Discussion Paper Series

IZA DP No. 18573

April 2026

Testing IV Validity and LATE Interpretation Using Flexible Covariate Specifications

Anna Krumme
FernUniversität in Hagen and
TU Dortmund

Matthias Westphal
FernUniversität in Hagen,
RWI Essen and IZA@LISER

IZA DP No. 18573 April 2026

Testing IV Validity and LATE Interpretation Using Flexible Covariate Specifications*

Abstract

Building on the testable implications for IV validity underlying local average treatment effect (LATE) estimation, we (i) propose a simple testing procedure that may accommodate high-dimensional covariates and (ii) demonstrate that it can also detect biases arising from misspecified IV regression models. While recent research has highlighted the importance of a correct covariate specification, existing IV validity tests are not designed to capture this source of bias. Simulation studies strongly suggest that the test performs well at detecting violations of conditional independence, violations of the exclusion restriction, and biases arising from covariate misspecification.

JEL classification

C12, C21, C26, C52

Keywords

testing instrument validity, local average treatment effects, covariate misspecification

Corresponding author

Matthias Westphal
matthias.westphal@fernuni-hagen.de

* We thank Andreas Fischeneder, Kai Miele, Hendrik Schmitz, Jeffrey Wooldridge, the participants in the Annual Conference of the International Association for Applied Econometrics 2025, the 3rd CINCH-dggö Academy in Health Economics, the internal RWI Health Economics Workshop, and the CESA Brown Bag Seminar for their helpful comments and discussions.

1 Introduction

Instrumental variables (IV) research designs are central to modern causal inference and provide a powerful lens for uncovering heterogeneity in economic behavior (Mogstad and Torgovitsky, 2024). Valid IVs must satisfy well-understood requirements: they must be exogenous, affect the outcome only through the treatment, and have a monotonic effect on the treatment in only one direction. In addition to the well-understood requirements, Blandhol et al. (2026) also document an additional assumption for IV regression models with covariates: "rich covariates", that is, specifications that approach perfectly saturated specifications. Although some tests for the general IV validity assumptions exist, they are rarely applied. Beyond the computational burden and technical challenges of non- or semiparametric approaches, a key reason for their limited uptake is their difficulty accommodating covariates flexibly without incurring the curse of dimensionality as the number of covariates increases.

This paper proposes the first test of instrumental-variables (IV) validity that accommodates covariates directly using flexible covariate specifications, without requiring a nonparametric pre-processing step. The test builds on the nonparametric testable implications derived by Huber and Mellace (2015), but implements them using linear-in-parameters conditional distribution regressions. Their use allows us to incorporate covariates transparently and to derive bounds for non-complying subpopulations within a tractable parametric framework. Our paper further links IV validity testing to recent work by Blandhol et al. (2026), who show that IV estimates may fail to be even weakly causal when covariates are misspecified, potentially reversing sign relative to the true effect.[1] We show that the same misspecification contaminates not only conventional IV estimates but also the testable implications underlying our procedure. This observation enables our test to detect specification errors that directly affect the Wald estimand. Moreover, because our approach can be combined with flexible estimation methods—such as partially linear IV models estimated via double/debiased machine learning (Chernozhukov et al., 2018)—it provides a practical framework for distinguishing between violations of IV validity and covariate misspecification. Overall, the paper contributes by integrating conditional distribution regressions with IV validity testing and incorporating recent insights on specification bias to provide a tractable, empirically implementable diagnostic for applied work.

The previous literature on IV validity has proposed two strands of tests: mean-based testable implications (Huber and Mellace, 2015) and the density-based conditions (Kitagawa, 2015). For mean effects such as the local average treatment effect (LATE), the test of Huber and Mellace (2015) is optimal to refute IV validity (Laffers and Mellace, 2017). However, this is a nonparametric test that may be applied to a specific cell of covariates,

[1] Blandhol et al. (2026) characterize a weakly causal two-stage least squares estimand as a weighted average of subgroup-specific treatment effects with non-negative weights.

but not to regression models with richer functional forms. Further literature mainly builds on the density-based approach. Mourifié and Wan (2017) reformulate the testable conditions and propose another testing procedure, Sun (2023) improves the Kitagawa (2015) test procedure and allows the treatment to be multivalued, and Arai et al. (2022) extends the density-based approach to fuzzy regression discontinuity designs. An alternative approach to test the density-based conditions is provided by Farbmacher et al. (2022), who uses causal forests to detect local violations of the LATE assumptions.[2]

All these tests are nonparametric, including the one in Farbmacher et al. (2022), which can, however, be applied flexibly across many different covariate cells. One exception is Carr and Kitagawa (2023) who extend Kitagawa's (2015) test to the marginal treatment effect framework, thereby accommodating a larger number of covariates. This test, however, requires semiparametric MTE estimation, which is computationally demanding with high-dimensional covariates and additionally requires either a strong continuous instrument, additional functional-form assumptions, or partial-identification approaches. Thus, the inability to control for a larger number of covariates in most proposed validity tests, while computationally feasible, limits their practical applicability. This is particularly true in settings where the exogeneity assumption holds only conditionally on various covariates (e.g., models with fixed effects). The traditional nonparametric tests suffer from the curse of dimensionality and quickly become impractical as the number of covariates increases. Moreover, non-parametric conditioning may not be feasible in every research design: models with two-way fixed effects, for example, are inherently based on an additive, non-saturated regression model (even in the simple 2×2 case). Thus, it is important to test IV validity within the same parametric model. However, parametric specifications amplify the risk of specification bias (De Chaisemartin and d'Haultfoeuille, 2020; Goldsmith-Pinkham et al., 2024), also in IV settings (Blandhol et al., 2026). A test that jointly assesses specification bias and traditional IV validity within conventional regression models thus represents an important methodological advance.

Our testing approach is most closely related to that of Huber and Mellace (2015) but differs in two main ways. First, we reduce the number of conditions tested from 4 to 2 by excluding non-binding conditions. Second, even though the testing conditions are based on mean potential outcomes, we make use of group-specific conditional density functions for which the covariates are held fixed at the mean. This allows us to point-identify mean potential outcomes for pure groups and partially identify bounds for the unobserved mean potential outcomes of mixed groups, conditional on covariates. Conditioning on covariates using the approach of Huber and Mellace (2015) is limited, as it requires running the procedure on covariate-specific subsamples.

[2]There are other papers in the literature that concentrate on violations of one or two of the validity assumptions, e.g. Angrist and Imbens (1995), Mogstad et al. (2021), Machado et al. (2019), De Chaisemartin (2017) and Kédagni and Mourifié (2020).

We conduct two conceptually distinct simulation settings to evaluate the test's performance: one in which either the conditional independence assumption or the exclusion restriction is violated, and another featuring specification bias. Across both designs, the results indicate good finite-sample performance in terms of size and power. We complement the simulation results with two empirical applications from the recent literature. The first uses a randomized training intervention as an instrument for the active use of mobile banking accounts to estimate the LATE of mobile banking adoption on rural household outcomes (Lee et al., 2021). The second instruments female leadership with the gender of the firstborn child of the previous monarch to estimate the effect of queenly rule on conflict outcomes in historical Europe (Dube and Harish, 2020). For comparison with the IV validity testing literature, we additionally apply our procedure to two widely used empirical settings—the draft-eligibility instrument proposed by Angrist (1991) and the college proximity instrument from Card (1993)—with the results reported in the Appendix.

This paper proceeds as follows. Section 2 introduces the general econometric setup, and presents the LATE assumptions and their testable implications. The testing procedure is detailed in Section 3, and Section 4 analyzes the impact of specification bias on our testable implications. Section 5 presents the simulation results, before the results for the two empirical applications are shown in Section 6. Section 7 concludes.

2 Setting and Assumptions

With a binary treatment D and a binary instrument Z, the key estimator for causal inference on the outcome Y is the so-called Wald estimator

$$\text{IV}_{Wald} = \frac{\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)}{\mathbb{E}(D \mid Z = 1) - \mathbb{E}(D \mid Z = 0)}. \tag{1}$$

Note that we suppress covariates X in this simple setup—in Section 4 we explicitly discuss specification bias as introduced by Blandhol et al. (2026). Angrist and Imbens (1995) show that with an additional set of assumptions, this ratio of mean differences has a causal interpretation as the local average treatment effect (LATE):

$$\text{IV}_{Wald} = \mathbb{E}(Y^1 - Y^0 \mid D^1 > D^0) := LATE \tag{2}$$

Here, Y^d is the potential outcome for treatment state $d \in \{0, 1\}$. Hence, for every individual, $Y^1 - Y^0$ is their specific treatment effect. The LATE averages this individual treatment effect across a particular group of individuals—those who take the treatment because of the instrument. To derive this, Angrist and Imbens introduce another potential outcome dimension for the treatment: D^z, indicating the potential treatment choice with a specific value of the instrument $z \in \{0, 1\}$. Correspondingly, this second dimension can be

added to the outcome. Y^{dz} then indicates the potential outcome for treatment state d and instrument value z.

We will now introduce the assumptions necessary to go from Eq. (1) to (2), which sets the path for testable implications on these assumptions.[3]

Assumption 1 (Mean independence):
$\mathbb{E}(Y^{dz} \mid Z = 1) = \mathbb{E}(Y^{dz}|Z = 0)$ and $\mathbb{E}(D^z \mid Z = 1) = \mathbb{E}(D^z|Z = 0) \quad \forall d, z\{0, 1\}$

Remark 1. *As we identify mean effects, not quantiles or probability densities, we only need this mean independence. In contrast, density-based testing approaches following Kitagawa (2015), base their tests on the stronger assumption of full independence:* $Y^{d1}, Y^{d0}, D^1, D^0 \perp\!\!\!\perp Z$.

By the independence assumption, we can write for the numerator of Eq. (1):

$$\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0) = \mathbb{E}(Y^{d1} - Y^{d0}),$$

which simply is the causal effect of Z on Y (also called intent-to-treat or reduced-form effect). The expression $\mathbb{E}(Y^{d1} - Y^{d0})$ means that the treatment state d is unrestricted and may vary from individual to individual in this difference, whereas the instrument state z is fixed. Analogously, we can rearrange the denominator of Eq. (1) through the independence assumption as follows:

$$\mathbb{E}(D \mid Z = 1) - \mathbb{E}(D \mid Z = 0) = \mathbb{E}(D^1 - D^0)$$

We can then decompose the average causal effect of Z on D based on counterfactual treatment behavior.

$$\begin{aligned}
\mathbb{E}(D^1 - D^0) &= \Pr(D^1 = 1) - \Pr(D^0 = 1) \\
&= \Pr(D^1 = 1, D^0 = 1) + \Pr(D^1 = 1, D^0 = 0) \\
&\quad - \Big[\Pr(D^0 = 1, D^1 = 1) + \Pr(D^0 = 1, D^1 = 0)\Big]
\end{aligned}$$

In principle, we can define and label the four possible types as always-takers (AT, defined by $D^1 = D^0 = 1$), compliers (C, $D^1 > D^0$), defiers (DF, $D^1 < D^0 = 1$) and never-takers (NT, $D^1 = D^0 = 0$). With this compact notation, we can simplify the equation above as:

$$\pi_{AT} + \pi_C - \big[\pi_{AT} + \pi_{DF}\big] \quad = \quad \pi_C - \pi_{DF}$$

[3]Note that the testable implications can be adapted to the density-based conditions basing on stronger assumptions (see remarks on assumptions 1 and 2) as shown by Huber and Mellace (2015) in section VI. This, however, increases the computational burden of testing drastically .

Using these types, we can also decompose the numerator of Eq. (1) as

$$\begin{aligned}\mathbb{E}(Y^{d1} - Y^{d0}) &= \pi_{NT} E\big(Y^{01} - Y^{00} | D^1 = D^0 = 0\big) + \pi_{AT} E\big(Y^{11} - Y^{10} | D^1 = D^0 = 1\big) \\ &\quad + \pi_C E\big(Y^{11} - Y^{00} | D^1 = 1, D^0 = 0\big) + \pi_{DF} E\big(Y^{01} - Y^{10} | D^1 = 0, D^0 = 1\big)\end{aligned}$$

Conditional on the type, we only need the value of the instrument to infer treatment take-up. Thus, we use δ^z_{type} to denote the corresponding expected outcome. Then we write the above equation as:

$$\mathbb{E}(Y^{d1} - Y^{d0}) = \pi_{NT}\big[\delta^1_{NT} - \delta^0_{NT}\big] + \pi_{AT}\big[\delta^1_{AT} - \delta^0_{AT}\big] + \pi_C\big[\delta^1_C - \delta^0_C\big] + \pi_{DF}\big[\delta^1_{DF} - \delta^0_{DF}\big]$$

Now, we use this notion to rewrite Eq. (2) as:

$$\text{IV}_{Wald} = \frac{\pi_{NT}\big[\delta^1_{NT} - \delta^0_{NT}\big] + \pi_{AT}\big[\delta^1_{AT} - \delta^0_{AT}\big] + \pi_C\big[\delta^1_C - \delta^0_C\big] + \pi_{DF}\big[\delta^1_{DF} - \delta^0_{DF}\big]}{\pi_C - \pi_{DF}} \quad (3)$$

This expression is more complicated than Eq. (2). To give it the desired interpretation, we need to make additional assumptions.

Assumption 2 (Mean exclusion restriction):
$\mathbb{E}(Y^{d,1}) = \mathbb{E}(Y^{d,0})$ for $d \in \{0, 1\}$.

Remark 2. *Again, we only need the exclusion restriction to hold in expectation for the identification of mean effects. IV validity conditions of Kitagawa (2015) require $Y^{d,1} = Y^{d,0}$ for $d \in \{0, 1\}$.*

By the exclusion restriction, the instrument only affects Y through D, such that effects for always-takers and never-takers are nonexistent:

$$\text{IV}_{Wald} = \frac{\pi_C E\big(Y^{01} - Y^{00} | D^1 = 1, D^0 = 0\big) - \pi_{DF} E\big(Y^{10} - Y^{01} | D^1 = 0, D^0 = 1\big)}{\pi_C - \pi_{DF}} \quad (4)$$

The last step uses

Assumption 3 (Monotonicity): $Pr(D^1 \geq D^0) = 1$

By monotonicity, $\pi_{DF} = 0$, and the above expression simplifies to Eq. (2). Although the testable implications discussed in this paper may detect violations of assumptions 1–3, we will assume that monotonicity holds for notational clarity and because violations of

the monotonicity assumptions must be substantial to be detected.[4] We refer the reader to De Chaisemartin (2017) and Słoczyński (2025) for more details of such a violation.

