Robustness?
Range Tests for Equality and Equivalence Across Specifications

David A. Jaeger*
University of St Andrews, CEPEO, CReAM, CESifo, IZA@LISER, and CEPR

July 2026

Latest version:
`https://www.djaeger.org/research/robustness`

Abstract

Applied economists routinely compare estimates across specifications, observe that they are "similar," and conclude that their results are "robust." This common practice makes an implicit inferential claim about the range of estimates, but the estimates are usually generated from the same data and have a joint sampling distribution that published tables do not show. I formalize this claim with two bootstrap statistics. The minimum equivalence bound, $R^*_{1-\alpha}$, is the smallest tolerance within which the estimates can be judged equivalent. The range-based equality p-value, p_R, tests whether the estimates are statistically distinguishable. Together they separate two ideas that informal robustness checks often conflate: failure to detect differences and affirmative evidence of agreement. Because the bootstrap resamples the same observations and re-estimates all specifications in each replication, it captures dependence across estimates without requiring inversion of a potentially near-singular variance-covariance matrix. Simulations show approximately correct size and demonstrate that common visual rules often provide false comfort. Applications to five prominent papers validate some conventional robustness claims while revealing cases in which apparent agreement reflects imprecision rather than stability. In a pre-registered survey of CEPR and NBER affiliates, expert judgments align with the framework in obvious cases but diverge in intermediate cases, where the verdict depends on the joint sampling distribution rather than on the visible spread of point estimates alone. I suggest that $R^*_{.95}$ and p_R be reported whenever multiple specifications are presented as evidence of robustness.

Keywords: robustness; specification sensitivity; equivalence testing; bootstrap inference; joint inference; model uncertainty

JEL classification: C12; C14; C52

*Department of Economics, University of St Andrews. Email: david.jaeger@st-andrews.ac.uk. Claude (Anthropic) was used for brainstorming, parsing the literature, editing, and coding. I thank David Autor, Rajeev Dehejia, Luca Favero, Neil Lloyd, Matthew Notowidigdo, Steve Pischke, seminar participants at Aalto University and the University of St Andrews, and participants at the Second Annual Fife Applied Microeconomics and Golf and the 2026 European Society for Population Economics conferences for comments. I thank Yann Calvó López and McKenna LeClear at `refine.ink` for help with the survey incentive. All errors are the responsibility of the author.

I Introduction

Applied economists routinely present estimates from multiple specifications and declare the results "robust" when the coefficients look similar. Referees routinely ask for "robustness checks," requesting additional specifications that vary the controls, sample, estimator, or functional form. Most commonly, a researcher presents specifications that add demographics, fixed effects, lagged outcomes, or other controls and interprets coefficient stability as evidence that the result is not driven by specification choice. Without formal inference, however, casual inspection leaves the robustness claim speculative. The informal practice encodes a familiar statistical object, the range of estimates across specifications. This paper gives that range an inferential interpretation.

"Eyeball tests" can mislead in two ways. The first problem is *conceptual*. Researchers often compare coefficients as though they were repeated estimates of one parameter. When treatment effects are heterogeneous, however, different estimators typically target distinct population objects, and several estimates casually labeled "the treatment effect" may correspond to different estimands. Numerical similarity across those estimates may be descriptively interesting, but it is not automatically evidence about one parameter. The appropriate response is not to abandon such comparisons, but to be precise about what they establish regarding the population targets under comparison.

The second problem is *inferential*. Estimates generated from the same data have a joint sampling distribution, and similarity across different specifications may arise because the same information is being viewed through different lenses. Eyeballing five estimates and declaring them "basically the same" does not constitute a statistical test. Judging each estimate against its own standard error also misses the dependence among estimates, which is typically neither shown nor estimated. These covariances can be substantial (Pei et al., 2019). The question is whether the observed range of estimates is larger than their joint sampling variation alone would generate, and how much tolerance is required before the estimates can be called equivalent.

I address both problems with the bootstrap. Drawing the same resampled observations and estimating all specifications in the comparison set at each replication captures the joint distribution of the estimates, with no need to derive analytic cross-specification covariances (Efron, 1979; Davison and Hinkley, 1997). Because the bootstrap imposes no structure on how the specifications differ, the framework accommodates broad collections of estimates, whether they target one object or several distinct ones, and whether they employ the same or different estimators. The computation of the statistics is the same, but their interpretation depends on the specifications' targets. With a common probability limit, the statistics assess agreement among estimates of one object. With distinct probability limits, the minimum equivalence bound assesses the spread across the population objects themselves. The framework requires only that the design admit a valid joint bootstrap and that the estimates be on a common scale so that their range is substantively meaningful.

I use the bootstrapped range distribution to answer two distinct questions, depending on whether the bootstrap is recentered. Without recentering, the $(1-\alpha)$ quantile of the bootstrapped range distribution is the *minimum equivalence bound*, $R^*_{1-\alpha}$, the smallest tolerance at which the data support a claim of equivalence. It answers *how much disagreement must one*

tolerate to call the estimates in the comparison set robust? Recentering the bootstrap estimates on the observed full-sample estimates imposes the null of a common probability limit and gives the *range-based equality p*-value, p_R, the share of recentered bootstrap ranges at least as large as the observed range. It answers *do the observed differences exceed what shared sampling variation alone would produce?* The equivalence bound reverses the burden of proof imposed by the equality test. Because noisy estimates are hard to distinguish from one another, imprecision supports the claim that all estimates are equal. Under the equivalence bound, imprecision weakens a robustness claim, because noisy estimates require a larger tolerance before equivalence can be established.

When specifications share a probability limit, $R^*_{1-\alpha}$ bounds the disagreement among estimators of a single quantity, and a small value is evidence that the estimators concur. When specifications' probability limits are distinct, as under treatment-effect heterogeneity, $R^*_{1-\alpha}$ bounds the dispersion across those objects, and a small value is evidence that the objects themselves are close. The framework does not require every specification to estimate the same parameter. Instead, it gives the robustness claim a precise meaning in terms of the range of the population objects being compared.

The existing literature has built a rich apparatus for testing whether estimates are statistically distinguishable. The Durbin–Wu–Hausman test (Durbin, 1954; Wu, 1973; Hausman, 1978) tests equality of a coefficient pair under the requirement that one estimator be efficient under the null, so that the variance of the contrast reduces to the difference of the marginal variances. Extensions (Clogg et al., 1995; Schaffer, 2024) relax the efficiency requirement, and Lu and White (2014) generalize the comparison to non-nested models and heterogeneous estimators. These are tests of whether estimates can be distinguished, so imprecision supports a claim of robustness.

A second family of papers uses contrasts among estimates to assess robustness relative to a privileged specification. Altonji et al. (2005) and Oster (2019) use movements across nested OLS specifications to assess the importance of selection on unobservables, while Masten and Poirier (2026) show that the amount of omitted-variable selection required to reverse the sign of a coefficient can be substantially smaller than that required to explain it away. Gelbach (2016) instead decomposes movements across specifications into the contributions of individual covariates.

Recent work has increasingly characterized robustness through a minimum perturbation required to overturn a conclusion. Cinelli and Hazlett (2020) quantify the minimum strength of omitted confounding needed to alter inference. Closest in spirit to my paper is Prallon (2026), who quantifies the minimum tolerance around a designated main specification that is consistent with the observed robustness checks and shares the diagnosis that claims about the spread of estimates cannot be assessed without their joint sampling distribution. Prallon's robustness radius is the smallest tolerance b for which the data cannot reject the hypothesis that every check lies within b of the designated main specification. Like the minimum equivalence bound, it accounts for correlation across estimates rather than relying on marginal precision alone. My framework differs in its object and interpretation. The minimum equivalence bound treats specifications symmetrically and characterizes the full range across them, while the robustness

radius focuses on deviations from a designated baseline. The two approaches also treat imprecision differently, as imprecision can make deviations from a baseline harder to reject, whereas it increases the tolerance required to establish equivalence across the comparison set. I also report the equality p-value p_R alongside the bound, so that lack of detectable difference and affirmative similarity are not conflated. In practical terms, Prallon's robustness radius answers how far the checks can differ from the designated main estimate, whereas my statistics answer how wide the comparison set itself is and whether its members can be treated as equivalent within a substantively meaningful tolerance.

A separate tradition summarizes estimates from many specifications rather than testing them in pairs. Leamer (1983) argues that a conclusion drawn from a single specification is fragile when many defensible alternatives are available, and his extreme bounds analysis treats the range of estimates as the object of interest. Specification-curve and multiverse analyses (Simonsohn et al., 2020) display the distribution of coefficients generated by defensible specification choices and test the sharp null of no effect against a permutation distribution. Similarly, Card et al. (2023) take the narrow envelope produced by many combinations of controls as evidence that their estimate is stable. These methods describe how far the estimates range but leave the reader to judge whether the spread is small enough to call the result robust. I make that judgment formal by attaching a coverage statement to the spread through $R^*_{1-\alpha}$ and pairing it with an equality test based on the same bootstrap estimates. The equivalence bound adapts the logic of two one-sided tests from biostatistics (Schuirmann, 1987) to multiple estimates in the comparison set.[1]

I evaluate the framework in three ways. I first evaluate it in simulations under known data-generating processes with varying degrees of agreement among the probability limits. In the core regular designs, the equivalence bound $R^*_{.95}$ covers the true spread in probability limits at or above its nominal rate, and the range-based equality test has approximately correct size and substantial power. Informal criteria for assessing similarity, such as confidence interval overlap or all estimates being statistically significant and sharing the same sign, frequently provide false comfort, classifying specifications as robust when the formal tests do not support that conclusion.

I then re-analyze five papers that span major identification strategies in modern applied economics to illustrate how formal inference confirms, qualifies, or overturns published robustness claims. I find clear evidence that the results in Lee (2008) on the incumbency advantage in U.S. House elections are robust, with no rejection of equality and a tight minimum equivalence bound. I also find robust results in a randomized experiment (Karlan and List, 2007), while two difference-in-differences applications (Autor, 2003; Dube et al., 2010) reveal varying degrees of fragility.

The most revealing illustration applies the framework to the returns-to-education results from Angrist and Krueger (1991) and Bound et al. (1995) presented by Angrist and Pischke (2009). Angrist and Pischke present OLS, two-stage least squares (TSLS), and limited information maximum likelihood (LIML), and recommend that researchers who find TSLS and LIML to be similar "be happy" about the apparent lack of finite-sample bias in TSLS. In specifications

[1] See Lakens (2017) for a practical introduction to two one-sided tests.

with weak identification, there is little to celebrate. The seeming agreement between TSLS and LIML reflects shared imprecision rather than establishing equivalence. The equality test fails to reject that the estimates are the same, but the minimum equivalence bound cannot establish that they agree. The conventional reading confuses absence of evidence with evidence of absence, and the two statistics together make the distinction clear.

Lastly, I surveyed CEPR and NBER affiliates to gauge how leading empirical economists assess robustness and to compare their responses to the framework's conclusions. I asked them to rate the robustness of the published tables underlying the five illustrations above, on a five-point scale from not at all robust to extremely robust, with around 290 economists rating each table. Professional judgment and the framework agree when estimates are obviously similar or obviously different, but diverge in intermediate cases. Among respondents, 85 percent rated the estimates of the electoral effects of incumbency in Lee (2008) very or extremely robust, a conclusion supported by the framework. Yet 52 percent said the same of the returns-to-education estimates from Angrist and Pischke (2009), where the minimum equivalence bound reaches five times the average estimate. Respondents rank the tables as the framework does but compress the magnitudes, rating four of the six tables similarly even though the corresponding robustness ratios span nearly an order of magnitude.

As a reporting norm, I suggest that the minimum equivalence bound, $R^*_{.95}$, and the range-based equality p-value, p_R, be reported whenever multiple specifications are presented as evidence of robustness. A researcher who reports $R^*_{.95} = 0.02$ and $p_R = 0.87$ on coefficients ranging from 0.07 to 0.08 has made a precise robustness claim, with the data supporting equivalence within one quarter of the coefficients themselves, and p_R indicating that the null of equality is not rejected. A researcher who reports coefficients ranging from 0.07 to 0.10 and claims robustness when the same estimates generate $R^*_{.95} = 0.43$ and $p_R = 0.79$ has revealed that the word "robust" is doing a great deal of work. The high p_R reflects only that imprecise estimates cannot be distinguished from one another, while the minimum equivalence bound indicates disagreement of 0.43, five times the estimates themselves, must be accommodated to claim robustness. Both statistics can be computed from the same bootstrap estimates, and the companion R and Stata packages (Appendix D) produce them directly.

II What Does "Robust" Mean?

Define the *comparison set* as a set of K specifications, each an estimator applied to the same data, producing estimates $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)'$. Let θ_k denote the probability limit of $\hat{\theta}_k$, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$, so that

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}.$$

This paper disciplines claims about the range of $\boldsymbol{\theta}$,

$$\Delta \equiv \max_{1 \le k \le K} \theta_k - \min_{1 \le j \le K} \theta_j.$$

If all specifications in the comparison set are consistent and target the same estimand, $\Delta = 0$ and any disagreement among the elements of $\hat{\boldsymbol{\theta}}$ is due to sampling variation. Δ can be positive

for two reasons, however. Specifications may target genuinely different estimands and, under consistency of each estimator for its intended target, Δ is the spread of population objects. Alternatively, specifications may target the same estimand but differ in whether their underlying assumptions hold. When $\Delta > 0$ in this case, not all estimators are consistent for the common target, and Δ is the spread produced by the failed assumptions. A researcher who presents multiple specifications and declares the results "robust" is therefore making a claim about Δ.

To gauge variability across estimates, given the data and comparison set, I propose the *minimum equivalence bound* $R^*_{.95}$, defined as the smallest tolerance at which equivalence can be established at the 5 percent significance level. I also propose a range-based equality test of whether the data can detect any difference among the estimates. The p-value from the equality test, p_R, tells the reader whether the estimates can be statistically distinguished from one another. Reporting $R^*_{.95}$ tells the reader how much tolerance is needed for robustness, an affirmative claim of similarity rather than a simple failure to detect a difference. Together, the two statistics give an assessment of whether the results are robust to the specification choices under consideration.[2]

Consider the comparison set that appears most often in applied work, a set of K OLS specifications that vary the included covariates:

$$Y_i \;=\; \theta_k\, D_i + X_{ik}'\gamma_k + u_{ik}, \qquad k = 1, \ldots, K,$$

where θ_k is the effect of interest and X_{ik} are the covariates included in specification k, such as demographics, fixed effects, or lagged outcomes. A researcher interprets stability of the $\hat{\theta}_k$ across specifications as evidence that omitted variables or modeling choices do not substantively alter the result. The framework provides metrics for judging that stability, but their interpretation depends on the specifications' targets. If the covariates in X_{ik} are uncorrelated with D_i, so that including them does not change the estimand, as with pre-treatment covariates in a randomized experiment, the specifications target the same parameter. A small $R^*_{.95}$ then indicates tight agreement, and p_R is a direct test of equality. If the covariates in X_{ik} are correlated with D_i and are included to address omitted-variable bias, specifications with fewer covariates may be inconsistent for the desired target. In this case, a large $R^*_{.95}$ indicates that the substantive conclusion depends on which covariates are included. Finally, if some covariates in X_{ik} are themselves affected by the treatment, adding them changes the estimand. In that case, $R^*_{.95}$ bounds the spread across distinct probability limits, while p_R tests whether equality among those limits is detectably false at the current sample size.

When specifications have different probability limits, as they can when treatment effects are heterogeneous (Imbens and Angrist, 1994), the null of equality is false, and equality tests measure whether the differences are detectable at the current sample size. Because the test is consistent, the null will be rejected mechanically with sufficiently large samples regardless of whether the differences between the probability limits are substantively important, while non-rejection in small samples rewards imprecision with a claim of robustness. Neither rejection

[2]The framework is not a specification test in the sense of Vuong (1989), Cox (1961, 1962), or the encompassing tradition (Mizon and Richard, 1986; Hendry, 1995). The comparison set is taken as given, and the framework addresses how close the elements of $\boldsymbol{\theta}$ are to one another.

nor non-rejection of equality in this case addresses the spread in probability limits. Under heterogeneous effects, the relevant question is the magnitude of that spread, exactly what $R^*_{1-\alpha}$ bounds.

The statistics are conditional on the comparison set. Adding a specification to a comparison set weakly increases $R^*_{.95}$. More generally an $R^*_{.95}$ reported across twelve specifications is not directly comparable to an $R^*_{.95}$ reported across three. The larger-K claim is more stringent. Primary comparison sets should be determined by the substantive claim and the estimand structure before the robustness statistics are examined. Additional subsets may be informative, but should be labeled as exploratory. As in the applications below, every table reporting $R^*_{.95}$ should also report K, so that the scope of the robustness claim, not only its magnitude, is apparent.

III Equality and Equivalence Tests of Robustness

The key inferential challenge when comparing estimates from non-nested specifications is lack of knowledge about their joint distribution. The bootstrap addresses this problem with a great deal of generality and without requiring an analytic variance-covariance matrix. It accommodates combinations of regular estimators or functions of estimators as long as each can be computed on a resample and a common resampling scheme consistently reproduces their joint distribution. Resampling should match the original sampling design and preserve the dependence structure of the data. When observations are independent, this means resampling individuals with replacement. When the data are clustered, for example by state or classroom, the block bootstrap resamples entire clusters. In applications with more complex dependence, the resampling scheme should use the principal clustering dimension that can be applied jointly across the comparison set, and any remaining mismatch with the original design should be stated explicitly.[3] All K specifications are then re-estimated from the same B resamples, producing

$$\hat{\boldsymbol{\theta}}^{(b)} \equiv (\hat{\theta}_1^{(b)}, \ldots, \hat{\theta}_K^{(b)})', \qquad \text{with } b = 1, \ldots, B.$$

Test statistics are computed directly from the B vectors of bootstrap estimates.

III.A Bootstrap range tests

The sample range of the estimates,

$$R(\hat{\boldsymbol{\theta}}) \;\equiv\; \max_{1\le k\le K} \hat{\theta}_k \;-\; \min_{1\le j\le K} \hat{\theta}_j,$$

is the sample analogue of the population range Δ. Focusing on $R(\hat{\boldsymbol{\theta}})$ reflects applied practice for assessing robustness and subsumes the union of all pairwise comparisons.[4] The range is interpretable in the units of the coefficient and requires no covariance matrix or matrix inversion,

[3] The same principle applies to subgroup comparisons. Resampling is from the full sample and subgroups are formed within each replicate, incorporating group membership into the sampling variation.

[4] The use of the sample range as a test statistic dates to Tukey (1949), whose honestly significant difference test uses the studentized range of group means in the ANOVA setting.

which keeps it numerically stable even when K is large or some specifications are nearly collinear. The bootstrap range, $R(\hat{\boldsymbol{\theta}}^{(b)})$, is calculated in each bootstrap resample b:

$$R(\hat{\boldsymbol{\theta}}^{(b)}) \;\equiv\; \max_{1\le k\le K} \hat{\theta}_k^{(b)} \;-\; \min_{1\le j\le K} \hat{\theta}_j^{(b)},$$

with the B bootstrap ranges forming the unrecentered reference distribution used to construct the minimum equivalence bound. With unique extrema this distribution reproduces the first-order sampling behavior of the range. With ties, it yields a conservative bound but need not reproduce the sampling distribution of the sample range itself. I use the bootstrap estimates in two ways, generating $R^*_{1-\alpha}$ and p_R.

The *minimum equivalence bound* is the $(1-\alpha)$ quantile of the unrecentered bootstrap range distribution:

$$R^*_{1-\alpha} \;\equiv\; \inf\!\big\{d : B^{-1}\textstyle\sum_{b=1}^{B} \mathbf{1}\big(R(\hat{\boldsymbol{\theta}}^{(b)}) \le d\big) \ge 1-\alpha\big\} \;\in\; [0,\infty),$$

directly answering the question *how much disagreement must one tolerate to call the estimates in the comparison set robust?* Put differently, $R^*_{1-\alpha}$ is the smallest tolerance at which the data support a claim of equivalence at level α. This is the $\lceil(1-\alpha)B\rceil$-th order statistic of the B bootstrap ranges and is the reported statistic throughout. Appendix A distinguishes it from its ideal-bootstrap counterpart $R^{*\infty}_{1-\alpha}$, the exact conditional quantile, where the theory requires the distinction. Because $R^*_{1-\alpha}$ is in the metric of the coefficient itself, its magnitude should always be reported alongside the mean estimate $\bar{\hat{\theta}} \equiv K^{-1}\sum_{k=1}^{K}\hat{\theta}_k$. Under joint asymptotic normality and joint bootstrap consistency, it is a valid upper confidence bound on Δ, with coverage at the nominal level when the maximal and minimal probability limits are unique and above it otherwise. Pre-specifying a tolerance τ would define the null $H_0 : \Delta \ge \tau$ against $H_A : \Delta < \tau$, where rejection of the null is evidence of robustness. Reporting $R^*_{1-\alpha}$ without committing to a tolerance in advance is analogous to reporting a confidence interval rather than testing a single null, and gives the smallest τ at which the null $H_0 : \Delta \ge \tau$ would reject at level α.

The *range-based equality test* for $\boldsymbol{\theta}$ uses the same B bootstrap estimates, preserving their variances and covariances but imposing $\Delta = 0$ by recentering each bootstrap estimate on its full-sample counterpart. Define

$$\dot{\theta}_k^{(b)} \;\equiv\; \hat{\theta}_k^{(b)} - \hat{\theta}_k,$$

giving the recentered bootstrap range

$$R(\dot{\boldsymbol{\theta}}^{(b)}) \;\equiv\; \max_{1\le k\le K} \dot{\theta}_k^{(b)} \;-\; \min_{1\le j\le K} \dot{\theta}_j^{(b)},$$

and yielding the p-value

$$p_R \;\equiv\; \frac{1+\sum_{b=1}^{B}\mathbf{1}(R(\dot{\boldsymbol{\theta}}^{(b)}) \ge R(\hat{\boldsymbol{\theta}}))}{B+1} \;\in\; (0,1]$$

for $H_0 : \Delta = 0$ against $H_A : \Delta > 0$, following the add-one convention for Monte Carlo tests (Davison and Hinkley, 1997). Rejecting the null means the observed disagreement is larger than sampling noise around a common probability limit would generate. The range-based

equality test has asymptotic size at the nominal level under joint asymptotic normality and joint bootstrap consistency.[5]

III.B Asymptotic validity

The minimum equivalence bound and the range-based equality test have asymptotically nominal size and at-least-nominal coverage when the estimates are jointly asymptotically normal across the K specifications (Assumption 1) and the bootstrap appropriate to the sampling design consistently reproduces that joint distribution, including cross-specification covariances (Assumption 2). Appendix A contains proofs of the propositions below.

Proposition 1 (Size)**.** *When Assumptions 1 and 2 hold, under the null* $\boldsymbol{\theta} = \theta\mathbf{1}$*,*

$$P(p_R \leq \alpha) \;\to\; \alpha$$

at every level α *for which* $(B+1)\alpha$ *is an integer, a set that includes the conventional levels at the* $B = 9{,}999$ *used throughout.*

Recentering imposes $\Delta = 0$ while leaving the bootstrap variance intact, so under the null the recentered range and the sample range share a limiting distribution and p_R is asymptotically uniform. When every specification in the comparison set estimates the same object, the range-based equality test is asymptotically exact.

Proposition 2 (Coverage)**.** *Under Assumptions 1 and 2:*

(i) If all K *specifications share a probability limit, so that* $\Delta = 0$*, then* $\lim_{n\to\infty} P(\Delta \leq R^*_{1-\alpha}) = 1$*.*

(ii) If the largest and the smallest probability limits are each attained by exactly one specification, then $\lim_{n\to\infty} P(\Delta \leq R^*_{1-\alpha}) = 1-\alpha$ *at every level* α *for which* $(B+1)\alpha$ *is an integer.*

The coverage of $R^*_{1-\alpha}$ is exact when the largest and smallest probability limits are each attained by a single specification. When $\Delta = 0$, $R^*_{1-\alpha}$ errs only toward larger tolerances.

Validity of the framework is inherited from its constituent estimators. The conditions for the validity of $R^*_{1-\alpha}$ and p_R reduce to joint asymptotic normality of $\hat{\boldsymbol{\theta}}$ and bootstrap consistency for its joint distribution. Where bootstrap consistency is well established for each estimator under the resampling method used, such as the pairs bootstrap for i.i.d. data or the block bootstrap for clustered or time-series data, the joint bootstrap inherits that consistency under standard conditions. Where bootstrap consistency is known to fail, as for matching estimators under the pairs bootstrap (Abadie and Imbens, 2008), IV under weak identification, where asymptotic normality itself fails (Staiger and Stock, 1997), or parameters on the boundary of the parameter space (Andrews, 2000), the framework inherits those failures. The bootstrap does not fix problems that the underlying estimators do not resolve.

[5] Appendix B discusses a generalized Wald test as a precision-weighted alternative to the range-based equality test. I also show that the bootstrap-based equality test reproduces the analytical Hausman (1978) test in the standard fixed effects versus random effects setting.

When all specifications in a comparison set share a probability limit, $R^*_{.95}$ measures their sampling variation. When all specifications in the comparison set have distinct probability limits, $R^*_{.95}$ bounds the spread in those population objects. A subtle case arises when a comparison set mixes the two situations, combining specifications that share a probability limit with others that do not. In that case the bootstrap no longer reproduces the sampling distribution of the range (Fang and Santos, 2019), and $R^*_{.95}$ loses its asymptotic exactness but retains its coverage validity. Whatever the mix, the bound remains a conservative upper confidence bound on Δ, erring only toward larger tolerances.[6] Because an inflated tolerance works against a robustness claim, a mixed set only makes such claims harder, not easier.[7] The practical solution is to report uniform sets, either all same-probability-limit or all different-probability-limit. When a mixed set must be summarized, the maximum of the constituent pairwise $R^*_{.95}$ values is a tighter valid bound than the joint bound. To the extent that mixed comparison sets are a concern, the concern attaches to the interpretation of the minimum equivalence bound and not to the range-based equality test, which retains its asymptotic validity, because all comparison sets are homogeneous under the null.

Heterogeneous effects make the uniform-sets rule easier to satisfy. Specifications that would share a target under constant effects generally do not under heterogeneous effects. For example, different instruments recover differently weighted averages of the underlying effects, different sets of control variables reweight those effects, and different samples describe different populations. Heterogeneity therefore changes comparison sets with a common probability limit to ones with distinct probability limits. This changes the interpretation of the minimum equivalence bound to measuring spread across related objects rather than sampling variation around one object. The equality test remains valid across specifications with heterogeneous targets, indicating whether the differences are detectable in the given sample, and leaving the minimum equivalence bound as the fundamental measure of robustness.

III.C Interpreting the two statistics

In some contexts a researcher can pre-specify a tolerance, $\tau > 0$, for the largest value of Δ considered consistent with robustness. When τ is specified, the equivalence test evaluates whether the data can reject the null that Δ exceeds this tolerance. Specifying τ permits calculation of a p-value for the equivalence test, based on the share of unrecentered bootstrap ranges that exceed τ. Rejecting the null establishes robustness within the stated tolerance.

When a tolerance, τ, and a level, α, are specified, the equivalence and equality hypotheses regarding Δ,

$$H_0 : \Delta \geq \tau \text{ versus } H_A : \Delta < \tau$$

and

$$H_0 : \Delta = 0 \text{ versus } H_A : \Delta > 0,$$

[6] See Corollary 1 in Appendix A.

