern about violations of strong monotonicity, I also replicate a sample of 988 instrumental variables regressions from 25 papers published in journals of the American Economic Association between 2006 and 2015. Every specification in my sample is based on a linear first-stage regression that restricts the effects of the instrument to be homogeneous. If strong monotonicity is violated but weak monotonicity is not, the homogeneous first stage will be misspecified and the conditional first stage will be positive for some covariate values but negative for others. First, I present strong suggestive evidence of the latter phenomenon, which directly translates to the incidence of negative weights in the usual application of IV. Then, I formally reject the null hypothesis of first-stage homogeneity in more than 70% of specifications in an average paper, despite accounting for multiple hypothesis testing.

In addition, I illustrate my findings in an application to the causal effects of pretrial detention on case outcomes (Stevenson, 2018). Here, I consider several saturated specifications, which allows me to compare the estimates of AI's specification with the usual application of IV. I can also formally test whether the conditional first stage is positive for some covariate values and negative for others, and I conclusively reject the null hypothesis of sign homogeneity. The estimates based on AI's specification are smaller than in the usual application of IV, and the difference is often statistically significant. Mikusheva and Sun (2022)'s pretest rejects in every case I consider, which supports the notion that the estimates based on AI's specification are preferable.

Finally, I supplement this paper with companion R and Stata packages, `fejiv`, available at the Comprehensive R Archive Network (CRAN) and the Statistical Software Components (SSC)

Archive, respectively.[2] These packages, based on the MATLAB code of Chao *et al.* (2023), can be used to implement the FEJIV estimator in practice. A Stata package to implement Mikusheva and Sun (2022)'s pretest is also available from Sun (2023).

Two papers closely related to this are Kolesár (2013) and Blandhol *et al.* (2022, 2025). Like this paper, Kolesár (2013) studies the interpretation of 2SLS estimands under weak monotonicity while also considering the probability limits of several jackknife-type estimators, limited information maximum likelihood (LIML), and other alternatives to 2SLS. Kolesár (2013)'s main result on 2SLS is not particular to any specification but instead represents a generic two-step IV estimand as a weighted average of conditional LATEs. The resulting weights are positive, subject to an additional condition that needs to be verified on a case-by-case basis.[3] In contrast, this paper focuses specifically on the usual application of IV and AI's interacted specification. The benefit is that this allows me to considerably simplify the representation and obtain results that are more transparent and easy to interpret. This includes the novel result that in the usual application of IV, the weights on some conditional LATEs may be negative under weak monotonicity. In another contribution, released after this paper first circulated, Blandhol *et al.* (2022, 2025) focus on the consequences of misspecification of the model for the instrument propensity score that is implicit in IV and 2SLS estimation. In this paper I focus instead on violations of strong monotonicity and their implications.

The remainder of the paper is organized as follows. Section 2 introduces my framework. Section 3 provides my theoretical contributions, a review of the literature on many instruments, and a simulation study. Section 4 studies negative first stages and first-stage heterogeneity in a sample of recent applications of IV methods and illustrates my findings in an analysis of the causal effects of pretrial detention on case outcomes. Section 5 concludes. The appendix contains my proofs as well as additional simulation and estimation results.

2 Framework

In this section I formally define the objects of interest, *i.e.* the conditional and unconditional IV and 2SLS estimands. I reserve the term "2SLS" for the appropriate estimand in a model with interacted instruments; see equation (3) below. When a single instrument is used instead, I use the term "IV" or "linear IV"; see equation (2). In what follows, I also review identification in the LATE framework with covariates (cf. Abadie, 2003). Throughout the paper I assume that the appropriate moments exist whenever necessary.

[2] To download the R package from CRAN, type `install.packages("fejiv")` in the R/RStudio console. To download the Stata package from SSC, type `ssc install fejiv` in the Command window.

[3] This condition essentially requires that the first stage postulated by the researcher provides a sufficiently good approximation to the true first stage (cf. Heckman and Vytlacil, 2005).

2.1 Notation and Estimands

Suppose we are interested in the causal effects of a treatment, $D \in \{0, 1\}$, on an outcome, $Y = Y(D)$, where $Y(1)$ and $Y(0)$ are potential outcomes. An instrument, $Z \in \{0, 1\}$, is also available, and it determines which of the potential treatment states, $D(1)$ and $D(0)$, is observed, $D = D(Z)$. In principle, we could let $Y = Y(Z, D)$, but we will rule out direct effects of Z on Y below. Finally, let $X = (1, X_1, \ldots, X_J)$ denote a row vector of covariates. In some cases I will allow for the possibility that additional instruments have been created by interacting Z with all elements of X; then, $Z_{\mathrm{C}} = (Z, ZX_1, \ldots, ZX_J)$ will be used to denote the resulting row vector of instruments.

To provide motivation for what follows, let us consider the standard single-equation linear model for Y:

$$Y = D\beta + X\rho + \upsilon, \tag{1}$$

where X and the instrument(s) are assumed to be uncorrelated with the error term υ. Also, β is the coefficient of interest. In this paper I do not assume that equation (1) is correctly specified; in particular, I allow the effect of D on Y to be correlated with both observables and unobservables.

In practice, however, many researchers act as if this model is correctly specified and use linear IV or 2SLS for estimation. In what follows, I will focus on the interpretation of the probability limits of the IV and 2SLS estimators of β when equation (1) is possibly misspecified. With a single instrument, the probability limit of linear IV or, simply, the (linear) IV estimand is

$$\beta_{\mathrm{IV}} = \left[(\mathrm{E}\left[Q'W\right])^{-1} \mathrm{E}\left[Q'Y\right] \right]_1, \tag{2}$$

where $W = (D, X)$, $Q = (Z, X)$, and $[\cdot]_k$ denotes the kth element of the corresponding vector. Clearly, when a single instrument is available, equation (2) characterizes the target of estimation in most empirical studies, which I also call the "usual" or "standard" estimand. This specification corresponds to reduced-form and first-stage regressions that project Y and D on X and Z, excluding any interactions between X and Z. Hence, I also refer to this specification as "noninteracted."

On the other hand, if a vector of interacted instruments, Z_{C}, is used in 2SLS estimation of equation (1), the relevant probability limit or, simply, the 2SLS estimand is

$$\beta_{\mathrm{2SLS}} = \left[\left(\mathrm{E}\left[W'Q_{\mathrm{C}}\right] (\mathrm{E}\left[Q_{\mathrm{C}}'Q_{\mathrm{C}}\right])^{-1} \mathrm{E}\left[Q_{\mathrm{C}}'W\right] \right)^{-1} \mathrm{E}\left[W'Q_{\mathrm{C}}\right] (\mathrm{E}\left[Q_{\mathrm{C}}'Q_{\mathrm{C}}\right])^{-1} \mathrm{E}\left[Q_{\mathrm{C}}'Y\right] \right]_1, \tag{3}$$

where $Q_{\mathrm{C}} = (Z_{\mathrm{C}}, X)$. In this specification, the corresponding reduced-form and first-stage regressions project Y and D on X and Z_{C}, which implies that the effects of Z on Y and D are allowed to vary with X due to the interactions between X and Z. Thus, I also refer to this specification as "interacted" or "fully interacted."

Regardless of the implicit restrictions on the effects of the instrument, the true first stage can be written as

$$\mathrm{E}\left[D \mid X, Z\right] = \psi(X) + \omega(X) \cdot Z, \tag{4}$$

where

$$\omega(x) = \mathrm{E}\left[D \mid Z = 1, X = x\right] - \mathrm{E}\left[D \mid Z = 0, X = x\right] \tag{5}$$

is the conditional first-stage slope coefficient or, equivalently, the coefficient on Z in the regression of D on 1 and Z in the subpopulation with $X = x$. Similarly, the conditional IV (or Wald) estimand can be written as

$$\beta(x) = \frac{\mathrm{E}\left[Y \mid Z = 1, X = x\right] - \mathrm{E}\left[Y \mid Z = 0, X = x\right]}{\mathrm{E}\left[D \mid Z = 1, X = x\right] - \mathrm{E}\left[D \mid Z = 0, X = x\right]}. \tag{6}$$

This parameter is equivalent to the coefficient on D in the IV regression of Y on 1 and D in the subpopulation with $X = x$, with Z as the instrument for D.

2.2 Local Average Treatment Effects

In the LATE framework of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996), the population consists of four latent groups: always-takers, for whom $D(1) = D(0) = 1$; never-takers, for whom $D(1) = D(0) = 0$; compliers, for whom $D(1) = 1$ and $D(0) = 0$; and defiers, for whom $D(1) = 0$ and $D(0) = 1$. As demonstrated by Imbens and Angrist (1994), if, among other things, we rule out the existence of defiers and assume that X is orthogonal to Z, the unconditional IV estimand, $\beta_{\mathrm{IV}} = \frac{\mathrm{E}[Y|Z=1]-\mathrm{E}[Y|Z=0]}{\mathrm{E}[D|Z=1]-\mathrm{E}[D|Z=0]}$, recovers the average treatment effect for compliers, also referred to as the local average treatment effect (LATE).

Some of my results will allow for the existence of both compliers and defiers, and hence throughout this paper I instead follow Kolesár (2013) in defining the LATE as

$$\tau_{\mathrm{LATE}} = \mathrm{E}\left[Y(1) - Y(0) \mid D(1) \neq D(0)\right], \tag{7}$$

i.e. the average treatment effect for individuals whose treatment status is affected by the instrument. This group includes both compliers and defiers; it will be restricted to compliers whenever the existence of defiers is ruled out. It is useful to note that this unconditional LATE parameter can also be written as

$$\tau_{\mathrm{LATE}} = \frac{\mathrm{E}\left[\pi(X) \cdot \tau(X)\right]}{\mathrm{E}\left[\pi(X)\right]}, \tag{8}$$

where

$$\tau(x) = \mathrm{E}\left[Y(1) - Y(0) \mid D(1) \neq D(0), X = x\right] \tag{9}$$

is the conditional LATE and

$$\pi(x) = \mathrm{P}\left[D(1) \neq D(0) \mid X = x\right] \tag{10}$$

is the conditional proportion of compliers and defiers. The following assumption, together with additional assumptions below, will be used to identify $\tau(x)$ and $\pi(x)$, and thereby also τ_{LATE}. Recall that $Y = Y(D)$ when direct effects of Z on Y are ruled out and $Y = Y(Z, D)$ otherwise.

Assumption IV.

