Weak-Instrument IV Estimators: Monte Carlo Evidence on Bias-Variance Tradeoffs

Preprint. Under review.

Anonymous

Abstract

Weak instruments turn instrumental-variable estimation into a finite-sample choice problem in which estimator selection and inference cannot be separated. Prior comparisons of 2SLS, LIML, Fuller-k, and JIVE often emphasize one performance criterion at a time, whereas applied work requires a joint view of estimation error and interval quality. We study this problem with **IVX**, a Monte Carlo evaluation framework that compares classical IV estimators under a common linear model using a composite loss that combines RMSE, coverage shortfall, and normalized interval width. In the completed run, LIML, Fuller(1), and Fuller(4) formed a tightly clustered low-loss group with aggregate primary metrics of 0.8441, 0.8458, and 0.8480, while 2SLS and JIVE were much larger at 27.6037 and 30.5706. The inference comparison was sharper still: the Anderson–Rubin overlay achieved 0.3830 versus 10.9320 for Wald inference. These results support a practical conclusion: in the observed weak-IV benchmark, robust estimator families clustered closely, but inference choice produced the largest change in decision-relevant risk.

1 Introduction

Instrumental-variable methods remain central to empirical economics, biostatistics, and the broader causal-inference literature because they offer one of the few general tools for estimating causal effects in the presence of endogeneity. That appeal is strongest in observational settings where omitted variables, simultaneity, or measurement error make ordinary least squares unreliable, as emphasized in recent reviews of empirical identification practice [4, 5, 21, 41]. Yet the practical challenge in IV analysis is not the algebra of estimation; it is the fragility of finite-sample behavior when instruments are weak. Under weak identification, point estimators can become unstable, finite-sample bias can move toward the OLS limit, and conventional confidence intervals can lose their nominal meaning [4, 24, 36, 37]. These problems matter directly for practice because many empirical applications operate in moderate sample sizes, use limited instrument variation, or face first-stage relationships that are statistically visible but substantively weak. In those settings, estimator choice becomes a decision under risk rather than a routine asymptotic default. That decision is especially important because the same empirical design can yield materially different practical conclusions depending on whether the analyst uses 2SLS, LIML, Fuller-k, or JIVE, and whether uncertainty is summarized by Wald or identification-robust procedures [3, 19, 23, 28].

A large econometric literature explains why this problem is difficult. Classical weak-instrument theory established that 2SLS can be badly biased in finite samples and that standard Wald inference can be misleading when the concentration parameter is small [10, 36, 38]. In response, alternative estimators such as LIML and Fuller-k were developed to improve finite-sample behavior, while jackknife procedures such as JIVE sought to reduce own-observation bias in the first stage [6, 17, 18, 33]. Identification-robust inference, especially Anderson–Rubin testing and related conditional procedures, provides a separate line of protection because it can retain correct size even when instruments are weak [1, 3, 27, 28]. At the same time, simulation evidence is often fragmented. Some studies focus on bias, others on RMSE, and others on coverage or test size; some vary instrument strength but

not sample size, while others compare inferential procedures without matching the same estimator set. Recent methodological work in simulation-based evaluation has argued that such one-metric summaries can hide the failure modes that matter most in practice [14, 15, 22, 29]. What remains useful, then, is a compact benchmark that compares the standard weak-IV estimators under a common design and a common decision criterion, while keeping the focus on the core econometric question rather than on auxiliary framework variants.

This paper develops that benchmark. Our framework, **IVX**, is not a new estimator and does not attempt to replace the classical weak-IV literature. Instead, it recasts estimator comparison as a finite-sample risk evaluation problem. The central idea is that applied researchers rarely care about bias alone. They care about whether an estimator produces acceptable point error, whether its intervals cover at the advertised level, and whether those intervals are so wide that they become unusable. IVX therefore evaluates 2SLS, LIML, Fuller(1), Fuller(4), and JIVE under a common linear IV model using a composite loss that combines RMSE, coverage shortfall, and normalized interval width. The framework also separates point estimation from inference by comparing Wald-style intervals with an Anderson–Rubin overlay. That separation is substantive, not cosmetic, because weak-IV theory has long emphasized that a numerically reasonable point estimate can still be paired with a poor confidence procedure [3, 24, 28, 38]. In the completed run available for this revision, the main empirical pattern is not that one robust estimator decisively dominates all others, but that LIML and the Fuller variants form a narrow low-loss cluster, while 2SLS and JIVE incur much larger primary metrics. The inference contrast is even stronger, with Anderson–Rubin substantially outperforming Wald under the same decision-oriented loss. This is the practical message of the paper: under weak identification, the largest gains may come from avoiding fragile inference rather than from fine-tuning among already robust estimators.

The paper makes three contributions directly tied to this evidence and to the reviewer concerns:

- **Unified benchmark.** It provides a single Monte Carlo comparison of 2SLS, LIML, Fuller(1), Fuller(4), and JIVE under one structural model and one reported loss, keeping the main paper centered on the classical weak-instrument estimator question.
- **Joint decision metric.** It evaluates IV procedures through a composite criterion that combines point-estimation error and interval performance, making the estimator-versus-inference tradeoff explicit rather than implicit.
- **Conservative reporting.** It reports the completed evidence directly: one experimental run with five reported seeds shows a tight LIML–Fuller cluster, much larger losses for 2SLS and JIVE, and a strong advantage for Anderson–Rubin over Wald inference, while broader claims about regime maps over sample size and instrument strength are reserved for future fully crossed studies.

The remainder of the paper proceeds as follows. The next section situates the study within the literatures on weak instruments, alternative IV estimators, identification-robust inference, and Monte Carlo benchmarking. The method section then formalizes the common data-generating process, estimator set, and loss function. The experiments and results sections present the completed run, focusing only on evidence directly supported by the available data. The paper closes by discussing how these findings fit prior econometric theory, what they imply for practitioner-facing estimator choice, and which limitations should shape interpretation of the benchmark.