[7] As I show in Appendix A, a mixed set costs sharpness (through overcoverage) and meaning (because one statistic is asked to measure both variability in the homogeneous set and distance between that set and the non-homogeneous specification). For example, a comparison set that pools instrumental-variables estimates sharing a target under homogeneous effects with an OLS estimate converging to a different population object yields a single number that conflates estimator agreement with estimand spread.

define a taxonomy for claiming robustness about the comparison set:

	Equality	
Equivalence	$p_R > \alpha$	$p_R \leq \alpha$
$R^*_{1-\alpha} \leq \tau$	**Robust:** the estimates are equivalent within tolerance and statistically indistinguishable.	**Robust:** differences are detectable but within tolerance. The differences are real but substantively small.
$R^*_{1-\alpha} > \tau$	**Robustness not established:** the data cannot distinguish the estimates, but neither do they support equivalence within tolerance.	**Not robust:** differences are detectable, but the data cannot establish that they are within tolerance.

This characterization makes clear that $R^*_{1-\alpha}$ is the fundamental determinant of whether a set of estimates is robust. While any comparison with $\Delta > 0$ will eventually move to the $p_R \leq \alpha$ column as n grows, $R^*_{1-\alpha}$ converges to Δ itself. *The minimum equivalence bound is therefore the more important and the more durable of the two statistics.*

The choice of τ is not arbitrary, but should be grounded in theory, policy, or the previous literature. A natural anchor is what the maintained assumptions imply about Δ, with τ set strictly above the implied value or bound so that the test has power to affirm the maintained case. Under constant effects, the assumptions imply $\Delta = 0$ and any defensible positive tolerance serves. With heterogeneous treatment effects, a sensible choice for τ is an upper bound on the range of the population LATEs targeted by the estimators (Imbens and Angrist, 1994). Referees should shift their focus from requesting "robustness checks" to asking whether τ is sensible given the context of the research. The second-best alternative to pre-specifying τ is to judge the magnitude of $R^*_{.95}$ relative to the criteria one would use in pre-specifying τ, the same exercise but *ex post* rather than *ex ante*.

In the absence of a natural choice for τ, I also offer a heuristic based on the *robustness ratio*, $R^*_{.95}/|\bar{\hat{\theta}}|$.[8] As a rule of thumb, a robustness ratio below one-half suggests the estimates are robust, with a bound less than half the average estimate. A ratio between one-half and one suggests the results are weakly robust. A ratio above one means that the minimum equivalence bound exceeds the average estimate itself, so that claiming robustness requires accepting a tolerance at least as large as the effect being estimated. This heuristic serves when a researcher is unwilling or unable to specify τ. $R^*_{.95}$ itself, in the units of the coefficient, remains the primary object, and the ratio is uninformative when $\bar{\hat{\theta}}$ is near zero, where the bound should be judged directly.

III.D Reporting norms and researcher discretion

Whenever multiple specifications are presented as evidence of robustness, I suggest reporting a standard set of statistics to increase the transparency of the resulting claims. The minimum

[8] I offer this heuristic in the spirit of the first-stage F rule-of-thumb value of 2 in Bound et al. (1995), later modified to 10 by Staiger and Stock (1997) and subsequently formalized by Stock and Yogo (2005).

equivalence bound, $R^*_{.95}$, and the range-based equality p-value, p_R, should be reported together, alongside the number of specifications K, the mean estimate $\bar{\hat{\theta}}$, and the observed range $R(\hat{\boldsymbol{\theta}})$, the quantities every robustness table in Section V displays. $R^*_{.95}$ and p_R answer different questions and both are informative about what robustness conclusions the data can support, although as noted above, $R^*_{.95}$ is the more important of the two. The R and Stata packages accompanying this paper also report $R^*_{.50}$, the median of the unrecentered bootstrap range distribution, as a matter of course. A large gap between $R^*_{.50}$ and $R^*_{.95}$ signals a heavy right tail in the range distribution. The packages also report the robustness ratio for each comparison as a rough guide. When a researcher pre-specifies an equivalence tolerance τ, the equivalence p-value at that tolerance should also be reported. All of these statistics are generated from the same set of bootstrap estimates.

A natural objection is that a researcher could choose specifications in the comparison set to obtain a small $R^*_{.95}$, reporting only those that agree. This discretion already exists, however, and the framework makes it visible rather than creating it. In current practice, researchers choose specifications and assess their agreement by eye with no statistics at all. Reporting the standard set of statistics makes robustness claims visible and verifiable. $R^*_{.95}$ quantifies agreement or disagreement across the specifications, but it is the researcher, and not the framework, who chooses what specifications to report. The framework's statistics are conditional on the comparison set in the way that any reported estimate is conditional on the specification that produced it.

Most importantly, selecting a comparison set that produces a small minimum equivalence bound narrows what can be claimed. A small $R^*_{.95}$ across nearly identical specifications supports only a narrow robustness claim. Eyeballing has the opposite property. A set of specifications chosen to yield similar point estimates gives the impression of robustness even when the joint sampling distribution would support a different conclusion. Indeed, the same set of estimates and standard errors could lead to either a claim of robustness or a claim of fragility, depending on the joint sampling distribution. The minimum equivalence bound makes such claims precise.

The comparison set should follow from the claim being evaluated. If the claim concerns all of the reported specifications, the range should be calculated over the full set. A claim that a preferred baseline survives a series of alternative specifications can be evaluated using baseline-anchored pairs or prespecified groups containing the baseline and related alternatives, as in Table VII, Part B. The two comparison sets answer different questions, and the relevant set should be specified before the statistics are computed rather than chosen afterward.

The comparison set is a natural object to pre-specify, even in observational studies, as a way to limit researcher discretion. A pre-analysis plan or registered report can commit the researcher in advance to the K specifications over which $R^*_{.95}$ will be computed, and, ideally, to τ for the equivalence claim. The framework's contribution is inference on the robustness claim being made about a given comparison set. It is not a panacea for researcher discretion.

IV Simulations

I first evaluate the framework using six simulation designs with known data-generating processes (DGPs). Each design defines K specifications, all estimating the coefficient on a treatment variable D_i. I evaluate each design over 10,000 Monte Carlo replications, with $n = 2{,}000$ observations and $B = 9{,}999$ bootstrap resamples per replication. The designs vary the source of disagreement across specifications. Designs 1 and 6 have a common estimand with all estimators consistent. Designs 2, 3, and 4 have estimators whose probability limits differ through an omitted covariate that both shifts the outcome and interacts with the treatment (Design 2), a bad control that creates a single outlier (Design 3), and small omitted-variable biases in alternating directions (Design 4). Design 5 has genuinely distinct estimands. Full DGP details are in Appendix C and in briefer form in the notes to Table I. The exercise evaluates whether the range-based equality test has the correct size and how the median unrecentered bound $R^*_{.50}$ compares to the spread in probability limits implied by each DGP. As a benchmark, I also calculate the generalized Wald equality statistic of Appendix B and its p-value, p_W.

In Table I, I present the share of Monte Carlo replications where $p_R \leq 0.05$ or $p_W \leq 0.05$ (as a size check in Designs 1 and 6 and a power check in Designs 2–5), the Monte Carlo median of $R^*_{.50}$, which tracks the population range from above, the Monte Carlo median of $R^*_{.95}$ as recommended above, and the empirical coverage of $R^*_{.95}$ as an upper confidence bound for Δ. The results establish that the range-based equality test has approximately correct size. Under the null of equal probability limits (Designs 1 and 6), the range test rejects at 4.4 and 3.7 percent, compared with 3.3 and 2.4 percent for the Wald test. Under the alternatives in Designs 2–5, both tests have high power, with rejection rates between 97.7 and 100 percent. Their modest differences reflect the alternatives against which each statistic is naturally powerful. The Wald test performs slightly better against the dense deviations in Design 4, while the range test performs better when a single specification is an outlier in Design 3. The range statistic loads entirely on the extreme coordinates, while the quadratic form spreads weight across all K contrasts.

To gauge the contribution of sampling variation to $R^*_{.50}$, I construct a version of each design that holds K, the sample size, and the covariance structure fixed but sets the coefficients so that $\Delta = 0$. I record the Monte Carlo median of $R^*_{.50}$ under each matched-null design, reported as ν_n in the last column of Table I. For the two designs in which the null is already imposed, ν_n reproduces the reported $R^*_{.50}$ up to Monte Carlo error. The statistic is larger when one of the specifications is intrinsically imprecise, as with the mediator specification in Design 3. It is therefore best understood as a benchmark for the scale of the sampling component, not as a constant that can be subtracted from $R^*_{.50}$ to recover Δ. Across Designs 2 through 4, the excess of the median $R^*_{.50}$ over Δ is no greater than ν_n. This is consistent with the matched-null statistic capturing the scale of the finite-sample contribution to the bound.[9] In Design 5, $\Delta = 0.400$, the same as the median $R^*_{.50}$ to three decimals.

The "Coverage" column of Table I reports the empirical coverage of $R^*_{.95}$ as an upper confidence bound for Δ. Coverage meets or exceeds the nominal 0.95 level in every design, with

[9]The gap between Δ and $R^*_{.50}$ closes at rate $n^{-1/2}$. Increasing the sample to $n = 10{,}000$ in Design 4 lowers the median $R^*_{.50}$ from 0.042 to 0.034, closing toward $\Delta = 0.028$.

TABLE I
Simulation Results: Equality and Equivalence Tests Across Six Designs

Design	K	Equality (rej. rate) p_R	Equality (rej. rate) p_W	Implied Δ	Equivalence (median R^*) $R^*_{.50}$	Equivalence (median R^*) $R^*_{.95}$	Coverage of $R^*_{.95}$	Matched null ν_n
D1: Same estimand	5	0.044	0.033	0	0.014	0.027	1.000	0.014
D2: Similar plims	5	1.000	1.000	0.120	0.123	0.144	0.974	0.013
D3: Single outlier	5	1.000	0.985	0.240	0.244	0.317	0.956	0.040
D4: Diffuse disagreement	6	0.977	0.994	0.028	0.042	0.055	1.000	0.021
D5: Distinctly different	5	1.000	1.000	0.400	0.400	0.402	0.998	0.002
D6: Near-singular	10	0.037	0.024	0	0.003	0.004	1.000	0.003

Notes: Results are based on 10,000 Monte Carlo replications, each with $n = 2{,}000$ observations and $B = 9{,}999$ bootstrap resamples. The equality columns report 5% rejection rates, measuring size under equal probability limits in Designs 1 and 6 and power under unequal probability limits in Designs 2–5. "Implied Δ" is the range of the probability limits generated by each DGP. $R^*_{.50}$ and $R^*_{.95}$ are the Monte Carlo medians of the corresponding unrecentered bootstrap range quantiles, and coverage is the share of replications in which $\Delta \leq R^*_{.95}$. ν_n is the Monte Carlo median of $R^*_{.50}$ in a matched-null version of each design that holds K, n, and the covariance structure fixed while setting all probability limits equal. It benchmarks the contribution of sampling variation to the bound.

DGPs:

- D1: $Y_i = \beta D_i + X_i'\gamma + u_i$, $D_i \perp X_i$, with $K = 5$ OLS specifications varying the control set and all targeting $\beta = 0.5$.
- D2: $Y_i = (0.5 + 0.15X_{i1})D_i + X_i'\gamma + u_i$, $D_i = 0.5X_{i1} + \nu_i$, with $K = 5$ OLS specifications. Omitting X_{i1} produces a probability-limit range of approximately 0.12 through omitted-variable bias.
- D3: $Y_i = \beta D_i + 0.4M_i + 0.2X_{i1} + u_i$, $\beta = 0.3$, with mediator $M_i = 0.6D_i + \eta_i$, $D_i \perp X_i$, and $K = 5$. Four specifications estimate the total effect, 0.54, while one includes M_i and estimates the direct effect, 0.30.
- D4: $Y_i = \beta D_i + X_i'\gamma + u_i$, $D_i = 0.3\sum_j X_{ij} + \nu_i$, $\beta = 0.5$, $\gamma_j = 0.05(-1)^j$, and $K = 6$. Each specification omits one X_{ij}, producing small alternating-sign biases.
- D5: Five outcomes satisfy $Y_{ik} = \beta_k D_i + 0.2X_{ik} + u_i$, with $\beta \in \{0.3, 0.5, 0.7, 0.4, 0.6\}$, so the estimands differ.
- D6: D1 with $K = 10$ specifications, each using $\tilde{X}_{ik} = X_{i1} + 0.1\epsilon_{ik}$. Near-identical controls produce a near singular matrix.

coverage of 1.000 by construction in the two designs with $\Delta = 0$. Coverage is nearest the nominal level, at 0.956, in Design 3, in which one probability limit is shared by four specifications and one specification is an outlier. Designs 2 and 3, as well as the positive-Δ runs of Design 7, include ties at an extremum among estimators with distinct influence functions, the configuration covered by Corollary 1; their coverage rates provide finite-sample corroboration of that result. In the other designs, coverage is above the nominal 0.95, reflecting the framework's conservatism (see Appendix A). Coverage is also stable as the sample grows, as increasing n to 10,000 in Design 4 leaves coverage at 1.000 while $R^*_{.95}$ tightens toward Δ.

Using the same designs and bootstrap samples as in Table I, in Table II I compare the formal range test with two informal criteria for robustness: whether all pairwise confidence intervals overlap, and whether all estimates have the same sign and are individually significant at the 5 percent level. Panels A and C report results for Designs 1 and 6, in which the probability limits are equal, while Panels B and D report results for Designs 2–5, in which the probability limits are unequal. In the latter panels, the $p_R \leq 0.05$ column reports the "false comfort" rate, the share of replications in which an informal criterion reports robustness despite statistically detectable

disagreement.[10] The same-sign-and-significant criterion reports robustness in every replication of every design, so its false comfort rate is between 97.7 and 100 percent in Designs 2–5. Given the effect sizes and sample sizes in these designs, that criterion is relatively easy to satisfy; the comparison remains informative because it still fails to distinguish equality from substantively small disagreement. Confidence-interval overlap performs better in Designs 2, 3, and 5 but fails nearly completely in Design 4, where the intervals overlap in every replication and detectable disagreement goes unreported in 97.7 percent. The informal criteria generally agree with the formal test only in Designs 1 and 6, where $\Delta = 0$. They also cannot distinguish detectable disagreement that is small relative to the effect from disagreement that is nearly as large as the effect. Design 4 rejects equality but has a robustness ratio of 0.11, while Designs 3 and 5 have ratios of 0.64 and 0.80. The same-sign-and-significant criterion reports robustness in all three cases, while confidence-interval overlap does so only in Design 4.

TABLE II
Simulation Results: Eyeballing vs. Formal Range Test

	Informal criterion judges robust		
Design	Judged robust	$p_R > 0.05$	$p_R \leq 0.05$
Panel A: Confidence-interval overlap, equal probability limits			
D1: Same estimand	1.000	0.956	0.044
D6: Near-singular	1.000	0.963	0.037
Panel B: Confidence-interval overlap, unequal probability limits			
D2: Similar plims	0.002	0.000	0.002
D3: Single outlier	0.014	0.000	0.014
D4: Diffuse disagreement	1.000	0.023	0.977
D5: Distinctly different	0.000	0.000	0.000
Panel C: Same sign and all significant, equal probability limits			
D1: Same estimand	1.000	0.956	0.044
D6: Near-singular	1.000	0.963	0.037
Panel D: Same sign and all significant, unequal probability limits			
D2: Similar plims	1.000	0.000	1.000
D3: Single outlier	1.000	0.000	1.000
D4: Diffuse disagreement	1.000	0.023	0.977
D5: Distinctly different	1.000	0.000	1.000

Notes: The Monte Carlo replications and DGPs are the same as in Table I. Panels A and B report the confidence-interval-overlap criterion for designs with equal and unequal probability limits, respectively. Panels C and D report the same-sign-and-all-significant criterion for designs with equal and unequal probability limits, respectively. "Confidence-interval overlap" classifies a set as robust when all pairwise 95% confidence intervals overlap. "Same sign and all significant" requires all estimates to have the same sign and to be individually significant at the 5% level. "Judged robust" is the share of replications classified as robust by the informal criterion. The next two columns divide that share according to whether the range test fails to reject or rejects $H_0 : \Delta = 0$, and therefore sum to "Judged robust." In Panels A and C, the $p_R \leq 0.05$ column measures formal-test size. In Panels B and D, it measures false comfort: the informal criterion judges the estimates robust even though the range test detects disagreement. The reverse case, in which the informal criterion rejects robustness while $p_R > 0.05$, does not occur in any of the 10,000 replications.

[10] Individual standard errors for the eyeballing criteria are computed from the diagonal of $\widehat{V}^*$, not from within-resample variance estimation. The bootstrap computations store only point estimates on each resample, although one could instead calculate standard errors within each bootstrap replicate and base the informal conclusions on those estimates.

I next consider a modification of Design 1 that adds a group-level error component to the data-generating process. Each group contains 50 observations, and the disturbance has intragroup correlation 0.3. I vary the number of groups, $G \in \{10, 15, 20, 30, 50, 100\}$, and resample clusters using either the block bootstrap or the wild cluster bootstrap (Cameron et al., 2008). This design is directly relevant to the Autor (2003) and Dube et al. (2010) illustrations below, where inference relies on state-level clustering with 50 or fewer states. Table III shows that the block-bootstrap range test remains close to nominal even with few clusters, while the Wald test substantially over-rejects. With $G = 10$, the rejection rates are 6.1 percent for the range test and 18.0 percent for the Wald test. The wild cluster bootstrap does not improve either test and is particularly poor for the Wald statistic. At $G = 50$, the setting most relevant to the applications, the block-bootstrap range test rejects in 5.3 percent of replications, compared with 7.0 percent for the Wald test. These results support using the range as the primary test statistic in the applications below.

The rejection-rate columns evaluate size, so I also vary Δ within the same design to evaluate coverage of the $R^*_{.95}$ bound. Adding an omitted-variable channel so that the specification omitting all covariates has probability limit 0.65 while the rest remain at 0.5 gives $\Delta = 0.15$ with the cluster structure unchanged. The equivalence bound covers at or above its nominal level at every cluster count, with coverage of $R^*_{.95}$ rising from 0.962 at $G = 10$ to 0.978 at $G = 100$, and the median $R^*_{.50}$ declines toward Δ as the number of clusters grows. The equality test rejects this Δ in 99.3 percent of replications even at $G = 10$.

TABLE III
Design 7: Few Clusters
Size at $\Delta = 0$ and Coverage at $\Delta = 0.15$

	$\Delta = 0$ (rej. rate)				$\Delta = 0.15$, block bootstrap			
	Block		Wild cluster		(rej. rate)	(median R^*)		Coverage
G	p_R	p_W	p_R	p_W	p_R	$R^*_{.50}$	$R^*_{.95}$	of $R^*_{.95}$
10	0.061	0.180	0.091	0.373	0.993	0.161	0.215	0.962
15	0.055	0.118	0.076	0.211	0.999	0.159	0.205	0.969
20	0.055	0.104	0.070	0.169	1.000	0.158	0.198	0.968
30	0.055	0.080	0.062	0.115	1.000	0.157	0.190	0.973
50	0.053	0.070	0.059	0.085	1.000	0.155	0.181	0.974
100	0.050	0.051	0.053	0.061	1.000	0.154	0.172	0.978

Notes: The DGP extends Design 1 to G clusters of $n_g = 50$ observations with an intracluster correlation of 0.3. In the $\Delta = 0$ columns, all $K = 5$ specifications have probability limit 0.5, and entries are rejection rates for $H_0 : \Delta = 0$ at the 5% level. In the $\Delta = 0.15$ columns, the specification omitting all covariates has probability limit 0.65, while the other four have probability limit 0.5. The p_R column reports rejection rates for $H_0 : \Delta = 0$, the $R^*_{.50}$ and $R^*_{.95}$ columns report Monte Carlo medians, and coverage is the share of replications in which $R^*_{.95} \geq 0.15$. Results are based on 10,000 Monte Carlo replications with $B = 9{,}999$ bootstrap replications each. "Block" resamples clusters with replacement, while "Wild cluster" uses the procedure of Cameron et al. (2008). The $\Delta = 0.15$ results use the block bootstrap.

In the final simulation, I examine how the framework performs under weak identification using a design with homogeneous treatment effects that mirrors the re-evaluation of results from Angrist and Pischke (2009). I estimate the effect of a treatment D_i by OLS, TSLS, and LIML, with one endogenous regressor, two excluded instruments, Z_i and $Z_i^2 - 1$, and $\text{Corr}(u_i, v_i) = 0.5$.

I vary identification strength through the concentration parameter $n\pi^2$, with targets from 1 to 500.[11] Both excluded instruments enter the estimated first stage, so the model is overidentified and TSLS and LIML differ in finite samples. Because the population first-stage coefficient on $Z_i^2 - 1$ is zero, the realized joint first-stage F statistic averages roughly $1 + n\pi^2/2$.

Table IV reports results for the three pairwise comparison sets rather than a single mixed one. The TSLS–LIML pair is a within-estimand comparison, with both estimators targeting the same structural parameter at every instrument strength. The TSLS–LIML comparison never rejects equality, even at $\mu^2 = 500$, but the equivalence bounds demonstrate why the non-rejections should not be interpreted as agreement under weak identification. At $\mu^2 = 1$, $R^*_{.95} = 6.604$, and even at $\mu^2 = 10$, $R^*_{.95} = 0.587$, both bounds exceed the true coefficient of 0.5. Only once $\mu^2 \geq 100$ does the TSLS–LIML bound become tight, with $R^*_{.95} = 0.030$ at $\mu^2 = 100$. By contrast, OLS has a different probability limit because of endogeneity, so the OLS–TSLS and OLS–LIML pairs are across-estimand power checks, bootstrap analogues of the Durbin–Wu–Hausman test. When identification is weak, the IV estimators are too imprecise to reject equality with OLS, although $R^*_{.95}$ is large, distinguishing "no evidence of a difference" from "evidence of no difference." Coverage rates are 1.000 for the TSLS–LIML comparison at every instrument strength, as they must be when $\Delta = 0$. For the OLS–IV pairs, coverage is close to nominal once identification is strong enough for Assumption 1 to be a reasonable approximation. The mild dips at intermediate strength, such as 0.932 for OLS–TSLS at $\mu^2 = 20$, reflect finite-sample non-normality of the IV estimates rather than a failure of the range.[12]

The simulations establish that the framework performs as the theory promises. The range-based equality test holds its size in the regular null designs, has power comparable to the generalized Wald benchmark where probability limits differ, remains stable with few clusters, and reveals the consequences of weak identification. The minimum equivalence bound covers the population range at or above its nominal level in every core design, dips no more than two percentage points below nominal in the weak-instrument comparisons, and its median tracks the population range from above. The informal criteria that current practice relies on cannot make these distinctions. The simulations do not show, however, how the framework operates when multiple specifications are used to answer substantive research questions. I take up this question next.

V Empirical Illustrations

To show how the framework functions in practice, I replicate five well-known papers that cover several of the most common identification strategies in applied economics: sharp regression discontinuity (Lee, 2008), randomized field experiments (Karlan and List, 2007), staggered-adoption difference-in-differences (Autor, 2003), spatial difference-in-differences (Dube et al., 2010), and instrumental variables (Angrist and Pischke, 2009). These papers were chosen be-

[11] The concentration parameter, the noncentrality of the joint first-stage F statistic, is the standard scalar measure of identification strength (Staiger and Stock, 1997; Stock and Yogo, 2005). With L excluded instruments, $E[F] \approx 1 + \mu^2/L$, and under this design's unit variances $\mu^2 = n\pi^2$ in expectation.

[12] Additional results for jackknife (Angrist et al., 1999; Kolesár, 2013) and split-sample (Angrist and Krueger, 1995) instrumental-variables estimators and for subsampling appear in Supplementary Tables S1–S3 in the replication package.

TABLE IV
Design 8: Weak Instruments

		Rejection rate (5%)		R^* (med.)	
μ^2	Comparison set	Range	Wald	$R^*_{.95}$	Cov. $R^*_{.95}$
1	TSLS, LIML	0.000	0.000	6.604	1.000
	OLS, TSLS	0.000	0.000	2.142	0.999
	OLS, LIML	0.000	0.000	8.219	1.000
5	TSLS, LIML	0.000	0.000	1.991	1.000
	OLS, TSLS	0.019	0.019	1.393	0.982
	OLS, LIML	0.003	0.003	3.177	0.995
10	TSLS, LIML	0.000	0.000	0.587	1.000
	OLS, TSLS	0.111	0.111	1.074	0.957
	OLS, LIML	0.031	0.031	1.594	0.981
20	TSLS, LIML	0.000	0.000	0.193	1.000
	OLS, TSLS	0.488	0.488	0.881	0.932
	OLS, LIML	0.311	0.311	1.025	0.960
50	TSLS, LIML	0.000	0.000	0.064	1.000
	OLS, TSLS	0.966	0.966	0.729	0.934
	OLS, LIML	0.956	0.956	0.767	0.953
100	TSLS, LIML	0.000	0.000	0.030	1.000
	OLS, TSLS	1.000	1.000	0.641	0.939
	OLS, LIML	1.000	1.000	0.657	0.948
500	TSLS, LIML	0.000	0.000	0.006	1.000
	OLS, TSLS	1.000	1.000	0.472	0.944
	OLS, LIML	1.000	1.000	0.474	0.949

Notes: The DGP is $Y_i = \beta D_i + u_i$ and $D_i = \pi Z_i + v_i$, with $\beta = 0.5$, $\text{Corr}(u_i, v_i) = 0.5$, and $n = 2{,}000$. The excluded instruments are Z_i and $Z_i^2 - 1$. The target concentration parameter is $\mu^2 = n\pi^2$, with $\pi \in \{0.022, 0.050, 0.071, 0.100, 0.158, 0.224, 0.500\}$ for $\mu^2 \in \{1, 5, 10, 20, 50, 100, 500\}$. With two excluded instruments, the realized joint first-stage F statistic averages approximately $1 + \mu^2/2$. TSLS and LIML share the same probability limit, while OLS differs because of endogeneity. Results are based on 10,000 Monte Carlo replications with $B = 9{,}999$ bootstrap resamples each. Rejection rates are for equality at the 5% level. The R^* column reports the Monte Carlo median of $R^*_{.95}$, and "Cov. $R^*_{.95}$" reports its empirical coverage of the population range Δ. With $K = 2$, the range and Wald equality tests coincide, so their rejection rates are identical by construction.

cause they have publicly available replication packages and include tables in which the authors present multiple specifications about which they make robustness claims. In each case, I follow the specification choices made by the original authors, testing the robustness of the full set of reported specifications as well as subsets motivated by how the authors present and discuss their results. These are the first five papers I replicated and for which I calculated the robustness statistics. They were not selected on the basis of those statistics.

