

A PRACTICAL GUIDE TO INSTRUMENTAL VARIABLES METHODS WITH HETEROGENEOUS TREATMENT EFFECTS*

TYMON SŁOCZYŃSKI[†] LIYANG SUN[‡] S. DERYA UYSAL[§]

MAY 13, 2026

Instrumental variables (IV) methods represent one of the most popular approaches to identifying and estimating causal effects in economics. Currie et al. (2020) establish that the popularity of IV methods in applied microeconomics has grown continuously since the 1980s, as indicated by the content of the National Bureau of Economic Research (NBER) working papers. Likewise, Goldsmith-Pinkham (2026) documents that roughly 30% of NBER working papers in applied microeconomics mention instrumental variables.

Standard textbook treatments of IV methods assume a linear model with constant effects. At the same time, the influential local average treatment effect (LATE) framework of Imbens and Angrist (1994), Angrist and Imbens (1995), and Angrist et al. (1996) allows for a very general form of treatment effect heterogeneity, which the textbook model rules out. In this paper, we will offer a nontechnical, practical guide to the literature on IV methods with heterogeneous treatment effects, which originates from the work of Angrist and Imbens. Instead of aiming to provide a comprehensive survey of the literature, our focus will be on highlighting several areas in which existing work in applied microeconomics deviates from what we regard as best practices motivated by the recent theoretical literature.

The first area we explore is the choice of the target parameter. Recent research highlights possible differences between the local average treatment effect (LATE), as defined by Imbens and Angrist (1994), and the probability limits of usual IV and two-stage least squares (2SLS) estimators, especially in cases where the instrument is only valid conditional on covariates. As an alternative, we will discuss an intuitive, general strategy to estimate the LATE when

*We thank Jonathan Roth and Jeff Wooldridge for helpful conversations and comments. Liyang Sun acknowledges generous support from Stone Centre's internal research support. Derya Uysal acknowledges financial support from Deutsche Forschungsgemeinschaft (CRC TRR 190, project number 280092119).

[†]Department of Economics, Brandeis University, tslocz@brandeis.edu

[‡]Department of Economics, University College London, liyang.sun@ucl.ac.uk

[§]Department of Economics, LMU Munich, derya.uysal@econ.lmu.de

covariates matter. The second area we consider is the flexibility of parametric specifications. Recent research shows that a causal interpretation of IV and 2SLS estimands requires that the conditional mean of the instrument given covariates be linear. Parametric misspecification is also possible when estimating the LATE with covariates. As a solution to such problems, we will discuss flexible estimation strategies based on machine learning. The third area we explore is possible violations of the assumptions underlying the LATE framework. These assumptions have testable implications that have spurred a sizable theoretical literature, yet have had limited influence on applied work. We will discuss the intuition behind the resulting tests and their implementation. We will also discuss estimation approaches that are more robust to violations of monotonicity, which is often a controversial assumption.

Throughout the paper, we focus on the standard LATE framework with a binary treatment and a binary instrument as the leading case. Interested readers should also consider several other papers that offer a comprehensive survey of the existing literature, or focus instead on a subset of possible applications of IV methods. Mogstad and Torgovitsky (2018) use the marginal treatment effect (MTE) framework of Heckman and Vytlacil (2005) to discuss IV identification and extrapolation under treatment effect heterogeneity. Mogstad and Torgovitsky (2024) review the literature on IV methods with heterogeneous treatment effects, including extensions to multivalued treatments, multivalued instruments, and multiple instruments. Borusyak et al. (2025) provide a guide to the literature on shift-share instruments. Chyn et al. (2025) and Goldsmith-Pinkham et al. (2025) survey the literature on “judge leniency” designs, which is another important application of IV methods.

Target Parameters

In this section we discuss the differences between three leading estimands (target parameters) in applications of IV methods: the probability limit in the usual linear IV regression, the 2SLS estimand in a saturated specification with multiple interacted instruments, and the “true” local average treatment effect. We argue that this last parameter should be more commonly estimated in relevant applied work—ideally as the main parameter of interest or at least as a robustness check. We also illustrate the possible differences between these parameters in two empirical applications.

Motivating Example: Stratified RCTs with Imperfect Compliance

We start by considering a leading application of instrumental variables methods: a randomized controlled trial (RCT) with imperfect compliance. Concretely, let Y_i denote the outcome, D_i denote the binary treatment, and Z_i denote the randomized assignment to

treatment. We will refer to Z_i as the instrument, following standard econometrics terminology. We want to leverage the randomness in Z_i to estimate the causal effect of D_i on Y_i . Let $D_i(1)$ and $D_i(0)$ denote the two potential treatments, i.e., the counterfactual treatment statuses that would be observed if a unit were assigned ($Z_i = 1$) or not assigned ($Z_i = 0$) to be treated. (Observed and assigned treatment statuses may differ due to imperfect compliance.) Following the usual terminology, we refer to units with $D_i(1) > D_i(0)$ as compliers. Under the standard assumptions underlying the local average treatment effect framework—namely, independence of the instrument, exclusion restriction, relevance, and monotonicity—the usual linear IV regression identifies the LATE, i.e., the average treatment effect for compliers (Imbens and Angrist, 1994). This is the so-called Wald estimand:

$$\beta = \frac{\mathbb{E}(Y_i | Z_i = 1) - \mathbb{E}(Y_i | Z_i = 0)}{\mathbb{E}(D_i | Z_i = 1) - \mathbb{E}(D_i | Z_i = 0)} = \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) > D_i(0)],$$

where $Y_i(1)$ and $Y_i(0)$ are potential outcomes.

Treatment assignment, however, is rarely *completely* random. Consider a stratified RCT where the assignment is only random once we condition on an appropriate set of covariates, X_i . For example, in the Oregon Health Insurance Experiment (OHIE), a randomized lottery that assigned low-income adults the opportunity to apply for Medicaid, it is necessary to control for household size and survey wave (Finkelstein et al., 2012). The independence assumption, $(Y_i(1), Y_i(0), D_i(1), D_i(0)) \perp Z_i$, no longer holds unconditionally because these covariates may be related to potential outcomes and treatments as well as the probability of treatment assignment. Specifically, household size had a direct effect on the chance of winning this particular lottery, while, at the same time, it is plausibly related to health expenditures, health care utilization, and baseline health status. Similarly, survey wave captures timing effects that may have influenced who was reached by the survey, while it also picks up time-varying shocks to outcomes. Consequently, Z_i is no longer independent of potential outcomes and treatments unless we control for these particular covariates.

There is not, however, a unique method to control for covariates, and this makes the LATE framework with covariates substantially more complicated than the baseline model without covariates. In what follows, we discuss three approaches to control for covariates in stratified RCTs with imperfect compliance. First, consider a *linear IV regression*, a population regression specification where covariates, X_i , are controlled in an additively separable fashion in the outcome equation,

$$Y_i = \gamma'X_i + \beta_{IV}D_i + e_i \tag{1}$$

as well as in the first-stage regression,

$$D_i = \delta' X_i + \pi Z_i + v_i. \quad (2)$$

In the case of the OHIE, the elements of X_i are discrete and consist of mutually exclusive indicators for different combinations of values of household size and survey wave. The method of analysis in Finkelstein et al. (2012) is equivalent to estimating β_{IV} using these covariates. To simplify the comparison with other approaches, we assume, for now, that $X_i = (X_{i1}, \dots, X_{iJ})$ does indeed saturate the model. That is, X_i is a vector of dimension J , each entry X_{ij} is a dummy variable, and for every observation i , $\sum_j X_{ij} = 1$.