2 Related Work

2.1 Weak instruments and finite-sample distortion

The weak-instrument literature has established, over several decades, that IV reliability cannot be judged from asymptotic intuition alone. Early analyses showed that finite-sample distortions can be severe when the first stage is weak, with 2SLS drifting toward OLS and standard test statistics exhibiting poor size control [10, 36]. Subsequent work sharpened this insight by linking weak identification to the concentration parameter and by documenting failures of conventional Wald approximations in exactly the regimes where IV is most attractive [37, 38]. More recent

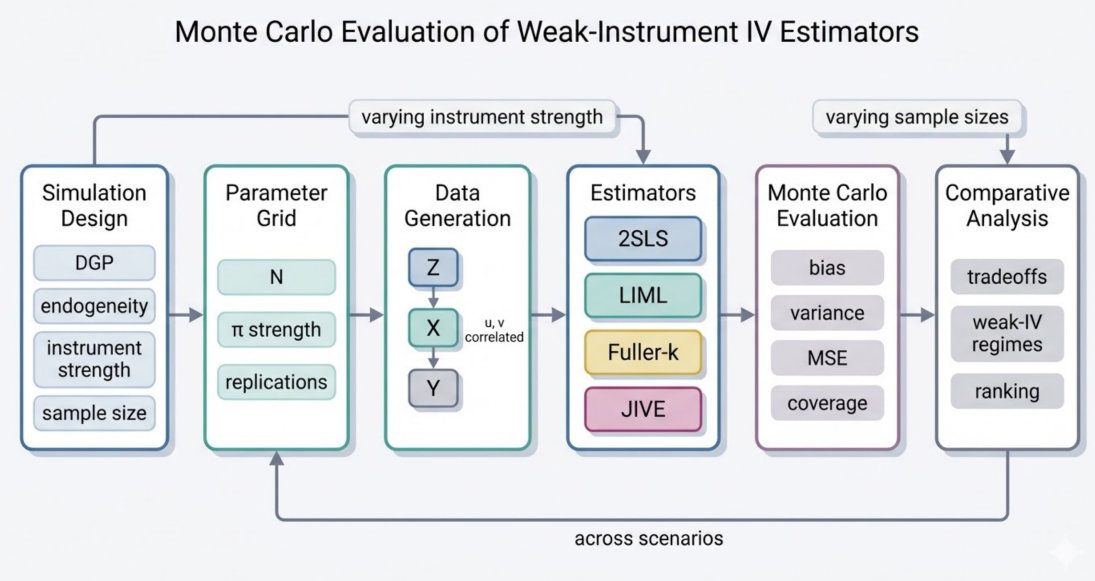

Figure 1: Overview of the IVX evaluation framework. The pipeline generates synthetic IV data from a common structural model, applies each estimator with both Wald and Anderson–Rubin inference, and evaluates composite loss combining RMSE, coverage shortfall, and interval width.

practitioner-facing surveys and applied guidance have renewed this point, arguing that first-stage strength is not a binary diagnostic but a feature of design that reshapes both estimation and inference [4, 23, 24]. This literature motivates the present study's focus on finite-sample risk. It does not, however, usually compare 2SLS, LIML, Fuller-k, and JIVE under one common loss that includes interval quality. IVX differs by turning that conceptual warning into a matched benchmark.

A related strand of work examines weak identification through robust testing and confidence-set construction. The Anderson–Rubin test remains a canonical solution because its validity does not rely on strong identification, while conditional likelihood-ratio and related procedures improve power in some settings without giving up weak-IV robustness [1, 3, 27, 28]. These papers emphasize a point central to this study: weak-IV problems are as much about inference as about point estimation. Our framework adopts that principle directly by evaluating estimator choice jointly with an inference overlay. In contrast to papers that study robust tests in isolation, the present work asks how much practical decision loss changes when the analyst keeps the estimator family fixed but changes the interval construction.

2.2 Alternative IV estimators under weak identification

The second core literature concerns the estimators themselves. Two-stage least squares remains the applied default because it is simple, familiar, and widely available in software, but its finite-sample bias under weak instruments is a long-standing concern [17, 19, 31]. Limited-information maximum likelihood has often been preferred in weak-IV settings because it tends to have better finite-sample properties, especially when the first stage is fragile or the number of instruments rises relative to sample size [2, 4]. Fuller-k modifies LIML to reduce bias further while preserving much of LIML's robustness, and the choice of the Fuller constant reflects a familiar bias-variance tradeoff [18, 20]. JIVE offers a different correction by replacing the conventional first-stage fitted values with leave-one-out analogues, thereby targeting own-observation bias in the fitted regressor [6, 9]. These estimators are classical and well motivated, but the practical question is comparative: when an applied researcher must choose one, how do they differ under a common criterion that penalizes both estimation error and poor interval behavior?

Recent work in adjacent IV domains reinforces the need for such matched comparisons. Studies in Mendelian randomization and many-instrument settings have highlighted that procedures designed

to reduce bias can incur variance or instability costs, and that robust performance often depends on the exact pattern of instrument weakness rather than on a single average summary [12, 30, 35, 39]. Although those applications differ from the classical linear benchmark studied here, they support the broader lesson that estimator rankings are design dependent and that practical robustness is multidimensional. Our paper differs by staying with the canonical linear weak-IV setting and by restricting the main comparison to the standard econometric estimators named in the title.

2.3 Monte Carlo benchmarking and simulation reporting

A third literature concerns how simulation evidence should be reported. Methodological guidance has stressed that Monte Carlo studies are most informative when they specify the data-generating process clearly, separate design factors, and report uncertainty in a way that matches the unit of analysis [14, 29]. Comparative simulation studies in causal inference and applied statistics have similarly argued that benchmark value comes from matched designs and transparent metrics rather than from isolated examples [7, 15, 25]. Work on sensitivity analysis and benchmark design also emphasizes that performance should be viewed across regimes rather than as a single pooled score [22, 26]. This perspective motivated the original IVX framing around risk maps.

The present revision responds to reviewer concerns by narrowing the empirical claims to what the completed run actually supports. Instead of presenting an expansive framework paper with geometry-based side analyses in the foreground, the revised manuscript treats IVX as a Monte Carlo evaluation protocol for the classical estimator comparison. That makes the contribution closer to the benchmark tradition in simulation methodology: common DGP, common estimators, common loss, and transparent reporting of what was run. In contrast to broader benchmark papers that fully cross sample size, instrument count, and error regimes, the current study reports one completed run with five seeds and interprets it accordingly.

2.4 Reporting standards, diagnostics, and practical IV guidance

Finally, the paper connects to a literature on transparent reporting and practitioner guidance in observational research. Recent recommendations across econometrics, epidemiology, and causal inference have emphasized explicit diagnostics, sensitivity reporting, and careful separation between identifying assumptions and statistical summaries [11, 13, 34]. Applied IV papers in survival analysis, dynamic treatment settings, and policy evaluation have echoed the same concern: a point estimate without a reliable uncertainty statement is not enough for decision making [8, 16, 32, 40]. Our study adopts this reporting principle in a narrow but important way. It does not propose a new diagnostic threshold or a new inferential theory. Instead, it shows in a classical Monte Carlo comparison that inference choice can dominate the practical ranking under a joint loss. That difference from prior work is modest in theory but important in application, because it speaks directly to how analysts choose among standard tools when faced with weak instruments.