(i) (Conditional independence) $(Y(0,0), Y(0,1), Y(1,0), Y(1,1), D(0), D(1)) \perp Z \mid X$;

(ii) (Exclusion restriction) $\mathrm{P}[Y(1,d) = Y(0,d) \mid X] = 1$ for $d \in \{0,1\}$ a.s.;

(iii) (Relevance) $0 < \mathrm{P}[Z = 1 \mid X] < 1$ and $\mathrm{P}[D(1) = 1 \mid X] \neq \mathrm{P}[D(0) = 1 \mid X]$ a.s.

Assumption IV is standard but not sufficient to identify $\tau(x)$ and $\pi(x)$. It is also necessary to restrict the existence of defiers (Imbens and Angrist, 1994). The following assumption, due to Abadie (2003), rules out the existence of defiers at any value of covariates.

Assumption SM (Strong monotonicity)**.** $\mathrm{P}[D(1) \geq D(0) \mid X] = 1$ a.s.

In many applications, Assumption SM may be too restrictive (cf. de Chaisemartin, 2017; Dahl, Huber, and Mellace, 2023). A testable implication of Assumption SM is that $\omega(x)$, the conditional first-stage slope coefficient, is always non-negative. If this is formally rejected or otherwise implausible, an alternative assumption is necessary to obtain point identification. One possibility is to restrict treatment effect heterogeneity, as discussed by Heckman and Vytlacil (2005) and Mogstad and Torgovitsky (2018), in which case we will be able to identify the average treatment effect rather than the unconditional LATE parameter. Another possibility is to replace Assumption SM with a weaker assumption that postulates the existence of compliers but no defiers at some covariate values and the existence of defiers but no compliers elsewhere. While the relative appeal of these two assumptions is context dependent, I will focus on the latter in what follows.

Assumption WM (Weak monotonicity)**.** There exists a subset of the support of X such that $\mathrm{P}[D(1) \geq D(0) \mid X] = 1$ on it and $\mathrm{P}[D(1) \leq D(0) \mid X] = 1$ on its complement.

To understand the difference between Assumptions SM and WM, consider a recent paper by Deryugina *et al.* (2019), who estimate the health effects of air pollution using an instrument based on changes in local wind direction. Imagine a pollution source located to the east of a particular city. When the wind also blows from the east, the city will experience relatively high levels of pollution; the opposite is true when the wind blows from the west. Assumption SM would require that every city reacts to a specific wind direction (say, east) in the same way (say, high pollution). This, however, is known not to be true. Deryugina *et al.* (2019) explain, for example, that air pollution is relatively high in San Francisco when the wind blows from the southeast, while the same is true in Boston when the wind blows from the southwest. Indeed, Assumption WM would allow for the possibility that different locations react to a specific wind direction in different ways.[4]

Importantly, Assumption WM, together with Assumption IV, is sufficient to identify $\tau(x)$ and $\pi(x)$. Before stating the relevant lemma, it is useful to define an auxiliary function

$$c(x) = \text{sgn}\Big(\text{P}\left[D(1) \geq D(0) \mid X = x\right] - \text{P}\left[D(1) \leq D(0) \mid X = x\right]\Big), \tag{11}$$

where $\text{sgn}(\cdot)$ is the sign function. Clearly, $c(x)$ equals 1 if there are only compliers at $X = x$ and -1 if there are only defiers at $X = x$.

The following lemma summarizes identification of the conditional LATE parameter and the conditional proportion of individuals whose treatment status is affected by the instrument.

Lemma 2.1.

(i) *Under Assumptions IV and SM, $\tau(x) = \beta(x)$ and $\pi(x) = \omega(x)$.*

(ii) *Under Assumptions IV and WM, $\tau(x) = \beta(x)$ and $\pi(x) = |\omega(x)| = c(x) \cdot \omega(x)$.*

Lemma 2.1 consists of well-known results and straightforward extensions of these results, and as such it is stated without proof (cf. Angrist *et al.*, 1996; Angrist and Pischke, 2009). Note that strong monotonicity implies weak monotonicity, which means that every statement that is true under weak monotonicity is also true under strong monotonicity as a special case. I will follow this logic in the statement of the theoretical results below.

3 Negative Weights in Linear IV

3.1 Angrist and Imbens (1995), Revisited

Let us begin by revisiting AI's representation of the 2SLS estimand. Recall that AI study a special case of the model in equation (1) where all covariates are binary and represent membership in disjoint groups or strata. In this case, each of the original covariates needs to be discrete or discretized, which means that the population can be divided into K groups, where K denotes the number of possible combinations of values of these variables. (For example, with six binary variables, we have $K = 2^6 = 64$.) Let $G \in \{1, \ldots, K\}$ denote group membership and $G_k = 1[G = k]$ denote the resulting group indicators. AI consider a model where original covariates are replaced with these group indicators, $X = (1, G_1, \ldots, G_{K-1})$, while reduced-form and first-stage regressions include a

[4] If we knew the pollution-inducing wind direction for every location, as we do in the case of Boston and San Francisco, Assumption SM might remain plausible for an appropriately redefined instrument. However, if this direction needs to be estimated, as is likely the case in practice, Assumption WM will be more appropriate.

full set of interactions between X and Z; that is, $Z_{\mathrm{C}} = (Z, ZG_1, \ldots, ZG_{K-1})$. The following lemma restates AI's and Kolesár (2013)'s interpretation of the 2SLS estimand in this context.

Lemma 3.1 (Angrist and Imbens, 1995; Kolesár, 2013)**.** *Suppose that* $X = (1, G_1, \ldots, G_{K-1})$ *and* $Z_{\mathrm{C}} = (Z, ZG_1, \ldots, ZG_{K-1})$*. Suppose further that Assumptions IV and WM hold. Then*

$$\beta_{\mathrm{2SLS}} = \frac{\mathrm{E}\left[\sigma^2(X) \cdot \tau(X)\right]}{\mathrm{E}\left[\sigma^2(X)\right]},$$

where $\sigma^2(X) = \mathrm{Var}\left[\mathrm{E}\left[D \mid X, Z\right] \mid X\right] = \mathrm{E}\left[(\mathrm{E}\left[D \mid X, Z\right] - \mathrm{E}\left[D \mid X\right])^2 \mid X\right]$*.*

Lemma 3.1 establishes that the 2SLS estimand in AI's interacted specification is a convex combination of conditional LATEs, with weights equal to the conditional variance of the first stage. This result is due to AI and has usually been interpreted as requiring that the existence of defiers is completely ruled out (*e.g.*, Angrist and Pischke, 2009). Kolesár (2013) demonstrates that it also holds under weak monotonicity.

It may not be immediately obvious how the 2SLS weights in Lemma 3.1 differ from the "desired" weights in equation (8). The following result facilitates this comparison.

Theorem 3.2. *Suppose that* $X = (1, G_1, \ldots, G_{K-1})$ *and* $Z_{\mathrm{C}} = (Z, ZG_1, \ldots, ZG_{K-1})$*. Suppose further that Assumptions IV and WM hold. Then*

$$\beta_{\mathrm{2SLS}} = \frac{\mathrm{E}\left[[\pi(X)]^2 \cdot \mathrm{Var}\left[Z \mid X\right] \cdot \tau(X)\right]}{\mathrm{E}\left[[\pi(X)]^2 \cdot \mathrm{Var}\left[Z \mid X\right]\right]}.$$

Theorem 3.2 shows that the 2SLS estimand in AI's interacted specification is a convex combination of conditional LATEs, with weights equal to the product of the squared conditional proportion of compliers or defiers and the conditional variance of Z.[5] Since the "desired" weights in equation (8) consist only of the conditional proportion of compliers or defiers, AI's specification overweights the effects in groups with strong first stages and with large variances of Z. Importantly, this result does not require strong monotonicity; weak monotonicity is sufficient.

Remark 3.1. Although Lemma 3.1 and Theorem 3.2 show that AI's specification can avoid negative weights, practitioners rarely use multiple interacted instruments. In a survey of recent applications of IV methods, Blandhol *et al.* (2022, 2025) determine that only 1 out of 99 applicable papers has used AI's specification. Specifications with many interactions between the instrument(s) and covariates were more common in earlier work using IV methods (*e.g.*, Angrist, 1990; Angrist and Krueger, 1991) but have since become rare, likely out of concern for the many instrument bias.[6]

[5] See also Walters (2018) for a related remark that focuses on "descriptive" estimands and does not use the LATE framework for interpretation.

[6] Indeed, Bound, Jaeger, and Baker (1995) write that their results "indicate that *the common practice* of adding

3.2 Usual Application of IV

Remark 3.1 suggests that Theorem 3.2 cannot be used directly to interpret most empirical studies because modern applications of IV methods avoid using many interacted instruments. A similar point is made by Angrist and Pischke (2009, p. 178), who maintain, however, that an indirect argument in Abadie (2003) implies that "some kind of covariate-averaged LATE" is estimated in noninteracted specifications as well. In what follows, I show that Angrist and Pischke (2009)'s assertion would be *false* under weak monotonicity. The claim is true under strong monotonicity, which I will be able to demonstrate directly, deriving the exact form of "covariate-averaged LATE" that linear IV estimates. I also revisit Abadie (2003)'s indirect argument later on.

To save space, I combine two extensions of AI's analysis in what follows. On the one hand, I am interested in the interpretation of the IV estimand when we retain AI's restriction that the model for covariates is saturated but no longer use the interacted instruments. This analysis does not require any additional assumptions. On the other hand, I am also interested in the interpretation of the IV estimand in nonsaturated specifications. This analysis proceeds under the assumption that the instrument propensity score, defined as

$$e(X) = \mathrm{E}\left[Z \mid X\right], \tag{12}$$

is linear in X. This assumption is standard and has been used by Kolesár (2013), Lochner and Moretti (2015), Evdokimov and Kolesár (2019), and Ishimaru (2024), among others.

Assumption PS (Instrument propensity score)**.** $e(X) = X\alpha$.

Assumption PS holds automatically when Z is randomized, and also when all covariates are discrete and the model for covariates is saturated. (This is why the statement of the theoretical results below only invokes Assumption PS and does not separately mention saturated specifications.) Assumption PS may also provide a good approximation to $e(X)$ in other situations, especially when X includes powers and cross-products of the original covariates. This assumption is critical. Blandhol *et al.* (2022, 2025) determine that Assumption PS is necessary for the IV and 2SLS estimands to maintain their interpretation as a convex combination of conditional LATEs.