I show the replicated point estimates and standard errors for each application. For each comparison set, I report K (the number of specifications in the set), $\bar{\hat{\theta}}$ (the average of the full-sample estimates in the comparison set), $R(\hat{\boldsymbol{\theta}})$ (the observed range of the full-sample estimates in the comparison set), $R^*_{.50}$ (the median of the unrecentered bootstrap range distribution), $R^*_{.95}$ (the minimum equivalence bound), p_R (the p-value for the bootstrap range-based equality test), and the robustness ratio $R^*_{.95}/|\bar{\hat{\theta}}|$. Where a sensible threshold τ can be determined from the literature, I also report p_τ, the p-value for hypothesis $H_0 : \Delta \geq \tau$ versus $H_A : \Delta < \tau$. I use $B = 9{,}999$

bootstrap resamples to generate the robustness statistics. When the sampling design uses clustering, I apply a common block bootstrap at the principal clustering level available across the comparison set and state any mismatch with a more complex original dependence structure. In Appendix F, I show the bootstrap correlation matrix for the point estimates for each application. Correlations between different specifications are often quite substantial, demonstrating the importance of accounting for the joint sampling variation in making inferences about the similarity of the estimates.

For each application, I also report how expert economists assessed the published table that I replicate. Respondents saw each table as it appeared in the original paper, with point estimates, standard errors, and table notes but without any of the statistics developed here.[13] Between 288 and 294 CEPR and NBER affiliates rated each table on a five-point scale from not at all robust to extremely robust. The tables were presented in random order.[14] The reader may find it useful to form a judgment about each table before reading the survey results and the framework statistics that follow it.

V.A Lee (2008): Sharp regression discontinuity

Lee (2008) examines the electoral advantage of incumbency in U.S. House elections using a sharp regression discontinuity design.[15] The "close elections" identification strategy relies on the discontinuity at 50 percent (in a two-party system) to generate quasi-random variation in the treatment. Using data from 6,558 elections from 1946 to 1998, he finds that incumbency raises a party's vote share in the next election by about 8 percentage points. In his Table 2, Lee presents results across multiple specifications that vary the inclusion of baseline covariates (prior vote share, officeholding experience, electoral experience, and all controls jointly) and notes that "the results are quite robust to various specifications... [and] the estimated effect also remains stable when a completely different method of controlling for pre-determined characteristics is utilized" (pp. 691–692).

I present the replicated results in Table V, Part A. All results from Lee (2008) Table 2 were reproduced exactly.[16] The first five "main" specifications use the Democratic vote share in the subsequent election as the dependent variable. Two supplemental specifications use residualized vote share (regressing the dependent variable on covariates first) and first-differenced vote share as dependent variables. All specifications target the same estimand, as the added covariates are pre-determined and, by the design, uncorrelated with treatment at the cutoff. The bootstrap resamples elections, matching the observation-level inference reported by Lee and preserving the joint dependence across specifications; it does not additionally cluster repeated elections within congressional districts.

[13] Full citations to each paper were also given, as the papers are prominent enough that it made little sense to try to conceal them. It should be noted that one of the columns in the example from Angrist and Pischke (2009) was not shown in the survey, as discussed below.

[14] Appendix E describes the survey design and instrument in detail.

[15] The data for this replication can be found in the *Mostly Harmless Econometrics* Data Archive: `https://economics.mit.edu/sites/default/files/inline-files/Lee2008%20(1).zip`, last accessed 16 April 2026.

[16] Lee also reports a placebo test in his Table 2, column 8, which I was able to replicate exactly but do not report here.

TABLE V, PART A
Lee (2008) Incumbency Advantage: Point Estimates

Lee Col:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\hat{\theta}$	0.0766	0.0777	0.0765	0.0765	0.0775	0.0808	0.0788
	(0.0114)	(0.0108)	(0.0114)	(0.0114)	(0.0109)	(0.0136)	(0.0134)
Controls							
Dem. vote share, $t-1$		✓			✓		
Dem. win, $t-1$		✓			✓		
Dem. political exp.			✓		✓		
Opp. political exp.			✓		✓		
Dem. electoral exp.				✓	✓		
Opp. electoral exp.				✓	✓		
Alternative outcome							
Residualized						✓	
First difference							✓

Notes: The dependent variable is Democratic vote share in election $t+1$. Standard errors are in parentheses. Point estimates replicate Table 2 of Lee (2008). Appendix Table F.I reports the bootstrap correlation matrix for these estimates.

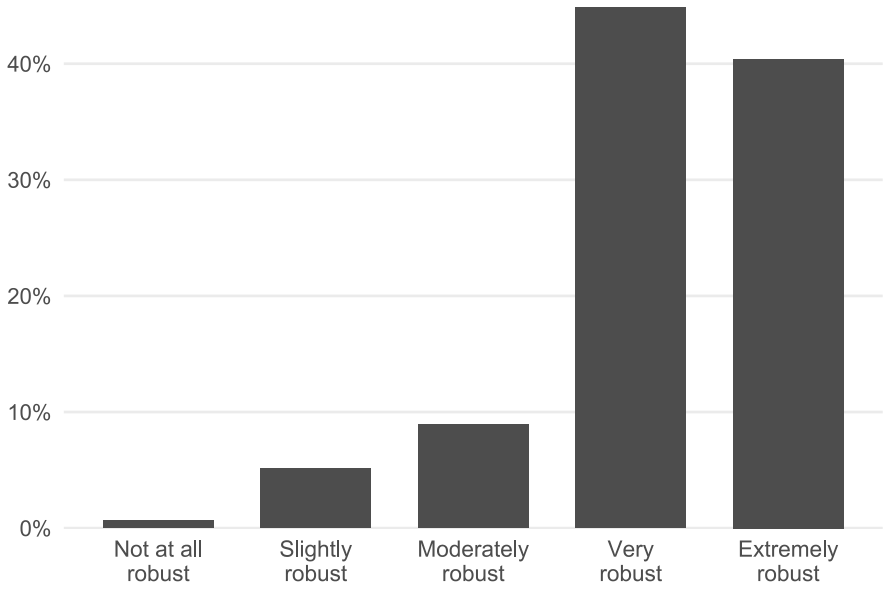
The five main specifications using Democratic vote share in $t+1$ as the dependent variable produce estimates ranging from 0.0765 to 0.0777. Adding the two alternative dependent variables to the comparison set increases the range to 0.0042, still less than half a percentage point on an average effect greater than 7 percentage points. Figure I shows how economists read this table. Of 292 CEPR and NBER affiliates shown Lee's Table 2, 85 percent rated the results very or extremely robust and 6 percent rated them not at all or slightly robust. The mean assessment, 4.19 on the five-point scale, is the highest of any table shown in the survey.

In Table V, Part B, I present the robustness statistics for the estimates in Part A. The bootstrap resamples elections, the observation unit of the original design. Asymptotic normality and joint bootstrap consistency hold here, so p_R has nominal asymptotic size and $R^*_{.95}$ is a valid upper confidence bound on Δ. For the "All specifications" result that corresponds to the comparison set rated by CEPR and NBER affiliates, I find $R^*_{.95} = 0.027$, relative to the average coefficient of 0.078, giving a robustness ratio of 0.347. Moreover, I cannot reject the null that the estimates are equal ($p_R = 0.921$). The robustness conclusion is even stronger if attention is limited to the five "main" specifications, with $R^*_{.95} = 0.008$, $p_R = 0.871$, and $R^*_{.95}/|\bar{\hat{\theta}}| = 0.105$. Including either the residual-based or first-difference-based specification with the five main specifications in the comparison set leads to similar conclusions of robustness.

The framework formally validates Lee's claim that the results are "quite robust to various specifications." The equality test cannot reject the null that the probability limits coincide, and the minimum equivalence bound supports equivalence within tight bounds relative to the estimated coefficients. Professional consensus concurs with the framework's conclusions.

To illustrate the value of a literature-based robustness tolerance τ, I borrow a benchmark from the incumbency-advantage literature, where modern estimates are typically within about two percentage points of one another (King, 1991; Cox and Katz, 1996; Ansolabehere and Snyder, 2002; Gelman and Huang, 2008). At $\tau = 0.02$, the equivalence test for the five main

FIGURE I
Survey Assessments of Robustness: Lee (2008) Incumbency Advantage



Notes: Distribution of responses from 292 CEPR and NBER affiliates asked to assess the results shown in Table V, Part A on a five-point Likert scale. Mean response is 4.19. Median response is 4.0.

TABLE V, PART B
Lee (2008) Incumbency Advantage: Robustness Tests

Comparison set	K	$\bar{\hat{\theta}}$	$R(\hat{\boldsymbol{\theta}})$	$R^*_{.50}$	$R^*_{.95}$	p_R	p_τ	$R^*_{.95}/\lvert\bar{\hat{\theta}}\rvert$
All specifications (1–7)	7	0.0778	0.0042	0.0128	0.0270	0.9208	0.1875	0.3471
Main (1–5)	5	0.0770	0.0012	0.0032	0.0081	0.8706	0.0001	0.1049
Main & residualized (1–6)	6	0.0776	0.0042	0.0088	0.0242	0.7690	0.1087	0.3123
Main & first difference (1–5, 7)	6	0.0773	0.0023	0.0076	0.0214	0.9191	0.0693	0.2766

Notes: Robustness statistics for estimates in Table V, Part A. $\bar{\hat{\theta}}$ is the mean of the full-sample point estimates and $R(\hat{\boldsymbol{\theta}})$ is their observed range in the comparison set. $R^*_{.50}$ and $R^*_{.95}$ are the median and 95th percentile of the unrecentered bootstrap range distribution. p_R is the add-one-adjusted p-value for the null hypothesis $H_0\colon \Delta = 0$, calculated from the recentered bootstrap distribution. At the literature-based tolerance $\tau = 0.02$, p_τ is the add-one-adjusted p-value for $H_0\colon \Delta \geq \tau$, calculated from the unrecentered bootstrap distribution. Results are based on $B = 9{,}999$ bootstrap resamples of elections with replacement.

specifications rejects $H_0 : \Delta \geq 0.02$, with $p_\tau = 0.0001$.[17] Lee's stability claim is therefore affirmatively established at a literature-based tolerance roughly one quarter of the average estimate. For the other comparison sets, which include the residualized and first-difference specifications, all equivalence p-values exceed 0.05, so equivalence within the two-percentage-point tolerance is not established at that level.

[17]This is the smallest p-value attainable with $B = 9{,}999$, as no unrecentered bootstrap range reached 0.02. Lemma 2 in Appendix A states the general duality between the equivalence p-value and $R^*_{.95}$.

V.B Karlan and List (2007): Randomized field experiment

Karlan and List (2007) conduct a large-scale natural field experiment to test whether matching grants affect the incidence of charitable giving.[18] Over 50,000 prior donors to a liberal US nonprofit received direct mail solicitations randomly assigned to a control group or one of several matching grant treatments. One of the paper's central findings is that the effects of the treatment are heterogeneous, with donors living in states that voted for George W. Bush in the 2004 presidential election ("red states") more likely to give than donors living in states that voted for John Kerry ("blue states"). In Table 6 of their paper, Karlan and List examine the robustness of this red-state finding by adding demographic interactions one at a time: the proportion of the state that is white, the proportion that is black, age, household size, median income, homeownership, college education, and proportion urban, plus an all-demographics specification that includes all interactions simultaneously. A tenth specification controls for prior donation history.[19] In every column, the coefficient of interest is the Treatment × Red State marginal effect from probit estimation.[20][21] Karlan and List write (p. 1785) that "[u]pon interacting treatment with education, income, racial composition, age, household size, home ownership, number of children and urban/rural indicator, the coefficient on the interaction term red*treatment [sic] remains robust." In footnotes 16 and 17 Karlan and List reiterate claims of robustness, including for specifications not present in the table.

I report replication results in Table VI, Part A, with the different specifications in rows.[22] The marginal effects range from 0.0067 (all demographics, in row 10) to 0.0098 (age and homeownership, in rows 4 and 7), with an observed range of 0.0032. All ten estimates are positive and all are individually significant at the 5 percent level with the exception of the all-demographics specification, which is significant only at the 10 percent level.

Survey responses of CEPR and NBER affiliates are shown in Figure II. Of 293 respondents, 66 percent rated the results very or extremely robust and 9 percent rated them not at all or slightly robust, with a mean assessment of 3.75, second only to Lee (2008) among the tables in the survey.

In Table VI, Part B, I report robustness tests for various comparison sets. The bootstrap resamples individuals, the observation unit of the original design. Asymptotic normality and joint bootstrap consistency hold here, so p_R has nominal asymptotic size and $R^*_{.95}$ is a valid upper confidence bound on Δ. Survey respondents saw the table as shown in the *American Economic*

[18] The data for this replication can be found on the Harvard Dataverse: `https://doi.org/10.7910/DVN/27853`, last accessed 16 April 2026.

[19] Karlan and List include an urban indicator as one of the specifications in their Table 6 but omit the results for space, reporting in the table note that the Treatment × Red State marginal effect remains 0.010. I present that regression as one of the 10 specifications.

[20] In a nonlinear probability model, the coefficient on an interaction term is not generally equal to the cross-partial effect of the interacted variables (Ai and Norton, 2003). I analyze the published marginal effects because they are the objects to which Karlan and List attach their robustness claim.

[21] Karlan and List used the Stata command `dprobit` to estimate the marginal effects in their Table 6. Although the command has been deprecated in Stata, I also use it here to replicate their results exactly. Using the more current `probit` with `margins, dydx(*) atmeans` produces similar but not identical marginal effects.

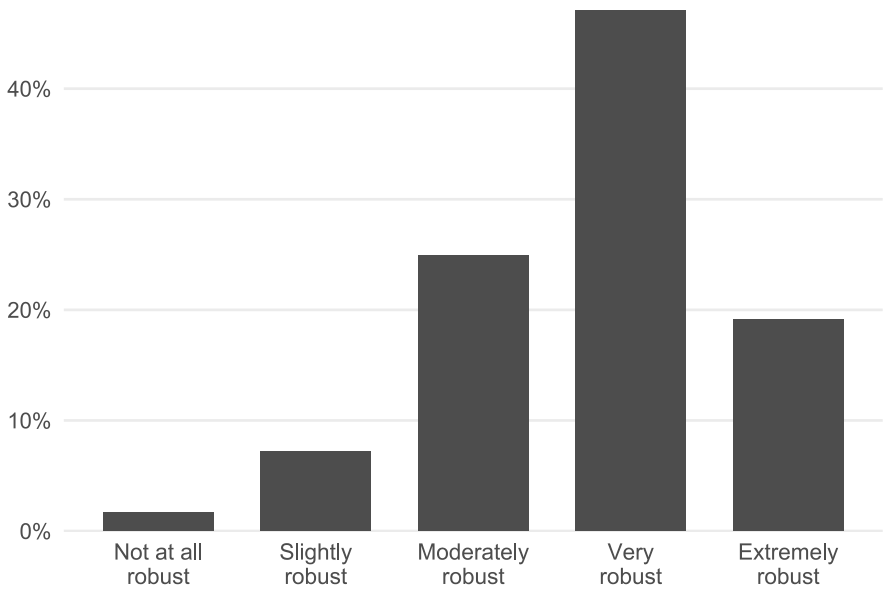
[22] Sample sizes vary across specifications because of missing demographic data. Within each bootstrap resample, I draw individuals with replacement from the full dataset and estimate each specification using the observations from the resample that have non-missing values for that specification's covariates. The joint dependence across specifications is preserved because the same resampled individuals enter every specification that their data support.

TABLE VI, PART A
Karlan and List (2007) Treatment × Red State: Point Estimates

			Controls included									
Spec.	$\hat{\theta}$	(s.e.)	Donation hist.	Prop. white	Prop. black	Age 18–39	HH size	Med. HH income	Home own.	College	Urban	$\bar{N}$
(1)	0.0093	(0.0034)	✓									50,048
(2)	0.0097	(0.0035)		✓								48,210
(3)	0.0095	(0.0035)			✓							48,040
(4)	0.0098	(0.0035)				✓						48,210
(5)	0.0095	(0.0035)					✓					48,214
(6)	0.0090	(0.0035)						✓				48,202
(7)	0.0098	(0.0035)							✓			48,207
(8)	0.0089	(0.0035)								✓		48,208
(9)	0.0095	(0.0035)									✓	48,210
(10)	0.0067	(0.0036)		✓	✓	✓	✓	✓	✓	✓	✓	48,032

Notes: The dependent variable is an indicator for donating within one month of the solicitation. All specifications are probit models, and the reported coefficients are marginal effects of Treatment × Red State estimated with `dprobit` in Stata. Each specification includes treatment, red state, their interaction, and the indicated demographic variables and treatment interactions. Standard errors are in parentheses. $\bar{N}$ is the average sample size across the 9,999 bootstrap resamples; sample sizes vary across specifications because of missing demographic data. Marginal effects replicate Table 6 of Karlan and List (2007). Appendix Table F.II reports the bootstrap correlation matrix for these estimates.

FIGURE II
Survey Assessments of Robustness: Karlan and List (2007)



Notes: Distribution of responses from 293 CEPR and NBER affiliates asked to assess the results shown in Table VI, Part A, as published (without the urban specification), on a five-point Likert scale. Mean response is 3.75. Median response is 4.0.

Review, without the urban specification, which Karlan and List estimated and discussed but did not display. The "As displayed" row of Table VI presents this comparison set, yielding a p_R of 0.077, $R^*_{.95}$ of 0.0058, and a robustness ratio of 0.633, essentially the same as the "All specifications" row because the urban estimate lies inside the range of the displayed estimates and its inclusion leaves the extrema of the joint bootstrap estimates essentially unchanged. The comparison sets that include all of the demographic variables are weakly robust while the individual demographics comparison set and the comparison set that includes those and the donation history are robust.

TABLE VI, PART B
Karlan and List (2007) Treatment × Red State: Robustness Tests

Comparison set	K	$\bar{\hat{\theta}}$	$R(\hat{\boldsymbol{\theta}})$	$R^*_{.50}$	$R^*_{.95}$	p_R	$R^*_{.95}/\lvert\bar{\hat{\theta}}\rvert$
All specifications (1–10)	10	0.0092	0.0032	0.0035	0.0058	0.0784	0.6315
As displayed (1–8, 10)	9	0.0091	0.0032	0.0035	0.0058	0.0775	0.6329
Indiv. demog. only (2–9)	8	0.0095	0.0009	0.0014	0.0027	0.6601	0.2841
Indiv. demog. & all demog. (2–10)	9	0.0092	0.0032	0.0034	0.0057	0.0572	0.6251
Donation & indiv. demog. (1–9)	9	0.0094	0.0009	0.0016	0.0028	0.8028	0.2983
Donation & all demog. (1 & 10)	2	0.0080	0.0026	0.0026	0.0052	0.1070	0.6512

Notes: Robustness statistics for estimates in Table VI, Part A. $\bar{\hat{\theta}}$ is the mean of the full-sample marginal effects and $R(\hat{\boldsymbol{\theta}})$ is their observed range in the comparison set. $R^*_{.50}$ and $R^*_{.95}$ are the median and 95th percentile of the unrecentered bootstrap range distribution. p_R is the add-one-adjusted p-value for the null hypothesis $H_0\colon \Delta = 0$, calculated from the recentered bootstrap distribution. Results are based on $B = 9{,}999$ bootstrap resamples of individuals with replacement. The "as displayed" set excludes the urban specification (row 9), which Karlan and List (2007) estimated and discussed but did not report, to match the table shown to survey respondents.

As in Lee (2008), the null of equal probability limits is not rejected at the 5 percent level and the minimum equivalence bounds indicate agreement across the estimates, although the classification depends on the comparison set. The individual-demographic specifications are robust by the heuristic, with robustness ratios below 0.30. The full displayed comparison set is weakly robust because the imprecise all-demographics specification widens the bound. Karlan and List's broad conclusion is therefore supported most clearly for the specifications that vary one demographic control at a time.

V.C Autor (2003): Staggered-adoption difference-in-differences

Autor (2003) examines the relationship between adoption of exceptions to "employment at will" and temporary help services (THS) employment between 1979 and 1995.[23] Autor uses a staggered difference-in-differences design, with states ruling at different times that workers are entitled to ongoing employment even in the absence of a contract if contractual rights are guaranteed by the context of the employment relationship. Identification in the design comes from controlling for state and year fixed effects and state-specific linear time trends.[24] He finds that states whose courts ruled that there was an implied contract exception to the employment-at-will doctrine experienced higher THS employment. In the baseline specification (Table 4,

[23]The data for this replication can be found on David Autor's data archive page: `https://economics.mit.edu/people/faculty/david-h-autor/data-archive`, last accessed 16 April 2026.

[24]I apply my framework to the "zoo" of two-way fixed-effects estimators in Jaeger (2026).

column 1), he reports a "baseline" result of an increase of 0.137 (standard error 0.062) in log THS employment.[25]

I present my replication of Autor's robustness checks from his Table 5 in Table VII, Part A, along with the baseline specification. As a rule, when one specification is maintained as preferred, I recommend that it be included in all of the comparison sets. The column labeled (0) is the baseline specification and the other columns correspond to the same-numbered column in Autor's Table 5. All results replicated exactly. The bootstrap resamples states (the level of policy variation and clustering), preserving within-state serial correlation. At 50 clusters, Design 7 in the simulations shows no size distortions for the range-based equality test and, when the population range is positive, the equivalence bound covers at or above its nominal level.

TABLE VII, PART A
Autor (2003) Implied Contract Exception: Point Estimates

Autor col:	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\hat{\theta}$	0.1374	0.1484	0.1318	0.1741	0.1414	0.1338	0.1450	0.1412
	(0.0618)	(0.0566)	(0.0628)	(0.0557)	(0.0681)	(0.0772)	(0.0560)	(0.0678)
Log nonfarm emp.		✓	✓	✓	✓	✓	✓	✓
Linear trends	✓	✓	✓	✓	✓		✓	✓
Quadratic trends			✓		✓			✓
Region × year				✓	✓			✓
Demographics						✓	✓	✓

Notes: The dependent variable is log state temporary help services employment. All specifications include state and year fixed effects and controls for good-faith and public-policy exceptions; columns (1)–(7) also include log nonfarm employment. State-clustered standard errors are in parentheses. The sample contains $N = 850$ state-year observations from 50 states. The baseline estimate in column (0) replicates column (1) of Table 4 in Autor (2003), and columns (1)–(7) replicate the corresponding columns of his Table 5. Appendix Table F.III reports the bootstrap correlation matrix for these estimates.

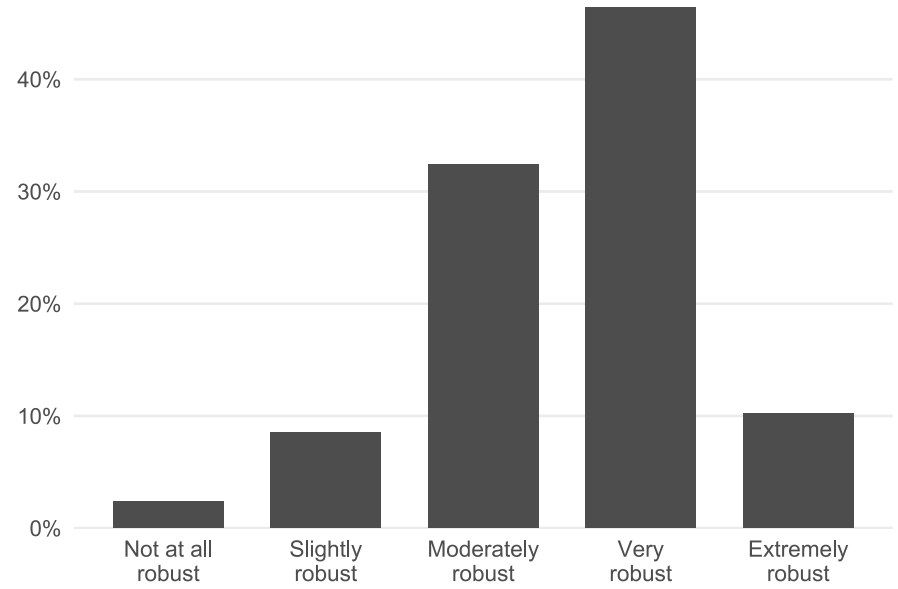
Autor states that in Table 5 of Autor (2003) he "test[s] the robustness of the results by controlling for a richer set of covariates, including state employment, quadratic state time trends, region-by-year effects, and labor-force demographics." The implied contract exception to the employment-at-will doctrine increases THS employment by between 0.132 and 0.174 log points, with a range of 0.042 log points. Autor concludes that when adding controls for "quadratic state time trends and interactions between year dummies and indicator[s] for each of the nine census regions that allow state THS employment to trend nonlinearly and also absorb region-specific shocks... the implied contract coefficient is largely insensitive to these additional controls."

Figure III shows that CEPR and NBER affiliates agree with Autor. Of 293 respondents shown the published table (without the baseline specification), 57 percent rated the results very or extremely robust and 89 percent rated them at least moderately robust. The visible range of the point estimates, 0.042 on a mean of approximately 0.145 appears to be tight.

For the two full comparison sets, labeled "All specifications" and "All specs. excl. baseline" in Table VII, Part B, p_R exceeds 0.9, meaning that the differences among the estimates are

[25]This estimate is also presented in column (8), Table 3 of Autor (2003).

FIGURE III
Survey Assessments of Robustness: Autor (2003)



Notes: Distribution of responses from 293 CEPR and NBER affiliates asked to assess the results shown in Table VII on a five-point Likert scale. Mean response is 3.54. Median response is 4.0.

not statistically detectable. Equality is not rejected for any of the other comparison sets.[26] Examining the minimum equivalence bounds here tells a different story than Autor and the survey respondents' "robust" conclusion, however. I estimate a minimum equivalence bound of 0.192 for all of the specifications in Table VII, Part A, about 33 percent larger than the average coefficient estimate. Removing the baseline specification from the comparison set has little effect, yielding an $R^*_{.95}$ of 0.189. This is the comparison set that survey respondents were shown and that a majority judged to be robust. The minimum equivalence bound adds information that neither the published table nor the survey respondents had, since non-rejection of equality does not establish tight agreement across the comparison set.

Removing the largest estimate, column (3), from the comparison set does little to change the minimum equivalence bound, which falls from 0.192 to 0.182. Excluding the least precise estimate, column (5), has a larger effect, reducing $R^*_{.95}$ to 0.138. Even then, $R^*_{.95}$ remains almost as large as the average estimated coefficient, placing the robustness ratio only marginally in the "weakly robust" range. The only comparison set classified as robust is also the narrowest, containing the baseline and a specification that adds log non-farm employment as a regressor. For that pair, $R^*_{.95} = 0.038$ and the average coefficient is 0.143.

While Autor's point estimates lie in a relatively narrow numerical range, the minimum equivalence bound shows that the joint sampling distribution is too imprecise to establish equivalence.

[26]These results are an example in which a baseline-relative robustness measure is natural because Autor explicitly privileges column (0). Since the baseline estimate lies near the center of the estimates, deviations from it mechanically produce a smaller one-sided distance than the full range. This illustrates the distinction between a baseline-relative measure and the comparison-set-level question addressed by my framework. Adding the baseline to the non-baseline comparison set therefore barely changes the equivalence bound, from 0.189 to 0.192.