Second, consider a *2SLS regression*, a population regression specification where the outcome equation is still specified as additively separable but covariates are interacted with the instrument in the first-stage regression:

$$Y_i = \gamma' X_i + \beta_{AI} D_i + e_i, \quad (3)$$

$$D_i = \delta' X_i + \sum_j \pi_j Z_i X_{ij} + v_i. \quad (4)$$

Angrist and Pischke (2009) refer to this specification as the “saturate and weight” approach. Słoczyński (2026) calls it “AI’s specification,” after Angrist and Imbens (1995). With saturated covariates, the first-stage regression in equation (4) is correctly specified by design, and each π_j may be interpreted as the causal effect of Z_i on D_i for the subgroup with covariate $X_{ij} = 1$, which is the same as the proportion of compliers in that subgroup.¹

Although equation (4) is correctly specified, equation (3) still imposes a constant effect of D_i on Y_i . Thus, the third approach we consider is grounded in a nonparametric identification result rather than a specific estimator. As shown by Frölich (2007), the average treatment effect for compliers is identified as

$$\beta_{LATE} = \frac{\mathbb{E}\{\mathbb{E}(Y_i | Z_i = 1, X_i) - \mathbb{E}(Y_i | Z_i = 0, X_i)\}}{\mathbb{E}\{\mathbb{E}(D_i | Z_i = 1, X_i) - \mathbb{E}(D_i | Z_i = 0, X_i)\}}, \quad (5)$$

which means that the LATE is the ratio of the average treatment effect of the instrument on the outcome and the average treatment effect of the instrument on the treatment. (As we will see below, β_{IV} and β_{AI} are *not* generally equal to the LATE.) In practice, this ratio can be estimated by a variety of approaches, which consist of using one’s favorite estimator of the average treatment effect twice—once to compute the numerator and again to compute the denominator of the expression in (5). With saturated covariates, this is

¹We return to the issue of misspecification in the next section.

particularly straightforward, because many standard estimators will be numerically identical and equivalent to a procedure that consists of computing the effect of the instrument on the outcome and the treatment in each covariate cell (using simple differences in means), and aggregating them together using the proportions of observations in those cells.

Under the LATE assumptions generalized to the case with covariates, as in Abadie (2003), we can derive a weighted average representation of the three estimands in terms of covariate-specific LATEs. Denote by $\tau_j = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0), X_{ij} = 1]$ the average treatment effect for compliers with covariate level j and by $p_j = \mathbb{P}(X_{ij} = 1)$ the share of individuals with that covariate level. From Słoczyński (2026), we have

$$\beta_{IV} = \frac{\sum_j p_j \pi_j \cdot \text{Var}[Z_i \mid X_{ij} = 1] \cdot \tau_j}{\sum_j p_j \pi_j \cdot \text{Var}[Z_i \mid X_{ij} = 1]} \quad (6)$$

and

$$\beta_{AI} = \frac{\sum_j p_j \pi_j^2 \cdot \text{Var}[Z_i \mid X_{ij} = 1] \cdot \tau_j}{\sum_j p_j \pi_j^2 \cdot \text{Var}[Z_i \mid X_{ij} = 1]}, \quad (7)$$

where the representation in (7) is also implied by Theorem 3 in Angrist and Imbens (1995). Also, from Frölich (2007), we have

$$\beta_{LATE} = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)] = \frac{\sum_j p_j \pi_j \cdot \tau_j}{\sum_j p_j \pi_j}. \quad (8)$$

As is clear from these expressions, the average treatment effect for compliers, β_{LATE} , may be substantially different from β_{IV} , β_{AI} , or both—as long as π_j , the conditional proportion of compliers, or $\text{Var}[Z_i \mid X_{ij} = 1]$, the conditional variance of the instrument, vary across covariate values.² Relative to the LATE, β_{IV} overweights the covariate cells with large values of $\text{Var}[Z_i \mid X_{ij} = 1]$, while β_{AI} overweights those with large values of the product of π_j and $\text{Var}[Z_i \mid X_{ij} = 1]$ (Słoczyński, 2026). In a model with homogeneous treatment effects, such weighting may be desirable in practice due to improved estimation efficiency, but with heterogeneous treatment effects, it would be difficult to substantiate an explicit interest in β_{IV} or β_{AI} . That being said, the weighted average representations in (6)–(8) only imply that the three estimands *may* be substantially different from each other, but it is an empirical question whether they will actually differ in a particular application.

To illustrate this empirical question, we use data from Finkelstein et al. (2012). In this setting, as mentioned above, the instrument (the ability to apply for Medicaid) is randomized conditional on household size and survey wave, and the identifying assumptions underlying

²The interpretation of β_{LATE} , β_{IV} , and β_{AI} will change when the monotonicity assumption is violated. We return to this issue below.

Table 1: Causal Effects of Medicaid on the Number of Outpatient Visits

	$\hat{\beta}_{\text{LATE}}$	$\hat{\beta}_{\text{IV}}$	$\hat{\beta}_{\text{AI}}$
A. Full Sample	1.001 (0.134)	1.001 (0.134)	1.010 (0.132)
B. Three-Person Households	4.022 (1.189)	2.790 (1.296)	2.268 (1.118)

Notes: The source of data is Finkelstein et al. (2012). The outcome is the number of outpatient visits in the last six months, as in Table V of Finkelstein et al. (2012). The sample sizes are 23,441 and 58 in Panels A and B, respectively. Standard errors are in parentheses and are clustered at the household level.

the LATE framework seem plausible. In our subsequent analysis, we focus on the impact of Medicaid coverage on the number of outpatient visits in the last six months. To simplify the illustration, unlike Finkelstein et al. (2012), we do not use survey weights.

Panel A of Table 1 reports the estimates of β_{LATE} , β_{IV} , and β_{AI} using the full sample from the OHIE. Although there is substantial heterogeneity in covariate-specific LATEs, the different estimands yield nearly identical point estimates and standard errors, suggestive of a precisely estimated increase in the number of outpatient visits in response to Medicaid enrollment by about 1 in six months on average. Apparently, the correlation between $\widehat{\text{Var}}[Z_i | X_{ij} = 1]$ and $\hat{\tau}_j$ as well as $\hat{\pi}_j \cdot \widehat{\text{Var}}[Z_i | X_{ij} = 1]$ and $\hat{\tau}_j$ is sufficiently weak among compliers to render the differences between $\hat{\beta}_{\text{LATE}}$, $\hat{\beta}_{\text{IV}}$, and $\hat{\beta}_{\text{AI}}$ meaningless.

The picture is very different, however, when we restrict our attention to the subsample of three-person households in Panel B of Table 1. This restriction does not alter the identifying assumptions, i.e., randomization continues to hold conditional on survey wave, but it changes the practical importance of the different components in the weighted average representations in (6)–(8). Indeed, the estimated LATE, 4.022, is now substantially larger than $\hat{\beta}_{\text{IV}}$ and $\hat{\beta}_{\text{AI}}$, which are equal to 2.790 and 2.268, respectively.

