Instrument-Hacking*

Michael Keane† Timothy Neal‡ Patrick Vu§

July 1, 2026

Abstract

In instrumental-variable (IV) studies, researchers often evaluate multiple candidate instruments and selectively report the specification with the most favorable first- or second-stage statistics. We show that this form of instrument selection ("instrument-hacking") induces median bias in IV estimators toward the OLS estimand, undermining the rationale for using IV. When all candidate instruments have the same true strength, median bias increases monotonically with the number of available instruments. More generally, when candidate instruments differ in true strength, monotonicity need not hold because increasing the number of instruments can lead researchers to select genuinely stronger instruments. Nonetheless, in simulations calibrated to the empirical distribution of instrument strengths in the IV literature, median bias increases monotonically with the number of available instruments. Furthermore, it is substantial in magnitude, even when only a few instruments are available.

Keywords: Instrumental variables, p-hacking, 2SLS, GMM, causal inference, size distortion, median bias

---

*We thank seminar participants at the Australian National University, Deakin University, Oxford University, University College London, University of Melbourne, and University of Sydney for helpful comments.

†Carey School of Business and Department of Economics, Johns Hopkins University. Email: m.keane@unsw.edu.au

‡Department of Economics, University of New South Wales. Email: timothy.neal@unsw.edu.au

§Department of Economics, University of New South Wales. Email: patrick.vu@unsw.edu.au

Instrumental-variable (IV) studies in economics exhibit clear evidence of p-hacking and publication bias, while for randomized controlled trials, difference-in-differences, and regression discontinuity designs, such evidence is more limited (Brodeur et al., 2020; Kranz and Pütz, 2022). A possible explanation is that observational IV creates a unique researcher degree of freedom: researchers can test multiple candidate instruments and selectively report the specification yielding the 'most favorable' result. We refer to this practice as *instrument-hacking*. Consistent with this account, IV estimates in randomized controlled trials with partial compliance – where treatment assignment provides a unique, non-selectable instrument – exhibit relatively little evidence of p-hacking (Brodeur et al., 2020, Table A17).

In this paper, we study the impact of instrument-hacking on the credibility of just-identified 2SLS. We consider two forms of instrument selection. In *first-stage instrument-hacking* one selects among candidate instruments to maximize the first-stage F-statistic, consistent with the over-representation of F-statistics just above the conventional threshold of 10 in IV studies (Andrews et al., 2019). In *second-stage instrument-hacking* one selects among candidate instruments to maximize the second-stage t-statistic, consistent with discontinuities or clumping in the distribution of t-statistics around conventional significance thresholds (Brodeur et al., 2020).

These forms of selection need not reflect intent to manipulate results: researchers may regard searching for stronger instruments as a way to improve statistical credibility and address weak-instrument concerns. We show, however, that this seemingly reasonable practice can have important adverse consequences.

In the first set of theoretical results, we consider a single-equation linear IV model with one endogenous regressor, and a researcher that searches over multiple valid instruments of equal true strength. This stylized context, which we later generalize, is designed to clearly illustrate the core problem. When the true effect β is zero, we show that both first- and second-stage instrument-hacking induce median bias in the 2SLS estimator toward the OLS estimand.[1] Moreover, the magnitude of this bias increases monotonically with the number of available instruments.[2]

These results are noteworthy for two reasons. First, they imply that instrument-hacking undermines the central purpose of IV, as it reintroduces bias toward the OLS

[1]More specifically, first-stage instrument-hacking induces median bias toward OLS for any true effect β, whereas second-stage instrument-hacking requires $\beta = 0$.

[2]We show that similar distortions arise under selection rules that jointly target the first-stage F-statistic and the second-stage t-statistic.

estimand. This contrasts with OLS settings without endogeneity, where p-hacking typically produces exaggeration bias. Second, it is well known that 2SLS with many instruments suffers from bias and size distortions. This problem is considered so severe that researchers are sometimes explicitly advised to informally select on instruments, as in Angrist and Kolesár (2024) and Angrist and Pischke (2009, p.157). In fact, the latter advise: "Pick your best single instrument and report just-identified estimates using this one only." Our results show that this type of selection generates the very same distortions that arise from using many instruments.

Instrument-hacking leads to median bias toward the OLS estimand because of the "power asymmetry" problem of two-stage least squares (2SLS) – see Keane and Neal (2023, 2024). This refers to the fact that 2SLS standard errors tend to be smaller in samples where the 2SLS estimate is close to the OLS estimate, a pattern that holds even when instruments are strong. Thus, when researchers using 2SLS engage in instrument-hacking, reported estimates tend to be those with smaller standard errors, and these estimates tend to be shifted toward the OLS estimand.

In our second set of theoretical results, we relax the stylized 'equal instrument strength' assumption by allowing instrument strengths to follow an arbitrary bounded distribution. This more realistic setting yields results that are more general, but necessarily less sharp. Under either form of instrument-hacking, we show the median of the 2SLS estimator is biased toward the OLS estimand, and the magnitude of the bias is bounded below by an increasing function of the number of available instruments. However, bias need not be monotonically increasing, particularly if the empirical distribution contains many weak instruments and one very strong instrument.

To address this theoretical ambiguity, we assess what instrument-hacking implies under an empirically plausible distribution of instrument strengths. To calibrate this distribution, we collect data on IV regressions published in 800 papers in top 5 economics journals (excluding *Econometrica*) from 2011–23. Notably, this is 22.5% of all papers, highlighting the practical importance of IV. We then use the Andrews and Kasy (2019) method to extract the underlying distribution of true instrument strengths from the observed distribution of first-stage sample F statistics. Finally, we simulate instrument-hacking when candidate instruments are drawn from this empirical distribution. In this setting, median bias increases monotonically with the number of available instruments for both first- and second-stage instrument-hacking.

The simulation exercise also allows us to assess the magnitude of bias induced by instrument hacking in a realistic setting. For second-stage instrument-hacking, we find bias is quite severe even if researchers only consider a small number of instruments. For example, when researchers selectively report the just-identified specification with the largest second-stage t-ratio among three available instruments, median bias reaches 0.12 standardized effect size units.[3] This magnitude is comparable to typical treatment effect sizes reported across a range of economics literatures (Evans and Yuan, 2022; Bartoš et al., 2024), making the implied bias economically substantial.[4] Targeting the first-stage F-statistic alone results in a smaller, though still economically meaningful, median bias of around 0.03 SD.

Notably, both first and second-stage instrument hacking also induce severe size distortions in 2SLS t-tests. Furthermore, the overwhelming majority of false rejections occur in the direction of the OLS bias due to the power asymmetry problem. Overall, our theoretical and empirical results show that instrument-hacking can pose a serious threat to credibility, even when only a small number of instruments are observable. In fact, our results show that the problems created by instrument-hacking can be at least as severe as those created by using 2SLS with many instruments.

Given these problems, we recommend that researchers refrain from instrument selection altogether, and instead use all available instruments in conjunction with robust estimation and inference. Under homoskedasticity, the optimal choice would be the limited information maximum likelihood (LIML) estimator of Anderson and Rubin (1949)) combined with the Conditional Likelihood Ratio (CLR) test for inference (Moreira, 2003). We show this approach delivers median-unbiased estimates, correct size, and symmetric rejection rates under the null.[5]

We also consider the impact of instrument hacking using the AR test of Anderson and Rubin (1949). In simulations where one targets significance of the AR test, rather than the second-stage t-statistic, median bias is ameliorated, though not eliminated. Size distortions remain, but rejections are roughly balanced on either side of the

[3] That is, given the true null $\beta = 0$, the median effect estimate is that a one SD increase in the endogenous variable x causes a 0.12 SD increase in the outcome y.

[4] Evans and Yuan (2022) find a median effect size of 0.1 SD across 234 studies of educational interventions and test scores. Similarly, Chetty et al. (2014) estimate that a one standard deviation increase in teacher quality raises end-of-year test scores by 0.13 standard deviations. More broadly, Bartoš et al. (2024) summarize 327 economics meta-analyses and find a median reported effect size of 0.2 SD across a range of literatures.

[5] Given heteroskedasticity, the analogous approach is the continuously updated GMM of Hansen et al. (1996), which generalizes LIML to the heteroskedastic case, combined with the Kleibergen (2005) test that generalizes CLR.

null hypothesis, as the AR test removes the major source of the correlation between effect sizes and standard errors (Keane and Neal, 2024). On the other hand, first-stage instrument-hacking breaks this desirable property of the AR test, compromising its key advantage over the t-statistic. Thus, while robust inference can ameliorate problems created by instrument hacking, it is not a panacea.

This paper connects two large literatures: the literature on p-hacking and publication bias (Franco et al., 2014; Brodeur et al., 2016, 2020; Andrews and Kasy, 2019; DellaVigna and Linos, 2022) and the literature on instrumental variables (Bound et al., 1995; Stock and Yogo, 2005; Donald and Newey, 2001). Existing work on instrument selection in econometrics has largely focused on the prescriptive question of what researchers *should* do when multiple instruments are available (Donald and Newey, 2001). By contrast, this paper studies the question of how actual researcher behavior affects estimation and inference, motivated by evidence of publication bias and p-hacking in observational IV (Brodeur et al., 2020). In particular, it provides the first systematic analysis of how p-hacking interacts with the power asymmetry of IV estimators to generate bias toward the OLS estimand.

Our work is also related to the literature on pre-screening based on the first-stage F statistic. Simulation studies have found that such screening can exacerbate weak-instrument bias and size distortions (Hall et al., 1996; Zivot et al., 1998; Angrist and Kolesár, 2024). Building on this literature, we provide theoretical conditions under which first-stage instrument-hacking generates median bias and identify the mechanism: the correlation between IV estimates and their standard errors. But, to our knowledge, the results we present for second-stage instrument-hacking are completely new to the literature.

This paper proceeds as follows. Section 1 introduces the model and Section 2 illustrates instrument-hacking in the context of estimating the returns to schooling using 2SLS. Section 3 presents theoretical results examining the impact of hacking on median bias. Section 5 presents results for the calibrated model. Section 6 makes recommendations for applied researchers, and Section 7 concludes.

1. Setup

In this section, we introduce the model used throughout the paper and provide a formal definition of instrument-hacking. Our main aim is to study the impact of instrument-hacking on the statistical credibility of IV estimators as the number of instruments to select from increases.