Let us first consider the case of weak monotonicity. The following result shows that the interpretation of the linear IV estimand is very unappealing in this context.

interaction terms as excluded instruments may exacerbate the problem" (emphasis mine). On the other hand, some recent applications of the wind instrument (Deryugina *et al.*, 2019; Bondy, Roth, and Sager, 2020) and the "judges design" (Aizer and Doyle, 2015; Mueller-Smith, 2015; Stevenson, 2018) interact the instrument with selected covariates, which is similar in spirit to AI's specification. However, quantitatively speaking, this is still very rare in practice: in a sample of 122 papers considered by Mogstad *et al.* (2021), only seven include specifications with some covariate interactions with a baseline instrument.

Theorem 3.3. *Suppose that Assumptions IV, WM, and PS hold. Then*

$$\beta_{\mathrm{IV}} = \frac{\mathrm{E}\left[c(X) \cdot \pi(X) \cdot \mathrm{Var}\left[Z \mid X\right] \cdot \tau(X)\right]}{\mathrm{E}\left[c(X) \cdot \pi(X) \cdot \mathrm{Var}\left[Z \mid X\right]\right]}.$$

Theorem 3.3 provides a new representation of the IV estimand in the standard specification, *i.e.* one that, perhaps incorrectly, restricts the effects of the instrument in the reduced-form and first-stage regressions to be homogeneous across covariate values. Unlike in AI's specification, the estimand in the standard specification is not necessarily a convex combination of conditional LATEs. This is because $c(x)$ takes the value -1 for every value of covariates where there exist defiers but no compliers, and hence the corresponding weights in Theorem 3.3 are negative as well. It follows that, when IV is applied in the usual way, the estimand may no longer be interpretable as a causal effect. It is even possible that this parameter may be negative (positive) when treatment effects are positive (negative) for everyone in the population.

The following result demonstrates that this problem disappears when we impose the strong version of monotonicity.

Corollary 3.4. *Suppose that Assumptions IV, SM, and PS hold. Then*

$$\beta_{\mathrm{IV}} = \frac{\mathrm{E}\left[\pi(X) \cdot \mathrm{Var}\left[Z \mid X\right] \cdot \tau(X)\right]}{\mathrm{E}\left[\pi(X) \cdot \mathrm{Var}\left[Z \mid X\right]\right]}.$$

Corollary 3.4 provides a direct argument for Angrist and Pischke (2009)'s assertion that the standard specification of IV recovers a convex combination of conditional LATEs. As noted previously, however, this statement is no longer true under weak monotonicity. If strong monotonicity holds, then the weights in Corollary 3.4 may be more desirable than those in AI's specification. Indeed, a comparison of Corollary 3.4 and equation (8) shows that the standard specification, like AI's specification, overweights the effects in groups with large variances of Z but not, unlike the latter, in groups with strong first stages.[7]

Remark 3.2. Abadie (2003) shows that, under Assumptions IV, SM, and PS, the IV estimand is equivalent to the coefficient on D in the linear projection of Y on D and X among compliers. In other words, IV is analogous to ordinary least squares (OLS), with the exception of its ability to implicitly condition the analysis on the (latent) subpopulation of compliers. Corollary 3.4 provides another argument that "IV is like OLS." Indeed, as shown by Angrist (1998), the only difference between the OLS estimand and the ATE is in the dependence of the OLS weights on $\mathrm{Var}\left[D \mid X\right]$. Similarly, Corollary 3.4 shows that, under strong monotonicity, the only difference between the

[7] To be clear, both specifications attach a greater weight to conditional LATEs in groups with strong first stages, as required by equation (8). But AI's specification places even more weight on such conditional LATEs than is necessary to recover the unconditional LATE parameter.

IV estimand and the LATE is in the dependence of the IV weights on $\mathrm{Var}[Z \mid X]$. However, this analogy between OLS and IV may be problematic for IV given the undesirable properties of the OLS estimand under treatment effect heterogeneity (cf. Słoczyński, 2022).

Remark 3.3. Bond, White, and Walker (2007) discuss the interpretation of interacted and noninteracted specifications in randomized experiments with noncompliance in which the existence of defiers is completely ruled out. In this case, the standard specification of IV recovers the unconditional LATE parameter but the interacted specification does not.[8] This is a special case of the difference between Theorem 3.2 and Corollary 3.4 where $\mathrm{Var}[Z \mid X]$ is constant. However, Theorem 3.3 makes it clear that under weak monotonicity the standard specification no longer recovers the unconditional LATE parameter or even a convex combination of conditional LATEs.

Remark 3.4. Theorem 3.3 and Corollary 3.4 are also related to Theorem 1 in Kolesár (2013), which provides a common representation of any two-step instrumental variables estimand in the case of a binary D, a discrete Z, and under conditions similar to Assumptions IV, WM, and PS. To present this result, it is necessary to introduce some additional notation. Let $P = \mathrm{E}[D \mid Z, X]$, $P^L = \mathrm{L}[D \mid Z_{\mathrm{G}}, X]$, and $\tilde{P}^L = P^L - \mathrm{L}[D \mid X]$, where $\mathrm{L}[\cdot]$ is the linear projection and $Z_{\mathrm{G}} = z_{\mathrm{G}}(X, Z)$ is the vector of constructed instruments, which may include (some) interactions between X and Z. Also, let $\mathcal{P}_x$ denote the support of P conditional on $X = x$ and J_x denote the number of support points, with $\mathcal{P}_x = \{p_{1,x} < \ldots < p_{J_x,x}\}$. Then, Kolesár (2013) shows that

$$\beta_{\mathrm{TSIV}} = \int \sum_{j=1}^{J_x - 1} \frac{\theta_j(x)}{\int \sum_{j=1}^{J_x-1} \theta_j(x)\,\mathrm{d}F^X(x)} \tau(p_{j,x}; x)\,\mathrm{d}F^X(x), \tag{13}$$

where β_{TSIV} is any two-step instrumental variables estimand (*e.g.*, 2SLS) which uses Z_{G} as instruments, $\theta_j(x) = \left(p_{j+1,x} - p_{j,x}\right) \cdot \mathrm{P}\left[P > p_{j,x} \mid X = x\right] \cdot \mathrm{E}\left[\tilde{P}^L \mid X = x, P > p_{j,x}\right]$, and $\tau(p_{j,x}; x) = \frac{\mathrm{E}\left[Y \mid P = p_{j+1,x}, X = x\right] - \mathrm{E}\left[Y \mid P = p_{j,x}, X = x\right]}{p_{j+1,x} - p_{j,x}}$ is the conditional LATE based on two adjacent elements of $\mathcal{P}_x$. Kolesár (2013)'s result is generic in the sense that it applies to any given vector of instruments $Z_{\mathrm{G}} = z_{\mathrm{G}}(X, Z)$. At the same time, Theorem 3.3 is specific to the IV estimand. However, its focus on that particular specification simplifies the result, making it more transparent and easier to interpret than equation (13).[9] In Appendix A, I also present an alternative proof of Theorem 3.3, which uses Kolesár (2013)'s representation of β_{TSIV}.

[8] Instead, the interacted specification recovers a convex combination of conditional LATEs, which is generally different from the unconditional LATE parameter. A similar point about models with fully independent instruments is made by Huntington-Klein (2020), who also revisits the link between the existence of defiers and negative weights in this context (cf. Imbens and Angrist, 1994; de Chaisemartin, 2017; Dahl *et al.*, 2023) and recommends interacted specifications.

[9] Using equation (13) to determine whether a given specification rules out the incidence of negative weights requires verifying the condition $\mathrm{P}\left[\theta_j(X) \geq 0\right] = 1$ on a case-by-case basis.

Remark 3.5. A testable implication of strong monotonicity is that $\omega(x)$, the conditional first-stage slope coefficient, is always non-negative. In a saturated specification with $X = (1, G_1, \ldots, G_{K-1})$, it is straightforward to construct a formal test based on this observation.[10] If we define

$$\omega = \Big(\mathrm{E}\left[D \mid Z = 1, G = k\right] - \mathrm{E}\left[D \mid Z = 0, G = k\right]\Big)_{k=1}^{K}, \tag{14}$$

then the null hypothesis can be written as

$$H_0: \quad (-1) \cdot \omega \leq 0 \tag{15}$$

and the test statistic as

$$T = \max_{1 \leq k \leq K} \frac{(-1) \cdot \hat{\omega}_k}{\hat{\sigma}_{\hat{\omega}_k}}. \tag{16}$$

One possible choice of critical values for this test statistic are the one-step self-normalized critical values of Chernozhukov, Chetverikov, and Kato (2019). Another is based on the Bonferroni procedure, which requires, however, that K is much smaller than the sample size.

Remark 3.6. Suppose we are interested in the estimand of Corollary 3.4, but we are only willing to assume weak monotonicity. If $\omega(x)$ were known, we could define a new, "reordered" instrument as $Z_{\mathrm{R}} = 1[\omega(X) > 0] \cdot Z + 1[\omega(X) < 0] \cdot (1 - Z)$ and subsequently use it in a noninteracted specification. In Appendix A, I show that this procedure would recover the estimand of interest. In practice, however, $\omega(x)$ is unknown and would need to be estimated. I leave the study of the properties of the resulting reordered IV estimator to future work.

3.3 Finite Sample Considerations

Given the theoretical results in Sections 3.1 and 3.2, it seems reasonable to consider AI's interacted specification whenever weak monotonicity is plausible but strong monotonicity is not. However, this approach has some limitations in finite samples: it requires dividing the sample into K groups, and when K is sufficiently large relative to the sample size, some groups will be small. With many groups and instruments, this situation leads to bias, which results from overfitting the first stage. In other words, the first-stage fitted values pick up the noise, not just the signal, and a large amount of noise, particularly likely with many small groups, translates to poor estimates of the first stage and bias in the second stage.

This phenomenon, known as the "many instrument" bias, has been extensively studied in the econometrics literature. Recent surveys include Anatolyev (2019) and Mikusheva and Sun

[10] See also Semenova (2025) for an analogous test in the context of endogenous sample selection.

(2024).[11] In the remainder of this section, I first review several solutions to this problem, which offer finite sample improvements over 2SLS when estimating specifications with many instruments (*e.g.*, AI's specification). Then, I review a recent pretest designed to evaluate whether, in a given dataset, the instruments are jointly strong enough to ensure consistency. I conclude with a simulation study.

3.3.1 Estimation with Many Instruments

The problem of the many instrument bias is usually studied using the asymptotic sequence of Kunitomo (1980), Morimune (1983), and Bekker (1994), which allows the number of instruments, K, to increase in proportion with the sample size, N. In the context of AI's interacted specification, fixing the ratio of K to N does not allow the group sizes to grow when the sample size grows, which reproduces the practical problem of small groups.