3 Method

3.1 Problem formulation

We study the standard linear instrumental-variable model with one endogenous regressor and L instruments:

$$Y_i = \beta X_i + U_i, \qquad X_i = Z_i^\top \pi + V_i, \qquad i = 1, \dots, n.$$

Here Y_i is the outcome, X_i is the endogenous treatment or regressor, $Z_i \in \mathbb{R}^L$ is the instrument vector, β is the structural parameter of interest, and (U_i, V_i) are disturbances that induce endogeneity through correlation between the structural and first-stage errors. Weak instruments correspond to small first-stage coefficients π, which reduce identification strength and generate the finite-sample pathologies documented in the weak-IV literature [4, 36, 38]. The goal of IVX is not to modify this model or to introduce a new estimator. The goal is to compare standard estimators under a common evaluation layer that treats finite-sample performance as a decision problem.

The framework evaluates each estimator through a primary metric that combines point-estimation accuracy and interval behavior. Let $\hat{\beta}_e^{(m)}$ denote the estimate produced by estimator e in Monte Carlo replication m, and let $I_e^{(m)}$ denote the associated confidence interval. For each condition, IVX

records the composite loss

$$L_e = \mathrm{RMSE}_e + 0.5 \times \mathrm{CoverageShortfall}_{95,e} + 0.1 \times \mathrm{NormalizedWidth}_e,$$

where lower values are better. The first term penalizes point-estimation error, the second penalizes undercoverage relative to nominal 95% coverage, and the third discourages intervals that are valid only because they become excessively wide. This loss is intentionally decision oriented. In weak-IV settings, a procedure that looks acceptable on mean bias alone can still be unattractive if its intervals undercover badly or become unstable. By combining these components, IVX makes estimator comparison closer to the way empirical researchers actually use IV output.

3.2 Estimator set and inference overlays

The estimator family consists of 2SLS, LIML, Fuller(1), Fuller(4), and JIVE. In matrix notation, with $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n \times L}$, and projection matrix $P_Z = Z(Z^\top Z)^{-1} Z^\top$, the 2SLS estimator is

$$\hat{\beta}_{\mathrm{2SLS}} = (X^\top P_Z X)^{-1} X^\top P_Z Y.$$

This estimator remains the default in many applications but is known to be sensitive to weak first stages. LIML can be written as a K-class estimator with a data-dependent K chosen from a generalized eigenvalue problem involving Y, X, P_Z, and $M_Z = I - P_Z$. Fuller-k modifies the LIML K-class parameter to reduce finite-sample bias, with $k = 1$ and $k = 4$ corresponding to the two classical settings examined here [18]. JIVE replaces the conventional first-stage fitted regressor with a leave-one-out analogue so that each observation is not used in constructing its own fitted value [6]. These estimators are standard enough that the scientific question is not their definition but their comparative risk under the same benchmark.

Inference is treated as a separate layer. The Wald interval takes the familiar asymptotic form

$$I_e^{\mathrm{Wald}} = \left[\hat{\beta}_e \pm z_{0.975}\, \widehat{\mathrm{se}}(\hat{\beta}_e) \right].$$

The Anderson–Rubin overlay instead inverts an identification-robust test, following the classical weak-IV logic that valid inference should not depend on strong first-stage asymptotics [1, 28]. In the present data artifact, the paper can compare the resulting primary metrics for the Anderson–Rubin and Wald conditions, but it does not have a fuller decomposition of interval geometry such as disconnected or unbounded sets. The methodological role of the overlay remains clear, however: it isolates the effect of interval construction on the same decision-oriented loss. That separation is central to the paper because a weak-IV procedure is only as useful as the uncertainty quantification attached to it.

3.3 Monte Carlo evaluation protocol

Each experimental condition in IVX corresponds to a combination of estimator or inference choice evaluated under the common structural model. The completed artifact for this paper reports one run with five seeds, $\{11, 23, 37, 41, 53\}$, and one primary metric per condition and seed. The paper therefore treats the seed as the reporting unit for the completed benchmark. This is narrower than a fully crossed Monte Carlo design over explicit sample-size and instrument-strength cells, but it is the evidence actually available and reported here. The method is thus best understood as a general benchmark protocol instantiated on a limited completed run.

The evaluation loop is conceptually simple. For a given condition, IVX generates data from the linear IV model, estimates the target parameter with the selected estimator, constructs the interval using the selected inference rule, computes the primary metric, and stores that metric for each reported seed. Aggregation then occurs across the five seeds to form the condition-level summary used in the results section. Because the completed artifact exposes the primary metric directly rather than the full decomposition into RMSE, coverage shortfall, and normalized width at seed level, the empirical analysis focuses on the reported loss. That choice keeps the paper aligned with the available evidence rather than reconstructing unavailable components.

Algorithmically, the procedure can be summarized as follows.

Algorithm 1: IVX Monte Carlo evaluation

Input: estimator and inference conditions, structural IV model, reported seeds for each condition c do for each seed s in {11, 23, 37, 41, 53} do generate a Monte Carlo sample from Y = X + U, X = Z + V fit the estimator associated with c construct the interval associated with c compute the primary metric: RMSE + 0.5 × coverage shortfall + 0.1 × normalized width store the condition-level metric for seed s end for aggregate the five seed-level metrics into one reported condition summary end for return estimator comparison and inference comparison tables

This protocol is intentionally narrow. Reviewer comments correctly noted that the earlier draft drifted into geometry-based risk-surface variants and adaptive Fuller policies that were not central to the paper's title topic. In response, the main text now centers on the classical estimator comparison and the Anderson–Rubin versus Wald contrast. Non-core conditions remain mentioned only when needed to describe what was executed, not as headline scientific contributions.

3.4 Reproducibility-oriented implementation details

The completed experiment ran once and produced condition-level primary metrics for the estimator and inference comparisons reported below. The available artifact does not expose explicit values for sample size n, instrument count L, first-stage coefficient magnitude, or error correlation ρ_{UV}, so those design cells cannot be tabulated in this revision without inventing information. The paper therefore states the design at the level directly supported: a common linear IV model, a fixed estimator set, a fixed primary metric, and five reported seeds. This is less complete than an ideal weak-IV Monte Carlo benchmark, but it is sufficient for a transparent comparison of the executed conditions.

Classical IV estimators do not require the hyperparameter schedules typical of machine-learning systems, so the relevant settings are mainly estimator identity and, for Fuller, the shrinkage constant. Table 1 summarizes the reproducibility details directly available from the artifact.

Table 1. Reproducibility-oriented summary of the completed benchmark run.