1.1 Data-Generating Process

Consider a standard single-equation linear IV model with one endogenous regressor:

$$\begin{aligned} y_i &= \alpha + \beta x_i + u_i \\ x_i &= \delta + \sum_{k=1}^{\bar{K}} \pi_k z_{ki} + \rho u_i \end{aligned} \tag{1}$$

The parameter $\rho \in (-1, 1)$ measures the degree of endogeneity. The case $\rho = 0$ corresponds to exogeneity. Since our interest is in bias arising from endogeneity, we focus on the case $\rho \neq 0$. In the equation for x_i, there are $\bar{K}$ instruments, where $\bar{K}$ is a fixed parameter. The researcher observes a subset of $K \leqslant \bar{K}$ instruments, while the remaining $(\bar{K} - K)$ instruments are unobserved and absorbed in the error term. We assume throughout that the instruments are valid. The true strength of instrument k depends jointly on its first-stage coefficient π_k, its variance and N. The true strength of an instrument can only be estimated in a sample, by the sample F.

1.2 Instrument-Hacking

Instrument-hacking is conceptualized as a selection rule that determines which instrument from a candidate set is reported. We focus on rules that select among just-identified IV specifications for two reasons. First, this restriction provides greater analytical tractability for our theoretical results. Second, just-identified specifications are often favored in applied work due to (i) their favorable finite-sample properties in the absence of p-hacking (Angrist and Kolesár, 2024), and (ii) the relative ease of justifying the exclusion restriction and interpreting the resulting LATE with a single instrument.[6]

Each just-identified specification associated with observed instrument z_k generates a 2SLS estimate, a first-stage sample F-statistic and a second-stage t-ratio: $(\hat{\beta}_k, F_k, t_k)$ for $k = 1, 2, ..., K$. Instrument-hacking is defined as a selection rule that determines which specification is reported, which we denoted by $(\hat{\beta}_{k*}, F_{k*}, t_{k*})$.

Definition 1 (Just-Identified Instrument-Hacking)**.** The reported just-identified specification $(\hat{\beta}_{k*}, F_{k*}, t_{k*})$ satisfies $k^* \in \arg\max_{k\in\{1,...,K\}} s_k$, where s_k is the score.

The choice of score (s_k) determines the selection rule and hence the form of

[6]Online Appendix OA.A shows that allowing for over-identified specifications for second-stage hacking worsens size distortion and slightly ameliorates median bias.

instrument-hacking. We consider two main forms: (i) *first-stage instrument-hacking*, where $s_k = F_k$; and (ii) *second-stage instrument-hacking*, where $s_k = |t_k|$. Incentives to select on F_k and $|t_k|$ are well understood by researchers, and empirical evidence of selection on both statistics is well documented (Ioannidis, 2005; Andrews and Kasy, 2019; Brodeur et al., 2020). In Subsection 3.2, we also consider selection rules that jointly target both statistics.

2. Illustrative Example

This section illustrates the mechanism through which instrument-hacking affects the statistical properties of 2SLS estimators in a simple numerical example. Using the model in (1), suppose the true effect is zero ($\beta = 0$), but the endogenous regressor is positively correlated with the structural error, with $\rho = 0.5$, so the OLS bias is positive. Normalize $\mathrm{Var}(x_i) = 1$, so the OLS estimand is .50. Furthermore, to isolate the effects of instrument selection, suppose all $\bar{K} = 10$ candidate instruments are equally strong binary instruments, each assigning treatment with probability 0.5 and having first-stage effect $\pi = \sqrt{0.3}$. The strength of the first stage, as summarized by the true F-statistic, is jointly determined by π, the assignment probability, and N. Letting N=350, the implied true first-stage F-statistic in each just-identified specification is 28.4, which is well above conventional weak-instrument thresholds.[7]

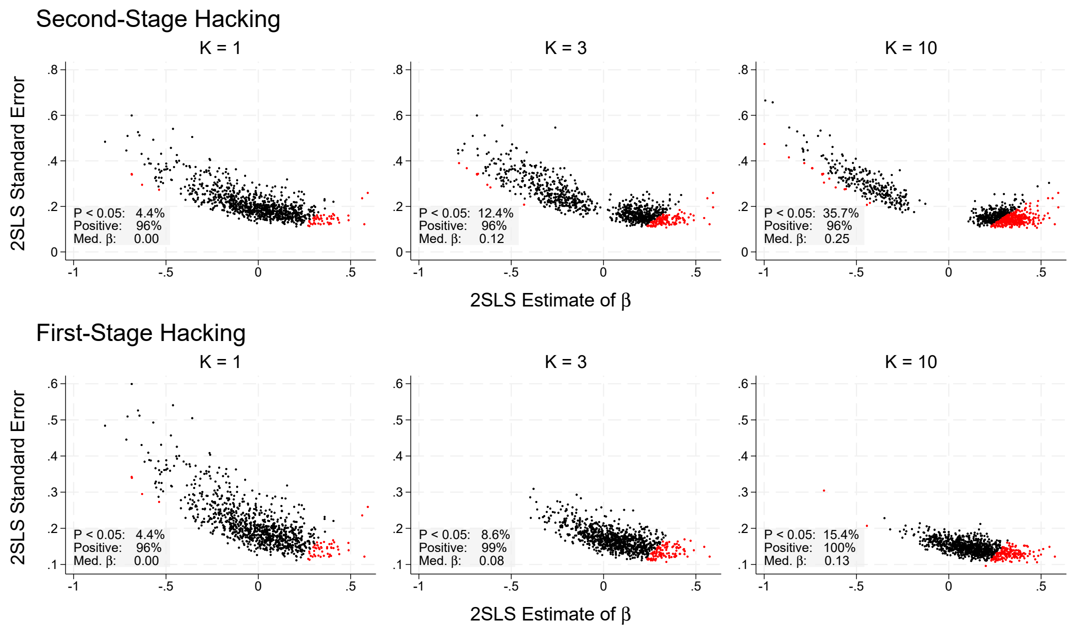
Now suppose a researcher can observe $K \geqslant 1$ valid instruments from the $\bar{K} = 10$ total available, and reports the just-identified specification to maximize either the sample second-stage t-statistic or the sample first-stage F-statistic. How does the number of available instruments, K, affect the median bias and other properties of the 2SLS estimator?

Figure 1 presents results for second-stage instrument-hacking ($s_k = |t_k|$) in the top panel and first-stage instrument-hacking ($s_k = F_k$) in the bottom. Each figure plots the joint distribution of the 2SLS estimate and its standard error for different numbers of available instruments (i.e. $K = 1, 3, 10$). Significant estimates according to a 5% level t-test are coded in red. In the bottom left-hand corner of each figure, we report the median, size and the fraction of rejections that are positive.

First consider the second-stage hacking results in the top panel. When $K = 1$, instrument-hacking cannot occur. And because instruments are strong, median bias

[7]In a just-identified regression, the true $F = N\frac{R^2}{1-R^2}$. Here, $R^2 = (\pi^2/4)/\,\mathrm{Var}(x) = \pi^2/4 = 0.075$; so with $N = 350$, we obtain $F = 350 \times \frac{0.075}{0.925} = 28.38$. Of course, the sample F is a random variable.

FIGURE 1. Instrument-Hacking: Joint Distribution of $\hat{\beta}$ and Standard Errors



Notes: 2SLS estimates and standard errors when a researcher engages in instrument-hacking among just-identified specifications over 100,000 replications. In all cases, $\beta = 0$, $\pi_k = \sqrt{0.3}$ and $\rho = 0.5$. Red dots indicate statistically significant estimates at the 5% level. 'P $<$ 0.05' refers to the rejection rate when the $\hat{\beta}$ p-value is below the 0.05 threshold. 'Positive' refers to the proportion of total rejections that feature a $\hat{\beta} > 0$ (i.e. the estimate is the same sign as the biased OLS estimate). 'Med. β' refers to the median estimated β, a measure of median bias in the estimator.

is negligible and the size of the 5% test is close to its nominal level. When $K \geqslant 1$ however, the researcher can select the just-identified specification with the largest second-stage t-statistic. With $K = 3$, this raises median bias to 0.12, in the direction of the OLS bias, and increases the size of the 5% t-test to 12.4%, with nearly all rejections on the positive side. As a result, researchers become more likely to reject the null hypothesis and incorrectly conclude that there is a positive causal effect. With $K = 10$, these effects are even more pronounced.

The mechanism driving these results is the correlation between $\hat{\beta}$ and $SE(\hat{\beta})$ inherent in 2SLS, which Keane and Neal (2023, 2024) term the 'power asymmetry problem.' In each plot, 2SLS standard errors tend to be smaller in samples where the 2SLS estimate is close to the OLS estimand.[8] Hence, if $K \geqslant 1$, results with small 2SLS estimates and large standard errors tend to be censored from the joint distribution of reported results, leading to bias in the direction of OLS.

[8]If OLS bias were of the same magnitude but negative ($\rho = -0.5$), the joint distribution would be reflected about $\hat{\beta} = 0$ and instrument-hacking would push median bias in a negative direction.

Next consider the first-stage hacking results in the bottom panel of Figure 1. This distorts results in a different way. In particular, as the number of available instruments increases, results with large 2SLS standard errors are censored from the distribution of reported results. This occurs because the 2SLS standard error is inversely proportional to the first-stage sample F. Combined with the power asymmetry of 2SLS estimators (i.e., estimates closer to OLS have smaller standard errors), this means that reported estimates of the causal effect are again biased toward the OLS estimand, leading to distortions similar to those under second-stage hacking.

Thus, Figure 1 illustrates the key mechanism studied in this paper. Both first and second-stage instrument-hacking systematically shift reported 2SLS estimates toward the OLS estimand, and this distortion becomes more pronounced as the number of available instruments increases. This pattern differs from more familiar forms of p-hacking in OLS settings, where selection on statistical significance typically generates exaggeration bias, pushing estimates away from zero. By contrast, instrument-hacking biases 2SLS estimates toward the OLS estimand, thereby undermining the primary purpose of IV. The next section formalizes these insights and establishes general results on the direction and magnitude of bias from instrument-hacking.