TABLE VII, PART B
Autor (2003) Implied Contract Exception: Robustness Tests

Comparison set	K	$\bar{\hat{\theta}}$	$R(\hat{\boldsymbol{\theta}})$	$R^*_{.50}$	$R^*_{.95}$	p_R	$R^*_{.95}/\lvert\bar{\hat{\theta}}\rvert$
All specifications (0–7)	8	0.1441	0.0422	0.0998	0.1920	0.9399	1.3320
All specs. excl. baseline (1–7)	7	0.1451	0.0422	0.0968	0.1894	0.9188	1.3056
All specs. excl. largest point est. (0–2, 4–7)	7	0.1399	0.0166	0.0869	0.1824	0.9975	1.3039
All specs. excl. largest std. err. (0–4, 6–7)	7	0.1456	0.0422	0.0754	0.1378	0.7799	0.9466
Baseline & employment (0, 1)	2	0.1429	0.0110	0.0142	0.0380	0.4912	0.2660
Baseline & no demographics (0–4)	5	0.1466	0.0422	0.0698	0.1339	0.6917	0.9130
Baseline & demographics (0, 5–7)	4	0.1393	0.0111	0.0785	0.1732	0.9951	1.2430
Baseline & all controls (0, 7)	2	0.1393	0.0037	0.0358	0.1046	0.9401	0.7513

Notes: Robustness statistics for estimates in Table VII, Part A. $\bar{\hat{\theta}}$ is the mean of the full-sample point estimates and $R(\hat{\boldsymbol{\theta}})$ is their observed range in the comparison set. $R^*_{.50}$ and $R^*_{.95}$ are the median and 95th percentile of the unrecentered bootstrap range distribution. p_R is the add-one-adjusted p-value for the null hypothesis $H_0\colon \Delta = 0$, calculated from the recentered bootstrap distribution. Results are based on $B = 9{,}999$ bootstrap resamples of states with replacement.

The equality tests do not reject that the estimates are equal, but the estimates are also too noisy to confirm similarity. The table therefore establishes descriptive stability of the point estimates, not affirmative equivalence at a tight tolerance. The joint imprecision behind $R^*_{.95} = 0.192$ for the all-specifications comparison set is invisible when only the point estimates and standard errors are displayed.

V.D Dube, Lester, and Reich (2010): Spatial difference-in-differences

Dube et al. (2010), henceforth DLR, estimate the effect of minimum wage increases on earnings and employment in the restaurant industry using a spatial difference-in-differences design.[27] Their design extends the 2×2 difference-in-differences design of Card and Krueger (1994) to multiple contiguous border county pairs that straddle state borders, using county-pair-by-period fixed effects to absorb common shocks within each pair. In their Table 2, DLR present estimates of the earnings and employment elasticities with respect to the minimum wage across twelve specifications. Eight specifications use an all-county sample, varying the fixed-effect structure (period, census-division by period, state trends plus census-division by period, and MSA by period) and whether total private-sector earnings are included as a control. Four specifications use a contiguous border county-pair sample, with either period fixed effects or county-pair-by-period fixed effects, again with and without the private-sector control.

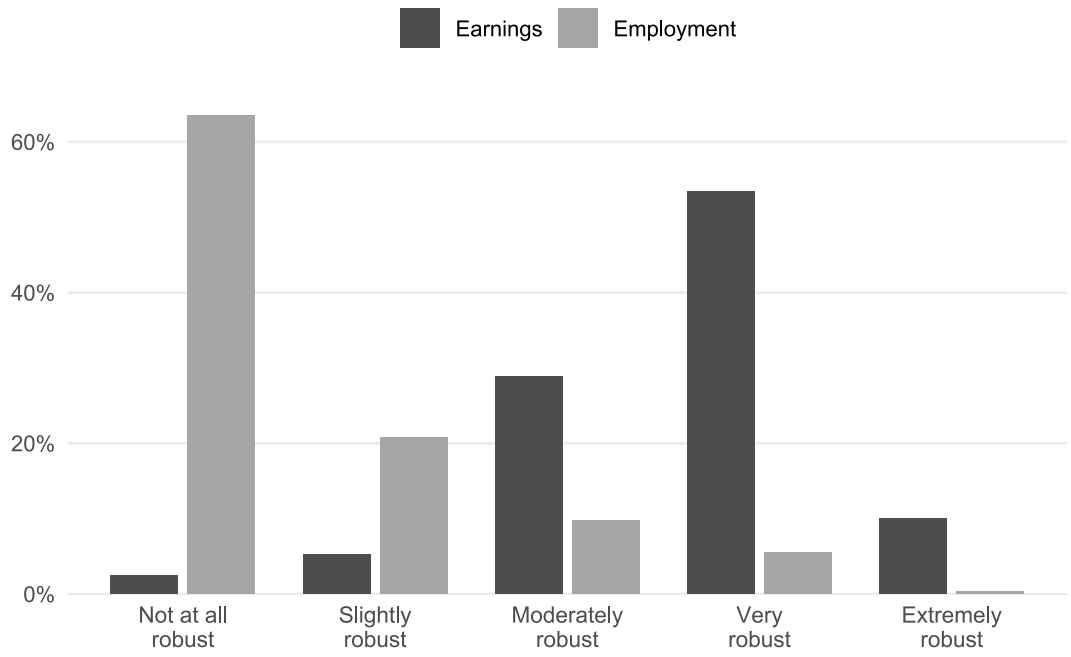
I report my replication of DLR's results in Table VIII, Part A. DLR number column pairs, and so my columns (1) and (2) correspond to their columns (1) in Table 2, etc. The bootstrap resamples states, preserving the principal clustering dimension that can be applied jointly across all specifications and both outcomes. The original border specifications also allow clustering by border segment, which state resampling does not reproduce. The affected border-sample bounds and equality tests should therefore be interpreted with this limitation in mind. For the border specifications, I report both state-clustered bootstrap standard errors and, in square brackets,

[27] The data for this replication are available in the Harvard Dataverse: `https://doi.org/10.7910/DVN/L4DUZ7`. Last seen 16 April 2026.

the two-way clustered standard errors from DLR. Where the latter are larger, particularly for the pair-by-period specifications, the state-level bootstrap may understate joint sampling variation and produce equivalence bounds that are too narrow.[28] All results replicate DLR's point estimates exactly. Estimates of the earnings elasticity range from 0.149 to 0.224 in the all-county sample and from 0.188 to 0.232 in the contiguous border county-pair sample, with all estimates statistically significantly different from zero at the 5 percent level. In contrast, the estimates of the employment elasticity range from −0.211 to 0.054 in the all-county sample and from −0.137 to 0.057 in the contiguous border county-pair sample. Among the employment elasticity estimates, only the estimate in column (1) is statistically different from zero at the 5 percent level, with those in columns (2) and (9) significant at the 10 percent level.

Survey respondents were asked to assess the earnings and employment results separately, with results shown in Figure IV. The responses to the two sets of results contrast sharply, with 64 percent of 288 respondents rating the earnings estimates very or extremely robust, while 84 percent of the same respondents rated the employment estimates not at all or slightly robust, with the majority of respondents (64 percent) seeing the results as "not at all robust."

FIGURE IV
Survey Assessments of Robustness: Dube, Lester, and Reich (2010)



Notes: Distribution of responses from 288 CEPR and NBER affiliates asked to assess the earnings and employment results shown in Table VIII on a five-point Likert scale. Mean response for earnings is 3.64 Median response for earnings is 4.0. Mean response for employment is 1.58. Median response for employment is 1.0

[28]Dube et al. (2010) cluster by state and border segment. I use a state-level cluster bootstrap so that the same resampled states are used across all specifications and both outcomes, preserving cross-specification and cross-equation covariance. Because border segments cross state lines, the bootstrap does not preserve dependence between observations in different states that share a segment. This omission appears negligible for the period fixed-effects specifications, where the state-only and two-way standard errors in Table VIII, Part A, are similar. See Menzel (2021) for a formal discussion of using the bootstrap for two-way cluster-robust variance estimation.

TABLE VIII, PART A
Dube, Lester, and Reich (2010): Point Estimates

	All-county								Border			
	Period FE		CensDiv		StateT+CD		MSA		Period FE		Pair×Period	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Log earnings												
	0.2242	0.2166	0.2037	0.1951	0.2187	0.2102	0.1530	0.1492	0.2318	0.2210	0.2001	0.1887
	(0.0330)	(0.0278)	(0.0381)	(0.0342)	(0.0371)	(0.0343)	(0.0298)	(0.0276)	(0.0337)	(0.0341)	(0.0479)	(0.0435)
									[0.032]	[0.032]	[0.065]	[0.060]
Log employment												
	−0.2109	−0.1765	−0.0277	−0.0233	0.0545	0.0392	0.0524	0.0321	−0.1367	−0.1118	0.0572	0.0165
	(0.0955)	(0.0965)	(0.0665)	(0.0682)	(0.0553)	(0.0501)	(0.0837)	(0.0780)	(0.0702)	(0.0764)	(0.0795)	(0.0617)
									[0.072]	[0.076]	[0.118]	[0.098]
Implied labor demand elasticity (median [2.5, 97.5] bootstrap percentiles)												
		−0.82		−0.14		0.16		0.31		−0.57		−0.01
		[−1.77, 0.24]		[−1.03, 0.88]		[−0.27, 1.01]		[−2.21, 2.74]		[−1.39, 0.10]		[−0.95, 0.96]
$\bar{N}$	91,249	91,230	91,249	91,230	91,249	91,230	48,462	48,445	70,510	70,472	70,510	70,472
Controls												
County FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Period FE	✓	✓	✓	✓	✓	✓			✓	✓		
Cens. div. × period			✓	✓	✓	✓						
State linear trends					✓	✓						
MSA × period							✓	✓				
County-pair × period											✓	✓
Pop (emp. only)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Total priv. sector		✓		✓		✓		✓		✓		✓

Notes: The dependent variables are log average weekly earnings and log employment in restaurants from the QCEW. The treatment variable is log minimum wage after Frisch–Waugh–Lovell projection. State-clustered standard errors are in parentheses; for the border specifications, square brackets report the original two-way-clustered standard errors (state and border segment). $\bar{N}$ is the average sample size across the 9,999 bootstrap resamples, which varies because states are drawn with replacement. The same resampled states are used for both outcomes. The implied labor-demand elasticity is the ratio of the employment coefficient to the earnings coefficient in each bootstrap resample; brackets report its 2.5th and 97.5th percentiles. Point estimates replicate Table 2 of Dube et al. (2010). Appendix Tables F.IV and F.V report the bootstrap correlation matrices for earnings and employment.

In Table VIII, Part B, I present the robustness statistics for DLR's estimates. Across all twelve of the earnings specifications, the observed range is less than half of the average coefficient (0.083 versus 0.201). The minimum equivalence bound for the comparison set of all 12 earnings specifications is 0.214, slightly above the average estimated coefficient, yielding a robustness ratio of 1.066. The all-county and border samples target different populations, so this full-table comparison bounds dispersion across distinct estimands rather than disagreement about a single object. Taking just the all-county specifications (1–8) does not change the conclusion much, since this comparison set has a robustness ratio of 1.008. DLR argue that the border specifications are most credible, however, because county pairs that straddle state boundaries provide a natural control group for the local shocks that can confound the all-county design. Across the four contiguous border county-pair specifications, the estimates are "weakly robust," with $R^*_{.95} = 0.144$ and an average coefficient of 0.210. Comparing similar specifications from the all-county and contiguous border county-pair samples either without (columns 1 and 9) or with (columns 2 and 10) the private-sector control yields small bounds relative to the coefficients, in both cases a robustness ratio of 0.241.

TABLE VIII, PART B
Dube, Lester, and Reich (2010): Robustness Tests

Comparison set	K	$\bar{\hat{\theta}}$	$R(\hat{\boldsymbol{\theta}})$	$R^*_{.50}$	$R^*_{.95}$	p_R	p_τ	$R^*_{.95}/\lvert\bar{\hat{\theta}}\rvert$
Earnings								
All specifications (1–12)	12	0.2010	0.0826	0.1137	0.2144	0.6060	–	1.0667
All-county (1–8)	8	0.1963	0.0751	0.0913	0.1979	0.4974	–	1.0080
Border (9–12)	4	0.2104	0.0431	0.0569	0.1440	0.5031	–	0.6847
Period FE: All-county & border (1 & 9)	2	0.2280	0.0075	0.0197	0.0549	0.7764	–	0.2406
Period FE & priv. sect.: All-cty & bord. (2 & 10)	2	0.2188	0.0044	0.0186	0.0528	0.8617	–	0.2411
Employment								
All specifications (1–12)	12	−0.0362	0.2681	0.3607	0.5748	0.3611	0.9981	15.8581
All-county (1–8)	8	−0.0325	0.2654	0.3441	0.5651	0.3089	0.9938	17.3780
Border (9–12)	4	−0.0437	0.1940	0.1939	0.3832	0.1067	0.8630	8.7683
Period FE: All-county & border (1 & 9)	2	−0.1738	0.0742	0.0630	0.1540	0.2693	0.2521	0.8859
Period FE & priv. sect.: All-cty & bord. (2 & 10)	2	−0.1442	0.0647	0.0552	0.1388	0.2872	0.1803	0.9629

Notes: Robustness statistics for estimates in Table VIII, Part A. $\bar{\hat{\theta}}$ is the mean of the full-sample point estimates and $R(\hat{\boldsymbol{\theta}})$ is their observed range in the comparison set. $R^*_{.50}$ and $R^*_{.95}$ are the median and 95th percentile of the unrecentered bootstrap range distribution. p_R is the add-one-adjusted p-value for the null hypothesis $H_0\colon \Delta = 0$, calculated from the recentered bootstrap distribution. For employment, at the literature-based tolerance $\tau = 0.10$, p_τ is the add-one-adjusted p-value for $H_0\colon \Delta \geq \tau$, calculated from the unrecentered bootstrap distribution. No tolerance is specified for earnings. Results are based on $B = 9{,}999$ bootstrap resamples of states with replacement, using the same resampled states for both outcomes. The robustness ratio should be interpreted cautiously when $\bar{\hat{\theta}}$ is near zero.

The picture is very different for the estimated employment elasticities, and DLR themselves are explicit about this. Across all twelve specifications, point estimates range from −0.211 (all-county, period FE) to 0.057 (border, pair-by-period). The large minimum equivalence bound for this comparison set, 0.575, reflects both this variability and the lack of precision of the estimates. DLR interpret the differences between the all-county and contiguous border county-pair results as evidence that the traditional all-county specifications suffer from omitted-variable bias and that the border-county specifications, which absorb local shocks through pair-by-period fixed effects, are the credible ones. Within the border specifications, however, $R^*_{.95} = 0.383$ on a mean

of -0.044. Even relative to the largest (in absolute value) estimate (-0.137 in column 9), the minimum equivalence bound is large.

A substantive benchmark is present in the traditional minimum-wage literature, which commonly places employment elasticities between -0.1 and -0.2 (Neumark and Wascher, 1992, 2007). I therefore set $\tau = 0.10$, which permits differences across specifications as large as the lower end of the conventional estimate of the entire employment response to a minimum-wage increase. None of the employment comparison sets rejects $H_0 \colon \Delta \geq 0.10$ at the 5 percent level. Even the two narrow comparisons between otherwise similar all-county and border specifications yield equivalence p-values of 0.252 and 0.180. The data therefore do not establish that the employment estimates are equivalent within a conventionally meaningful employment response.

The framework extends naturally to derived quantities. As an example, in Table VIII, Part A, I also report the median and 95 percent confidence interval for the implied labor demand elasticity, $\hat{\eta} = \hat{\theta}_{\text{emp}}/\hat{\theta}_{\text{earn}}$, that DLR report in their Table 2 for the specifications that include the private-sector control. Both the elasticity and the confidence interval are computed from the bootstrap distribution of the ratio, while DLR's reported elasticities are computed from the jointly estimated SUR system, so their point estimates differ slightly from the ratio medians reported here. The implied elasticities are imprecisely estimated. The median ranges from -0.82 in the simplest all-county specification to 0.31 in the MSA-by-period specification, and every confidence interval includes zero.

Taken as a whole, the framework yields conclusions broadly consistent with DLR's own interpretation, although they depend somewhat on the comparison set. The set containing all 12 earnings specifications has a robustness ratio of 1.066 and is therefore not robust by the heuristic, while the preferred contiguous-border county-pair comparison set is weakly robust. The employment estimates, which DLR describe as fragile, are not robust, and equivalence is never established at the literature-based tolerance $\tau = 0.10$. Survey respondents assessed the full published earnings table and generally rated the earnings estimates robust, while rating the employment estimates fragile. The survey and the framework therefore agree clearly on employment, while the earnings results illustrate the importance of the comparison set.

V.E Angrist and Krueger (1991)/Angrist and Pischke (2009): Instrumental variables

The final illustration should be familiar to anyone who has studied labor economics or applied microeconometrics in the last 30 years. I apply the framework to some of the most replicated results in economics and re-examine a robustness argument involving weak instruments presented in *Mostly Harmless Econometrics* (Angrist and Pischke, 2009, henceforth AP). In their original paper, Angrist and Krueger (1991) exploit the interaction of school entry-age laws and compulsory schooling laws, using quarter of birth interacted with year of birth as instruments for years of education to estimate the returns to schooling. Beginning with Bound et al. (1995) and Staiger and Stock (1997), the data from Angrist and Krueger (1991) have been reanalyzed many times in the weak- and many-instruments literature.

My focus here is on the results and arguments regarding comparing two-stage least squares (TSLS) and limited information maximum likelihood (LIML) estimates of the returns to edu-

cation presented in AP's discussion of the finite-sample bias of TSLS (pp. 205–218). AP offer the following practical advice for researchers concerned about this bias:

> Check overidentified 2SLS with LIML. LIML is less precise than 2SLS but also less biased. If the results come out similar, be happy. If not, worry, and try to find stronger instruments or reduce the degree of overidentification. (p. 213)

This is a textbook (literally) example of informal robustness reasoning. What does "similar" mean? How similar is similar enough? And does failure to detect a difference between two noisy estimates constitute evidence that they agree and that we can "be happy"? The heuristic relies on LIML remaining approximately median unbiased at weaker identification than TSLS. Similarity between LIML and TSLS is then taken as evidence that finite-sample bias in TSLS is small, even when the first-stage F statistic on the excluded instruments is low.

I replicate five of the six specifications from AP Table 4.6.2, omitting column 2 (where the first-stage F statistic indicates the specification is essentially unidentified), using the data from Angrist and Krueger (1991). In Table IX, Part A, I present OLS, TSLS, and LIML estimates of the returns to education for the same sample of 329,509 men born 1930–1939 from Angrist and Krueger (1991).[29] Column numbers refer to AP Table 4.6.2. All specifications control for year of birth. Column 1 presents results from a specification with three quarter of birth instruments (for TSLS and LIML) similar to the one estimated by Bound et al. (1995). Columns 3 and 4 present specifications with 30 and 28 quarter of birth $\times$ year of birth instruments, respectively (without and with a quadratic in age measured in quarters), while columns 5 and 6 present specifications with 180 and 178 quarter of birth $\times$ year of birth and quarter of birth $\times$ state of birth instruments, respectively (again, without and with a quadratic in age). All estimates reproduce those in AP exactly. With 329,509 observations, the OLS estimates have minuscule standard errors, with t-ratios over 200. The TSLS and LIML estimates are also in a relatively narrow band, with estimates from 0.0760 (TSLS in column 4) to 0.1100 (LIML in column 6) with a range of 0.034. Within columns, the TSLS and LIML estimates are even closer, and casual inspection would readily classify them as "similar."

AP make this assessment themselves. About the results in columns 3 and 4 (the 30/28-instrument case), they write:

> The 2SLS estimates are a bit lower than in column 1 and hence closer to OLS. But LIML is not too far away from 2SLS. Although the LIML standard error is pretty big in column 4, it is not so large that the estimate is uninformative. On balance, there seems to be little cause for worry about weak instruments in 30-instrument models, even with the age quadratic included. (p. 215)

They also consider the 180/178-instrument case:

> The most worrisome specifications are those reported in columns 5 and 6.... The first-stage F for these models are only 2.6 and 2.0. On the plus side, the LIML estimates again look fairly similar to 2SLS. Moreover, the LIML standard errors are

[29] The data for this replication are available from the author at `http://www.djaeger.org/data/ak3039.dta`. They are also available on Josh Angrist's data archive `https://economics.mit.edu/people/faculty/josh-angrist/angrist-data-archive`, last accessed 16 April 2026.

TABLE IX, PART A
Angrist and Pischke (2009): Point Estimates

AP Column	(1)	(3)	(4)	(5)	(6)
OLS	0.0711	0.0711	0.0711	0.0673	0.0673
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
TSLS	0.1052	0.0891	0.0760	0.0928	0.0907
	(0.0201)	(0.0161)	(0.0290)	(0.0093)	(0.0107)
LIML	0.1064	0.0929	0.0811	0.1064	0.1100
	(0.0205)	(0.0177)	(0.0415)	(0.0116)	(0.0147)
Controls					
Year of birth	✓	✓	✓	✓	✓
State of birth				✓	✓
Age, age^2			✓		✓
Instruments					
QOB	✓				
QOB × YOB		✓	✓	✓	✓
QOB × SOB				✓	✓
Number of instruments	3	30	28	180	178
First-stage F	32.27	4.91	1.61	2.58	1.97

Notes: Columns correspond to Table 4.6.2 of Angrist and Pischke (2009), omitting column 2 ($F = 0.42$). The dependent variable is log weekly wage. Standard errors are in parentheses. The sample contains $N = 329{,}509$ observations. Point estimates replicate Table 4.6.2 of Angrist and Pischke (2009), using data from Angrist and Krueger (1991). Appendix Table F.VI reports the bootstrap correlation matrix for these estimates.

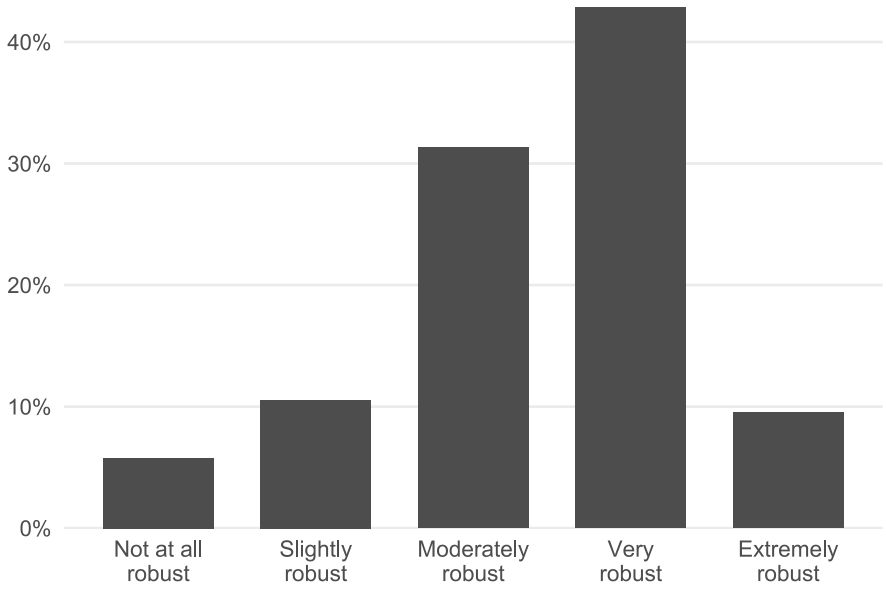
> not too far above the 2SLS standard errors in this case. This suggests you can't always determine instrument relevance using a mechanical rule... in some cases a low F may not be fatal. (p. 215)

Figure V shows that survey respondents share AP's assessment. Shown Table 4.6.2 with column (2) whited out and the TSLS and LIML estimates in columns (1) and (3)–(6) outlined, 52 percent of 294 respondents rated the results very or extremely robust and 84 percent rated them at least moderately robust. The published table reports the first-stage F statistics, so the information about the strength of identification in the IV models was known to the respondents, and the 16 percent of respondents who rated the results not at all or slightly robust form the largest skeptical minority among the tables without visible disagreement in the survey. The median respondent judged the IV estimates as a group to be very robust.

In Table IX, Part B, I present robustness statistics for the point estimates in Table IX, Part A. The table reports uniform comparison sets: the within-column TSLS–LIML pairs, which isolate estimator agreement and correspond to AP's heuristic, the within-column OLS–TSLS and OLS–LIML pairs, which bound the OLS-versus-IV gap in each column, and the within-estimator across-column sets, which abstract from estimator choice and focus on specification variation. I do not report any comparison sets that bundle the common-probability-limit TSLS–LIML pair with an OLS specification.[30] The comparison set highlighted for survey re-

[30]Corollary 1 in the Appendix shows the validity of a bound on a mixed comparison set under its stated conditions, but a single number would conflate estimator agreement with the separate question of the OLS-versus-IV gap, which the within-column OLS–TSLS and OLS–LIML pairs bound separately.

FIGURE V
Survey Assessments of Robustness: Returns to Education



Notes: Distribution of responses from 294 CEPR and NBER affiliates asked to assess the results shown in Table IX on a five-point Likert scale. Mean response is 3.40. Median response is 4.0.

spondents, the ten TSLS and LIML estimates across the five columns, has a joint minimum equivalence bound of $R^*_{.95} = 0.460$, nearly five times the average estimate of 0.095. This is driven primarily by the LIML estimate in column 4. Across columns, the bound for the five LIML estimates alone is 0.452, compared with 0.086 for the five TSLS estimates. Under the constant-effects assumption that AP's exercise maintains, all ten estimates target a single return to schooling, and the joint bound summarizes their dispersion whatever the identification strength.[31] An important caveat applies to the weakly identified columns. Neither TSLS nor LIML is asymptotically normal at the conventional rate under weak identification (Staiger and Stock, 1997), and the pairs bootstrap is generally not valid in this setting, so the formal interpretation of $R^*_{.95}$ as a 95% upper confidence bound on Δ, established in Appendix A under standard regularity conditions, does not apply there. The reported statistics should be read as descriptive summaries of how dispersed the joint bootstrap estimates of TSLS and LIML are in this sample.

Mincer wage equations are among the most frequently estimated relationships in economics. As a benchmark for the spread of estimated returns to schooling, I use the controlled OLS–IV differences summarized in Tables 4 and 5 of Card (1999). Excluding the estimates from Angrist and Krueger (1991), the median absolute difference in those tables is approximately three percentage points, and I therefore set $\tau = 0.03$. The five OLS estimates provide a particularly clear example of statistically detectable but substantively negligible differences: equality is rejected with $p_R = 0.0001$, while the minimum equivalence bound is only $R^*_{.95} = 0.0039$.

[31] See Proposition 2(i) in Appendix A. The maximum over the constituent pairwise bounds, 0.445, corroborates the joint figure from below, per Corollary 1.