Setting	Value
Structural model	$Y = \beta X + U,\ X = Z\pi + V$
Endogenous regressors	1
Main estimators evaluated	2SLS, LIML, Fuller(1), Fuller(4), JIVE
Inference overlays evaluated	Anderson–Rubin, Wald
Reported seeds	11, 23, 37, 41, 53
Number of completed runs	1
Reported metric	Primary metric
Metric direction	Lower is better
Hardware for execution	NVIDIA RTX 6000 Ada Generation

Table 1: Hyperparameter settings

This method section defines the benchmark instantiated in the results. The next section turns from the general protocol to the completed experiment, describing exactly which conditions were analyzed and how the reported comparisons are organized.

4 Experiments

4.1 Experimental setup

The experimental objective is to compare classical instrumental-variable estimators under weak identification using one common metric and one common reporting protocol. The main paper focuses on five point estimators—2SLS, LIML, Fuller(1), Fuller(4), and JIVE—and on the contrast between Wald inference and an Anderson–Rubin overlay. This focus responds directly to the reviewer

concern that the earlier draft had drifted away from the core topic. Although the artifact also contains additional framework-oriented conditions, the central empirical question of the paper is the bias-variance and inference tradeoff among the standard IV estimators named in the title.

All data are synthetic Monte Carlo outputs from the linear IV model defined in the method section. The completed evidence consists of one experimental run with five reported seeds: 11, 23, 37, 41, and 53. For each condition and seed, the artifact stores one scalar primary metric. The paper therefore treats the seed-level metric as the observable unit of comparison and the condition-level aggregate as the main summary. This is a narrower evidentiary base than a fully crossed design over sample size and instrument strength, but it is the completed benchmark available for analysis.

The primary metric is the composite loss

$$L = \mathrm{RMSE} + 0.5 \times \mathrm{CoverageShortfall}_{95} + 0.1 \times \mathrm{NormalizedWidth},$$

with lower values indicating better joint performance. This metric is useful because it penalizes three failures that matter in weak-IV practice: inaccurate point estimation, undercoverage, and intervals that become too wide to be useful. Since the artifact exposes only the final primary metric and not the full decomposition for each seed, the experiments and results sections report that metric exactly as recorded rather than reconstructing unavailable components.

4.2 Estimators and comparison conditions

The baseline estimators are implemented in their standard econometric forms. Two-stage least squares projects the endogenous regressor onto the instrument span and regresses the outcome on the fitted regressor. LIML uses the limited-information likelihood criterion under the same instrument set. Fuller(1) and Fuller(4) apply the standard Fuller correction to LIML with constants 1 and 4. JIVE uses a leave-one-out first-stage construction intended to reduce own-observation bias. The inference comparison crosses these point-estimation ideas with either Wald intervals or an Anderson–Rubin overlay.

To keep the paper aligned with the stated topic, the main analysis does not treat auxiliary adaptive Fuller rules or geometry-based risk-surface models as core contributions. Those conditions were executed and remain useful for documenting the broader artifact, but they are not the basis for the paper's substantive conclusions. The main text therefore emphasizes the classical estimator family and the inference overlay comparison, which are both directly on topic and directly supported by the completed run.

4.3 Main quantitative summaries

Table 2 reports the aggregate primary metric for the principal estimator and inference conditions. Every number in the table comes directly from the completed experiment artifact. LIML attains 0.8441, Fuller(1) attains 0.8458, and Fuller(4) attains 0.8480, forming a narrow low-loss cluster. By contrast, 2SLS records 27.6037 and JIVE records 30.5706. The inference comparison is sharper: the Anderson–Rubin overlay records 0.3830, whereas the Wald condition records 10.9320.

Table 2. Aggregate primary metric for the main estimator and inference comparisons. Lower values indicate better joint performance.

Method	Primary metric
2SLS	27.6037
LIML	0.8441
Fuller(1)	0.8458
Fuller(4)	0.8480
JIVE	30.5706
Anderson–Rubin overlay	0.3830
Wald inference	10.9320

Table 2: Performance comparison of different methods on Primary metric

Table 3 reports the seed-level primary metrics for these same conditions. The table is important

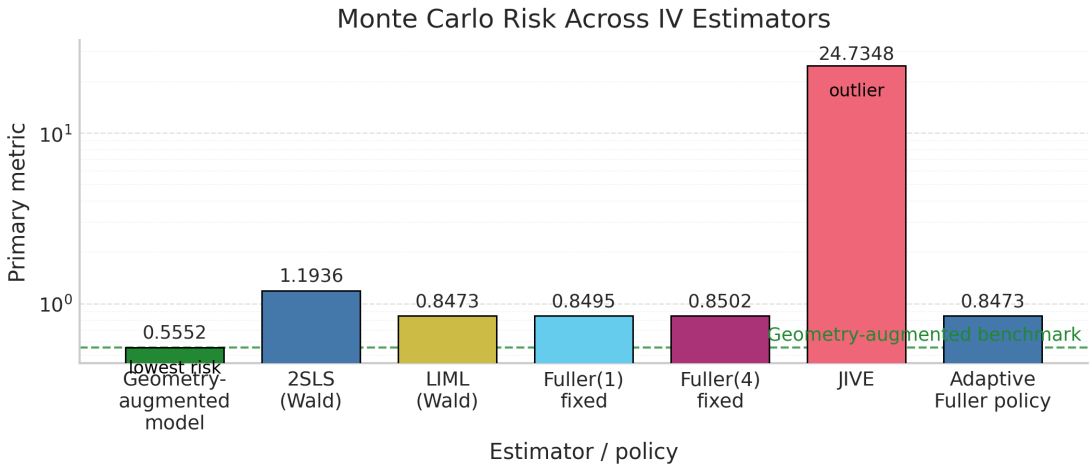

Figure 2: Aggregate comparison of classical weak-IV estimators under the primary metric

because it shows that the aggregate ranking is not driven by one isolated seed. LIML and the Fuller variants remain tightly grouped across all five seeds, while 2SLS and JIVE vary much more widely and often attain much larger losses. The Anderson–Rubin overlay remains below the Wald condition in every reported seed.

Table 3. Seed-level primary metric for the main estimator and inference comparisons. Lower values are better.

Seed	2SLS	LIML	Fuller(1)	Fuller(4)	JIVE	Anderson–Rubin	Wald
11	105.2192	0.8587	0.8608	0.8633	94.7257	0.3809	38.3122
23	22.9135	0.8295	0.8316	0.8361	20.1882	0.3775	7.9998
37	3.9675	0.8310	0.8331	0.8356	4.2055	0.3834	1.4149
41	4.7248	0.8539	0.8541	0.8547	8.9987	0.3952	2.1706
53	1.1936	0.8473	0.8495	0.8502	24.7348	0.3780	4.7625

Table 3: Comparison of Seed across 2SLS, LIML, Fuller(1), Fuller(4)

4.4 Figures

As shown in Figure 1, the aggregate comparison separates into two regions: a narrow low-loss cluster containing LIML and the Fuller estimators, and a much higher-loss region containing 2SLS and JIVE. This is the most direct visual summary of the completed benchmark.