The different types are not directly distinguishable in the data, but conditioning on both D and Z yields expectations in which only one or two types contribute. This insight, first used by Imbens and Rubin (1997), is the first step to seeing the consequences when an assumption is violated. For instance, if we condition on $D = 1$ and $Z = 1$, always-takers and compliers enter the expectation:

$$\mathbb{E}(Y \mid D = 1, Z = 1) = \frac{\pi_C}{\pi_C + \pi_{AT}} \delta^1_C + \frac{\pi_{AT}}{\pi_C + \pi_{AT}} \delta^1_{AT} \tag{5}$$

If $Z = 0$ in the treated case, always-takers exclusively enter the expectation:

$$\mathbb{E}(Y \mid D = 1, Z = 0) = \delta^0_{AT}$$

For the untreated case with $Z = 1$, only never-takers must contribute to the expectation:

$$\mathbb{E}(Y \mid D = 0, Z = 1) = \delta^1_{NT}$$

If $D = 0$ and $Z = 0$, the expectation is mixed with never-takers and compliers:

$$\mathbb{E}(Y \mid D = 0, Z = 0) = \frac{\pi_C}{\pi_C + \pi_{NT}} \delta^0_C + \frac{\pi_C}{\pi_C + \pi_{NT}} \delta^0_{NT} \tag{6}$$

3 Testable Implications, Testing Procedure, and Estimation

By assumptions 1–3, we take the always-takers' mean when $Z = 0$, δ^0_{AT}, and use Eq. (5) to infer the mean for the treated compliers, δ^1_C. This works, because the assumptions imply $\delta^0_{AT} = \delta^1_{AT}$. Analogously, we can use the never-takers' mean δ^1_{NT}, equate it to δ^0_{NT}, and infer the mean of the untreated compliers according to Eq. (6).

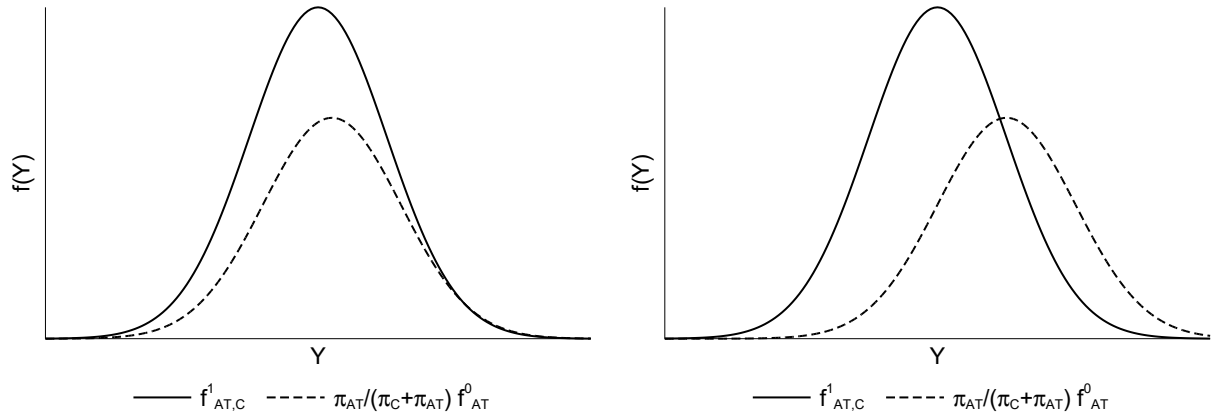
If either one of the assumptions does not hold, $\delta^1_{AT} \neq \delta^0_{AT}$ and/or $\delta^1_{NT} \neq \delta^0_{NT}$. We can test whether this is likely to be fulfilled by using the type and Z-specific probability distribution functions, $f^z_{type}(Y)$, together with the fact that the equations do not only need to hold in expectation but also in distribution. Hence, the two treated expectations become:

$$\begin{aligned} f(Y \mid D = 1, Z = 1) &= \frac{\pi_C}{\pi_C + \pi_{AT}} f^1_C(Y) + \frac{\pi_{AT}}{\pi_C + \pi_{AT}} f^1_{AT}(Y) \\ &:= f^1_{AT,C}(Y) \end{aligned}$$

[4] The presence of defiers can shift the distributions and potential outcome means for the groups with $D = 1$ and $Z = 0$ or $D = 0$ and $Z = 1$, thereby preventing identification of δ^0_{AT} and δ^1_{NT} within the testing procedure. However, if the proportion of defiers is sufficiently large, the test detects these shifts by revealing contamination in the estimates of δ^0_{AT} and δ^1_{NT}.

$$f(Y \mid D = 1, Z = 0) = f^0_{AT}(Y)$$

The implication of IV validity is $f^0_{AT}(Y) = f^1_{AT}(Y)$ for the treated case with $D = 1$ and $f^1_{NT}(Y) = f^0_{NT}(Y)$ for the untreated case with $D = 0$. This equivalence has testable implications for the observed distributions: the joint distributions of $f^1_{AT,C}(Y)$ and $f^0_{NT,C}(Y)$ must nest the normalized single-type distributions $\frac{\pi_{AT}}{\pi_{AT}+\pi_C} f^0_{AT}(Y)$ and $\frac{\pi_{NT}}{\pi_{NT}+\pi_C} f^1_{NT}(Y)$, respectively.



(a) No violation (b) Violation

Figure 1: Graphical test of IV validity

Notes: Own illustration.

Figure 1 visualizes the two possible scenarios in Panels (a) and (b). Panel (a) displays a type-specific density (dashed line) that is compatible with the joint density (solid line). The two densities do not cross. Panel (b) displays the case where the two densities do cross. Testing whether the densities cross is the idea behind the Kitagawa (2015) test, which assumes full conditional independence.

A similar (but not equivalent) implication of an incompatible distribution is that the mean of f^0_{AT} (δ^0_{AT}) must lie within the lower and upper bound of extreme-case scenarios. To keep notation and testing simple, we assume the outcome to be continuous.[5] The lowest possible mean of the unobserved δ^1_{AT} results from $f(Y|D = 1, Z = 1)$ if the unobserved always-takers are placed in the lowest possible ranks of the distribution. As we know the share of always-takers, the lowest possible ranks are the first $q = \frac{\pi_{AT}}{\pi_{AT}+\pi_C}$ quantiles. This extreme-case scenario assumes the always-takers occupy the entire quantile range from 0 to q in the joint distribution, resulting in the lowest possible mean, $\delta^{1,LB}_{AT}$. Formally, this reads

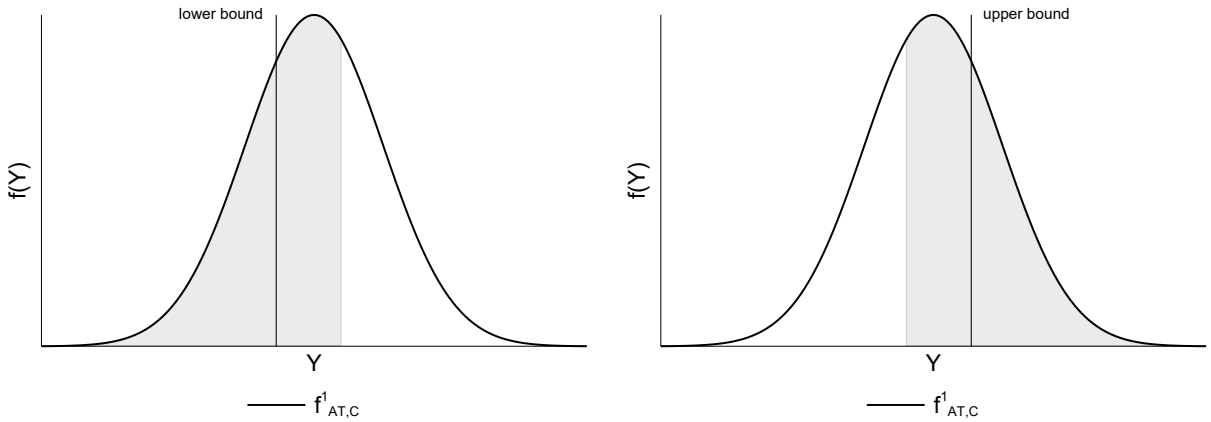
$$\delta^{1,LB}_{AT} = \int_0^{\frac{\pi_{AT}}{\pi_{AT}+\pi_C}} y \, dF(Y = y \mid D = 1, Z = 1). \tag{7}$$

[5]The approach can be adapted to discrete outcomes similarly to the Huber and Mellace (2015) test (see online appendix A.4 of Huber and Mellace, 2015).

The converse extreme case scenario is when the always-takers are placed in the highest q quantiles for the upper bound. This yields

$$\delta_{AT}^{1,UB} = \int_{\frac{\pi_C}{\pi_{AT}+\pi_C}}^{1} y \, dF(Y = y \mid D = 1, Z = 1). \tag{8}$$

Figure 2 visualizes the lower and upper bounds for the joint treated distribution, which is the mean produced by the gray part of the distribution.



(a) Lower bound (b) Upper bound

Figure 2: Graph of upper and lower bound

Notes: Own illustration. Shaded areas equal the q's proportion of the integral located in the lower (a) or upper (b) tail of the distribution. The vertical solid lines indicate the lower and upper bounds of $\mathbb{E}(Y^{1,1}|T^Z = AT^1, X)$.

For the untreated distributions, the extreme-case scenarios form if the never-takers place in the lowest or highest $r = \frac{\pi_{NT}}{\pi_{NT}+\pi_C}$ ranks of the joint $f_{NT,C}^0$ distribution.

$$\delta_{NT}^{0,LB} = \int_{0}^{\frac{\pi_{NT}}{\pi_{NT}+\pi_C}} y \, dF(Y = y \mid D = 0, Z = 0). \tag{9}$$

$$\delta_{NT}^{0,UB} = \int_{\frac{\pi_C}{\pi_{NT}+\pi_C}}^{1} y \, dF(Y = y \mid D = 0, Z = 0) \tag{10}$$

We now have the two admissible intervals, which we use to compare the pure always-takers and never-taker means δ_{AT}^0 and δ_{NT}^0. The means are either

- compatible if $\delta_{AT}^0 \in \left[\delta_{AT}^{1,LB}, \delta_{AT}^{1,UB}\right]$ and $\delta_{NT}^1 \in \left[\delta_{NT}^{0,LB}, \delta_{NT}^{0,UB}\right]$. Then, we cannot reject IV validity. Or
- incompatible if either $\delta_{AT}^0 \notin \left[\delta_{AT}^{1,LB}, \delta_{AT}^{1,UB}\right]$ or $\delta_{NT}^1 \notin \left[\delta_{NT}^{0,LB}, \delta_{NT}^{0,UB}\right]$. Then, we can reject IV validity as one of assumptions 1–3 must be violated.

Only one of the two conditions can be tested in settings with one-sided non-compliance that rule out the existence of always or never-takers.

These testing equations are equivalent to, but expressed differently than, the testable implications derived by Huber and Mellace (2015). They are optimal to refute IV validity defined by assumptions 1-3 as long as the outcome is continuous (see Laffers and Mellace, 2017). Yet, just as the Kitagawa (2015) testing conditions, they cannot verify IV validity. The probability of detecting a violation increases as the bounds narrow. Greater shares of always or never-takers compared to complier shares correspond to tighter bounds. Additionally, conditioning on covariates, especially those that explain most of the variation in the treatment selection or outcome, can tighten the bounds (Lee, 2009; Semenova, 2026). Huber and Mellace (2015) show that imposing mean dominance assumptions—namely, that the potential outcome means of always-takers (never-takers) are greater (smaller) than or equal to those of compliers in the treated (untreated) state—can tighten the bounds and, when both assumptions hold, yield equality constraints. This holds likewise for our approach, as we test the same identifying assumptions (conditional on covariates), which can help increase testing power. However, this might not be relevant in many applied settings, where mean dominance assumptions are less plausible than the IV validity conditions.

With this notation, we can define the parameters that we test as

$$\theta_1 = \begin{cases} \delta^0_{AT} - \delta^{1,UB}_{AT} & \text{if } \delta^{1,LB}_{AT} < \delta^0_{AT} \\ \delta^{1,LB}_{AT} - \delta^0_{AT} & \text{else.} \end{cases}$$

for the treated case and

$$\theta_0 = \begin{cases} \delta^1_{NT} - \delta^{0,UB}_{NT} & \text{if } \delta^{0,LB}_{NT} < \delta^1_{NT} \\ \delta^{0,LB}_{NT} - \delta^1_{NT} & \text{else.} \end{cases}$$

for the untreated case. If IV validity is violated, θ_1 and/or θ_0 are structurally larger than zero, meaning that the δ^0_{AT} and/or δ^1_{NT} lie outside their corresponding bounds. This defines our hypothesis as

$$H_0 : \begin{pmatrix} \theta_1 \\ \theta_0 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{11}$$

A positive θ indicates that δ^0_{AT} or δ^1_{AT} lie outside the admissible bounds, i.e., the means are incompatible with IV validity.

Estimation

Now, for the estimation approach covariates are explicitly expressed, as their implementation into an easy testing procedure is the key contribution of this paper. To determine θ_1 and θ_0, we need to estimate the type shares π_{AT}, π_{NT} and π_C. We do so by estimating the following first-stage equation,

$$D_i = \pi_{AT} + \pi_C Z_i + \widetilde{X}_i'\delta + U_{Di} \tag{12}$$

where $\widetilde{X}$ indicates the demeaned covariate vector X. Since the covariates are held constant at their means and both D and Z are binary, the constant can be interpreted as the share of always-takers (always $D = 1$), and π_C as the share of compliers (D varies with Z).[6] Consequently, as the shares sum up to one, the share of never-takers is given by $\pi_{NT} = 1 - \pi_{AT} - \pi_C$.

Furthermore, for the conditional expected values entering θ_0 and θ_1, we estimate the conditional densities $f^1_{AT,C}(Y)$, $f^0_{AT}(Y)$, $f^1_{NT}(Y)$, and $f^0_{NT,C}(Y)$. To derive the conditional pdfs, we start by estimating the conditional cdfs for each observable group (determined by possible combinations of D and Z) given covariates with a distribution regression approach. These regressions are standard tools in the IV quantile treatment-effects literature (Chernozhukov and Hansen, 2005; Frandsen et al., 2012) and in the construction of Horowitz–Manski–Lee bounds in IV settings (Dong, 2019; Westphal et al., 2022). $F(y) = Pr(Y \leq y|D = d, Z = z, \widetilde{X})$ is a binary choice model with the dependent variable $\mathbb{1}[Y \leq y]$ for an arbitrary threshold y.[7] Therefore, we run repeated linear probability models of the form

$$\begin{aligned}\mathbb{1}[Y \leq y] = {} & F^0_{NT,C}(y)\mathbb{1}[D = 0]\mathbb{1}[Z = 0] + F^0_{AT}(y)\mathbb{1}[D = 1]\mathbb{1}[Z = 0] \\ & + F^1_{NT}(y)\mathbb{1}[D = 1]\mathbb{1}[Z = 0] + F^1_{AT,C}(y)\mathbb{1}[D = 1]\mathbb{1}[Z = 1] + \widetilde{X}'\lambda + v\end{aligned} \tag{13}$$

with various thresholds y in the support of Y. Note that $F^0_{NT,C}$, F^0_{AT}, F^1_{NT}, and $F^1_{AT,C}$ are parameters estimated by this regression. They measure the share of observations conditional on $D = d$ and $Z = z$ below the threshold y, while all $\widetilde{X}$ are set to zero (and are, hence, fixed).[8] Repeating this regression for many y on the support of Y approximates the group-specific conditional CDF. By choosing a finer grid of values for y, one can improve the chance to describe $F(y)$ accurately. Since the PDF is the derivative of the CDF, we estimate the slope of the conditional CDFs at each y. The slopes at each evaluation point can be estimated using kernel-weighted local polynomial regression. This requires

[6] This interpretation is valid as long as Assumptions 1 and 3 hold. Without covariates, the (sum of) shares can easily be calculated with $\pi_{AT} = Pr(D = 1|Z = 0)$, $\pi_{AT} + \pi_C = Pr(D = 1|Z = 1)$, $\pi_{NT} = Pr(D = 0|Z = 1)$, and $\pi_{NT} + \pi_C = \Pr(D = 0|Z = 0)$.