3. Theory: Median Bias with Homogeneous Instruments

Under what conditions do the patterns illustrated in the example hold more generally? We address this question with two sets of theoretical results. In this section we consider a homogeneous-instrument setting, in which candidate instruments are binary and have identical first-stage effects, $\pi_k = \pi$ for $k = 1, \dots, \bar{K}$. Then later, in Section 4, we allow instrument strengths to differ across candidate instruments, yielding more general results that characterize a lower bound on the evolution of median bias. Proofs are provided in the Appendix, unless otherwise stated.

To begin, we state assumptions common to both sets of results:

Assumption 1 (Structural Error Term)**.** Let u_i be i.i.d. and follow a continuous, unbounded distribution that is symmetric about zero and strictly log-concave on its support.

Assumption 2 (Gaussian Approximation for Reduced-Form and First Stage)**.** For each instrument $k = 1, \dots, \bar{K}$, let $\hat{\gamma}_k$ and $\hat{\pi}_k$ denote the OLS coefficients from the reduced-form regression of y_i on $(1, z_{ki})$ and the first-stage regression of x_i on $(1, z_{ki})$, respectively. Assume $\pi_k \neq 0$ and that

$$\sqrt{N}\begin{pmatrix} \hat{\gamma}_1 - \gamma_1 \\ \hat{\pi}_1 - \pi_1 \\ \vdots \\ \hat{\gamma}_{\bar{K}} - \gamma_{\bar{K}} \\ \hat{\pi}_{\bar{K}} - \pi_{\bar{K}} \end{pmatrix} \sim \mathcal{N}(0, \Sigma),$$

where Σ is a known positive-definite covariance matrix.

Assumption 3 (Sign Screening on Estimated First Stage)**.** Researchers screen for IV results where the sign of the estimated first-stage matches the sign of the true first-stage.

Assumption 4 (Designed Orthogonality of Candidate Instruments)**.** The set of candidate instrument assignment vectors is constructed so that, for every pair $k \neq k'$, $\widehat{\mathrm{Cov}}(z_k, z_{k'}) = 0$.

Assumption 1 is relatively mild and is satisfied by many common error distributions (e.g., normal, logistic, Gumbel). Assumption 2 follows Andrews et al. (2019) and Angrist and Kolesár (2024) by imposing a Gaussian approximation for the reduced-form and first-stage estimators and treating their covariance matrix as known.[9] Given this, the asymptotic variance of the 2SLS estimator for the just-identified specification using instrument z_k can be obtained by applying the delta method to the ratio $\hat{\beta}_k = \hat{\gamma}_k / \hat{\pi}_k$. This variance is a quadratic function of $\hat{\beta}_k$ and is minimized at the population OLS estimand. Mechanically, this implies that the 2SLS standard error tends to be smaller when the 2SLS estimate is closer to the OLS estimand. This has important implications for instrument-hacking because any selection rule that favors small standard errors will tend to select estimates closer to the OLS estimand.

Assumption 3 follows Angrist and Kolesár (2024) and Andrews and Armstrong (2017) in assuming researchers screen the estimated first-stage to have the same sign as the true first-stage. This is a mild assumption in practice. For example, consider a design where the instrument assigns individuals to receive an encouragement to take up a treatment, such as a subsidy. If the estimated first-stage coefficient is negative in

[9] Relative to the standard single-instrument setup, our formulation extends the approximation jointly across the full set of candidate instruments. This extension is natural in our setting because the object of interest is selection across many candidate IV specifications.

a given sample due to sampling variation, researchers would typically conclude that the encouragement failed and discard the corresponding specification.

Assumption 4 can be understood as a design restriction that forces instruments to be independent in each sample. Its role is to eliminate mechanical cross-instrument interactions arising from correlations among the candidate instruments, so that each just-identified specification depends only on its own instrument vector and the common structural error. This assumption can be weakened, but we maintain it for the sake of analytical transparency and parsimony.[10]

The empirical model in Section 5 does not impose Assumptions 2 and 4.[11] Its covariance structure is estimated from the data, and candidate instruments are allowed to be correlated in finite samples. As a result, it provides a direct assessment of the relevance of the theoretical mechanism under more realistic conditions, where the simplifying assumptions outlined here are relaxed. The fact that the main qualitative patterns persist in these settings suggests that the results are not driven by these assumptions, but instead reflect a more general feature of instrument-hacking arising from the power asymmetry of the 2SLS estimator.

3.1 Median Bias with Homogeneous Instruments

The first set of theoretical results relies on the following additional assumption:

Assumption 5 (Homogeneous Instruments)**.** Let all $k = 1, 2, \ldots, \bar{K}$ instruments be binary, $z_{ki} \in 0, 1$, with common first-stage effect $\pi_k = \pi \neq 0$ and common assignment probability $q_k = q \in (0, 1)$, where $q_k \equiv \frac{1}{N} \sum_{i=1}^{N} z_{ki}$. Finally, assume $z_{ki} \perp\!\!\!\perp u_i$ for all k.

Under Assumption 5 each candidate instrument can be interpreted as a distinct but equally strong encouragement design for treatment x_i. This is a strong assumption that approximates settings where several instruments of similar strength are available. Nevertheless, it allows us to clearly demonstrate the mechanisms through which instrument hacking can generate bias. This assumption is later relaxed in Section 4, where we allow for unequal instrument strengths.

With this we can state the first set of results:

[10] The same result can be obtained by assuming that finite-sample correlations across instruments are sufficiently small (e.g. due to independent instruments and a sufficient large sample size) so as to never overturn the ordering of t-ratios or F-statistics across just-ID specifications.

[11] The estimation procedure does, however, retain first-stage sign screening. For more details, see Section 5.1.

Proposition 1 (Second-Stage Instrument-Hacking and Median Bias)**.** Let $\beta = 0$. Under Assumptions 1–5 and second-stage instrument-hacking, the median of the reported 2SLS estimator is biased toward the OLS estimand for any $K \geqslant 1$, and is strictly increasing in absolute value with the number of available instruments K.

Proposition 2 (First-Stage Instrument-Hacking and Median Bias)**.** Under Assumptions 1–5 and first-stage instrument-hacking, the median of the reported 2SLS estimator is biased toward the OLS estimand for any $K \geqslant 1$, and is strictly increasing in absolute value with the number of available instruments K.

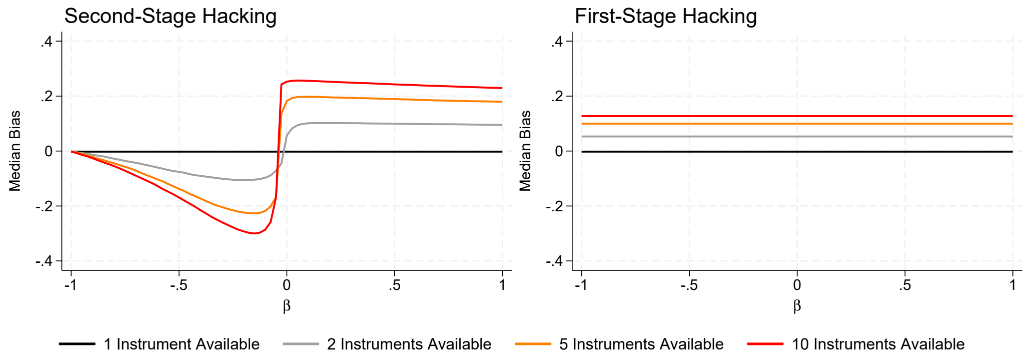
That is, with homogeneous instruments, instrument-hacking generates bias in the direction of OLS, and the magnitude of that bias increases monotonically with the number of instruments available to select from. This generalizes patterns observed in the illustrative example, by showing that this outcome occurs irrespective of instrument strength (π); the degree of endogeneity ($\rho \in (-1, 1)$); and, for first-stage instrument-hacking, any true treatment effect ($\beta \in \mathbb{R}$).

The intuition for these results is simple: If instruments have equal true strength, the one that appears to be 'strongest' in a finite sample is that with the greatest finite sample correlation with the structural error u. If that correlation is relatively large, the sample F is relatively large and the 2SLS estimate is shifted toward $\hat{\beta}_{OLS}$. Hence first-stage hacking generates median bias towards OLS. This 'strongest' instrument also generates a relatively small $se(\hat{\beta}_{2SLS})$, and hence a relatively large t-statistic. So second-stage hacking generates median bias towards OLS as well.

It is noteworthy that the second-stage instrument hacking result in Proposition 1 assumes $\beta = 0$. Unlike first-stage hacking, the direction of bias under second-stage hacking depends on both the OLS estimand and the true β. We illustrate the distinction in Figure 2, which is based on simulations of the DGP in Section 2, except here we vary true β, so the OLS estimand is $\beta + 0.50$. The figure plots bias from instrument hacking as a function of the true β. Under first-stage instrument-hacking, we see that bias is always positive (as the OLS bias is positive), independent of true β, and increasing in the number of instruments available.

For second-stage hacking the situation is more complicated. As shown in Figure 2, there exists a critical negative value of β at which the direction of median bias reverses. For values of β above this threshold, median bias is toward the OLS estimand, whereas for sufficiently negative values below it, median bias becomes negative. This occurs

FIGURE 2. Median Bias under Instrument-Hacking as β Varies



Notes: Median bias of the 2SLS estimator when the researcher instrument-hacks among just-identified specifications. The DGP is outlined in Section 2, with $\rho = 0.5$, and $\pi = \sqrt{0.3}$.

because when $\beta < 0$ the estimates that are shifted towards OLS are also closer to zero, so they are likely to have small $|t_k|$ statistics, making them unlikely to be selected. Consequently, for true β below the critical value, the conventional exaggeration bias effect dominates, and the bias is negative.

3.2 Joint Preferences

Propositions 1 and 2 analyze instrument-hacking when selection targets the second-stage t-ratio or the first-stage F-statistic in isolation. In practice, however, researchers and journals may evaluate IV specifications based on both criteria simultaneously. Consistent with this view, empirical evidence documents an over-representation of studies reporting first-stage F-statistics above 10 together with second-stage t-ratios exceeding 1.96 (Andrews et al., 2019; Brodeur et al., 2016, 2020).

This subsection shows that the results in the previous subsection extend naturally to settings with joint preferences. Recall that instrument-hacking in Definition 1 is defined as a selection rule that reports the just-identified specification maximizing a target score s_k. We consider two forms of joint preferences, each corresponding to a different specification of the score function s_k.