Under this asymptotic sequence, 2SLS is inconsistent unless the concentration parameter, a measure of instrument strength, grows faster than the number of instruments. The classic alternatives include the limited information maximum likelihood (LIML) estimator of Anderson and Rubin (1949) and the bias-corrected two-stage least squares (B2SLS) estimator of Nagar (1959), both of which are consistent under homoskedasticity when the concentration parameter grows faster than the square root of the number of instruments (Chao and Swanson, 2005). However, homoskedasticity of first-stage errors is impossible when the treatment is binary. Under heteroskedasticity, LIML and B2SLS require the same (stronger) condition as 2SLS (Chao *et al.*, 2012).

Under heteroskedasticity, the weaker condition that the concentration parameter grows faster than the square root of the number of instruments is sufficient for the consistency of the jackknife IV estimator (JIVE) of Angrist, Imbens, and Krueger (1999), as also shown by Chao *et al.* (2012). The basic idea underlying jackknife-type estimators is that using a "leave-one-out" predictor of the treatment—effectively a separate first stage for each unit—will reduce the noise and bias.

At the same time, however, most of the estimators discussed so far are inconsistent under the asymptotic sequence that allows the number of covariates, alongside the number of instruments, to increase in proportion with the sample size. This is potentially a major limitation because, in AI's specification, the number of covariates and the number of instruments are the same and equal to the number of groups. Still, several modifications to JIVE and B2SLS are robust to many instruments and many covariates, including the improved jackknife IV estimator (IJIVE) of Ackerberg and Devereux (2009), the modified bias-corrected two-stage least squares (MB2SLS) estimator of Anatolyev (2013), the unbiased jackknife IV estimator (UJIVE) of Kolesár (2013), and three jackknife-type estimators of Chao *et al.* (2023), referred to as the fixed effect jackknife

[11] The classic literature on many instruments has focused on the homogeneous effects model, but I interpret its results through the lens of the framework in Section 2.

IV (FEJIV) estimator, the fixed effect limited information maximum likelihood (FELIM) estimator, and the fixed effect Fuller (1977) (FEFUL) estimator. Although the performance of LIML is not additionally affected by many covariates (Anatolyev, 2013), both LIML and MB2SLS rely on the homoskedasticity assumption. Furthermore, LIML does not even share the estimand with two-step IV estimators, such as 2SLS, MB2SLS, JIVE, IJIVE, UJIVE, and FEJIV, making it inappropriate in settings with treatment effect heterogeneity (Kolesár, 2013). FELIM and FEFUL do not belong to the class of two-step IV estimators either. Finally, Chao *et al.* (2023) discuss the limitations of IJIVE and, to a lesser extent, UJIVE, making FEJIV the likely estimator of choice.

While the framework of Chao *et al.* (2023) does not explicitly allow for treatment effect heterogeneity, the suitability of the FEJIV estimator in my framework follows from Kolesár (2013), who shows that any member of a broad class of two-step IV estimators has a common weighted average representation under treatment effect heterogeneity (cf. Remark 3.4). Because both 2SLS and FEJIV fall into this class, their estimands have the same interpretation under treatment effect heterogeneity and standard asymptotics. The difference is that under many instrument asymptotics, 2SLS becomes inconsistent for this estimand, whereas FEJIV remains consistent.

3.3.2 Weak Identification

Specifications with many instruments require that they are sufficiently strong as a group, although they can be individually weak or even irrelevant (cf. Anatolyev, 2019). In the context of AI's specification, the original instrument can be weak in some groups as long as it is sufficiently strong in others. But how strong is strong enough?

Mikusheva and Sun (2022) study weak identification in linear models with many instruments, which is a situation where the concentration parameter divided by the square root of the number of instruments remains bounded as the sample size grows. They also develop a pretest for this phenomenon to evaluate whether identification is strong in a given dataset. (Their test statistic $\widetilde{F}$ should be compared to a cutoff of 4.14.) Under the null of weak identification, no consistent estimator exists, and inference can instead be based on a jackknifed version of the AR test statistic. When the pretest rejects, Mikusheva and Sun (2022) recommend the jackknife IV estimator, which is consistent under the alternative (Chao *et al.*, 2012).

3.3.3 Simulations

In what follows, I study the finite sample performance of several two-step IV estimators of AI's specification, with a focus on settings with many small groups, treatment effect heterogeneity, and violations of Assumption SM. I adapt the data-generating process from Blandhol *et al.* (2022), which was designed to mimic the college proximity study in Card (1995). The simulation design also originally assumed homogeneous treatment effects and no monotonicity violations. As we will see, these restrictions are responsible for Blandhol *et al.* (2022)'s conclusion that the usual application of IV is easier to estimate without bias than the interacted specification.

In the baseline data-generating process, as in Blandhol *et al.* (2022), I draw X uniformly from a Halton sequence $\mathcal{X}$ on $[0, 1]$, subsequently drawing Z, $D(Z)$, and $Y(D)$ as

$$\begin{aligned} \mathrm{P}[Z = 1 \mid X] &= 0.119 + 1.785X - 1.534X^2 + 0.597X^3, & (17) \\ D(Z) &= 1[\Phi(V) \leq p(Z)], & (18) \\ Y(D) &= \log\left(129.7 + 1247.7X - 2149X^2 + 1515.7X^3\right) + 1.2D + U, & (19) \end{aligned}$$

where (U, V) are standard multivariate normal with correlation 0.527, drawn independently of (X, Z). I also set $|\mathcal{X}| = 250$, $p(0) = \mathrm{P}[D = 1 \mid Z = 0] = 0.22$, and $p(1) = \mathrm{P}[D = 1 \mid Z = 1] = 0.29$. In this setting, treatment effects are homogeneous and equal to 1.2. Strong monotonicity is satisfied even though the instrument is relatively weak, with the proportion of compliers independent of X and equal to $p(1) - p(0) = 0.07$. Again, these parameters are calibrated to the data in Card (1995).

In subsequent modifications of this data-generating process, I introduce treatment effect heterogeneity by specifying $Y(1)$ and $Y(0)$ as

$$\begin{aligned} Y(1) &= \log\left(129.7 + 1247.7X - 2149X^2 + 1515.7X^3\right) + 1.2 + U, & (20) \\ Y(0) &= \log\left(1 \cdot 129.7 + 2 \cdot 1247.7X - 3 \cdot 2149X^2 + 4 \cdot 1515.7X^3\right) + U, & (21) \end{aligned}$$

while also allowing for violations of strong (but not weak) monotonicity. This is accomplished by switching the values of $p(0)$ and $p(1)$ for some groups. Specifically, to generate what I refer to as "moderate" monotonicity violations, I reverse the values of $p(0)$ and $p(1)$ if $X > 0.75$. For "large" monotonicity violations, the threshold value of X is 0.5. I also consider a setting with "weak cells," that is, values of X where the proportion of compliers and defiers is zero. Here, I set $p(0) = p(1) = 0.22$ if $1/3 < X < 2/3$ and reverse the original values of $p(0)$ and $p(1)$ if $X > 2/3$.

Two final modifications involve the number and relative sizes of groups and the instrument strength. So far, the groups were equal sized. To reproduce the likely scenario that some groups are large while others are small, I also consider a setting with $|\mathcal{X}| = 20$, but where X is not drawn uniformly. Specifically, I set $\mathrm{P}[G = k]$ to be proportional to 1.3^k, making the largest group 1.3^{19} times larger than the smallest. As in Blandhol *et al.* (2022), I also consider a scenario where the instrument is stronger than the "weak" case above, with 0.52 replacing 0.29 as the larger value of $p(Z)$ whenever $p(0) \neq p(1)$ conditional on X. This sets the conditional proportion of compliers or defiers equal to 0.3, except in the "weak cells" design, where it is either 0.3 or 0.

The total number of simulation designs is sixteen, with $|\mathcal{X}| = 20$ or $|\mathcal{X}| = 250$, two levels of

Table 1: Simulation Results for $K = 250$, “Weak” IV, and No Monotonicity Violations

	$N = 3,000$			$N = 10,000$			$N = 50,000$		
A. Estimator Performance	Bias	Median Bias	MSE	Bias	Median Bias	MSE	Bias	Median Bias	MSE
OLS	–0.735	–0.735	1.080	–0.736	–0.735	1.303	–0.735	–0.735	3.457
IV	0.095	0.001	0.941	0.013	–0.002	0.225	0.001	–0.001	0.106
2SLS	–0.698	–0.701	1.000	–0.635	–0.636	1.000	–0.387	–0.387	1.000
MB2SLS	–59.10	–0.966	6.4e+06	4.546	–0.025	7.1e+04	0.053	0.036	0.313
JIVE	–0.803	–0.805	1.355	–0.901	–0.892	2.029	–9.473	–3.842	1.5e+05
IJIVE	0.385	–0.430	629.8	–0.068	–0.112	37.97	0.013	–0.004	0.256
UJIVE	–1.540	–0.482	516.7	1.073	–0.062	4.8e+03	0.017	0.000	0.261
FEJIV	–1.396	–0.512	3.5e+03	0.395	–0.027	132.0	0.017	0.000	0.261
B. Pretest for Weak Identification									
Average $\widetilde{F}$		1.83			2.27			10.37	
$q_{0.05}$		–0.07			0.12			7.54	
$q_{0.95}$		3.83			4.40			13.43	

Notes: “OLS” is the OLS estimator in the regression of the outcome on the treatment indicator and group indicators. “IV” is the IV estimator in the noninteracted specification. The remaining estimators are based on the interacted specification. JIVE, IJIVE, and UJIVE are computed after dropping all groups with fewer than two observations in either (X, Z) combination. FEJIV is computed after dropping all groups with fewer than three observations in either (X, Z) combination. The pretest for weak identification follows Mikusheva and Sun (2022); see also the Stata implementation in Sun (2023). Bias and median bias are reported as the proportion of the target parameter. MSE is normalized by the MSE of 2SLS. Results are based on 1,000 replications. Pretest results are based on 250 replications.

instrument strength (“weak” or “strong”), and four scenarios of violations of strong monotonicity, referred to as no violations, moderate violations, large violations, and violations with weak cells. Treatment effects are homogeneous when strong monotonicity holds and heterogeneous otherwise. The target parameter is the estimand in Theorem 3.2, which is, except in the “weak cells” design, equal to that in Corollary 3.4, making monotonicity violations the only reason why the estimands of the interacted and noninteracted specifications may be different. I consider two sample sizes, $N = 3,000$ and $N = 10,000$, when $|\mathcal{X}| = 20$, and additionally $N = 50,000$ when $|\mathcal{X}| = 250$. The smallest sample size, $N = 3,000$, is similar to the sample size in Card (1995).