The differences between $\hat{\beta}_{\text{LATE}}$, $\hat{\beta}_{\text{IV}}$, and $\hat{\beta}_{\text{AI}}$ follow directly from their distinct weighting schemes. The estimated LATE aggregates covariate-specific treatment effects in proportion to each cell’s estimated share among the compliers, $\hat{p}_j \hat{\pi}_j / \sum_k \hat{p}_k \hat{\pi}_k$. By contrast, the IV and 2SLS estimates additionally weight each cell by the conditional variance of the instrument, $\widehat{\text{Var}}[Z_i | X_{ij} = 1]$, and the latter again by $\hat{\pi}_j$. As a result, these estimates place relatively more weight on covariate cells in which treatment effects can be estimated more precisely, which need not coincide with those that are most representative of the complier population.

Table 2 illustrates this mechanism. For each of the three covariate cells in the subsample of three-person households, we report its sample proportion, \hat{p}_j , the conditional variance of

Table 2: Estimated Weights

Survey Wave	\hat{p}_j	$\widehat{\text{Var}}[Z_i X_{ij} = 1]$	$\hat{\pi}_j$	$\hat{w}_{\text{LATE},j}$	$\hat{w}_{\text{IV},j}$	$\hat{w}_{\text{AI},j}$	$\hat{\tau}_j$
1	0.241	0.168	0.455	0.351	0.600	0.723	1.067
2	0.448	0.037	0.280	0.401	0.151	0.112	6
3	0.310	0.099	0.250	0.248	0.249	0.165	5

Notes: The source of data is Finkelstein et al. (2012). The table is restricted to the subsample of 58 individuals in three-person households. \hat{p}_j is the proportion of observations in covariate cell (survey wave) j . $\widehat{\text{Var}}[Z_i | X_{ij} = 1]$ is the variance of Z_i in cell j . $\hat{\pi}_j$ is the estimated proportion of compliers in cell j . $\hat{w}_{\text{LATE},j}$, $\hat{w}_{\text{IV},j}$, and $\hat{w}_{\text{AI},j}$ are the estimated weights of cell j in β_{LATE} , β_{IV} , and β_{AI} , respectively. $\hat{\tau}_j$ is the estimated LATE on the number of outpatient visits in cell j . Each of $\hat{\beta}_{\text{LATE}}$, $\hat{\beta}_{\text{IV}}$, and $\hat{\beta}_{\text{AI}}$, as reported in Panel B of Table 1, can be obtained as the dot product of $\hat{\tau}_j$ and the respective weights.

Z_i , as well as the estimated proportion of compliers, $\hat{\pi}_j$, and conditional LATE, $\hat{\tau}_j$. We also report the estimated weights underlying each of the three estimands, which are simply the sample counterparts of the expressions in (6)–(8). For example, $\hat{w}_{\text{LATE},j} = \hat{p}_j \hat{\pi}_j / \sum_k \hat{p}_k \hat{\pi}_k$ and $\hat{w}_{\text{IV},j} = \hat{p}_j \hat{\pi}_j \cdot \widehat{\text{Var}}[Z_i | X_{ij} = 1] / \sum_k \hat{p}_k \hat{\pi}_k \cdot \widehat{\text{Var}}[Z_i | X_{ik} = 1]$. In the first covariate cell, $\hat{w}_{\text{LATE},1} = (0.241 \cdot 0.455) / (0.241 \cdot 0.455 + 0.448 \cdot 0.280 + 0.310 \cdot 0.250) \approx 0.351$ and $\hat{w}_{\text{IV},1} = (0.241 \cdot 0.168 \cdot 0.455) / (0.241 \cdot 0.168 \cdot 0.455 + 0.448 \cdot 0.037 \cdot 0.280 + 0.310 \cdot 0.099 \cdot 0.250) \approx 0.600$. (All values are rounded for illustration.) Each of the estimates reported in Panel B of Table 1 can be obtained as the dot product of $\hat{\tau}_j$ and the respective weights. For example, $\hat{\beta}_{\text{LATE}} = 0.351 \cdot 1.067 + 0.401 \cdot 6 + 0.248 \cdot 5 \approx 4.022$ and $\hat{\beta}_{\text{IV}} = 0.600 \cdot 1.067 + 0.151 \cdot 6 + 0.249 \cdot 5 \approx 2.790$. In this subsample, because $\widehat{\text{Var}}[Z_i | X_{ij} = 1]$ and $\hat{\pi}_j$ are large when $\hat{\tau}_j$ is small, $\hat{\beta}_{\text{IV}}$ and $\hat{\beta}_{\text{AI}}$ underweight the covariate cells with large estimated treatment effects, producing a downward bias relative to the LATE. As a result, the three estimands answer different causal questions despite relying on the same experimental variation.³ A cautious researcher should investigate whether a similar pattern arises in their specific setting.

Beyond Stratified RCTs

While strata indicators necessarily saturate the model in stratified RCTs, saturated specifications may be infeasible in other applications of instrumental variables methods, especially when there are multiple covariates or some covariates are continuously distributed. If saturation is infeasible, so is the specification in equations (3)–(4). The linear IV regression in

³One way to see this is through the lens of the framework developed by Poirier and Słoczyński (2025) to quantify the internal validity and representativeness of various estimands. If we treat the population of compliers—the largest subpopulation for which the average treatment effect is nonparametrically identified in an IV setting under standard assumptions—as our target, then β_{LATE} is associated with the largest possible value of the measure of internal validity, equal to 1, while the measures associated with β_{IV} and β_{AI} will generally be less than 1. This means that these estimands correspond to average treatment effects for some (possibly small) subsets of the complier subpopulation, whereas β_{LATE} is the average over all compliers.

equations (1)–(2) remains feasible, but a weighted average representation of β_{IV} , analogous to equation (6), additionally requires the parametric restriction that the conditional mean of Z_i given X_i is linear in X_i . Blandhol et al. (2026) refer to this assumption as “rich covariates” and prove that it is necessary for a causal interpretation of β_{IV} .

As in the case of stratified RCTs, we recommend explicitly estimating β_{LATE} in every relevant non-RCT application, even if only as a robustness check alongside $\hat{\beta}_{IV}$. The two parameters, β_{IV} and β_{LATE} , may often be similar, but when they are different, β_{LATE} is arguably of much greater interest. As discussed above, the expression in (5) implies that the LATE can be estimated as the ratio of estimates of the average treatment effect of the instrument on the outcome and treatment. In the absence of saturation, there are many distinct estimators of the average treatment effect with good properties, which implies that there are also many suitable estimators of the LATE.⁴

One such estimator is based on inverse probability weighting (IPW), which requires first-step estimation of the conditional mean of Z_i given X_i , often referred to as the instrument propensity score. This is the same conditional mean that is central to the “rich covariates” assumption. Here, however, we focus on estimating β_{LATE} and will likely use a logit or probit model, not the linear probability model implicit in β_{IV} . In any case, the IPW estimator uses the first-step estimates of the instrument propensity score to reweight the units with $Z_i = 1$ and $Z_i = 0$ so that both subsamples become comparable in terms of their covariate values. This, in turn, enables straightforward estimation of the average treatment effects of the instrument on the outcome and treatment using simple differences in means of reweighted data. The ratio of these differences is the IPW estimator of the LATE, implied by equation (5) and proposed by Frölich (2007). Słoczyński et al. (2025) emphasize the importance of using normalized weights in this context, explore the connections to “kappa weighting” of Abadie (2003), and develop the Stata package `kappalate`. The package allows both binary and non-binary (discrete or continuous) treatments; the instrument, however, must be binary.

