Ashoka University
Economics Discussion Paper 170

What No First Stage Can Detect: Functional-Form Contamination in Linear IV

September 2026

Parush Arora, Ashoka University

https://ashoka.edu.in/economics-discussionpapers

What No First Stage Can Detect: Functional-Form Contamination in Linear IV

Parush Arora[1]

Abstract

Applied instrumental variables (IV) practice reports a first-stage F, now often the conditional F of Sanderson and Windmeijer (2016), and reads a large value as a license to interpret the second stage. We show that no first-stage diagnostic can deliver that license. With a scalar instrument, a scalar treatment, and covariates entered linearly, the 2SLS estimand splits into a signal that a saturated specification would target and a contamination, the covariance between curvature in the instrument propensity and a level function of the covariates. The same nuisance sits in both terms, so it biases the estimand and inflates the reported strength at once. When the instrument is nearly collinear with the covariates the signal vanishes and the strength is manufactured. When the strength is honest the curvature still biases the estimand through the outcome, where no first-stage number reaches it. We prove that no functional of the joint distribution of instrument, treatment, and covariates can detect this second bias, and we build a directed test from reduced-form regressions, a corrected estimator, and a reportable contamination share. Two applications show both failures. A husband's insurance instrument (Olson, 1998) has a conditional F above 36,000, yet correcting a linear income control more than doubles the estimate. The instrument of Nunn and Wantchekon (2011) loses most of its first stage once geography enters flexibly.

Keywords: Instrumental variables, two-stage least squares, local average treatment effect, first-stage strength, contamination bias, functional form, specification testing, conditional moment test

JEL: C26, C21, C52

[1]Department of Economics, Ashoka University, Sonipat, India.

1. Introduction

The instrumental-variable (IV) estimator is the workhorse of applied microeconomics whenever treatment is suspected to be endogenous. Its credibility rests on two pillars, the exclusion restriction and the strength of the first stage. The first is fundamentally untestable and is defended by argument and by sensitivity analysis (Conley et al., 2012; van Kippersluis and Rietveld, 2018). The second is treated as a statistical object to be certified. A large literature, including Bound et al. (1995), Staiger and Stock (1997), Stock and Yogo (2005), Montiel Olea and Pflueger (2013), and Andrews et al. (2019), has produced a battery of rule-of-thumb and size-based procedures for diagnosing weak instruments, and reporting a first-stage F is now routine, indeed mandatory in most journals. When the model carries a rich set of controls, Sanderson and Windmeijer (2016) show that the relevant object is the conditional first-stage F, computed after partialling the covariates out of the excluded instrument, and applied work increasingly reports it. Whether a large value of it should reassure is a separate question, and Stevenson (2026) argues on behavioural grounds that it need not, a first-stage F above 200 coexisting with substantial bias once researcher discretion enters. We give a mechanical counterpart that binds even for a single disciplined analyst running the standard specification.

Running in parallel, a second literature has clarified what a valid, strong instrument identifies. Since Imbens and Angrist (1994) and Angrist et al. (1996), IV is understood to recover a local average treatment effect (LATE) under treatment-effect heterogeneity. Heckman and Vytlacil (2005) formalize essential heterogeneity, de Chaisemartin (2017) treats failures of monotonicity, and Mogstad and Torgovitsky (2018, 2024) synthesize the modern picture. Most consequential for practice is Blandhol et al. (2022). Once covariates enter the second stage linearly, the near-universal specification, 2SLS is in general not a convex average of LATEs unless the covariates are entered nonparametrically. Zhao et al. (2025) show that consistent interacted 2SLS requires the product of the IV propensity score and the covariates to be linear in the covariates. It is by now understood that linearity of the instrument propensity $\mathbb{E}[Z \mid X]$ in the covariates is the pivotal condition, and Blandhol et al. (2022) recommend assessing it directly with a Ramsey RESET on $\mathbb{E}[Z \mid X]$, a diagnostic on (Z, X) alone. Słoczyński et al. (2026) synthesize these diagnostics.

We study linear IV with a scalar instrument Z, a scalar treatment D, and covariates entered linearly, under conditional-on-X validity. The population 2SLS coefficient on D is the Frisch–Waugh–Lovell ratio $\beta_{2SLS} = \mathbb{E}[\tilde{Z}Y]/\mathbb{E}[\tilde{Z}D]$, where $\tilde{Z} = Z - L(Z \mid 1, X)$ is the covariate-residualized instrument. Writing the residual as $\tilde{Z} = (Z - e(X)) + h(X)$, with $e(x) = \mathbb{E}[Z \mid X = x]$ the instrument propensity and $h = e - \ell$ its component orthogonal to the linear covariate span, delivers an exact decomposition,

$$\beta_{2SLS} = \frac{\overbrace{\mathbb{E}[\mathrm{Cov}(Z, Y \mid X)]}^{\text{signal } S_Y} + \overbrace{\mathrm{Cov}(h(X), m_Y(X))}^{\text{contamination } \theta_Y}}{\underbrace{\mathbb{E}[\mathrm{Cov}(Z, D \mid X)]}_{\text{signal } S_D} + \underbrace{\mathrm{Cov}(h(X), m_D(X))}_{\text{contamination } \theta_D}}, \tag{1}$$

where $m_D(x) = \mathbb{E}[D \mid X = x]$ and $m_Y(x) = \mathbb{E}[Y \mid X = x]$. The signal is the average within-cell covariance of the instrument with the outcome and with the treatment, the estimand a fully interacted (saturated) specification would target. Each contamination term is a covariance between

the propensity curvature h and a covariate level function, and so vanishes exactly when h is orthogonal to that level function, which a propensity linear in X ($h \equiv 0$) or a linear level function delivers as a sufficient case. The binary-treatment LATE model is the leading special case, where $\mathrm{Cov}(Z, D \mid X) = e(X)\{1 - e(X)\}p(X)$ with p the conditional first stage and the ratio S_Y/S_D is a convex average of conditional LATEs. The contamination, the strength–bias identity, the impossibility theorem, the directed test, the correction, and the reportable share developed below are properties of linear IV with a linear covariate specification, and none of them turns on the support of Z or D.

The bias throughout is measured against $\bar{L} = S_Y/S_D$, the estimand of the saturated specification that Blandhol et al. (2022) recommend, so $\beta_{2SLS} - \bar{L}$ is the price of the linear covariate shortcut relative to that alternative. In the binary case $\bar{L}$ is a convex average of conditional LATEs and the shortcut can push the estimand outside the LATE hull, which is why we run that case in full. Outside the binary case $\bar{L}$ is a functional-form benchmark rather than a causal average, and the results below concern this gap and hold whatever $\bar{L}$ means.

We ask what the first-stage F, the diagnostic every IV paper reports, says about this bias. The answer is nothing, and worse than nothing. The contamination is the same quantity that enters the first-stage covariance the conditional F is built from, so the F co-moves with the bias and can be manufactured by the very curvature that produces it (Proposition 1). We sharpen this into an impossibility. No functional of the law of (Z, D, X), hence no first-stage strength statistic, can detect the contamination, because the bias loads on the conditional law of the outcome (Proposition 3). This matters because the F is read as the near-mandatory certificate licensing the second stage, and a propensity-only check such as the RESET of Blandhol et al. (2022) cannot repair it, since whether nonzero curvature is harmful turns on the outcome such a check never sees.

The knife-edge that Blandhol et al. (2022) identify, linearity of the propensity in X, is the case $h \equiv 0$ in which our contamination vanishes, and we characterize the gap that opens once it fails. Three papers are adjacent but study different objects. Słoczyński (2022) analyzes the sign of the LATE weights in the just-identified design, a question about the causal reading of $\bar{L}$ itself, while our object is the gap $\beta_{2SLS} - \bar{L}$ and whether any first-stage number can detect it. Goldsmith-Pinkham et al. (2024) treat contamination across several treatments, a channel that survives a saturated propensity, while ours is a single-treatment functional-form channel that vanishes under saturation, so the two compound in a design with both. Krumme and Westphal (2026), in concurrent work, also direct a diagnostic at the outcome rather than at the propensity alone, extending testable-inequality implications for IV validity to a joint null of validity and correct specification under a strong first stage, whereas our restriction conditions on validity, ties to the first-stage strength, and stays valid under weak identification.

Because the contamination sits in both the numerator and the denominator of (1), it produces two distinct failure modes. In **mode (a)**, manufactured strength, the instrument is nearly collinear with the covariates, the signal S_D vanishes, θ_D dominates the denominator, and the conditional F is inflated by the very curvature that biases the estimand (Section 3). In **mode (b)**, invisible outcome-side bias, the strength is honest, θ_D is small and the F large, yet the estimand is biased through the numerator term θ_Y, which loads on the outcome where no first-stage number can reach

it.

First, Proposition 2 characterizes collinearity between instrument and covariates, the pure form of mode (a). Along any sequence in which $\mathrm{Var}(Z \mid X) \to 0$, the signal vanishes while the contamination does not, and the estimand converges to $\mathrm{Cov}(h^*, m_Y^*)/\mathrm{Cov}(h^*, m_D^*)$, a ratio of covariate contrasts with no causal content. Second, Corollary 1 shows the conditional first-stage slope stays bounded away from zero along the same sequence, so the conditional F diverges in n rather than collapsing, and Proposition 3 turns this into the impossibility result above, which covers the conditional F, the effective F, and the Cragg–Donald statistic in particular. Third, we give a directed conditional-moment test of the null of no contamination, built from reduced-form regressions alone (Section 4). It has power exactly where the F is silent, and its size is controlled even under weak identification, in both the screening form and a signal-cleared bias-targeting form whose validity holds uniformly across the collinearity range, where a saturate and compare contrast collapses. Fourth, the same identity yields a bias-corrected estimator with an exact no-free-lunch cost, the identifying variation it surrenders equalling the manufactured share, and a reportable contamination share ϱ to be read beside the conditional F (Section 5). The share measures the first-stage channel alone, and like the F it stays silent about the outcome-side bias, so a small ϱ narrows but does not certify a design, a point the health insurance application makes concrete.

Each result is stated for the general model, with the binary LATE case noted as it arises. The general model is a scalar treatment of arbitrary support. A vector of treatments, where the contamination studied here compounds with the cross-treatment channel of Goldsmith-Pinkham et al. (2024), is the one boundary we leave to future work. Section 6 confirms them in Monte Carlo designs spanning a multivalued (ordered) treatment, a continuous instrument with a continuous treatment, and an over-identified design in which the screen stacks moments across instruments. In each, as $\mathrm{Var}(Z \mid X) \to 0$ the estimand leaves the hull of the within-cell IV slopes while the conditional F diverges, and the directed test fires. Section 7 works one application in full and adds two further designs. The flagship, the effect of spousal health insurance coverage on wives' labour supply (Olson, 1998), is a clean instance of mode (b). A conditional F above 36,000 and a contamination share of 0.09 are both reassuring, yet entering husband income linearly, the standard choice, more than doubles the estimated effect once corrected, and the corrected estimator stays fully identified. The slave trade instrument of Nunn and Wantchekon (2011), with a continuous instrument and a continuous treatment, is the real data face of mode (a), where correcting the propensity collapses a first stage that looked powerful, and the 401(k) eligibility design corroborates the mode (b) pattern in the most studied strong-instrument application in the literature.

Section 2 sets up the general model and proves the decomposition and the strength–bias identity. Section 3 develops the collinearity limit, the failure of the conditional F, and the impossibility theorem. Section 4 constructs the test. Section 5 presents the bias-corrected estimator and the identification trade-off. Section 6 reports the Monte Carlo evidence and Section 7 the applications. Section 8 gives recommendations for practice and Section 9 concludes. Proofs are collected in the Appendix.

2. Framework and the strength–bias identity

Let $Z \in \mathbb{R}$ be a scalar instrument, $D \in \mathbb{R}$ a scalar treatment, $Y \in \mathbb{R}$ an outcome, and $X \in \mathbb{R}^{d_x}$ a vector of covariates. The decomposition that follows uses only the two maintained assumptions below, a linear specification and regularity, and no structure on treatment assignment, so in particular no binariness or monotonicity. For the interpretation of the signal ratio as a convex average of local average treatment effects, and hence for the binary-treatment specialization we return to throughout, we additionally invoke the standard conditional-on-X LATE assumptions. We state them first, so that later results may reference them, but they are used only for that interpretation. In the LATE model potential treatments are $D(0), D(1)$ and potential outcomes $Y(d, z)$, with observed $D = D(Z)$ and $Y = Y(D, Z)$.

Assumption 1 (Conditional validity)**.** $\{Y(d, z), D(z)\}_{d,z} \perp\!\!\!\perp Z \mid X$*, and* $Y(d, 1) = Y(d, 0) \equiv Y(d)$ *(exclusion), almost surely.*

Assumption 2 (Conditional monotonicity)**.** $D_i(1) \geq D_i(0)$ *almost surely.*

Assumption 3 (Conditional relevance)**.** *The within-cell covariance of instrument and treatment* $\mathrm{Cov}(Z, D \mid X = x)$ *is nonzero on a set of* x *of positive probability, and its average* $S_D = \mathbb{E}[\mathrm{Cov}(Z, D \mid X)]$ *is nonzero, so that the saturated specification benchmark* $\bar{L} = S_Y / S_D$ *is well defined. In the binary specialization of Remark 1 this covariance is* $e(X)\{1 - e(X)\}p(X)$ *with* $p(x) = \mathbb{E}[D \mid Z = 1, X = x] - \mathbb{E}[D \mid Z = 0, X = x]$*, and instrument overlap* $0 < e(x) < 1$ *together with* $p(x) > 0$ *on a positive-probability set makes* S_D *strictly positive. Relevance confined to strata where* $\mathrm{Var}(Z \mid X) = 0$ *would leave* $S_D = 0$*, and the estimand, though still defined through* $\mathbb{E}[\tilde{Z} D] = \theta_D$ *whenever* $\theta_D \neq 0$*, would be pure contamination with no* $\bar{L}$ *to benchmark against.*

Assumption 4 (Maintained linear specification)**.** *The analyst estimates the just-identified linear IV model* $Y = \beta D + X'\gamma + \varepsilon$*, using* Z *as the single excluded instrument and* $(1, X')$ *as included instruments.*

Assumption 5 (Regularity)**.** Y*,* D*,* Z *and* X *have finite second moments, and* $\mathbb{E}[(1, X')'(1, X')]$ *is nonsingular. In addition,* $\mathbb{E}[\tilde{Z} D] \neq 0$*, where* $\tilde{Z} \equiv Z - L(Z \mid 1, X)$ *and* $L(\cdot \mid 1, X)$ *denotes the population linear projection on* $(1, X')$*.*

Under Assumption 4, the population 2SLS coefficient on D admits the familiar FWL / partialling-out form

$$\beta_{2SLS} = \frac{\mathbb{E}[\tilde{Z}\, Y]}{\mathbb{E}[\tilde{Z}\, D]}. \tag{2}$$

Define the instrument propensity $e(x) = \mathbb{P}(Z = 1 \mid X = x) = \mathbb{E}[Z \mid X = x]$, its linear projection $\ell(x) = L(Z \mid 1, X)(x)$, and the propensity curvature

$$h(x) \;\equiv\; e(x) - \ell(x), \qquad \mathbb{E}[h(X)] = 0, \quad \mathbb{E}[h(X)\, r(X)] = 0 \text{ for all affine } r. \tag{3}$$

Let $m_D(x) = \mathbb{E}[D \mid X = x]$, $m_Y(x) = \mathbb{E}[Y \mid X = x]$, and let $\mathrm{LATE}(x) = \mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0), X = x]$ denote the conditional LATE.

Lemma 1 (Contamination decomposition)**.** *Under Assumptions 4–5,*

$$\beta_{2SLS} = \frac{S_Y + \theta_Y}{S_D + \theta_D}, \tag{4}$$

where the signal *terms are the averaged within-cell covariances*

$$S_Y = \mathbb{E}\big[\operatorname{Cov}(Z, Y \mid X)\big], \qquad S_D = \mathbb{E}\big[\operatorname{Cov}(Z, D \mid X)\big],$$

and the contamination *terms are the covariances of the propensity curvature with the covariate level functions,*

$$\theta_Y = \operatorname{Cov}\big(h(X), m_Y(X)\big) = \mathbb{E}[h(X)Y], \quad \theta_D = \operatorname{Cov}\big(h(X), m_D(X)\big) = \mathbb{E}[h(X)D].$$

Consequently, writing $\bar{L} \equiv S_Y/S_D$ for the saturated specification estimand,

$$\beta_{2SLS} - \bar{L} = \frac{\theta_Y - \bar{L}\,\theta_D}{S_D + \theta_D}, \qquad \text{with } S_D + \theta_D = \mathbb{E}[\tilde{Z}D]. \tag{5}$$

The signal S_Y/S_D is the estimand of a fully interacted (saturated) specification. Writing $\tau(x) = \operatorname{Cov}(Z, Y \mid X = x)/\operatorname{Cov}(Z, D \mid X = x)$ for the within-cell IV slope, $\bar{L} = \mathbb{E}[\operatorname{Cov}(Z, D \mid X)\,\tau(X)]/\mathbb{E}[\operatorname{Cov}(Z, D \mid X)]$ is a weighted average of within-cell slopes with weights proportional to the within-cell covariance of instrument and treatment. We say the sign condition holds when $\operatorname{Cov}(Z, D \mid X = x)$ has a single sign across x. Under it these weights are non-negative, $\bar{L}$ is a genuine convex average, and it lies in the hull of the within-cell slopes, the reading we use in the Monte Carlo of Section 6 and in the applications of Section 7. Without it the weights can change sign and $\bar{L}$ need not lie in that hull. In the binary specialization of Remark 1 monotonicity (Assumption 2) delivers the sign condition, and for a general treatment it is an additional restriction we flag where it is used. The contamination terms are covariances between the propensity curvature h and the covariate level functions. By (3) they equal the covariance between the nonlinear part of the propensity and the nonlinear part of the level function, so each vanishes exactly under orthogonality, $\theta_D = 0 \iff h \perp m_D$ and $\theta_Y = 0 \iff h \perp m_Y$. A linear propensity ($h \equiv 0$) kills both terms and reproduces Blandhol et al. (2022), while a linear level function kills only its own. Orthogonality can also hold with neither function linear.