TABLE IX, PART B
Angrist and Pischke (2009): Robustness Tests

Comparison set	K	$\bar{\hat{\theta}}$	$R(\hat{\boldsymbol{\theta}})$	$R^*_{.50}$	$R^*_{.95}$	p_R	p_τ	$R^*_{.95}/\lvert\bar{\hat{\theta}}\rvert$
Within column, TSLS and LIML								
Col 1	2	0.1058	0.0012	0.0013	0.0066	0.3390	0.0006	0.0626
Col 3	2	0.0910	0.0038	0.0067	0.0221	0.5905	0.0160	0.2433
Col 4	2	0.0786	0.0052	0.0347	0.4025	0.9099	0.5472	5.1229
Col 5	2	0.0996	0.0136	0.0202	0.0416	0.3060	0.2109	0.4175
Col 6	2	0.1003	0.0193	0.0268	0.0659	0.3461	0.4330	0.6571
Within column, OLS and TSLS								
Col 1	2	0.0882	0.0342	0.0336	0.0686	0.1007	0.5668	0.7783
Col 3	2	0.0801	0.0180	0.0159	0.0419	0.2669	0.1776	0.5233
Col 4	2	0.0735	0.0049	0.0177	0.0534	0.8561	0.2540	0.7263
Col 5	2	0.0801	0.0255	0.0181	0.0335	0.0252	0.0994	0.4189
Col 6	2	0.0790	0.0233	0.0152	0.0322	0.0711	0.0748	0.4081
Within column, OLS and LIML								
Col 1	2	0.0888	0.0354	0.0355	0.0727	0.1091	0.6022	0.8188
Col 3	2	0.0820	0.0218	0.0232	0.0625	0.3473	0.3747	0.7624
Col 4	2	0.0761	0.0101	0.0584	0.4297	0.9039	0.7186	5.6461
Col 5	2	0.0869	0.0391	0.0388	0.0730	0.0569	0.6793	0.8400
Col 6	2	0.0887	0.0427	0.0431	0.0940	0.1522	0.6811	1.0600
Within estimator, across columns								
All OLS	5	0.0696	0.0038	0.0038	0.0039	0.0001	0.0001	0.0561
All TSLS	5	0.0908	0.0293	0.0388	0.0864	0.5287	0.6533	0.9516
All LIML	5	0.0994	0.0288	0.0774	0.4519	0.8751	0.8909	4.5472
All TSLS and LIML estimates (set shown to survey respondents)								
All TSLS & LIML	10	0.0951	0.0340	0.0830	0.4597	0.8549	0.9594	4.8350

Notes: Robustness statistics for estimates in Table IX, Part A. $\bar{\hat{\theta}}$ is the mean of the full-sample point estimates and $R(\hat{\boldsymbol{\theta}})$ is their observed range in the comparison set. $R^*_{.50}$ and $R^*_{.95}$ are the median and 95th percentile of the unrecentered bootstrap range distribution. p_R is the add-one-adjusted p-value for the null hypothesis $H_0\colon \Delta = 0$, calculated from the recentered bootstrap distribution. At the literature-based tolerance $\tau = 0.03$, p_τ is the add-one-adjusted p-value for $H_0\colon \Delta \geq \tau$, calculated from the unrecentered bootstrap distribution. Results are based on $B = 9{,}999$ bootstrap resamples of individuals with replacement. Weak identification places the low-F columns outside Assumption 1, so the statistics there are descriptive and formal inference applies only where the first stage is strong. Within-column TSLS–LIML equality tests are conservative under strong identification because the pairs share an influence function (Proposition 4).

Although the hypothesis of equality between TSLS and LIML is not rejected in any of the five columns ($p_R \geq 0.306$), the test sharply distinguishes among them. Column 1 rejects $H_0\colon \Delta \geq 0.03$, with an equivalence p-value below 0.001, formally establishing agreement within three percentage points. Column 3 also yields a small equivalence p-value of 0.016, but its first-stage F statistic of 4.91 places it outside the secure range of the regular asymptotic approximation, and this result, like those in Columns 4, 5, and 6, should be interpreted descriptively. Those columns do not establish equivalence at this tolerance, with $p_\tau = 0.547$, 0.211, and 0.433, respectively. AP's advice to "be happy" is therefore formally supported in column 1 and descriptively supported in column 3, but not in columns 4, 5, or 6, even though equality is not rejected anywhere. The largest equivalence bound is not in columns 5 or 6, which AP

describe as worrisome, but in column 4, where $R^*_{.95} = 0.403$. The within-column bounds line up inversely with the first-stage F statistics, so the large joint bound for the survey comparison set is driven primarily by the weakly identified column 4.[32]

The same tolerance also gives the OLS–IV comparisons a substantive interpretation. None of the within-column OLS–TSLS or OLS–LIML pairs establishes equivalence within three percentage points at the 5 percent level. The OLS–TSLS differences in columns 5 and 6 are closest, with $p_\tau = 0.099$ and 0.075, while the OLS–LIML comparisons remain further from equivalence. The table therefore separates two claims that are often blurred together. TSLS and LIML may agree with one another even when the IV estimates are not equivalent to OLS.

As noted above, the bootstrap-based statistics in Table IX, Part B, should be interpreted with caution, because the pairs bootstrap is generally not valid under weak identification. An independent check on whether the dispersion summarized by $R^*_{.95}$ reflects genuine joint imprecision rather than a bootstrap artifact is to compute weak-identification-robust confidence sets for β in each column. Table X reports Anderson–Rubin (AR) (Anderson and Rubin, 1949) and conditional likelihood ratio (CLR) (Moreira, 2003) confidence bands. In the linear IV setting, these procedures yield confidence sets with correct coverage even when instruments are weak. The pattern across columns corroborates the descriptive reading of $R^*_{.95}$ in Table IX, Part B. In columns 1, 3, 5, and 6, the identification-robust regions are wider than the TSLS and LIML confidence intervals, but qualitatively consistent with them, and informative in the sense of not including zero. In column 4, however, the CLR 95 percent region is $[-0.181, 0.389]$ and the AR region is $[-0.296, 0.555]$. These bounds suggest the returns to schooling could range from substantially negative to substantially positive, and each is roughly five to seven times wider than the TSLS confidence interval. From these intervals, which respect weak-identification asymptotics, it is difficult to draw meaningful conclusions about the return to education. The marginal agreement between $\hat{\beta}_{\text{TSLS}} = 0.076$ and $\hat{\beta}_{\text{LIML}} = 0.081$ that AP's heuristic relies on is an artifact of the point estimates landing close together in this sample, not a feature of the data's evidence about the return to education or the lack of finite-sample bias in TSLS.

TABLE X
Conventional and Weak-Identification-Robust 95% Confidence Regions for β

AP Column	TSLS CI	LIML CI	CLR	AR
(1)	[0.066, 0.145]	[0.066, 0.147]	[0.069, 0.149]	[0.066, 0.152]
(3)	[0.058, 0.121]	[0.058, 0.128]	[0.056, 0.133]	[0.014, 0.178]
(4)	[0.019, 0.133]	[0.000, 0.162]	[−0.181, 0.389]	[−0.296, 0.555]
(5)	[0.075, 0.111]	[0.084, 0.129]	[0.078, 0.136]	[0.024, 0.203]
(6)	[0.070, 0.112]	[0.081, 0.139]	[0.072, 0.152]	[0.005, 0.248]

Notes: CLR is the conditional likelihood ratio test of Moreira (2003); AR is the Anderson and Rubin (1949) test. Both are weak-identification-robust. In the linear IV setting they yield confidence regions with correct coverage even when instruments are weak. Confidence regions are inverted from the corresponding test statistics on a grid for $\beta \in [-5, 5]$ with 2,500 points. Columns refer to AP Table 4.6.2. First-stage F values are reported in Table IX, Part A.

[32]The ordering is what the k-class expansion of Appendix A.VI predicts. The TSLS–LIML contrast scales with the inverse concentration per instrument, roughly $1/(F-1)$. Column 4, with a first-stage F of 1.61, is the weakest specification in the table.

The bootstrap distribution underlying $R^*_{.95} = 0.403$ in column 4 summarizes the same fundamental imprecision that CLR and AR reveal through different methods. CLR and AR characterize what the data say about the underlying parameter using the structural model. The equivalence bound instead characterizes the dispersion of the TSLS and LIML estimates while treating the estimators as a black box. The procedures answer different questions, but they reach the same substantive conclusion. Column 4 is fundamentally uninformative despite the proximity of the two point estimates.

This illustration supports a general conclusion about checking TSLS against LIML. The comparison is informative in one direction only. A large divergence is a genuine warning because, under many-instrument asymptotics, the estimators can have different probability limits (Bekker, 1994; Chao and Swanson, 2005; Hahn and Hausman, 2002). The converse does not hold. With a fixed number of instruments, TSLS and LIML share a probability limit at every instrument strength. Under strong identification their closeness is largely mechanical, while under weak identification two realized estimates may be close by coincidence even when both are highly dispersed. In this table, column 1 illustrates the strong-identification case, column 3 is intermediate, and columns 4–6 illustrate the weak-identification case. Comparing TSLS and LIML is therefore mechanically reassuring exactly where reassurance is not needed and least trustworthy exactly where it is sought. When the regular asymptotic approximation is unreliable, the minimum equivalence bound should be interpreted descriptively and alongside identification-robust confidence regions.

VI Conclusion

I start from the simple observation that comparing coefficients across specifications and claiming robustness is a widespread practice in applied economics. These comparisons may be informative, but casual inspection does not provide a statistical basis for the resulting robustness claim. Researchers are often unclear whether the effects they are estimating across specifications target the same object, implicitly treating the underlying θ_k as equal. Casual "eyeballing" does not account for the joint sampling distribution of estimates generated from the same data, and a lack of detectable difference between estimates is not the same as affirmative evidence of equivalence.

In this paper I have proposed a statistical framework that treats robustness as an inferential claim about multiple estimates, using the bootstrap to capture their joint sampling distribution. The fundamental inferential object is the minimum equivalence bound, $R^*_{.95}$, which directly measures how much tolerance the specifications require to establish equivalence and, under the conditions developed in the paper, covers the population range at or above its nominal level. The framework also provides a range-based test of equality for detecting disagreement across estimates that simulations show has correct size and power under a variety of data-generating processes common in empirical research. The statistics are easy to calculate with the R and Stata software that accompanies the paper once bootstrap estimates of the quantities in question have been generated. Reporting both costs little and communicates a great deal.

I applied the framework to five well-known empirical studies spanning commonly used identification strategies: regression discontinuity (Lee, 2008), a randomized experiment (Karlan and List, 2007), staggered-adoption difference-in-differences (Autor, 2003), spatial difference-in-differences (Dube et al., 2010), and instrumental variables (Angrist and Pischke, 2009). In most cases, formal equality tests and casual inspection would support the authors' robustness claims. The minimum equivalence bound tells a more complete story. For Lee (2008), $R^*_{.95}$ is small and the affirmative claim of equivalence is well supported. For Autor (2003) and Angrist and Pischke (2009), apparent robustness is not supported by the data. In the familiar weakly identified returns-to-education example in which Angrist and Pischke suggest that we "be happy" because TSLS and LIML look similar, $R^*_{.95}$ is several times larger than the estimates themselves. The framework supports some of the published robustness claims, but not all of them.

The survey responses from CEPR and NBER affiliates show that practitioners' informal assessments mirror the framework. Professional judgment and the framework agree at the extremes. Among respondents shown Lee (2008), 85 percent rated the results very or extremely robust, and the framework agrees. The Dube et al. (2010) employment estimates, which disagree on sign in the published table, drew the opposite response, with only 16 percent rating them moderately robust or above. The agreement extends beyond these two cases. In addition to the mean ratings shown with the histograms, I calculate Borda scores (Saari, 1994) to aggregate respondents' assessments.[33] Across the six tables, respondents' assessments are strongly correlated with both dispersion statistics (the robustness ratio and the normalized observed range, $R(\hat{\boldsymbol{\theta}})/|\bar{\hat{\theta}}|$). The mean rating and mean Borda score rank the tables identically, so each robustness statistic has the same Spearman rank correlation with both survey measures. The robustness ratio has a Spearman rank correlation of -1.00, while the normalized observed range has a Spearman rank correlation of -0.714.[34] The distinction between the two dispersion statistics is clearest where their orderings differ. Autor (2003) has the second-tightest normalized observed range among the six tables but the fourth-highest mean rating, the same position assigned by the robustness ratio. By contrast, the equality p-value is only weakly related to the assessments, with a Spearman correlation of 0.257.[35] Respondents' assessments therefore track the dispersion of the estimates rather than the statistical evidence regarding equality. Where the normalized observed range and the robustness ratio rank the tables differently, the assessments follow the ratio.

The agreement lies in the ordering rather than the magnitudes. For the four applications between the extremes, 84 to 92 percent of respondents rated the results at least moderately robust, even though the corresponding robustness ratios span nearly an order of magnitude. Respondents rated Autor (2003) and the Angrist and Pischke (2009) table nearly as robust as

[33]For each respondent, the examples are ranked according to the reported robustness rating. If a respondent assessed J examples, the lowest-ranked example receives a Borda score of 0, the highest-ranked example receives $J-1$, and tied examples receive the average of the scores for the positions they occupy. Missing responses are omitted from that respondent's ranking. The reported Borda score for each example is the average across respondents who rated it.

[34]The Pearson correlations of the robustness ratio with the mean rating and mean Borda score are -0.976 and -0.937, respectively. The corresponding correlations for the normalized observed range are -0.964 and -0.907.

[35]The equality p-value has Pearson correlations with the mean rating and mean Borda score are 0.356 and 0.341, respectively.

Karlan and List (2007), whereas the framework classifies the full Karlan and List comparison as weakly robust and the Autor and Angrist and Pischke comparison sets as not robust. This difference does not imply that professional judgment is poorly calibrated. The respondents recovered the framework's ordering from the point estimates and standard errors that the tables display, but the tables do not report the joint sampling distribution needed to determine whether the estimates are equivalent. The framework therefore does not replace professional judgment, but provides a formal inferential foundation for current informal practice.

Pre-analysis plans have become common for experiments, but these plans rarely specify what counts as "robust" across specifications and rarely extend to non-experimental analyses. The framework provides a natural way to bring pre-analysis commitments to robustness claims. A researcher can specify a tolerance τ on the basis of a literature review or theoretical bounds, commit to that tolerance before seeing the data, and then report $R^*_{.95}$ to see whether the data support a claim of robustness. The difficulty of establishing τ for a particular question is not a weakness of the framework but a clarification of a choice that researchers already make implicitly every time they eyeball a set of specifications and call them robust. Declaring which specifications will be estimated and what standard will be used to judge them equivalent also eliminates the temptation to search over subsets for a favorable $R^*_{.95}$.

A researcher whose results are not robust will ask what to do next. When $R^*_{.95}$ is large relative to the estimates, the appropriate response is not simply to report that the results are "not robust," but to investigate the source of the variation. Researchers should identify which specification choices account for the spread, determine whether those choices alter the underlying estimand or instead reveal sensitivity to modeling assumptions, and explain whether some specifications are more credible than others on substantive or design-based grounds. If the estimates correspond to different well-defined estimands, the variation may itself be informative, reflecting treatment-effect heterogeneity or differences in the populations, comparisons, or margins being studied rather than mere sensitivity to specification choice. If the specifications in the comparison set are intended to estimate the same parameter, a large $R^*_{.95}$ indicates that the conclusion depends materially on the choices underlying those specifications and that a single preferred estimate requires additional justification. In either case, the framework is best viewed as a diagnostic that quantifies and makes transparent the degree of robustness, directs attention toward the assumptions generating instability, and encourages researchers to report the range of substantively plausible conclusions supported by the data.

The credibility revolution brought discipline to identification and inference about individual parameters. The same discipline should apply to the robustness claims about the similarity of results across different specifications that appear in nearly every empirical paper. The question should no longer be whether the results "look similar," but whether the data support a well-founded claim of equivalence.

References

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6), 1537–1557.

Abrevaya, J., & Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*, 73(4), 1175–1204.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, 113(1), 151–184.

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80(1), 123–129.

Anderson, T. W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 20(1), 46–63.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2), 399–405.

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., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4), 979–1014.

Angrist, J. D., & Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2), 225–235.

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1), 57–67.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ansolabehere, S., & Snyder, J. M., Jr. (2002). The incumbency advantage in U.S. elections: An analysis of state and federal offices, 1942–2000. *Election Law Journal*, 1(3), 315–338.

Autor, D. H. (2003). Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of Labor Economics*, 21(1), 1–42.

Babu, G. J. (1984). Bootstrapping statistics with linear combinations of chi-squares as weak limit. *Sankhyā: The Indian Journal of Statistics, Series A*, 46(1), 85–93.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, 62(3), 657–681.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3), 457–468.

Bickel, P. J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6), 1196–1217.

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(430), 443–450.

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Vol. 3A, 1801–1863. Amsterdam: Elsevier.

Card, D., Fenizia, A., & Silver, D. (2023). The health impacts of hospital delivery practices. *American Economic Journal: Economic Policy*, 15(2), 42–81.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772–793.

Casella, G. & R. L. Berger (2002). *Statistical Inference*. Duxbury.

Chao, J. C., & Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5), 1673–1692.

Cinelli, C., and C. Hazlett (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B*, 82(1), 39–67.

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, 100(5), 1261–1293.

Cox, D. R. (1961). Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 105–123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, 24(2), 406–424.

Cox, G. W., & Katz, J. N. (1996). Why did the incumbency advantage in U.S. House elections grow? *American Journal of Political Science*, 40(2), 478–497.

Davidson, R. & MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–2996.

Dube, A., Lester, T. W., & Reich, M. (2010). Minimum wage effects across state borders: Estimates using contiguous counties. *Review of Economics and Statistics*, 92(4), 945–964.

Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute*, 22(1/3), 23–32.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28(1), 181–187.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1), 1–26.

Fang, Z., & Santos, A. (2019). Inference on directionally differentiable functions. *Review of Economic Studies*, 86(1), 377–412.

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, 34(2), 509–543.

Gelman, A., & Huang, Z. (2008). Estimating incumbency advantage and its variation, as an example of a before–after study. *Journal of the American Statistical Association*, 103(482), 437–446.

Gonçalves, S., & White, H. (2004). Maximum likelihood and the bootstrap for nonlinear dynamic models. *Journal of Econometrics*, 119(1), 199–219.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Hahn, J., & Hausman, J. (2002). A new specification test for the validity of instrumental variables. *Econometrica*, 70(1), 163–189.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6), 1251–1271.

Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman & E. Leamer (Eds.), *Handbook of Econometrics* (Vol. 5, pp. 3159–3228). Elsevier.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford University Press.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467–475.

Jaeger, D. A. (2026). Taming of the Zoo: A Practical Practitioner's Guide to Staggered-Adoption Difference-in-Differences. Working paper in preparation, University of St Andrews.

Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, 97(5), 1774–1793.

Kim, J., & Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, 18(1), 191–219.

King, G. (1991). Constituency service and incumbency advantage. *British Journal of Political Science*, 21(1), 119–128.

Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. Working paper, Cowles Foundation, Yale University.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73(1), 31–43.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, 142(2), 675–697.

Lu, X., & White, H. (2014). Robustness checks and robustness tests in applied economics. *Journal of Econometrics*, 178, 194–206.

Mariano, R. S. (1982). Analytical small-sample distribution theory in econometrics: The simultaneous-equations case. *International Economic Review*, 23(3), 503–533.

Masten, M. A. and Poirier, A. (2026). The effect of omitted variables on the sign of regression coefficients. *American Economic Review*, 116(7), 2685–2710.

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89(5), 2143–2188.

Mizon, G. E., & Richard, J.-F. (1986). The encompassing principle and its application to testing non-nested hypotheses. *Econometrica*, 54(3), 657–678.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4), 1027–1048.

Moreira, M. J., J. R. Porter, and G. A. Suarez (2009). Bootstrap validity for the score test when instruments may be weak. *Journal of Econometrics*, 149(2009), 52–64.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, 27(4), 575–595.

Neumark, D., & Wascher, W. (1992). Employment effects of minimum and subminimum wages: Panel data on state minimum wage laws. *Industrial and Labor Relations Review*, 46(1), 55–81.

Neumark, D., & Wascher, W. (2007). Minimum wages and employment. *Foundations and Trends in Microeconomics*, 3(1–2), 1–182.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of Econometrics*, Vol. 4 (pp. 2111–2245). Elsevier.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2), 187–204.

Pei, Z., Pischke, J.-S., and Schwandt, H. (2019). Poorly measured confounders are more useful on the left than on the right. *Journal of Business & Economic Statistics*, 37(2), 205–216.

Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling*. Springer.

Prallon, B. (2026). How robust are robustness checks? Working paper, Cornell University.

Saari, D. G. (1994). *Geometry of Voting*. Springer, New York.

Schaffer, M. E. (2024). Null hypothesis misspecification testing revisited: How (not) to test orthogonality conditions. Invited paper, Eighteenth International Conference of The Thailand Econometrics Society, Chiangmai University, January 2025.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4, 1208–1214.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics*, 9(6), 1187–1195.

Słoczyński, T. (2022). Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *Review of Economics and Statistics*, 104(3), 501–509.

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–586.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge University Press.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2), 99–114.

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

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), 307–333.

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

Wu, D.-M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica*, 41(4), 733–750.

Appendices

A Asymptotic Validity

This appendix states the conditions under which the equivalence bound and the equality tests are valid, proves their limiting properties, and establishes power against fixed alternatives. It also characterizes the tie configurations under which the equivalence bound becomes non-regular and gives conditions under which the resulting bound remains conservative for the population range of probability limits. I close with the conditions under which the assumptions can fail.

A.I Notation and conditions

The following notation is used throughout.

$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_K)'$**:** the vector of full-sample estimates in the comparison set.

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$**:** the vector of probability limits of the full-sample estimates in the comparison set, with $\hat{\theta}_k \xrightarrow{p} \theta_k$.

$g(\mathbf{x}) = \max_{1 \le k \le K} x_k - \min_{1 \le k \le K} x_k$**:** the range function, mapping $\mathbb{R}^K$ to $\mathbb{R}_{\ge 0}$.

$\Delta \equiv g(\boldsymbol{\theta})$**:** the range of probability limits in the comparison set.

$R(\hat{\boldsymbol{\theta}}) = g(\hat{\boldsymbol{\theta}})$**:** the observed range of the full-sample estimates in the comparison set.

$\hat{\boldsymbol{\theta}}^{(b)} = (\hat{\theta}_1^{(b)}, \ldots, \hat{\theta}_K^{(b)})'$**:** the estimates on bootstrap replication b.

$\dot{\theta}_k^{(b)} = \hat{\theta}_k^{(b)} - \hat{\theta}_k$**:** the recentered bootstrap estimate.

P^***,** P**:** probability under one bootstrap replication conditional on the data, and the corresponding unconditional probability over the sample and the B bootstrap replications together.

$\hat{q}_p^*$**:** the p-quantile of $\sqrt{n}(R(\hat{\boldsymbol{\theta}}^{(b)}) - R(\hat{\boldsymbol{\theta}}))$ conditional on the data.

$\mathcal{M}^+(\boldsymbol{\theta}) = \{k : \theta_k = \max_{1 \le \ell \le K} \theta_\ell\}$**,** $\mathcal{M}^-(\boldsymbol{\theta}) = \{j : \theta_j = \min_{1 \le \ell \le K} \theta_\ell\}$**:** the population active sets, the coordinates attaining the maximum and the minimum.

A.II Assumptions and definition of statistics

The regular case maintains two assumptions:

Assumption 1 (Joint asymptotic normality)**.** $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$ for some positive-definite Σ.

Assumption 2 (Bootstrap consistency)**.** $\sqrt{n}(\hat{\boldsymbol{\theta}}^{(b)} - \hat{\boldsymbol{\theta}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$ in probability conditional on the data.

Assumption 2 is consistency for the joint distribution, including all cross-specification covariances, not merely consistency of each marginal.

The theory distinguishes each reported statistic from its ideal-bootstrap counterpart, the object defined by the exact conditional distribution P^* that the B bootstrap estimates approximate. Each reported statistic converges to its ideal counterpart (denoted with an ∞ superscript) as $B \to \infty$.

The **ideal range equality p-value** is the conditional tail probability

$$p_R^{\infty} \;=\; P^*\Big(R(\dot{\boldsymbol{\theta}}^{(b)}) \geq R(\hat{\boldsymbol{\theta}})\Big),$$

and the reported p_R is its add-one estimate from the B bootstrap estimates,

$$p_R \;=\; \frac{1 + \sum_{b=1}^{B} \mathbf{1}\Big\{R(\dot{\boldsymbol{\theta}}^{(b)}) \geq R(\hat{\boldsymbol{\theta}})\Big\}}{B+1}.$$

The **ideal minimum equivalence bound** $R^{*\infty}_{1-\alpha}$ is the $(1-\alpha)$ quantile of the unrecentered bootstrap range $R(\hat{\boldsymbol{\theta}}^{(b)})$ under the exact conditional distribution:

$$R^{*\infty}_{1-\alpha} \;=\; \inf\Big\{t : P^*\Big(R(\hat{\boldsymbol{\theta}}^{(b)}) \leq t\Big) \geq 1-\alpha\Big\} \;=\; R(\hat{\boldsymbol{\theta}}) + n^{-1/2}\hat{q}^*_{1-\alpha}.$$

The second equality expresses the bound as the observed range plus an allowance for sampling variation, taken from the bootstrap and shrinking at the $n^{-1/2}$ rate. In the regular case, a unique maximum and a unique minimum, the median of the bootstrap range distribution differs from the observed range by $o_p(n^{-1/2})$, and the allowance is, to first order, the distance from that median to the $(1-\alpha)$ quantile. Both pieces converge: $R^*_{1-\alpha} \xrightarrow{p} \Delta$ under Assumptions 1 and 2, whatever the tie configuration, so the bound is a consistent and conservatively biased estimator of the population range.

The reported $R^*_{1-\alpha}$ replaces P^* by the empirical distribution of the B bootstrap ranges. With $R_{(1)} \leq \cdots \leq R_{(B)}$ the ordered values of $R(\hat{\boldsymbol{\theta}}^{(b)})$,

$$R^*_{1-\alpha} \;=\; \inf\left\{t : B^{-1}\sum_{b=1}^{B} \mathbf{1}\Big(R(\hat{\boldsymbol{\theta}}^{(b)}) \leq t\Big) \geq 1-\alpha\right\} \;=\; R_{(\lceil (1-\alpha)B \rceil)},$$

is the corresponding type-1 order statistic of the B unrecentered bootstrap range estimates. The propositions are proved for the ideal objects and carried to the reported statistics by the rank lemma that opens the proofs.

A.III Proofs

The propositions that follow are proved for the ideal statistics, but size and coverage guarantees for the reported statistics cannot be achieved by letting B grow. This yields only an approximation, with each reported statistic approaching its ideal counterpart as $B \to \infty$ and the stated guarantees reached only in the double limit as both $n \to \infty$ and $B \to \infty$. An exact limiting distribution can instead be guaranteed by using the rank of the observed statistic among the B bootstrap replications rather than approximating a target, and the reported statistics achieve their stated size and coverage with B held fixed, which the following lemma establishes once

for both. In the proofs that use the lemma, T_0 is the scaled observed statistic and $T_1, \ldots, T_B$ are its recentered bootstrap counterparts.

Lemma 1 (Finite-B rank lemma)**.** *Let $(T_0, T_1, \ldots, T_B)$ converge jointly in distribution, as $n \to \infty$ with B fixed, to $B+1$ i.i.d. random variables whose common distribution is continuous. Then the rank of T_0 among the $B+1$ values is asymptotically uniform on $\{1, \ldots, B+1\}$, and:*

(i) the add-one p-value $p = \big(1 + \sum_{b=1}^{B} \mathbf{1}(T_b \geq T_0)\big)/(B+1)$ is asymptotically uniform on the grid $\{1/(B+1), \ldots, 1\}$, and $\lim_{n\to\infty} P(p \leq \alpha) = \lfloor (B+1)\alpha \rfloor/(B+1)$, which equals α at every level for which $(B+1)\alpha$ is an integer and lies below α by less than $(B+1)^{-1}$ otherwise;

(ii) $\lim_{n\to\infty} P\big(T_0 \leq T_{(\lceil (1-\alpha)B \rceil)}\big) = \lceil (1-\alpha)B \rceil/(B+1)$, where $T_{(k)}$ denotes the k-th order statistic of $(T_1, \ldots, T_B)$, which equals $1-\alpha$ under the same grid condition and differs from it by less than $(B+1)^{-1}$ otherwise.