The first is *smooth preferences*, where researchers seek to maximize a weighted score of both statistics: $s_k = \lambda F_k + (1 - \lambda)\,|t_k|$, where $\lambda \in (0, 1)$. The second is *lexicographic-threshold preferences*, where researchers aim initially target a first-stage threshold, such as $F_k \geqslant c$. If no available specification clears the threshold, the specification with the largest first-stage F-statistic is reported. If one or more specifications do clear the threshold, they restrict attention to those specifications and select the one with the largest second-stage t-ratio. This can be expressed as the score $s_k = \big(\mathbb{1}\{F_k \geqslant c\}, |t_k|\big)$, with lexicographic ordering, and where $c > 0$ denotes the first-stage threshold. The threshold c may take any value; for example, the commonly used benchmark of $c = 10$ in applied work (Stock and Watson, 2015), or, alternatively, more stringent thresholds proposed in recent work (Lee et al., 2022).[12]

The following proposition extends the homogeneous-instrument-strength results from the previous subsection to these forms of joint preferences.

Proposition 3 (Median Bias with Joint Preferences)**.** Let $\beta = 0$. Under Assumptions 1–5 and either smooth or lexicographic-threshold preferences, the median of the reported 2SLS estimator is biased toward the OLS estimand for any $K \geqslant 1$, and strictly increasing in absolute value with the number of available instruments K.

The proof follows the same general strategy as Proposition 1, and is therefore provided only in the Online Appendix OA.B.

For intuition behind Proposition 3, Figure 1 shows that first- and second-stage instrument-hacking censor different regions of the joint distribution, with both forms of selection leading to increased median bias. Combining these selection rules therefore naturally leads to similar outcomes. Given empirical evidence consistent with joint selection on both statistics, Proposition 3 suggests that instrument-hacking may, in practice, generate size distortions that systematically favor rejections in the direction of the OLS estimand, at least when instruments have similar strength. We now turn to the more general case in which instrument strengths may vary.

4. Theory: Median Bias with Heterogeneous Instruments

Thus far, the theoretical results have assumed all instrument are binary with the same strength and assignment probabilities (Assumption 5). This assumption provides an important benchmark case for isolating the central mechanism through which instrument-hacking generates bias. In general, however, we might expect successive instruments obtained by researchers to vary both in their first-stage coefficients (π_k) and assignment probabilities (q_k). In this subsection, we relax the assumption of homogeneous instruments and allow for heterogeneity across candidate instruments.

[12] An alternative rule in which researchers first target statistical significance (e.g. $|t_k| \geqslant 1.96$) and then maximize first-stage strength yields similar results, but seems less plausible as a selection rule.

4.1 Non-Monotonicity Example

Before introducing the more general results, we first illustrate that monotonicity of median bias in K, which holds in the homogeneous case, may not extend to settings with heterogeneous instrument strengths. This motivates the need for more general results that accommodate heterogeneity in instrument strength.

Consider a simple two-point distribution of instrument strengths. Each candidate instrument is either weak, with true first-stage coefficient $\pi_k = 0.05$ (probability 0.9), or strong, with $\pi_k = 0.15$ (probability 0.1). Instruments are independent, with equal assignment probabilities $q_k = 0.5$. We set $N = 5,000$, $u_i \sim N(0,1)$ and $\rho = 0.5$, and normalize $\mathrm{Var}(x_i) = 1$. Under this normalization, weak and strong instruments correspond to true first-stage F-statistics of approximately 3.13 and 28.28, respectively. We then simulate the impact of first-stage instrument hacking, where researchers select the instrument that maximizes the sample first-stage F-statistic. For further details on the setup and simulation, see Online Appendix OA.C.

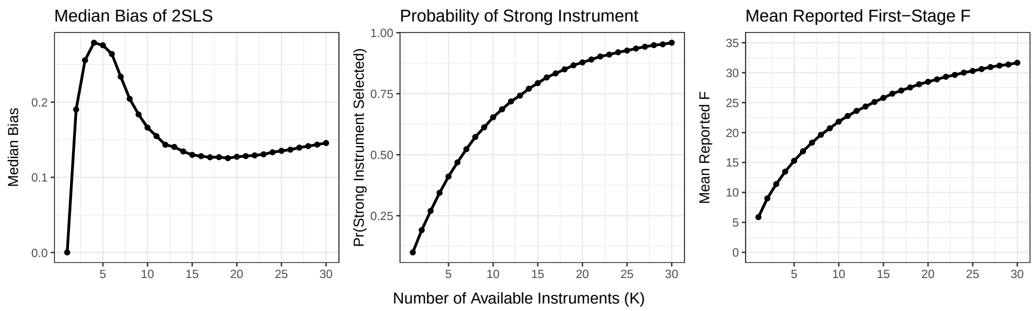
Figure 3 presents the main results under first-stage instrument-hacking. The left panel shows that median bias of the 2SLS estimator is non-monotonic in the number of available instruments K. In particular, median bias increases sharply for small values of K, reaching a peak at around $K = 4$, after which it declines before eventually flattening and slightly increasing again. Notably, the magnitude of bias is large: at its peak, median bias is close to the normalized OLS bias of 0.5, indicating that the selected IV estimator can be nearly as biased as OLS.

The intuition is as follows. When K is small, there is a high probability that all available instruments are weak. Selection among weak instruments generates large median bias in the direction of OLS. As K increases, the probability that at least one strong instrument is available rises rapidly, and the selection rule increasingly picks this strong instrument. This is reflected in the middle panel, where the probability of selecting a strong instrument increases with K, and in the right panel, where the mean reported first-stage F-statistic rises. Since stronger instruments attenuate bias, this leads to a decline in median bias.[13]

The stark two-point distribution used here—with many weak instruments and a

[13] As K becomes large, selection is driven not only by differences in signal but also by extreme realizations of noise. Even conditional on selecting strong instruments, those with favorable noise realizations are disproportionately chosen. This induces selection on noise that reintroduces bias in the direction of OLS, causing median bias to increase despite continued 'improvements' in the estimated first-stage strength.

FIGURE 3. Non-Monotonicity in First-Stage Hacking Example



small fraction of strong ones—is chosen for transparency. However, in settings where one instrument is clearly much stronger than the others, researchers may be able to identify and prioritize it ex ante, making a model of random selection from such a pool less plausible. It remains an open question whether a more empirically plausible distribution of instrument strengths can generate non-monotonicity, an issue we turn to in Section 5. We next consider results under very weak distributional assumptions.

4.2 General Results: Median Bias with Heterogeneous Instruments

The example illustrates that monotonicity of median bias in K need not hold once the restriction of homogeneous instruments is relaxed. We therefore replace Assumption 5 with a weaker condition that allows for heterogeneous instrument strength:

Assumption 5′ (Heterogeneous Instruments)**.** For each candidate instrument, define $\sigma^2_{z_k} \equiv \frac{1}{N}\sum_{i=1}^{N} \tilde{z}^2_{ki}$ and $\lambda_k \equiv \pi_k^2 \sigma^2_{z_k}$. The sequence of instrument strengths $\{\lambda_k\}_{k=1}^{\bar{K}}$ is i.i.d. from a fixed distribution μ_λ. Moreover, there exists a finite constant $\bar{\lambda} > 0$ such that the support of μ_λ is contained in $(0, \bar{\lambda}]$. Finally, assume $z_{ki} \perp\!\!\!\perp u_i$ for all k.

For general instruments, strength depends jointly on the first-stage coefficient π_k and the variability of the instrument, as summarized by $\lambda_k = \pi_k^2 \sigma^2_{z_k}$. The assumption of i.i.d. draws from the distribution of instrument strengths is substantive. It corresponds to a setting in which researchers lack knowledge of instrument strengths in advance of observing them, and instead treat instrument search as random draws from a fixed distribution. This assumption would be violated if, for example, researchers could systematically target instruments that are known in advance to be stronger.

Definition 2. Given the upper bound on instrument strength $\bar{\lambda}$, let the finite constants $t^-_{\max} > 0$ and $F^-_{\max} > 0$ denote the largest values that the t and F statistics can attain when the 2SLS estimate lies in the direction opposite to the OLS estimand.

These maximums or "caps" are illustrated in Figure A1 in the Appendix. By contrast, t and F statistics linked with estimates that are aligned with the OLS bias can exceed these caps. Intuitively, this asymmetry drives our results. Once the median of the selected t or F statistic exceeds the cap, the upper half of realizations must come from the direction aligned with the OLS estimand. Since the lower bounds on median bias are increasing functions of the selected statistic, and since the median selected statistic is weakly increasing in K, these lower bounds are also weakly increasing in K. Formally, given these definitions and the weaker Assumption $5'$, we can state the following more general results:

Proposition 4 (Second-Stage Instrument-Hacking with Heterogeneous Instruments)**.** Let $\beta = 0$. Suppose Assumptions 1–4 and $5'$ hold and that researchers engage in second-stage instrument-hacking. Let $\hat{\beta}^*_K$ denote the reported 2SLS estimator when K instruments are available, and $|t^*_K|$ be the corresponding absolute t-ratio. Then

$$\text{Median}\Big(\text{sign}(\rho)\hat{\beta}^*_K\Big) \geqslant 0$$

for all K. Moreover, there exists a finite constant $t^-_{\max} > 0$ and an increasing function $g : (t^-_{\max}, \infty) \to (0, 1/|\rho|)$ such that whenever $\text{Median}(|t^*_K|) > t^-_{\max}$, it follows that

$$\text{Median}\Big(\text{sign}(\rho)\hat{\beta}^*_K\Big) \geqslant g(\text{Median}(|t^*_K|)) > 0$$

where $\text{Median}(|t^*_K|)$ is weakly increasing in K.