Remark 1 (Binary-treatment LATE specialization)**.** Suppose $Z, D \in \{0, 1\}$ and Assumptions 1–3 hold. Then $\operatorname{Cov}(Z, D \mid X) = e(X)\{1 - e(X)\}p(X)$ and, by the conditional Wald identity, $\operatorname{Cov}(Z, Y \mid X) = e(X)\{1-e(X)\}p(X)\mathrm{LATE}(X)$. Hence $S_D = \mathbb{E}[e(1-e)p]$, $S_Y = \mathbb{E}[e(1-e)p\,\mathrm{LATE}]$, the within-cell slope $\tau(x)$ is the conditional LATE, and $\bar{L} = S_Y/S_D$ is a convex average of conditional LATEs, the saturated specification estimand of Blandhol et al. (2022). Monotonicity is what single-signs $\operatorname{Cov}(Z, D \mid X)$ and so makes the average convex, and it plays no role in the decomposition itself.

The convex-average reading of $\bar{L}$ rests on the sign condition alone, and the primitive cases in which that condition holds are worth collecting, so that it reads as a bounded scope statement

rather than a caveat invoked in passing. The condition, that $\mathrm{Cov}(Z, D \mid X = x)$ has a single sign across x, holds in each of the following. (i) Binary Z and binary D under conditional monotonicity (Assumption 2), where $\mathrm{Cov}(Z, D \mid X) = e(X)\{1 - e(X)\}p(X)$ and overlap with $p(x) \geq 0$ fixes the sign, the case of Remark 1. (ii) An ordered treatment $D = \sum_j \mathbf{1}\{D \geq j\}$ under stagewise monotonicity, where each threshold crossing is monotone in Z and $\mathrm{Cov}(Z, D \mid X)$ is a sum of nonnegative stage covariances, the ordered design of Section 6. (iii) A continuous Z and D with a within-cell first stage of constant sign, $\partial_z \mathbb{E}[D \mid Z = z, X = x]$ of one sign in (z, x), which holds for a monotone structural first stage. Outside these cases the weights can change sign and $\bar{L}$ need not lie in the hull of within-cell slopes. The condition governs the interpretation of $\bar{L}$ only. The decomposition of Lemma 1, the strength–bias identity, the impossibility result, the directed test, and the correction hold without it.

The signal ratio $\bar{L}$ is the object a saturated specification targets, and we take it as the benchmark. Our subject is a feature of the decomposition that has gone unremarked, which we state as the paper's organizing result.

Proposition 1 (Strength–bias identity)**.** *The denominator of the 2SLS estimand is the population first-stage covariance $\mathbb{E}[\tilde{Z}D] = S_D + \theta_D$ from which the Sanderson–Windmeijer conditional F is constructed. The contamination θ_D is the common term. It enters the denominator of the bias $\beta_{2SLS} - \bar{L} = (\theta_Y - \bar{L}\theta_D)/(S_D + \theta_D)$ and, being a signed covariance, inflates the reported first-stage covariance when $\theta_D > 0$ and deflates it when $\theta_D < 0$. The displacement of β_{2SLS} from $\bar{L}$ requires the numerator $\theta_Y - \bar{L}\theta_D \neq 0$, so strength and interpretability need not move together. The point of the identity is the weaker but consequential one that they* can *be driven by the same nuisance θ_D, and under collinearity (Section 3) they are.*

Proof. This is immediate from Lemma 1. The denominator of (4) is $S_D + \theta_D$, and $\mathbb{E}[\tilde{Z}D] = \mathbb{E}[(h(X) + (Z - e(X)))D] = \theta_D + S_D$ since $\mathbb{E}[(Z - e(X))D \mid X] = \mathrm{Cov}(Z, D \mid X)$ integrates to S_D. □

The identity organizes the two failure modes. Through the denominator the contamination inflates the reported strength, dominant under collinearity (mode a, Section 3). Through the numerator, via θ_Y, it biases the estimand even when the reported strength is honest, a bias that loads entirely on the outcome (mode b, the health insurance design of Section 7.1).

The leading binary case in one pass

Before developing the results for a treatment of arbitrary support, we run the whole argument once in the binary-treatment LATE model of Remark 1, where every object has a causal reading. It is the case most applied designs occupy, and it fixes the intuition that the general theory then extends.

Here the benchmark $\bar{L} = S_Y/S_D$ is a genuine convex average of conditional LATEs, so a saturated specification recovers a causal parameter and the only question is what the 2SLS estimate does to it. The bias is $\beta_{2SLS} - \bar{L} = (\theta_Y - \bar{L}\theta_D)/(S_D + \theta_D)$, and both contamination terms are covariances of the propensity curvature h with a covariate level function. The strength–bias identity (Proposition 1) is the observation that the denominator $S_D + \theta_D$ is exactly the first-stage covariance

the conditional F is built from, so the nuisance θ_D that shifts the estimand is the same one that moves the reported strength.

That single fact produces the two modes. In mode a the instrument is nearly a nonlinear function of the covariates, the signal S_D vanishes, and θ_D manufactures the conditional F out of the very curvature that biases the estimand. In mode b the strength is honest, yet θ_Y displaces the estimand through the outcome, where no first-stage number can reach, so here the estimand can leave the hull of conditional LATEs while the F stays large (Section 6.3). The impossibility result, the directed test, and the correction previewed above all hold in this binary model with a causal reading of $\bar{L}$, and the remainder of the paper establishes them for a scalar treatment of arbitrary support, with the binary case as the special instance in which $\bar{L}$ is a convex average of LATEs.

3. Collinearity and the failure of the first-stage diagnostic

We now formalize what happens as the instrument becomes collinear with the covariates. Intuitively, "Z is collinear with X" means Z is close to a (possibly nonlinear) function of X, i.e. the conditional variance $\mathrm{Var}(Z \mid X)$ is small. In the binary case this variance equals $e(X)\{1 - e(X)\}$. We index a sequence of data-generating processes (DGPs) by t.

Assumption 6 (Collinearity sequence)**.** *Along $t = 1, 2, \ldots$ the following hold. (i) $\mathrm{Var}_t(Z \mid X) \to 0$ for almost every x and $\mathbb{E}[\mathrm{Var}_t(Z \mid X)] \to 0$. (ii) The propensity curvature converges, $h_t \to h^*$ in L^2, with h^* non-degenerate. (iii) The within-cell slope and the level functions converge in L^2 to limits τ^*, m_D^*, m_Y^*, with $\sup_t \mathbb{E}[\mathrm{Var}_t(D \mid X)] < \infty$ and $\{\mathrm{Var}_t(Y \mid X)\}$ uniformly integrable. (iv) $\mathrm{Cov}(h^*, m_D^*) \neq 0$.*

Part (i) is the collinearity limit. Part (ii) says the propensity retains curvature as Z is pushed toward a deterministic function of X, and a nonlinear index is the leading example, of which Section 6 uses several. Part (iv) is the genericity condition that makes the limit interesting. It holds whenever the nonlinear part of $m_D = \mathbb{E}[D \mid X]$ is not orthogonal to h^*, and it rules out a knife-edge cancellation of the propensity curvature by curvature in the baseline conditional mean rather than following from relevance alone.

Proposition 2 (Collinearity contamination limit)**.** *Under Assumptions 4–5 and 6, the signal terms vanish, $S_{D,t} \to 0$ and $S_{Y,t} \to 0$, while the contamination terms converge to $\mathrm{Cov}(h^*, m_D^*) \neq 0$ and $\mathrm{Cov}(h^*, m_Y^*)$. Hence*

$$\beta_{2SLS,t} \longrightarrow \frac{\mathrm{Cov}(h^*, m_Y^*)}{\mathrm{Cov}(h^*, m_D^*)}. \tag{6}$$

The limit is a ratio of covariate-level contrasts and carries no causal content. In the binary LATE specialization of Remark 1 it may lie outside $[\operatorname{ess\,inf}_x \mathrm{LATE}^*(x),\ \operatorname{ess\,sup}_x \mathrm{LATE}^*(x)]$ and take the opposite sign of every $\mathrm{LATE}^*(x)$.

The reduced-form intuition is transparent. As $\mathrm{Var}(Z \mid X) \to 0$, the within-cell experimental variation in the instrument disappears. All that survives in $\tilde{Z} = Z - \ell(X)$ is the projection error $h(X)$, which is a deterministic function of X. The estimator then compares Y and D across covariate cells using weights $h(X)$, exactly the comparison an IV design is meant to avoid. The next result is the payoff for practice.

Corollary 1 (The conditional F does not warn)**.** *The population Sanderson–Windmeijer conditional first-stage slope is* $\pi_t = \mathbb{E}[\tilde{Z}_t D]/\operatorname{Var}(\tilde{Z}_t) = (S_{D,t} + \theta_{D,t})/\operatorname{Var}(\tilde{Z}_t)$*, and under Assumption 6,*

$$\pi_t \;\longrightarrow\; \frac{\operatorname{Cov}(h^*, m_D^*)}{\mathbb{E}[(h^*)^2]} \neq 0.$$

Consequently the conditional first-stage F statistic, whose population counterpart is proportional to $n\,\pi_t^2 \operatorname{Var}(\tilde{Z}_t)/\sigma_v^2$, is bounded away from zero and diverges with the sample size along the sequence, even though by Proposition 2 the estimand is pure contamination. A large conditional F therefore places no bound on the contamination bias (5)*.*

Corollary 1 is the formal statement of the applied warning. The conditional F of Sanderson and Windmeijer (2016) correctly certifies that $\tilde{Z}$ is a strong predictor of D. Under collinearity, however, $\tilde{Z}$ predicts D through the contamination channel $h(X)$, not through the within-cell signal $\operatorname{Cov}(Z, D \mid X)$. Strength and interpretability are distinct, and can even be driven by the same nuisance.

Two qualifications keep the warning honest and locate where it bites. First, the collinearity that drives mode (a) is itself visible without the outcome. As $\operatorname{Var}(Z \mid X) \to 0$ the instrument is nearly a function of the covariates, so a regression of Z on a flexible basis in X returns a fit approaching one, and the within-cell signal announces itself as weak even while the conditional F is large. A diligent analyst can therefore see manufactured strength coming in mode (a), from the fit of the instrument on the covariates. The sharp content of Corollary 1 is that the F points the wrong way exactly when that fit is what should be checked. Second, and by consequence, the impossibility bites hardest in mode (b), where the signal is honest, the fit of the instrument on the covariates is unremarkable, and only the outcome reveals the bias. The bias-targeting test of Section 4 inherits this division. Its power against a fixed alternative is full when the signal is bounded away from zero, the mode (b) regime, and declines as the signal vanishes, the mode (a) regime. This is the Anderson–Rubin trade-off of Proposition 8. The two regimes do not leave a gap, because the regime in which the test loses power is the one the fit of the instrument on the covariates already flags.

The following result shows the failure is not special to the conditional F but is shared by every strength diagnostic, and explains why. Such diagnostics are computed from the first stage, whereas the bias lives in the outcome.

Proposition 3 (Strength diagnostics cannot detect contamination)**.** *Let T be any diagnostic that is a measurable functional of the joint distribution of (Z, D, X), in particular any first-stage relevance or strength statistic, including the conditional F, the effective F of Montiel Olea and Pflueger (2013), and the Cragg–Donald statistic. Fix any such law with nonzero curvature, $h \neq 0$. Then T is uninformative about the contamination bias. There exist data-generating processes with that common joint distribution of (Z, D, X), hence identical values of T, but arbitrarily different biases $\beta_{2SLS} - \bar{L}$. Detecting the bias requires the conditional distribution of Y. The restriction to $h \neq 0$ is not a weakening. When $h \equiv 0$ the first stage already certifies $\theta_D = \theta_Y = 0$ and hence no contamination, by Lemma 1, so the interesting case is exactly the one T cannot resolve.*

Proposition 3 is the deep reason the applied bridge fails. No refinement of the first-stage report, however sophisticated, can be made to carry information about the contamination. By (5) the bias

depends on $\theta_Y = \mathbb{E}[h(X)Y]$ and on $\bar{L} = S_Y/S_D$, both functionals of the outcome that are left completely free once the law of (Z, D, X) is fixed. This is an instance of the level dependence of Blandhol et al. (2022). Concretely, the free direction is the outcome level of the units the instrument does not move. Shifting the conditional outcome mean m_Y in the curvature direction h changes θ_Y, and hence the bias, while leaving the law of (Z, D, X), the within-cell slopes, and every strength diagnostic unchanged. A strength statistic summarizes how the instrument moves the treatment, whereas the bias lives with the level of the units it does not move, so a valid diagnostic must use Y.

One thing the first stage can certify is the absence of curvature. By Lemma 1, $h \equiv 0$, a propensity linear in X and testable from (Z, X) alone by a RESET, forces $\theta_D = \theta_Y = 0$ and hence no contamination. The impossibility is therefore conditional on nonzero curvature. Once $h \neq 0$, no first-stage functional can say whether that curvature is harmful, because harm runs through θ_Y and $\bar{L}$, which only the outcome pins down. This is the precise content of the blanket statements elsewhere that a strong first stage cannot license a causal reading. This cuts in the opposite direction to the familiar impossibility that instrument validity cannot be confirmed even with infinite data (Słoczyński et al., 2026). There the outcome is of no help, while here it is exactly what a diagnostic requires. The observation both motivates and disciplines the test of the next section, whose central moment $\theta_Y = \mathbb{E}[h(X)Y]$ is precisely such an outcome functional. It also implies, a fortiori, that a Ramsey RESET applied to the first stage, a functional of (Z, X) alone, cannot distinguish a biased design from an unbiased one. We confirm this in Section 6.

4. A directed test for contamination

A correction is already available. Saturating the covariates, or equivalently residualizing the instrument on a flexibly estimated propensity, removes the contamination and returns the estimand to the signal ratio $\bar{L}$ (Section 5). But the correction does not remove the need to test, because it is not free. It surrenders exactly the identifying variation the contamination supplied, so in the collinearity regime where the bias is largest it leaves the design weakly identified or unidentified (part (ii) of Proposition 8). An analyst therefore needs to know whether the curvature actually reaches the estimate before paying that cost, and by Proposition 3 no first-stage number answers this. This section builds a test that does, in two forms, a cheap screen for curvature that loads on D and Y and a bias-targeting refinement that separates the curvature which biases the estimand from the curvature which does not.

4.1. The null and its interpretation

By Lemma 1, $\beta_{2SLS} = \bar{L}$ whenever $\theta_D = \theta_Y = 0$. We therefore test

$$H_0 : \ \theta_D = 0 \ \text{ and } \ \theta_Y = 0 \qquad \text{against} \qquad H_1 : \ (\theta_D, \theta_Y) \neq (0, 0). \tag{7}$$

Under H_0 the 2SLS estimand retains its interpretation as a convex average of LATEs. Because $\theta_D = \mathbb{E}[h(X)D]$ and $\theta_Y = \mathbb{E}[h(X)Y]$ depend only on the propensity curvature and on observables, H_0 is a statement about the reduced form. The consequential misspecification is nonlinearity of the instrument propensity $\mathbb{E}[Z \mid X]$ in the directions that also load on D and Y. This is what

distinguishes the test from a Ramsey RESET applied to the first stage $D \sim (Z, X)$. RESET rejects for propensity curvature in any direction, including curvature orthogonal to the outcome that leaves (5) unaffected. This screening statistic instead probes only curvature that loads on D and Y, and does so without touching an endogenous regressor. It is still a sufficient-condition test. It can reject when the loadings satisfy $\theta_Y = \bar{L}\theta_D$ and the estimand is in fact unbiased (harmless curvature), which is why Section 4.3 adds the necessary-and-sufficient bias-targeting variant that probes the single direction $Y - \bar{L}D$.

4.2. Estimator and limiting distribution

Let $b(X) = (b_1(X), \ldots, b_K(X))'$ be a vector of basis functions (for example polynomials or splines) orthogonalized against $(1, X')$ in the population, so that $\mathbb{E}[b(X)(1, X')] = 0$. These span the curvature directions to be probed. Model the propensity as $e(X) = \alpha_0 + X'\alpha_1 + b(X)'\delta + \rho(X)$ with $\mathbb{E}[\rho(X)b(X)] = 0$, so that the component of the curvature captured by the basis is $h_K(X) = b(X)'\delta$. Given an i.i.d. sample $\{(Y_i, D_i, Z_i, X_i)\}_{i=1}^n$, the test is computed in three steps.

(i) Regress Z_i on $(1, X_i', b(X_i)')$ by OLS and form $\hat{h}_i = b(X_i)'\hat{\delta}$.

(ii) Form the sample contamination moments $\hat{\theta}_D = n^{-1}\sum_i \hat{h}_i D_i$ and $\hat{\theta}_Y = n^{-1}\sum_i \hat{h}_i Y_i$, stacked as $\hat{\theta} = (\hat{\theta}_D, \hat{\theta}_Y)'$.

(iii) Compute the Wald statistic $W_n = n\,\hat{\theta}'\hat{\Sigma}^{-1}\hat{\theta}$, where $\hat{\Sigma}$ is a consistent estimator of the asymptotic variance Σ of $\sqrt{n}\,\hat{\theta}$.