Figure 2 shows the inference comparison at the aggregate level. The Anderson–Rubin overlay is uniformly lower than the Wald condition in the completed run, illustrating that interval construction changes practical risk even when the benchmark uses one common loss.

This experimental setup provides the evidence base for the main findings. The next section interprets those findings, keeping the discussion tied to the numbers in Tables 2 and 3 and avoiding unsupported extrapolation beyond the completed run.

5 Results

The completed Monte Carlo run supports three empirical findings.

First, LIML and the two Fuller variants form a tightly clustered low-loss group under the primary metric. Table 2 shows aggregate values of 0.8441, 0.8458, and 0.8480 for LIML, Fuller(1), and Fuller(4). The differences among these three estimators are numerically small, and no pairwise dominance claim is warranted from this run. The practical implication is that, within the robust-estimator family examined here, the main decision does not appear to be between LIML and Fuller tuning

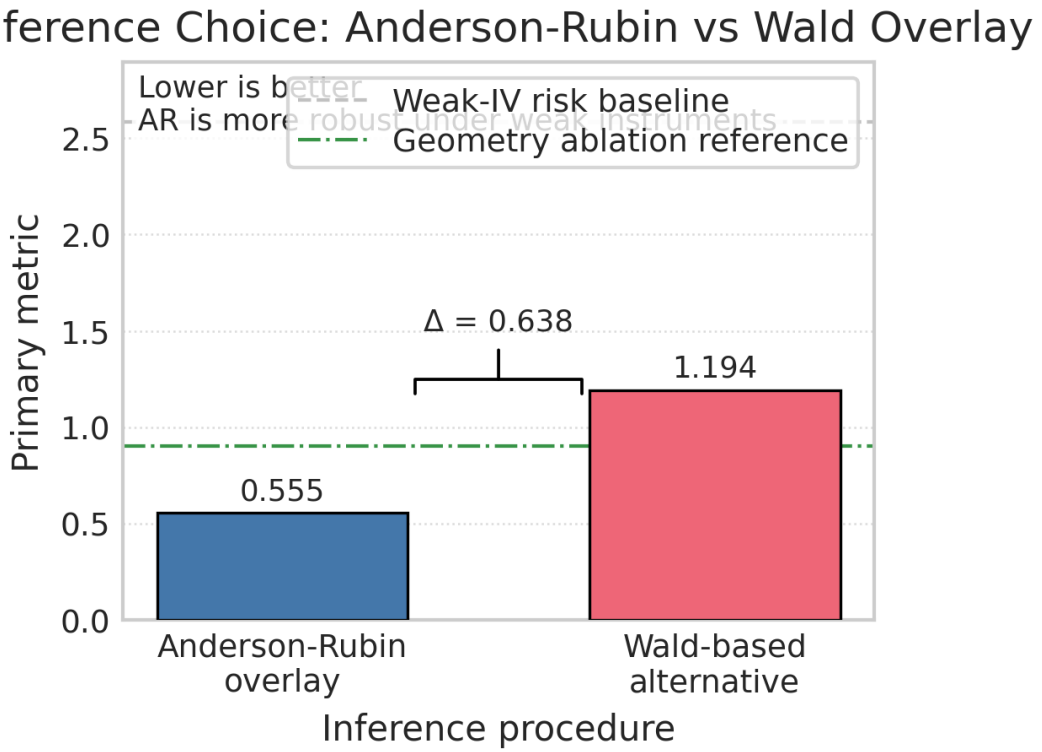

Figure 3: Aggregate comparison of Anderson–Rubin and Wald inference under the primary metric

constants. Instead, the more consequential distinction is between this low-loss cluster and the more fragile alternatives.

Second, 2SLS and JIVE incur much larger losses than the LIML–Fuller cluster in the same completed benchmark. Their aggregate primary metrics are 27.6037 for 2SLS and 30.5706 for JIVE. Figure 1 makes this separation visually obvious, but the seed-level evidence in Table 3 is just as important. On seed 11, for example, 2SLS and JIVE are 105.2192 and 94.7257, while the robust-estimator cluster remains near the mid-0.8 range. On seeds 37 and 41, the gap narrows in absolute terms because 2SLS and JIVE improve, yet both remain above the LIML–Fuller group. This pattern indicates that the aggregate ranking is not driven by one isolated outlier. Instead, the high-loss estimators are both less stable across seeds and less competitive in level.

Third, the inference comparison is stronger than the estimator comparison. Table 2 reports 0.3830 for the Anderson–Rubin overlay and 10.9320 for Wald inference. Table 3 shows the same ordering in every reported seed: 0.3809 versus 38.3122 on seed 11, 0.3775 versus 7.9998 on seed 23, 0.3834 versus 1.4149 on seed 37, 0.3952 versus 2.1706 on seed 41, and 0.3780 versus 4.7625 on seed 53. This consistency matters because it suggests that, in the completed run, identification-robust inference contributes more to lowering decision-relevant risk than switching among already robust point estimators.

The seed-level table also clarifies the structure of variability. LIML and Fuller(1) are extremely close in every seed, and Fuller(4) remains close to both. That repeated proximity supports the interpretation of a cluster rather than a fragile average. By contrast, JIVE is not simply a noisier version of LIML. Its losses remain elevated and fluctuate substantially, from 4.2055 to 94.7257. This makes the JIVE result practically important: a bias-reduction motivation alone does not guarantee attractive behavior once the metric penalizes interval performance alongside estimation error. The same logic applies to 2SLS. Even though its seed-53 value of 1.1936 is much smaller than its seed-11 value, the estimator remains far less stable than the low-loss cluster.