[7] Without further indication, it is implicit that all cdfs are given conditional on covariates.

[8] One could, instead of linear models, run, e.g., repeated logit models and use predictive margins for each group.

selecting a kernel function and a bandwidth. One can follow Mourifié and Wan (2017) and use the rule-of-thumb choice by Fan and Gijbels (1996)[9], apply bandwidths that minimize the mean integrated squared error, or choose your own bandwidth.

Calculating the θs based on the estimated density function yields the estimated parameters $\widehat{\theta}_1$ and $\widehat{\theta}_0$. Still, bootstrap-based inference is needed to test the H_0 at given significance levels. Therefore, we generate B bootstrap samples of size N (number of observations) randomly drawn from the original sample with replacement and indicated with $b \in \{1, 2, ..., B\}$. $\widehat{\theta}_{1,b}$ and $\widehat{\theta}_{0,b}$ denote the estimates calculated within every sample. Our p-value-based test is very similar to the simple bootstrap test with the Bonferroni adjustment applied by Huber and Mellace (2015), except that we reduce the number of constraints from the outset when defining the test parameters. To obtain p-values, we recenter the parameter from each bootstrap sample, such that $\widetilde{\theta}_{1,b} = \widehat{\theta}_{1,b} - \widehat{\theta}_1$ and $\widetilde{\theta}_{0,b} = \widehat{\theta}_{0,b} - \widehat{\theta}_0$. This step, suggested by Hall and Wilson (1991), increases testing power if bootstrap samples are drawn from populations that do not satisfy H_0. To test the constraints of the H_0 against an upper-tailed alternative hypothesis separately, the bootstrap p-values for the treated and untreated cases are then given by

$$\begin{aligned} p_{\widehat{\theta}_1} &= \tfrac{1}{B} \textstyle\sum_{b=1}^{B} \mathbb{1}[\widetilde{\theta}_{1,b} > \widehat{\theta}_1] \\ p_{\widehat{\theta}_0} &= \tfrac{1}{B} \textstyle\sum_{b=1}^{B} \mathbb{1}[\widetilde{\theta}_{0,b} > \widehat{\theta}_0]. \end{aligned} \tag{14}$$

[10]

However, we want to perform a joint test on θ_1 and θ_0. The more conditions are tested, the higher the probability of obtaining an unusually high test statistic at random. Therefore, we apply the Šidák or Dunn-Šidák correction where the significance level for each test is set to $\alpha' = 1 - (1 - \alpha)^{\frac{1}{m}}$ with m being the number of tests and α the overall significance level (Šidák, 1967). For the p-value of the joint test follows that $\widehat{p} = 1 - (1 - \min(p_{\widehat{\theta}_1}, p_{\widehat{\theta}_0}))^m$. Although slightly less conservative than the Bonferroni correction, the Šidák correction can still be too conservative when m is large, and the test statistics are positively correlated (MacKinnon, 2009). With $m = 2$ in our case, we have the fewest conditions tested simultaneously. If the test statistics are not independent, the resulting p-value $\widehat{p}$ is still an upper bound and $\min(p_{\widehat{\theta}_1}, p_{\widehat{\theta}_0})$ the lower bound in the extreme case of perfectly correlated statistics (MacKinnon, 2009). Hence, consulting $p_{\widehat{\theta}_1}$ and $p_{\widehat{\theta}_0}$ as well as the Šidák corrected p-value for the joint test $\widehat{p}$ should be enough to judge on the H_0 in most settings.

[9] This rule-of-thumb bandwidth choice is implemented in several Stata packages; for example, it is the default of the *lpoly* package.

[10] This follows from the fact that we want to reject our H_0 when the observed value of our test statistic $\widehat{T}$ is in the upper tail of $F(T)$, the cdf of T under the H_0. The distribution of the bootstrap test statistics $\widehat{T}_b$ gives the empirical distribution function $\widehat{F}$, i.e., the asymptotic approximation of F. Then, the bootstrap p-value is $p_{\widehat{\theta}} = 1 - \widehat{F}(\widehat{T}) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}[\widehat{T}_b > \widehat{T}]$ (see MacKinnon, 2009). Plugging in $\widehat{T}_b = \sqrt{N}\,(\widehat{\theta}_b - \widehat{\theta})/\sigma_{\widehat{\theta}}$ and $\widehat{T} = \sqrt{N}\,\widehat{\theta}/\sigma_{\widehat{\theta}}$ yields the simplified version in Eq. (14).

To summarize, we conduct the following step-by-step implementation[11]:

1. Demean covariates to get $\widetilde{X}$.
2. Estimate shares of types with first stage regression (Eq. (12)).
3. Set a grid for evaluation points within the support of Y (e.g., quantiles of the observed distribution of Y).[12]
4. Estimate conditional cdfs with repeated regressions of binary choice models (Eq. (13)).
5. Determine the slopes at the evaluation points to get conditional pdfs (e.g., with local linear regression).[13]
6. Calculate conditional means δ^0_{AT} and δ^1_{NT} as well as lower and upper bounds $\delta^{1,LB}_{AT}$, $\delta^{1,UB}_{AT}$, $\delta^{0,LB}_{NT}$ and $\delta^{0,UB}_{NT}$ according to equations (7–10).
7. Determine θ_1 and θ_0 by plugging in results from step 6.
8. Conduct inference on both parameters, i.e., derive bootstrapped inference by repeating steps 1 to 7 with B bootstrap samples of size N of the original sample (B =number of bootstrap repetitions, N =number of observations), derive the corresponding p-values according to Eq. (14) and apply the Šidák method to obtain one p-value for the joint test.

4 Extending the Framework to Misspecified Covariates: Detecting IV Specification Bias

In this section, we focus on detecting potential specification bias. For simplicity, we assume that the IV validity assumptions—defined in Section 2—hold. To assess how misspecified covariates affect our approach, we return to Eq. (13), where the parameters of the group-specific conditional cumulative distribution functions (CDFs) are estimated through a sequence of OLS regressions for each evaluation point y, conditioning on covariates X. These parameters are $\widehat{F^0_{NT,C}(y)}$, $\widehat{F^1_{NT}(y)}$, $\widehat{F^0_{AT}(y)}$, and $\widehat{F^1_{AT,C}(y)}$. For notational convenience, define four binary variables $D_dZ_z := \mathbb{1}\{D = d\} \cdot \mathbb{1}\{Z = z\}$ for each combination of D and Z. We define

$$\widetilde{D_dZ_z} = D_dZ_z + \Delta_{dz}(X),$$

[11]For implementation, see the Stata replication files for the empirical applications from section 6 in the Online Supplementary Material.

[12]As quantiles use to bunch in the middle of a unimodal distribution, one might want to use more dense evaluation points in the tails of the distribution.

[13]All results shown in the paper are based on local linear regressions with Epanechnikov kernel.

where

$$\Delta_{dz}(X) \equiv \mathbb{E}(D_d Z_z \mid X) - \mathbb{L}(D_d Z_z \mid X)$$

captures potential specification error. The first term, $\mathbb{E}(D_d Z_z \mid X)$, denotes the true (nonparametric) conditional mean—that is, the expected value within each cell defined by the full set of covariates in X.

The second term, $\mathbb{L}(D_d Z_z \mid X)$, is the linear projection of $D_d Z_z$ onto X, given by

$$\mathbb{L}(D_d Z_z \mid X) = X(X'X)^{-1}X'(D_d Z_z),$$

and reflects the potentially restrictive or parsimonious functional form imposed by the covariate specification when the model is not fully saturated. The misspecification error $\Delta_{dz}(X)$ therefore arises whenever the true conditional expectation cannot be represented by the chosen linear specification in X.

Following Blandhol et al. (2026), but extending their results to our separate estimation approach, the coefficients in Eq. (13) can be written as

$$\frac{\mathbb{E}\left(Y\,\widetilde{D_d Z_z}\right)}{\mathbb{E}\left(\widetilde{D_d Z_z}\right)^2} = \frac{\mathbb{E}\left[\mathbb{E}(Y\,\widetilde{D_d Z_z} \mid X)\right]}{\mathbb{E}\left[\mathbb{E}((\widetilde{D_d Z_z})^2 \mid X)\right]},$$

where, in our approach, we substitute Y with an indicator variable $\mathbb{1}\left[Y \leq y\right]$ for a specific point y from the support of Y, and estimate many such regressions across different values of y. This yields pointwise estimates that together characterize the entire CDF. The right-hand side applies the law of iterated expectations, thereby evaluating the expectation operator (nonparametrically) for each X cell and aggregating across these cells.

Using the definition of $\widetilde{D_d Z_z}$ and rearranging, the coefficients can be written as

$$\begin{aligned}
\frac{\mathbb{E}\left(Y\,\widetilde{D_d Z_z}\right)}{\mathbb{E}\left(\widetilde{D_d Z_z}\right)^2} &= \frac{\mathbb{E}(Y D_d Z_z) + \mathbb{E}\left[\mathbb{E}(Y \mid X)\Delta_{dz}(X)\right]}{\mathbb{E}\left[\mathbb{E}(D_d Z_z + 2 D_d Z_z \Delta_{dz}(X) + \Delta_{dz}(X)^2 \mid X)\right]} \\
&= \frac{\mathbb{E}(Y \mid D = d, Z = z)\mathbb{E}(D_d Z_z) + \mathbb{E}\left[\mathbb{E}(Y \mid X)\Delta_{dz}(X)\right]}{\mathbb{E}(D_d Z_z) + 2\mathbb{E}\left[\mathbb{E}(D_d Z_z \mid X)\Delta_{dz}(X)\right] + \mathbb{E}\left[\mathbb{E}\left((\Delta_{dz}(X))^2\right)\right]}
\end{aligned} \tag{15}$$

If partialing out with respect to X is exact—that is, if $\mathbb{L}(D_d Z_z|X) = E[D_d Z_z|X]$—there is no misspecification error ($\Delta_{dz}(X) = 0$). In this case, OLS yields unbiased estimates of group-specific outcomes $\mathbb{E}(Y \mid D = d, Z = z)$.

Bias arises when the linear projection fails to capture nonlinear or interactive effects among covariates that influence treatment or instrument assignment. For instance, suppose age and work experience enter linearly, but their joint influence on treatment probability is nonlinear—such as when treatment likelihood increases rapidly with experience only at certain ages. In that case, although both variables are included in X, their interaction term is omitted, so the linear projection cannot reproduce the curvature of $\mathbb{E}(D_d Z_z \mid X)$. This misspecification implies that $\Delta_{dz}(X) \neq 0$ for some combinations of covariate values (e.g., age=a, experience=e). Consequently, the residuals from this approximation have nonzero means within certain cells of X, introducing systematic bias.

To analyse the sources of the bias, we simplify Eq. (15) by applying a first-order Taylor expansion to get:

$$\begin{aligned}
\text{bias}(D=d, Z=z) &= \frac{\mathbb{E}\left(Y\,\widetilde{D_d Z_z}\right)}{\mathbb{E}\left(\widetilde{D_d Z_z}\right)^2} - \mathbb{E}(Y \mid D=d, Z=z) \\
&\approx \frac{\mathbb{E}\left[\mathbb{E}(Y|X)\Delta_{dz}\right]}{\mathbb{E}(D_d Z_z)} - 2\frac{\mathbb{E}(Y \mid D=d, Z=z)}{\mathbb{E}(D_d Z_z)}\mathbb{E}\left[\mathbb{E}(D_d Z_z \mid X)\Delta_{dz}(X)\right]
\end{aligned}$$

See Appendix B for a formal derivation.

As our test compares $F^0_{NT,C}(y)$ to $F^1_{NT}(y)$ and $F^0_{AT}(y)$ to $F^1_{AT,C}(y)$—that is, it evaluates the effects of Z on the distributions of Y conditional on D and X—it is informative to consider the relative bias by shifting Z while holding D fixed:

$$\Delta\text{bias}(D=d) \equiv \text{bias}(D=d, Z=1) - \text{bias}(D=d, Z=0).$$

In Appendix B, we show that the bias can be rearranged to

$$\begin{aligned}
\Delta\text{bias}(D=d) \approx\ & \mathbb{E}\Bigg\{ \overbrace{\left[\frac{\mathbb{E}(\Delta_{d1}(X))}{\mathbb{E}(D_d Z_1)} - \frac{\mathbb{E}(\Delta_{d0}(X))}{\mathbb{E}(D_d Z_0)}\right]}^{①} \mathbb{E}(Y \mid X) \Bigg\} \\
& -2\mathbb{E}(Y \mid D=d, Z=1)\overbrace{\mathbb{E}\left[\omega_{d1}(X)\Delta_{d1}(X) - \omega_{d0}(X)\Delta_{d0}(X)\right]}^{②} \\
& -2\Big[\mathbb{E}(Y \mid D=d, Z=1) - \mathbb{E}(Y \mid D=d, Z=0)\Big]\mathbb{E}\underbrace{\left[\omega_{d0}(X)\Delta_{d0}(X)\right]}_{③},
\end{aligned}$$

with $\omega_{dz}(X) = \mathbb{E}(D_d Z_z \mid X)/\mathbb{E}(D_d Z_z)$ being weighting factors that sum to one. This expression shows that the bias depends on three components. These components are:

1) **Spillover contamination factor:**

$$\left[\frac{\mathbb{E}(\Delta_{d1}(X))}{\mathbb{E}(D_d Z_1)} - \frac{\mathbb{E}(\Delta_{d0}(X))}{\mathbb{E}(D_d Z_0)}\right]$$

This factor captures the extent to which the conditional outcome mean is contaminated by inadmissible compliance types, such as never-takers in $D = 1$ or always-takers in $D = 0$. More generally, this is the source of bias that also affects joint estimation, as shown by Blandhol et al. (2026).