Proposition 5 (First-Stage Instrument-Hacking with Heterogeneous Instruments)**.** Let $\beta \in \mathcal{R}$. Suppose Assumptions 1–4 and $5'$ hold and that researchers engage in first-stage instrument-hacking. Let $\hat{\beta}^*_K$ denote the reported 2SLS estimator when K instruments are available, and F^*_K be the corresponding first-stage F-statistic. Then

$$\text{Median}\Big(\text{sign}(\rho)(\hat{\beta}^*_K - \beta)\Big) \geqslant 0$$

for all K. Moreover, there exists a finite constant $F^-_{\max} > 0$ and an increasing function $h : (F^-_{\max}, \infty) \to (0, \infty)$ such that whenever $\text{Median}(F^*_K) > F^-_{\max}$, it follows that

$$\text{Median}\Big(\text{sign}(\rho)(\hat{\beta}^*_K - \beta)\Big) \geqslant h(\text{Median}(F^*_K)) > 0$$

where $\text{Median}(F^*_K)$ is weakly increasing in K.

Together, these results show that, for both first- and second-stage instrument-hacking, the median reported 2SLS estimator is weakly biased in the direction of the OLS estimand for any K. They also provide lower bounds on the magnitude of this bias once the relevant t or F statistic exceeds its cap.

The informativeness of the lower bounds depends on the maximum instrument strength, $\bar{\lambda}$. When $\bar{\lambda}$ is small, the lower bounds are more likely to be strictly positive even for relatively small values of K. In contrast, when $\bar{\lambda}$ is large, the lower bounds may remain close to zero for empirically relevant values of K, and therefore provide little additional information beyond the direction of the bias.[14]

5. Empirical Results

As we saw in Section 4, in heterogeneous-strength environments, median bias need not increase monotonically with K; it depends on the distribution of instrument strengths. This section assesses how median bias and other properties of exactly-identified 2SLS vary with K, given a realistic distribution of instrument strengths derived from the applied IV literature. To begin, we collected information on first-stage F statistics from all IV regressions published in the *American Economic Review*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and *Journal of Political Economy* from 2011 to 2023.[15] We found 800 papers (22.5% of the total) that used IV in some way, of which 496 contained usable IV results in the main body of the article. 80% of these articles reported at least one first-stage F statistic. This search culminated in 5,881 F-statistics and associated second-stage t-statistics.

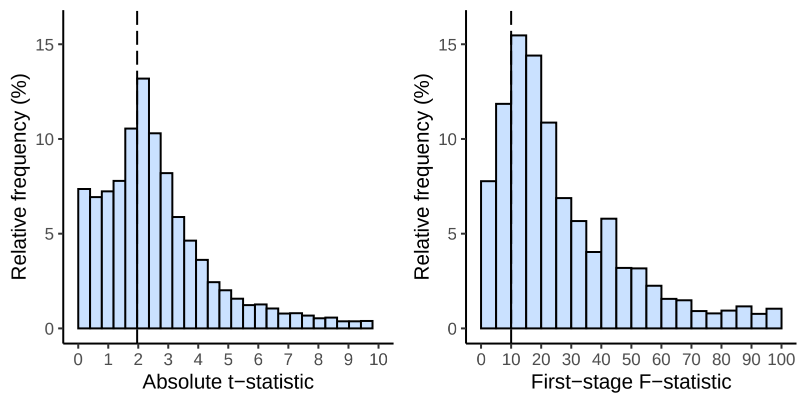
Figure 4 plots the empirical distributions of the absolute second-stage t-statistic and the first-stage F-statistics. Both exhibit visible bunching around conventional reporting thresholds ($t = 1.96$ and $F = 10$), consistent with selective reporting based on both second-stage statistical significance and first-stage strength.

To assess this formally, we consider a symmetric 15% bandwidth around each threshold: $F \in (8.5, 11.5)$ and $|t| \in (1.666, 2.254)$. Classifying observations by whether they fall above or below each cutoff, local smoothness implies the four resulting cells should contain approximately equal numbers of observations. However, the jointly

[14] Formally, the thresholds $t^{-}_{\max}$ and $F^{-}_{\max}$ are increasing in $\bar{\lambda}$.

[15] To find IV regressions, we searched for the word ‘instrument’ in all articles over the sample period. We then manually searched within these articles for any IV regression in the main body of the paper, and recorded the first-stage F statistic, the second-stage t statistic, and other information. Occasionally, the first-stage F was not reported but the first-stage t statistic was. In the case of IV regressions with single instruments, we squared this t-statistic to derive the F.

Figure 4. Descriptive Statistics



favorable cell ($F \geqslant 10, |t| \geqslant 1.96$) contains 44.3% of observations, significantly above the 25% benchmark ($p = 0.00015$). The excess mass is also visible in the marginals: 64.6% of observations lie above the $F = 10$ threshold $F = 10$ ($p = 0.0064$) while 60.8% lie above the $|t| = 1.96$ threshold ($p = 0.0356$). It is interesting that the evidence for first-stage hacking is stronger than for second-stage hacking.

5.1 Method

Given selective reporting of results, the observed empirical distribution of instrument strengths does not reflect the true underlying distribution of instrument strengths available to researchers. Hence, we use the Andrews and Kasy (2019) selection model to infer the underlying true distribution of instrument strengths. Their model assumes that observed instrument strengths in published papers are generated by i.i.d. draws from the following DGP:

1. Draw a true first-stage F-statistic from a distribution μ_F, i.e., $F_i \sim \mu_F$.
2. Given F_i, assume the sample F statistic obeys $\sqrt{\hat{F}_i} \mid F_i \sim |N(\sqrt{F_i}, 1)|$.[16]
3. Let D_i be a Bernoulli random variable equal to one if the estimate is reported and published, and zero otherwise. Assume that $\mathbb{P}(D_i = 1 \mid \hat{F}_i) = p(\hat{F}_i)$ where $p(\hat{F}_i)$ is a weakly increasing function of $\hat{F}_i$.

[16] Letting t_i denote the true first-stage t-ratio, Andrews and Kasy (2019) adopt the Gaussian approximation $\hat{t}_i \mid t_i \sim N(t_i, 1)$. Given that $|\hat{t}_i| = (\hat{F}_i)^{1/2}$ and $|t_i| = \sqrt{F_i}$, it follows that $|\hat{t}_i| \mid t_i \sim |N(|t_i|, 1)|$, which is equivalent to Step 2 above. This approach is consistent with the first-stage sign screening used in the theory: since the likelihood is written in terms of $|\hat{t}_i| = (\hat{F}_i)^{1/2}$, it treats the reported first-stage coefficient as oriented to have the same sign as the true first-stage coefficient.

Following the empirical application in Andrews and Kasy (2019), we assume that μ_F follows a gamma distribution with location and scale parameters (λ, κ). Consistent with the descriptive evidence, we model selective reporting using a step function at the conventional first-stage threshold of 10: $p(\hat{F}_i) = 1 - (1-\gamma) \cdot \mathbb{1}\{\hat{F}_i < 10\}$, where $\gamma < 1$. Thus, when the first-stage F statistic falls below the $F = 10$ threshold the researcher reports the result with probability $\gamma < 1$, while reporting with probability 1 otherwise. Of course, other distributions and selection rules are possible.

It is worth noting that we cannot estimate the models of first-stage, second-stage or joint instrument hacking that we defined in Section 1.2 because we do not observe the set of instruments that a researcher had to choose from in each (i.e., we only see reported results). Instead, we interpret the Andrews and Kasy (2019) model as a flexible reduced-form approximation to a general model of instrument selection, and we seek to be agnostic about the true structural selection model.

5.2 Estimation and Model Fit

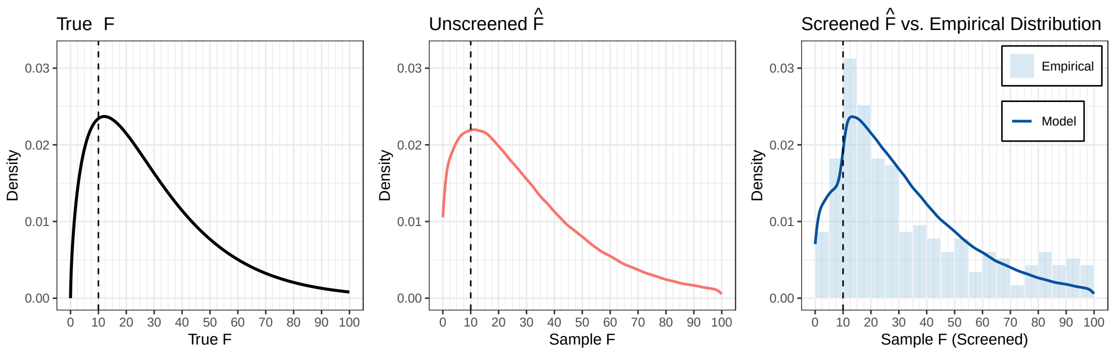
For estimation, we restrict attention to the first just-identified IV result reported in each paper. We also restrict the sample to published findings with reported F-statistics below 100 to reduce the influence of outliers.[17] Estimating the model by maximum likelihood yields $(\hat{\lambda}, \hat{\kappa}, \hat{\gamma}) = (1.66, 18.42, 0.615)$, where $\hat{\lambda}$ and $\hat{\kappa}$ are the estimated shape and scale parameters of the true F distribution, and $\hat{\gamma}$ is the estimated selection parameter.

The true F distribution implied by the estimates is plotted in the first panel of Figure 5. It is right-skewed, with a median of 24.7, and 17.5% of its mass is below the conventional threshold of $F = 10$. The sample F distribution is obtained by adding sampling noise, as in Step 2 of the DGP. This distribution, shown in the middle panel, has 20.1% of the mass below 10. The third panel then applies the estimated selection rule (Step 3) and compares the resulting screened distribution to the empirical distribution of reported first-stage F-statistics. The model generates a reasonably close correspondence between the simulated and the empirical distributions.[18]

[17] The estimated distribution itself still permits realizations above 100 since the support of the Gamma distribution is the positive real line.

[18] Note that in the model $P(\hat{F} < 10|selection) = (.201)(.615)/[(.201)(.615)+(1-0.201)(1)]=0.134$, compared to 13.5% in the data.

Figure 5. First-Stage F-Statistic Distributions: True, Unscreened, and Screened



5.3 Simulating Instrument-Hacking

Given the estimated distribution of true instrument strengths, we can now simulate the properties of the exactly identified 2SLS under any form of instrument hacking. To simulate instrument-hacking, we draw a true first-stage F-statistic from the estimated distribution for each candidate instrument, convert it into the corresponding first-stage coefficient, simulate the model, and then estimate the resulting just-identified IV specifications. We then apply the relevant instrument-hacking rule to the set of estimated specifications. We consider first-stage and second-stage hacking as well as a lexicographic rule. It is important to note that we do not assert that researchers engage in any particular form of hacking. The goal here is simply to see how specific forms of instrument hacking affect properties of 2SLS estimates, given a realistic distribution of instrument strengths. For further details on the simulation procedure, see Online Appendix OA.D.