To illustrate the possible differences between β_{IV} and β_{LATE} in a nonsaturated setting, we estimate both parameters using Abadie’s (2003) sample from the 1991 Survey of Income and Program Participation (SIPP). In this application, based on Poterba et al. (1994), one of the causal effects of interest is that of participation in a 401(k) retirement plan on participation in an individual retirement account (IRA). While 401(k) participation is likely endogenous, Poterba et al. (1994) and Abadie (2003) argue that 401(k) eligibility, which is determined by the employer, can be used as an instrument for 401(k) participation. Because 401(k)

⁴Some of these estimators include those in Tan (2006), Frölich (2007), Uysal (2011), Heiler (2022), Sant’Anna et al. (2022), Słoczyński et al. (2022, 2025), and Ma et al. (2026). Related machine learning approaches have been developed by Belloni et al. (2017), Chernozhukov et al. (2018), Sun and Tan (2022), Singh and Sun (2024), and others.

Table 3: Causal Effects of 401(k) on IRA Participation

$\hat{\beta}_{IV}$	$\hat{\beta}_{LATE}$ (logit)	$\hat{\beta}_{LATE}$ (probit)
0.0274	0.0165	0.0175
(0.0132)	(0.0135)	(0.0136)

Notes: The data are Abadie’s (2003) subsample of the 1991 Survey of Income and Program Participation (SIPP). The sample size is 9,275. The outcome is an indicator for participation in IRAs, as in Table 3 of Abadie (2003). The covariate specification includes family income, age, age squared, marital status, and family size. $\hat{\beta}_{LATE}$ corresponds to the normalized IPW estimator of Uysal (2011) and Słoczyński et al. (2025), with logit or probit instrument propensity scores estimated using maximum likelihood. Heteroskedasticity-robust standard errors are in parentheses.

eligibility is not randomized or “as good as randomly assigned,” Abadie (2003) controls for family income, age, age squared, marital status, and family size to estimate the causal effects of 401(k) participation using this instrument.

Table 3 replicates the linear IV estimate of the effect of 401(k) on IRA participation, as reported in column (2) of Table 3 in Abadie (2003). This estimate, statistically significant at the 5% level, suggests that participation in a 401(k) retirement plan increases the probability of IRA participation by about 2.74 percentage points. However, Table 3 also reports two IPW estimates of the LATE. These estimates, based on logit and probit instrument propensity scores, are much smaller than the linear IV estimate, with p -values equal to 0.221 and 0.197, respectively. This calls into question the initial conclusion about the positive impact of 401(k) on IRA participation. If we ignore estimation uncertainty and the possibility of misspecification, the difference between $\hat{\beta}_{IV}$ and $\hat{\beta}_{LATE}$ must be driven by the different weights underlying the corresponding estimands. In our view, the possibility that such differences may materialize necessitates explicit estimation of β_{LATE} in relevant applied work—either as the main parameter of interest or at least as a robustness check.

Panel A of Appendix Table A1 lists several R and Stata packages, including `kappalate`, which can be used to estimate the LATE when covariates matter.

Parametric Misspecification

In this section we discuss the possibility that estimates of β_{IV} and β_{LATE} may be biased due to parametric misspecification. If all the requisite covariates are observed and controlled for, the leading case of misspecification occurs when important interaction and higher-order terms are omitted. First, we discuss how this concern applies to estimation of both β_{IV} and β_{LATE} , and how it is possible to test for parametric misspecification. Second, we recommend modern

estimation approaches based on machine learning as a general solution to this problem.

Rich Covariates

A large literature on weighted average representations of β_{IV} , β_{AI} , and related estimands has used the assumption that the conditional mean of Z_i given X_i is linear in X_i . As mentioned in the previous section, a recent paper by Blandhol et al. (2026) shows that this assumption, termed “rich covariates,” is not only sufficient but also *necessary* for the resulting estimand to represent a non-negatively weighted average of complier causal effects. If covariates are not rich, the estimand is not “weakly causal,” which means that when all conditional average treatment effects have the same sign, the estimand may assume the opposite sign. Poirier and Słoczyński (2025) show that this is equivalent to the absence of any subpopulation of compliers whose average treatment effect is recovered by the estimand of interest regardless of the pattern of treatment effect heterogeneity.

An immediate implication is that applied researchers targeting simple instrumental variables estimands should only use covariate specifications that are plausibly rich. To examine whether a given specification is rich, Blandhol et al. (2026) recommend the regression specification error test (RESET) of Ramsey (1969). In its canonical form, RESET, as applied to this problem, consists of estimating a regression of Z_i on X_i using OLS, obtaining fitted values, \hat{Z}_i , and, finally, estimating a regression of Z_i on X_i and several higher-order terms of \hat{Z}_i , such as \hat{Z}_i^2 , \hat{Z}_i^3 , and \hat{Z}_i^4 (again using OLS). In this setup, RESET is simply the F test of joint significance of the higher-order terms in the second-step regression, and should be interpreted as a test for neglected nonlinearity (Wooldridge, 2010, Section 6.3.3). If we reject the null hypothesis, we conclude that the original covariate specification is not rich.

It would be inaccurate to suggest that estimators of β_{LATE} cannot suffer from similar specification problems. For example, the IPW estimator discussed in the previous section relies on the assumption that the conditional mean of Z_i given X_i is correctly specified, *however it is specified*. This offers more flexibility than the result in Blandhol et al. (2026), because we are no longer tied to the linear probability model and can instead use, say, the logit or probit model. On the other hand, the logit or probit specification will not be correctly specified if we fail to include relevant interaction and higher-order terms, as in the case of “rich covariates.” To test for parametric misspecification of the logit or probit model for the instrument propensity score, one may use the extension of RESET proposed by Papke and Wooldridge (1996). Here, we need to augment the original specification with several higher-order terms of the fitted linear index and, again, test for their joint significance.

Table 4 illustrates the problem of parametric misspecification. Panel A revisits the estimates of β_{IV} and β_{LATE} in Table 3, and reports the RESET p -values associated with each.