2) **Between instrument-state contamination factor:**

$$\left[\omega_{d1}(X)\Delta_{d1}(X) - \omega_{d0}(X)\Delta_{d0}(X)\right]$$

This contamination arises from differences in the relative specification errors between $Z = 1$ and $Z = 0$.

3) **Contamination factor due to unobserved heterogeneity:**

$$\left[\omega_{d0}(X)\Delta_{d0}(X)\right]$$

This factor determines the degree to which unobserved heterogeneity—that is, the difference between the complying and non-complying groups $\mathbb{E}(Y \mid D = d, Z = 1) - \mathbb{E}(Y \mid D = d, Z = 0)$)—affects the bias.

For a numerical illustration of misspecification bias and its implications for our testing procedure, see Section 5.2.

Remark 3 (Omitted-variable interpretation)**.** *Suppose the specification error is*

$$\Delta_{dz}(X) = \mathbb{E}(D_d Z_z \mid X) - \mathbb{L}(D_d Z_z \mid X_{\text{restr}}),$$

where X denotes the full covariate vector sufficient for mean independence and $X_{\text{restr}} \subset X$ is the restricted covariate set used in estimation. Let $X_{\text{omit}} := X \setminus X_{\text{restr}}$ denote the omitted covariates. Then our bias formula quantifies the bias arising from omitting X_{omit}, i.e., the omitted-variable bias induced by using X_{restr} instead of X—demonstrating that specification error and violations of conditional mean independence are closely connected.

Blandhol et al. (2026) propose using Ramsey's RESET test to assess covariate misspecification under the null hypothesis that $\mathbb{E}(Z \mid X) = \mathbb{L}(Z \mid X)$, that is, when the (nonparametric) conditional expectation of the instrument is correctly captured by its linear projection. Their approach, therefore, provides an indirect diagnostic by testing whether the regression of the instrument on the covariates is correctly specified. In contrast, our procedure is targeted: it is designed to detect only misspecifications that directly affect the Wald estimand.

Rather than testing the correctness of the functional form with Z as the dependent variable per se, our test focuses on deviations that affect the LATE more directly. Moreover, the power of the RESET test depends on auxiliary modeling choices—such as the polynomial order—that are external to the identifying framework.

When our test rejects the joint null of IV validity and correct covariate specification, the first diagnostic step should be to increase the flexibility of the covariate adjustment. In practice, this amounts to moving toward a more saturated specification—for example, by introducing higher-order polynomials, interaction terms, or richer sets of fixed effects. Such refinements are straightforward to implement within our framework. If rejection persists and p-values do not increase after these adjustments, the remaining evidence is unlikely to be driven by covariate misspecification. Importantly, this comparison should be interpreted jointly with the width of the bounds, as p-values may also increase mechanically when changes in the specification alter type shares and thereby widen the bounds. When the bounds remain similar across specifications, however, a lack of improvement in the p-values suggests that the rejection is unlikely to be caused by covariate misspecification. In that case, the rejection should be interpreted as pointing to violations of IV validity rather than functional-form errors in the control structure.

At first glance, a potential limitation of our procedure is that it may fail to detect violations when biases induced by covariate misspecification and IV invalidity offset one another. In such cases, however, the implied distortion of the LATE itself is likely to be limited, since the two components cancel at the level of the estimand. A non-rejection should therefore be interpreted as evidence that any remaining bias in the causal parameter is quantitatively small rather than as proof that each underlying assumption holds exactly.[14] That said, it is possible to reduce the scope for such offsetting biases. The magnitude—and potentially even the sign—of $\Delta_{dz}(X)$ generally differs depending on whether $d = 1$ or $d = 0$ is evaluated. By contrast, violations of the exclusion restriction are invariant to relabeling the treatment. The same logic can apply—albeit more weakly—to deviations from conditional independence. This asymmetry can be exploited empirically: implementing the test both with D and with $\widetilde{D} = 1 - D$ limits the possibility that misspecification bias and IV invalidity cancel in both parameterizations simultaneously, thereby increasing the test's sensitivity.

[14] In principle, one could combine our approach with more flexible nuisance-estimation methods, such as the double machine learning framework of Chernozhukov et al. (2018). However, this may not be appropriate in research designs that deliberately impose covariate restrictions for identification—for example, additive structures in two-way fixed-effects settings.

5 Simulation

5.1 Testing violations of conditional independence and the exclusion restriction

We perform Monte Carlo simulations to assess the size and power of our testing procedure—that is, the probabilities of falsely and correctly rejecting H_0. The data-generating process (DGP) follows the designs in Huber and Mellace (2015) and Carr and Kitagawa (2023). We simulate $S = 1000$ times, allowing random parameters to vary across simulations. For each simulated data set, we draw $B = 499$ bootstrap replications (holding the parameters fixed) to obtain a p-value. The DGP reads

$$Y = X'\beta_X + \beta_D D + \beta_Z Z + U$$

$$D = \mathbb{1}[\pi_0 + \pi_1 Z + U_D \geq 0]$$

$$\text{with } \pi_0 = \Phi^{-1}(0.45) \text{ and } \pi_1 = \Phi^{-1}(0.55) - \Phi^{-1}(0.45)$$
$$\text{(implying } \pi_{AT} = 0.45, \pi_C = 0.1, \text{ and } \pi_{NT} = 0.45)$$

$$Z = \mathbb{1}[X'\gamma + U_Z \geq 0]$$

$$X = (X_1, X_2, X_3);\ X_j \sim N(0, I) \forall j \in \{1, 2, 3\}$$

$$U_Z \sim N(0, 1)$$

$$U, U_D \sim N(0, \Sigma) \qquad \text{with } \Sigma = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix},$$

where $\Phi^{-1}()$ is the inverse of the standard normal distribution to better control the (non)complier shares. Because X and U are independent, this DGP satisfies the "rich covariates" condition of Blandhol et al. (2026). Hence, issues concerning the interpretation of 2SLS as LATE in the presence of covariates do not arise. We set each component of β_X ($\beta_{X,1}, \beta_{X,2}, \beta_{X,3}$) to 1, and the treatment effect to $\beta_D = 1$. Monotonicity (Assumption 3) holds by construction, since π_0 and π_1 are constant across individuals.

Our simulation focuses on potential violations of Assumptions 1 and 2. We distinguish settings in which independence holds only conditional on X and settings in which the exclusion restriction fails because the instrument directly affects Y. We vary γ and β_Z as follows:

- **Violation of the independence assumption (Assumption 1):**
  - Independence holds unconditionally: $\gamma_1 = \gamma_2 = \gamma_3 = 0$

- Independence assumption violated when not conditioning on X: $\gamma_1 = \gamma_2 = \gamma_3 = 0.22$

- **Violation of the exclusion restriction (Assumption 2)**
  - Exclusion restriction holds: $\beta_Z = 0$
  - Exclusion restriction violated: $\beta_Z = 1$

Setting all elements of γ to 0.22 induces an omitted variable bias for Z without conditioning on X of roughly 1. Likewise, β_Z represents a direct effect of Z on Y, violating the exclusion restriction. Under both violations, the distribution of Y shifts upwards when $Z = 1$.

We apply our test procedure to each simulated or replicated sample, both with and without conditioning on X. Rejection rates are computed as

$$\text{Rejection rate} = \frac{1}{S} \sum_{s \in S} \mathbb{1}\Big[p_{\hat{\theta}} \leq \text{Nominal size}\Big],$$

for $S = 1000$ simulations and nominal sizes $\{0.1, 0.05, 0.01\}$. Šidák adjusted p-values $p_{\hat{\theta}}$ are based on $B = 499$ bootstrap replications. Table 1 reports the rejection rates We present the results for sample sizes of 250 and 1000, with and without covariates, and compares them to the Huber and Mellace (2015) procedure without covariates.[15]

The white cells of Table 1 correspond to cases in which IV is valid ($\beta_Z = 0$ and either with covariates included or $\gamma_j = 0$). With covariates, our test delivers rejection rates at or below the nominal size, even with the smaller sample ($N = 250$). For the larger sample, the rejection rate is zero whenever the instrument is truly valid. Rejection rates may fall below nominal sizes because the DGP is not at the boundary of the test condition. Without covariates, the test also performs well so long as X does not affect Z (columns 1–3), with one minor exception (for $N = 250$ at the 5% level. The lower rejection rates of the Huber and Mellace (2015) procedure can be attributed to its greater precision when no covariates are included, as it does not rely on estimating conditional distributions. When $\gamma \neq 0$ (columns 4–6), X and Z are not correlated, violating unconditional mean independence (Assumption 1). These violations appear in rejection rates above the nominal size for the test without covariates (light gray cells) across both sample sizes. By contrast, conditioning on X corrects for this confounding, yielding rejection rates that are substantially lower and already at or below nominal size for $N = 250$. Thus, whenever the assignment of Z is not unconditionally random and relevant covariates are observed, the test should include these covariates to avoid falsely rejecting IV validity.

[15] We adapt the Stata code by Huber and Mellace (2014). For comparability, we report rejection rates based on simple bootstrap tests with Šidák-adjusted p-values under two binding constraints. These are weakly below those obtained under the Bonferroni adjustment for four constraints used in Huber and Mellace (2015). Although they show that some alternative tests outperform the Bonferroni-adjusted bootstrap in parts of their study, overall patterns are similar.

Table 1: Simulation results

	Z and X are independent ($\gamma_1 = \gamma_2 = \gamma_3 = 0$)			Z depends on X ($\gamma_1 = \gamma_2 = \gamma_3 = 0.22$)		
Nominal size:	0.1	0.05	0.01	0.1	0.05	0.01
Exclusion restriction holds: $\beta_Z = 0$						
w/ covariates (only our approach feasible)						
N=250	0.075	0.050	0.010	0.050	0.020	0.010
N=1000	0.000	0.000	0.000	0.000	0.000	0.000
w/o covariates						
– Our approach						
N=250	0.095	0.060	0.001	0.525	0.430	0.210
N=1000	0.000	0.000	0.000	0.680	0.570	0.345
–Huber & Mellace (2015)						
N=250	0.010	0.005	0.000	0.210	0.175	0.050
N=1000	0.000	0.000	0.000	0.540	0.370	0.175
Exclusion restriction violated: $\beta_Z = 1$						
w/ covariates (only our approach feasible)						
N=250	0.580	0.495	0.290	0.680	0.600	0.425
N=1000	0.705	0.610	0.425	0.800	0.710	0.535
w/o covariates						
– Our approach						
N=250	0.500	0.380	0.220	0.970	0.950	0.830
N=1000	0.620	0.510	0.355	1.000	1.000	1.000
–Huber & Mellace (2015)						
N=250	0.285	0.190	0.065	0.865	0.790	0.570
N=1000	0.485	0.355	0.205	1.000	1.000	1.000

Notes: The rejection rates are based on the Šidák adjusted p-values. The bandwidths minimize the mean integrated squared error for Gaussian data (the default of the Stata package *locpoly3*). When $\beta_Z = 0$, the instrument is valid (white cells) except in columns 4-6 without including conditioning covariates (light gray cells), where X and Z are not independent. When $\beta_Z \neq 0$, the exclusion restriction does not hold; hence, the instrument is invalid (medium gray cells). Additionally, X and Z are not independent in columns 4-6 without conditioning on covariates (dark gray cells).

When the exclusion restriction fails ($\beta_Z = 1$), the instrument is invalid and H_0 should be rejected (medium- and dark-gray cells). When only the exclusion restriction is violated (medium-gray cells), rejection rates exceed the nominal size for both sample sizes. In columns 1–3, where no confounders affect Z, conditioning on covariates increases rejection rates by tightening the bounds on the unobserved potential outcomes, thereby increasing power to detect violations. Somewhat unexpectedly, without covariates, our procedure yields higher rejection rates than Huber and Mellace (2015) (indicating a better test performance in the gray cells). However, this may also reflect imprecision from our choice of evaluation points and bandwidths, which tends to increase sensitivity to violations (as evidenced by slightly higher false-rejection rates in the white cells). When both the exclusion restriction and the unconditional independence fail (dark gray cells), higher rejection rates are expected because both violations shift outcome distributions for $Z = 1$ in the same direction.

Overall, the results show that our procedure performs well in size and power and clearly outperforms the version without covariates when observed confounders are correlated with Z and Y. Moreover, including covariates can be beneficial even when Z and X are independent, as it may tighten the bounds on the means of the unobserved potential outcomes. Naturally, when X and Y are independent, the test of Huber and Mellace (2015) is by construction as least as precise as ours.

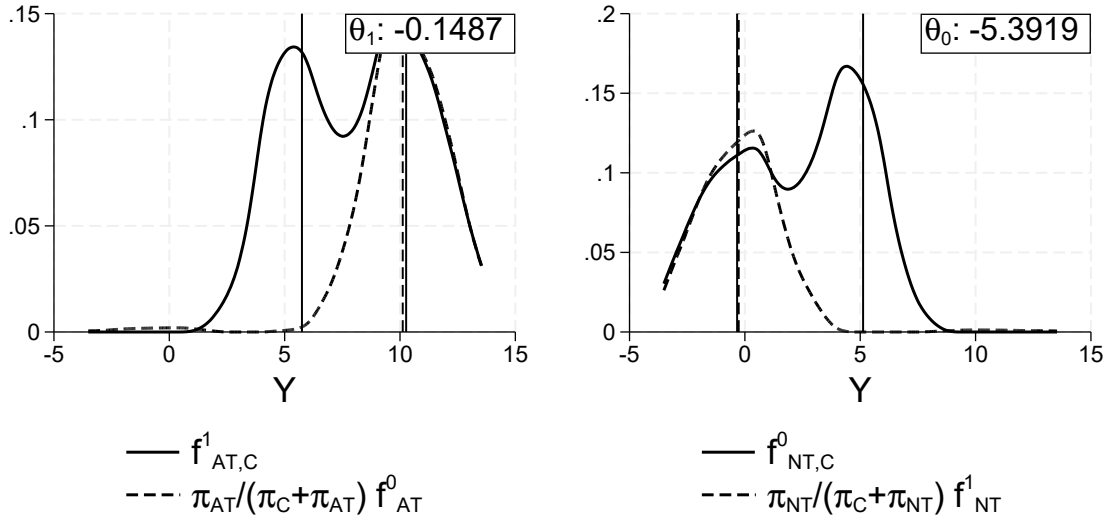
5.2 Simulated numerical example of the specification bias

To demonstrate that our testing procedure can detect not only violations of the IV validity assumptions but also specification bias arising from under-specified models, we present results for a simple simulated data example. The design closely follows the numerical illustration in Blandhol et al. (2026). The corresponding DGP is given by