Table 1 reports simulation results for a number of estimators in the “weak” IV case with 250 groups and no monotonicity violations. The first three columns, setting $N = 3,000$, correspond to the baseline results in Blandhol *et al.* (2022). Even though I consider a larger number of estimators than Blandhol *et al.* (2022), I reach the same conclusion: all estimators are severely biased, with the only exception of IV in the noninteracted specification, whose bias is less than 10% and median bias is practically zero. However, panel B of Table 1 reveals that this conclusion is predictable: the average value of Mikusheva and Sun (2022)’s test statistic, $\widetilde{F}$, is 1.83, well below the cutoff of 4.14, which means that consistent estimation of the interacted specification is impossible. The remaining columns report simulation results for $N = 10,000$ and $N = 50,000$. Here, the strength of identification gradually increases, with the average value of $\widetilde{F}$ exceeding 10 when $N = 50,000$.

Table 2: Simulation Results for $K = 250$, "Weak" IV, and Moderate Monotonicity Violations

	$N = 3{,}000$			$N = 10{,}000$			$N = 50{,}000$		
A. Estimator Performance	Bias	Median Bias	MSE	Bias	Median Bias	MSE	Bias	Median Bias	MSE
OLS	–1.159	–1.158	1.168	–1.160	–1.159	1.316	–1.160	–1.159	3.357
IV	0.543	0.181	3.418	0.219	0.176	0.440	0.180	0.175	0.239
2SLS	–1.059	–1.056	1.000	–0.999	–0.997	1.000	–0.621	–0.623	1.000
MB2SLS	–2.382	–1.589	452.1	0.174	–0.113	140.2	–0.019	–0.037	0.213
JIVE	–1.150	–1.152	1.215	–1.403	–1.391	1.997	–9.380	–5.865	6.9e+04
IJIVE	–2.173	–0.672	1.7e+03	4.456	–0.161	1.8e+04	0.020	0.001	0.229
UJIVE	–1.436	–0.793	2.1e+03	–0.735	–0.076	812.2	0.028	0.007	0.234
FEJIV	–0.991	–0.715	895.6	0.807	–0.072	261.1	0.028	0.007	0.234
B. Pretest for Weak Identification									
Average $\widetilde{F}$		1.92			2.30			10.08	
$q_{0.05}$		–0.01			0.27			7.48	
$q_{0.95}$		4.04			4.73			12.87	

Notes: "OLS" is the OLS estimator in the regression of the outcome on the treatment indicator and group indicators. "IV" is the IV estimator in the noninteracted specification. The remaining estimators are based on the interacted specification. JIVE, IJIVE, and UJIVE are computed after dropping all groups with fewer than two observations in either (X, Z) combination. FEJIV is computed after dropping all groups with fewer than three observations in either (X, Z) combination. The pretest for weak identification follows Mikusheva and Sun (2022); see also the Stata implementation in Sun (2023). Bias and median bias are reported as the proportion of the target parameter. MSE is normalized by the MSE of 2SLS. Results are based on 1,000 replications. Pretest results are based on 250 replications.

Indeed, when this is the case, the best-performing estimators of AI's specification—IJIVE, UJIVE, and FEJIV—are practically unbiased, in line with the results in Mikusheva and Sun (2022).

Table 2 introduces moderate monotonicity violations. With $N = 3{,}000$, the average value of $\widetilde{F}$ is again below 2. Now, however, every estimator is severely biased, including IV in the noninteracted specification. (Estimation of this specification is biased because of monotonicity violations. Estimation of the interacted specification is biased because of insufficient instrument strength.) With larger sample sizes, $N = 10{,}000$ and $N = 50{,}000$, identification gets stronger. Specifically, when $N = 50{,}000$, the average value of $\widetilde{F}$ again exceeds 10, and IJIVE, UJIVE, and FEJIV perform very well. IV estimation of the noninteracted specification remains biased; however, it is competitive with the best-performing estimators in terms of MSE.

Tables 3 and 4 consider large monotonicity violations and "weak cells." It remains the case that IJIVE, UJIVE, and FEJIV are nearly unbiased whenever the average value of $\widetilde{F}$ is large enough. This includes the "weak cells" design in Table 4, which underscores the notion that the instrument can be weak in some groups as long as it is sufficiently strong in others.[12] On the other hand, unlike in Table 2, IV estimation of the noninteracted specification is not only biased in Tables 3 and 4, but

[12] Intuitively, if $\pi(x) = 0$ when $X = x$, $\tau(x)$ is not identified. However, because $\tau_{\text{LATE}} = \frac{\text{E}[\pi(X)\cdot\tau(X)]}{\text{E}[\pi(X)]}$, the weight on $\tau(x)$ in τ_{LATE} would have been zero anyway, and analogously for the estimands in Theorem 3.2, Theorem 3.3, and Corollary 3.4. That is, as long as the overall instrument strength is sufficient (cf. Mikusheva and Sun, 2022), it does not matter that some conditional LATEs cannot be well estimated due to a conditional-on-X weak IV problem, because those conditional LATEs are irrelevant for the target estimand.

Table 3: Simulation Results for $K = 250$, “Weak” IV, and Large Monotonicity Violations

	$N = 3{,}000$			$N = 10{,}000$			$N = 50{,}000$		
A. Estimator Performance	Bias	Median Bias	MSE	Bias	Median Bias	MSE	Bias	Median Bias	MSE
OLS	–1.165	–1.166	1.187	–1.166	–1.166	1.329	–1.166	–1.166	3.347
IV	0.839	–0.252	723.7	1.498	0.459	252.2	0.670	0.547	3.322
2SLS	–1.056	–1.057	1.000	–1.001	–1.006	1.000	–0.627	–0.624	1.000
MB2SLS	0.073	–0.751	625.9	–0.052	–0.291	15.78	–0.070	–0.082	0.186
JIVE	–1.154	–1.151	1.239	–1.413	–1.399	2.025	–2.757	–5.762	1.6e+04
IJIVE	–1.341	–0.753	456.7	0.085	–0.212	158.8	0.018	0.004	0.216
UJIVE	–8.712	–0.842	7.0e+04	0.011	–0.113	99.13	0.026	0.010	0.221
FEJIV	–1.745	–0.686	6.3e+03	1.626	–0.057	721.5	0.026	0.010	0.221
B. Pretest for Weak Identification									
Average $\widetilde{F}$		2.01			2.24			9.74	
$q_{0.05}$		–0.06			0.37			7.15	
$q_{0.95}$		4.30			4.41			12.41	

Notes: “OLS” is the OLS estimator in the regression of the outcome on the treatment indicator and group indicators. “IV” is the IV estimator in the noninteracted specification. The remaining estimators are based on the interacted specification. JIVE, IJIVE, and UJIVE are computed after dropping all groups with fewer than two observations in either (X, Z) combination. FEJIV is computed after dropping all groups with fewer than three observations in either (X, Z) combination. The pretest for weak identification follows Mikusheva and Sun (2022); see also the Stata implementation in Sun (2023). Bias and median bias are reported as the proportion of the target parameter. MSE is normalized by the MSE of 2SLS. Results are based on 1,000 replications. Pretest results are based on 250 replications.

Table 4: Simulation Results for $K = 250$, “Weak” IV, and Monotonicity Violations with Weak Cells

	$N = 3{,}000$			$N = 10{,}000$			$N = 50{,}000$		
A. Estimator Performance	Bias	Median Bias	MSE	Bias	Median Bias	MSE	Bias	Median Bias	MSE
OLS	–1.196	–1.195	1.136	–1.193	–1.194	1.185	–1.194	–1.194	2.269
IV	0.945	–0.310	975.8	0.053	0.380	1.2e+03	0.805	0.646	3.480
2SLS	–1.108	–1.111	1.000	–1.082	–1.080	1.000	–0.779	–0.784	1.000
MB2SLS	–6.127	–1.278	1.2e+04	0.183	–0.479	246.3	–0.109	–0.132	0.308
JIVE	–1.129	–1.122	1.084	–1.315	–1.319	1.493	–3.229	–2.828	23.50
IJIVE	–11.17	–0.900	2.1e+05	9.651	–0.480	5.2e+04	0.026	–0.015	0.396
UJIVE	–0.814	–1.063	3.3e+03	3.604	–0.411	1.1e+04	0.039	–0.001	0.409
FEJIV	3.766	–0.778	2.3e+04	–6.777	–0.476	1.6e+04	0.039	0.001	0.408
B. Pretest for Weak Identification									
Average $\widetilde{F}$		1.73			1.48			6.18	
$q_{0.05}$		0.03			–0.58			4.10	
$q_{0.95}$		3.57			3.54			8.52	

Notes: “OLS” is the OLS estimator in the regression of the outcome on the treatment indicator and group indicators. “IV” is the IV estimator in the noninteracted specification. The remaining estimators are based on the interacted specification. JIVE, IJIVE, and UJIVE are computed after dropping all groups with fewer than two observations in either (X, Z) combination. FEJIV is computed after dropping all groups with fewer than three observations in either (X, Z) combination. The pretest for weak identification follows Mikusheva and Sun (2022); see also the Stata implementation in Sun (2023). Bias and median bias are reported as the proportion of the target parameter. MSE is normalized by the MSE of 2SLS. Results are based on 1,000 replications. Pretest results are based on 250 replications.

also noisy, which leads to very high values of MSE.

The remaining simulation results, for "weak" IV with $|\mathcal{X}| = 20$ and for "strong" IV with both values of $|\mathcal{X}|$, are reported in Tables B.1–B.12 in Appendix B. The bottom line is still that Mikusheva and Sun (2022)'s pretest does a great job differentiating between cases where IJIVE, UJIVE, and FEJIV perform well or very well, and cases where all estimators of the interacted specification perform badly. Roughly speaking, values of $\widetilde{F}$ exceeding the recommended cutoff of 4.14 are associated with low bias, even when, with $|\mathcal{X}| = 250$ and $N = 3{,}000$, there are only 12 units in each group; values of $\widetilde{F}$ exceeding 10–15 are associated with negligible or no bias, at least in the data-generating process under consideration.