Aside from the distribution of instrument strength, we set other aspects of the DGP to the same values used in Section 4.1. First we set $\rho = 0.5$ to capture a moderate degree of endogeneity, although below we also consider $\rho = 0.2$ and 0.8. We set $\beta = 0$ to focus on size distortion, set N=5000, and normalize $\text{Var}(x_i) = 1$ and $\sigma_u^2 = 1$ wlog so β has a natural interpretation as an effect size – i.e., the effect of a one standard deviation ('SD') increase in x_i on y_i in standard deviations.

5.4 Results for $\rho = 0.50$

Figure 6 and Table 1 present the results. When there is no instrument-hacking ($K = 1$), the median bias of the 2SLS estimator is zero, but the size of the 5% t-test is 4.3%. The t-test is underpowered in some runs, as the unscreened distribution of instrument strength contains some weak instruments. Furthermore, 93% of rejections of H_0:$\beta = 0$ occur when $\hat{\beta} > 0$, illustrating the power asymmetry problem.

The first row of Figure 6 and left-side columns of Table 1 show the impact of second-stage instrument-hacking. As we see in Table 1, median bias is monotonically increasing with the number of available instruments. If researchers choose from only K=2 available instruments, median bias is 0.062, and if they choose from 3 instruments the median bias increases to 0.116.

Given that β is an effect size, these biases are substantial. For example, Bartoš et al. (2024) find a median treatment effect size of 0.2 SD across a 327 meta-analyses in economics, and Evans and Yuan (2022) find a median effect size of 0.1 SD in 234 articles on the effect of educational interventions on test scores.[19] Thus, if the true effect were in fact zero, selectively reporting the just-identified specification that maximizes the t-ratio among three available instruments would generate a median bias about as large as numerous estimated effect sizes in economics.

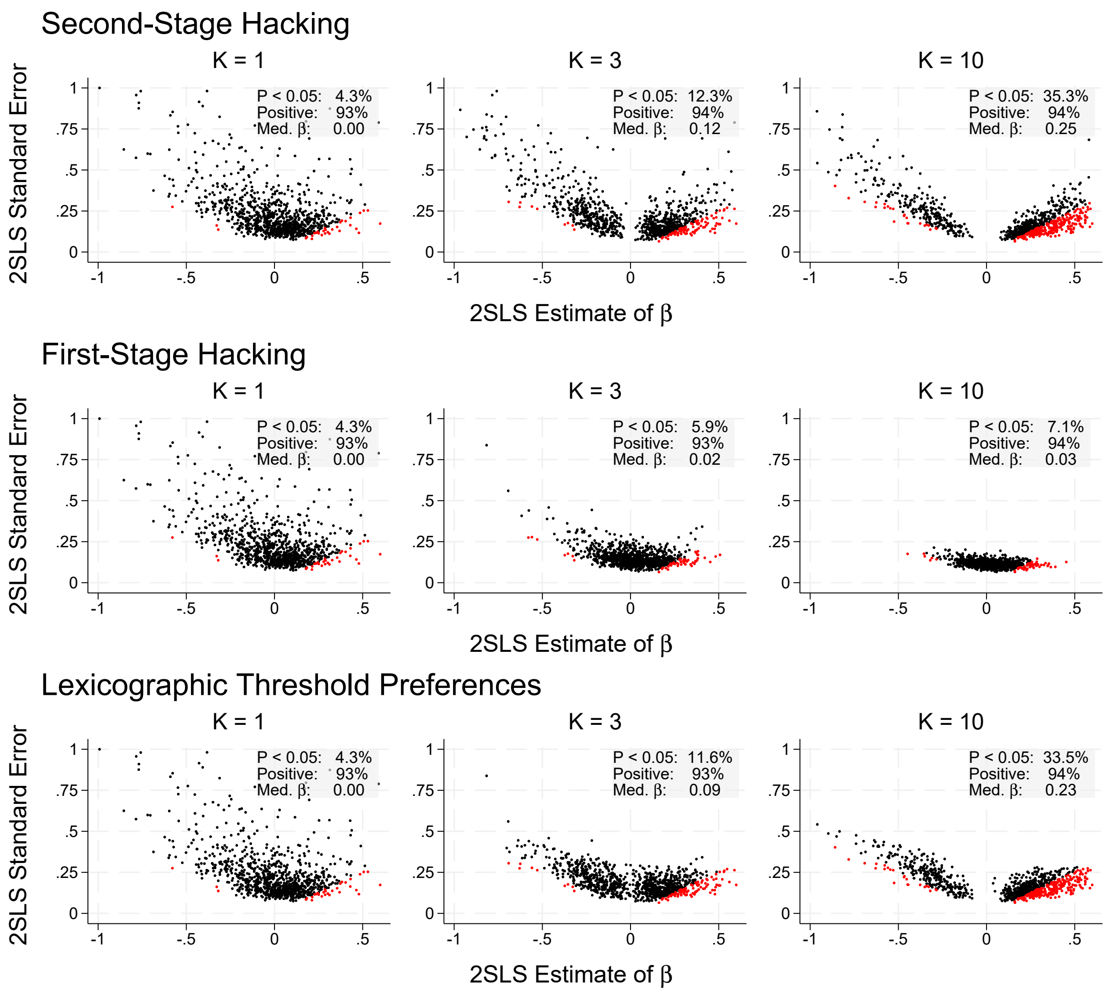
Size inflation also appears when there are 2 or more instruments to choose from, and it increases monotonically with K. For instance, when $K = 3$ the 5% t-test has size of 12.3%. Furthermore, the overwhelming majority of false rejections occur when $\hat{\beta} > 0$, due to the power asymmetry. As a result, conventional meta-analytic methods may incorrectly conclude that the literature provides consistent evidence for a positive effect even when the true effect is zero or even somewhat negative.

The middle row of Figure 6 and middle columns of Table 1 show the impact of first-stage instrument-hacking. If researchers choose from K=2 available instruments median bias is 0.020 standard deviations, and when K=3 it is .025. These are still fairly large biases relative to typical effect size estimates in the literature, but they are much more modest than we found for second stage hacking. However, even though median bias is small under first-stage selection, the resulting false rejections continue to be highly asymmetric, with about 93% occurring on the side of the OLS bias

Under first-stage hacking, increasing the number of available instruments generates two opposing forces. On the one hand, selecting the instrument with the highest first-stage F statistic tends to favor specifications with positive sampling error in the first stage, which pushes the 2SLS estimator toward the OLS estimand. On the other hand, when instrument strengths vary across specifications, increasing K also

[19] Chetty et al. (2014) report that "a one standard deviation increase in teacher quality raises end-of-year scores by 0.13 standard deviations of the student test score distribution on average across grades and subjects."

Figure 6. Instrument-Hacking with Instrument Strengths Drawn from Data: Joint Distribution of $\hat{\beta}$ and Standard Errors



Notes: 2SLS estimates and standard errors when researchers p-hack among just-identified specifications from 100,000 replications. In all cases, $\beta = 0$, $\rho = 0.5$, and true first-stage $F \sim \text{Gamma}(1.66, 18.42)$. Red dots indicate significant estimates at 5% level. "Lexicographic-threshold preferences" refers to first targeting the first-stage F-statistic when no specification exceeds $F = 10$, and otherwise minimizing the second-stage p-value among specifications with first-stage F-statistics greater than 10. 'P $\leqslant$ 0.05' refers to the size (rejection rate) of the estimator, "Positive" refers to the proportion of rejections on the positive side of the true null (i.e. in the direction of the biased OLS estimate), and 'Med β' refers to the median estimate of β which is also the bias in this case.

raises the probability that at least one genuinely strong instrument is available, which mitigates the increase in median bias. The relatively modest increase in bias in the empirical simulations suggests that these two forces partially offset one another.

Earlier we presented evidence of selection on *both* first-stage strength and second-

Table 1 – Properties of the 2SLS t-Test Under Instrument Selection

Type of Hacking:	Second-Stage			First-Stage			Lexicographic		
Instruments Available	Size	Pos. Rej.	Median Bias	Size	Pos. Rej.	Median Bias	Size	Pos. Rej.	Median Bias
1	0.043	93%	0.002	0.043	93%	0.002	0.043	93%	0.002
2	0.084	93%	0.062	0.054	93%	0.020	0.080	93%	0.046
3	0.123	94%	0.116	0.059	93%	0.025	0.116	93%	0.093
5	0.196	94%	0.177	0.064	93%	0.030	0.185	93%	0.155
8	0.293	94%	0.225	0.069	94%	0.033	0.277	94%	0.204
10	0.353	94%	0.246	0.071	94%	0.035	0.335	94%	0.226

Notes: This table presents rejection rates and median bias from Monte Carlo simulations of the DGP outlined in Section 5. Pos Rej. refers to the proportion of rejections that feature a $\hat{\beta} > 0$. "Lexicographic-threshold preferences" refers to first targeting the first-stage F-statistic when no specification exceeds $F = 10$, and otherwise minimizing the second-stage p-value among specifications with first-stage F-statistics greater than 10.

stage significance, and indeed it seems implausible that applied researchers would focus exclusively on instrument strength and not also consider statistical significance in a context where two or more instruments pass the F=10 threshold. We therefore turn next to the lexicographic-threshold preferences discussed in Section 3.2. Under this selection rule, researchers minimize the second-stage p-value among all specifications with a first-stage F-statistic greater than 10; if no such specifications exist, they instead select the specification with the largest first-stage F-statistic. These results are reported in the right columns of Table 1 and the bottom row of Figure 6.