Assumption 7 (Test regularity). *(i) $\{(Y_i, D_i, Z_i, X_i)\}_{i=1}^n$ are i.i.d. with $\mathbb{E}[Y^4] < \infty$, $\mathbb{E}[D^4] < \infty$ and $\mathbb{E}\|b(X)\|^4 < \infty$. (ii) $G = \mathbb{E}[b(X)b(X)']$ is nonsingular and the basis dimension K is fixed. (iii) $(\alpha_0, \alpha_1, \delta)$ are the unique population least-squares coefficients in $e(X) = \alpha_0 + X'\alpha_1 + b(X)'\delta + \rho(X)$ with $\mathbb{E}[\rho(X)(1, X', b(X)')'] = 0$. (iv) $\Sigma = \mathrm{Var}(\psi_i)$ is nonsingular.*

Proposition 4 (Limiting distribution and weak-identification robustness). *(i) Under Assumption 7, $\sqrt{n}\,(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma)$, where $\Sigma = \mathrm{Var}(\psi_i)$ is the asymptotic variance of the two-step moment estimator. The influence function ψ_i is the stacked contamination moment $(b(X_i)'\delta\, W_i - \theta_W)_{W \in \{D,Y\}}$ plus a generated-regressor correction for the first-step estimation of δ, given explicitly in the Appendix. Under H_0, $W_n \xrightarrow{d} \chi^2_2$. Against a fixed alternative with $(\theta_D, \theta_Y) \neq 0$, $W_n \to \infty$, so the test is consistent, and against local alternatives $\theta = \eta/\sqrt{n}$ it has nontrivial power with noncentrality $\eta'\Sigma^{-1}\eta$.*

(ii) Call $\{P_n\}$ a weak-identification null sequence *if the fourth moments of $(Y, D, \|b(X)\|)$ are bounded uniformly in n with $\Sigma_n \to \Sigma_* \succ 0$, so that Assumption 7 holds uniformly, the first-stage strength drifts to zero, $S_{D,n} \to 0$ (as when $\mathrm{Var}(Z \mid X) \to 0$ or $p(X) \to 0$), the curvature stays nondegenerate, $h_n \to h^* \neq 0$, and the null is maintained, $\theta_{D,n} = \theta_{Y,n} = 0$ for all n. Along any such sequence $W_n \xrightarrow{d} \chi^2_2$ under H_0, and likewise the orthogonalized statistic W_n^o of Proposition 5, so the test has asymptotic size α regardless of the degree of identification. By contrast the contaminated collinearity sequence of Assumption 6, with $\mathrm{Cov}(h^*, m_D^*) \neq 0$ and hence $\theta_{D,n} \not\to 0$, is an* alternative *rather than a null, along which the noncentrality $n\,\theta_n'\Sigma_n^{-1}\theta_n$ diverges and the power tends to one even as every first-stage strength diagnostic diverges with the sequence (Corollary 1).*

The estimator $\hat{\theta}$ is a two-step GMM estimator, a first-step OLS for $(\alpha_0, \alpha_1, \delta)$ followed by the second-step sample moments (ii). Under Assumption 7, Newey and McFadden (1994, Theorem 6.1) give asymptotic linearity with the influence function above and $\Sigma = \mathrm{Var}(\psi_i)$. The second term of ψ_i is the generated-regressor correction for first-step estimation of δ, and it is nonzero precisely because $b(X)$ enters through $\hat{\delta}$. The plug-in $\hat{\Sigma} = n^{-1}\sum_i \hat{\psi}_i\hat{\psi}_i'$, formed from the sample analogues $\hat{\psi}_i$, is consistent by the law of large numbers and the continuous mapping theorem, so $W_n \xrightarrow{d} \chi^2_2$ under H_0. Because $\hat{\theta}$ is a Hadamard-differentiable functional of the empirical measure, the nonparametric pairs bootstrap is first-order valid by the functional delta method (van der Vaart, 1998, Theorem 23.9). Resampling (Y_i, D_i, Z_i, X_i) jointly reproduces the generated-regressor term automatically, and we use it throughout Section 6.

The null (7) is a statement about the reduced form, and the estimator $\hat{\theta}$ never divides by the first-stage covariance $\mathbb{E}[\tilde{Z}D]$. The test is therefore immune to the weak-instrument problem, which is a small-denominator problem for β_{2SLS}. This is what separates the diagnostic from any procedure built on an IV estimate, and because the collinearity limit is precisely where it matters, part (ii) of Proposition 4 states it formally.

The mechanism is that $\theta = (\theta_D, \theta_Y)$ and its influence function ψ are functionals of the reduced form (Z, D, Y, X), estimated at rate $\sqrt{n}$ by the regression of Z on $(1, X', b(X)')$, whose behaviour is governed by $G = \mathbb{E}[bb']$ rather than by the relevance of the instrument. Because the curvature stays nondegenerate, $h_n \to h^* \neq 0$, the leading influence term $b(X)'\delta\, W - \theta_W$ stays nondegenerate and Σ_n remains nonsingular even as the generated-regressor correction shrinks with the residual variance of Z. The screening test thus retains its size in exactly the regime, $S_D \to 0$, in which every first-stage strength diagnostic diverges (Corollary 1) and a saturate and compare Hausman contrast collapses. The point is that weak identification and contamination are separate. The null sequence has $S_D \to 0$ without contamination, whereas Assumption 6 couples $S_D \to 0$ with $\theta_D \not\to 0$ and is precisely the design the test is meant to reject. The bias-targeting refinement of Section 4.3 does require S_D bounded away from zero, for the separate reason that it plugs in $\hat{\tilde{L}}$. That boundary is discussed there.

Proposition 4 fixes the basis dimension K, so the curvature is a parametric first step. Applied users will often prefer to estimate the propensity flexibly or by machine learning, which breaks Assumption 7(ii). The first-step error is then no longer $\sqrt{n}$-negligible, and a naive plug-in is biased at rate faster than $\sqrt{n}^{-1}$. The following result restores valid inference by replacing the plug-in moment with a Neyman-orthogonal one, so that only $n^{1/4}$-consistent, cross-fitted nuisances are required.

Proposition 5 (Orthogonalized test with a data-driven propensity)**.** *Let the nuisances be the propensity $e(\cdot)$ and the level functions $m_W(\cdot) = \mathbb{E}[W \mid X]$, $W \in \{D, Y\}$, and write $\tilde{m}_W(X) = m_W(X) - L(m_W \mid 1, X)$ for the level function detrended of its affine part. For $W \in \{D, Y\}$ define the orthogonalized moment*

$$\psi^o_W(O; e, m_W) = h(X)\, W + \tilde{m}_W(X)\,\big(Z - e(X)\big) - \theta_W, \qquad h = e - L(e \mid 1, X), \tag{8}$$

which has mean zero at the truth and is Neyman-orthogonal. Its pathwise derivative with respect to e

and with respect to m_W vanishes. Estimate (e, m_D, m_Y) by any method with K-fold cross-fitting, and form $\hat{\theta}^o = (\hat{\theta}^o_D, \hat{\theta}^o_Y)'$, $\hat{\theta}^o_W = n^{-1}\sum_i\{\hat{h}(X_i)W_i + \hat{\tilde{m}}_W(X_i)(Z_i - \hat{e}(X_i))\}$, and $W^o_n = n\,\hat{\theta}^{o\prime}\hat{\Sigma}_o^{-1}\hat{\theta}^o$. Suppose the cross-fitted nuisances satisfy $\|\hat{e} - e\|_{L^2} = o_p(n^{-1/4})$ and $\|\hat{m}_W - m_W\|_{L^2} = o_p(n^{-1/4})$, hence the product rate $\|\hat{e} - e\|_{L^2}\|\hat{m}_W - m_W\|_{L^2} = o_p(n^{-1/2})$, with the fourth moments of Assumption 7(i) and $\Sigma_o = \mathrm{Var}(\psi^o_i)$ nonsingular. Then $\sqrt{n}(\hat{\theta}^o - \theta) \xrightarrow{d} N(0, \Sigma_o)$ and, under H_0, $W^o_n \xrightarrow{d} \chi^2_2$.

The fixed-basis test of Proposition 4 and this orthogonalized test are two ends of a spectrum, with an undersmoothed fixed-form series ($K = o(n^{1/2})$ with the two-step correction of Newey and McFadden (1994)) as the intermediate route. It is orthogonalization (8) that removes the $n^{1/4}$ bottleneck and licenses off-the-shelf machine learning for the propensity. We use a fixed cubic basis throughout and treat K as a robustness dimension.

A word on what a non-rejection certifies. The fixed-basis screen of Proposition 4 has power only against contamination that projects onto the span of $b(X)$. Its noncentrality loads on $\mathbb{E}[b(X)D]$ and $\mathbb{E}[b(X)Y]$, so any curvature h orthogonal to b leaves the statistic unaffected. A non-rejection therefore certifies only that there is no detectable contamination in the probed directions, not that $\theta_D = \theta_Y = 0$ in every direction. This is the price of directing the test. The growing-basis and orthogonalized versions (Proposition 5) restore consistency against all directions as $K \to \infty$. With the fixed cubic basis used below, the empirical "clears" of Section 7 should accordingly be read as "no contamination detectable at cubic order," and K as a robustness dimension one can raise.

A natural alternative to our test is to saturate and compare, estimating β under a saturated covariate specification and comparing it to the linear-specification estimate via a Hausman contrast. Under collinearity this is exactly the wrong tool. Saturating drives the effective instrument toward the vanishing signal term $S_D \to 0$ (Proposition 2), so the saturated estimator is weakly identified and its sampling variability swamps the contrast. Our test sidesteps this because it never forms an IV ratio, and its size is controlled along the collinearity sequence (part (ii) of Proposition 4), the property the saturate and compare contrast cannot have.

4.3. A bias-targeting variant

The null (7) is sufficient for $\beta_{2SLS} = \bar{L}$ but not necessary. By (5) the bias vanishes whenever $\theta_Y = \bar{L}\,\theta_D$, which can hold with $(\theta_D, \theta_Y) \neq 0$. The leading example is harmless curvature. Here the propensity is nonlinear in X, so $h \neq 0$ and, since m_D inherits the curvature, $\theta_D \neq 0$, yet the outcome level is linear in X and the effect is homogeneous, so $\theta_Y = \bar{L}\theta_D$ and the estimand is unbiased. Testing (7) then rejects a design that is in fact fine. The screening test of Section 4 is therefore best read as a strongly identified, conservative first pass.

The exact no-bias restriction is a single moment. Since $\theta_Y - \bar{L}\theta_D = \mathrm{Cov}\big(h(X),\, m_Y(X) - \bar{L}\,m_D(X)\big) = \mathbb{E}\big[h(X)\{Y - \bar{L}D\}\big]$,

$$H_0^* :\ \mathbb{E}\big[h(X)\{Y - \bar{L}\,D\}\big] = 0, \tag{9}$$

which is necessary and sufficient for $\beta_{2SLS} = \bar{L}$ whenever $S_D \neq 0$. Written as a ratio, H_0^* invites a plug-in of $\bar{L} = S_Y/S_D$, fragile precisely where contamination is worst. Clearing the denominator gives an equivalent restriction in objects that are all $\sqrt{n}$-estimable irrespective of the signal. With

$\psi \equiv \theta_Y S_D - \theta_D S_Y$,

$$b = \beta_{2SLS} - \bar{L} = \frac{\psi}{S_D\,\mathbb{E}[\tilde{Z}D]}, \qquad H_0^* \iff \psi = 0. \tag{10}$$

The two forms test the same null but differ in where the small signal S_D appears, and (10) forces a genuine trade-off. We build both, a plug-in that tracks b directly and a signal-cleared form that tracks ψ.

For the plug-in we estimate H_0^* with the bias-corrected $\hat{\bar{L}}$ of (15) (Section 5) and the flexible curvature $\hat{h}_i$. Writing $\hat{g} = n^{-1}\sum_i \hat{h}_i\{Y_i - \hat{\bar{L}}D_i\} = \hat{\theta}_Y - \hat{\bar{L}}\hat{\theta}_D = \hat{\psi}/\hat{S}_D$, the statistic is $T_n^* = \hat{g}^2/\hat{v}$ with $\hat{v}$ the pairs-bootstrap variance of $\hat{g}$, obtained by recomputing $\hat{h}$ and $\hat{\bar{L}}$ on each resample. Suppose H_0^* and Assumption 7 hold and S_D is bounded away from zero, so that $\hat{\bar{L}}$ is $\sqrt{n}$-consistent for $\bar{L}$ (Proposition 8). Then $\sqrt{n}\,\hat{g} \xrightarrow{d} N(0, v)$ for a finite $v > 0$, the pairs-bootstrap variance satisfies $n\hat{v} \xrightarrow{p} v$, and $T_n^* = n\hat{g}^2/(n\hat{v}) \xrightarrow{d} \chi^2_1$, consistent against any fixed alternative with nonzero bias. The representation $\hat{g} = \hat{\psi}/\hat{S}_D$ shows the weakness. As $S_D \to 0$ the plug-in $\hat{\bar{L}}$ is no longer $\sqrt{n}$-consistent and T_n^* divides by a vanishing $\hat{S}_D$, so it is no longer asymptotically pivotal (Proposition 8). This is the size side of the trade-off in (10).

The ψ-form removes exactly that boundary problem by testing $\hat{\psi}$ itself, which never divides by the signal. Estimate the four reduced-form means from the flexible first-step regression of Z on $(1, X', b(X)')$, with $\hat{\theta}_W = n^{-1}\sum_i \hat{h}_i W_i$ and $\hat{S}_W = n^{-1}\sum_i (Z_i - \hat{e}(X_i))W_i$ for $W \in \{D, Y\}$, and form

$$\hat{\psi} = \hat{\theta}_Y \hat{S}_D - \hat{\theta}_D \hat{S}_Y, \qquad A_n = \hat{\psi}^2/\hat{v}_\psi, \tag{11}$$

where $\hat{v}_\psi$ is a consistent estimator of $\mathrm{Var}(\hat{\psi})$, either the delta-method plug-in $n^{-1}\widehat{\nabla\psi}'\hat{\Omega}\,\widehat{\nabla\psi}$ with $\nabla\psi = (-S_Y, S_D, \theta_Y, -\theta_D)'$ and $\hat{\Omega}$ the joint variance of $\sqrt{n}\,(\hat{\theta}_D, \hat{\theta}_Y, \hat{S}_D, \hat{S}_Y)$, or the pairs bootstrap. Because $\hat{\psi}$ is a smooth function of averages that are all $\sqrt{n}$-estimable without dividing by the first stage, A_n stays pivotal at the collinearity boundary.

Proposition 6 (Weak-identification-robust bias targeting)**.** *Let $\{P_n\}$ satisfy Assumption 7 uniformly, with the curvature nondegenerate, $h_n \to h^* \neq 0$ and hence $(\theta_D^*, \theta_Y^*) \neq 0$, and let $\hat{v}_\psi$ be ratio-consistent for $\mathrm{Var}(\hat{\psi})$ along the sequence. Then under H_0^*, $A_n \xrightarrow{d} \chi^2_1$ whether or not $S_{D,n}$ is bounded away from zero, so the test that rejects when $A_n > \chi^2_{1,1-\alpha}$ has asymptotic size α uniformly over the collinearity range. Against a fixed alternative with $b = \beta_{2SLS} - \bar{L} \neq 0$ and S_D bounded away from zero the power tends to one. As $S_D \to 0$ with b fixed, $\psi = b\,S_D\,\mathbb{E}[\tilde{Z}D] \to 0$ and the power against that fixed b declines toward α, the unavoidable weak-identification trade-off familiar from Anderson–Rubin inference.*

The two statistics are complementary by construction, not by hedge. A_n delivers size uniformly, including the collinearity boundary where T_n^* divides by a vanishing signal, and T_n^* buys higher power where the design is strongly identified. This sits alongside the earlier axis of the "report both" protocol, the screening test of Section 4, whose size is likewise controlled under weak identification (Proposition 4) and which serves as a cheap sufficient first pass, against the bias-targeting test that adjudicates harmless curvature. The protocol is to report the screen and the bias-targeting test alongside the contamination share $\hat{\varrho}$, which by Proposition 8(iii) says how much identification a correction would cost, using A_n for a size guarantee that holds across the collinearity range and

T_n^* where strong identification makes its extra power worth having. In the finite-sample designs of Section 6.6 the two forms agree when strongly identified and A_n holds size as $\mathrm{Var}(Z \mid X) \to 0$ (Table B.10), confirming Proposition 6. Section 6.6 compares both against a RESET.

4.4. The over-identified case

The bias-targeting restriction has an exact over-identified analogue, at the cost of the single reduced form. Let $Z = (Z_1, \ldots, Z_J)'$ be a vector of instruments with residualized components $\tilde{Z}_j = Z_j - L(Z_j \mid 1, X)$, and collect the second moments $Q_{ZD} = \mathbb{E}[\tilde{Z}D]$, $Q_{ZY} = \mathbb{E}[\tilde{Z}Y]$, and $Q_{ZZ} = \mathbb{E}[\tilde{Z}\tilde{Z}']$. Each inherits the decomposition of Lemma 1, $Q_{ZD} = S_D + \theta_D$ and $Q_{ZY} = S_Y + \theta_Y$ with per-instrument signal and contamination vectors $S_{W,j} = \mathbb{E}[\mathrm{Cov}(Z_j, W \mid X)]$ and $\theta_{W,j} = \mathbb{E}[h_j(X)W]$, while $Q_{ZZ} = \Sigma_S + H$ with $\Sigma_S = \mathbb{E}[\mathrm{Cov}(Z \mid X)]$ and $H = \mathbb{E}[h(X)h(X)']$ the curvature Gram matrix. The population 2SLS coefficient on D is the ratio of quadratic forms

$$\beta_{2SLS} = \frac{Q_{ZD}'\, Q_{ZZ}^{-1}\, Q_{ZY}}{Q_{ZD}'\, Q_{ZZ}^{-1}\, Q_{ZD}}, \tag{12}$$

which is just-identified IV with the combined instrument $\hat{D} = \pi'\tilde{Z}$, $\pi = Q_{ZZ}^{-1}Q_{ZD}$. Let $\bar{L}$ be the benchmark of Section 5, the 2SLS estimand built from the clean instruments $Z_j - e_j(X)$, which specializes (12) with (S_D, S_Y, Σ_S) in place of (Q_{ZD}, Q_{ZY}, Q_{ZZ}) and reduces to S_Y/S_D when $J = 1$.