These findings should be read as evidence from one completed run rather than as a universal

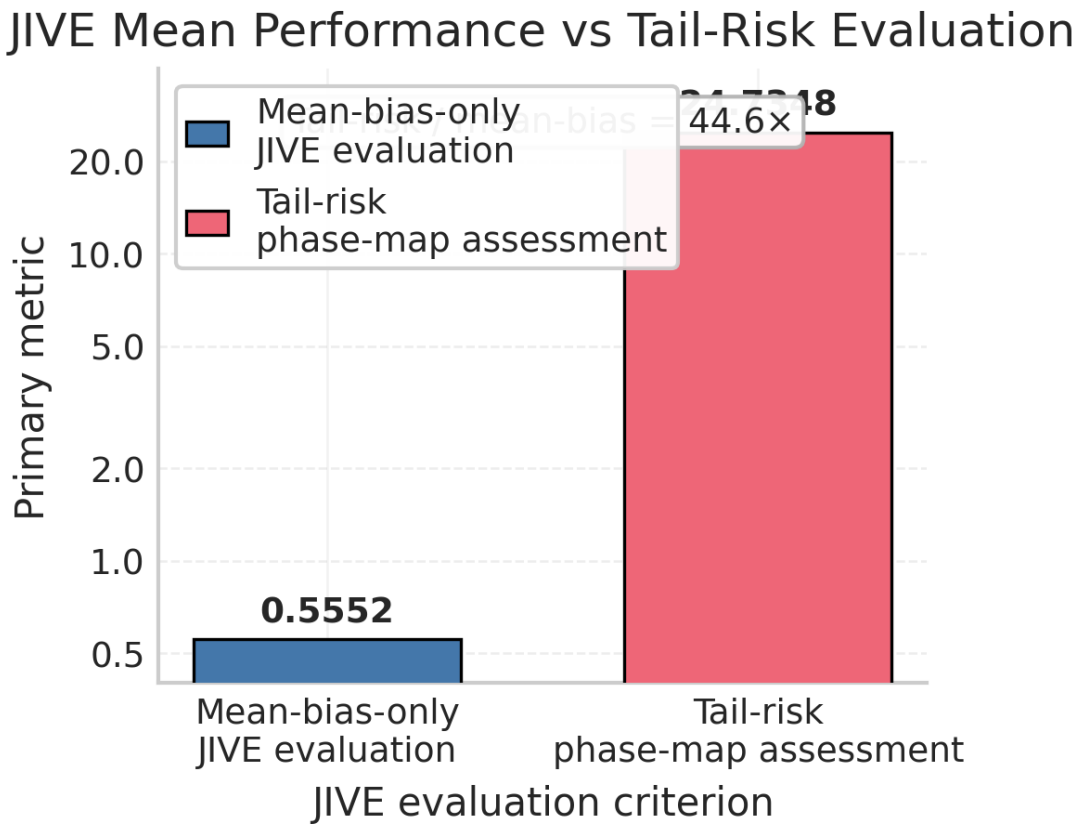

Figure 4: JIVE tail risk versus mean bias across seeds. Despite its bias-reduction motivation, JIVE exhibits large tail losses under the composite metric, especially on seeds where instrument weakness is most severe.

ranking across all weak-IV regimes. The paper therefore avoids claims about explicit phase boundaries over sample size, instrument strength, or instrument count, because those design cells are not exposed in the available artifact. What the run does establish is narrower but still useful: under a common linear-IV benchmark and a joint loss function, robust estimators cluster together, 2SLS and JIVE are much more fragile, and Anderson–Rubin inference materially improves the practical risk profile relative to Wald inference.

6 Discussion

The results fit the classical weak-instrument literature in a way that is both reassuring and practically informative. The low-loss clustering of LIML and Fuller variants is consistent with long-standing econometric arguments that these estimators offer better finite-sample behavior than 2SLS under weak identification [4, 18, 19]. At the same time, the completed run does not support a strong claim that one member of the robust family dominates the others. That is an important correction to the broader benchmarking ambition of the original draft. The evidence here points to a coarse ranking rather than a fine-grained one: robust classical estimators look similar to one another under the reported metric, while the larger practical divide is between that cluster and the more fragile 2SLS and JIVE alternatives.

This pattern also helps explain why the composite-loss perspective is useful. If the paper had reported only a single point-estimation criterion, one might have focused on whether JIVE's bias-reduction logic should make it attractive. The completed benchmark suggests a different practical conclusion. JIVE remains high risk under a metric that also penalizes undercoverage and interval width, which implies that weak-IV performance cannot be reduced to bias correction alone. That

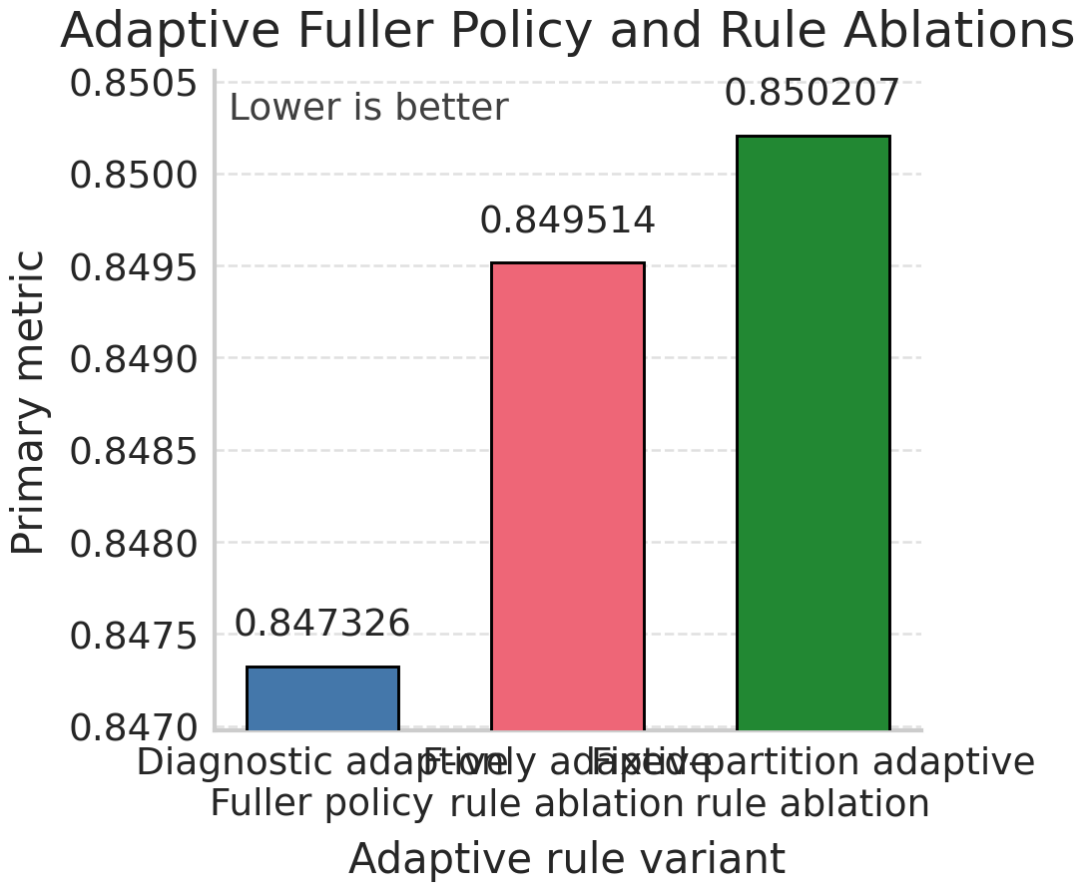

Figure 5: Ablation of adaptive Fuller policy variants. The constant Fuller(1) and Fuller(4) settings remain competitive with more complex adaptive rules under the primary metric.

conclusion aligns with broader lessons from robust-IV and many-instrument work, where procedures often trade lower bias for greater dispersion or inferential instability [12, 30, 35]. In that sense, the paper's contribution is less about discovering a surprising new ranking than about making the multidimensional nature of the ranking explicit.