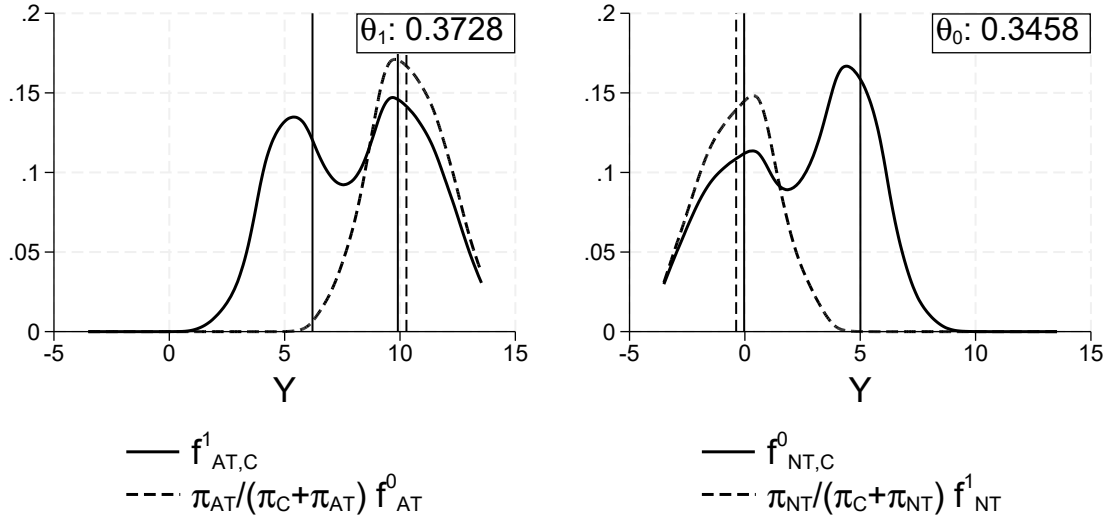
$$
\begin{aligned}
\mathcal{Z}, \mathcal{D} &\sim U(0,1) \\
U &\sim N(0, 0.33) \\
X &\in \{-1, 0, 1\} \text{ with equal probability} \\
Z &= \begin{cases} \mathbb{1}[\mathcal{Z} \leq 0.8] & \text{if } |X| = 1 \\ \mathbb{1}[\mathcal{Z} \leq 0.4] & \text{if } X = 0 \end{cases} \\
D^1 &= \begin{cases} \mathbb{1}[\mathcal{D} \leq \frac{2}{3}] & \text{if } |X| = 1 \\ \mathbb{1}[\mathcal{D} \leq \frac{5}{6}] & \text{if } X = 0 \end{cases} \\
D^0 &= \begin{cases} \mathbb{1}[\mathcal{D} \leq \frac{1}{3}] & \text{if } |X| = 1 \\ \mathbb{1}[\mathcal{D} \leq \frac{1}{2}] & \text{if } X = 0 \end{cases} \\
D &= D^0 + Z(D^1 - D^0) \\
Y &= 0 \cdot \mathbb{1}[D^0 = 0, D^1 = 0] + 10 \cdot \mathbb{1}[D^0 = 1, D^1 = 1] + (4.5 + Z) \cdot \mathbb{1}[D^0 = 0, D^1 = 1] + U.
\end{aligned}
$$

This setup implies expected group shares of $\pi_{AT} = \frac{1}{3}\mathbb{1}[|X| = 1] + \frac{1}{2}\mathbb{1}[X = 0]$, $\pi_C = \frac{1}{3}$, and $\pi_{NT} = \frac{1}{3}\mathbb{1}[|X| = 1] + \frac{1}{6}\mathbb{1}[X = 0]$. If X enters linearly in the regression model, then $E(\widetilde{D_d Z_z} \mid X) \neq 0$, implying $\Delta_{dz} \neq 0$ for each combination of D and Z. Instrument validity holds irrespective of covariate specification in this example.

Figure 3 displays the graphical test results. Both panels show the estimated densities and (bounds of) mean potential outcomes our test is based on. In each panel the left graph



(a) Fully saturated



(b) Misspecified covariates

Figure 3: Graphs for the college proximity instrument

Notes: Own illustration. The bandwidths minimize the mean integrated squared error for Gaussian data (the default of the Stata package *locpoly3*). PDFs for the mixed groups are shown by the solid curves, and for the single groups, by the dashed curves. The vertical solid lines indicate the lower and upper bounds $\delta_{AT}^{1,LB/UB}$ (left) and $\delta_{NT}^{0,LB/UB}$ (right), and the vertical dashed lines display the conditional mean δ_{AT}^{0} (left) and δ_{NT}^{1} (right). f_{AT}^{0} and f_{NT}^{1} are down-weighted by their relative shares. This does not affect the mean potential outcome given by the dashed vertical line. In panel (a) the model is fully saturated, i.e., including a dummy variable for each value of X, and in (b) the model linearly conditions on X.

corresponds to the treated state ($D = 1$) and the right graph to the untreated state ($D = 0$). The level-shifted outcome distributions for always-takers and never-takers relative to compliers—implied by the outcome definition in the DGP—are visible as bimodal joint density functions (solid CDFs). These outcome-level differences introduce bias under covariate misspecification because each type-specific outcome distribution is contaminated by the overall outcome distribution. Misspecification bias additionally depends on the specification errors $\Delta_{dz}(X)$ (see Section 4). Figure A.1 in Appendix A illustrates the

specification errors for each combination of D and Z in this setting, revealing non-zero specification errors that vary across observable groups for each value of X.

Panel (a) shows results for a fully saturated model; hence, the set of covariates is rich in the sense of Blandhol et al. (2026), and the 2SLS estimator can be interpreted as identifying the LATE. The estimated values of both testing parameters (θ_1 and θ_0) are negative, already indicating no bias. Moreover—as reported in column (1) of Table 2—both individual p-values and their Šidák correction are close to one, so there is no evidence against either instrument validity or correct covariate specification. By contrast, panel (b) presents results for a model with covariate misspecification—specifically one that conditions linearly on categorical variable X. While instrument validity remains intact, the 2SLS estimator is now subject to misspecification bias for the LATE interpretation. While instrument validity remains intact, the 2SLS estimator now suffers from specification bias affecting its LATE interpretation. For treated (untreated) units, the potential outcome means for always-takers (never-takers) differ across instrument states such that $\theta_1 > 0$ ($\theta_0 > 0$). All corresponding p-values—zero for both θ_1, θ_0, and the Šidák correction (see Table 2, column 2)—clearly support rejection of H_0. Because all three IV validity assumptions hold by construction, these rejections correctly identify bias originating from covariate misspecification rather than invalid instruments.

Table 2: Test results for model misspecification

	fully saturated	misspecified covariates
θ_1	-0.149	0.373
$p_{\hat{\theta}_1}$	0.988	0.000
θ_0	-5.392	-0.346
$p_{\hat{\theta}_0}$	1.000	0.000
Šidák corrected $\hat{p}$	1.000	0.000
Shares		
π_C	0.336	0.270
π_{AT}	0.389	0.433
π_{NT}	0.275	0.297
Observations	10,000	10,000

Notes: Test results are based on 499 bootstrap samples. In column (1) the model is fully saturated, i.e., including a dummy variable for each value of X, and in column (2) the model linearly conditions on X.

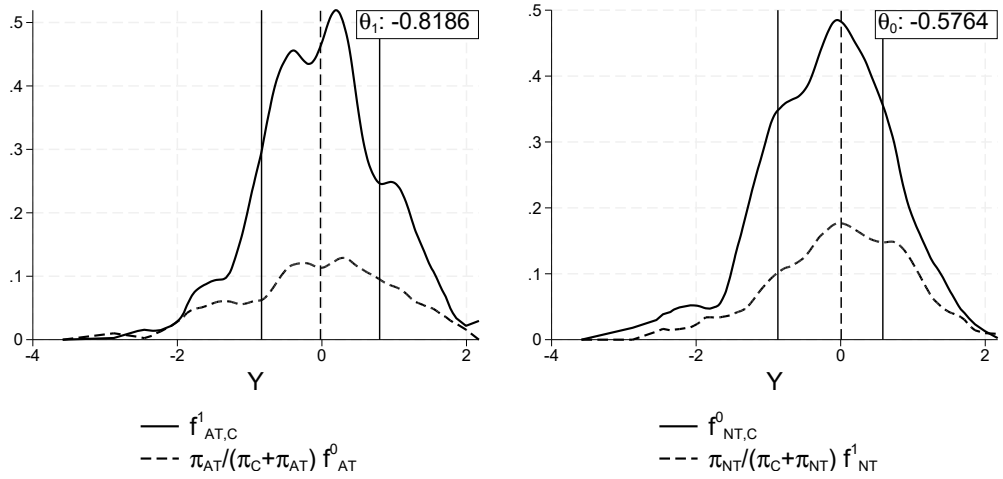
Overall, this simulation illustrates that our test effectively detects problems that compromise interpretation of 2SLS estimates as LATE when they arise from misspecified covariates rather than violations of IV validity.

6 Applications

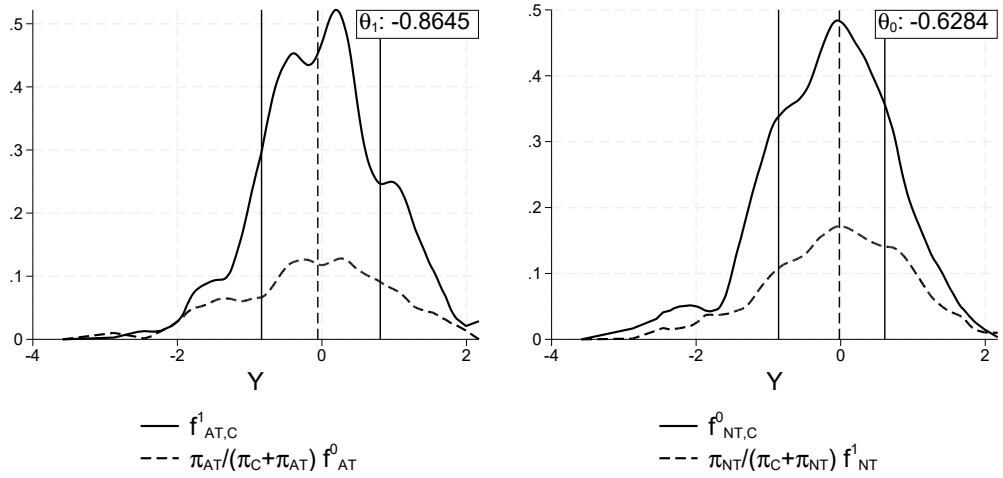
In this section, we apply our procedure to two modern applications from the literature. The first studies the effects of mobile banking training that incentivizes money transfers from urban to rural areas in Bangladesh (Lee et al., 2021), and the second uses the gender of the firstborn child as an instrument for female leadership in fifteenth- to twentieth-century Europe (Dube and Harish, 2020). For comparison, we additionally apply our testing approach to two well-known settings that have been widely used in the IV validity testing literature: the Vietnam-era draft lottery instrument from Angrist (1991) and the college proximity instrument from Card (1993). These results are reported in Appendix C. In line with previous evidence, we do not find indications of IV invalidity for the draft eligibility instrument. For the college proximity instrument, by contrast, the results depend on the inclusion of covariates: without controls the test suggests violations of the validity conditions, whereas including the covariates used in the original study removes this indication.

6.1 Mobile Banking—Training Instrument

In the first application, we revisit the randomized field experiment by Lee et al. (2021), which studies whether facilitating mobile banking can strengthen financial links between urban migrants and their rural families in Bangladesh. The intervention provided training and assistance to adopt mobile banking accounts, enabling migrants to transfer money to rural relatives more easily. The empirical strategy instruments active mobile banking use with the randomized training assignment to estimate a local average treatment effect of mobile banking adoption on rural household outcomes. Because the training assignment was randomized, the conditional independence assumption is highly credible. The original paper reports strong balance in baseline characteristics between treatment and control groups, supporting the validity of the randomization. Monotonicity also appears plausible, as the intervention lowers barriers to adoption and it is difficult to imagine individuals becoming less likely to use mobile banking because they received training. A potential concern is the exclusion restriction, since the training could affect outcomes through channels other than mobile banking use. However, the intervention was deliberately narrow—consisting only of a short training and assistance with account setup and not providing financial transfers or other support—suggesting that such effects are likely limited. The original specification also includes a small set of covariates. Although the number of controls is limited, functional-form misspecification cannot be ruled out—for example if variables such as age or household size enter the outcome equation nonlinearly. At the same time, because treatment assignment is randomized, these covariates are not required for identification of IV validity.



(a) Without covariates



(b) With covariates[a]

Figure 4: Graphs for the mobile banking training instrument

Notes: Own illustration based on data from Lee et al. (2021). Bandwidth=0.20. PDFs for the mixed groups are given by the solid curves, and for the single groups, they are given by the dashed curves. The vertical solid lines indicate the lower and upper bounds $\delta_{AT}^{1,LB/UB}$ (left) and $\delta_{NT}^{0,LB/UB}$ (right), and the vertical dashed lines display the conditional mean δ_{AT}^{0} (left) and δ_{NT}^{1} (right). f_{AT}^{0} and f_{NT}^{1} are down-weighted by their relative shares, which does not affect the corresponding means. [a]Covariates include indicators for female and primary school completion of the household head, as well as age and household size.

The data come from the experiment conducted by Lee et al. (2021), which follows 815 migrant–household pairs consisting of an urban migrant in Dhaka and a rural household in Gaibandha district.[16] The instrument Z equals one if the pair was randomly assigned to receive the mobile banking training. The endogenous treatment variable D indicates active use of a mobile banking account after the intervention. As outcome Y, we use the consumption index from the original study, corresponding to column (3) of Table 6 in Lee et al. (2021). We apply our test once without covariates and once including the baseline controls used by Lee et al. (2021): gender, age, and primary school completion

[16]The data and variable construction are taken from the replication files (Lee et al., 2020) accompanying Lee et al. (2021). We use the dataset prepared for the estimation of Table 6 in the original paper and employ the same variables and sample. The code implementing our test is provided in the Supplementary Material to facilitate replication.

of the household head, as well as household size. As in the original study, we employ an Analysis of Covariance specification that additionally conditions on the baseline value of the outcome.

Figure 4 presents the graphical results. Panel (a) shows the densities and bounds of the mean potential outcomes without covariates, and panel (b) shows the corresponding results with covariates. In each panel, the left graph corresponds to the treated state (relevant for θ_1) and the right graph to the untreated state (relevant for θ_0). The graphical evidence suggests that the validity conditions hold, as the dashed vertical lines lie within the bounds indicated by the solid vertical lines in both specifications. As there are no major changes when conditioning on covariates, problems with covariate misspecification are unlikely. This pattern is confirmed by the first two columns of Table 3. Both θ_1 and θ_0 are negative with and without covariates, and the corresponding p-values, as well as the Šidák-corrected p-values, equal one. Hence, we cannot reject IV validity (or correct covariate specification when covariates are included) in this application.