Proposition 7 (Over-identified screen and exact restriction)**.** *(i) If $\theta_D = \theta_Y = 0$ then $\mathbb{E}[\hat{D}\,W] = \pi' S_W$ for $W \in \{D, Y\}$, so the combined instrument's curvature part $\pi' h(X)$ is orthogonal to D and Y and β_{2SLS} is a contamination-free instrument combination whatever the curvature H carried by the weighting matrix. (ii) With $\mathbb{E}[\hat{D}D] \neq 0$,*

$$\beta_{2SLS} = \bar{L} \quad \Longleftrightarrow \quad H_0^{\dagger}:\ \mathbb{E}\big[\hat{D}\,(Y - \bar{L}D)\big] = 0. \tag{13}$$

(iii) At $J = 1$, $\hat{D} \propto \tilde{Z}$ and $S_Y - \bar{L}S_D = 0$, so (13) becomes $\mathbb{E}[h(X)\{Y - \bar{L}D\}] = 0$ of (9).

Proof. For (i), $\mathbb{E}[\hat{D}W] = \pi' Q_{ZW} = \pi'(S_W + \theta_W) = \pi' S_W$ when $\theta_W = 0$, and since $\hat{D} = \pi'(Z - e(X)) + \pi' h(X)$ the second term contributes $\pi'\theta_W = 0$ to each moment. For (ii), $\beta_{2SLS} = \mathbb{E}[\hat{D}Y]/\mathbb{E}[\hat{D}D]$ by (12), so $\beta_{2SLS} - \bar{L} = \mathbb{E}[\hat{D}(Y - \bar{L}D)]/\mathbb{E}[\hat{D}D]$. For (iii), at $J = 1$ the scalar π cancels and $\bar{L} = S_Y/S_D$ gives $S_Y - \bar{L}S_D = 0$, leaving $\theta_Y - \bar{L}\theta_D = \mathbb{E}[h(X)\{Y - \bar{L}D\}]$. □

The screen $\theta_D = \theta_Y = 0$ therefore stays a sufficient test for the absence of contamination, now for a transparent reason. Any combination of instruments whose per-component curvature is orthogonal to D and Y is itself a clean instrument, so the weighting π may be contaminated by H without biasing the estimand, and the finite-sample behaviour of the stacked $\chi^2_{2,J}$ screen is the design (C) evidence of Section 6.4. What the screen cannot do, exactly as in the just-identified case, is separate harmful from harmless curvature, and (13) is the restriction that does. Written out it is $\pi'(Q_{ZY} - \bar{L}Q_{ZD}) = 0$, and unlike the just-identified null $\mathbb{E}[h(X)\{Y - \bar{L}D\}] = 0$ it does not reduce to a statement about the curvature alone. The combination $\pi = Q_{ZZ}^{-1}Q_{ZD}$ carries the contamination through both the moment vector Q_{ZD} and the inverse second-moment matrix Q_{ZZ}^{-1}, so the exact analogue must invert Q_{ZZ} and couples the signal-side over-identification weighting

with the contamination in a way the single reduced form does not. This is the sense in which over-identification is the less mechanical direction.

The sample analogue plugs the residualized second moments and the corrected $\hat{\bar{L}}$ of (15) into (13), forming $\hat{g} = n^{-1}\sum_i \hat{D}_i(Y_i - \hat{\bar{L}}D_i)$ and $T_n^\dagger = \hat{g}^2/\hat{v}$ with $\hat{v}$ the pairs-bootstrap variance. As a smooth function of sample second moments $\hat{g}$ is asymptotically normal when the combined first stage $\mathbb{E}[\hat{D}D]$ is bounded away from zero, so $T_n^\dagger \xrightarrow{d} \chi^2_1$ under $H_0^\dagger$, and clearing $\mathbb{E}[\hat{D}D]$ from the denominator gives a signal-cleared form valid at the collinearity boundary exactly as A_n does for the just-identified ψ in (10).

4.5. Reporting the contamination share

Beyond a hypothesis test, the decomposition yields a single interpretable number. Define the contamination share

$$\varrho \;\equiv\; \frac{\theta_D}{S_D + \theta_D} \;=\; \frac{\theta_D}{\mathbb{E}[\tilde{Z}D]}, \tag{14}$$

the fraction of the observed first-stage covariance $\mathbb{E}[\tilde{Z}D]$ attributable to propensity curvature rather than to the within-cell signal. It is a share of the covariance, not of the conditional F or the concentration parameter, which are quadratic in it. In the nonlinear-propensity family of Section 6, (14) reduces to $\varrho = \mathbb{E}[h^2]/\operatorname{Var}(\tilde{Z}) \in [0,1]$. In general ϱ is a signed decomposition of the first-stage covariance. Since $S_D > 0$ always but θ_D is a signed covariance, $\varrho \in [0,1]$ and curvature inflates the F only under $\theta_D > 0$, the sign the mode-(a) designs here satisfy. When $\theta_D < 0$ the curvature instead deflates the F and gives $\varrho < 0$, and ϱ is unstable when signal and curvature nearly cancel ($\mathbb{E}[\tilde{Z}D] \approx 0$). We therefore report ϱ with its sign. It is estimated by $\hat{\varrho} = \hat{\theta}_D/\widehat{\mathbb{E}[\tilde{Z}D]}$ with $\widehat{\mathbb{E}[\tilde{Z}D]} = n^{-1}\sum_i \tilde{Z}_i D_i$, and a percentile bootstrap furnishes a confidence interval. We recommend reporting $\hat{\varrho}$ next to the conditional F. A large F paired with a non-negligible $\hat{\varrho}$ signals that the strength is partly manufactured, the signature of mode (a). Yet ϱ is a first-stage gauge and measures only mode (a). It is not a bias estimate, because by Proposition 3 the bias also loads on the outcome-side term θ_Y. A small $\hat{\varrho}$ therefore does not certify a design. When it is paired with a directed-test rejection, that combination is itself the signature of mode (b), namely honest strength with outcome-side bias, as in the health insurance design, where $\hat{\varrho} = 0.09$ yet the estimand more than doubles once corrected. The share thus complements, and does not replace, the outcome-based tests of Sections 4–4.3. Read together, a large F, the share $\hat{\varrho}$, and the directed-test p-value locate which mode, if either, is in play.

5. Bias-corrected estimation and the identification trade-off

The decomposition suggests an immediate correction, namely to subtract the estimated contamination from numerator and denominator. Using $\mathbb{E}[\tilde{Z}Y] - \theta_Y = \mathbb{E}[(Z - e(X))Y]$ and likewise for D, the corrected estimand equals the signal ratio $\bar{L}$. It is implemented by replacing the linear-projection residual $\tilde{Z}$ with the residual from a flexibly estimated propensity $\hat{e}(\cdot)$,

$$\hat{\bar{L}} = \frac{\sum_i \big(Z_i - \hat{e}(X_i)\big)Y_i}{\sum_i \big(Z_i - \hat{e}(X_i)\big)D_i}. \tag{15}$$

Proposition 8 (Corrected estimand and the cost of correction)**.** *(i) Suppose $\hat{e}(\cdot)$ satisfies $\|\hat{e}-e\|_{L^2} = o_p(1)$ and, in the fixed data-generating process, $S_D = \mathbb{E}[\mathrm{Cov}(Z, D \mid X)] \neq 0$. Then under Assumptions 4–5, $\hat{\bar{L}} \xrightarrow{p} \bar{L} = S_Y/S_D$, the saturated specification estimand, a convex average of within-cell IV slopes and, in the binary specialization of Remark 1, a convex average of conditional LATEs. Consistency requires no orthogonality or rate condition. Asymptotic normality and inference do (see the Appendix and Proposition 5), and, since S_D may be small, weak-instrument-robust confidence sets are appropriate.*

(ii) The corrected estimator's own first-stage covariance is exactly the signal term, $\mathbb{E}[(Z - e(X))D] = S_D$. Under the collinearity sequence of Assumption 6, $S_D \to 0$, so the corrected estimator is weakly identified in precisely the regime in which the correction matters most. No specification of the covariates both removes the contamination and preserves the apparent first-stage strength, because the strength was the contamination.

(iii) The corrected and reported (contaminated) first-stage covariances obey the exact identity

$$\underbrace{S_D}_{\text{corrected first stage}} = (1-\varrho) \quad \underbrace{\mathbb{E}[\tilde{Z}D]}_{\text{reported first stage}}, \qquad \varrho = \frac{\theta_D}{\mathbb{E}[\tilde{Z}D]},$$

with ϱ the contamination share (14)*. De-contamination scales the identifying covariance down by exactly ϱ, and the corrected concentration parameter $S_D^2/\{\sigma_v^2\,\mathbb{E}[\mathrm{Var}(Z \mid X)]\}$ carries a factor $(1-\varrho)^2$ in its numerator, so the trade-off of part (ii) is quantitative and known in advance. The fraction of identifying variation surrendered by correcting is the same ϱ that measured how much of the apparent strength was contamination.*

Estimator (15) is the residualized-propensity, or partialling-out, IV estimator, equivalent to the saturation that Blandhol et al. (2022) recommend and here written in influence-function form. What is new is the cost of that correction, stated in parts (ii)–(iii) and made precise by the collinearity limit. A design with small ϱ, such as the health insurance instrument of Section 7.1 ($\hat{\varrho} = 0.09$), is corrected at negligible cost, whereas a design whose strength is mostly curvature ($\varrho \to 1$) cannot be both corrected and strongly identified.

The practical message follows. The analyst facing a strong conditional F under collinearity between instrument and covariates has not found a strong design that merely needs a better second-stage specification. The strength is an artifact of the very nonlinearity that biases the estimand, and removing it returns the design to (correctly) weak. The response is to diagnose with our test, followed by either a redesigned instrument or explicit weak-IV-robust inference on the corrected estimand (Andrews et al., 2019; Montiel Olea and Pflueger, 2013).

6. Monte Carlo evidence

6.1. Designs

Throughout, $X \sim \mathrm{Uniform}(-2, 2)$ and $g(X) = X^2 - \frac{4}{3}$ (mean zero), the per-unit structural effect is $\tau(X) = 1 + 0.25X \in [0.5, 1.5]$, so the hull of within-cell IV slopes is $[0.5, 1.5]$, and the outcome is $Y = A_\mu\, g(X) + \tau(X)D + \mathcal{N}(0,1)$ with $A_\mu = 1.5$. We use four designs that span the support the estimator meets in practice. **(A)** An ordered treatment $D \in \{0, 1, 2, 3\}$, built

from three independent monotone stages $D = \sum_{j=1}^{3} \mathbf{1}\{V_j \le \pi_{0j} + Z\,\delta_j\}$ with a binary instrument $Z \sim \text{Bernoulli}(e(X))$ and $e(X) = \Lambda(\kappa\, g(X))$, where the knob κ drives $\text{Var}(Z \mid X) \to 0$. **(B)** A continuous instrument $Z = e(X) + \sigma\,\mathcal{N}(0,1)$ with $e(X) = 0.3X + 0.6\,g(X)$ and a continuous treatment $D = 0.5 + 0.5\,g(X) + (Z - e(X)) + 0.3\,\mathcal{N}(0,1)$, so $\text{Var}(Z \mid X) = \sigma^2$ is the collinearity knob and $\text{Cov}(Z, D \mid X) = \sigma^2$ the signal. **(C)** An over-identified design with two binary instruments whose propensities e_1, e_2 are nonlinear in X and a binary treatment monotone in both. **(D)** A binary instrument and binary treatment, $e(X) = \Lambda(\kappa g(X))$ with a binary complier structure, the case of Remark 1, used for the finite-sample size, power, and head-to-head exercises below. The within-cell covariance $\text{Cov}(Z, D \mid X)$ is of a single sign by construction in every design, so the sign condition of Section 2 holds and the within-cell hull $[0.5, 1.5]$ is the correct convex benchmark. All numbers are from the `R` scripts released with the paper.

6.2. The collinearity limit

Table 1 traces designs (A) and (B) as $\text{Var}(Z \mid X) \to 0$, with $n = 200\,000$ to approximate the population. In both, the 2SLS estimand leaves the within-cell slope hull $[0.5, 1.5]$ already at mild collinearity and converges monotonically to the theoretical contamination limit $\text{Cov}(h, m_Y)/\,\text{Cov}(h, m_D)$ of Proposition 2 (which the ordered design approaches from below as its own limit falls, and the continuous design hits exactly at 4.0). Throughout, the conditional first-stage F stays enormous, on the order of 12 000 in (A) and in the millions in (B), never warning, confirming Corollary 1. Figure 1 plots both. On a single contaminated dataset from each design ($n = 100\,000$) the directed test rejects decisively, with screening and bias-targeting p-values below 10^{-4}, and the bias-corrected estimator recovers the within-cell benchmark, $0.99\,[0.94, 1.05]$ in (A) and $1.00\,[0.96, 1.04]$ in (B).

6.3. The binary case, where the estimand leaves the LATE hull

Design (D) makes the collinearity limit concrete in the model where the benchmark has a causal reading. Here the within-cell slopes are the conditional LATEs, so the hull $[0.5, 1.5]$ is the LATE hull, and $\bar{L}$ is a convex average of conditional LATEs. Table 2 traces the binary design as κ rises, with $n = 200\,000$. The 2SLS estimand leaves the LATE hull already at mild collinearity and climbs to 8.86, far above $\sup_x \text{LATE}(x) = 1.5$, even though every conditional LATE is positive and bounded by 1.5. It converges to the contamination limit $\text{Cov}(h, m_Y)/\,\text{Cov}(h, m_D)$, which is itself outside the hull, while the conditional first-stage F stays near 40 000 throughout, four orders of magnitude above the threshold of 10. This is the construction behind Proposition 2, an instance in which all $\text{LATE}(x) > 0$ yet the limit exceeds $\sup_x \text{LATE}(x)$. Figure 2 plots the estimand leaving the band of all conditional LATEs against the unmoved F.

6.4. Over-identification and the stacked screen

Design (C) tests the over-identified extension. With two instruments the screen stacks the four contamination moments $(\theta_{D,1}, \theta_{Y,1}, \theta_{D,2}, \theta_{Y,2})$ and refers the Wald statistic to χ^2_4. Over 2000 replications at $n = 3000$, under linear propensities (the null) it rejects 1.5% of the time at the 5% level, and under curved propensities its power is 1.00, while the joint first-stage F of the two

Table 1: The collinearity limit ($n = 200\,000$). In panel (A) the treatment is ordered, $D \in \{0, 1, 2, 3\}$, and the instrument binary. In panel (B) both the instrument and the treatment are continuous. In each the 2SLS estimand leaves the within-cell slope hull $[0.5, 1.5]$ and converges to the contamination limit $\mathrm{Cov}(h, m_Y)/\mathrm{Cov}(h, m_D)$, while the conditional F stays orders of magnitude above the threshold of 10.

Panel (A). Ordered treatment $D \in \{0, 1, 2, 3\}$, binary instrument				
κ	β_{2SLS}	cond. F	contam. limit	$\mathbb{E}[\mathrm{Var}(Z \mid X)]$
0.10	1.530	12 570	151.4	0.249
0.50	3.539	12 635	32.8	0.230
1.00	5.377	12 821	19.1	0.189
2.00	7.241	12 890	13.1	0.119
4.00	8.415	12 666	10.7	0.057
8.00	8.844	12 187	9.78	0.028
32.0	8.857	12 473	9.10	0.007
Panel (B). Continuous instrument and continuous treatment				
σ	β_{2SLS}	cond. F	contam. limit	$\mathbb{E}[\mathrm{Var}(Z \mid X)]$
1.00	1.897	2.8×10^6	4.00	1.000
0.60	2.623	1.5×10^6	4.00	0.360
0.40	3.176	1.1×10^6	4.00	0.160
0.25	3.616	9.1×10^5	4.00	0.062
0.15	3.859	8.2×10^5	4.00	0.022
0.05	3.974	7.9×10^5	4.00	0.003
0.02	3.993	8.0×10^5	4.00	0.000

instruments averages 131 under the null and 246 under the alternative. As in the just-identified case the screen separates designs the first-stage F cannot.

The same design tests the exact restriction (13). Alongside the contaminated design and a linear-propensity clean design, we add a harmless over-identified design, curved propensities with a linear outcome level and a homogeneous effect, where $\theta_D, \theta_Y \neq 0$ yet $\beta_{2SLS} = \bar{L}$. Table 3 reports rejection rates. The stacked screen fires on both the contaminated and the harmless designs, because curvature is present in each, while the exact test $T_n^\dagger$ of (13) has power one against the contaminated design ($\beta_{2SLS} = 9.5$ against $\bar{L} = 1.0$) and holds size on the harmless (4.3%) and clean (5.7%) designs, where $\beta_{2SLS} = \bar{L}$. It is the over-identified counterpart of the head-to-head in Table 4, separating harmful from harmless curvature exactly where the stacked screen and the joint first-stage F cannot.