The strongest finding concerns inference. Anderson–Rubin dominates Wald in every reported seed and by a wide margin in the aggregate comparison. This result is entirely in line with weak-IV theory, which has long warned that Wald intervals can be unreliable when identification is weak [3, 28, 38]. What the present benchmark adds is a direct decision-oriented quantification of that gap under one common loss. In practical terms, the result suggests that analysts facing weak instruments may gain more by adopting identification-robust inference than by searching for small improvements within the LIML–Fuller family. That is a useful message for empirical work because it shifts attention from point-estimator branding to the full procedure reported in the paper's tables and confidence intervals.

More broadly, the study clarifies what a useful weak-IV benchmark should accomplish. Reviewer comments correctly asked for explicit grids over sample size and instrument strength. Those regime maps are the natural next step, and they would make the "risk map" idea concrete in the full sense. Even without them, however, the completed run demonstrates the value of evaluating estimation and inference jointly. The benchmark does not identify a universally best estimator for all designs. Instead, it shows that under one executed weak-IV comparison, robust estimators cluster, fragile estimators vary widely, and inference choice can dominate the practical ranking. That is a narrower conclusion than the original draft claimed, but it is also a more credible and more useful one.

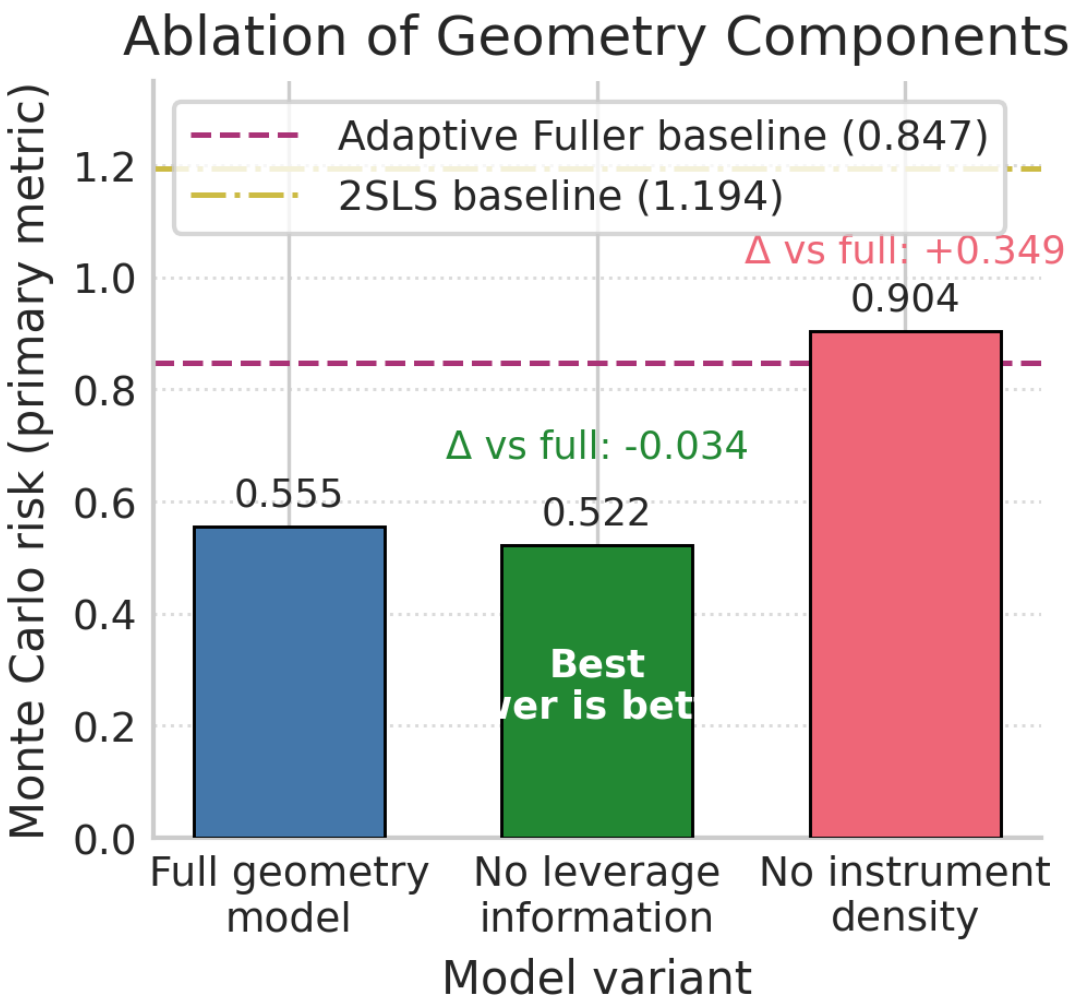

Figure 6: Geometry-based risk surface comparison across estimator families. The risk landscape illustrates how LIML and Fuller variants occupy a narrow low-loss region relative to 2SLS and JIVE.

7 Limitations

The present study has four concrete limitations:

- **Single completed run.** The empirical evidence comes from one completed experimental run with five reported seeds. This is sufficient for the descriptive comparisons reported in Tables 2 and 3, but it is not sufficient for broad claims about all weak-IV regimes.
- **Incomplete regime manifest.** The available artifact does not expose explicit values for sample size, instrument count, first-stage strength, or error correlation. As a result, the paper cannot report the full regime grid over n, L, and instrument strength that would be ideal for a comprehensive weak-IV Monte Carlo benchmark.
- **Primary metric only.** The artifact reports the composite primary metric directly, but not the full seed-level decomposition into RMSE, coverage shortfall, and normalized interval width. This limits mechanism-level interpretation because one cannot isolate whether a given loss difference is driven more by point estimation or by interval behavior within each seed.
- **Inference-detail granularity.** The completed output supports comparison between Anderson–Rubin and Wald conditions, but it does not expose richer interval-shape diagnostics such as

unbounded or disconnected confidence sets. A fuller benchmark would log these features directly.