Table 4: Revisiting Causal Effects of 401(k) on IRA Participation

A. Original Specification	$\hat{\beta}_{IV}$	$\hat{\beta}_{LATE}$ (logit)	$\hat{\beta}_{LATE}$ (probit)
	0.0274 (0.0132)	0.0165 (0.0135)	0.0175 (0.0136)
RESET p -Value	0.000	0.000	0.000
B. Flexible Specification	$\hat{\beta}_{IV}$	$\hat{\beta}_{LATE}$ (logit)	$\hat{\beta}_{LATE}$ (probit)
	0.0167 (0.0132)	0.0177 (0.0128)	0.0178 (0.0128)
RESET p -Value	0.346	0.153	0.245

Notes: The data are Abadie’s (2003) subsample of the 1991 Survey of Income and Program Participation (SIPP). The sample size is 9,275. The outcome is an indicator for participation in IRAs, as in Table 3 of Abadie (2003). The original covariate specification in Panel A includes family income, age, age squared, marital status, and family size. The flexible covariate specification in Panel B includes a quintic in family income, a quintic in age, an indicator for marital status, and indicators for each value of family size. $\hat{\beta}_{LATE}$ corresponds to the normalized IPW estimator of Uysal (2011) and Słoczyński et al. (2025), with logit or probit instrument propensity scores estimated using maximum likelihood. RESET is based on a quartic in fitted values from each initial model of the instrument propensity score. Heteroskedasticity-robust standard errors are in parentheses.

With the original covariate specification, we decidedly reject the “rich covariates” assumption but also the correct specification of the logit and probit models for the instrument propensity score. Can a more flexible specification salvage a causal interpretation of our estimates? Panel B of Table 4 uses a quintic in family income, a quintic in age, an indicator for marital status, and indicators for each value of family size. With this set of covariates, RESET does not reject any of the null hypotheses of correct specification of the linear probability, logit, and probit models. $\hat{\beta}_{IV}$ and $\hat{\beta}_{LATE}$ are now very similar to each other, although $\hat{\beta}_{LATE}$ has barely changed relative to Panel A while $\hat{\beta}_{IV}$ has decreased by about 40%.

An interesting question is whether this empirical illustration correctly indicates that $\hat{\beta}_{IV}$ may be less robust to parametric misspecification than estimators of β_{LATE} . While we are not aware of any such results for generic IPW estimators, there are other parametric approaches to estimate β_{LATE} with improved robustness properties, such as the covariate balancing estimators of Heiler (2022), Sant’Anna et al. (2022), and Słoczyński et al. (2025), and the “doubly robust” estimators of Tan (2006), Uysal (2011), Słoczyński et al. (2022), and Ma et al. (2026). Several of these estimators are implemented in the Stata package `drlate` (Słoczyński et al., 2022). We recommend these estimators for wider use, together with their machine learning counterparts discussed below.

Double Machine Learning

The previous subsection explained that a “weakly causal” interpretation of β_{IV} and β_{LATE} may fail when the conditional mean of the instrument given covariates is misspecified, perhaps due to omitted interaction or higher-order terms. One practical response is to manually enrich the covariate specification and then apply RESET. In many empirical applications, however, the number of potential interactions and nonlinear transformations is large, making manual specification and specification testing cumbersome.

A complementary approach is to use double/debiased machine learning (DML) to flexibly approximate the relevant conditional means. DML methods for estimating β_{LATE} have been developed by Belloni et al. (2017), Chernozhukov et al. (2018), Sun and Tan (2022), and Singh and Sun (2024), whereas Chernozhukov et al. (2018) also applied the DML methodology to estimating β_{IV} . An application of double/debiased machine learning in an IV context proceeds as follows. The researcher supplies a rich dictionary of covariates to a machine learning algorithm (e.g., lasso, ridge, or regression tree). This method then approximates the relevant nonlinear relationships—the conditional means of potential outcomes and potential treatments, as well as that of the instrument—in a data-driven way, with tuning parameters typically chosen by cross-validation. Finally, DML combines the resulting flexible predictions with orthogonal scores and cross-fitting to decrease the underlying regularization bias. In this sense, DML can be viewed as an automated approach to constructing “rich” covariate specifications, reducing the need for ad hoc choices of interaction and higher-order terms.

Relative to manually specified parametric models, restrictions of DML are less explicit and arise implicitly from the choice of the machine learning algorithm. For example, lasso assumes that the target functions can be well approximated by a sparse subset of the dictionary, while tree-based methods rely on recursive partitioning and local averaging. Thus, even in the DML framework, the validity of the researcher’s conclusions continues to hinge on the chosen approximation. Accordingly, it is still recommended to compare multiple machine learning algorithms to assess robustness and to reduce reliance on the approximation properties, and thus the implicit parametric restrictions, of any single method.

Table 5 illustrates these considerations by reporting DML estimates of β_{IV} and β_{LATE} in the 401(k) application considered above. In the case of both parameters, we estimate the relevant conditional means using six different machine learning algorithms: lasso, elastic net, ridge, random forest, regression tree, and XGBoost. The resulting estimates are generally smaller than the initial linear IV estimate in Table 3 but larger than the corresponding LATE estimates. At the same time, the DML estimates remain somewhat sensitive to the choice of the machine learning algorithm. Thus, while DML relaxes explicit functional form assumptions, it is still advisable to verify robustness across alternative approximation

Table 5: DML Estimates of Causal Effects of 401(k) on IRA Participation

	$\hat{\beta}_{IV}$	$\hat{\beta}_{LATE}$
Lasso	0.0182 (0.0132)	0.0194 (0.0128)
Elastic Net	0.0184 (0.0132)	0.0224 (0.0152)
Ridge	0.0213 (0.0132)	0.0212 (0.0130)
Random Forest	0.0201 (0.0132)	0.0223 (0.0127)
Regression Tree	0.0251 (0.0134)	0.0245 (0.0131)
XGBoost	0.0366 (0.0134)	0.0330 (0.0126)

Notes: The data are Abadie’s (2003) subsample of the 1991 Survey of Income and Program Participation (SIPP). The sample size is 9,275. The outcome is an indicator for participation in IRAs, as in Table 3 of Abadie (2003). $\hat{\beta}_{IV}$ is based on the partially linear IV model in the `DoubleML` package in R. $\hat{\beta}_{LATE}$ is based on the interactive IV model. The covariate dictionary corresponds to the flexible specification in Panel B of Table 4. All models use 3 folds. Instrument propensity scores are trimmed at 0.01 and 0.99 to avoid near-zero denominators. The compliance structure assumes no always-takers and the presence of never-takers, consistent with the 401(k) setting. Lasso, elastic net, and ridge select the penalty parameter λ by 5-fold cross-validation. For random forest, regression tree, and XGBoost, hyperparameters are set separately for each nuisance function following the `DoubleML` 401(k) example (docs.doubleml.org). Standard errors are in parentheses.

methods (as we do in Table 5). We recommend researchers routinely report estimates using at least two or three different algorithms to assess the sensitivity of their conclusions.

Panel B of Appendix Table A1 lists the leading R and Stata packages that can be used to implement DML methods.

Assumption Violations

In this section we advise practitioners to systematically assess whether the assumptions underlying the LATE framework may be violated. We begin by reviewing several testable implications of these assumptions, and explain how the resulting statistical tests can be implemented in practice. We argue that using such tests can give more credence to IV applications. We also discuss an intuitive estimation approach that is more robust to violations of monotonicity, which is a controversial assumption in many applications. The validity of this assumption may also be rejected by one of the statistical tests that we review, in which case such alternative estimation approaches are particularly useful.