6.5. Size and power

In design (D), the binary instrument and treatment, two exercises confirm the test's finite-sample behaviour. Both are collected in Appendix B. On a single large contaminated dataset ($n = 100\,000$, $\kappa = 4$) the directed test rejects at $p < 10^{-4}$ while the conditional F is 19 737, and the bias-corrected estimator recovers the true convex average of LATEs (Table B.7). Over 2000 replications at $n = 3000$ it has a conservative size of 1.9% against the 5% level under a linear propensity, and power essentially one under a curved one, at a mean conditional F near 585 in both cases. The F cannot discriminate the two designs the test separates almost perfectly. A power sweep (Table B.9, Figure B.5) raises the induced bias from 0 to 2.64 while the mean conditional F stays pinned near

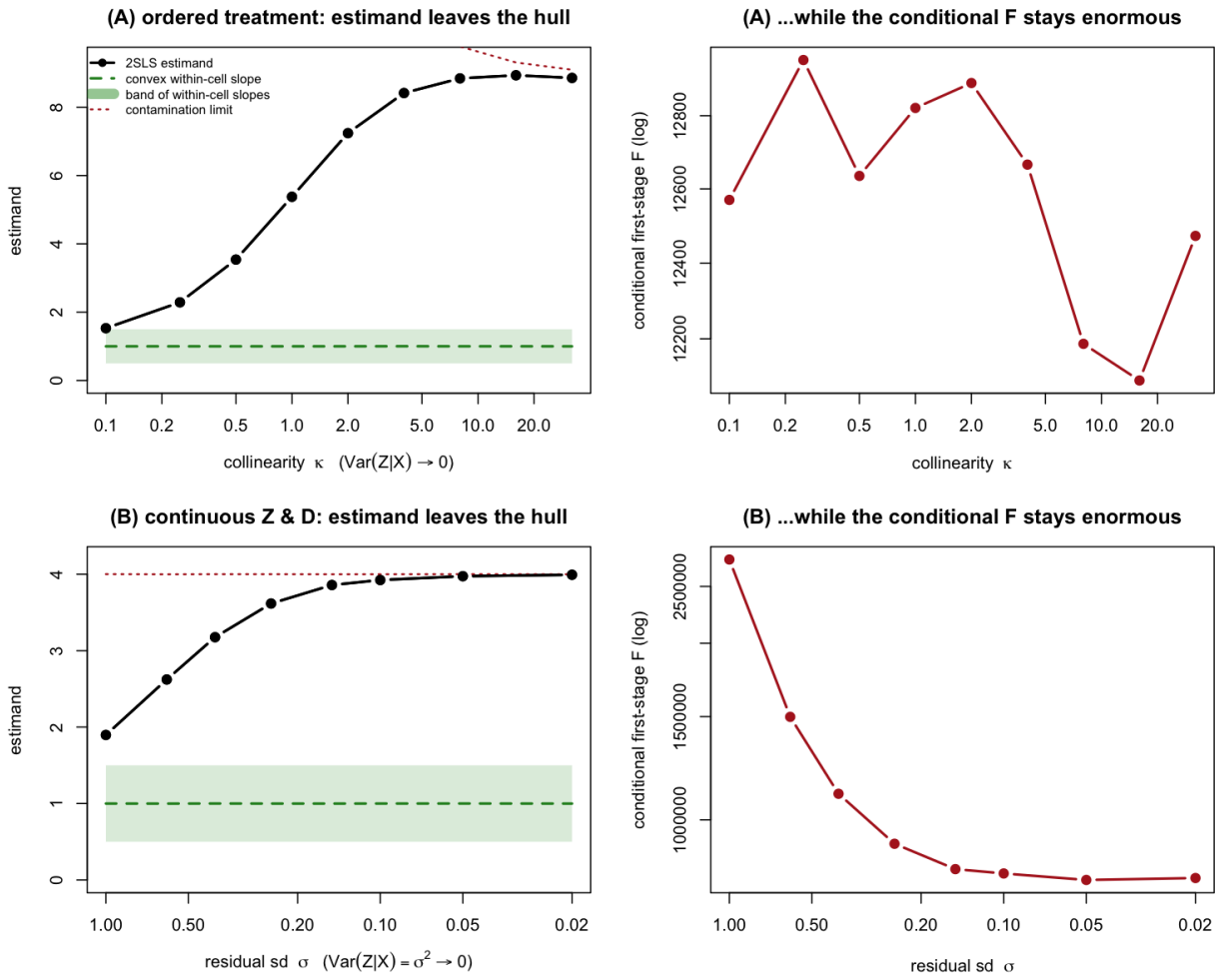

Figure 1: Top row, ordered treatment (A). Bottom row, continuous instrument and treatment (B). In the left panels, as $\mathrm{Var}(Z \mid X) \to 0$ the 2SLS estimand (solid) leaves the band of within-cell IV slopes (shaded, $[0.5, 1.5]$) and departs from the convex within-cell slope (dashed), converging to the contamination limit (dotted). In the right panels the conditional first-stage F (log scale) stays enormous throughout, far above the weak-IV threshold of 10.

1740.

6.6. Head-to-head comparison across three designs

Still in design (D), Table 4 and Figure 3 compare three procedures across three designs at $n = 3000$. The first procedure is a Ramsey RESET testing curvature of the propensity $Z \sim X$, a functional of (Z, X) alone. The second is the strongly identified screening test of Section 4 (H_0 : $\theta_D = \theta_Y = 0$), which we label "Directed-S". The third is the bias-targeting test (9) of Section 4.3, labelled "Directed-X". The contaminated design has a nonlinear propensity and a nonlinear outcome, so it carries real bias. The harmless design has a nonlinear propensity but a linear outcome and a homogeneous effect, so the bias is zero despite strong curvature. The clean design has a linear propensity and zero bias. All three designs carry a mean conditional F near 600.

The pattern illustrates Proposition 3. In the harmless design the RESET rejects 100% of the time. It detects the propensity curvature but, being blind to the outcome, cannot tell that the cur-

Table 2: The binary case ($n = 200\,000$, $A_\mu = 1.5$), design (D). True LATE(X) $\in [0.5, 1.5]$ throughout. The 2SLS estimand leaves the LATE hull and converges to the contamination limit $\mathrm{Cov}(h, m_Y)/\mathrm{Cov}(h, m_D)$, while the conditional F stays near 40 000.

κ	β_{2SLS}	cond. F	contam. limit	$\mathbb{E}[e(1-e)]$
0.10	1.527	39 808	151.4	0.249
0.25	2.316	39 888	61.9	0.245
0.50	3.497	40 459	32.8	0.230
1.00	5.408	38 776	19.1	0.189
2.00	7.325	38 904	13.1	0.118
4.00	8.487	37 790	10.7	0.057
8.00	8.794	37 881	9.77	0.027
16.0	8.877	37 979	9.32	0.013
32.0	8.860	38 215	9.11	0.007

Table 3: Over-identified head-to-head (two instruments, 2000 replications, $n = 3000$, cubic basis, $B = 199$). The stacked screen tests $\theta_D = \theta_Y = 0$ (χ^2_4), and the exact test $T_n^\dagger$ tests $\beta_{2SLS} = \bar{L}$ via (13) (χ^2_1). Rejection rates at the 5% level. The harmless design has curvature present ($\theta \neq 0$) but no bias ($\beta_{2SLS} = \bar{L}$), the case only the exact test resolves.

design	stacked screen	exact $T_n^\dagger$	β_{2SLS}	$\bar{L}$
CONTAMINATED (bias $\neq 0$)	1.00	1.00	9.51	1.00
HARMLESS (bias $= 0$)	1.00	**0.043**	1.00	0.99
CLEAN (bias $= 0$)	0.016	0.057	1.01	1.00

vature is inconsequential, and the same $F \approx 600$ that accompanies a large bias in the contaminated design accompanies zero bias here. Directed-S also raises false alarms (100%), because it tests the sufficient, not the necessary, condition. Only Directed-X separates the three designs as the theory requires. It has full power against genuine contamination, holds size in the harmless design (4.7% in the plug-in form T_n^* and 5.0% in the signal-cleared form A_n) that the other two procedures and the conditional F cannot certify, and sits at 4.1% in the clean design, within sampling error of the nominal level. The two forms coincide here because all three designs are strongly identified, as Proposition 6 anticipates. The clean-design size holds across the sample sizes of Table B.8, from 5.8% at $n = 1000$ to 4.4% at $n = 20\,000$, and is correctly calibrated rather than merely asymptotic, with the sensitivity to the number of studentizing bootstrap draws detailed in that table.

Table B.10 in Appendix B pushes the same contrast along the collinearity knob. Both bias-targeting forms keep full power against genuine bias as $\mathrm{Var}(Z \mid X)$ falls by an order of magnitude, so neither loses power to collinearity. They part company from the screen only in the harmless design. There Directed-S raises a false alarm every time, because its sufficient condition $\theta = 0$ is false whenever the propensity is nonlinear, while both the plug-in Directed-X and the signal-cleared A_n control size throughout, the size guarantee for A_n holding by Proposition 6 even at the smallest $\mathrm{Var}(Z \mid X)$ in the sweep. This is the content of the "report both" recommendation. Directed-S is a cheap screen that, at the cubic basis used here, misses no contamination in the probed directions but cannot vindicate harmless curvature, and Directed-X is the adjudicator that, in those same directions, separates harmful from harmless. Both guarantees are relative to the fixed basis, as the

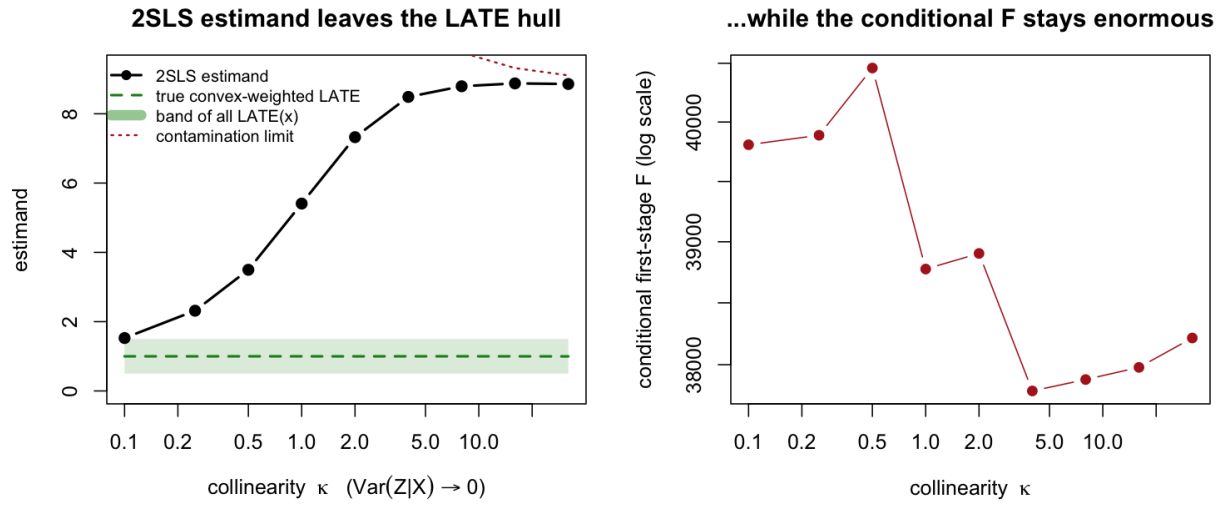

Figure 2: The binary design (D). Left panel. As collinearity κ rises, the 2SLS estimand (solid) leaves the band of all conditional LATEs (shaded) and departs from the true convex average of LATEs (dashed), converging to the contamination limit (dotted). Right panel. The conditional first-stage F (log scale) stays near 40 000 throughout, far above the weak-IV threshold of 10.

Table 4: Head-to-head rejection rates (5% level, 3000 replications, $n = 3000$, cubic basis, $B = 999$). Directed-X is the plug-in T_n^* and Directed-X (AR) the signal-cleared A_n of (11). "mean cond. F" is the average Sanderson–Windmeijer conditional first-stage F. Monte Carlo standard errors are below 0.004.

design	RESET	Directed-S	Directed-X	Directed-X (AR)	mean cond. F
CONTAMINATED (bias $\neq 0$)	1.00	1.00	1.00	1.00	605
HARMLESS (bias $= 0$)	1.00	1.00	**0.047**	**0.050**	607
CLEAN (bias $= 0$)	0.045	0.014	0.041	0.041	573

directed-power discussion of Section 4 notes. A growing or orthogonalized basis extends them to all directions.

6.7. Estimating the contamination share

Table B.12 and Figure B.6 in Appendix B validate the share estimator (14). The plug-in $\hat{\varrho}$ at $n = 3000$ tracks the population share from $\varrho \approx 0.02$ to $\varrho \approx 0.75$ with 93–95% bootstrap coverage. Because the bias also loads on the outcome-side term, a modest share can coincide with substantial bias. At $\varrho = 0.02$ the estimand is already displaced by 1.3 here. Thus $\hat{\varrho}$ flags manufactured strength, while the test of Section 4.3 adjudicates its consequences.

7. Empirical application

We work one application in full and corroborate it with two further designs. The flagship is the effect of spousal health insurance coverage on wives' labour supply (Section 7.1), a binary-instrument, binary-treatment design in which an overwhelming first stage and a small contamination share are both reassuring, yet a standard linear control for husband income leaves the estimate biased by more than a factor of two, a contamination the directed test flags and the correction repairs

Figure 3: Rejection rates of the RESET, the sufficient (Directed-S) and the bias-targeting (Directed-X) tests across the contaminated, harmless, and clean designs. In the harmless design (middle group) the RESET and Directed-S raise false alarms while Directed-X holds size, and the conditional F (near 600 in all groups) cannot discriminate.

without loss of identification. The 401(k) eligibility design (Section 7.2) is a second binary instance of the same mode (b). The slave trade design of Nunn and Wantchekon (2011) (Section 7.3), with a continuous instrument and a continuous treatment, is the real data face of mode (a), manufactured strength, where correcting the propensity collapses a first stage that looked powerful. Throughout, the flexible propensity uses a low-order polynomial (or spline) basis and inference uses the pairs bootstrap of Section 4.

One interpretive point governs how these results should be read, made in the introduction and specialized here to the applications. The curvature $h = e - L(e \mid 1, X)$, the moments $\theta_D = \mathbb{E}[hD]$ and $\theta_Y = \mathbb{E}[hY]$, the share ϱ, and the two directed tests are functionals of $\mathbb{E}[Z \mid X]$ and the level functions alone, so by Sections 4–5 they retain their size and consistency across these applications whatever the support of Z and D. The binary-treatment LATE structure of Remark 1 additionally reads the signal ratio $\bar{L} = S_Y/S_D$ as a convex average of conditional LATEs through the Wald–LATE identity, which holds for the health insurance and 401(k) designs. For the continuous slave trade design $\bar{L}$ is the saturated-propensity linear-IV estimand, and the test there detects functional-form contamination of the reported estimand relative to that benchmark. The convex decomposition for multivalued treatments is the open frontier of Section 9.

7.1. Spousal health insurance and wives' labour supply

Our headline application is the effect of spousal health insurance coverage on the weekly hours worked by married women, the design of Olson (1998). The treatment D is whether a wife is covered by her husband's health insurance, the outcome Y is her weekly hours, and the instrument Z is whether the husband holds insurance through his own job, a precondition for that coverage that shifts it without bearing on the wife's hours except through it. We use the cross-section of $n = 22\,272$ married women assembled by Olson (1998) and distributed with the `Ecdat` package. The instrument is overwhelming, with a conditional first-stage F above 36 000. But husbands hold

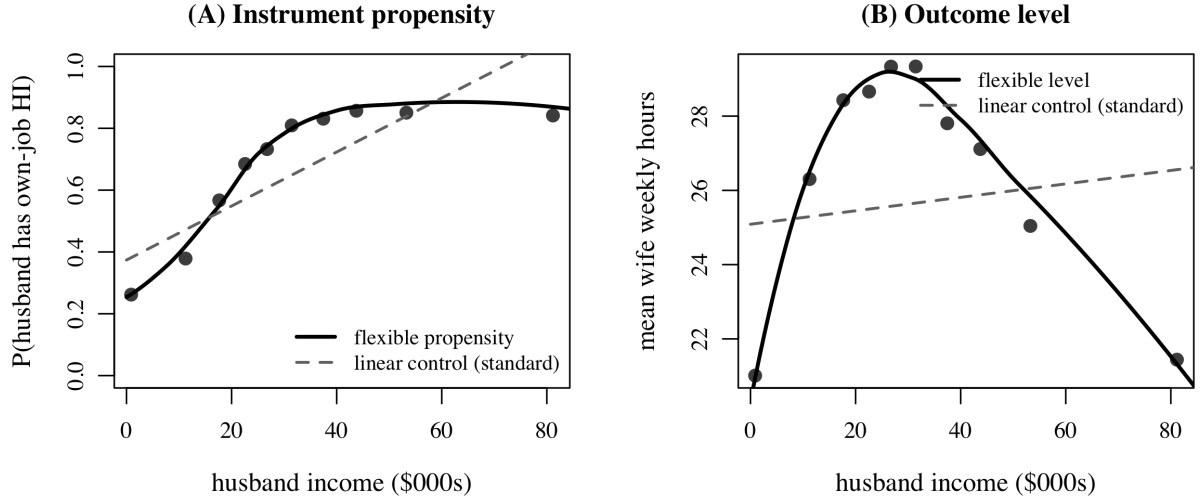

Figure 4: Husband income enters both the instrument propensity and the outcome nonlinearly, while the standard specification enters it linearly. Panel A, the probability that the husband holds own-job insurance (the instrument) against husband income, with binned means, a flexible fit, and the linear control the standard specification imposes. Panel B, mean wife weekly hours (the outcome) against husband income. The propensity curvature a linear control leaves behind (panel A) covaries with the curved outcome level (panel B), and that covariance is the outcome-side contamination θ_Y no first-stage number sees.

own-job coverage in a manner strongly nonlinear in their earnings, and a husband's earnings are a first-order, nonlinear determinant of his wife's hours, while husband income enters the standard specification linearly. This is precisely the configuration of Lemma 1 in which linear covariate adjustment contaminates the estimand.