Table 3: Results of the empirical applications

	Training mobile banking (Lee et al., 2021)		Firstborn gender (Dube and Harish, 2020)		
	w/o covariates	w/ covariates[a]	w/o covariates	w/ covariates[b]	w/ add. cov.[c]
θ_1	-0.819	-0.576	0.088	-3.545	-3.654
$p_{\hat{\theta}_1}$	1.000	1.000	0.447	0.625	0.651
θ_0	-0.865	-0.628	0.059	-13.453	-10.156
$p_{\hat{\theta}_0}$	1.000	1.000	0.503	1.000	1.00
Šidák corrected $\hat{p}$	1.000	1.000	0.859	0.859	0.878
Shares					
π_C	0.482	0.483	0.164	0.239	0.203
π_{AT}	0.219	0.218	0.078	0.041	0.059
π_{NT}	0.299	0.298	0.758	0.720	0.738
Bandwidth	0.20			3.00	
Observations	813			3,586	

Notes: Tests are based on 499 bootstrap replications. For the firstborn-gender application, the bootstrap procedure clusters at the same level as in the original paper. We use 314 and 51 evaluation points for the mobile banking and the firstborn-gender applications, respectively. [a]Dummies for female and primary school completion of the household head, age and household size used as covariates. [b]Covariates include dummy variables for missing gender of the firstborn child, indicators for legitimate children with and without missing birth year, and unrelated co-rulers among previous monarchs, as well as polity and decade fixed effects. [c]Additionally, 12 lasso-selected (with Stata program *ivlasso*) interaction terms between pairs of the original dummy variables are included.

6.2 Female Leadership—Firstborn Gender Instrument

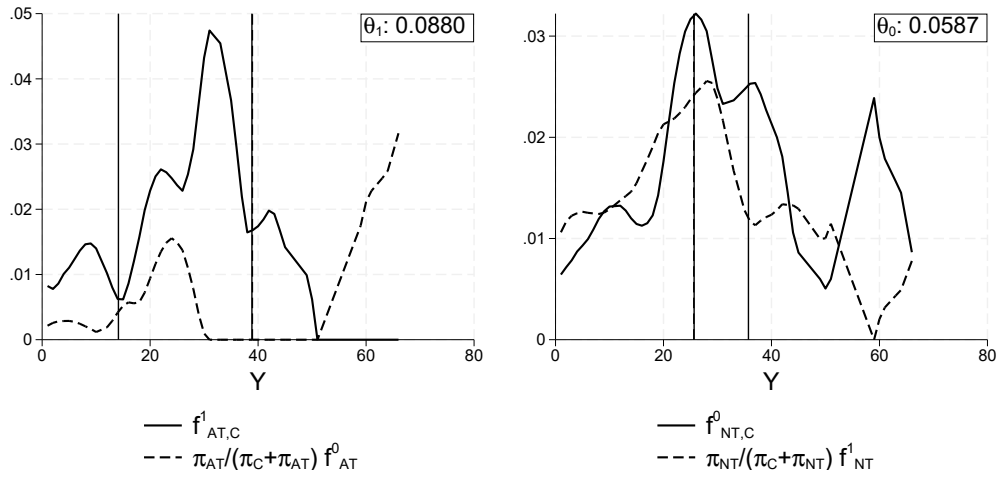
In the second application, we revisit the historical analysis by Dube and Harish (2020), which studies whether European states experienced more conflict under female rule between the fifteenth and twentieth centuries. The authors exploit institutional features of hereditary succession to identify the effect of female leadership on war participation.

Because older male children typically had priority in succession, the gender of the firstborn child of the previous monarch affected the probability that a queen would come to power. The empirical strategy therefore instruments female rule with the gender of the firstborn child to estimate a local average treatment effect of queenly rule on conflict outcomes. Identification relies on the assumption that the gender of the firstborn child is as good as random and therefore independent of potential outcomes. The original paper argues that this assumption is plausible because the gender of the first child is biologically determined and shows in several falsification tests that it does not predict war participation in the contemporaneous reign or in polities that never experienced female rule. However, the empirical specification conditions on a set of covariates related to the family structure of previous monarchs, such as the number of siblings and indicators for missing gender information. These variables are included to address alternative channels through which succession patterns could affect conflict. Consequently, the IV assumptions must hold conditional on these controls. If their functional form is misspecified—for example because the relationship between family structure and outcomes is more complex than captured by the baseline specification—this may affect the causal interpretation of the IV estimand, as emphasized by Blandhol et al. (2026), who also revisit this setting to examine potential bias from covariate misspecification.

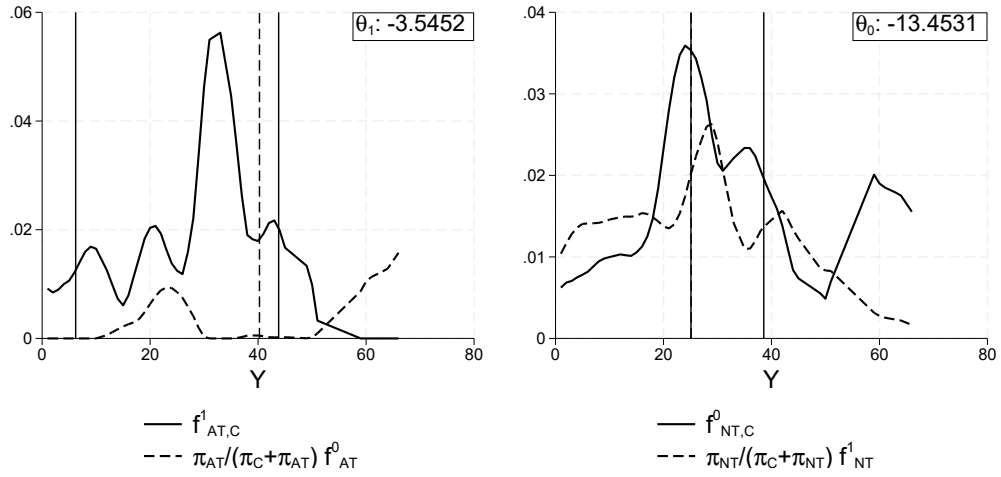
The dataset contains 193 reigns across 18 European polities between 1480 and 1913, corresponding to 3,586 polity–year observations, with queens ruling in about 18% of reigns.[17] The instrument Z indicates whether the previous monarch had a firstborn child who was male. The endogenous treatment variable D is an indicator for whether the current ruler is a queen. As outcome variable Y, we use the length of the reign in years, which is the only non-binary outcome available in the paper and therefore compatible with our testing procedure. The empirical specification follows column (3) of Table 3 in Dube and Harish (2020), which includes polity and decade fixed effects as well as baseline controls such as the number of siblings of the previous monarch and indicators for missing gender information among siblings and children. To examine the role of covariate specification, we apply our test under three specifications. First, we estimate the model without covariates. Second, we include the control variables used in the original paper. Third, we augment this specification with interaction terms between pairs of the original controls, selected via a lasso procedure (Stata command *ivlasso*) to allow for more flexible covariate relationships.

Figure 5 presents the graphical results. Panel (a) shows the densities and bounds of mean potential outcomes without covariates, panel (b) includes the original controls, and panel (c) additionally conditions on lasso-selected interaction terms between pairs of these controls. The estimates for the θs are displayed in the upper right corner of each graph.

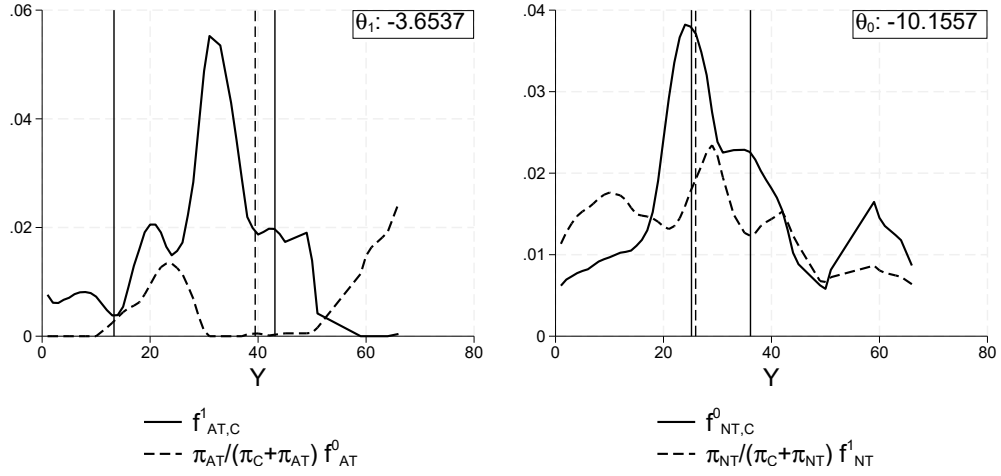
[17] The data are taken from the accompanying replication files in the Supplemental Material of Dube and Harish (2020). We use the dataset prepared for the estimation of Table 3 in the original paper and employ the same sample and variables, replacing the original war outcome with the reign length variable. The code implementing our test is provided in the Supplementary Material to facilitate replication.



(a) Without covariates



(b) With original covariates[b]



(c) With additional covariates[c]

Figure 5: Graphs for the firstborn gender instrument

Notes: Own illustration based on the data from Dube and Harish (2020). Bandwidth=3.00. PDFs for the mixed groups are shown by solid curves, and for the single groups by dashed curves. The vertical solid lines indicate the lower and upper bounds $\delta_{AT}^{1,LB/UB}$ (left) and $\delta_{NT}^{0,LB/UB}$ (right), and the vertical dashed lines display the conditional mean δ_{AT}^{0} (left) and δ_{NT}^{1} (right). f_{AT}^{0} and f_{NT}^{1} are down-weighted by their relative shares. This does not affect the mean potential outcome given by the dashed vertical line. [b]Covariates include dummy variables for missing gender of the firstborn child, indicators for legitimate children with and without missing birth year, and unrelated co-rulers among previous monarchs, as well as polity and decade fixed effects. [c]Additionally, 12 lasso-selected (with Stata program *ivlasso*) interaction terms between pairs of the original dummy variables are included.

Without covariates (panel a), the conditional means for the pure always- and never-taker groups lie slightly outside the bounds implied by the testing conditions, suggesting a potential violation of IV validity. Including the original covariates widens the bounds, which can be explained by larger estimated complier shares. The wider bounds now encompass the dashed lines, which is compatible with IV validity and correct covariate specification. However, the bounds—especially in the treated case—are relatively wide, and the result for the untreated case remains close to the boundary. Panel (c) therefore considers a more flexible specification including additional interaction terms. Visually, even though the bounds tighten, the deviation becomes slightly smaller as the dashed line moves somewhat farther away from the bounds.

Formal inference is reported in Table 3. Without covariates, the p-values for both θs and the Šidák correction exceed 0.1, so the null hypothesis of IV validity cannot be rejected. Including the original covariates yields somewhat larger p-values. However, this change must be interpreted jointly with the width of the bounds. Conditioning on the original controls also increases the estimated complier share and therefore widens the bounds. Hence, the higher p-values may reflect either improved support for the IV assumptions or simply weaker restrictions implied by the wider bounds. Allowing for additional interaction terms provides a clearer comparison. In this specification, the bounds become slightly tighter while the p-values increase marginally further. This pattern suggests that a more flexible covariate specification fits the data somewhat better. At the same time, the change is small, indicating that covariate misspecification is unlikely to play a major role in this application. Overall, the results do not indicate a rejection of IV validity. Because the bounds remain relatively wide, detecting violations is inherently difficult in this setting. Nevertheless, the evidence suggests that any bias of the LATE due to covariate misspecification is likely limited. This interpretation is broadly consistent with the findings of Blandhol et al. (2026), who also analyze this setting. While their RESET test rejects the condition for a weakly causal interpretation of the specification with a binary war outcome, the difference between their original IV estimate and a more flexible double/debiased machine learning estimator amounts to about 20.4% and is not statistically significant. Moreover, the magnitude of such differences may vary across outcomes.

7 Conclusion

This paper introduces an easily implementable testing procedure, based on distribution regressions, that evaluates the identifying assumptions underlying the LATE while flexibly incorporating covariates. We make two main contributions.

First, our test avoids reliance on fully nonparametric methods, which quickly suffer from the curse of dimensionality as the number of covariates increases. In many empirical

designs, a restricted set of covariates is an inherent and desirable feature—for example, to sustain a common trends assumption in difference-in-differences settings. However, such restrictions can induce specification bias, as emphasized by Blandhol et al. (2026). Second, we show that our test is able to detect precisely this form of specification bias when it matters for the IV estimand. In this sense, our approach unifies the literature on tests of IV validity with the emerging literature on specification errors in IV models (Blandhol et al., 2026; De Chaisemartin and d'Haultfoeuille, 2020; Goldsmith-Pinkham et al., 2024). Moreover, our framework naturally accommodates extensions based on the double machine-learning approach of Chernozhukov et al. (2018), thereby reducing the risk that IV estimates exhibit incorrect signs due to covariate misspecification (Blandhol et al., 2026).

Our procedure uses group-specific conditional distribution estimates to construct bounds on unobserved mean potential outcomes, which we compare to their observed counterparts to test the mean-based implications derived in Huber and Mellace (2015). Monte Carlo simulations that separately vary (i) violations of the instrument's conditional independence assumption, (ii) violations of the exclusion restriction, and (iii) the covariate-induced specification bias demonstrate that the test performs well in finite samples. We illustrate the empirical relevance of the test using two recent IV applications from the literature. The first uses a randomized training intervention as an instrument for mobile banking adoption in Bangladesh (Lee et al., 2021), and the second instruments female leadership with the gender of the firstborn child of the previous monarch to study conflict outcomes in historical Europe (Dube and Harish, 2020). In both settings, the test does not reject IV validity, although the results highlight the role of covariate specification and the difficulty of detecting violations when the implied bounds are wide. For comparison, we also revisit two classic IV applications—draft eligibility and college proximity—in the Appendix. For draft eligibility (Angrist, 1991), the test finds no evidence against IV validity. For college proximity (Card, 1993), the results depend on the inclusion of covariates: the test suggests violations without controls but not when the original covariates are included. Both findings are consistent with conclusions in the existing validity testing literature on these instruments. Overall, the paper demonstrates the usefulness of the proposed test both conceptually and in empirically relevant applications.

References

Angrist, J. D. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *American Economic Review*, 80(3):313–336.

Angrist, J. D. (1991). The Draft Lottery and Voluntary Enlistment in the Vietnam Era. *Journal of the American Statistical Association*, 86(415):584–595.

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.

Arai, Y., Hsu, Y.-C., Kitagawa, T., Mourifié, I., and Wan, Y. (2022). Testing identifying assumptions in fuzzy regression discontinuity designs. *Quantitative Economics*, 13(1):1–28.

Blandhol, C., Bonney, J., Mogstad, M., and Torgovitsky, A. (2026). When Is TSLS Actually LATE? *Review of Economic Studies*. Forthcoming.