Testing Instrument Validity

As first observed by Balke and Pearl (1997), the standard LATE assumptions—specifically, independence of the instrument, exclusion restriction, and monotonicity—have implications that can be used to test instrument validity. If the LATE assumptions hold, then it must be the case that

$$\mathbb{P}(Y_i \in A, D_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i \in A, D_i = 1 \mid Z_i = 0) \geq 0 \quad (9)$$

and

$$\mathbb{P}(Y_i \in A, D_i = 0 \mid Z_i = 0) - \mathbb{P}(Y_i \in A, D_i = 0 \mid Z_i = 1) \geq 0 \quad (10)$$

for any subset A of outcome values (e.g., any interval or range).⁵ Because the inequalities in (9) and (10) are defined in terms of moments of observed variables, they can be examined with empirical data. Kitagawa (2015) derived a formal statistical test based on these inequalities, and proved that they are sharp (i.e., cannot be improved upon) as well as necessary but not sufficient for instrument validity. In other words, we may be able to reject invalid instruments, but it is fundamentally impossible to confirm instrument validity, even with infinite data. In a subsequent contribution, Huber and Mellace (2015) developed a sharp test of weaker assumptions that still suffice to identify the LATE. Mourifié and Wan (2017) reformulated the inequalities in (9) and (10) as two conditional moment inequalities, with Y_i entering as a conditioning variable, which facilitates implementation. Sun (2023) and Kwon and Roth (2026) extended this framework to the case of multivalued treatments.

A careful reader might ask: *Why* is it the case that the inequalities in (9) and (10) must hold whenever an instrument is valid? One answer to this question follows from Imbens and Rubin (1997), who proved that the difference in (9) identifies the joint distribution of the treated outcome and complier status, $\mathbb{P}(Y_i(1) \in A, D_i(1) > D_i(0))$, while the difference in (10) identifies $\mathbb{P}(Y_i(0) \in A, D_i(1) > D_i(0))$, the joint distribution of the untreated outcome and being a complier. Because these probabilities, like any other probabilities, must be nonnegative, the differences in (9) and (10) must be nonnegative, too.

Useful intuition for these testable implications is also provided by Kwon and Roth (2026). Note that the difference in (9) is simply the coefficient on Z_i in the population regression of the compound outcome $1[Y_i \in A, D_i = 1]$ on Z_i . If the instrument Z_i is valid and there are no defiers, any difference can only be driven by always-takers, never-takers, and compliers. In the case of always-takers and never-takers, Z_i has no effect on D_i . If so, then the exclusion restriction implies that it also has no effect on Y_i , ensuring that the effect of Z_i on the

⁵Formally, the inequalities in (9) and (10) must hold for any Borel set A .

compound outcome is zero for both groups. It follows that the entire difference in (9) is driven by compliers. However, in the case of compliers, $Z_i = 0$ implies that $D_i = 0$, which means that $\mathbb{P}(Y_i \in A, D_i = 1 \mid Z_i = 1) - \mathbb{P}(Y_i \in A, D_i = 1 \mid Z_i = 0) = \mathbb{P}(Y_i \in A, D_i = 1 \mid Z_i = 1)$. This probability, too, cannot be negative, which is exactly what the inequality in (9) requires. A similar intuition justifies the inequality in (10).

Statistical tests based on the Balke and Pearl conditions may seem impractical in some applications, especially when the instrument is only valid conditional on covariates.⁶ In such cases, the covariates enter the inequalities in (9) and (10) as conditioning variables, which increases the computational burden. To facilitate implementation, Farbmacher et al. (2022) proposed testing the Balke and Pearl conditions using machine learning. Specifically, their procedure uses causal forests to estimate the conditional average treatment effects of Z_i on the compound outcomes $1[Y_i \in A, D_i = 1]$ and $-1[Y_i \in A, D_i = 0]$, and subsequently selects groups with probable assumption violations using regression trees. With sample splitting, group selection in one subsample still enables “honest” testing in another.

One limitation of the procedure in Farbmacher et al. (2022) is that it only considers the Balke and Pearl conditions, which require discretizing the outcome variable to construct the compound outcomes. This would not be necessary, for example, when testing the Mourifié and Wan conditions. It also seems reasonable to use a similar machine-learning-based approach to construct a formal test of the implication of independence and monotonicity that the conditional first stage, $\mathbb{E}(D_i \mid Z_i = 1, X_i) - \mathbb{E}(D_i \mid Z_i = 0, X_i)$, must be nonnegative at all covariate values.⁷ These testable implications of the LATE assumptions, as well as other features, are implemented in the R package `montest` (Andresen, 2026).

Table 6 reports the resulting p -values for three IV applications: the Oregon Health Insurance Experiment (OHIE), as in Tables 1 and 2, the 401(k) application of Abadie (2003), as in Tables 3–5, and the study of causal effects of pretrial detention on case outcomes in Stevenson (2018). This last application uses data on more than 300,000 Philadelphia arrests between 2006 and 2013, and instruments for pretrial detention with indicators for randomly assigned judges. Following the replication of Stevenson (2018) in Słoczyński (2026), we use incarceration length as the outcome variable as well as a saturated covariate specification with indicators for each combination of offense type, race and gender of the defendant, and three time periods considered by Stevenson (2018). We also focus on a binary instrument

⁶As an alternative, Mao and Sant’Anna (2020) and Carr and Kitagawa (2023) show how to incorporate covariates to test the validity of the identifying assumptions in the MTE framework of Carneiro et al. (2011).

⁷This implication is sometimes tested by practitioners, especially in applications of the “judge leniency” design, but the choice of groups to examine is typically ad hoc. See, for example, Maestas et al. (2013), Dobbie et al. (2018), and Autor et al. (2019). An appropriate procedure based on machine learning can systematically search for groups where violations of monotonicity are most likely to be found.

Table 6: p -Values for Tests of Instrument Validity

	BP	MW	FS
Finkelstein et al. (2012)	0.99996	0.971	0.9999998
Abadie (2003)	0.997	0.9999996	N/A
Stevenson (2018)	0.000102	0.0241	0.000226

Notes: The source of data and variable choice for Finkelstein et al. (2012) are the same as in Tables 1–2. The source of data and variable choice for Abadie (2003) are the same as in Tables 3–5. For Stevenson (2018), the data are a sample of 331,971 arrests in Philadelphia. The outcome is incarceration length, defined as the maximum days of an incarceration sentence. The treatment is pretrial detention. The instrument is whether a given case was heard by the most lenient judge, referred to as “Judge C” in Słoczyński (2026). The covariate specification is saturated in the 17 most common offense types, race and gender of the defendant, and three time periods considered by Stevenson (2018). Groups with fewer than three cases heard by “Judge C” or not heard by “Judge C” are dropped. “BP” refers to the Balke and Pearl conditions. “MW” refers to the Mourifié and Wan conditions. “FS” refers to the requirement that the conditional first stage is nonnegative at all covariate values. All tests are implemented using the `montest` command in R.

that indicates whether a given case was assigned to the most lenient judge or not.

The p -values in Table 6 clearly indicate that there is fundamentally no evidence against instrument validity in the first two applications, regardless of whether we consider the Balke and Pearl (“BP”) conditions, the Mourifié and Wan (“MW”) conditions, or the nonnegativity of the conditional first stage (“FS”). In fact, this last condition is trivially satisfied in the 401(k) application of Abadie (2003), which is characterized by one-sided noncompliance. At the same time, we clearly reject the null hypothesis of instrument validity using the data from Stevenson (2018). The conclusion is essentially the same whether we consider the testable implications of independence, exclusion, and monotonicity (BP and MW) or the testable implications of independence and monotonicity (FS). Assuming that independence is satisfied, which is plausible, this implies that either monotonicity is violated but exclusion is satisfied, or both monotonicity and exclusion are violated. Our discussion in the next subsection will implicitly assume the former possibility.