Other estimators are clearly not competitive with IJIVE, UJIVE, and FEJIV. When there are violations of monotonicity, the usual application of IV is biased and often unstable. 2SLS estimation of the interacted specification is generally biased, as expected. MB2SLS is usually dominated by IJIVE, UJIVE, and FEJIV, especially on bias. JIVE is generally biased and unstable.

To be clear, the purpose of this simulation study is not to claim that AI's specification can be estimated without bias in a typical application of IV methods. Instead, the simulations show that IJIVE, UJIVE, and FEJIV estimation of AI's specification is reliable *if* the instruments are jointly strong enough, which can be verified using Mikusheva and Sun (2022)'s pretest. Future research should examine whether the number of instruments in AI's specification could be reduced using appropriate regularization techniques, perhaps a modification of the existing approaches in Chernozhukov, Hansen, and Spindler (2015a,b) and Wiemann (2024).

4 Empirical Applications

The results so far underscore the importance of using the interacted specification when weak monotonicity is plausible but strong monotonicity is not. In this section I present evidence of violations of strong monotonicity and first-stage homogeneity in a sample of recent applications of IV methods. Then, I revisit a study of the effects of pretrial detention on case outcomes in Philadelphia, where violations of strong monotonicity are particularly evident (Stevenson, 2018).

4.1 Review of Applications of Instrumental Variables

In what follows, I use a sample of 1,309 instrumental variables regressions previously analyzed by Young (2022), which corresponds to the universe of IV estimates reported in the main text of 30 papers published in journals of the American Economic Association between 2006 and 2015.[13]

[13] Young (2022)'s goal was to cover the universe of replicable IV applications in this period subject to a small number of additional inclusion criteria reported in his paper.

After dropping specifications with multiple instruments, without additional covariates, or based on panel data, I obtain my final sample of 988 regressions in 25 papers.[14] The number of regressions per paper is highly uneven in this sample, with the mean equal to $988/25 = 39.52$ and the quartiles equal to 8, 14, and 40. The list of papers under consideration is provided in Appendix C.

Given the inclusion criteria above, every specification in my sample is based on a linear first-stage regression of D on Z and X, without any interactions between Z and X. In my first exercise, I implicitly include these interactions by means of separate regressions of D on X given $Z = 1$ and $Z = 0$.[15] This simple approach allows me to estimate the conditional first stage at every value of X as the difference in conditional means, $\hat{\omega}(x) = \hat{\mathrm{E}}[D \mid Z = 1, X = x] - \hat{\mathrm{E}}[D \mid Z = 0, X = x]$. Subsequently, I report the fraction of these estimates that are opposite in sign ("negative") to the estimate in the original first stage, which is equivalent to the fraction of observations with negative weights in the usual application of IV (cf. Theorem 3.3). This is analogous to the recommendation of de Chaisemartin and D'Haultfœuille (2020) to report the fraction of units with negative weights in two-way fixed effects regressions. Similarly, Semenova (2025) reports the fraction of observations with negative predictions in a sample selection context related to mine.

Panel A of Table 5 indicates that negative first stages are a common occurrence in recent applications of IV. The average fraction of observations with a negative first stage is 21.8% when using the linear probability model (LPM) to estimate the conditional means in $\omega(x)$ and 17.6% when using the probit model. After weighting by the inverse of the number of applicable regressions associated with a given paper, these averages increase to 28.5% and 28.0%, respectively, giving the average of the within-paper averages.

It may be the case that a portion of the estimated negative first stages is due to noise. However, the regressions in my sample are usually not saturated, which means that the formal test of violations of monotonicity in Remark 3.5 is not appropriate. Instead, in my second exercise, I explicitly add interaction terms to each original (linear) first stage and test whether the corresponding coefficients are jointly equal to zero. Under the alternative, the true first stage is heterogeneous, which is a necessary condition for strong monotonicity being false but weak monotonicity being true.

Panel B of Table 5 reports the results of this exercise. Using the Bonferroni procedure to account for multiple hypothesis testing separately for each paper, I conclude that 22 of 25 papers have at least one first stage that is heterogeneous. Using the Holm correction, I reject an average of 71.5% of homogeneous first stages per paper. The last column demonstrates that these conclusions

[14] Because Young (2022) only considered papers with replication code in Stata, I define "specifications based on panel data" as those using Stata's `xtivreg` or `xtivreg2` commands in the original replication package. The number of applicable regressions in several papers would decrease substantially if we eliminated not only duplicate IV regressions—which Young (2022) already did—but also duplicate first stages. However, my preliminary attempt to do so did not meaningfully change any of the results reported in this section.

[15] If the original treatment or instrument are not binary, I replace them with indicators for whether these variables are above their medians. I demonstrate robustness to other binarizations in Tables C.1 and C.2 in Appendix C.

Table 5: Main Results on Negative First Stages and First-Stage Heterogeneity

A. Negative First Stages	LPM	Probit
Average Share	0.218	0.176
Weighted Average Share	0.285	0.280

B. First-Stage Heterogeneity	LPM	Probit
Rejected Papers	22/25	19/21
Average Share of Rejections	0.715	0.749

Notes: Panel A reports summary statistics on the fraction of observations for which $\hat{\mathrm{E}}[D \mid Z = 1, X = x] - \hat{\mathrm{E}}[D \mid Z = 0, X = x]$ is negative. "Average Share" treats every applicable regression equally. "Weighted Average Share" weights by the inverse of the number of applicable regressions associated with a given paper. Panel B reports results of Wald tests that the coefficients on the interaction terms in regressions of D on Z, X, and ZX are jointly equal to zero. "Rejected Papers" reports the number of papers for which the Bonferroni p-value is less than or equal to 0.05. "Average Share of Rejections" reports the average share (across papers) of regressions associated with a given paper for which the corresponding Holm p-value is less than or equal to 0.05. D and Z are defined as either the original endogenous explanatory variable and instrument (if they are binary) or indicators for whether these variables are above their medians, subject to a normalization that Z is always associated with a positive estimated coefficient in the linear first stage. Sampling weights and clustered standard errors are used in line with the original papers. Paper-specific results are reported in Table C.3 in Appendix C.

are robust to using the probit instead of the linear probability model (LPM).[16]

4.2 Reanalysis of Stevenson (2018)

Now, I turn to a reanalysis of Stevenson (2018)'s study of the effects of pretrial detention on case outcomes. In this application, recently reanalyzed by Cunningham (2021), Coulibaly, Hsu, Mourifié, and Wan (2024), and Mogstad and Torgovitsky (2024), violations of strong monotonicity are evident, which I will be able to formally demonstrate.

The data are based on the Philadelphia court records and cover 331,971 arrests between 2006 and 2013. The "treatment" of interest is pretrial detention or, in other words, whether the defendant was incarcerated in the period between their arrest and disposition; the purpose of such detention is that they appear in court and do not commit another crime. The empirical question is whether pretrial detention has a causal effect on case outcomes, such as conviction and incarceration length. Naturally, pretrial detention is endogenous, and Stevenson (2018)'s identification strategy is based on random assignment of bail magistrates (judges) to cases. These judges have broad authority to set bail—the amount required for pretrial release—at a level they choose. Thus, being assigned a strict judge makes the defendant less likely to be able to pay bail and more likely to be detained.

[16] The smaller number of papers under consideration when using the probit model reflects convergence and other estimation problems in the missing specifications.

Table 6: Eight Judges in Stevenson (2018)

	N	Detention Rate
Judge A	21,523	0.402
Judge B	13,087	0.432
Judge C	56,585	0.395
Judge D	33,690	0.413
Judge E	55,038	0.432
Judge F	41,475	0.413
Judge G	56,301	0.398
Judge H	54,272	0.418

Notes: The data are Stevenson (2018)'s sample of 331,971 arrests in Philadelphia. N is the number of cases heard by a given judge. "Detention Rate" is the proportion of cases heard by a given judge such that the defendant is subsequently detained pretrial.

In Philadelphia, bail hearings usually last one or two minutes, which made it possible for only eight judges to hear all the cases in Stevenson (2018)'s data. Table 6 reports the number of cases and the detention rate for each judge. The magistrate I refer to as "Judge C" is the most lenient, with a relatively low detention rate of 0.395. In what follows—unlike Stevenson (2018), who uses the full set of judge indicators as instruments—I focus on a single instrument defined as whether a given case was heard by Judge C. A simple regression of pretrial detention on the "Judge C" dummy reveals a first stage of –0.0195 with a standard error of 0.0023.

In the present context, strong monotonicity requires that every defendant detained by Judge C would also have been detained by other judges. However, this condition seems implausible, with the likely dimensions of monotonicity violations including the offense type (Stevenson, 2018) and the defendant's race (Abrams, Bertrand, and Mullainathan, 2012). As in Stevenson (2018), I focus on the seventeen most common offense types.[17] I also consider three racial categories: Black, White, and other. The offense types are not mutually exclusive, which means that, in principle, the sample could be divided into $3 \cdot 2^{17}$ groups based on the defendant's race and the offense type. However, most of these groups are empty, and I also drop nonempty groups with fewer than three cases heard by Judge C or not heard by Judge C. As a result, for this specification, the final sample consists of 431 groups and 327,560 cases.

With such a saturated specification, a formal test of violations of strong monotonicity is straightforward, as discussed in Remark 3.5. Given that Judge C is more lenient than others, the overall

[17] These include drug possession, drug sale, aggravated assault, robbery, first offense DUI, simple assault, drug purchase, burglary, shoplifting, theft, marijuana possession, murder, motor vehicle theft, prostitution, third-degree felony firearm possession, second-degree felony firearm possession, and vandalism.

first stage is negative—not positive, as assumed previously—and the null hypothesis requires that all the group-specific first stages are also non-positive. In other words, the null can be written as

$$H_0: \ \omega \leq 0 \tag{22}$$

while the test statistic equals

$$T = \max_{1 \leq k \leq K} \frac{\hat{\omega}_k}{\hat{\sigma}_{\hat{\omega}_k}}. \tag{23}$$

When I implement this test, I obtain a test statistic of 5.637 and a p-value of 3.7e-06, despite accounting for multiple hypothesis testing.[18] In this application, strong monotonicity is clearly rejected. Further details on the group-specific impact of Judge C on pretrial detention are provided in Table 7. Because presenting estimates for 431 groups is impractical, I restrict my attention to twenty groups with the largest (most positive) and ten groups with the smallest (most negative) z statistics. For each group, I report the number of cases, the conditional first stage and its standard error, and the corresponding Holm p-value. At any conventional significance level, we can reject that the first stage is non-positive in two groups: defendants charged with burglary and vandalism who are neither Black nor White and Black defendants charged with robbery, simple assault, and theft. In general, a common feature of many of the groups with the largest z statistics is a combination of being charged with a property crime (*e.g.*, theft or burglary) and a violent crime (*e.g.*, simple assault or aggravated assault). In fact, seven of these groups are charged with robbery, which is simultaneously a violent crime and a crime against property.[19] Many of these groups comprise of defendants who are neither Black nor White. On the other hand, the groups with the smallest z statistics are universally Black or White, and charged with nonviolent crimes.