Figure 4 makes the mechanism visible. The probability that the husband holds own-job insurance rises steeply and then saturates in his income (panel A), so the propensity curvature h that a linear control leaves behind is large, and mean wife hours are single-peaked in husband income (panel B), so the outcome level m_Y that curvature multiplies is itself far from linear. Their covariance is the outcome-side contamination θ_Y, and no first-stage number sees it.

Table 5 reports the just-identified 2SLS estimate under a linear husband income control, with the diagnostic.[2] Under a linear control, the common default, spousal coverage lowers weekly hours by 1.87. Both the screening and the bias-targeting tests reject decisively ($p < 10^{-3}$) despite $F > 36\,000$, and the bias-corrected estimand is -4.45 ($[-5.14, -3.81]$), more than double the naive figure with an interval that excludes it. The correction is validated the way it is constructed. Controlling husband income by a natural spline in an ordinary 2SLS returns -4.51, which the correction from the linear specification reproduces to the second decimal. The finding is invariant to the basis, moving from

[2]Treatment D is coverage by the husband's health insurance, instrument Z is the husband holding insurance through his own job, outcome Y is the wife's weekly hours worked. The linear control vector is (education, potential experience, husband income, children under six, children six to eighteen, race, Hispanic origin, region). The curvature basis $b(X)$ for the naive specification is a cubic in husband income $(\text{inc}^2, \text{inc}^3)$ residualized against the linear controls, and the flexible benchmark controls husband income by a natural spline with six degrees of freedom. Inference is a pairs bootstrap with $B = 999$ resamples that refits the propensity, the basis, and the level functions on each resample. Data ship with the `Ecdat` package, code in `hi_full.R`.

Table 5: Spousal health insurance and wives' weekly hours ($n = 22\,272$). Effect of coverage by the husband's insurance, instrumented by the husband holding own-job insurance. A conditional F above 36 000 does not prevent contamination under a linear husband income control, and the correction recovers the flexible-income answer without loss of identification. Test p-values from the pairs bootstrap. For the corrected row, cond. F is the corrected estimator's own first stage (regressing D on the residualized instrument $Z - \hat{e}(X)$) and the bracketed interval is the pairs bootstrap.

husband income control	2SLS effect	cond. F	$\hat{\varrho}$	Directed-S p	Directed-X p
linear (naive)	-1.87	36301	0.09	< 0.0001	< 0.0001
natural spline (benchmark)	-4.51	n/a	n/a	n/a	n/a
contamination-corrected	-4.45 $[-5.14, -3.81]$	*31314*			

-3.97 under a quadratic control to -4.51 under a six-knot spline, and both tests reject at every basis.

This is the paper's thesis on real data, in near-pure mode (b). The first stage is honest and immense, and the contamination share is only $\hat{\varrho} = 0.09$, so both numbers a first stage reports, the conditional $F > 36\,000$ and the small share, describe a clean, strongly identified design. They miss the bias because it loads on the outcome side, through θ_Y rather than the first-stage term θ_D. By Proposition 3 no first-stage diagnostic could reveal this, and none does. A propensity RESET would register the income curvature as present, as the screening test does, but presence is not harm, and only the outcome-directed test and the correction establish that the curvature reaches the estimate here and return the specification-robust answer of -4.45. The design in which every first-stage number is reassuring is the one that is biased.

Unlike a design in which correction destroys identification (Proposition 8), here that caveat does not bind. Because the contamination share is small, the corrected estimator retains almost all of the first-stage covariance, and its own first-stage F, regressing D on the residualized instrument $Z - \hat{e}(X)$, is 31 314. The corrected estimate is therefore strongly identified and its bootstrap interval is valid. This is the same eligibility structure as the 401(k) design below, a treatment attainable only where an instrument grants eligibility, but it is the cleaner illustration for two reasons. The linear control at issue is husband income, which a labour-supply equation enters linearly as a matter of course rather than a specification a careful analyst would already avoid, and Olson (1998) set the design up precisely to compare parametric and semiparametric estimates, having suspected the functional form that the diagnostic now localizes and tests against the outcome.

7.2. A second binary instance: the 401(k) design

The 401(k) eligibility design reproduces the pattern in the most studied strong-instrument application in the literature. The effect of 401(k) participation on household net financial assets is identified using 401(k) eligibility as an instrument (Poterba et al., 1995; Abadie, 2003), on the cross-section of $n = 9{,}275$ households of Wooldridge (2010). Eligibility is overwhelming, with a conditional first-stage F exceeding 7,000, and firms offer it in a manner strongly nonlinear in income while income is a first-order nonlinear determinant of wealth, the configuration of Lemma 1.

Table 6 reports the just-identified 2SLS estimate (participation on net financial assets, in thou-

Table 6: The 401(k) design (n = 9,275). Effect of participation on net financial assets (\$000s), instrumented by eligibility. A first-stage $F > 7{,}000$ does not prevent contamination under linear income control, and the correction recovers the flexible-income answer. Test p-values from the pairs bootstrap. For the corrected row, cond. F is the corrected estimator's own first stage (regressing D on the residualized instrument $Z - \hat{e}(X)$), the bracketed interval is the pairs bootstrap, and its identification-robust Anderson–Rubin 95% interval is $[8.2, 17.0]$.

income control	2SLS effect	cond. F	$\hat{\varrho}$	Directed-S p	Directed-X p
linear (naive)	8.40	7401	0.03	< 0.0001	0.0001
quadratic	13.66	7423	0.00	0.121	0.113
natural splines (benchmark)	13.05	n/a	n/a	n/a	n/a
contamination-corrected	*12.6 [9.0, 16.0]*	*7295*			

sands of dollars) under three income specifications, with the diagnostic.[3] Under a linear income control, the common default, the estimate is 8.40. Both directed tests reject decisively ($p < 10^{-3}$) despite $F > 7{,}000$, and the correction returns 12.6 ($[9.0, 16.0]$), reproduced by a quadratic income control (13.7, which also clears the flag at $p_X = 0.11$) and a natural-spline control (13.1). The naive linear-income estimate is the outlier, understating the effect by roughly a third. The contamination share is $\hat{\varrho} = 0.03$, so the displacement is almost entirely the outcome-side term θ_Y, and the corrected estimator's own first stage, 7295, confirms that de-contamination costs no identification here.

Here too a first-stage F above seven thousand offers no protection, and by Proposition 3 none could. The corrected estimand's identification-robust Anderson–Rubin 95% interval, $[8.2, 17.0]$, coincides with the normal-theory one, so the corrected estimate is strongly identified, and controlling income flexibly is exactly what the correction reproduces.

7.3. Manufactured strength: the slave trade instrument

Mode (a) also has a real data face. Nunn and Wantchekon (2011) instrument an ethnic group's historical slave exports with its distance from the coast to estimate the effect of the slave trade on present-day trust, one of the most cited designs in the field, on $n = 17\,371$ individuals in 156 ethnic-group clusters. We state at the outset what the exercise claims and what it does not. It shows that the reported first-stage strength is fragile to how the geographic covariates enter, not that the slave-trade effect on trust is spurious. The corrected estimates below stay negative and economically meaningful, and the object of study is the diagnostic, not the magnitude of the effect. The instrument looks powerful, with an unclustered first-stage F above 1200. Clustering at the ethnic-group level at which it varies already cuts this to 17, and the diagnostic explains why even the honest number is generous. The contamination share is $\hat{\varrho} = 0.52\,[0.13, 1.00]$. About half of the residualized instrument is smooth geographic variation, the curvature of latitude and longitude that distance-to-coast inherits, rather than identifying variation. Entering those coordinates flexibly

[3]Full specification, from the released `iv_401k.R`. Treatment D is 401(k) participation, instrument Z is 401(k) eligibility, outcome Y is net financial assets. The linear control vector is (income, age, married, male, family size). The curvature basis $b(X)$ for the linear specification is $(\text{inc}^2, \text{inc}^3, \text{age}^2, \text{inc} \times \text{age})$ residualized against the linear controls. The quadratic specification adds $\text{inc}^2, \text{age}^2$ to the controls with basis $(\text{inc}^3, \text{inc}^4, \text{age}^3, \text{inc} \times \text{age})$. The spline robustness check uses a natural-spline income basis with six degrees of freedom, and the flexible benchmark controls income and age by natural splines with five and four degrees of freedom. Inference is a pairs bootstrap with $B = 1499$ resamples that refits the propensity, the curvature basis, and the level functions on each resample.

rather than linearly, as Nunn and Wantchekon (2011) do, moves the trust-in-neighbours estimate from -0.243 to -0.190 and attenuates the related trust outcomes by as much as a half, and it collapses the corrected instrument's own first stage from 17 to 2.4. This is manufactured strength in the sense of Proposition 2. Much of the apparent identification was the covariate curvature the conditional F cannot separate from signal, and removing it leaves too little variation to identify the effect. The bias-targeting test does not reject here, as Proposition 4 leads one to expect under weak identification, since once the manufactured component is stripped the design no longer speaks, while the screening test, which needs no identification, still registers the curvature. The instrument's apparent strength is thus in material part geographic functional form, invisible to the first stage and visible to the share. Because distance-to-coast is nearly a deterministic function of the coordinates, controlling them flexibly absorbs much of any geographic instrument, so what the design establishes is the fragility of the reported strength to how geography enters, not the magnitude of the effect itself. Appendix Table B.11 shows the collapse does not depend on the cubic basis. Entering the coordinates quadratically, cubically, or through natural splines leaves the contamination share near one half and drives the corrected first stage to between 2.4 and 3.0 in every case, for trust in relatives as well as trust in neighbours, while the cluster-robust first stage of the original instrument stays at 17.0 throughout.

Taken together the three applications exhibit both failure modes on canonical data. The health insurance and 401(k) designs are mode (b), where an honest first stage, above thirty-six thousand and above seven thousand, conceals an outcome-side bias the directed test detects and the correction repairs. The slave trade design is mode (a), where the apparent strength is in material part covariate curvature and correcting it collapses the first stage. In each the conditional F is silent, whether because the bias loads on the outcome or because the strength itself is manufactured, and the directed test and the reportable share speak where it cannot.

8. Recommendations for practice

Our results imply a short protocol that adds little cost to a standard IV analysis. (1) Report the conditional first-stage F of Sanderson and Windmeijer (2016), but do not treat it as evidence for a causal reading of the second stage, and report the estimated contamination share $\hat{\varrho}$ (14) beside it. (2) Estimate the instrument propensity $\mathbb{E}[Z \mid X]$ flexibly and test its curvature against D and Y using the screening statistic of Section 4 and, to distinguish harmful from harmless curvature, the bias-targeting test of Section 4.3. A screening rejection signals curvature that loads on D and Y even when the conditional F is large, but it does not by itself establish bias. Use the bias-targeting test to decide, reading its signal-cleared form A_n (11) for a size guarantee that holds even when the design is weakly identified and the plug-in T_n^* for its extra power when it is not. (3) If the bias-targeting test rejects, report the bias-corrected estimand (15) together with weak-IV-robust inference, and recognize (Proposition 8) that a collapse in the corrected first stage means the apparent strength was contamination rather than identification. A screening rejection that the bias-targeting test does not confirm is the harmless-curvature case and calls for no correction. (4) Where feasible, prefer an instrument whose propensity is approximately linear in the controls, or saturate the controls at the design stage while monitoring the resulting conditional strength.

9. Conclusion

The applied ritual of reporting a first-stage F answers whether an instrument moves the treatment, not whether the second-stage estimate is a meaningful causal average. Under collinearity between instrument and covariates these questions come apart in a specific, detectable way. The same propensity curvature that contaminates the 2SLS estimand also inflates the conditional F, so the strength diagnostic cannot see the problem and the natural saturation fix trades the bias for weak identification. A directed conditional-moment test, built entirely from reduced-form regressions, has power exactly where the conditional F is silent.

The analysis covers a scalar instrument and a scalar treatment of arbitrary support, entered in a just-identified linear model with linear covariates, with the binary-treatment LATE model as the leading special case. We now say what survives beyond that boundary and what does not. Over-identification is handled in Section 4.4, the less mechanical direction. With a vector of instruments the numerator and denominator of Lemma 1 become signal-plus-contamination vectors and 2SLS combines them through the residualized-instrument second-moment matrix, so the population coefficient is a weighted combination rather than a clean term-by-term ratio. The stacked screen $\theta_D = \theta_Y = 0$ stays sufficient for no contamination, the impossibility result (Proposition 3) holds verbatim since the bias still loads on the outcome, and the exact bias-targeting restriction, which inverts that second-moment matrix, is the single moment (13), confirmed against the stacked screen in Section 6.4. Multiple, that is vector, treatments are the substantive frontier. A multivalued single treatment is already covered above, but with a vector of treatments our single-treatment functional-form contamination compounds with the cross-treatment contamination of Goldsmith-Pinkham et al. (2024), a conceptually distinct channel that persists even under a saturated propensity. Disentangling the two, where the object to be tested is a matrix rather than a scalar contamination, is the main extension we leave to future work. What extends cleanly in every case is the reason the first stage cannot help. The bias always loads on the conditional law of the outcome, so Proposition 3 is the general fact, of which the binary model is one instance. The remaining loose end, a data-driven curvature basis, is closed by Proposition 5.

References

Andrews, I., Stock, J.H., Sun, L., 2019. Weak instruments in instrumental variables regression: theory and practice. Annual Review of Economics 11, 727–753.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–455.

Abadie, A., 2003. Semiparametric instrumental variable estimation of treatment response models. Journal of Econometrics 113, 231–263.

Bierens, H.J., 1990. A consistent conditional moment test of functional form. Econometrica 58, 1443–1458.

Nunn, N., Wantchekon, L., 2011. The slave trade and the origins of mistrust in Africa. American Economic Review 101, 3221–3252.

Olson, C.A., 1998. A comparison of parametric and semiparametric estimates of the effect of spousal health insurance coverage on weekly hours worked by wives. Journal of Applied Econometrics 13, 543–565.

Poterba, J.M., Venti, S.F., Wise, D.A., 1995. Do 401(k) contributions crowd out other personal saving? Journal of Public Economics 58, 1–32.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, 2nd ed. MIT Press.

Blandhol, C., Bonney, J., Mogstad, M., Torgovitsky, A., 2022. When is TSLS actually LATE? NBER Working Paper 29709.

Goldsmith-Pinkham, P., Hull, P., Kolesár, M., 2024. Contamination bias in linear regressions. American Economic Review 114, 4015–4051.

Zhao, A., Ding, P., Li, F., 2025. Two-stage least squares with treatment–covariate interactions for treatment effect heterogeneity. Working paper, arXiv:2502.00251.

Krumme, A., Westphal, M., 2026. Testing IV validity and LATE interpretation using flexible covariate specifications. IZA Discussion Paper 18573.

Dossani, A., Dotson, J.P., Schonlau, R., 2024. Contaminated control variables in 2SLS models. Working paper.

Słoczyński, T., Sun, L., Uysal, S.D., 2026. A practical guide to instrumental variables methods with heterogeneous treatment effects. Working paper, arXiv:2605.15115.

Young, A., 2024. Nearly collinear robust procedures for 2SLS estimation. The Stata Journal 24, 3–28.

van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press.

Bound, J., Jaeger, D.A., Baker, R.M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association 90, 443–450.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21, C1–C68.

Conley, T.G., Hansen, C.B., Rossi, P.E., 2012. Plausibly exogenous. Review of Economics and Statistics 94, 260–272.

de Chaisemartin, C., 2017. Tolerating defiance? Local average treatment effects without monotonicity. Quantitative Economics 8, 367–396.

Heckman, J.J., Vytlacil, E., 2005. Structural equations, treatment effects, and econometric policy evaluation. Econometrica 73, 669–738.

Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–475.

van Kippersluis, H., Rietveld, C.A., 2018. Beyond plausibly exogenous. The Econometrics Journal 21, 316–331.

Mogstad, M., Torgovitsky, A., 2018. Identification and extrapolation of causal effects with instrumental variables. Annual Review of Economics 10, 577–613.

Mogstad, M., Torgovitsky, A., 2024. Instrumental variables with unobserved heterogeneity in treatment effects, in: Handbook of Labor Economics. Elsevier.

Montiel Olea, J.L., Pflueger, C., 2013. A robust test for weak instruments. Journal of Business & Economic Statistics 31, 358–369.

Newey, W.K., 1985. Maximum likelihood specification testing and conditional moment tests. Econometrica 53, 1047–1070.

Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing, in: Handbook of Econometrics, vol. 4. Elsevier, pp. 2111–2245.

Ramsey, J.B., 1969. Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society B 31, 350–371.

Sanderson, E., Windmeijer, F., 2016. A weak instrument F-test in linear IV models with multiple endogenous variables. Journal of Econometrics 190, 212–221.

Słoczyński, T., 2022. When should we (not) interpret linear IV estimands as LATE? Working paper, Brandeis University.

Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65, 557–586.

Stevenson, M.T., 2026. Humans in the loop: the next frontier in the credibility revolution. Journal of Economic Literature, forthcoming.

Stock, J.H., Yogo, M., 2005. Testing for weak instruments in linear IV regression, in: Identification and Inference for Econometric Models. Cambridge University Press, pp. 80–108.