Card, D. (1993). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. Working Paper 4483, National Bureau of Economic Research.

Carr, T. and Kitagawa, T. (2023). Testing Instrument Validity with Covariates. Papers, arXiv.org.

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. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V. and Hansen, C. (2005). An IV Model of Quantile Treatment Effects. *Econometrica*, 73(1):245–261.

De Chaisemartin, C. (2017). Tolerating Defiance? Local Average Treatment Effects without Monotonicity. *Quantitative Economics*, 8(2):367–396. Publisher: Wiley Online Library.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–2996.

Dong, Y. (2019). Regression Discontinuity Designs with Sample Selection. *Journal of Business & Economic Statistics*, 37(1):171–186.

Dube, O. and Harish, S. P. (2020). Queens. *Journal of Political Economy*, 128(7):2579–2652.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press.

Farbmacher, H., Guber, R., and Klaassen, S. (2022). Instrument Validity Tests With Causal Forests. *Journal of Business & Economic Statistics*, 40(2):605–614.

Frandsen, B. R., Frölich, M., and Melly, B. (2012). Quantile Treatment Effects in the Regression Discontinuity Design. *Journal of Econometrics*, 168(2):382–395.

Goldsmith-Pinkham, P., Hull, P., and Kolesár, M. (2024). Contamination Bias in Linear Regressions. *American Economic Review*, 114(12):4015–4051.

Hall, P. and Wilson, S. R. (1991). Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics*, 47(2):757–762.

Huber, M. and Mellace, G. (2014). Stata Code for "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints".

Huber, M. and Mellace, G. (2015). Testing Instrument Validity for LATE Identification Based on Inequality moment Constraints. *The Review of Economics and Statistics*, 97(2):398–411.

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.

Kédagni, D. and Mourifié, I. (2020). Generalized Instrumental Inequalities: Testing the Instrumental Variable Independence Assumption. *Biometrika*, 107(3):661–675.

Laffers, L. and Mellace, G. (2017). A note on testing instrument validity for the identification of LATE. *Empirical Economics*, 53:1281–1286.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies*, 76(3):1071–1102.

Lee, J. N., Morduch, J., Ravindran, S., Shonchoy, A., and Zaman, H. (2021). Poverty and migration in the digital age: Experimental evidence on mobile banking in bangladesh. *American Economic Journal: Applied Economics*, 13(1):38–71.

Lee, J. N., Morduch, J., Ravindran, S., Shonchoy, A. S., and Zaman, H. (2020). Data and code for: Poverty and migration in the digital age: Experimental evidence on mobile banking in bangladesh. ICPSR - Interuniversity Consortium for Political and Social Research.

Machado, C., Shaikh, A. M., and Vytlacil, E. J. (2019). Instrumental Variables and the Sign of the Average Treatment Effect. *Journal of Econometrics*, 212(2):522–555.

MacKinnon, J. G. (2009). *Bootstrap Hypothesis Testing*, chapter 6, pages 183–213. John Wiley & Sons, Ltd.

Mogstad, M. and Torgovitsky, A. (2024). Chapter 1—Instrumental Variables with Unobserved Heterogeneity in Treatment Effects. In Dustmann, C. and Lemieux, T., editors, *Handbook of Labor Economics*, volume 5, pages 1–114. Elsevier.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2021). The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables. *American Economic Review*, 111(11):3663–98.

Mourifié, I. and Wan, Y. (2017). Testing Local Average Treatment Effect Assumptions. *The Review of Economics and Statistics*, 99(2):305–313.

Semenova, V. (2026). Generalized Lee Bounds. *Journal of Econometrics*. Forthcoming.

Šidák, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Słoczyński, T. (2025). When Should We (not) Interpret Linear IV Estimands as LATE? *Review of Economic Studies*. Forthcoming.

Sun, Z. (2023). Instrument Validity for Heterogeneous Causal Effects. *Journal of Econometrics*, 237(2, Part A):105523.

Wan, Y. and Mourifie, I. (2016). Replication data for: "Testing Local Average Treatment Effect Assumptions". Harvard Dataverse, V1.

Westphal, M., Kamhöfer, D. A., and Schmitz, H. (2022). Marginal College Wage Premiums under Selection into Employment. *The Economic Journal*, 132(646):2231–2272.

Appendix

A Additional Figures

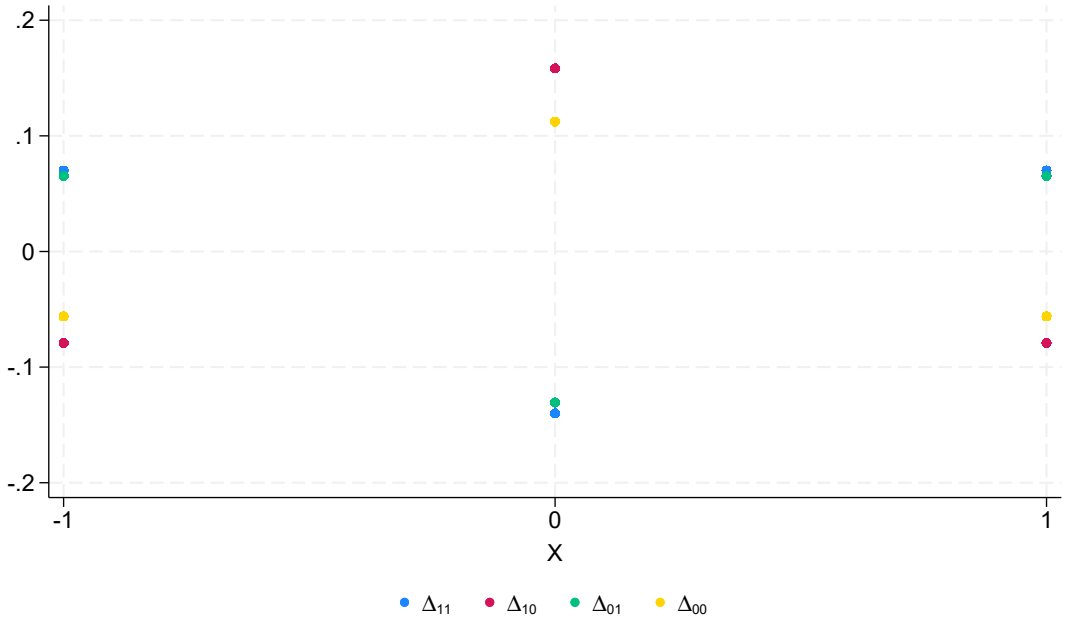


Figure A.1: Specification errors in the numerical example of Section 5.2

Notes: Own illustration based on simulated data. The graph shows specification errors $\Delta_{dz} = \mathbb{E}(D_d Z_z \mid X) - \mathbb{L}(D_d Z_z \mid X)$ and how they vary over the values of X for each observable group, defined by the possible combinations of D and Z.

B Taylor expansion

We want to simplify Eq. (15), that is,

$$\frac{\mathbb{E}\left(Y\,\widetilde{D_d Z_z}\right)}{\mathbb{E}\left(\widetilde{D_d Z_z}\right)^2} = \frac{\mathbb{E}(Y \mid D=d, Z=z)\mathbb{E}(D_d Z_z) + \mathbb{E}\left[\mathbb{E}(Y \mid X)\Delta_{dz}(X)\right]}{\mathbb{E}(D_z Z_z) + 2\mathbb{E}\left[\mathbb{E}(D_d Z_z \mid X)\Delta_{dz}(X)\right] + \mathbb{E}\left[\mathbb{E}\left((\Delta_{dz}(X))^2\right)\right]}.$$

For this, first, define

$$\begin{aligned}
R_{dz}(\Delta) &\equiv \frac{\mu_{dz}\, p_{dz} + \mathbb{E}\big[g(X)\Delta_{dz}(X)\big]}{p_{dz} + 2\mathbb{E}\big[m_{dz}(X)\Delta_{dz}(X)\big] + \mathbb{E}\big[\Delta_{dz}(X)^2\big]}, \text{ with} \\
\mu_{dz} &\equiv \mathbb{E}(Y \mid D=d, Z=z), \\
p_{dz} &\equiv \mathbb{E}(D_d Z_z), \\
g(X) &\equiv \mathbb{E}(Y \mid X), \\
m_{dz}(X) &\equiv \mathbb{E}(D_d Z_z \mid X).
\end{aligned}$$

The goal is to apply a first-order Taylor expansion around $R_{dz}(0)$, which is a correct approximation for small deviations $|\Delta| > 0$. To see that it can be applied, we write

$$R_{dz}(\Delta) = \frac{A_0 + a}{B_0 + b + O(||\Delta^2||)},$$

with $A_0 = \mu_{dz} p_{dz}$ and $B_0 = p_{dz}$, that is, the respective values of the numerator and denominator when $\Delta = 0$, and a and b, the respective numerator- and denominator-specific linear impacts of Δ. We then use the fact that $O(||\Delta^2||)$ can be ignored in a first-order Taylor expansion around $\frac{A_0}{B_0}$, yielding approximately

$$\frac{A_0 + a}{B_0 + b} \approx \frac{A_0}{B_0} + \frac{a}{B_0} - \frac{A_0}{B_0^2} b.$$

The first term, $\frac{A_0}{B_0}$, is the baseline level when $a = b = 0$ (i.e., $\Delta = 0$). The second term, $\frac{a}{B_0}$, is the partial impact of Δ on the numerator, while the last term, $-\frac{A_0}{B_0^2} b$, is the partial impact of Δ on the denominator. Applied back to $R_{dz}(\Delta)$, this expansion reads

$$R_{dz}(\Delta) \approx \mu_{dz} + \tfrac{1}{p_{dz}}\mathbb{E}\big[g(X)\Delta_{dz}(X)\big] - \tfrac{2\mu_{dz}}{p_{dz}}\mathbb{E}\big[m_{dz}(X)\Delta_{dz}(X)\big].$$

Hence, the first-order bias satisfies

$$\text{bias}(D=d, Z=z) \;\equiv R_{dz}(\Delta) - \mu_{dz} \approx \tfrac{1}{p_{dz}}\Big(\mathbb{E}\big[g(X)\Delta_{dz}(X)\big] - 2\mu_{dz}\mathbb{E}\big[m_{dz}(X)\Delta_{dz}(X)\big]\Big).$$

We aim to assess how misspecification bias affects the test statistic for mean differences between $Z=1$ and $Z=0$ when $D=d$. Therefore, we define $\Delta\text{bias}(D=d) \equiv \text{bias}(D=d, Z=1) - \text{bias}(D=d, Z=0)$ and plug in the expression derived above:

$$\begin{aligned}\Delta\text{bias}(D=d) \;\approx\; & \frac{\mathbb{E}[g(X)\Delta_{d1}(X)] - 2\mu_{d1}\mathbb{E}[m_{d1}(X)\Delta_{d1}(X)]}{p_{d1}} \\ & - \frac{\mathbb{E}[g(X)\Delta_{d0}(X)] - 2\mu_{d0}\mathbb{E}[m_{d0}(X)\Delta_{d0}(X)]}{p_{d0}}.\end{aligned}$$

Using the original expressions again:

$$\begin{aligned}&\Delta\text{bias}(D=d) \\ &\quad\approx \frac{\mathbb{E}[\mathbb{E}(Y \mid X)\Delta_{d1}(X)] - 2\mathbb{E}(Y \mid D=d, Z=1)\mathbb{E}[\mathbb{E}(D_d Z_1 \mid X)\Delta_{d1}(X)]}{\mathbb{E}(D_d Z_1)} \\ &\qquad - \frac{\mathbb{E}[\mathbb{E}(Y \mid X)\Delta_{d0}(X)] - 2\mathbb{E}(Y \mid D=d, Z=0)\mathbb{E}[\mathbb{E}(D_d Z_0 \mid X)\Delta_{d0}(X)]}{\mathbb{E}(D_d Z_0)}.\end{aligned}$$

We aim to simplify this expression to identify the most important forces that govern it. We first define normalized weights

$$\omega_{dz}(X) = \frac{\mathbb{E}(D_d Z_z \mid X)}{\mathbb{E}(D_d Z_z)},$$

with $\mathbb{E}(\omega_{dz}(X)) = 1$. Using this definition, we can write the bias as

$$\begin{aligned}\Delta\text{bias}(D=d) \;\approx\; & \mathbb{E}\left\{\left[\frac{\mathbb{E}[\Delta_{d1}(X)]}{\mathbb{E}(D_d Z_1)} - \frac{\mathbb{E}[\Delta_{d0}(X)]}{\mathbb{E}(D_d Z_0)}\right]\mathbb{E}(Y \mid X)\right\} \\ & -2\mathbb{E}\Big[\mathbb{E}(Y \mid D=d, Z=1)\omega_{d1}(X)\Delta_{d1}(X) \\ & \qquad -\mathbb{E}(Y \mid D=d, Z=0)\omega_{d0}(X)\Delta_{d0}(X)\Big].\end{aligned}$$

To further simplify, we add and subtract $2\mathbb{E}\big[\mathbb{E}(Y \mid D = d, Z = 1)\omega_{d0}(X)\Delta_{d0}(X)\big]$. With this trick, we can identify three components of bias, as the expression becomes:

$$\begin{aligned}\Delta\text{bias}(D = d) \;\approx\; & \mathbb{E}\left\{\left[\frac{\mathbb{E}[\Delta_{d1}(X)]}{\mathbb{E}(D_d Z_1)} - \frac{\mathbb{E}[\Delta_{d0}(X)]}{\mathbb{E}(D_d Z_0)}\right]\mathbb{E}(Y \mid X)\right\} \\ & -2\mathbb{E}(Y \mid D = d, Z = 1)\mathbb{E}\Big[\omega_{d1}(X)\Delta_{d1}(X) - \omega_{d0}(X)\Delta_{d0}(X)\Big] \\ & -2\Big[\mathbb{E}(Y \mid D = d, Z = 1) - \mathbb{E}(Y \mid D = d, Z = 0)\Big]\mathbb{E}\big[\omega_{d0}(X)\Delta_{d0}(X)\big].\end{aligned}$$

C Two Classic Applications

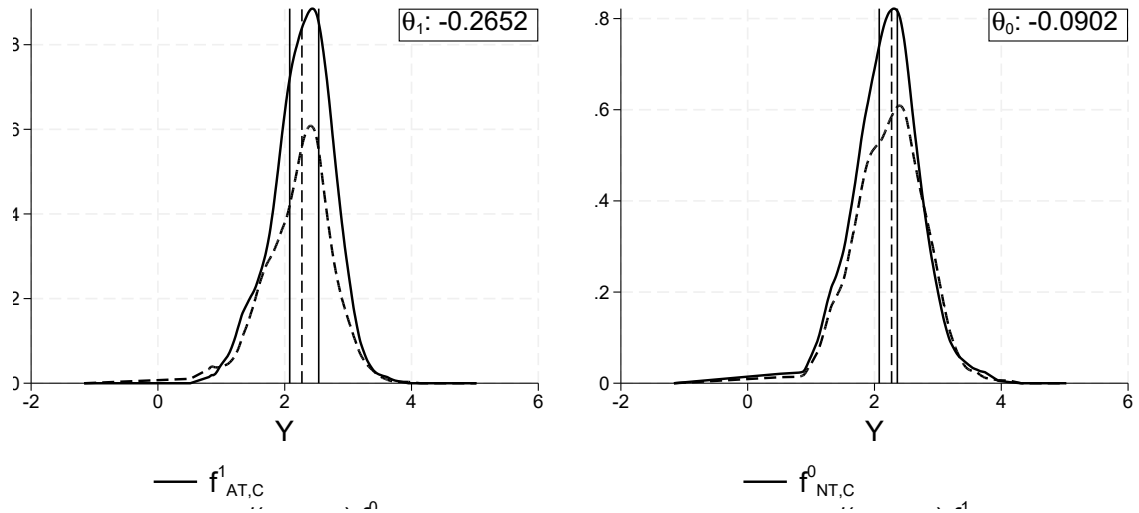
We apply our testing approach to two well-known empirical settings that have also been analyzed by Mourifié and Wan (2017), Kitagawa (2015), Sun (2023), Carr and Kitagawa (2023), and Huber and Mellace (2015). Using these applications allows us to compare our results with existing IV validity tests. The first relies on the Vietnam-era draft lottery instrument from Angrist (1991), and the second uses college proximity as an instrument following Card (1993).