Under this joint selection rule, the results are very similar to those for pure second-stage instrument-hacking. The median bias and size distortions are only slightly smaller. The reason is that under our estimated distribution of true instrument strengths, most first-stage sample F-statistics exceed 10, so the screen on the second-stage t-ratio drives the results. Importantly, this shows that the problems created by instrument hacking are not just a weak-instrument phenomenon.[20]

In summary, it is interesting that the median bias and size distortions created by second-stage instrument-hacking are considerably more severe than those created by first-stage hacking. This is surprising given that the consequences of first-stage screening in IV have been recognized for decades (Hall et al., 1996). By contrast,

[20] Keane and Neal (2024) show that the mechanical relationship between estimated effect sizes and standard errors remains quantitatively important at substantially higher levels of instrument strength than the conventional $F > 10$ threshold.

although selection on statistical significance in IV studies has received considerable attention, to our knowledge no prior work has examined or quantified how such selection (i.e., second stage hacking) interacts with the power asymmetry of IV estimators to generate median bias toward the OLS estimand and size distortions.

5.5 Robustness: Weaker Endogeneity, Stronger Instruments

In Online Appendix OA.E we consider robustness to the key DGP assumptions. First, we consider a case where $\rho = 0.20$, so endogeneity is much weaker.[21] Remarkably, we find the problems induced by second-stage hacking are still severe. For instance, median bias is 0.082 at $K = 3$ compared to .114 when ρ=0.5. The reason is that, when K=1, 72% of rejections of H_0: $\beta = 0$ occur when $\hat{\beta} > 0$. Thus, the 2SLS power asymmetry is severe even at this modest level of ρ, and as we have explained, the power asymmetry is what drives the bias from instrument hacking. In contrast, median bias from first-stage hacking is almost eliminated when $\rho = 0.20$. As we explained earlier, bias in this case is the outcome of two competing factors, and at this low level of endogeneity the gain from finding stronger instruments almost cancels out the contamination of instruments with endogeneity induced by instrument search. The appendix also reports with $\rho = 0.8$, which are similar to $\rho = 0.5$.

In Online Appendix OA.F, we also consider a calibration where instruments are much stronger, by drawing first-stage F-statistics from a Gamma$(10, 14)$ distribution, which implies a median first-stage F of 135 and a first percentile of 48. Even in this environment, second-stage instrument-hacking generates meaningful median bias. For instance, when $K = 3$ it generates median bias of 0.04 and rises to 0.11 when $K = 10$. In contrast, first-stage instrument-hacking causes bias up to 0.03 when $K = 10$.

Rejections are highly asymmetric even with very strong instruments. At $K = 1$ the percentage of rejections towards the OLS bias is 70%. This remains roughly constant as K increases under second-stage instrument-hacking, but rises sharply with K under first-stage hacking up to 93% rejections towards the OLS bias when $K = 10$. The reason for this is, again, first-stage hacking trims estimates with large standard errors which tend to be on the opposite side of the OLS bias.

Finally, in Online Appendix OA.G we repeated the empirical analysis using distributions of instrument strengths estimated from articles in different sub-fields of

[21] Angrist and Kolesár (2024) argue using three examples that $|\rho|$ in microeconometric applications seldom exceeds 0.4. In contrast, Keane and Neal (2025) estimate a value of ρ in their microeconometric application of -0.7.

economics.[22] The results are very similar across subfields as the distributions of instrument strengths are very similar.

6. Alternative Estimation and Inference Strategies

Finally we consider alternative approaches to estimation and inference in contexts with multiple candidate instruments. One obvious approach is to simply estimate over-identified models that utilize all available instruments. We consider doing this with 2SLS estimation and the t-test, as well as using LIML estimation with the Conditional Likelihood Ratio (CLR) test of Moreira (2003).[23] We continue to use the DGP from Section 5, that is based on the empirical distribution of instruments strengths, and we focus on the ρ=0.5 case. Results are reported in Table 2.

Table 2 – 2SLS and LIML Using All Available Instruments: No Instrument Hacking

Estimation Strategy:	2SLS with t-test			LIML with CLR test		
Instruments Available	Size	Pos. Rej.	Median Bias	Size	Pos. Rej.	Median Bias
1	0.043	93%	0.002	0.051	52%	0.002
2	0.051	88%	0.010	0.051	50%	0.000
3	0.055	85%	0.012	0.051	50%	0.000
5	0.057	84%	0.014	0.051	50%	0.000
10	0.064	85%	0.015	0.050	51%	0.000

Notes: This table presents rejection rates and median bias from Monte Carlo simulations of the DGP outlined in Section 5 when using all available instruments. There is no instrument hacking. 'Pos. Rej.' refers to the proportion of rejections that feature a $\hat{\beta} > 0$.

Using 2SLS with all available instruments, both median bias and size inflation are very small compared to just-identified IV with instrument hacking in Table 1. For example, when K=5, the median bias of 2SLS using all instruments is .015 SD, compared to .177 SD when one chooses the 'best' instrument based on the t-test result, and .030 SD when one chooses based on first-stage F. Thus, using just-identified IV with the 'best' instrument in an attempt to avoid the problems created by using many instruments actually makes median bias and size inflation worse.

[22] We classified articles according to the most common JEL topic categories: Labor and Demographic Economics (J), Microeconomics (D), Economic Development, Innovation, Technological Change, and Growth (O), Industrial Organization (L), and International Economics (F).

[23] When $K = 1$, the regression is just identified, so LIML coincides with 2SLS and the CLR test is equivalent to the Anderson and Rubin (1949) test (AR test). Under heteroskedasticity or clustered dependence, more efficient estimation would be obtained using continuously updated GMM with robust inference, such as the GMM extension to CLR introduced in Kleibergen (2005). See Hausman et al. (2012) and Mikusheva and Sun (2022) for related many-instrument settings.

The greatest problem with 2SLS using all instruments is that the power asymmetry problem remains severe, with 82% of rejections featuring $\hat{\beta} > 0$.

LIML performs much better. As we see in Table 2, it exhibits no median bias regardless of K. Furthermore, the CLR test has the correct size, and rejections are always symmetric around the true null. Based on our results, we regard LIML using all instruments, combined with CLR for inference, as the best approach.

Another interesting question is whether the use of the Anderson and Rubin (1949) test for inference instead of the t-test can reduce the problems created by instrument hacking. Keane and Neal (2023, 2024) show the AR test mitigates the power asymmetry problem that plagues the t-test.[24] And as we have shown both in the theory and simulations, the power asymmetry drives the median bias created by instrument hacking. Thus, it is natural to expect that use of AR will mitigate these problems. Accordingly, in Table 3 and Figure 7 we present results where researchers instrument-hack while relying on the AR-test p-value for inference, again using the same DGP.

Table 3 – 2SLS with AR-test Under Instrument Selection

Type of Hacking:	Second-Stage			First-Stage			Lexicographic		
Instruments Available	Size	Pos. Rej.	Median Bias	Size	Pos. Rej.	Median Bias	Size	Pos. Rej.	Median Bias
1	0.051	52%	0.002	0.051	52%	0.002	0.051	52%	0.002
2	0.099	52%	0.021	0.051	65%	0.020	0.079	60%	0.035
3	0.144	51%	0.047	0.052	70%	0.025	0.112	61%	0.070
5	0.226	51%	0.091	0.054	75%	0.030	0.179	60%	0.120
8	0.337	52%	0.123	0.058	79%	0.033	0.270	61%	0.160
10	0.401	52%	0.137	0.059	81%	0.035	0.326	61%	0.178

Notes: This table presents rejection rates and median bias from Monte Carlo simulations of the DGP outlined in Section 5 using the AR test for inference. Pos Rej. refers to the proportion of rejections that feature a $\hat{\beta} > 0$. "Lexicographic-threshold preferences" refers to first targeting the first-stage F-statistic when no specification exceeds $F = 10$, and otherwise minimizing the second-stage p-value among specifications with first-stage F-statistics greater than 10.

Comparing Tables 1 and 3, it is useful to first consider the K=1 case. We see that the AR test generates almost balanced rejections of H_0:β=0; i.e., 52% occur when $\hat{\beta} > 0$ compared to 93% for the t-test. This illustrates how the AR test mitigates the

[24] As Keane and Neal (2024) Section 2.1.4 explains, there are two sources of the power asymmetry, due to the fact that the 2SLS standard error depends on $\hat{cov}(z, x)$ and $\hat{\sigma}_{2sls}$. First, $\hat{\sigma}_{2sls}$ is minimized in samples where the 2sls estimate coincides with OLS. Second, $\hat{cov}(z, x)$ is larger in samples where $\hat{cov}(z, u)$ is positive. Hence positive realizations of $\hat{cov}(z, u)$ both drive the estimate towards OLS and drive down the standard error. The AR test solves the first problem but not the second.

power asymmetry problem.[25] Now consider what happens as we increase K, and use either the second-stage t or AR test to choose the 'best' just-identified specification. AR test rejections of β=0 remain roughly balanced, in contrast to t-test rejections where 94% occur when $\hat{\beta}$>0.[26]

Comparing the top panels of Figures 6 and 7 makes this advantage of the AR test visually apparent: Given second-stage hacking, the t test generates a preponderance of significant estimates on the positive side. In contrast, while the AR test generates severe size inflation, the excess rejections are not systematically concentrated on the positive side of the true null. Hence, a literature that relies on the AR test would be much less likely to come to a (false) consensus that the effect is positive.

The middle columns of Table 3 and middle row of Figure 7 show the case where the researcher engages in first-stage instrument hacking and then reports the AR test result for the 'best' just-identified specification in the second stage. The striking result here is that first-stage instrument hacking breaks the balanced rejections property of the AR test.[27] The share of positive rejections is increasing with K.