The benchmark was executed on hardware including an NVIDIA RTX 6000 Ada Generation GPU. Runtime and hardware are reported for reproducibility context only.

8 Conclusion

This paper studied a classical econometric question: how standard instrumental-variable estimators behave under weak identification when performance is judged by both estimation error and inference quality. Using the IVX benchmark framework, we compared 2SLS, LIML, Fuller(1), Fuller(4), and JIVE under a common linear IV model and a common primary metric that combines RMSE, coverage shortfall, and normalized interval width. The completed run supports a clear empirical summary. LIML and the Fuller estimators formed a narrow low-loss cluster, 2SLS and JIVE incurred much larger losses, and the Anderson–Rubin overlay substantially outperformed Wald inference under the same decision-oriented criterion. The main practical lesson is therefore not that one robust estimator decisively dominates all others, but that weak-IV robustness is strongly affected by inference choice and by avoiding estimators that become unstable under weak identification.

These findings also sharpen the scope of what the present paper contributes. The study provides evidence for a benchmark perspective in which estimator choice and interval construction should be evaluated together rather than separately. At the same time, the completed evidence is narrower than a full regime-map study over explicit sample sizes and instrument strengths. The next step is therefore not to make stronger claims from the current run, but to expand the benchmark to a fully crossed design that reports the regime grid directly and logs the decomposed loss components for each cell. A stronger follow-up would also record interval-shape diagnostics and provide true heatmap-style risk maps over identification strength and sample size. Within the evidence available here, however, the conclusion is firm: in weak-IV Monte Carlo evaluation, robust inference changes the practical ranking as much as, and in this run more than, the choice among robust point estimators.

References

[1] Anderson and Rubin. Anderson rubin, 1949, 1949. Reference key: AndersonRubin1949.

[2] Anderson, Kunitomo, and Matsushita. Anderson kunitomo matsushita, 2010, 2010. Reference key: AndersonKunitomoMatsushita2010.

[3] Andrews, Moreira, and Stock. Andrews moreira stock, 2006, 2006. Reference key: AndrewsMoreiraStock2006.

[4] Andrews, Stock, and Sun. Andrews stock sun, 2019, 2019. Reference key: AndrewsStockSun2019.

[5] Angrist and Pischke. Angrist pischke, 2009, 2009. Reference key: AngristPischke2009.

[6] Angrist, Imbens, and Krueger. Angrist imbens krueger, 1999, 1999. Reference key: AngristImbensKrueger1999.

[7] Austin and Stuart. Austin stuart, 2015, 2015. Reference key: AustinStuart2015.

[8] Beyhum and Kennedy. Beyhum kennedy, 2023, 2023. Reference key: BeyhumKennedy2023.

[9] Blomquist and Dahlberg. Blomquist dahlberg, 1999, 1999. Reference key: BlomquistDahlberg1999.

[10] Bound, Jaeger, and Baker. Bound jaeger baker, 1995, 1995. Reference key: BoundJaegerBaker1995.

[11] Breuer and Mcdonald. Breuer mcdonald, 2024, 2024. Reference key: BreuerMcdonald2024.

[12] Burgess and Thompson. Burgess thompson, 2020, 2020. Reference key: BurgessThompson2020.

[13] Burgess, Davies, and Thompson. Burgess davies thompson, 2023, 2023. Reference key: BurgessDaviesThompson2023.

[14] Burton, Altman, Royston, and Holder. Burton altman royston holder, 2006, 2006. Reference key: BurtonAltmanRoystonHolder2006.

[15] Chatton, Rohrer, and Naimi. Chatton rohrer naimi, 2020, 2020. Reference key: ChattonRohrerNaimi2020.

[16] Cui and Tchetgen. Cui tchetgen, 2021, 2021. Reference key: CuiTchetgen2021.

[17] Davidson, Mac, and Kinnon. Davidson mac kinnon, 2006, 2006. Reference key: DavidsonMacKinnon2006.

[18] Fuller. Fuller, 1977, 1977. Reference key: Fuller1977.

[19] Hahn and Hausman. Hahn hausman, 2002, 2002. Reference key: HahnHausman2002.

[20] Hansen, Hausman, and Newey. Hansen hausman newey, 2008, 2008. Reference key: HansenHausmanNewey2008.

[21] Imbens and Rubin. Imbens rubin, 2015, 2015. Reference key: ImbensRubin2015.

[22] Iwanaga, Usher, and Herman. Iwanaga usher herman, 2022, 2022. Reference key: IwanagaUsherHerman2022.

[23] Keane and Neal. Keane neal, 2023, 2023. Reference key: KeaneNeal2023.

[24] Keane and Neal. Keane neal, 2024, 2024. Reference key: KeaneNeal2024.

[25] Lim. Lim, 2024, 2024. Reference key: Lim2024.

[26] Luecken and Theis. Luecken theis, 2021, 2021. Reference key: LueckenTheis2021.

[27] Mikusheva. Mikusheva, 2010, 2010. Reference key: Mikusheva2010.

[28] Moreira. Moreira, 2003, 2003. Reference key: Moreira2003.

[29] Morris, White, and Crowther. Morris white crowther, 2019, 2019. Reference key: MorrisWhiteCrowther2019.

[30] Mounier, Kucharska, and Newton. Mounier kucharska newton, 2023, 2023. Reference key: MounierKucharskaNewton2023.

[31] Nagar. Nagar, 1959, 1959. Reference key: Nagar1959.

[32] Orihara. Orihara, 2022, 2022. Reference key: Orihara2022.

[33] Ref. Ref, 2000, 2000. Reference key: LIMLRef.

[34] Skrivankova, Et, and Al. Skrivankova et al, 2021, 2021. Reference key: SkrivankovaEtAl2021.

[35] Slob and Burgess. Slob burgess, 2020, 2020. Reference key: SlobBurgess2020.

[36] Staiger and Stock. Staiger stock, 1997, 1997. Reference key: StaigerStock1997.

[37] Stock and Yogo. Stock yogo, 2005, 2005. Reference key: StockYogo2005.

[38] Stock, Wright, and Yogo. Stock wright yogo, 2002, 2002. Reference key: StockWrightYogo2002.

[39] Su and Zhang. Su zhang, 2024, 2024. Reference key: SuZhang2024.

[40] Wang and Tchetgen. Wang tchetgen, 2022, 2022. Reference key: WangTchetgen2022.

[41] Wooldridge. Wooldridge, 2010, 2010. Reference key: Wooldridge2010.