Because strong monotonicity is rejected, the noninteracted specification cannot be used to estimate a convex combination of conditional LATEs (cf. Theorem 3.3). However, if weak monotonicity is plausible, the interacted specification will be appropriate, at least as long as the interacted instruments are sufficiently strong. In the present context, weak monotonicity seems quite sensible. Given that bail hearings in Philadelphia are extremely short, it is unlikely that more than a handful of factors—such as the offense type and the demographic characteristics of the defendant—could determine the amount of bail and the resulting likelihood of detention.

To incorporate additional factors into the analysis, I also consider two alternative specifications. First, I define the groups based on the offense type, the defendant's race, and their gender (male or female). In theory, the number of groups could be as large as $3 \cdot 2^{18}$ in this specification, but only

[18] In this application, the approach of Chernozhukov *et al.* (2019) produces almost identical critical values as the Bonferroni procedure.

[19] This is consistent with Stevenson (2018, p. 525)'s account that "[t]he magistrate that is most lenient overall is actually strictest when it comes to robbery." However, the test discussed in Remark 3.5 and the results in Table 7 are otherwise different from the analysis in Stevenson (2018).

Table 7: Conditional First Stages and Violations of Strong Monotonicity in Stevenson (2018)

Offense Type	Race	N	$\hat{\omega}_k$	$\hat{\sigma}_{\hat{\omega}_k}$	Holm p-value
burglary and vandalism	other	200	0.432***	0.077	3.73e-06
robbery, simple assault, and theft	Black	5,033	0.058***	0.014	0.00869
aggravated assault, robbery, simple assault, and theft	other	279	0.215**	0.085	1
first offense DUI and marijuana possession	other	135	0.143**	0.057	1
drug possession, robbery, simple assault, and theft	Black	93	0.253**	0.104	1
drug purchase	White	16	0.333**	0.140	1
drug purchase and marijuana possession	Black	111	0.218**	0.092	1
robbery, simple assault, theft, and third-degree felony firearm possession	other	221	0.163**	0.068	1
aggravated assault, drug possession, drug sale, and simple assault	Black	76	0.297**	0.134	1
aggravated assault, simple assault, and theft	other	30	0.432**	0.195	1
aggravated assault and simple assault	Black	14,262	0.024**	0.011	1
aggravated assault, robbery, simple assault, and third-degree felony firearm possession	other	11	0.500**	0.247	1
theft and vandalism	other	236	0.151**	0.076	1
burglary, theft, and vandalism	other	406	0.123**	0.062	1
aggravated assault, first offense DUI, and simple assault	Black	94	0.222*	0.114	1
burglary, robbery, theft, and third-degree felony firearm possession	Black	216	0.062*	0.033	1
robbery, simple assault, and theft	other	865	0.079*	0.043	1
third-degree felony firearm possession	White	416	0.116*	0.064	1
shoplifting and vandalism	Black	41	0.302*	0.169	1
burglary and theft	other	342	0.119*	0.067	1
. . .	. . .	. . .	. . .	. . .	. . .
drug possession and marijuana possession	Black	8,599	–0.049***	0.011	1
shoplifting	White	4,132	–0.088***	0.020	1
motor vehicle theft	White	890	–0.193***	0.041	1
drug possession and drug purchase	Black	6,885	–0.070***	0.014	1
motor vehicle theft	Black	2,183	–0.138***	0.026	1
drug possession	White	10,035	–0.052***	0.010	1
drug possession and drug purchase	White	7,692	–0.061***	0.011	1
prostitution	Black	2,967	–0.120***	0.022	1
theft	Black	5,886	–0.098***	0.017	1
shoplifting	Black	8,065	–0.115***	0.014	1

Notes: The data are Stevenson (2018)'s sample of 331,971 arrests in Philadelphia. The first two columns identify one of the 431 groups based on the offense type and the defendant's race. N is the number of cases in a given group. $\hat{\omega}_k$ is the conditional first stage, that is, the group-specific effect of Judge C on pretrial detention. $\hat{\sigma}_{\hat{\omega}_k}$ is the first-stage standard error. Entries in the table are sorted in descending order of $z = \hat{\omega}_k/\hat{\sigma}_{\hat{\omega}_k}$ and are restricted to twenty groups with the largest (most positive) and ten groups with the smallest (most negative) z statistics. Holm p-value equals $\min(1, p^*)$, where p^* is the product of the group-specific p-value for a one-sided test, based on $\hat{\omega}_k/\hat{\sigma}_{\hat{\omega}_k}$, and $r_k + 1$, where r_k is the number of group-specific p-values greater than that for a given group.

*Statistically different from zero at the 10% level; **at the 5% level; ***at the 1% level.

563 groups remain after I drop those that are empty or otherwise too small—requiring, as above, that there are at least three observations for every (G, Z) combination. Second, I define the groups based on the offense type, the defendant's race and gender, and three time periods considered by Stevenson (2018). The relevance of these specific time periods—divided by February 23, 2009 and February 23, 2011—results from concurring changes in the composition of magistrates other than Judge C. This sets the maximum number of groups in this specification at $3^2 \cdot 2^{18}$; in practice, the number of groups that are nonempty and sufficiently large is 981.

Table 8 reports the main results of my analysis. In panels C and D, for each of the three specifications described above, I report the Bonferroni/Chernozhukov *et al.* (2019) p-value for the test of violations of strong monotonicity as well as Mikusheva and Sun (2022)'s test statistic, $\widetilde{F}$, for the test of weak identification. The test results leave little doubt that strong monotonicity is violated while identification is strong. The p-values for the former test never exceed 0.0015, despite accounting for simultaneous testing of up to 981 hypotheses.[20] The values of $\widetilde{F}$ range between 19.32 and 21.56. If the simulations in Section 3.3.3 are any guide, we should expect negligible bias when estimating the interacted specification, at least when using the jackknife-type estimators such as IJIVE, UJIVE, and FEJIV.

Panels A and B of Table 8 report OLS, IV, 2SLS, IJIVE, UJIVE, and FEJIV estimates of the effects of pretrial detention on conviction and incarceration length. The noninteracted specification, marked as "IV," suggests that pretrial detention leads to a 17–19 p.p. increase in the likelihood of being convicted and an increase in incarceration length of 670–720 days. Such effects would have been substantial, but the validity of these estimates is questionable given the clear rejection of strong monotonicity in this application. When we turn to the interacted specification, the estimates become smaller. The effects on conviction are closer to zero—in the range of 4 to 15 p.p.—but often remain significant.[21] The effects on incarceration length are much smaller than in the noninteracted specification and suggestive of an effect of 50–160 days. These estimates are also usually not significantly different from zero, except for 2SLS. For both outcomes and each specification, the IJIVE, UJIVE, and FEJIV estimates are practically indistinguishable from each other but also clearly different from the 2SLS estimate.

My conclusions are generally in line with Stevenson (2018), whose paper includes a relatively rare recent example of using specifications with interacted instruments (cf. Remark 3.1), although

[20] In the second and third specifications, the largest z statistic is obtained in very small groups, which makes the normal approximation questionable. However, the rejection of strong monotonicity remains solid. The smallest Holm p-values in groups with at least 100 cases are 0.0109 and 0.0105 in the second and third specifications, respectively. The corresponding smallest Holm p-values in groups with at least 500 cases are 0.0408 and 0.0105. Note that these p-values are conservative, because they implicitly penalize hypothesis testing in groups smaller than 100 or 500 cases, even though such groups are ignored in this context.

[21] In the case of FEJIV, I report the standard errors derived by Chao *et al.* (2023). Practitioners should also consider a recent alternative proposed by Boot and Nibbering (2024), which explicitly accounts for treatment effect heterogeneity.

Table 8: Causal Effects of Pretrial Detention on Conviction and Incarceration Length

	Specification #1		Specification #2		Specification #3	
A. Effects on Conviction	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$
OLS	0.0591***	0.0019	0.0567***	0.0019	0.0530***	0.0019
IV	0.1852*	0.1070	0.1920*	0.1022	0.1704	0.1039
2SLS	0.1116***	0.0360	0.1193***	0.0325	0.0610**	0.0271
IJIVE	0.1283**	0.0585	0.1433***	0.0540	0.0429	0.0559
UJIVE	0.1304**	0.0616	0.1466**	0.0582	0.0419	0.0589
FEJIV	0.1292**	0.0604	0.1451**	0.0569	0.0417	0.0574
B. Effects on Incarceration Length	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$
OLS	184***	3	176***	2	173***	3
IV	689***	249	723***	237	666***	233
2SLS	158***	47	159***	41	130***	43
IJIVE	95	76	124*	68	58	79
UJIVE	93	83	123	75	56	99
FEJIV	92	78	122*	70	51	91
C. Test of Violations of Monotonicity						
p-value	3.73e-06		0.00133		1.83e-08	
D. Pretest for Weak Identification						
$\widetilde{F}$	21.38		21.56		19.32	
Number of Groups	431		563		981	
Number of Observations	327,560		325,915		319,573	

Notes: The data are Stevenson (2018)'s sample of 331,971 arrests in Philadelphia. The outcomes are conviction (Panel A) or incarceration length (Panel B), defined as the maximum days of an incarceration sentence. The treatment is pretrial detention. The instrument is whether a given case was heard by Judge C. Each specification is based on a division of the sample into a number of mutually exclusive groups, with a separate group for each combination of values of selected variables. Specification #1 uses the offense type and race (Black, White, or other) of the defendant. Specification #2 uses the offense type, race, and gender (male or female) of the defendant. Specification #3 uses the offense type, race and gender of the defendant, and three time periods considered by Stevenson (2018). Groups with fewer than three observations in either (G,Z) combination are dropped. "OLS" is the OLS estimator in the regression of the outcome on the treatment indicator and group indicators. "IV" is the IV estimator in the noninteracted specification. The remaining estimators are based on the interacted specification and are described in Section 3.3. The test of violations of monotonicity is described in Remark 3.5 and reports the Bonferroni/Chernozhukov *et al.* (2019) p-values. The pretest for weak identification follows Mikusheva and Sun (2022) and reports their test statistic, $\widetilde{F}$. The cutoff for this test is 4.14. See also the Stata implementation in Sun (2023).