Tauchen, G., 1985. Diagnostic testing and evaluation of maximum likelihood models. Journal of Econometrics 30, 415–443.

Appendix A. Proofs

Proof of Lemma 1

By (2), $\beta_{2SLS} = \mathbb{E}[\tilde{Z}Y]/\mathbb{E}[\tilde{Z}D]$ with $\tilde{Z} = Z - \ell(X)$. Write $\tilde{Z} = (Z - e(X)) + h(X)$ using $h = e - \ell$. For any conditionally square-integrable $W \in \{D, Y\}$,

$$\mathbb{E}[\tilde{Z}W] = \mathbb{E}[(Z - e(X))W] + \mathbb{E}[h(X)W].$$

For the first term, condition on X. Since $\mathbb{E}[Z - e(X) \mid X] = 0$, $\mathbb{E}[(Z - e(X))W \mid X] = \mathbb{E}[(Z - e(X))(W - m_W(X)) \mid X] = \mathrm{Cov}(Z, W \mid X)$, and taking expectations over X gives S_W. For the second term, $\mathbb{E}[h(X)W] = \mathbb{E}[h(X)\,\mathbb{E}[W \mid X]] = \mathbb{E}[h(X)m_W(X)] = \mathrm{Cov}(h(X), m_W(X))$, the last equality because $\mathbb{E}[h(X)] = 0$. This gives θ_W and (4). Equation (5) follows by subtracting $\bar{L} = S_Y/S_D$ from (4) and simplifying, and $S_D + \theta_D = \mathbb{E}[\tilde{Z}D]$ by construction. The binary specialization (Remark 1) follows because for binary Z, $\mathrm{Cov}(Z, W \mid X) = e(X)\{1-e(X)\}\{\mathbb{E}[W \mid Z = 1, X] - \mathbb{E}[W \mid Z = 0, X]\}$, which is $e(1-e)p$ for $W = D$ and, by the conditional Wald identity under the LATE assumptions, $e(1-e)p$ LATE for $W = Y$. □

Proof of Proposition 2

By Assumption 6(i), $\mathrm{Var}_t(Z \mid X) \to 0$ a.e. and in L^1. By Cauchy–Schwarz, $|\mathrm{Cov}(Z, W \mid X)| \le \sqrt{\mathrm{Var}_t(Z \mid X)\,\mathrm{Var}_t(W \mid X)}$. For $W = D$, with $\sup_t \mathbb{E}[\mathrm{Var}_t(D \mid X)] < \infty$, the Cauchy–Schwarz inequality applied to the outer expectation gives $|S_{D,t}| \le \sqrt{\mathbb{E}[\mathrm{Var}_t(Z \mid X)]\,\mathbb{E}[\mathrm{Var}_t(D \mid X)]} \to 0$. For $W = Y$, the integrand $|\mathrm{Cov}_t(Z, Y \mid X)| \le \sqrt{\mathrm{Var}_t(Z \mid X)}\sqrt{\mathrm{Var}_t(Y \mid X)} \to 0$ a.e., and the family is uniformly integrable because $\{\mathrm{Var}_t(Y \mid X)\}$ is and $\mathrm{Var}_t(Z \mid X)$ is bounded, so Vitali's theorem gives $S_{Y,t} \to 0$. By Assumption 6(ii)–(iii), $h_t \to h^*$ and $m_{D,t}, m_{Y,t}$ converge in L^2, so by Cauchy–Schwarz $\theta_{D,t} \to \mathrm{Cov}(h^*, m_D^*)$ and $\theta_{Y,t} \to \mathrm{Cov}(h^*, m_Y^*)$. The denominator converges to $\mathrm{Cov}(h^*, m_D^*) \neq 0$ by Assumption 6(iv), so (6) follows from the continuous mapping theorem. That the limit can fall outside the hull of within-cell slopes, and in the binary specialization outside the LATE hull, is confirmed by the binary design of Section 6.3 (Table 2, Figure 2), where every conditional LATE is positive and bounded by 1.5 yet the estimand converges to a limit near 9. □

Proof of Corollary 1

The population conditional first-stage slope from regressing D on $\tilde{Z}$ is $\pi_t = \mathbb{E}[\tilde{Z}_t D]/\mathrm{Var}(\tilde{Z}_t)$. By Lemma 1, $\mathbb{E}[\tilde{Z}_t D] = S_{D,t} + \theta_{D,t} \to \mathrm{Cov}(h^*, m_D^*)$, and $\mathrm{Var}(\tilde{Z}_t) = \mathbb{E}[\mathrm{Var}_t(Z \mid X)] + \mathbb{E}[h_t^2] \to \mathbb{E}[(h^*)^2]$ by Assumption 6(i)–(ii), using $\mathrm{Var}(\tilde{Z}_t) = \mathbb{E}[\mathrm{Var}_t(Z \mid X)] + \mathrm{Var}(h_t(X))$ and $\mathbb{E}[h_t] = 0$. Hence $\pi_t \to \mathrm{Cov}(h^*, m_D^*)/\mathbb{E}[(h^*)^2] \neq 0$. The conditional F has population analogue $n\,\pi_t^2\,\mathrm{Var}(\tilde{Z}_t)/\sigma_v^2$ with $\sigma_v^2 = \mathrm{Var}(D - \pi_t\tilde{Z} - X'c)$ bounded, so it is bounded away from zero and diverges in n. □

Proof of Proposition 3

Fix the joint law of (Z, D, X) with $h \neq 0$. Then $e(\cdot)$, $\ell(\cdot)$, $h(\cdot)$, $m_D(\cdot)$, $S_D = \mathbb{E}[\mathrm{Cov}(Z, D \mid X)]$, $\theta_D = \mathbb{E}[h(X)D]$ and $\mathrm{Var}(\tilde{Z})$ are all determined, so any functional T of this law takes a fixed value. By (5), $\beta_{2SLS} - \bar{L} = (\theta_Y - \bar{L}\theta_D)/(S_D + \theta_D)$ with $\theta_Y = \mathbb{E}[h(X)Y]$ and $\bar{L} = S_Y/S_D$, $S_Y = \mathbb{E}[\mathrm{Cov}(Z, Y \mid X)]$. Both θ_Y and S_Y are determined by the conditional law of Y given (Z, X), which is not restricted by the marginal law of (Z, D, X). Construct two outcome laws sharing that

marginal but with conditional means m_Y and m'_Y that differ on the span orthogonal to $(1, X')$, so that $\mathrm{Cov}(h, m_Y) \neq \mathrm{Cov}(h, m'_Y)$. Because $h \neq 0$ such a perturbation exists. The perturbation changes only the level $\mathbb{E}[Y \mid X]$ and leaves the law of (Z, D, X) and the within-cell slopes untouched, so both DGPs are admissible. The two biases then differ while T is identical, and since the gap $\mathrm{Cov}(h, m_Y) - \mathrm{Cov}(h, m'_Y)$ can be made arbitrarily large, the bias is unbounded as a function of the outcome law at fixed T. Hence T carries no information about the bias. When $h \equiv 0$ the construction is empty because $\mathrm{Cov}(h, \cdot) \equiv 0$, consistent with the first stage then certifying no contamination. □

Proof of Proposition 4, part (i)

$\hat{\theta}$ is a two-step estimator, a first-step OLS $\hat{\delta}$ from regressing Z on $(1, X', b(X)')$, and second-step sample means of $\hat{h}_i W_i$, $W \in \{D, Y\}$. Under the stated moment conditions and fixed K, $\hat{\delta} \xrightarrow{p} \delta$ and is asymptotically linear with influence $G^{-1} b(X_i) u_i$, $G = \mathbb{E}[bb']$. A mean-value expansion of $\hat{\theta}_W = n^{-1} \sum_i b(X_i)' \hat{\delta}\, W_i$ around δ gives $\sqrt{n}(\hat{\theta}_W - \theta_W) = n^{-1/2} \sum_i \{b(X_i)' \delta\, W_i - \theta_W\} + \mathbb{E}[W b(X)'] \sqrt{n}(\hat{\delta} - \delta) + o_p(1)$. Stacking $W \in \{D, Y\}$ and substituting the first-step influence gives the influence function

$$\psi_i = \begin{pmatrix} b(X_i)' \delta\, D_i - \theta_D \\ b(X_i)' \delta\, Y_i - \theta_Y \end{pmatrix} + \Big(\mathbb{E}[D\, b(X)'],\ \mathbb{E}[Y\, b(X)'] \Big)' G^{-1} b(X_i) u_i, \qquad u_i = Z_i - \alpha_0 - X_i' \alpha_1 - b(X_i)' \delta.$$

The central limit theorem yields $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma)$ with $\Sigma = \mathrm{Var}(\psi_i)$. Under H_0, $\theta = 0$ and $W_n = n \hat{\theta}' \hat{\Sigma}^{-1} \hat{\theta} \xrightarrow{d} \chi^2_2$ provided $\hat{\Sigma} \xrightarrow{p} \Sigma \succ 0$. Consistency against fixed alternatives and the local power statement are immediate from $\hat{\theta} \xrightarrow{p} \theta \neq 0$ and $\theta = \eta / \sqrt{n}$, respectively. Consistency of the pairs bootstrap for Σ follows from Hadamard differentiability of $\hat{\theta}$ in the empirical measure. □

Proof of Proposition 4, part (ii)

Fix a sequence $\{P_n\}$ as stated. The estimator $\hat{\theta}$ and its influence function $\psi_i = (b(X_i)' \delta\, W_i - \theta_W)_{W \in \{D,Y\}} + \big(\mathbb{E}[W b(X)']\big)_W G^{-1} b(X_i) u_i$ from part (i) above involve the first-stage strength nowhere. They are determined by the reduced-form regression of Z on $(1, X', b(X)')$ and by the cross-moments $\mathbb{E}[W b(X)']$, none of which is the first-stage covariance $\mathbb{E}[\tilde{Z} D]$. The weak-instrument problem is a small-denominator problem for $\beta_{2SLS} = \mathbb{E}[\tilde{Z} Y] / \mathbb{E}[\tilde{Z} D]$, and that denominator does not appear. The uniform fourth-moment bound and $\Sigma_n \to \Sigma_* \succ 0$ posited for the null sequence supply a Lyapunov condition, so the Lindeberg–Feller central limit theorem for triangular arrays applies to $n^{-1/2} \sum_i \psi_{i,n}$, giving $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_*)$ along the sequence. Under H_0, $\theta = 0$ and $\hat{\Sigma} \xrightarrow{p} \Sigma_*$, so $W_n \xrightarrow{d} \chi^2_2$. The generated-regressor term $\mathbb{E}[W b(X)'] G^{-1} b(X_i) u_i$ has variance of order $\mathrm{Var}(u_i) \to 0$ as Z becomes collinear, but the leading term $b(X_i)' \delta\, W_i - \theta_W$ has variance bounded below because the curvature stays nondegenerate, $h_n \to h^* \neq 0$ and hence $\delta \to \delta^* \neq 0$. This delivers $\Sigma_* \succ 0$. The identical argument applied to the orthogonalized influence ψ_i^o, which is likewise free of the first-stage covariance, gives $W_n^o \xrightarrow{d} \chi^2_2$. Finally, along the contaminated sequence of Assumption 6 the same central limit theorem gives $\sqrt{n}(\hat{\theta} - \theta_n) \xrightarrow{d} N(0, \Sigma_*)$ while $\theta_n \to \theta^*$ with $\theta^*_D = \mathrm{Cov}(h^*, m^*_D) \neq 0$. The noncentrality $n\, \theta_n' \Sigma_n^{-1} \theta_n \to \infty$, so $W_n \to \infty$ in probability and the power tends to one. □

Proof of Proposition 5

Mean zero at the truth is immediate, because $\mathbb{E}[\psi_W^o] = \mathbb{E}[h(X)W] + \mathbb{E}[\tilde{m}_W(X)(Z - e(X))] - \theta_W = \theta_W + 0 - \theta_W = 0$, using $\mathbb{E}[Z - e(X) \mid X] = 0$. For orthogonality with respect to e, perturb $e \to e + t\Delta$. Because $h = e - L(e \mid 1, X)$ is linear in e and the linear projection is self-adjoint, $\partial_t \mathbb{E}[h(X)W]\big|_{t=0} = \mathbb{E}[(\Delta - L(\Delta \mid 1, X))\, m_W(X)] = \mathbb{E}[\Delta\, \tilde{m}_W(X)]$, while $\partial_t \mathbb{E}[\tilde{m}_W(X)(Z - e(X))]\big|_{t=0} = -\mathbb{E}[\tilde{m}_W(X)\, \Delta(X)]$. The two terms cancel. Perturbing $m_W \to m_W + t\,\Xi$ moves only the correction term, with derivative $\mathbb{E}[\tilde{\Xi}(X)(Z - e(X))] = \mathbb{E}[\tilde{\Xi}(X)\, \mathbb{E}(Z - e(X) \mid X)] = 0$. Hence ψ_W^o is Neyman-orthogonal in both nuisances. The affine projection $L(\cdot \mid 1, X)$ is not a third nuisance but a linear functional of e. The e-perturbation above moves $h = e - L(e \mid 1, X)$ through the projection, and the same self-adjointness that makes the two derivatives cancel annihilates its first-order contribution to the pathwise derivative. In the estimator the projection is a parametric $\sqrt{n}$ step, an OLS of Z on $(1, X)$. Its estimation error contributes a regular $O_p(n^{-1/2})$ term to the influence function, not orthogonalized away but asymptotically linear, which the pairs bootstrap captures by recomputing the projection, the curvature basis, and the level functions on each resample. With K-fold cross-fitting the plug-in decomposes as $\hat{\theta}_W^o - \theta_W = n^{-1} \sum_i \psi_W^o(O_i; e, m_W) + (\text{bias}) + (\text{remainder})$. Orthogonality bounds the bias by the product $\|\hat{e} - e\|_{L^2} \|\hat{m}_W - m_W\|_{L^2} = o_p(n^{-1/2})$, and cross-fitting makes the remainder $o_p(n^{-1/2})$ under the stated L^2 rates and fourth moments (Chernozhukov et al., 2018). The leading term is a sample average of the fixed influence ψ_i^o, so $\sqrt{n}(\hat{\theta}^o - \theta) \xrightarrow{d} N(0, \Sigma_o)$ by the central limit theorem, and $W_n^o \xrightarrow{d} \chi_2^2$ under H_0 follows as in the proof of Proposition 4. □

Proof of Proposition 8

Write the numerator of (15) as $n^{-1} \sum_i (Z_i - \hat{e}(X_i)) Y_i = n^{-1} \sum_i (Z_i - e(X_i)) Y_i - n^{-1} \sum_i (\hat{e}(X_i) - e(X_i)) Y_i$. The first term converges in probability to $\mathbb{E}[(Z - e(X))Y] = S_Y$ by the argument in the proof of Lemma 1. For the second, by Cauchy–Schwarz its magnitude is at most $(n^{-1} \sum_i (\hat{e}(X_i) - e(X_i))^2)^{1/2} (n^{-1} \sum_i Y_i^2)^{1/2} = \|\hat{e} - e\|_n \cdot O_p(1)$, where $\|\hat{e} - e\|_n^2 = n^{-1} \sum_i (\hat{e}(X_i) - e(X_i))^2$ is the in-sample empirical norm. This is $o_p(1)$ whenever $\|\hat{e} - e\|_n = o_p(1)$. For the fixed, low-dimensional basis used throughout, the empirical norm converges to the population norm $\|\hat{e} - e\|_{L^2}$ by a uniform law of large numbers, so mere L^2-consistency of the propensity estimator suffices, with no orthogonality or rate requirement. For a flexible, high-dimensional or machine-learning $\hat{e}$ trained in-sample, population L^2-consistency need not control the empirical norm, and one should then cross-fit $\hat{e}$, evaluating it on observations not used to fit it, so that the displayed average is out-of-sample and the argument goes through unchanged. The denominator converges to $\mathbb{E}[(Z - e(X))D] = S_D$ identically, and provided $S_D \neq 0$ in the fixed data-generating process the ratio converges to $S_Y / S_D = \bar{L}$, proving part (i). (Asymptotic normality and inference for $\hat{\bar{L}}$, as opposed to consistency, do require controlling the first-step influence, for example a Neyman-orthogonalized moment with cross-fitting, and, because S_D may be small, weak-instrument-robust confidence sets. See Proposition 5 and Section 5.) For part (ii), the corrected estimator's first-stage covariance is exactly this denominator $\mathbb{E}[(Z - e(X))D] = S_D = \mathbb{E}[e(X)\{1 - e(X)\}p(X)]$, which by Proposition 2 tends to 0 under Assumption 6. The identity of part (iii) is immediate. Since $\varrho = \theta_D / \mathbb{E}[\tilde{Z}D]$ by (14) and $S_D = \mathbb{E}[\tilde{Z}D] - \theta_D$, we have $S_D = (1 - \varrho)\mathbb{E}[\tilde{Z}D]$. The concentration-parameter statement follows because the corrected instrument $Z - e(X)$ has covariance S_D with D and variance $\mathbb{E}[\mathrm{Var}(Z \mid X)]$. □