Proof. Write $(T_0^\infty, \ldots, T_B^\infty)$ for the limit vector, whose coordinates are i.i.d. with a continuous common distribution. Ties among its coordinates are a probability-zero event, so with probability one the limit vector has pairwise-distinct coordinates, where the rank of the first coordinate is a continuous, indeed locally constant, function of the vector. The continuous mapping theorem (van der Vaart, 1998, pp. 8–9) therefore gives convergence of the rank of T_0 to the rank of T_0^∞ among $B+1$ i.i.d. continuous random variables, which is uniform on $\{1, \ldots, B+1\}$ by exchangeability (Dwass, 1957; Davison and Hinkley, 1997). Parts (i) and (ii) restate rank events. The event $p \leq \alpha$ holds exactly when the rank of T_0 from the top is at most $\lfloor (B+1)\alpha \rfloor$, the event $T_0 \leq T_{(\lceil (1-\alpha)B \rceil)}$ holds exactly when the rank of T_0 from the bottom is at most $\lceil (1-\alpha)B \rceil$, and a uniform rank assigns each event its share of the $B+1$ positions. □

Lemma 2 (Duality of the bound and the equivalence p-value)**.** *Fix $\tau > 0$ and $\alpha \in (0,1)$, and let*

$$p_\tau \;=\; \frac{1 + \sum_{b=1}^{B} \mathbf{1}\big\{R(\hat{\boldsymbol{\theta}}^{(b)}) \geq \tau\big\}}{B+1}$$

be the add-one equivalence p-value, the share of unrecentered bootstrap ranges at or above the tolerance. Suppose $\alpha(B+1)$ is an integer and no bootstrap range equals τ exactly, an event of probability zero when τ is specified in advance and the range distribution is continuous. Then

$$R^*_{1-\alpha} \;\leq\; \tau \quad\Longleftrightarrow\quad p_\tau \;\leq\; \alpha.$$

Proof. Let N be the number of bootstrap replications with $R(\hat{\boldsymbol{\theta}}^{(b)}) \geq \tau$, and write $m = \lceil (1-\alpha)B \rceil$, so that $R^*_{1-\alpha}$ is the m-th order statistic of the B unrecentered ranges. The bound satisfies $R^*_{1-\alpha} \leq \tau$ exactly when at least m of the ranges are at most τ. Because no range equals τ, this holds exactly when $N \leq B - m$. Since $0 < \alpha < 1$, we have $B - m = B - \lceil B - \alpha B \rceil = \lfloor \alpha B \rfloor$. The p-value satisfies $p_\tau \leq \alpha$ exactly when $N \leq \alpha(B+1) - 1$. When $\alpha(B+1)$ is an integer, $\alpha(B+1) - 1 = \lfloor \alpha B + \alpha \rfloor - 1 = \lfloor \alpha B \rfloor$. The two conditions are therefore the same condition, $N \leq \lfloor \alpha B \rfloor$. □

The convention maintained throughout the paper, $B = 9{,}999$ with $\alpha = 0.05$ or 0.10, makes $\alpha(B+1)$ an integer by construction, so the conventional significance levels lie exactly on the

attainable Monte Carlo grid. The equivalence test and the minimum equivalence bound are therefore exactly dual at those levels. Reporting $R^*_{1-\alpha}$ and reporting the family of equivalence tests are the same act. The bound is the cutoff for the family of level-α equivalence tests, in the sense that the test rejects at every $\tau > R^*_{1-\alpha}$ and at no $\tau < R^*_{1-\alpha}$. At $\tau = R^*_{1-\alpha}$, the result depends only on the stated treatment of equality in the tail count.

A.III.a Size for p_R

Proposition 1. *When Assumptions 1 and 2 hold, under the null $\boldsymbol{\theta} = \theta\mathbf{1}$, $P(p_R \le \alpha) \to \alpha$ at every level α for which $(B+1)\alpha$ is an integer.*

Proof of Proposition 1. The equality test compares the sample range $R(\hat{\boldsymbol{\theta}})$ to the bootstrap distribution of the range of the recentered vector, $R(\dot{\boldsymbol{\theta}}^{(b)})$. Recentering imposes the equality null $\boldsymbol{\theta} = \theta\mathbf{1}$ on the bootstrap estimates by construction, subtracting each estimate's own full-sample value so that the resampled ranges measure sampling variation around 0 rather than around the observed range. Under that null, $g(\boldsymbol{\theta}) = 0$, and the continuous mapping theorem (van der Vaart, 1998, pp. 8-9) applied to Assumption 1 with the continuous range function $g(\mathbf{x})$, gives

$$\sqrt{n}\, R(\hat{\boldsymbol{\theta}}) \;=\; \sqrt{n}\, g(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \;\xrightarrow{d}\; g(\mathbf{Z}), \qquad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{1}$$

using the homogeneity and shift-invariance of $g(\cdot)$. For the bootstrap analogue, the recentered vector is $\dot{\boldsymbol{\theta}}^{(b)} = \hat{\boldsymbol{\theta}}^{(b)} - \hat{\boldsymbol{\theta}}$, so Assumption 2 and the continuous mapping theorem give $\sqrt{n}\, R(\dot{\boldsymbol{\theta}}^{(b)}) \xrightarrow{d} g(\mathbf{Z})$ conditional on the data, with the same limit as (1).

Let $F(\cdot)$ denote the cumulative distribution function of $g(\mathbf{Z})$. F is continuous because $g(\mathbf{Z}) = t$ requires one of the finitely many pairwise differences $Z_k - Z_j$ to equal t exactly, and under positive-definite Σ each difference is a normal random variable with positive variance, hence equal to t with probability zero. The ideal p-value p^∞_R is the conditional probability that $\sqrt{n}\, R(\dot{\boldsymbol{\theta}}^{(b)})$ is at least $\sqrt{n}\, R(\hat{\boldsymbol{\theta}})$, an evaluation at a random point that depends on the same data as the conditional distribution. The uniform convergence of distribution functions to a continuous limit (van der Vaart, 1998, p. 12) implies that the conditional convergence holds uniformly over evaluation points, so $p^\infty_R = 1 - F(\sqrt{n}\, R(\hat{\boldsymbol{\theta}})) + o_p(1)$. By (1) and the continuous mapping theorem this converges in distribution to $1 - F(g(\mathbf{Z}))$, which is Uniform$(0,1)$ by the probability integral transform (Casella and Berger, 2002, p. 54). Conditional on the data the B bootstrap estimates are i.i.d., each drawn from the same conditional distribution P^*, so bootstrap consistency gives joint convergence of $(\sqrt{n}\, R(\hat{\boldsymbol{\theta}}), \sqrt{n}\, R(\dot{\boldsymbol{\theta}}^{(1)}), \ldots, \sqrt{n}\, R(\dot{\boldsymbol{\theta}}^{(B)}))$ to $B+1$ independent copies of the continuous limit $g(\mathbf{Z})$. Applying Lemma 1(i) proves $P(p_R \le \alpha) \to \alpha$ at every level α for which $(B+1)\alpha$ is an integer. □

A.III.b Coverage of $R^*_{1-\alpha}$

Proposition 2. *Under Assumptions 1 and 2:*

(i) If all K specifications share a probability limit, so that $\Delta = 0$, then $\lim_{n\to\infty} P(\Delta \le R^*_{1-\alpha}) = 1$.

(ii) If the largest and the smallest probability limits are each attained by exactly one specification, so that $\mathcal{M}^+(\boldsymbol{\theta})$ and $\mathcal{M}^-(\boldsymbol{\theta})$ are singletons, then $\lim_{n\to\infty} P(\Delta \le R^*_{1-\alpha}) = 1-\alpha$ at every level α for which $(B+1)\alpha$ is an integer.

Proof of Proposition 2. Part (i). $\Delta = 0$ and $R(\hat{\boldsymbol{\theta}}^{(b)}) \ge 0$ on every bootstrap replication, so $R^*_{1-\alpha} \ge 0$ and the event $\Delta \le R^*_{1-\alpha}$ holds trivially with probability one for every n.

Part (ii). Let $\mathcal{M}^+(\boldsymbol{\theta}) = \{k^*\}$, $\mathcal{M}^-(\boldsymbol{\theta}) = \{j^*\}$ be the active sets indexing the largest and smallest probability limits. Since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, sample active sets coincide with the population active sets with probability approaching one, on which event $g(\mathbf{x}) = x_{k^*} - x_{j^*}$ in a neighborhood of both $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$. When the maximum and minimum are each attained by a single, well-separated coordinate, the range is locally the difference between the k^*-th and j^*-th coordinates, a smooth function of the estimates with no complication from the maximum or minimum operators, so the ordinary delta method (Davidson and MacKinnon, 2004) applied to Assumption 1 yields

$$\sqrt{n}(R(\hat{\boldsymbol{\theta}}) - \Delta) \xrightarrow{d} (\mathbf{e}_{k^*} - \mathbf{e}_{j^*})'\mathbf{Z} \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

with $\mathbf{e}_k$ the k-th standard basis vector in $\mathbb{R}^K$ and $\sigma^2 = \Sigma_{k^*k^*} + \Sigma_{j^*j^*} - 2\Sigma_{k^*j^*} > 0$. Because $\hat{\boldsymbol{\theta}}^{(b)} - \hat{\boldsymbol{\theta}} = o_{p^*}(1)$ conditional on the data, the bootstrap estimates lie in the same neighborhood with conditional probability approaching one, so $R(\hat{\boldsymbol{\theta}}^{(b)}) - R(\hat{\boldsymbol{\theta}})$ reduces to the same linear contrast evaluated at $\hat{\boldsymbol{\theta}}^{(b)} - \hat{\boldsymbol{\theta}}$. Assumption 2 applied to that contrast gives $\sqrt{n}(R(\hat{\boldsymbol{\theta}}^{(b)}) - R(\hat{\boldsymbol{\theta}})) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ conditional on the data. The conditional limit therefore coincides with the limit in (2). By the location-scale identity $R^{*\infty}_{1-\alpha} = R(\hat{\boldsymbol{\theta}}) + n^{-1/2}\hat{q}^*_{1-\alpha}$ and continuity of the Gaussian cumulative distribution function, $\hat{q}^*_{1-\alpha} \xrightarrow{p} \sigma z_{1-\alpha}$,

$$P(\Delta \le R^{*\infty}_{1-\alpha}) = P(\sqrt{n}(R(\hat{\boldsymbol{\theta}}) - \Delta) \ge -\hat{q}^*_{1-\alpha}) \longrightarrow P(\sigma Z \ge -\sigma z_{1-\alpha}) = 1-\alpha,$$

where $Z \sim \mathcal{N}(0,1)$. The deviations $\sqrt{n}(R(\hat{\boldsymbol{\theta}}) - \Delta)$ and $\sqrt{n}(R(\hat{\boldsymbol{\theta}}^{(b)}) - R(\hat{\boldsymbol{\theta}}))$, $b = 1, \dots, B$, converge jointly to $B+1$ i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables by the same conditional-independence factorization, and the coverage event $\Delta \le R^*_{1-\alpha}$ is the event $-T_0 \le T_{(\lceil (1-\alpha)B \rceil)}$ for $T_0 = \sqrt{n}(R(\hat{\boldsymbol{\theta}}) - \Delta)$ and $T_b = \sqrt{n}(R(\hat{\boldsymbol{\theta}}^{(b)}) - R(\hat{\boldsymbol{\theta}}))$. Because $-T_0$ shares the $\mathcal{N}(0, \sigma^2)$ limit and is independent of the bootstrap deviations, Lemma 1(ii) applied to $(-T_0, T_1, \dots, T_B)$ gives $\lim_{n\to\infty} P(\Delta \le R^*_{1-\alpha}) = 1-\alpha$ at every level α for which $(B+1)\alpha$ is an integer. □

A.IV Power against fixed alternatives

Proposition 3. *Under Assumptions 1 and 2, the range equality test is consistent against any fixed alternative with $\Delta > 0$. The ideal p-value satisfies $p^{\infty}_R \xrightarrow{p} 0$, the reported p_R converges in probability to its floor, $1/(B+1)$, and $P(p_R \le \alpha) \to 1$ for any $\alpha \ge 1/(B+1)$.*

Proof. Under $\Delta > 0$, $R(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \Delta$. Recentering gives $R(\dot{\boldsymbol{\theta}}^{(b)}) = O_{p^*}(n^{-1/2})$ conditional on the data, so $p^{\infty}_R \xrightarrow{p} 0$ and the reported $p_R \xrightarrow{p} 1/(B+1)$. □

A.V Validity of $R^*_{1-\alpha}$ in case of ties

Nothing in this section restricts the comparison sets a researcher may form. As long as a common resampling scheme consistently reproduces the joint distribution of the estimates, any comparison set is admissible, and the worst that tied or mixed configurations can do is make $R^*_{1-\alpha}$ conservative, never invalid. Sharpness, not validity, is what depends on knowing which objects the estimators target, and the results below say exactly when the bound is sharp, when it is merely conservative, and why conservatism is the only direction in which it can err.

The minimum equivalence bound, $R^*_{1-\alpha}$, depends on the range function g, which fails to be Hadamard differentiable at $\boldsymbol{\theta}$ when an active set $\mathcal{M}^+(\boldsymbol{\theta})$ or $\mathcal{M}^-(\boldsymbol{\theta})$ is not a singleton (Fang and Santos, 2019). The issue is somewhat intuitive to see. Imagine two coordinates tied at the maximum, each perturbed by small Gaussian noise. The range awards the maximum to only one of the tied pair rather than taking a linear combination of the two perturbations, and the standard bootstrap, which relies on local linear approximations, does not capture the kink at the tie.

Only ties that hold at the true $\boldsymbol{\theta}$ can produce the bootstrap non-regularity and it is essential to understand which ties are problematic and which are not. An *identity* tie holds by algebra in every sample. In a just-identified model TSLS and LIML coincide numerically, nothing in the contrast is random, and the within-pair comparison should not be reported at all. A *hypothesis* tie holds under a null and fails under the alternative. OLS and IV share a probability limit when the regressor is exogenous, the null of a Hausman test (Durbin, 1954; Wu, 1973; Hausman, 1978). Random effects and fixed effects likewise share a probability limit under the null that the effects are uncorrelated with the regressors, the comparison of Appendix B. A *maintained-assumption* tie holds conditional on a restriction the researcher imposes. OLS estimates with different sets of exogenous controls share a probability limit under homogeneous effects only when each specification remains consistent for the same target, so that omitted controls do not induce omitted-variable bias. Under heterogeneity, each set of control variables can produce a differently weighted average of the treatment effects (Angrist and Pischke, 2009, Section 3.3), yielding distinct estimands rather than a common one. The staggered-adoption estimators of the recent difference-in-differences literature likewise share a probability limit under homogeneous effects while targeting differently weighted averages under heterogeneity (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021; de Chaisemartin and D'Haultfœuille, 2020), the configuration I analyze in Jaeger (2026).

When $K = 2$ the problem cannot arise. With two specifications the active set is either a singleton or the whole comparison set, so a tie that binds some coordinates and not others does not exist. When the estimands are distinct, both active sets are singletons, the comparison is regular, and Propositions 1 and 2 apply directly. When the specifications share an estimand but not an influence function, the pair is a regular full null, regular in the sense that Propositions 1 and 2 apply directly, not that the range is smooth there. When they share an influence function, the pair is degenerate, and as discussed below both statistics err only in the conservative direction.[36] A hypothesis-tied pair occupies a covered configuration whichever way the hypothesis

[36]The influence function of an asymptotically linear estimator is the weight it places on each observation in the estimator's first-order expansion, $\sqrt{n}(\hat{\theta}_k - \theta_k) = n^{-1/2}\sum_i \psi_k(w_i) + o_p(1)$, where w_i denotes the data for

resolves. Under the null it is a regular full null, and under the alternative it is two singletons, so a hypothesis tie is harmless in any $K = 2$ comparison. This is why the Hausman test has an ordinary χ^2 limit. Every pairwise comparison is valid under Assumptions 1 and 2, and the partial-tie non-regularity is only a $K \geq 3$ phenomenon.

When $K \geq 3$, an active set can bind some coordinates and not others, and the configurations this creates are not covered by the cases discussed above. Here I establish coverage for three tied configurations, distinguished by when the tie holds. At configurations where a non-singleton active set pairs estimators with distinct influence functions and $\Delta > 0$, g is non-differentiable and the standard bootstrap is generally inconsistent (Fang and Santos, 2019). Proposition 5 in Section A.VI shows that a tie through a shared influence function escapes this failure. The reason is that influence-function-tied coordinates do not merely share a limit, they move together in every sample, collapsing to a single Gaussian coordinate in the limit, so the tie leaves no two distinct increments to take a maximum over and the kink never forms.

The binding configuration when $K \geq 3$ is a cluster tied in probability limit but not in influence function, bundled with a distinct-limit specification. Not every limit tie carries a common influence function. Differently weighted GMM estimators of the same overidentified moment conditions, for example, share a probability limit with distinct influence functions, so that cluster alone is a regular full null under Proposition 1. If these are paired with an OLS specification that targets a different population object, however, the bundle is the irregular case. For the binding case, exactness can be restored at the cost of a tuning parameter, for example the numerical-derivative bootstrap of Fang and Santos (2019, Section 4), subsampling (Politis et al., 1999), or prepivoting (Beran, 1987), but Corollary 1 shows validity is never at stake, so these remedies recover sharpness rather than coverage.

Maintaining homogeneity *enlarges* the set of at-the-truth ties, since every consistent estimator then targets the single structural parameter, while allowing heterogeneity *shrinks* it toward the algebraic identities, separating even TSLS and LIML because the overidentifying restrictions their tie requires are the restrictions heterogeneity violates. To know which specifications are tied, the researcher must first commit to what they target.

The simplest remedy for the mixed configuration is to report the internally homogeneous constituent comparisons rather than the mixed whole. In the shared-limit comparison $R^*_{1-\alpha}$ bounds finite-sample estimator disagreement, and in the distinct-limit comparisons it bounds estimand dispersion, so the decomposition forfeits nothing but the omnibus bound on the mixed set. In the OLS-TSLS-LIML comparison set the population range is exactly the OLS-versus-IV gap.

A supplementary table reports a bootstrap-versus-subsampling comparison for this mixed set: bootstrap coverage is near nominal at both sample sizes under strong identification, as Proposition 5 predicts, while subsampling is conservative except at the strongest identification.

observation i (Wooldridge, 2010, pp. 406-407). Two estimators share an influence function when they are, to first order, the same estimator: every observation moves both estimates by the same amount, their contrast is $o_p(n^{-1/2})$, and their joint asymptotic variance-covariance matrix is singular.

A.VI Influence-function-tied pairs

The influence-function tied-pairs configuration is an edge case that arises only when a comparison set contains two asymptotically equivalent estimators of the same parameter, TSLS and LIML under valid instruments and strong identification (Davidson and MacKinnon, 2004), OLS and feasible GLS under homoskedasticity, one-step and fully iterated MLE, and readers whose comparison sets vary controls, samples, or functional forms encounter no influence-function ties and can proceed to the tied-cluster case. The shared-limit comparison among influence-function-tied estimators deserves its own treatment, however, because it violates Assumption 1 outright. TSLS and LIML on the same instruments share more than a probability limit under strong identification with a fixed number of instruments. They share an influence function, because every k-class estimator with $\hat{\kappa} = 1 + O_p(n^{-1})$ has the same first-order expansion and $n(\hat{\kappa}_{\text{LIML}} - 1) = O_p(1)$ under these conditions (Nagar, 1959). The joint asymptotic covariance matrix of the pair is therefore singular, the contrast $\sqrt{n}(\hat{\theta}_1 - \hat{\theta}_2)$ is degenerate, and the relevant scale for the comparison is n^{-1} rather than $n^{-1/2}$.

Formally, let $K = 2$ and $\hat{\theta}_j = h_j(M_n)$, where M_n is a vector of sample moments with $\sqrt{n}(M_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z}_M \sim \mathcal{N}(\mathbf{0}, V)$ with V positive definite, each h_j is three times continuously differentiable in a neighborhood of $\boldsymbol{\mu}$, and the bootstrap is consistent for the moments, with $\sqrt{n}(M_n^{(b)} - M_n) \xrightarrow{d} \mathbf{Z}_M^* \sim \mathcal{N}(\mathbf{0}, V)$ in probability conditional on the data and $\mathbf{Z}_M^*$ independent of $\mathbf{Z}_M$. Suppose the pair is tied with a common influence function, so that the contrast $h \equiv h_1 - h_2$ satisfies $h(\boldsymbol{\mu}) = 0$ and $\nabla h(\boldsymbol{\mu}) = \mathbf{0}$. Define $Q(\mathbf{x}) \equiv \frac{1}{2}\mathbf{x}'\mathbf{H}\mathbf{x}$ with $\mathbf{H} \equiv \nabla^2 h(\boldsymbol{\mu}) \neq \mathbf{0}$, so that the tie does not extend to second order.

Proposition 4 (Influence-function-tied pairs)**.** *Under the conditions above:*

(i) $nR(\hat{\boldsymbol{\theta}}) \xrightarrow{d} |Q(\mathbf{Z}_M)|$, *with* $E[Q(\mathbf{Z}_M)] = \frac{1}{2}\operatorname{tr}(\mathbf{H}V)$.

(ii) Conditional on the data, $nR(\dot{\boldsymbol{\theta}}^{(b)}) \xrightarrow{d} |Q(\mathbf{Z}_M + \mathbf{Z}_M^*) - Q(\mathbf{Z}_M)|$.

(iii) p_R^{∞} *converges in distribution to* $P^*\big(|Q(\mathbf{Z}_M + \mathbf{Z}_M^*) - Q(\mathbf{Z}_M)| \geq |Q(\mathbf{Z}_M)| \mid \mathbf{Z}_M\big)$, *a non-degenerate random variable rather than a uniform.*

(iv) $R_{1-\alpha}^* = O_p(n^{-1})$ *and* $P(\Delta \leq R_{1-\alpha}^*) = 1$ *for every* n.

Convergence in parts (ii) and (iii) is convergence of the conditional distribution given the data, in probability. A fully formal treatment via stable convergence is beyond the scope of this paper.

Proof.

(i). Write $h = h_1 - h_2$, so that $\hat{\theta}_1 - \hat{\theta}_2 = h(M_n)$ with $h(\boldsymbol{\mu}) = 0$ and $\nabla h(\boldsymbol{\mu}) = \mathbf{0}$. A second-order Taylor expansion of h about $\boldsymbol{\mu}$ gives

$$n(\hat{\theta}_1 - \hat{\theta}_2) \;=\; Q\big(\sqrt{n}(M_n - \boldsymbol{\mu})\big) + o_p(1).$$

Since $\sqrt{n}(M_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z}_M$ and Q is continuous, the continuous mapping theorem gives $n(\hat{\theta}_1 - \hat{\theta}_2) \xrightarrow{d} Q(\mathbf{Z}_M)$, and with $R(\hat{\boldsymbol{\theta}}) = |\hat{\theta}_1 - \hat{\theta}_2|$ at $K = 2$, $n\,R(\hat{\boldsymbol{\theta}}) \xrightarrow{d} |Q(\mathbf{Z}_M)|$. The mean in the

statement is a property of the limiting variable, $E[Q(\mathbf{Z}_M)] = \frac{1}{2}\,E[\mathbf{Z}_M'\mathbf{H}\mathbf{Z}_M] = \frac{1}{2}\,\mathrm{tr}(\mathbf{H}V)$, and requires no interchange of limit and expectation.

(ii). For $K = 2$ the recentered range is $R(\dot{\boldsymbol{\theta}}^{(b)}) = |(\hat{\theta}_1^{(b)} - \hat{\theta}_2^{(b)}) - (\hat{\theta}_1 - \hat{\theta}_2)|$. Expanding $h(M_n^{(b)})$ about $\boldsymbol{\mu}$ and writing $\sqrt{n}(M_n^{(b)} - \boldsymbol{\mu}) = \sqrt{n}(M_n^{(b)} - M_n) + \sqrt{n}(M_n - \boldsymbol{\mu})$ gives $n\,h(M_n^{(b)}) \xrightarrow{d} Q(\mathbf{Z}_M + \mathbf{Z}_M^*)$ jointly with $n\,h(M_n) \xrightarrow{d} Q(\mathbf{Z}_M)$, so the recentered contrast converges to the difference.

(iii). Immediate from parts (i) and (ii) and continuity of the conditional distribution of the limit, which holds because $Q(\mathbf{Z}_M + \mathbf{Z}_M^*) - Q(\mathbf{Z}_M) = \mathbf{Z}_M'\mathbf{H}\mathbf{Z}_M^* + \frac{1}{2}\mathbf{Z}_M^{*\prime}\mathbf{H}\mathbf{Z}_M^*$ is, given $\mathbf{Z}_M$, a non-constant polynomial in the non-degenerate Gaussian $\mathbf{Z}_M^*$ whenever $\mathbf{H} \neq \mathbf{0}$.

(iv). The unrecentered bootstrap range satisfies $nR(\hat{\boldsymbol{\theta}}^{(b)}) = n|h(M_n^{(b)})| \xrightarrow{d} |Q(\mathbf{Z}_M + \mathbf{Z}_M^*)|$ conditional on the data, so the $(1-\alpha)$ quantile of $R(\hat{\boldsymbol{\theta}}^{(b)})$ is $O_p(n^{-1})$. Coverage is exact because $R(\hat{\boldsymbol{\theta}}^{(b)}) \geq 0$ on every bootstrap replication, so $R_{1-\alpha}^* \geq 0 = \Delta$.[37] □

The mean is exactly the approximate bias of Nagar (1959), the expectation of the leading terms of the stochastic expansion. With M_n a vector of sample averages, $E[M_n] = \boldsymbol{\mu}$ and $\mathrm{Var}(\sqrt{n}(M_n - \boldsymbol{\mu})) = V$ hold exactly, and $\nabla h(\boldsymbol{\mu}) = \mathbf{0}$ leaves no n^{-1} contribution from the linear term, so the approximate bias equals $\frac{1}{2n}\,\mathrm{tr}(\mathbf{H}V)$ with no approximation error of its own. Exact moments are another matter. LIML has no finite integer moments (Mariano, 1982), so the proposition's mean is a statement about the approximate bias, which always exists, and not about a finite-sample expectation, which for LIML does not.

The direction of the failure is more important than the failure itself. In the scalar case the event in part (iii) reduces to $|Z_M + Z_M^*| \geq \sqrt{2}\,|Z_M|$, and direct calculation shows the ideal test is conservative at every level below roughly 0.37, with a limiting rejection rate near 7×10^{-5} at the nominal 5 percent level.[38] In plain terms, when two estimators are first-order identical the equality test almost never rejects their equality, which is the correct behavior, and the equivalence bound rather than the p-value carries the information about their finite-sample gap. In Design 8 for the TSLS–LIML comparison, neither the range-based equality test nor the Wald-based equality test rejects in any of 10,000 Monte Carlo replications at any instrument strength.

Part (iv) assigns each statistic its correct role in the degenerate pair. The equality test almost never rejects, and the equivalence bound is more informative about the finite-sample gap. The bound covers $\Delta = 0$ in every sample and shrinks at rate n^{-1}, the scale at which TSLS and LIML actually disagree, so $R_{1-\alpha}^*$ serves as a descriptive upper quantile for the realized finite-sample disagreement that the Angrist and Pischke (2009) heuristic concerns. The empirical content of the within-column TSLS–LIML comparisons therefore resides in $R_{.95}^*$, not in p_R.