*Statistically different from zero at the 10% level; **at the 5% level; ***at the 1% level.

not of AI's interacted specification, which is implemented here. I also provide additional results in Appendix D. Table D.1 reports MB2SLS and JIVE estimates of the effects of pretrial detention. While the MB2SLS estimates are largely similar to the results in Table 8, the JIVE estimates are noisy and appear unreliable. Table D.2 reports the results of a bootstrap test for comparisons between the IV estimates in the noninteracted specification and various estimates of the interacted specification.[22] At the 5% level, I nearly always reject the null that the estimands are the same in the case of incarceration length but not conviction. Finally, Table D.3 reports the results of a similar bootstrap test for comparisons between 2SLS and other estimators of the interacted specification. These differences are often highly statistically significant, which reaffirms the importance of correcting for the many instrument bias.

5 Conclusion

In this paper I studied the interpretation of linear IV and 2SLS estimands when both the treatment and the instrument are binary, and when additional covariates are required for identification. Using the LATE framework of Imbens and Angrist (1994) and Angrist *et al.* (1996), I argued that the common practice of interpreting standard IV estimands as a convex combination of conditional LATEs, or even as the (unconditional) local average treatment effect, is substantially more problematic than previously thought. I showed that the interpretation of the usual application of IV, which limits the effects of the instrument in the reduced-form and first-stage regressions to be homogeneous, hinges critically on the specific variant of the monotonicity assumption that the researcher is willing to entertain. Under "weak monotonicity," some of the IV weights may be negative and the IV estimand may no longer be interpretable as a causal effect.

What should applied researchers do in practice? In this paper I argued that it might be worthwhile to revisit the interacted specification of Angrist and Imbens (1995), which is guaranteed to eliminate negative weights under the same assumptions that are problematic for the usual application of IV. Specifications with many interacted instruments were used in influential papers by Angrist (1990) and Angrist and Krueger (1991) but appear to have been largely abandoned in subsequent work out of concern for the many instrument bias. Unsurprisingly, however, the modern tools to estimate such specifications are substantially better than in the 1990s, as I also demonstrate in an extensive simulation study. A pretest for weak identification developed by Mikusheva and Sun (2022) can be used to determine whether consistent estimation of the interacted specification is possible. When the pretest rejects, several jackknife-type estimators can be used, including the FEJIV estimator of Chao *et al.* (2023), which I also implement in the companion R and Stata packages, `fejiv`.

There are at least two important situations when this recommendation will not be satisfactory. First, in some applications in which strong monotonicity is rejected, weak monotonicity will be implausible, too. If this is the case, it may be worthwhile to instead consider the partial identification approach of Noack (2021), which evaluates the sensitivity of what can be learned about the

[22] I perform this test for 2SLS, MB2SLS, JIVE, and UJIVE, but not IJIVE and FEJIV, because the latter estimators are very computationally demanding in the specifications that I consider, at least in my implementation.

local average treatment effect under violations of (weak) monotonicity. Second, a convex combination of conditional LATEs, which Angrist and Imbens (1995)'s specification is guaranteed to produce under weak monotonicity (and the usual application of IV under strong monotonicity), may be considered an imperfect substitute for the (unconditional) local average treatment effect.[23] If this is the case, there are many existing estimators that are consistent for the LATE under strong monotonicity (see, *e.g.*, Słoczyński, Uysal, and Wooldridge, 2025, and the references therein). An important avenue for future research is to develop estimators of the LATE that are also robust to weak monotonicity.

References

Abadie, Alberto (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Abrams, David S., Marianne Bertrand, and Sendhil Mullainathan (2012): "Do Judges Vary in Their Treatment of Race?" *Journal of Legal Studies*, 41, 347–384.

Ackerberg, Daniel A. and Paul J. Devereux (2009): "Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity," *Review of Economics and Statistics*, 91, 351–362.

Aizer, Anna and Joseph J. Doyle (2015): "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges," *Quarterly Journal of Economics*, 130, 759–804.

Anatolyev, Stanislav (2013): "Instrumental Variables Estimation and Inference in the Presence of Many Exogenous Regressors," *Econometrics Journal*, 16, 27–72.

——— (2019): "Many Instruments and/or Regressors: A Friendly Guide," *Journal of Economic Surveys*, 33, 689–726.

Anderson, Theodore Wilbur and Herman Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20, 46–63.

Angrist, Joshua D. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313–336.

——— (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249–288.

Angrist, Joshua D. and Guido W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67.

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, 444–455.

[23] A similar argument is made by Callaway, Goodman-Bacon, and Sant'Anna (2024) in the context of difference-in-differences designs.

Angrist, Joshua D. and Alan B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton–Oxford: Princeton University Press.

Bekker, Paul A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62, 657–681.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky (2022): "When Is TSLS Actually LATE?" NBER Working Paper no. 29709.

——— (2025): "When Is TSLS Actually LATE?" NBER Working Paper no. 29709.

Bond, Simon J., Ian R. White, and A. Sarah Walker (2007): "Instrumental Variables and Interactions in the Causal Analysis of a Complex Clinical Trial," *Statistics in Medicine*, 26, 1473–1496.

Bondy, Malvina, Sefi Roth, and Lutz Sager (2020): "Crime Is in the Air: The Contemporaneous Relationship between Air Pollution and Crime," *Journal of the Association of Environmental and Resource Economists*, 7, 555–585.

Boot, Tom and Didier Nibbering (2024): "Inference on LATEs with Covariates," arXiv:2402.12607.

Bound, John, David A. Jaeger, and Regina M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna (2024): "Difference-in-Differences with a Continuous Treatment," NBER Working Paper no. 32117.

Card, David (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky, Toronto: University of Toronto Press, 201–222.

Chao, John C. and Norman R. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692.

Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen (2012): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," *Econometric Theory*, 28, 42–86.

Chao, John C., Norman R. Swanson, and Tiemen Woutersen (2023): "Jackknife Estimation of a Cluster-Sample IV Regression Model with Many Weak Instruments," *Journal of Econometrics*, 235, 1747–1769.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2019): "Inference on Causal and Structural Parameters Using Many Moment Inequalities," *Review of Economic Studies*, 86, 1867–1900.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler (2015a): "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," *American Economic Review: Papers & Proceedings*, 105, 486–490.

——— (2015b): "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688.

Coulibaly, Mohamed, Yu-Chin Hsu, Ismael Mourifié, and Yuanyuan Wan (2024): "A Sharp Test for the Judge Leniency Design," NBER Working Paper no. 32456.

Cunningham, Scott (2021): *Causal Inference: The Mixtape*, New Haven–London: Yale University Press.

Dahl, Christian M., Martin Huber, and Giovanni Mellace (2023): "It Is Never Too LATE: A New Look at Local Average Treatment Effects with or without Defiers," *Econometrics Journal*, 26, 378–404.

de Chaisemartin, Clément (2017): "Tolerating Defiance? Local Average Treatment Effects without Monotonicity," *Quantitative Economics*, 8, 367–396.

de Chaisemartin, Clément and Xavier D'Haultfœuille (2020): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110, 2964–2996.

Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif (2019): "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction," *American Economic Review*, 109, 4178–4219.

Evdokimov, Kirill S. and Michal Kolesár (2019): "Inference in Instrumental Variables Analysis with Heterogeneous Treatment Effects." Unpublished.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group (2012): "The Oregon Health Insurance Experiment: Evidence from the First Year," *Quarterly Journal of Economics*, 127, 1057–1106.

Fuller, Wayne A. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–953.

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

Huntington-Klein, Nick (2020): "Instruments with Heterogeneous Effects: Bias, Monotonicity, and Localness," *Journal of Causal Inference*, 8, 182–208.

Imbens, Guido W. and Joshua D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Ishimaru, Shoya (2024): "Empirical Decomposition of the IV-OLS Gap with Heterogeneous and Nonlinear Effects," *Review of Economics and Statistics*, 106, 505–520.

Kitagawa, Toru (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

Kolesár, Michal (2013): "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity." Unpublished.

Kunitomo, Naoto (1980): "Asymptotic Expansions of the Distributions of Estimators in a Linear Functional Relationship and Simultaneous Equations," *Journal of the American Statistical Association*, 75, 693–700.

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

Mikusheva, Anna and Liyang Sun (2022): "Inference with Many Weak Instruments," *Review of Economic Studies*, 89, 2663–2686.

——— (2024): "Weak Identification with Many Instruments," *Econometrics Journal*, 27, C1–C28.

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

——— (2024): "Instrumental Variables with Unobserved Heterogeneity in Treatment Effects," in *Handbook of Labor Economics, Vol. 5*, ed. by Christian Dustmann and Thomas Lemieux, Amsterdam: Elsevier, 1–114.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters (2021): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*, 111, 3663–3698.

Morimune, Kimio (1983): "Approximate Distributions of k-Class Estimators When the Degree of Overidentifiability Is Large Compared with the Sample Size," *Econometrica*, 51, 821–841.

Mueller-Smith, Michael (2015): "The Criminal and Labor Market Impacts of Incarceration," Unpublished.

Nagar, Anirudh L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595.

Noack, Claudia (2021): "Sensitivity of LATE Estimates to Violations of the Monotonicity Assumption," arXiv:2106.06421.

Semenova, Vira (2025): "Generalized Lee Bounds," *Journal of Econometrics*, 251, 106055.

Słoczyński, Tymon (2022): "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *Review of Economics and Statistics*, 104, 501–509.

Słoczyński, Tymon, S. Derya Uysal, and Jeffrey M. Wooldridge (2025): "Abadie's Kappa and Weighting Estimators of the Local Average Treatment Effect," *Journal of Business & Economic Statistics*, 43, 164–177.

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

Sun, Liyang (2023): "manyweakiv: Weak-Instrument Robust Test for Linear IV Regressions with Many Instruments," https://github.com/lsun20/manyweakiv.

Walters, Christopher R. (2018): "The Demand for Effective Charter Schools," *Journal of Political Economy*, 126, 2179–2223.

Wiemann, Thomas (2024): "Optimal Categorical Instrumental Variables," arXiv:2311.17021.

Young, Alwyn (2022): "Consistency without Inference: Instrumental Variables in Practical Application," *European Economic Review*, 147, 104112.