In the first application, we use the draft eligibility instrument from Angrist (1991) to study the effect of veteran status on earnings. Military service may be endogenous due to self-selection, while the draft lottery provides a plausibly exogenous source of variation because eligibility was randomly assigned based on birth dates. This supports the conditional independence assumption, and monotonicity is also highly plausible. A potential concern is the exclusion restriction, as draft-eligible men might have attempted to defer or avoid service, for instance by staying longer in education, which could affect wages. We use data from the 1984 Survey of Income and Program Participation (SIPP).[18] The final sample without missings consists of 3,071 individuals. The treatment indicator D equals one for veteran status, the instrument Z indicates draft eligibility, and the outcome Y is the logarithm of weekly wages. Following Angrist (1990), we include birth-cohort dummies and a race indicator as covariates. As this specification is rather sparse, the type of covariate misspecification discussed in Blandhol et al. (2026) is unlikely to be a major concern in this setting.

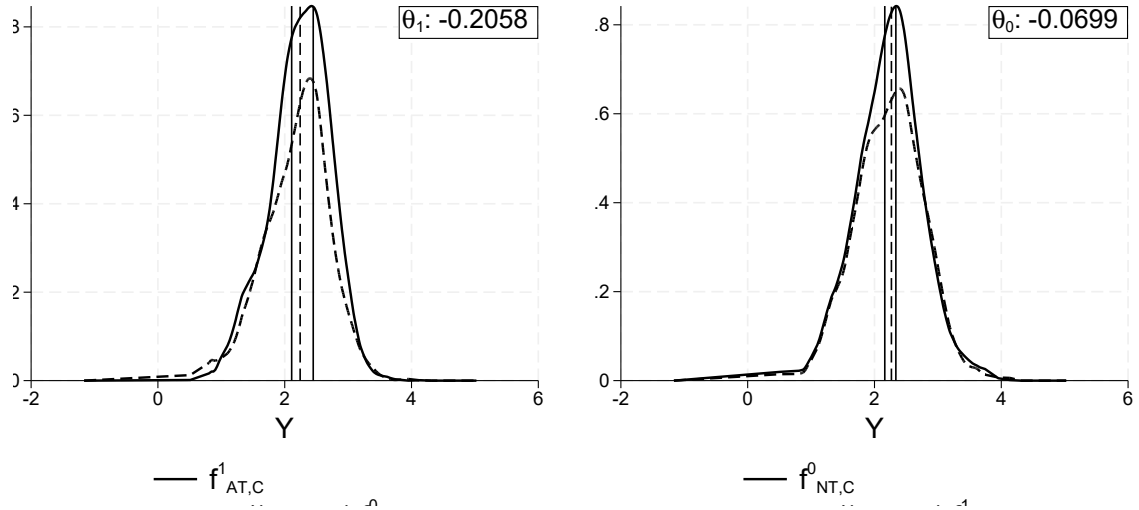
Figure C.1 presents graphical results without and with covariates. In both cases, the dashed vertical lines lie within the bounds indicated by the solid vertical lines, suggesting that the validity conditions hold. Conditioning on covariates slightly narrows the bounds. Table C.1 confirms this graphical evidence. The estimates of θ_1 and θ_0 are negative both with and without covariates, and the corresponding p-values as well as the Šidák-corrected p-values equal one. Hence, we cannot reject IV validity (or correct covariate specification when covariates are included). This finding is consistent with Kitagawa (2015) and Mourifié and Wan (2017).

The second application follows Card (1993), who study the returns to college education using college proximity as an instrument. Educational attainment may be endogenous due to unobserved factors such as ability that affect both schooling and wages. The instrument exploits the idea that living near a college lowers the cost of attending, while a key identifying assumption is that unobserved determinants of earnings are independent of adolescents' residential location. The data are drawn from the National Longitudinal

[18]The data set is available in the Review of Economics and Statistics Dataverse (Wan and Mourifie, 2016) as replication data for Mourifié and Wan (2017). Stata files for replication of our results are provided in the supplementary material.



(a) Without covariates



(b) With covariates[a]

Figure C.1: Graphs for the draft lottery instrument

Notes: Own illustration based on SIPP data. Bandwidth=0.15. Pdfs for the mixed groups are given by the solid curves, and for the single groups, they are given by the dashed curves. The vertical solid lines indicate the lower and upper bounds $\delta_{AT}^{1,LB/UB}$ (left) and $\delta_{NT}^{0,LB/UB}$ (right), and the vertical dashed lines display the conditional mean δ_{AT}^{0} (left) and δ_{NT}^{1} (right). f_{AT}^{0} and f_{NT}^{1} are down-weighted by their relative shares. This does not affect the mean potential outcome given by the dashed vertical line. [a]Dummies for the birth cohort and a dummy for being non-white are used as covariates.

Survey of Young Men (NLSYM).[19] Following Kitagawa (2015), we define a binary treatment for having at least 16 years of education in 1976.[20] The instrument indicates whether the individual grew up near a four-year college, and the outcome is the logarithm of weekly earnings in 1976. We first apply the test without covariates and then include the pre-

[19]The prepared dataset is available in the Review of Economics and Statistics Dataverse (Wan and Mourifie, 2016) as replication data for Mourifié and Wan (2017). Stata files for replication of our results are provided in the supplementary material.

[20]Coarsening the treatment variable may affect instrument validity through coarsening bias (see Sun, 2023). However, Carr and Kitagawa (2023) argue that this is unlikely to be problematic for the college proximity instrument in this dataset.

Table C.1: Results of the empirical applications

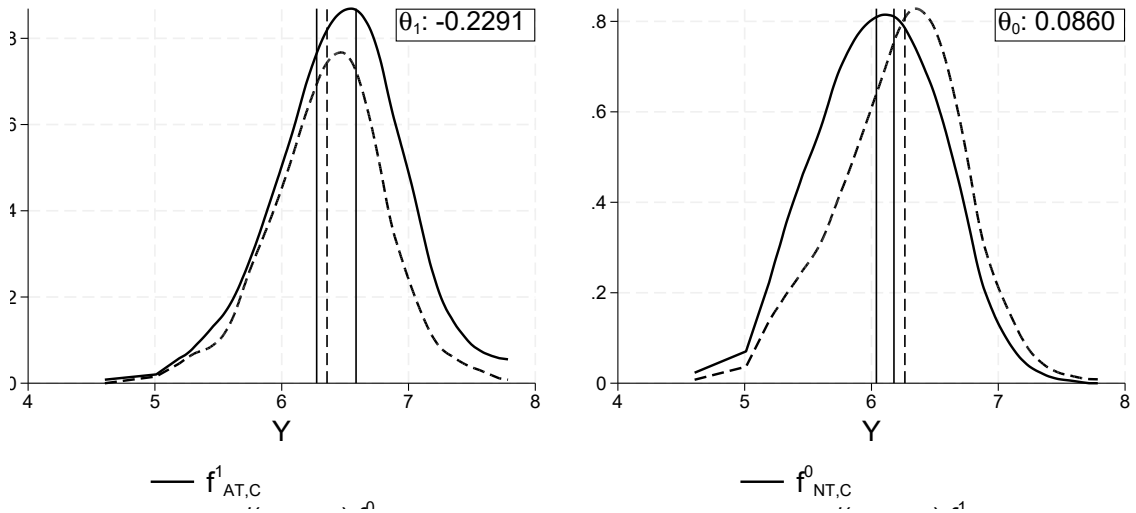
	Draft lottery (Angrist, 1991)		College proximity (Card, 1993)	
	w/o covariates	w/ covariates[a]	w/o covariates	w/ covariates[b]
θ_1	-0.295	-0.219	-0.233	-0.110
$p_{\hat{\theta}_1}$	1.000	1.000	1.000	0.996
θ_0	-0.109	-0.086	0.086	0.016
$p_{\hat{\theta}_0}$	1.000	1.000	0.002	0.323
Šidák corrected $\hat{p}$	1.000	1.000	0.004	0.541
Shares				
π_C	0.139	0.088	0.069	0.035
π_{AT}	0.265	0.288	0.225	0.248
π_{NT}	0.596	0.623	0.707	0.718
Bandwidth	0.15		0.20	
Observations	3,027		3,010	

Notes: Tests are based on 499 bootstrap samples.We use 260 and 360 evaluation points for the draft lottery and the college proximity application, respectively. [a]Dummies for birth cohorts and a dummy for non-white. [b]Dummy variables indicating race being black, residence in a standard metropolitan area (SMSA) in 1966 and 1976, region of residence in 1966, living in the south in 1976, living with both parents at age 14, and living with the mother only at age 14. Variables representing parents' years of education take on the overall mean when missing. Dummies for missing fathers' and mothers' education have also been added.

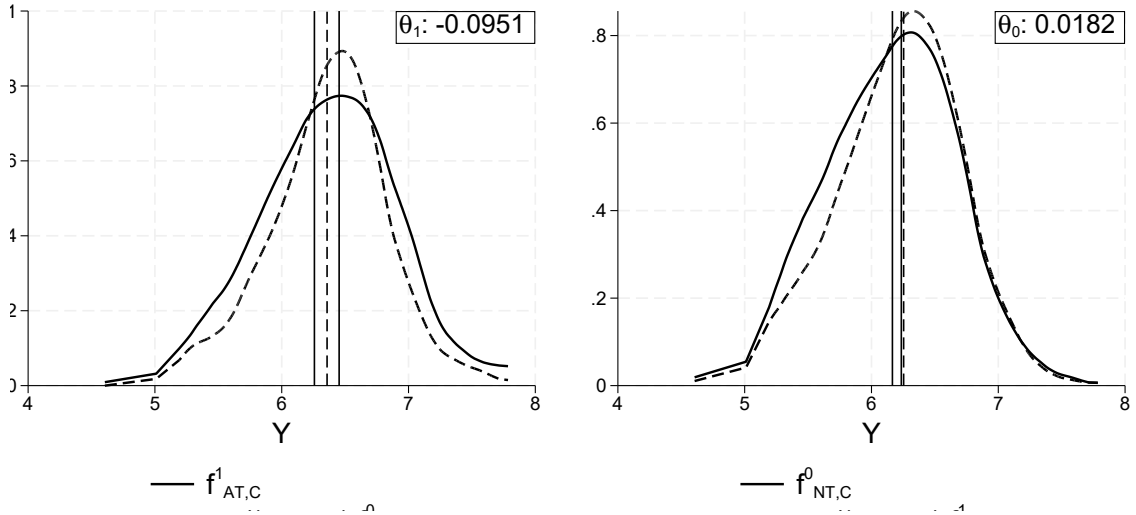
treatment covariates used by Card (1993), excluding interactions of parental education.[21] As shown in Blandhol et al. (2026) for a similar specification with years of education as a non-binary treatment, the results for this instrument are not strongly affected by potential covariate misspecification. The final sample consists of 3,010 observations.

Figure C.2 shows that without covariates the validity condition for the untreated state is violated, as the estimated mean δ^1_{NT} lies outside the bounds for δ^0_{NT}. Including covariates narrows the bounds and substantially reduces this deviation. Table C.1 reports the corresponding inference results. Without covariates, the p-value for θ_0 equals 0.002 and the Šidák-corrected p-value equals 0.004, leading to a rejection of IV validity. After conditioning on covariates, the p-values increase to 0.323 and 0.541, so that the null hypothesis cannot be rejected. Importantly, this increase in the p-values coincides with tighter bounds, indicating that the improved test results are not driven by mechanically weaker restrictions but are consistent with the identifying assumptions holding more plausibly once the original controls are included. Nevertheless, a small deviation from the testing conditions remains visible in the sample. Although this deviation is not statistically significant, it suggests that the identifying assumptions may not hold exactly. Taken together, the results therefore do not provide clear evidence against IV validity (or the correct specification of covariates) once covariates are included, but they also indicate that the conditions may only hold approximately in the sample. This pattern suggests that any resulting bias in the LATE is likely small. This interpretation aligns with the conclusions of Kitagawa (2015),

[21]The detailed list of covariates is reported in the notes of Table C.1.



(a) Without covariates



(b) With covariates[a]

Figure C.2: Graphs for the college proximity instrument

Notes: Own illustration based on NLSYM data. Bandwidth=0.20. PDFs for the mixed groups are shown by solid curves, and for the single groups by dashed curves. The vertical solid lines indicate the lower and upper bounds $\delta_{AT}^{1,LB/UB}$ (left) and $\delta_{NT}^{0,LB/UB}$ (right), and the vertical dashed lines display the conditional mean δ_{AT}^{0} (left) and δ_{NT}^{1} (right). f_{AT}^{0} and f_{NT}^{1} are down-weighted by their relative shares. This does not affect the mean potential outcome given by the dashed vertical line. [a]Dummy variables indicating race being black, residence in a standard metropolitan area (SMSA) in 1966 and 1976, region of residence in 1966, living in the south in 1976, living with both parents at age 14, and living with the mother only at age 14. Variables representing parents' years of education assume the overall mean when missing. Additionally, dummy variables for missing fathers' and mothers' education are added.

Huber and Mellace (2015), and Carr and Kitagawa (2023), while Mourifié and Wan (2017) find evidence against IV validity when testing across subsamples with a more limited set of controls. It is also consistent with the findings of Blandhol et al. (2026), who report only a modest relative difference between the baseline IV estimate and more flexible estimators designed to address potential covariate misspecification.