Asymptotics of the bias-targeting statistic T_n^*

Let $\hat{g} = n^{-1}\sum_i \hat{h}_i(Y_i - \hat{\bar{L}}D_i)$ estimate $g(\bar{L}) = \mathbb{E}[h(X)(Y - \bar{L}D)]$, which equals $\theta_Y - \bar{L}\theta_D$ and is zero under H_0^* by (5). Expanding in $(\hat{\delta}, \hat{\bar{L}})$ about $(\delta, \bar{L})$, $\sqrt{n}\,\hat{g} = n^{-1/2}\sum_i\{h(X_i)(Y_i - \bar{L}D_i) - g(\bar{L})\} + \mathbb{E}[b(X)'(Y - \bar{L}D)]\,\sqrt{n}(\hat{\delta} - \delta) - \theta_D\,\sqrt{n}(\hat{\bar{L}} - \bar{L}) + o_p(1)$. We require, beyond Assumption 7 and S_D bounded away from zero, that $\hat{\bar{L}}$ be $\sqrt{n}$-asymptotically linear with an influence function ϕ_i, $\sqrt{n}(\hat{\bar{L}} - \bar{L}) = n^{-1/2}\sum_i \phi_i + o_p(1)$. This is a genuine additional condition. Proposition 8 delivers only consistency of $\hat{\bar{L}}$, so its asymptotic linearity must be supplied separately, for instance by the Neyman-orthogonalized, cross-fitted estimator of Proposition 5 under those rate conditions, which is why we invoke that construction rather than the plug-in when we need an interval. Given it, $\sqrt{n}(\hat{\delta} - \delta)$ and $\sqrt{n}(\hat{\bar{L}} - \bar{L})$ are both asymptotically linear, so $\sqrt{n}\,\hat{g} \xrightarrow{d} N(0, v)$ for a finite $v > 0$. In this fixed, strongly identified data-generating process the pairs bootstrap is first-order valid for the smooth functional $\hat{g}$ by the same Hadamard-differentiability argument as in Proposition 4, so the bootstrap variance $\hat{v}$ satisfies $n\hat{v} \xrightarrow{p} v$. Then $T_n^* = \hat{g}^2/\hat{v} = (\sqrt{n}\,\hat{g})^2/(n\hat{v}) \xrightarrow{d} \chi_1^2$ under H_0^*, and $\hat{g} \xrightarrow{p} g(\bar{L}) \neq 0$ gives consistency against any fixed alternative with nonzero bias. (Equivalently, one may write the statistic as $n\hat{g}^2/\hat{v}_1$ with $\hat{v}_1 \xrightarrow{p} v$ a consistent estimate of the asymptotic variance of $\sqrt{n}\,\hat{g}$, and the two forms coincide since $\hat{v}_1 = n\hat{v}$.) As $S_D \to 0$ the term $-\theta_D\sqrt{n}(\hat{\bar{L}} - \bar{L})$ ceases to be $O_p(1)$, $\hat{\bar{L}}$ is no longer $\sqrt{n}$-consistent, and both the Gaussian limit and the bootstrap approximation for T_n^* fail. Size control at that boundary is recovered not by T_n^* but by the signal-cleared statistic A_n of (11), whose validity along the collinearity sequence is established in Proposition 6 below, which is why the protocol reads A_n for size and reserves T_n^* for the extra power it buys under strong identification. □

Proof of Proposition 6

Write $m = (\theta_D, \theta_Y, S_D, S_Y)'$ and $\hat{m}$ for its plug-in from the flexible first-step regression, with $\psi = q(m) = \theta_Y S_D - \theta_D S_Y$ and gradient $\nabla\psi = (-S_Y, S_D, \theta_Y, -\theta_D)'$. Each coordinate of $\hat{m}$ is a two-step sample mean of the kind analysed in Proposition 4. For $\hat{\theta}_W$ the influence function is the ψ_i of part (i) there, and for $\hat{S}_W = n^{-1}\sum_i(Z_i - \hat{e}(X_i))W_i$ it is $(Z_i - e(X_i))W_i - S_W$ plus the same generated-regressor correction, so along $\{P_n\}$ the stacked vector obeys $\sqrt{n}\,(\hat{m} - m_n) = n^{-1/2}\sum_i \zeta_{i,n} + o_p(1)$ with $\Omega_n = \mathrm{Var}(\zeta_{i,n})$, and a Lyapunov condition holds under the uniform fourth moments of Assumption 7, exactly as in part (ii) of Proposition 4. A first-order expansion gives $\hat{\psi} - \psi_n = \nabla\psi_n'(\hat{m} - m_n) + R_n$, where the remainder R_n is a sum of products of two centered coordinates of $\hat{m} - m_n$. Since each coordinate is $O_p(n^{-1/2}\sigma_{\cdot,n})$ with $\sigma_{\cdot,n}$ the sd of the relevant influence, $\sqrt{n}\,R_n = O_p(n^{-1/2}\max_k \sigma_{k,n}^2) = o_p(\sigma_{\psi,n})$, where $\sigma_{\psi,n}^2 = \nabla\psi_n'\Omega_n\nabla\psi_n$ is the asymptotic variance of $\sqrt{n}\,\hat{\psi}$, so the remainder is negligible after self-normalization. Under H_0^*, $\psi_n = 0$, hence $\sqrt{n}\,\hat{\psi}/\sigma_{\psi,n} = \nabla\psi_n' n^{-1/2}\sum_i \zeta_{i,n}/\sigma_{\psi,n} + o_p(1) \xrightarrow{d} N(0,1)$ by the Lindeberg–Feller theorem for triangular arrays, provided $\sigma_{\psi,n}^2$ stays bounded away from zero after normalization, which holds because as $S_{D,n}, S_{Y,n} \to 0$ the gradient tends to $(0, 0, \theta_Y^*, -\theta_D^*)$ with $(\theta_D^*, \theta_Y^*) \neq 0$, so $\sigma_{\psi,n}^2 \to \mathrm{Var}(\theta_Y^*\zeta^{S_D} - \theta_D^*\zeta^{S_Y}) > 0$ under the nondegeneracy of the reduced-form scores. This last step uses that the statistic is self-normalized, so only the ratio $\sqrt{n}\,\hat{\psi}/\sigma_{\psi,n}$ matters and the common vanishing of ψ and its standard error cancels. With $\hat{v}_\psi$ ratio-consistent, $n\hat{v}_\psi/\sigma_{\psi,n}^2 \xrightarrow{p} 1$, so $A_n = \hat{\psi}^2/\hat{v}_\psi = (\sqrt{n}\,\hat{\psi}/\sigma_{\psi,n})^2\,(\sigma_{\psi,n}^2/n\hat{v}_\psi) \xrightarrow{d} \chi_1^2$, whether or not $S_{D,n} \to 0$. For power, at a fixed law with $b \neq 0$

and S_D bounded away from zero, $\psi = b\, S_D\, \mathbb{E}[\tilde{Z}D] \neq 0$ is fixed while $\hat{v}_\psi = O_p(n^{-1})$, so $A_n \to \infty$ and the power tends to one. Holding b fixed while $S_D \to 0$ sends $\psi = b\, S_D\, \mathbb{E}[\tilde{Z}D] \to 0$, so the noncentrality $\psi_n^2 / \operatorname{Var}(\hat{\psi})$ need no longer diverge and the power against that fixed b declines toward α, the Anderson–Rubin trade-off. □

Appendix B. Additional Monte Carlo evidence

This appendix collects the size-and-power, power-curve, collinearity-sweep, and share-estimation exhibits summarized in Section 6. The design is that of Section 6.

Table B.7: Directed contamination test. Panel A, one dataset, $n = 100\,000$. Panel B, rejection rates, 5% level, 2000 replications, $n = 3000$, cubic basis.

Panel A. Single contaminated dataset

$\hat{\theta}_D$	$\hat{\theta}_Y$	p-value	naive β_{2SLS}	cond. F	corrected $\hat{\bar{L}}$
0.067	0.800	$< 10^{-4}$	8.52	19 737	1.03

Panel B. Size and power

design	rejection rate (5%)	mean cond. F
SIZE (linear propensity, H_0 true)	0.019	569
POWER (curved propensity, H_0 false)	1.000	606

Table B.8: Clean design (linear propensity, no bias), Directed-X size against n at the 5% nominal level, 2000 replications, $B = 999$. "Dir-X" is the plug-in T_n^* and "AR" the signal-cleared A_n. Monte Carlo standard errors are near 0.005. Studentizing T_n^* with too few bootstrap draws ($B = 99$) inflates the clean-design rejection rate through a noisy variance estimate, and at $B = 999$ it is at nominal.

n	Dir-X size	AR size	mean cond. F
1000	0.058	0.058	192
3000	0.049	0.048	570
9000	0.044	0.045	1704
20 000	0.044	0.045	3783

The power curve sweeps the linear-probability curvature knob b for three sample sizes. The top axis of Figure B.5 maps b to the induced population bias $\beta_{2SLS} - \bar{L}$. At $b = 0$ the propensity is exactly linear and the test controls size, rejecting at 1.4%, 1.7% and 1.3% for $n = 1000, 3000, 9000$ against the 5% nominal level (the $b = 0$ row of Table B.9), so the pairs bootstrap is mildly conservative rather than exactly calibrated in this design. Power then rises monotonically in b and in n, while the mean conditional F stays pinned near 1740, so the same F accompanies a bias of 0 and a bias of 2.64.

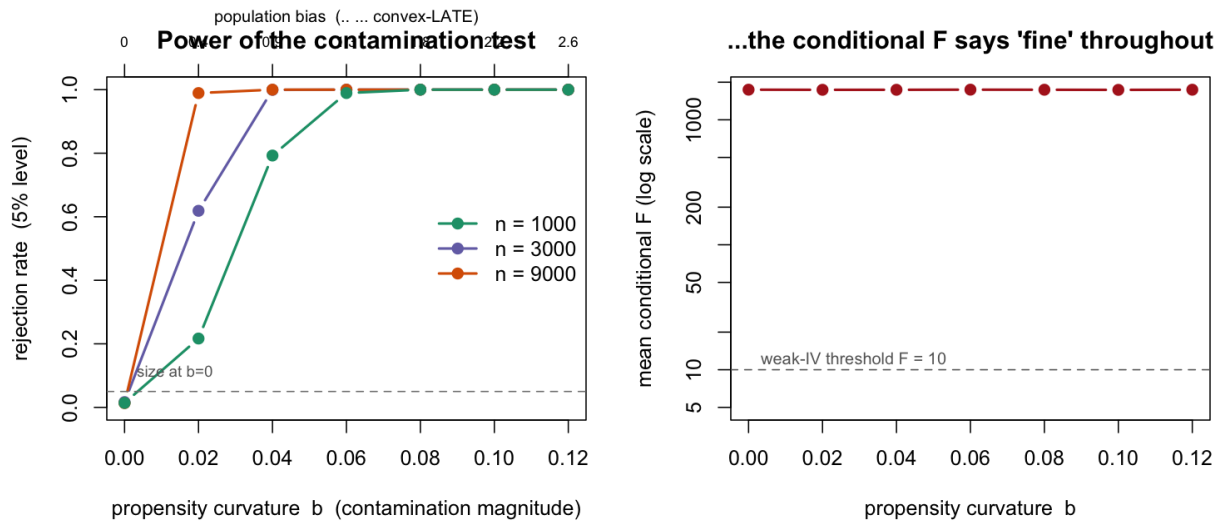

Figure B.5: Left panel. Rejection rate of the directed test vs. propensity curvature b (bottom axis) and induced population bias (top axis), for $n \in \{1000, 3000, 9000\}$, with the dashed line at the 5% level. Right panel. The mean conditional F (log scale) is invariant to b, sitting near 1740 throughout.

Table B.9: Power curve. Rejection rates (5% level, 1500 replications, $B = 299$) by curvature knob b and sample size. Induced population bias $\beta_{2SLS} - \bar{L}$ at $n = 200\,000$. Mean conditional F at $n = 9000$.

b	bias	$n = 1000$	$n = 3000$	$n = 9000$	mean cond. F
0.00	−0.02	0.014	0.017	0.013	1742
0.02	0.40	0.217	0.619	0.989	1738
0.04	0.87	0.793	0.999	1.000	1738
0.06	1.32	0.989	1.000	1.000	1743
0.08	1.80	1.000	1.000	1.000	1740
0.10	2.18	1.000	1.000	1.000	1736
0.12	2.64	1.000	1.000	1.000	1741

Table B.10: Directed tests as collinearity between instrument and covariates rises, in the contaminated design (rejection = power, the bias it faces in the third column) and the harmless design (rejection = size, 5% nominal). 2000 replications, $n = 3000$, cubic basis, $B = 999$. "Dir-X" is the plug-in T_n^* and "AR" the signal-cleared A_n. Directed-S raises a false alarm in every harmless row because it tests the sufficient condition $\theta = 0$. Both bias-targeting forms keep full power against genuine bias and hold size across the collinearity range, the size guarantee for A_n being that of Proposition 6.

		Contaminated (power)				Harmless (size)		
κ	$\mathrm{Var}(Z \mid X)$	bias	Dir-S	Dir-X	AR	Dir-S	Dir-X	AR
0.5	0.230	2.56	1.00	1.00	1.00	1.00	0.044	0.044
1.0	0.190	4.44	1.00	1.00	1.00	1.00	0.051	0.051
2.0	0.119	6.25	1.00	1.00	1.00	1.00	0.044	0.045
4.0	0.058	7.45	1.00	1.00	1.00	1.00	0.050	0.052
8.0	0.028	7.78	1.00	1.00	1.00	1.00	0.044	0.050
16.0	0.014	7.83	1.00	1.00	1.00	1.00	0.040	0.043

Table B.11: Robustness of the slave trade finding (Section 7.3) to the geography basis and the clustering level. The sample is that of Section 7.3, in 156 ethnic groups. Whatever the basis, the instrument has an unclustered first-stage F near 1258 and a cluster-robust F of 17.0, and the naive 2SLS estimate with geography entered linearly is -0.243 for trust in neighbours and -0.257 for trust in relatives. Each row refits the flexible propensity in the stated basis in latitude and longitude. Here $\hat{\varrho}$ is the contamination share with a 95% cluster bootstrap interval ($B = 999$), the corrected $\hat{\beta}$ the residualized propensity estimand, the corrected F its own cluster-robust first stage, and p_X the bias-targeting test of Section 4.3. The table documents the fragility of the reported strength to the geography basis, not a claim that the effect is spurious, and the corrected $\hat{\beta}$ stays negative throughout.

outcome	basis	$\hat{\varrho}$ [95%]	corrected $\hat{\beta}$	corrected F	p_X
trust in neighbours	quadratic	0.52 [0.08, 0.92]	−0.205	2.7	0.91
	cubic	0.52 [0.12, 0.96]	−0.190	2.4	0.98
	spline	0.52 [0.09, 1.03]	−0.200	3.0	0.98
trust in relatives	quadratic	0.51 [0.08, 0.92]	−0.137	2.7	0.89
	cubic	0.52 [0.11, 0.95]	−0.127	2.4	0.74
	spline	0.51 [0.10, 1.04]	−0.170	3.0	0.95

Table B.12: Contamination share. Population value, plug-in estimate at $n = 3000$, 95% percentile-bootstrap coverage of the population share (1000 replications, $B = 499$), and the induced 2SLS bias.

κ	ϱ (pop.)	$\hat{\varrho}$ (n=3000)	95% coverage	bias
0.25	0.021	0.022	0.949	1.306
0.50	0.082	0.080	0.953	2.549
1.00	0.239	0.240	0.949	4.365
2.00	0.494	0.497	0.950	6.340
4.00	0.682	0.683	0.933	7.539
8.00	0.735	0.741	0.942	7.659
16.0	0.752	0.754	0.940	7.814

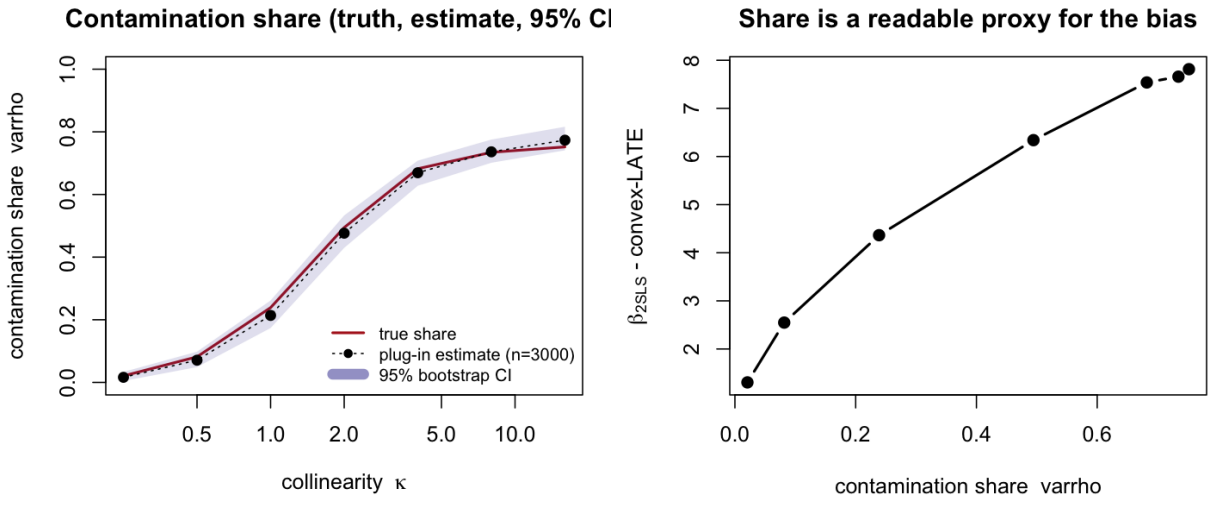

Figure B.6: Left panel. The population contamination share (solid), the plug-in estimate at $n = 3000$ (points) and its 95% bootstrap band, against the collinearity knob κ. Right panel. Along this one-parameter family the population bias is monotone in the share, so here $\hat{\varrho}$ orders these designs by danger. This ordering is a feature of the family swept, not a general property. Across unrelated designs bias also depends on θ_Y and $\bar{L}$, so equal shares can accompany different biases.