Panel C of Appendix Table A1 lists several R packages, including `montest`, which can be used to implement various tests of the LATE assumptions.

Robustness to Monotonicity Violations

If the monotonicity assumption is rejected or otherwise implausible, as in Stevenson (2018), estimation of β_{LATE} will be difficult, and the previously discussed estimators of β_{LATE} and β_{IV} will not consistently estimate any “weakly causal” target parameter. At the same time, appropriate estimators of β_{AI} will not be subject to such concerns, at least when a weaker assumption, termed “weak monotonicity” in Słoczyński (2026), is assumed to hold. Weak monotonicity, unlike strong monotonicity (i.e., the usual assumption), allows both compliers

and defiers to exist, but they must not coexist at any value of covariates; that is, weak monotonicity requires that there be no defiers at some covariate values and no compliers elsewhere. This assumption will be plausible in some applications, but not in others. In Stevenson (2018), it seems reasonable that only a small number of observed characteristics of cases and defendants would determine the relative leniency of any particular judge.

To see why standard estimators of β_{LATE} and β_{IV} will not work under weak monotonicity, note that the term π_j in the weighted average representations in (6)–(8) only retains its interpretation as the conditional proportion of compliers under strong monotonicity. In the absence of any assumptions, it is simply equal to $\mathbb{E}(D_i \mid Z_i = 1, X_i) - \mathbb{E}(D_i \mid Z_i = 0, X_i)$, that is, the conditional first stage. (Under strong monotonicity, the conditional first stage identifies the conditional proportion of compliers.) However, under weak monotonicity, the conditional first stage will be positive at some covariate values and negative at others, which directly translates to the incidence of “negative weights” in (6) and (8). On the other hand, π_j enters the expression in (7) as a quadratic rather than linearly, which ensures that the resulting weights are positive even if some conditional first stages are not. Słoczyński (2026) argues that this makes focusing on β_{AI} a useful compromise under weak monotonicity.

However, focusing on β_{AI} is not without drawbacks. Because, by construction, β_{AI} uses multiple instruments, its estimation will be subject to many-instrument bias, which results from overfitting the first stage. Recent surveys of the literature on many instruments include Anatolyev (2019) and Mikusheva and Sun (2024). The standard response to this problem is to estimate the specification in equations (3)–(4) with an alternative estimator, such as the fixed effect jackknife IV (FEJIV) estimator of Chao et al. (2023) or the unbiased jackknife IV estimator (UJIVE) of Kolesár (2013) and Goldsmith-Pinkham et al. (2025). Both of these estimators, unlike 2SLS, are consistent under the asymptotic sequence that allows the number of instruments and the number of covariates to increase in proportion with the sample size. In addition, Mikusheva and Sun (2022) propose a pretest for weak identification that can be used to determine whether multiple instruments are jointly strong enough for consistent estimation. While the pretest is not formally designed for settings with many covariates, the simulation results in Słoczyński (2026) suggest that it effectively distinguishes between settings in which FEJIV and UJIVE perform well and those in which all estimators of the specification in equations (3)–(4) perform poorly.

Table 7 replicates several of the estimates of causal effects of pretrial detention on incarceration length in Table 8 in Słoczyński (2026). The specification is the same as in Table 6 above. The linear IV estimate suggests that pretrial detention leads to a large, statistically significant increase in incarceration length of about 666 days. We know, however, that the underlying estimand is not “weakly causal” because of violations of (strong) monotonicity.

Table 7: Causal Effects of Pretrial Detention on Incarceration Length

IV	2SLS	FEJIV	UJIVE
666	130	51	56
(233)	(43)	(91)	(99)

Notes: The data are Stevenson’s (2018) sample of 331,971 arrests in Philadelphia. The outcome is incarceration length, defined as the maximum days of an incarceration sentence. The treatment is pre-trial detention. The instrument is whether a given case was heard by the most lenient judge, referred to as “Judge C” in Słoczyński (2026). The covariate specification is saturated in the 17 most common offense types, race and gender of the defendant, and three time periods considered by Stevenson (2018). Groups with fewer than three cases heard by “Judge C” or not heard by “Judge C” are dropped. Standard errors are in parentheses.

Indeed, the 2SLS estimate of β_{AI} , which eliminates the negative weights, is 80% smaller than the linear IV estimate; it also remains statistically significant. While this estimate might be more believable, it suffers from many-instrument bias, unlike the associated FEJIV and UJIVE estimates. These, however, only suggest a small increase in incarceration length of less than 2 months, and are not statistically different from zero. This conclusion contrasts sharply with the initial IV and 2SLS estimates.⁸ It is also supported by the fact that Mikusheva and Sun (2022)’s pretest rejects weak identification, as reported by Słoczyński (2026), which indicates that consistent estimation is possible.

Panel D of Appendix Table A1 lists several MATLAB, R, and Stata packages that can be used to implement FEJIV, UJIVE, and other relevant estimators, as well as the pretest in Mikusheva and Sun (2022).

Conclusion

The local average treatment effect framework of Angrist and Imbens has fundamentally reshaped how applied microeconomists think about instrumental variables. It is now common to acknowledge that IV models only identify average treatment effects for compliers, examine the characteristics of that subpopulation, and develop careful arguments in favor of instrument validity. In this paper, we have explored three areas in which we believe empirical practice could further benefit from closer engagement with the recent theoretical literature.

⁸To be clear, Stevenson (2018) does not use either of the problematic estimation approaches as her method of choice. Instead, she uses a nonsaturated specification with a large number of interacted instruments, and estimates the model using the jackknife IV estimator (JIVE) of Angrist et al. (1999). Unlike FEJIV and UJIVE, this estimator is not formally appropriate for settings with many covariates.

As we outline above, we recommend that practitioners explicitly target the “true” LATE rather than the usual IV estimand, consider flexible functional forms to avoid parametric misspecification, and use formal tools in response to possible assumption violations, including statistical tests of instrument validity and specifications with many interacted instruments.

References

- Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263.
- Anatolyev, S. (2019). Many instruments and/or regressors: A friendly guide. *Journal of Economic Surveys*, 33(2):689–726.
- Andresen, M. E. (2026). montest: Testing LATE assumptions and monotonicity using machine learning. <https://github.com/martin-andresen/montest>.
- Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.
- Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.
- Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, Princeton–Oxford.
- Autor, D., Kostøl, A., Mogstad, M., and Setzler, B. (2019). Disability benefits, consumption insurance, and household labor supply. *American Economic Review*, 109(7):2613–2654.
- Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.
- Blandhol, C., Bonney, J., Mogstad, M., and Torgovitsky, A. (2026). When is TSLS actually LATE? *Review of Economic Studies*, forthcoming.
- Borusyak, K., Hull, P., and Jaravel, X. (2025). A practical guide to shift-share instruments. *Journal of Economic Perspectives*, 39(1):181–204.
- Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, 101(6):2754–2781.
- Carr, T. and Kitagawa, T. (2023). Testing instrument validity with covariates. arXiv:2112.08092.
- Chao, J. C., Swanson, N. R., and Woutersen, T. (2023). Jackknife estimation of a cluster-sample IV regression model with many weak instruments. *Journal of Econometrics*, 235(2):1747–1769.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68.
- Chyn, E., Frandsen, B., and Leslie, E. (2025). Examiner and judge designs in economics: A practitioner’s guide. *Journal of Economic Literature*, 63(2):401–439.