[37] For the TSLS-LIML pair the mean in part (i) is the higher-order bias that grows with the degree of overidentification and that LIML removes (Nagar, 1959), and part (iii) is the failure of the bootstrap for statistics whose first-order term vanishes (Babu, 1984; Horowitz, 2001).

[38] The event is scale invariant, so normalize $\mathrm{Var}(Z_M) = \mathrm{Var}(Z_M^*) = 1$. Conditional on $Z_M = z$, the rejection probability is $P(|z + Z_M^*| \geq \sqrt{2}\,|z|) = \bar{\Phi}\big((\sqrt{2}-1)|z|\big) + \bar{\Phi}\big((\sqrt{2}+1)|z|\big)$, where $\bar{\Phi} = 1 - \Phi$ is the standard normal survival function, which decreases from one at $z = 0$ toward zero. The ideal test therefore rejects at level α exactly when $|Z_M|$ exceeds the cutoff z_α solving $\bar{\Phi}\big((\sqrt{2}-1)z_\alpha\big) + \bar{\Phi}\big((\sqrt{2}+1)z_\alpha\big) = \alpha$, and the limiting rejection rate is $2\bar{\Phi}(z_\alpha)$. At $\alpha = 0.05$ the cutoff is $z_\alpha \approx 3.98$ and the rejection rate is $2\bar{\Phi}(3.98) \approx 7 \times 10^{-5}$. The rejection rate equals the nominal level at $\alpha \approx 0.37$ and lies below it at every smaller level.

Proposition 5 (Influence-function-tied clusters)**.** *In the smooth-function-of-moments setting of this section, with V positive definite, suppose the specifications partition into clusters within which estimators share both a probability limit and an influence function. If the largest and smallest cluster-level probability limits are each attained by exactly one cluster, those clusters are distinct, and their influence functions differ, then* $\lim_{n\to\infty} P(\Delta \le R^{*\infty}_{1-\alpha}) = 1-\alpha$.

Proof. Write $\hat\theta_k = h_k(M_n)$ and define the clusters as the equivalence classes of the influence function, so that j and k belong to the same cluster if and only if $h_j(\boldsymbol{\mu}) = h_k(\boldsymbol{\mu})$ and $\nabla h_j(\boldsymbol{\mu}) = \nabla h_k(\boldsymbol{\mu})$, both equalities being what a shared influence function requires. For j and k in the same cluster, the second-order expansions in the proof of Proposition 4 give $\hat\theta_j - \hat\theta_k = O_p(n^{-1})$ and, conditional on the data, $\hat\theta_j^{(b)} - \hat\theta_k^{(b)} = O_{p^*}(n^{-1})$ as well. Because the maximal cluster-level limit is attained by a single cluster and strictly exceeds the rest, symmetrically for the minimum, and the estimates are consistent, the sample maximum lies in the max cluster and the sample minimum in the min cluster with probability approaching one, and within a cluster the identity of the attaining member perturbs the range only at order n^{-1}. Hence, for any representatives k^* and j^* of the max and min clusters,

$$\sqrt{n}\big(R(\hat{\boldsymbol{\theta}}) - \Delta\big) \;=\; \sqrt{n}\big[(\hat\theta_{k^*} - \theta_{k^*}) - (\hat\theta_{j^*} - \theta_{j^*})\big] + o_p(1) \;\xrightarrow{d}\; \mathcal{N}(0,\sigma^2),$$

with $\sigma^2 = (\nabla h_{k^*}(\boldsymbol{\mu}) - \nabla h_{j^*}(\boldsymbol{\mu}))' V (\nabla h_{k^*}(\boldsymbol{\mu}) - \nabla h_{j^*}(\boldsymbol{\mu})) > 0$, positive because the gradients of the max and min representatives differ by hypothesis and V is positive definite. The identical reduction applies to the unrecentered bootstrap range conditional on the data, using bootstrap consistency for the moments and the delta method on the representative contrast, so $\sqrt{n}(R(\hat{\boldsymbol{\theta}}^{(b)}) - R(\hat{\boldsymbol{\theta}}))$ converges conditionally to the same $\mathcal{N}(0,\sigma^2)$ limit. The location-scale identity and the coverage argument in the proof of part (ii) of Proposition 2 then apply verbatim. □

Validity without regularity. Proposition 5 restores exactness for influence-function ties. For the residual configurations, a cluster sharing only its probability limit or a bundle mixing tied and distinct specifications, Corollary 1 establishes conservative coverage when $\Delta = 0$ or when an active pair attaining Δ satisfies the stated pairwise normality, bootstrap-consistency, and positive-variance conditions. The result does not extend beyond those conditions.

Corollary 1 (Conservative validity for arbitrary comparison sets)**.** *Suppose that either* $\Delta = 0$*, or* $\Delta > 0$ *and at least one pair* (k^*, j^*) *with* $|\theta_{k^*} - \theta_{j^*}| = \Delta$ *is jointly asymptotically normal with the bootstrap consistent for the pair's joint distribution and with the contrast* $\hat\theta_{k^*} - \hat\theta_{j^*}$ *having positive asymptotic variance. Then* $\liminf_{n\to\infty} P(\Delta \le R^{*\infty}_{1-\alpha}) \ge 1-\alpha$*, whatever the tie configuration elsewhere in the comparison set.*

Proof. If $\Delta = 0$, then $R(\hat{\boldsymbol{\theta}}^{(b)}) \ge 0$ on every bootstrap replication, so the conditional quantile $R^{*\infty}_{1-\alpha} \ge 0 = \Delta$ and coverage is one for every n. Suppose $\Delta > 0$ and let $R^{*\infty}_{1-\alpha}(k^*, j^*)$ denote the exact conditional $(1-\alpha)$ quantile of $|\hat\theta^{(b)}_{k^*} - \hat\theta^{(b)}_{j^*}|$. The pair has distinct probability limits, so its active sets are singletons, and the argument of part (ii) of Proposition 2, which uses only the stated pairwise conditions, gives $P(\Delta \le R^{*\infty}_{1-\alpha}(k^*, j^*)) \to 1-\alpha$. On every bootstrap replication the range over all K specifications is at least the pairwise difference, $R(\hat{\boldsymbol{\theta}}^{(b)}) \ge |\hat\theta^{(b)}_{k^*} - \hat\theta^{(b)}_{j^*}|$,

and a conditional quantile is monotone under pointwise ordering of conditional distributions, so $R^{*\infty}_{1-\alpha} \geq R^{*\infty}_{1-\alpha}(k^*, j^*)$ in every sample. Hence $P(\Delta \leq R^{*\infty}_{1-\alpha}) \geq P(\Delta \leq R^{*\infty}_{1-\alpha}(k^*, j^*)) \to 1 - \alpha$. □

Any pair attaining $\Delta > 0$ has distinct probability limits, so an influence-function-tied pair is never the active pair, and under Assumptions 1 and 2 every pair qualifies. The positive-variance clause rules out only the knife-edge case of two estimators with different limits and identical influence functions. At non-regular configurations the inequality is typically strict. The bootstrap's failure to reproduce the sampling distribution of the range at a kink (Fang and Santos, 2019) surfaces here as conservatism rather than undercoverage.

The corollary also identifies the tighter of two valid summaries for a mixed set that must nonetheless be reported as one number, as when a published table or a survey stimulus presents one. For any comparison set, let $R^*_{1-\alpha}(k, j)$ denote the equivalence bound for the pair (k, j). The maximum of the constituent pairwise bounds, $\max_{k<j} R^*_{1-\alpha}(k, j)$, covers Δ by the same active-pair argument, each constituent being a $K = 2$ comparison and therefore individually valid, and never exceeds the joint bound, since on every bootstrap replication the range weakly dominates each pairwise difference. Both bounds are valid, but neither gives the mixed set a single-estimand interpretation. When the set mixes estimands, the bound summarizes total dispersion rather than robustness of a common parameter, and internally homogeneous comparison sets remain the recommended default.

In Table A.I, I evaluate the corollary in the two partial-tie configurations it covers, namely a pair of estimators with distinct influence functions sharing the maximal probability limit, and ties of that kind at both extremes. The data-generating processes are described as Design 9 in Appendix C. The oracle pair (that attaining Δ) covers at the nominal rate, between 0.949 and 0.955 across configurations and sample sizes, consistent with the exactness of part (ii) of Proposition 2. The joint bound covers at 0.985 to 0.999 and the maximum of the pairwise bounds at 0.979 to 0.997, and the conservatism persists at $n = 10{,}000$. The medians show the size of the slack. In the single-tie configuration the joint bound exceeds $\Delta = 0.5$ by 0.065 at $n = 2{,}000$ and by 0.029 at $n = 10{,}000$, shrinking at the $n^{-1/2}$ rate, and the max-pairwise bound sits between the joint bound and the oracle's throughout. The pointwise ordering behind the proof held in all 40,000 replications.

A.VII Conditions for validity and known limitations

It is useful to discuss when Assumptions 1 and 2 hold and when they may fail.

Joint asymptotic normality. Assumption 1 requires that $\hat{\boldsymbol{\theta}}$ is jointly asymptotically normal. This holds for M-estimators, GMM estimators, and most semiparametric estimators and smooth functions thereof (via the delta method) under standard regularity conditions (Newey and McFadden, 1994). It also holds for many combinations of heterogeneous estimators (e.g., OLS in one specification, IV in another) provided each estimator individually satisfies a central limit theorem and the joint distribution is determined by the common sampling variation. The condition can fail in several empirically relevant settings.

TABLE A.I
Design 9: Coverage of the Equivalence Bound at Partial-Tie Configurations

		Coverage of $R^*_{.95}$			Median $R^*_{.95}$		
Configuration	n	Joint	Max pairwise	Oracle pair	Joint	Max pairwise	Oracle pair
Tie at maximum	2,000	0.985	0.979	0.949	0.565	0.563	0.552
Tie at maximum	10,000	0.990	0.987	0.955	0.529	0.528	0.523
Ties at both extremes	2,000	0.998	0.996	0.951	0.578	0.573	0.552
Ties at both extremes	10,000	0.999	0.997	0.955	0.535	0.533	0.524

Notes: Entries report empirical coverage of $R^*_{.95}$ as an upper confidence bound for the population range $\Delta = 0.5$ and the corresponding median bound in the partial-tie configurations of Design 9. "Joint" uses the full comparison set, "Max pairwise" is the maximum of the constituent pairwise bounds, and "Oracle pair" uses the pair attaining Δ, which is known only in the simulation. Corollary 1 implies coverage of at least 0.95 for the two feasible bounds, while part (ii) of Proposition 2 implies exact coverage for the oracle pair. The pointwise ordering, joint $\geq$ max pairwise $\geq$ oracle pair, held in all 40,000 replications. Each cell uses 10,000 Monte Carlo replications with $B = 9{,}999$ bootstrap resamples. The data consist of n independent draws from a Gaussian vector with unit-variance coordinates, and each estimator is a coordinate sample mean. For $K = 3$, the coordinate means are $(0.5, 0.5, 0)$, with correlation 0.3 between the two tied coordinates and the third coordinate independent of both. For $K = 4$, a fourth coordinate with mean 0 and correlation 0.3 with the third creates ties at both extremes. The bootstrap resamples the n draws.

Weak instruments. When instruments are weak, the IV estimator is not asymptotically normal under conventional asymptotics (Staiger and Stock, 1997), and the pairs bootstrap does not consistently approximate its sampling distribution (Moreira, Porter, and Suarez, 2009). Andrews et al. (2019) survey the identification-robust methods developed in response. Proposition 1 does not apply in this case, and the statistics reported for the weakly identified columns of the Angrist and Pischke (2009) illustration are therefore interpreted descriptively and alongside identification-robust confidence regions.

Non-normal limits. Estimators with non-standard limiting distributions fall outside Assumption 1. The leading case is cube-root asymptotics. The maximum score estimator converges at rate $n^{-1/3}$ to a non-Gaussian limit (Kim and Pollard, 1990), and the standard bootstrap is inconsistent for it (Abrevaya and Huang, 2005). Subsampling remains valid under weak side conditions whenever a limiting distribution exists (Politis et al., 1999), at the cost of a subsample-size choice. It is no remedy for instrumental variables, however, as the subsample concentration parameter scales with m/n, and in the Angrist and Pischke (2009) application the Design 8 subsample sizes give median first-stage F statistics between 1.17 and 3.18, so a subsampling bound there would reflect weak-instrument artifacts rather than tie behavior.

Bootstrap consistency. Assumption 2 requires conditional convergence of the bootstrap distribution to the joint Gaussian limit. This holds for the pairs bootstrap under i.i.d. sampling (Bickel and Freedman, 1981; Singh, 1981; Hall, 1992) and for the block bootstrap under clustered sampling (Gonçalves and White, 2004; Cameron et al., 2008). Two settings deserve particular attention:

(a) Few clusters. When the number of clusters G is small (e.g., fewer than 20), the unrecentered bootstrap range in Design 7 is large relative to the population range, so the equivalence bound errs in the safe direction there. A large $R^*_{.95}$ makes the researcher less likely to declare

robustness when it should not be declared. The equality test is less protected. In Design 7 the range test over-rejects mildly at the smallest cluster counts and the Wald test over-rejects substantially, and the wild cluster bootstrap of Cameron et al. (2008), whose resamples the framework can consume without any change to the test statistics, does not repair that distortion. Equality p-values computed from very few clusters should therefore be interpreted with caution.

(b) Near-singular covariance. When K is large and some specifications are nearly identical, the bootstrap covariance matrix may be near-singular, and the contrast covariance that the Wald statistic inverts may be poorly conditioned. Simulation Design 6 ($K = 10$, nearly collinear specifications) shows that the Wald becomes conservative while the range maintains nominal size. For large K, the range statistic and the associated $R^*_{1-\alpha}$ are more reliable than the generalized Wald statistic.

Unstable specifications. Some estimators may fail to produce an estimate on certain bootstrap resamples. Dropping such resamples conditions the reference distribution on successful estimation, so it is defensible only when the failures are numerical rather than intrinsic and their probability vanishes with n. A non-negligible failure rate is not an ancillary diagnostic but evidence that the bootstrap approximation fails for that specification, raising the question whether the specification belongs in the comparison set. The companion software flags resamples with undefined estimates, reports the fraction dropped, and stops when it exceeds one percent.

B The Wald Benchmark and the Durbin–Wu–Hausman Test

This appendix defines the bootstrap Wald statistic reported alongside p_R in the simulation tables, establishes its validity under one additional assumption, and then relates both equality tests to the Durbin–Wu–Hausman test, verifying them against the analytical benchmark in the one setting with a closed-form target.

The precision-weighted alternative to the range-based equality test is a bootstrap-covariance quadratic form in the contrasts across specifications. Let

$$\widehat{V}^* \equiv \frac{1}{B-1}\sum_{b=1}^{B}\big(\hat{\boldsymbol{\theta}}^{(b)}-\bar{\boldsymbol{\theta}}^*\big)\big(\hat{\boldsymbol{\theta}}^{(b)}-\bar{\boldsymbol{\theta}}^*\big)', \qquad \bar{\boldsymbol{\theta}}^* \equiv B^{-1}\sum_{b=1}^{B}\hat{\boldsymbol{\theta}}^{(b)},$$

be the bootstrap covariance matrix, and let C be the $(K-1)\times K$ contrast matrix with $C_{jk} \equiv \mathbf{1}(k=j)-1/K$. The Wald statistic is

$$W \equiv (C\hat{\boldsymbol{\theta}})'\big(C\,\widehat{V}^*C'\big)^{-1}(C\hat{\boldsymbol{\theta}}),$$

which weights deviations in precisely estimated directions more heavily. At $K=2$, it is the direct-variance analogue of the analytical Hausman (1978) statistic, asymptotically equivalent to it under the efficiency condition the analytical test maintains. A p-value follows either by referring W to the χ^2_{K-1} distribution or by using the bootstrap reference distribution computed on recentered estimates $\dot{\boldsymbol{\theta}}^{(b)}$. Define

$$W^{(b)} \equiv (C\dot{\boldsymbol{\theta}}^{(b)})'\big(C\,\widehat{V}^*C'\big)^{-1}(C\dot{\boldsymbol{\theta}}^{(b)}),$$

and

$$p_W \equiv \frac{1+\sum_{b=1}^{B}\mathbf{1}(W^{(b)}\geq W)}{B+1} \in (0,1].$$

The validity of p_W requires the bootstrap to be consistent not only for the distribution of the estimates, Assumption 2, but for their second moment.

Assumption 3 (Bootstrap covariance consistency)**.** $nV^* \xrightarrow{p} \Sigma$, where $V^* \equiv \mathrm{Var}^*(\hat{\boldsymbol{\theta}}^{(b)})$ is the exact bootstrap covariance of $\hat{\boldsymbol{\theta}}^{(b)}$ conditional on the data.

Assumption 3 strengthens Assumption 2 from convergence of the conditional distribution to convergence of its second moment, and $\widehat{V}^*$ above is the B-replication sample estimate of V^*. The ideal Wald statistic W^∞ replaces $\widehat{V}^*$ by V^* in the quadratic form, and the ideal p-value p_W^∞ is the corresponding conditional tail probability.

Proposition 6 (Validity and consistency of the Wald benchmark)**.** *Under Assumptions 1, 2, and 3:*

(i) under the null $\Delta=0$, $W^\infty \xrightarrow{d} \chi^2_{K-1}$ *and* $p_W^\infty \xrightarrow{d} \mathrm{Uniform}(0,1)$*;*

(ii) given the data, $\widehat{V}^* \to V^*$ *almost surely as* $B\to\infty$ *and, under finite conditional fourth moments, has error of order* $B^{-1/2}$*, so the reported* W *and* p_W *converge to their ideal counterparts;*

(iii) under any fixed alternative $\Delta > 0$, the reported p_W converges in probability to its floor, $1/(B+1)$, and $P(p_W \le \alpha) \to 1$ for any $\alpha \ge 1/(B+1)$.

Proof. (i). Slutsky's theorem combined with Assumptions 1 and 3, the linear map $\mathbf{x} \mapsto C\mathbf{x}$, and matrix-inversion continuity gives $W^\infty = n(C\hat{\boldsymbol{\theta}})'(nCV^*C')^{-1}(C\hat{\boldsymbol{\theta}}) \xrightarrow{d} \boldsymbol{\eta}'(C\Sigma C')^{-1}\boldsymbol{\eta} \sim \chi^2_{K-1}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, C\Sigma C')$ is the limit of $\sqrt{n}\, C\hat{\boldsymbol{\theta}}$. The recentered statistics use $C\dot{\boldsymbol{\theta}}^{(b)} = C(\hat{\boldsymbol{\theta}}^{(b)} - \hat{\boldsymbol{\theta}})$, so Assumption 2 gives the same limit conditional on the data, and the probability-integral-transform argument in the proof of Proposition 1, with the continuous χ^2_{K-1} cumulative distribution function in place of F, gives $p_W^\infty \xrightarrow{d} \text{Uniform}(0,1)$.

(ii). Conditional on the data the $\hat{\boldsymbol{\theta}}^{(b)}$ are i.i.d. with covariance V^*, so the strong law of large numbers gives $\widehat{V}^* \to V^*$ almost surely as $B \to \infty$. Under finite conditional fourth moments, the error is $O_{p^*}(B^{-1/2})$, and the reported statistics inherit the ideal limits by continuity.

(iii). Under $\Delta > 0$, $\|\sqrt{n}\, C\hat{\boldsymbol{\theta}}\| \xrightarrow{p} \infty$ while the recentered $W^{(b)}$ remain $O_{p^*}(1)$ conditional on the data, so the count in the numerator of p_W converges to zero and p_W reaches its add-one floor. □

Part (ii) is a weaker guarantee than Lemma 1 provides for the range statistics. Every recentered $W^{(b)}$ is built from the same $\widehat{V}^*$, estimated once from all B replications jointly, so the $B+1$ Wald values are coupled through a shared ingredient and fail the conditional independence the lemma requires. The reported p_W therefore carries no fixed-B grid exactness, only the $B \to \infty$ approximation of part (ii), while p_R and $R^*_{1-\alpha}$, whose replications share nothing given the data, achieve their stated size and coverage at fixed B.

The bootstrap permits this construction for any K and any estimators for which the bootstrap is jointly valid and $C\Sigma C'$ is nonsingular, without an efficiency ranking, generalizing the seemingly-unrelated-regressions treatment of the IV and generalized method of moments case in Schaffer (2024). In theory the Wald test has a power advantage against dense alternatives, where many specifications differ slightly but no single pairwise difference dominates, but the simulations in Section IV show this advantage is small (in Design 4 the Wald test rejects alone in only 1.8 percent of replications), while the Wald test over-rejects substantially with few clusters and the range test remains much closer to nominal size. Against these properties, and the range test's avoidance of a covariance-matrix inversion that can be ill-conditioned when K is large or specifications are highly correlated, the minor power advantage does not justify reporting both, and p_R is the recommended equality test.

I next verify that in the tractable panel benchmark, the fixed-effects (FE) versus random-effects (RE) comparison of Hausman (1978), where the analytical p-value is available in closed form, p_{W,χ^2} nonetheless mimics the analytical test, alongside p_R. The exercise performs an analytical Hausman test as well as the two bootstrap-based equality tests on the same simulated panels, so the three rejection rates are directly comparable.

The data-generating process is

$$Y_{it} = \alpha_i + \beta\, X_{it} + \varepsilon_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

with $T = 10$, $\beta = 1$, $\alpha_i \sim N(0,1)$, and $\varepsilon_{it} \sim N(0,1)$. The regressor is $X_{it} = \rho\,\alpha_i + \nu_{it}$ with $\nu_{it} \sim N(0,1)$ i.i.d. and independent of α_i and ε_{it}. The $\rho = 0$ columns in Table B.I give a size check and the $\rho = 0.1$ columns a power check, with RE inconsistent. I run the exercise at six sample sizes, $N \in \{30, 50, 100, 500, 1{,}000, 2{,}000\}$, to trace how the bootstrap approximation improves with N. On each of 2,000 Monte Carlo replications, I estimate both FE and RE on the same panel, compute the analytical Hausman χ^2_1 p-value, and compute the bootstrap range and Wald p-values from $B = 999$ resamples of individuals, preserving the within-individual time series.

Table B.I reports rejection rates at the 5% level for all three tests. Figure B.I complements the table at $N = 1{,}000$ by plotting the bootstrap range-test p-value against the analytical Hausman p-value replication by replication. Although the range and the analytical Hausman statistic are constructed differently, at this sample size they test the same null with near-identical results, and the scatter falls along the 45-degree line.

TABLE B.I
Hausman Calibration: Analytical versus Bootstrap Rejection Rates

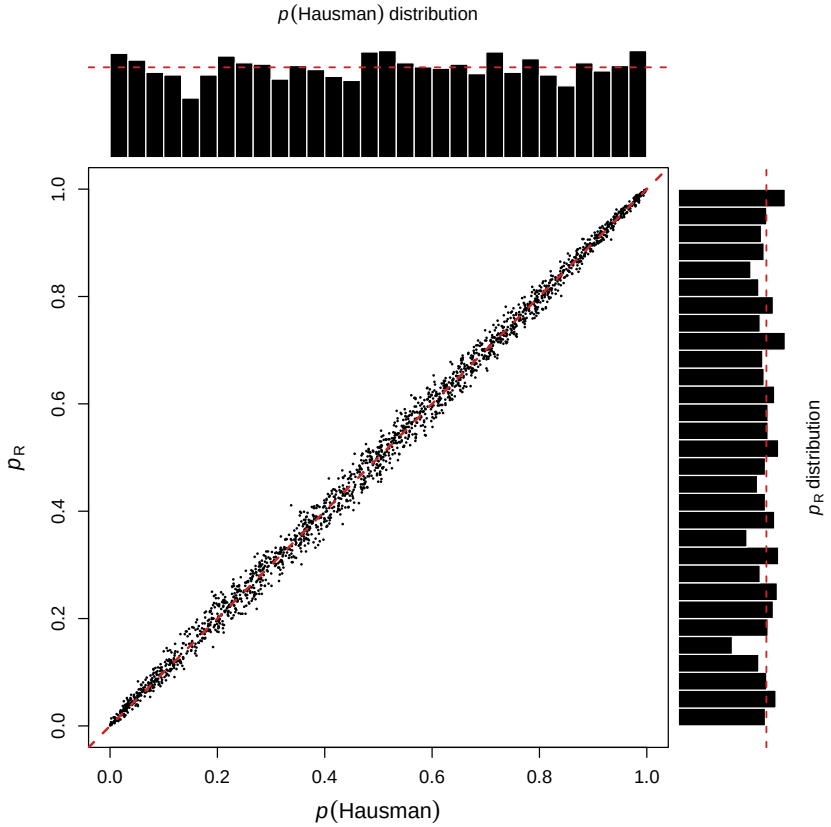
	$\rho = 0$			$\rho = 0.1$		
N	p_H	p_R	p_{W,χ^2}	p_H	p_R	p_{W,χ^2}
30	0.061	0.005	0.006	0.384	0.050	0.066
50	0.053	0.018	0.021	0.553	0.277	0.301
100	0.053	0.034	0.035	0.853	0.760	0.764
500	0.048	0.047	0.048	1.000	1.000	1.000
1,000	0.057	0.054	0.054	1.000	1.000	1.000
2,000	0.054	0.052	0.051	1.000	1.000	1.000

Notes: DGP is $Y_{it} = \beta X_{it} + \alpha_i + \varepsilon_{it}$ with $T = 10$, $\beta = 1$, $\alpha_i \sim N(0,1)$, $X_{it} = \rho\,\alpha_i + \nu_{it}$ with $\nu_{it} \sim N(0,1)$, and $\varepsilon_{it} \sim N(0,1)$. Under $\rho = 0$, both FE and RE are consistent (size check). Under $\rho = 0.1$, RE is inconsistent (power check). 2,000 Monte Carlo replications per cell, each with $B = 999$ bootstrap resamples of individuals; all three tests are computed on the same replications. p_H is the analytical Durbin–Wu–Hausman test, p_R the bootstrap range test, and p_{W,χ^2} the bootstrap-covariance Wald statistic referred to the χ^2_1 distribution. Rejection rates at the 5% level. Because p_{W,χ^2} refers the bootstrap-variance Wald statistic to the χ^2_1 distribution rather than to its resampled reference distribution, it need not coincide with p_R at $K = 2$ the way the bootstrap-reference Wald test of Section III does.

At $N \geq 500$ both bootstrap tests reproduce the analytical benchmark almost exactly. Under the null they reject at roughly the nominal rate (p_R at 0.047 to 0.054, p_{W,χ^2} within a few thousandths of p_R), and under the alternative all three tests reject in every replication. The range and Wald tests are nearly indistinguishable throughout, as expected at $K = 2$.

As N falls below 500 both bootstrap tests become conservative relative to the analytical Hausman. At $N = 100$ the range test rejects at 3.4% under the null against 5.3% for the analytical test, and the conservatism deepens at $N = 50$ (1.8%) and $N = 30$ (0.5%). It carries over to power: at $N = 100$ and $\rho = 0.1$ the analytical test rejects 85.3% of the time against 76.0% for the range test, and at $N = 30$ the figures are 38.4% and 5.0%. The Wald test tracks the range test to within about three percentage points in every cell. The bootstrap thus errs in the safe direction, under-rejecting the null in small samples at the cost of some power, and it does so whether the test statistic is the range or the Wald.