- Currie, J., Kleven, H., and Zwiers, E. (2020). Technology and big data are changing economics: Mining text to track methods. *AEA Papers and Proceedings*, 110:42–48.
- Dobbie, W., Goldin, J., and Yang, C. S. (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–240.
- Farbmacher, H., Guber, R., and Klaassen, S. (2022). Instrument validity tests with causal forests. *Journal of Business & Economic Statistics*, 40(2):605–614.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Oregon Health Study Group (2012). The Oregon Health Insurance Experiment: Evidence from the first year. *Quarterly Journal of Economics*, 127(3):1057–1106.
- Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75.
- Goldsmith-Pinkham, P. (2026). Tracking the credibility revolution across fields. NBER Working Paper no. 35051.
- Goldsmith-Pinkham, P., Hull, P., and Kolesár, M. (2025). Leniency designs: An operator’s manual. NBER Working Paper no. 34473.
- Heckman, J. J. and Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738.
- Heiler, P. (2022). Efficient covariate balancing for the local average treatment effect. *Journal of Business & Economic Statistics*, 40(4):1569–1582.
- Huber, M. and Mellace, G. (2015). Testing instrument validity for LATE identification based on inequality moment constraints. *Review of Economics and Statistics*, 97(2):398–411.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.
- Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, 64(4):555–574.
- Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83(5):2043–2063.
- Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. Unpublished.
- Kwon, S. and Roth, J. (2026). Testing mechanisms. *Review of Economic Studies*, forthcoming.
- Ma, Y., Sant’Anna, P. H. C., Sasaki, Y., and Ura, T. (2026). Doubly robust estimators with weak overlap. arXiv:2304.08974.
- Maestas, N., Mullen, K. J., and Strand, A. (2013). Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review*, 103(5):1797–1829.
- Mao, M. and Sant’Anna, P. H. C. (2020). Testing instrument validity in marginal treatment effects models. Unpublished.
- Mikusheva, A. and Sun, L. (2022). Inference with many weak instruments. *Review of Economic Studies*, 89(5):2663–2686.
- Mikusheva, A. and Sun, L. (2024). Weak identification with many instruments. *Econometrics Journal*, 27(2):C1–C28.
- Mogstad, M. and Torgovitsky, A. (2018). Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics*, 10:577–613.

- Mogstad, M. and Torgovitsky, A. (2024). Instrumental variables with unobserved heterogeneity in treatment effects. In Dustmann, C. and Lemieux, T., editors, *Handbook of Labor Economics, Vol. 5*, pages 1–114. Elsevier, Amsterdam.
- Mourifié, I. and Wan, Y. (2017). Testing local average treatment effect assumptions. *Review of Economics and Statistics*, 99(2):305–313.
- Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11(6):619–632.
- Poirier, A. and Słoczyński, T. (2025). Quantifying the internal validity of weighted estimands. arXiv:2404.14603.
- Poterba, J. M., Venti, S. F., and Wise, D. A. (1994). 401(k) plans and tax-deferred saving. In Wise, D. A., editor, *Studies in the Economics of Aging*, pages 105–142. University of Chicago Press, Chicago–London.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B*, 31(2):350–371.
- Sant’Anna, P. H. C., Song, X., and Xu, Q. (2022). Covariate distribution balance via propensity scores. *Journal of Applied Econometrics*, 37(6):1093–1120.
- Singh, R. and Sun, L. (2024). Double robustness for complier parameters and a semi-parametric test for complier characteristics. *Econometrics Journal*, 27(1):1–20.
- Słoczyński, T. (2026). When should we (not) interpret linear IV estimands as LATE? *Review of Economic Studies*, forthcoming.
- Słoczyński, T., Uysal, S. D., and Wooldridge, J. M. (2022). Doubly robust estimation of local average treatment effects using inverse probability weighted regression adjustment. arXiv:2208.01300.
- Słoczyński, T., Uysal, S. D., and Wooldridge, J. M. (2025). Abadie’s kappa and weighting estimators of the local average treatment effect. *Journal of Business & Economic Statistics*, 43(1):164–177.
- Stevenson, M. T. (2018). Distortion of justice: How the inability to pay bail affects case outcomes. *Journal of Law, Economics, and Organization*, 34(4):511–542.
- Sun, B. and Tan, Z. (2022). High-dimensional model-assisted inference for local average treatment effects with instrumental variables. *Journal of Business & Economic Statistics*, 40(4):1732–1744.
- Sun, Z. (2023). Instrument validity for heterogeneous causal effects. *Journal of Econometrics*, 237(2, Part A):105523.
- Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618.
- Uysal, S. D. (2011). Doubly robust IV estimation of the local average treatment effect. Unpublished.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge–London, 2nd edition.

Appendix Table A1: Selected Software Packages in MATLAB, R, and Stata

Package Name	Software	Authors	Description
A. Estimating LATE			
<code>causalweight</code>	R	H. Bodory and M. Huber	IPW estimation of the LATE using function <code>lateweight</code>
<code>drlate</code>	Stata	T. Słoczyński, S. D. Uysal, and J. M. Wooldridge	Regression adjustment, IPW, and doubly robust estimation of the LATE
<code>kappalate</code>	Stata	T. Słoczyński, S. D. Uysal, and J. M. Wooldridge	IPW estimation of the LATE
<code>lateffects</code>	Stata	StataCorp	IPW and doubly robust estimation of the LATE
B. Double Machine Learning			
<code>ddml</code>	Stata	A. Ahrens, C. B. Hansen, M. E. Schaffer, and T. Wiemann	Double/debiased machine learning estimation of IV and LATE parameters
<code>DoubleML</code>	R	P. Bach, M. S. Kurz, V. Chernozhukov, M. Spindler, and S. Klaassen	Double/debiased machine learning estimation of IV and LATE parameters
C. Testing LATE Assumptions			
<code>LATEtest</code>	R	H. Farbmacher	Tests of the LATE assumptions based on machine learning
<code>montest</code>	R	M. E. Andresen	Tests of monotonicity and LATE assumptions based on machine learning
<code>TestMechs</code>	R	S. Kwon and J. Roth	Tests of mediation mechanisms and related testable implications of LATE assumptions
D. Estimation with Many Instruments			
<code>fejiv</code>	MATLAB, R, and Stata	Q. Lei and T. Słoczyński	Fixed-effect jackknife IV estimation with many instruments
<code>manyiv</code>	Stata	M. Caceres Bravo, P. Goldsmith-Pinkham, P. Hull, and M. Kolesár	IV estimation and inference with many instruments
<code>manyweakiv</code>	Stata	L. Sun	Pretest for weak identification with many instruments using command <code>manyweakivpretest</code>