FIGURE B.I
Bootstrap range test versus analytical Hausman p-values



Notes: DGP is $Y_{it} = \beta X_{it} + \alpha_i + \varepsilon_{it}$ with $N = 1{,}000$, $T = 10$, $\beta = 1$, $\alpha_i \sim N(0,1)$, $X_{it} \sim N(0,1)$, $\varepsilon_{it} \sim N(0,1)$, and $\rho = 0$ (both FE and RE consistent). 2,000 Monte Carlo replications, each with $B = 999$ bootstrap resamples of individuals. Each point is one replication; the dashed red line is the 45-degree line. Marginal histograms show the distributions of the analytical Hausman (top) and bootstrap range-test (right) p-values; the dashed red lines mark the uniform density.

C Simulation Data-Generating Processes

This appendix provides full details of the six core data-generating processes used in the simulations. The six core designs, Designs 1 through 6, use $n = 2{,}000$ observations, $B = 9{,}999$ bootstrap estimates per Monte Carlo replication, and 10,000 Monte Carlo replications, with the bootstrap resampling individuals with replacement. The additional designs state their own sample structures and resampling schemes below. The coefficient of interest is always the coefficient on the treatment variable D_i. The designs are constructed to span the cases an applied researcher meets, one true null, four kinds of departure from it, and the near-collinear, few-cluster, and weak-instrument regimes that probe the machinery, and they draw on the omitted-variables and heterogeneous-effects structures of Lu and White (2014), Gelbach (2016), Słoczyński (2022), and Oster (2019). Each design below states the data-generating process, names the applied situation it represents, and derives the population range Δ from the design parameters. Every Δ is analytic, the spread of the estimator probability limits rather than a value estimated from the simulation, so the Monte Carlo results can be read against a known target. In Designs 2, 3, and 4 the estimators target a single structural parameter and Δ is the spread of their biases; in Design 5 they target distinct parameters and Δ is the spread of estimands.

Design 1: Same estimand ($K = 5$, null is true). *The textbook controls table with an exogenous regressor, where adding covariates cannot move the estimand.*

$$Y_i = \beta\, D_i + X_i'\gamma + u_i, \qquad D_i \sim N(0,1), \qquad X_i \sim N(0, I_4), \qquad u_i \sim N(0,1),$$

where D_i is independent of X_i, $\beta = 0.5$, and $\gamma = (0.30, 0.24, 0.19, 0.15)'$. Because $D_i \perp X_i$, the OLS coefficient on D_i has the same probability limit regardless of which elements of X_i are included as controls. The five specifications are OLS regressions of Y_i on D_i with $0, 1, 2, 3$, or 4 elements of X_i included. All five specifications target $\beta = 0.5$, so the implied population range is $\Delta = 0$.

Design 2: Similar probability limits ($K = 5$, null is false). *One control is a confounder, so the specification that omits it is biased and converges to a slightly different quantity.*

$$Y_i = (0.5 + 0.15\, X_{i1})\, D_i + X_i'\gamma + u_i, \qquad D_i = 0.5\, X_{i1} + \nu_i, \qquad \nu_i \sim N(0,1),$$

with $X_i \sim N(0, I_4)$, $\gamma = (0.30, 0.20, 0.10, 0.05)'$, and $u_i \sim N(0,1)$. Treatment effects are heterogeneous, with $\beta(X_{i1}) = 0.5 + 0.15\, X_{i1}$. The five specifications include $0, 1, 2, 3$, or 4 controls. Specifications that control for X_{i1} target $E[\beta(X_{i1})] = 0.5$. The specification that omits X_{i1} picks up omitted-variable bias from the main effect of X_{i1} on Y_i. The main effect of X_{i1} has coefficient 0.30 (the first element of γ), and $\mathrm{Cov}(D_i, X_{i1})/\mathrm{Var}(D_i) = 0.5/1.25 = 0.4$, so the bias is $0.30 \times 0.4 = 0.12$ and the probability limit is $0.5 + 0.12 = 0.62$.[39] Numerically, the population probability limits are $(0.620, 0.500, 0.500, 0.500, 0.500)$, giving an implied population range of $\Delta = 0.12$.

[39] Under $X_{i1} \sim N(0,1)$, the third moment is zero, which implies $\mathrm{Cov}(D_i, X_{i1} D_i) = 0$. The interaction therefore contributes nothing to the probability limit gap between specifications; all of the spread comes from main-effect omitted variable bias. A skewed distribution for X_{i1} would activate the interaction channel as well.

Design 3: Single outlier ($K = 5$, null is false). *One specification conditions on a post-treatment mediator, the classic bad control, and lands far from the rest.*

$$Y_i = \beta\, D_i + 0.4\, M_i + 0.2\, X_{i1} + u_i, \qquad M_i = 0.6\, D_i + \eta_i, \qquad \eta_i \sim N(0, 0.09),$$

with $\beta = 0.3$, $D_i \sim N(0,1)$, $X_i \sim N(0, I_3)$, and $u_i \sim N(0,1)$, where D_i is independent of X_i. Here M_i is a mediator caused by D_i that also affects Y_i. The total effect of D_i on Y_i is $\beta + 0.4 \times 0.6 = 0.54$. Because $D_i \perp X_i$, all four specifications that exclude M_i target the total effect regardless of which elements of X_i are included. The fifth specification includes M_i as a control, so the coefficient on D_i estimates the direct effect (0.30). This creates a sparse alternative with one specification far from the others. The implied population range is $\Delta = 0.54 - 0.30 = 0.24$.

Design 4: Diffuse disagreement ($K = 6$, null is false). *Each specification omits a different weak confounder, so all disagree slightly and none stands out.*

$$Y_i = \beta\, D_i + \sum_{j=1}^{6} \gamma_j X_{ij} + u_i, \qquad D_i = 0.3 \sum_{j=1}^{6} X_{ij} + \nu_i,$$

with covariates $X_{i1}, \ldots, X_{i6}$ drawn independently as $N(0,1)$, coefficients $\gamma_j = 0.05\,(-1)^j$, $\beta = 0.5$, and $\nu_i, u_i \sim N(0,1)$. Writing the outcome as an explicit sum, rather than $X_i'\gamma$, keeps the single covariate X_{ij} that specification j omits visible in the notation. Each of the six specifications includes five of the six covariates and omits one, with specification j omitting X_{ij}. Because D_i is correlated with every covariate and each X_{ij} has a small direct effect on Y_i whose sign alternates with j, omitting one covariate leaves a small omitted-variable bias whose sign follows that covariate's coefficient. All six specifications are therefore slightly biased, in directions that alternate, so the estimates spread out with none an outlier, the dense alternative that contrasts with the single outlier of Design 3. The size of each bias is the usual omitted-variable formula, the omitted coefficient times the regression coefficient of the omitted covariate on D_i holding the five included covariates fixed. That partial regression coefficient is $\mathrm{Cov}(D_i, X_{ij} \mid X_{i,-j}) / \mathrm{Var}(D_i \mid X_{i,-j}) = 0.3/1.09$, since D_i loads on each covariate with weight 0.3 and the remaining five covariates explain part of its variance, so the bias from dropping X_{ij} is $\gamma_j \times 0.3/1.09 = 0.05\,(-1)^j \times 0.3/1.09 \approx 0.014\,(-1)^j$. The population probability limits therefore alternate between 0.486 and 0.514, and the population range is $\Delta = 0.514 - 0.486 = 0.028$.

Design 5: Distinctly different estimands ($K = 5$, null is false). *The specifications estimate the effect on five different outcomes, genuinely distinct parameters.*
Five specifications estimate effects of D_i on five different outcome transformations:

$$Y_{ik} = \beta_k\, D_i + 0.2\, W_{ik} + u_{ik}, \qquad k = 1, \ldots, 5,$$

where each outcome k has its own coefficient, with $\beta_1 = 0.3$, $\beta_2 = 0.5$, $\beta_3 = 0.7$, $\beta_4 = 0.4$, $\beta_5 = 0.6$, and its own control W_{ik} and error u_{ik}, all drawn independently as $N(0,1)$, and $D_i \sim N(0,1)$. Each specification targets a genuinely different estimand. The implied population range is $\Delta = \beta_{\max} - \beta_{\min} = 0.4$.

Design 6: Near-singular covariance ($K = 10$, null is true). *Ten near-identical specifications, the case that strains the Wald statistic's matrix inverse.*
The DGP is the same as Design 1 with $\beta = 0.5$, but now with $K = 10$ specifications, each including a slightly different version of the same control variable:

$$\tilde{X}_{ik} = X_{i1} + 0.1\,\epsilon_{ik}, \qquad \epsilon_{ik} \sim N(0,1), \qquad k = 1, \ldots, 10.$$

Each specification regresses Y_i on D_i, X_{i2}, X_{i3}, and $\tilde{X}_{ik}$. Because the control variables differ only by small noise, the 10 estimates are nearly identical and the bootstrap covariance matrix is near-singular. The implied population range is $\Delta = 0$.

Design 7: Few clusters ($K = 5$, Δ varied). *Design 1 with clustered errors and few clusters, the setting where the block bootstrap is delicate.*
The DGP is Design 1 with $\beta = 0.5$, but observations are grouped into G clusters of $n_g = 50$ each. The error term has a group component and an individual component,

$$u_{ig} = \mu_g + \varepsilon_{ig}, \qquad \mu_g \sim N(0, \tfrac{3}{7}), \qquad \varepsilon_{ig} \sim N(0,1),$$

so the intracluster correlation is $\rho_{\text{ICC}} = \frac{3/7}{3/7+1} = 0.3$. I vary $G \in \{10, 15, 20, 30, 50, 100\}$. The five specifications are the same as Design 1. In the size runs the treatment is independent of the covariates and the implied population range is $\Delta = 0$. In the coverage runs the treatment is $D_{ig} = 0.5X_{1ig} + v_{ig}$ with $\text{Var}(D_{ig}) = 1$, so the specification omitting all covariates has probability limit 0.65, the four including X_1 have probability limit 0.5, and $\Delta = 0.15$ with the minimum attained by four specifications. The coverage runs use the block bootstrap.

Design 8: Weak instruments ($K = 3$, null is false for OLS vs. IV). *OLS beside TSLS and LIML as the instrument weakens, the returns-to-schooling table of the AK91 illustration.*

$$Y_i = \beta\, D_i + u_i, \qquad D_i = \pi Z_i + v_i,$$

with $Z_i \sim N(0,1)$, $\text{Corr}(u_i, v_i) = 0.5$, $\text{Var}(u_i) = \text{Var}(v_i) = 1$, and $n = 2{,}000$. TSLS and LIML are estimated using two excluded instruments, Z_i and $Z_i^2 - 1$, which is orthogonal to Z_i under normality. The pair (u_i, v_i) is jointly normal and independent of Z_i, so both excluded instruments are exogenous with respect to the structural error. The overidentification is necessary for TSLS and LIML to differ in finite samples; with a single instrument they are algebraically identical. I vary π so that the concentration parameter $n\pi^2$ takes values in $\{1, 5, 10, 20, 50, 100, 500\}$, the design labels F in Table IV. With two excluded instruments the realized joint first-stage F statistic averages roughly $1 + n\pi^2/2$, about 6 at the label of 10 and about 251 at 500. The three estimators are OLS, TSLS, and LIML, and Table IV reports the three pairwise comparison sets. TSLS and LIML target the same structural parameter ($\beta = 0.5$); their comparison traces how $R^*_{1-\alpha}$ behaves as identification strengthens, not a size check (Designs 1 and 6 provide the formal size checks). OLS has probability limit $\beta + 0.5/(1+\pi^2)$ because of the endogeneity, so the OLS-TSLS and OLS-LIML population range is $0.5/(1+\pi^2)$ and those comparisons are power checks. The three-way OLS-TSLS-LIML set is retained for the supplementary subsampling exercise.

Design 9: Partial ties ($K = 3$ and $K = 4$, coverage of the equivalence bound). *Distinct-influence-function estimators sharing an extreme probability limit, the non-regular configurations of Corollary 1.* Estimators are column means of constructed Gaussian variables, the minimal smooth-function-of-moments case, so probability limits and influence functions are controlled exactly. With $W_{1i}, \ldots, W_{4i}$ independent standard normals and $\rho = 0.3$, the three-estimator configuration uses $X_{1i} = 0.5 + W_{1i}$, $X_{2i} = 0.5 + \rho W_{1i} + \sqrt{1-\rho^2}\, W_{2i}$, and $X_{3i} = W_{3i}$, so the maximal limit is shared by two estimators with distinct influence functions, the minimal limit is unique, and $\Delta = 0.5$ analytically. The four-estimator configuration adds $X_{4i} = \rho W_{3i} + \sqrt{1-\rho^2}\, W_{4i}$, placing ties of the same kind at both extremes. Coverage of the joint bound $R^*_{.95}$, of the maximum of the pairwise bounds, and of the designated oracle pair $(1, 3)$ is evaluated at $n \in \{2{,}000,\ 10{,}000\}$ with 10,000 Monte Carlo replications, $B = 9{,}999$, and one L'Ecuyer-CMRG stream per replication from seed 42 (Table A.I).

D Companion Software

The statistics in this paper are implemented in two packages, one for R and one for Stata, that produce identical results from the same bootstrap estimates. Both are available from the author's repositories. The R package is at `https://github.com/davidajaeger/robustness` and the Stata package is available at `https://github.com/davidajaeger/robustness-stata`.[40]

The design separates generation from computation. A first step, specific to each application, re-estimates every specification on a common set of resampled units and saves the unrecentered bootstrap estimates. A second step, the packages documented here, reads those estimates and returns, for each comparison set, the observed range, the equality p-values p_R and p_W, and the equivalence bounds $R^*_{1-\alpha}$ at the chosen levels. The package requires, but cannot verify, that all of the specifications are estimated on the same bootstrap sample at every replication. This requirement is what permits the bootstrap to capture the joint sampling distribution across specifications. The generation scripts in the replication archive enforce and illustrate the shared resample.

In R, the core call takes the full-sample estimates for each specification, a $B \times K$ matrix of the bootstrap estimates, and a list describing the comparison sets:

```
robustness(theta, draws, comparisons = list(...)),
```

and returns an object whose `as.data.frame` method yields the reporting table directly. In Stata, the parallel call reads three files written by the generation step (the bootstrap estimates, metadata containing the full-sample estimates for each specification, and a file describing the comparison sets):

```
robustness using draws.dta, meta(meta.dta) comps(comps.dta),
```

and prints the two-panel table used for the empirical illustrations. Both default to reporting the median bound $R^*_{.50}$ and the 95th-percentile bound $R^*_{.95}$, and both return the full set of statistics, including the Wald-based quantities, for programmatic use.

[40] After publication the Stata package through SSC and the R package will be available through CRAN. The replication archive for all five illustrations, including the bootstrap-generation scripts, is at `https://github.com/davidajaeger/robustness-replications`.

E Survey Design and Pre-registration

This appendix documents the design of the survey whose responses are reported alongside the empirical illustrations in the main text. The pre-registration specifies a third DLR rating pooling the earnings and employment panels. I removed that item after the registration was locked but before fielding to limit respondent burden. This was a pre-fielding deviation from the registered instrument, and each respondent therefore provided six ratings rather than the seven described in the registration.

Pre-registration. The survey was pre-registered on the Open Science Framework on 28 April 2026, before any responses were collected.[41] The pre-registration specifies the five published tables, the outlined comparison set within each, the five-point response scale, and the target population. It defines the primary analysis as a comparison of respondents' robustness ratings to the minimum equivalence bound $R^*_{.95}$ and equality p-value p_R for the same comparison sets, testing whether ratings track the dispersion of point estimates or the joint dispersion measured by $R^*_{.95}$. The survey instrument differed from the pre-registration only in the removal of the pooled DLR item before fielding. The comparison of ratings to $R^*_{.95}$ and p_R reported in the conclusion implements the pre-registered primary analysis, with the aggregation of the ratings and the form of the comparison chosen after registration as described there.

Population and recruitment. The target population is research economists affiliated with CEPR and NBER who were likely to have encountered robustness claims in their own or others' research. The contact list is the union of publicly listed CEPR affiliates whose CEPR Programme Area is one of Development Economics, Economic History, Labour Economics, Organizational Economics, Political Economy, or Public Economics and publicly listed NBER affiliates whose NBER Program is one of Children and Families, Development Economics, Economics of Aging, Economics of Education, Economics of Health, Labor Studies, or Public Economics. Researchers affiliated with both networks are included if their affiliation in either network falls within these programs.

Email addresses for the listed affiliates were collected from publicly available sources outside the network directories themselves, namely individuals' institutional faculty pages or personal academic homepages, and only institutional email addresses were used. The invitation described the study as a short survey on how economists evaluate robustness and did not disclose the comparison to a formal statistic, so that responses reflect unprompted professional judgment.

Timing. The survey opened on 5 May 2026 and closed on 2 June 2026, a field period of four weeks. Of the approximately 1,650 individuals solicited, 324 consented, a response rate of about 20 percent, and between 288 and 294 assessed each table.

Instrument. After a consent screen, each respondent was shown the five tables in random order, one per screen. Each table was reproduced from its published source with the relevant

[41] "How Do Economists Evaluate Robustness? An Anonymous Survey," `https://doi.org/10.17605/OSF.IO/FYDU3`.

estimates outlined and was accompanied by the question "How do you assess the robustness of the outlined results? Please base your assessment on the results shown in the table," answered on the scale not at all robust (1), slightly robust (2), moderately robust (3), very robust (4), and extremely robust (5). The tables shown were the original published exhibits, which in some cases differ from the replication tables in the main text. In particular, the Angrist and Krueger (1991) table was shown as reproduced in Angrist and Pischke (2009), with column (2) whited out, the two-stage least squares and LIML estimates in columns (1) and (3)–(6) outlined, and the OLS estimates reported only in the table notes. The outlined comparison set is therefore the ten two-stage least squares and LIML estimates reported in the final row of Table IX. The Dube et al. (2010) table reported both the earnings and employment outcomes, and the robustness question was asked separately for each, with the earnings estimates and the employment estimates outlined in turn.[42] The outlined comparison set in each table corresponds to a comparison set reported in the main text, which permits each response to be paired with $R^*_{.95}$ and p_R.[43]

Ethics and anonymity. The study received ethics approval from the University of St Andrews Business School Ethics Committee. Responses are anonymous and contain no identifying information. Except for the email addresses of those who entered the prize draw, no data beyond the responses were recorded. Responses were separated from prize-draw emails immediately after downloading. Once separated, responses could not be linked to individual respondents, so a respondent who had submitted the survey could not subsequently be identified for withdrawal. Respondents who completed the survey could enter a prize draw for one of fifty manuscript reviews on `https://www.refine.ink`. Email addresses for the prize draw were collected on a separate screen and destroyed once the draw was completed.

[42] As noted above, an "overal" comparison for Dube et al. (2010) was pre-registered but removed before the survey was fielded.

[43] For Autor (2003), the outlined estimates are the seven specifications of the original table, corresponding to the comparison set labeled "All specs. excl. baseline" in Table VII, Part B.

F Bootstrap Correlation Structure of the Applications

This appendix reports the full Pearson correlation matrix of the bootstrap estimates for each of the five applications. These matrices display the linear dependence among the estimates. The joint distribution that the range statistics summarize combines these correlations with the marginal standard errors, and the range is a nonlinear functional of that joint distribution, so the matrices are descriptive companions to the tables rather than sufficient statistics for the bounds. The matrices differ markedly across applications, from the near-uniform high correlation of Karlan and List (2007) to the blocked and partly negative structure of the Dube et al. (2010) employment estimates. I report Pearson correlations here as the standard linear summary. For Angrist and Pischke (2009), the Spearman rank correlation of the column 4 TSLS-LIML draws, which is robust to the heavy tails that weak identification induces, is 0.914 against a Pearson correlation of 0.100, and that divergence is itself a symptom of the non-normal sampling distribution under weak instruments. Each matrix is symmetric and is shown in upper-triangular form.

TABLE F.I
Bootstrap Correlation Matrix: Lee (2008)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	1.000	.950	.999	.999	.947	.592	.653
(2)		1.000	.947	.947	.995	.781	.823
(3)			1.000	1.000	.947	.580	.656
(4)				1.000	.947	.579	.656
(5)					1.000	.768	.831
(6)						1.000	.696
(7)							1.000

Notes: Pearson correlations among the 7 specification estimates across $B = 9{,}999$ bootstrap resamples, in the order they appear in Table V. Leading zeros are omitted. Specifications: (1) no controls, (2) previous vote share, (3) office experience, (4) electoral experience, (5) all controls, (6) residualized outcome, (7) first-differenced outcome.

TABLE F.II
Bootstrap Correlation Matrix: Karlan and List (2007)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1)	1.000	.974	.975	.979	.977	.953	.973	.962	.976	.898
(2)		1.000	.985	.995	.994	.962	.993	.973	.993	.919
(3)			1.000	.995	.993	.973	.988	.982	.991	.919
(4)				1.000	.997	.971	.995	.983	.996	.916
(5)					1.000	.973	.988	.976	.993	.917
(6)						1.000	.954	.986	.972	.942
(7)							1.000	.977	.993	.921
(8)								1.000	.984	.933
(9)									1.000	.919
(10)										1.000

Notes: Pearson correlations among the 10 specification estimates across $B = 9{,}999$ bootstrap resamples, in the order they appear in Table VI. Leading zeros are omitted. Specifications: (1) donation history, (2) proportion white, (3) proportion black, (4) age 18–39, (5) household size, (6) median household income, (7) home ownership, (8) college, (9) urban, (10) all demographics.

TABLE F.III
Bootstrap Correlation Matrix: Autor (2003)

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(0)	1.000	.954	.836	.725	.639	.400	.933	.649
(1)		1.000	.858	.782	.684	.449	.980	.687
(2)			1.000	.727	.828	.368	.841	.802
(3)				1.000	.865	.400	.750	.829
(4)					1.000	.345	.670	.967
(5)						1.000	.467	.356
(6)							1.000	.695
(7)								1.000

Notes: Pearson correlations among the 8 specification estimates across $B = 9{,}999$ bootstrap resamples, in the order they appear in Table VII. Leading zeros are omitted. Specifications: (0) baseline, (1) linear trends, (2) quadratic trends, (3) linear + region×year, (4) quadratic + region×year, (5) demographics + state fixed effects, (6) demographics + linear, (7) demographics + quadratic + region×year.

TABLE F.IV
Bootstrap Correlation Matrix: Dube, Lester, and Reich (2010), Log Earnings

		All-county								Border			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
All-county	(1)	1.000	.978	.559	.536	.521	.509	.057	.031	.754	.679	−.001	−.031
	(2)		1.000	.601	.593	.542	.535	.022	−.008	.789	.746	.024	−.004
	(3)			1.000	.990	.775	.769	.027	−.004	.557	.548	.157	.156
	(4)				1.000	.759	.756	−.001	−.030	.544	.549	.157	.159
	(5)					1.000	.997	−.030	−.046	.478	.466	.048	.036
	(6)						1.000	−.035	−.052	.482	.475	.050	.039
	(7)							1.000	.993	.202	.174	.407	.406
	(8)								1.000	.165	.135	.393	.396
Border	(9)									1.000	.981	.167	.148
	(10)										1.000	.186	.171
	(11)											1.000	.991
	(12)												1.000

Notes: Pearson correlations among the 12 specification estimates across $B = 9{,}999$ bootstrap resamples, in the order they appear in the log earnings panel of Table VIII. Leading zeros are omitted. Specifications: All-county specifications: (1) period FE, population; (2) period FE, population + private; (3) census-division × period, population; (4) census-division × period, population + private; (5) state trend + census-division, population; (6) state trend + census-division, population + private; (7) MSA × period, population; (8) MSA × period, population + private. Border specifications: (9) period FE, population; (10) period FE, population + private; (11) pair × period, population; (12) pair × period, population + private.

TABLE F.V
Bootstrap Correlation Matrix: Dube, Lester, and Reich (2010), Log Employment

		All-county								Border			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
All-county	(1)	1.000	.975	.653	.616	−.057	−.032	−.054	−.047	.739	.763	.064	.213
	(2)		1.000	.660	.661	−.094	−.041	−.133	−.133	.726	.793	.000	.184
	(3)			1.000	.956	−.430	−.411	−.025	−.024	.451	.442	.080	.087
	(4)				1.000	−.502	−.419	−.077	−.079	.410	.444	.016	.064
	(5)					1.000	.955	−.026	−.015	−.065	−.094	.027	.026
	(6)						1.000	−.080	−.069	−.075	−.077	−.004	.038
	(7)							1.000	.970	−.014	−.102	.315	.180
	(8)								1.000	−.020	−.114	.345	.198
Border	(9)									1.000	.952	.107	.240
	(10)										1.000	.011	.231
	(11)											1.000	.756
	(12)												1.000

Notes: Pearson correlations among the 12 specification estimates across $B = 9{,}999$ bootstrap resamples, in the order they appear in the log employment panel of Table VIII. Leading zeros are omitted. Specifications: All-county specifications: (1) period FE, population; (2) period FE, population + private; (3) census-division × period, population; (4) census-division × period, population + private; (5) state trend + census-division, population; (6) state trend + census-division, population + private; (7) MSA × period, population; (8) MSA × period, population + private. Border specifications: (9) period FE, population; (10) period FE, population + private; (11) pair × period, population; (12) pair × period, population + private.

TABLE F.VI
Bootstrap Correlation Matrix: Angrist and Pischke (2009), Table 4.6.2

		Col. 1			Col. 3			Col. 4			Col. 5			Col. 6		
		OLS	TSLS	LIML	OLS	TSLS	LIML	OLS	TSLS	LIML	OLS	TSLS	LIML	OLS	TSLS	LIML
Col. 1	OLS	1.000	.015	.015	1.000	.021	.012	1.000	.017	−.001	.972	.045	.025	.972	.044	.020
	TSLS		1.000	.995	.015	.699	.706	.001	.057	.010	.017	.385	.395	.002	.055	.053
	LIML			1.000	.015	.693	.705	.001	.058	.009	.016	.379	.394	.002	.053	.054
Col. 3	OLS				1.000	.021	.012	1.000	.017	−.001	.972	.045	.025	.972	.044	.020
	TSLS					1.000	.981	.010	.673	.072	.025	.510	.495	.013	.266	.236
	LIML						1.000	.001	.650	.072	.016	.491	.496	.004	.251	.236
Col. 4	OLS							1.000	.017	−.002	.972	.038	.019	.972	.044	.019
	TSLS								1.000	.100	.021	.311	.290	.021	.341	.291
	LIML									1.000	−.001	.052	.056	−.001	.054	.058
Col. 5	OLS										1.000	.046	.025	1.000	.045	.018
	TSLS											1.000	.951	.039	.928	.816
	LIML												1.000	.018	.873	.880
Col. 6	OLS													1.000	.044	.018
	TSLS														1.000	.873
	LIML															1.000

Notes: Pearson correlations among the 15 specification estimates across $B = 9{,}999$ bootstrap resamples, in the order they appear in Table IX. Leading zeros are omitted. Each of the five specifications in Table 4.6.2 (columns 1, 3, 4, 5, and 6) contributes its OLS, TSLS, and LIML estimate.