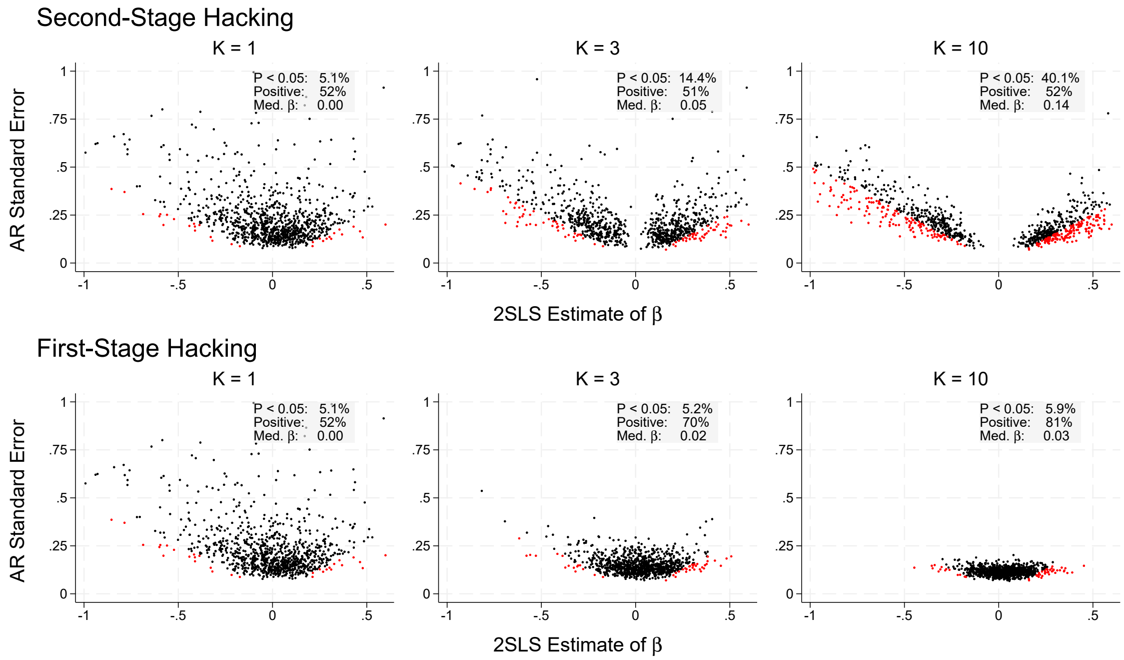
In summary, these results show that use of the AR test for inference does mitigate the problems created by second-stage instrument hacking, but it is far less useful if researchers also engage in first-stage hacking. The first-best solution is simply to use LIML with all available instruments for estimation and then use CLR for inference.

[25]The Spearman correlation between the 2SLS estimates and conventional 2sls standard errors in these runs is -.345, while that between the 2SLS estimates and the AR standard errors is only -.185. [We explain why the correlation drops in footnote 24]. Why does the AR test generate almost balanced rejections, despite the fact that the AR standard error tends to be smaller on positive estimates? The reason is that the distribution of 2SLS estimates is skewed left, meaning that positive estimates tend to be smaller than negative estimates. The combination of these two forces means that positive and negative estimates have similar AR test values, leading to nearly balanced rejections of β=0. [Note: The reason for skewness is that positive realizations of $\hat{cov}(z,u)$ both drive the estimate towards OLS and shrink the estimate, as positive realizations of $\hat{cov}(z,u)$ drive up $\hat{cov}(z,x)$ which appears in the denominator of the estimator.]

[26]The median bias induced by instrument hacking remains substantial when the researcher uses the AR test, but it is only about half as great as when using the t-test. For example, when K=5, median bias is .091 using AR vs .177 using t. A proof that second-stage hacking using the AR test generates median bias towards OLS is available on request, as is a proof for first-stage hacking.

[27]There is a negative correlation between 2SLS estimates and their AR standard errors because positive realizations of $\hat{cov}(z,u)$ both drive the estimate towards OLS and drive down the AR standard error. Selection for large $\hat{cov}(z,x)$ screens out cases with $\hat{cov}(z,u) < 0$ ($\hat{\beta} < 0$) as these tend to have relatively small $\hat{cov}(z,x)$ – i.e., weaker first stage. As AR would have found some of these negative estimates to be significant, the AR test results become unbalanced.

Figure 7. Instrument-Hacking with the AR Test and Instrument Strengths Drawn from Data: Joint Distribution of $\hat{\beta}$ and AR Standard Errors



Notes: 2SLS estimates and AR standard errors when researchers p-hack just-identified specifications (the AR test p-value in the case of second-stage hacking) from 100,000 replications. In all cases, $\beta = 0$, $\rho = 0.5$, and π_k is drawn from a first-stage F distribution of $Gamma(1.66, 18.42)$ (for more details see text). Red dots indicate statistically significant estimates from the AR test at 5% level.

7. Conclusion

This paper examines the statistical consequences of first and second-stage instrument-hacking in exactly-identified 2SLS regression. Theoretically, we show that instrument-hacking generates median bias in the direction of the OLS estimand. When candidate instruments are of equal strength, median bias increases monotonically with the number of available instruments. However, when instrument strengths are heterogeneous, monotonicity need not hold in general as increasing the number of candidate instruments can have offsetting effects: amplifying selection on favorable noise realizations, but also increasing the probability of finding a genuinely strong instrument. If we calibrate to the distribution of instrument strengths in IV studies published in leading economics journals, we find that median bias does increase monotonically with the number of available instruments. Furthermore, the magnitude of bias is economically significant even when researchers search over only a few instruments. The size distortion is also severe, and rejections of the true null $\beta = 0$ are heavily concentrated in the direction of the OLS bias.

The desire to pick a 'best' instrument and report only just-identified results is not necessarily motivated by a desire to 'hack' results. Rather, it can be motivated by the fact that 2SLS using multiple instruments suffers from bias towards OLS, while the 2SLS t-test suffers from size inflation and a severe power asymmetry problem – i.e., conventional Wald-based inference generates rejections of the true null $\beta = 0$ heavily concentrated in the direction of the OLS bias. Unfortunately, using instrument selection (either first or second-stage hacking) to arrive at a just-identified model exacerbates median bias, size distortion and power asymmetry problems compared to simply using all instruments and running 2SLS.

Thus, rather than responding to these problems by selecting down to a single instrument, we argue the preferred approach should be to use all instruments, in conjunction with robust estimation and inference. In the homoskedastic case this means LIML for estimation and the CLR test for inference.[28] We show this approach results in median unbiased estimates, no size distortion and no power asymmetry, regardless of the number of available instruments. As an alternative, we find the use of the AR test for inference mitigates the problems created by second-stage instrument hacking, but unfortunately the AR test is compromised by first-stage hacking.

[28] For heteroskedastic datasets, the equivalent approach would be the continuously updated GMM estimator of Hansen et al. (1996) coupled with a robust test such as the GMM extension of the CLR test in Kleibergen (2005).

References

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

Andrews, I. and T. B. Armstrong (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics 8*(2), 479–503.

Andrews, I. and M. Kasy (2019). Identification of and Correction for Publication Bias. *American Economic Review 109*(8), 2766–2794.

Andrews, I., J. H. Stock, and L. Sun (2019). Weak instruments in iv regression: Theory and practice. *Annual Review of Economics 11*, 727–753.

Angrist, J. D. and M. Kolesár (2024, March). One instrument to rule them all: The bias and coverage of just-id iv. *Journal of Econometrics 240*(2).

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bartoš, F., M. Maier, E.-J. Wagenmakers, F. Nippold, H. Doucouliagos, J. P. Ioannidis, W. M. Otte, M. Sladekova, T. K. Deresssa, S. B. Bruns, et al. (2024). Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics. *Research Synthesis Methods 15*(3), 500–511.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association 90*(430), 443–450.

Brodeur, A., N. Cook, and A. Heyes (2020). Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review 110*(11), 3634–3660.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9), 2633–2679.

DellaVigna, S. and E. Linos (2022). RCTs to Scale: Comprehensive Evidence From Two Nudge Units. *Econometrica 90*(1), 81–116.

Donald, S. G. and W. K. Newey (2001). Choosing the Number of Instruments. *Econometrica 69*(5), 1161–1191.

Evans, D. K. and F. Yuan (2022). How big are effect sizes in international education studies? *Educational evaluation and policy analysis 44*(3), 532–540.

Franco, A., N. Malhotra, and G. Simonovits (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science 345*(6203), 1502–1505.

Hall, A. R., G. D. Rudebusch, and D. W. Wilcox (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review*, 283–298.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics 14*(3), 262–280.

Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics 3*(2), 211–255.

Ioannidis, J. P. (2005). Why Most Published Research Findings Are False. *PLoS Med 2*(8).

Keane, M. and T. Neal (2023). Instrument strength in IV estimation and inference: A guide to theory and practice. *Journal of Econometrics 235*, 1625–1653.

Keane, M. and T. Neal (2024). A Practical Guide to Weak Instruments. *Annual Review of Economics 16*, 185–212.

Keane, M. and T. Neal (2025). Robust inference for the frisch labor supply elasticity. *Journal of Labor Economics 43*(S1), S179–S219.

Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica 73*(4), 1103–1123.

Kranz, S. and P. Pütz (2022). Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment. *American Economic Review 112*(9), 3124–3136.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022). Valid t-Ratio Inference for IV. *American Economic Review 112*(10), 3260–3290.

Mikusheva, A. and L. Sun (2022). Inference with many weak instruments. *The Review of Economic Studies 89*(5), 2663–2686.

Moreira, M. J. (2003). A Conditional Likelihood Ratio Test for Structural Models. *Econometrica 71*(4), 1027–1048.

Stock, J. H. and M. W. Watson (2015). *Introduction to Econometrics* (3rd ed.). Harlow, England: Pearson.

Stock, J. H. and M. Yogo (2005). Testing for weak instruments in linear iv regression. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models*, pp. 80–108. Cambridge University Press.

Zivot, E., R. Startz, and C. R. Nelson (1998). Valid confidence intervals and inference in the presence of weak instruments. *International Economic Review*, 1119–1144.

A. Appendix: Proofs of Propositions

Throughout the proofs, we use the following notation. When K instruments are available, the reported result is denoted by $\hat{\beta}_K^*$. Formally, $\hat{\beta}_K^* \equiv \hat{\beta}_{k*}$ where $k^* = \arg\max_{k=1,\ldots,K} s_k$ and s_k is the score function; other random variables follow the same notational convention. We denote demeaned variables as $\tilde{z}_{ki} \equiv z_{ki} - \frac{1}{N}\sum_i z_{ki}$.

The following identities and definitions are used throughout for clarity:

$$\hat{\beta}_k = \beta + \frac{\frac{1}{N}\sum_i \tilde{z}_{ki}u_i}{\frac{1}{N}\sum_i \tilde{z}_{ki}\tilde{x}_i} \iff \hat{\beta}_k - \beta = \frac{M_k}{C_k + \rho M_k} \tag{2}$$

where

$$C_k \equiv \pi_k \sigma_{z_k}^2 + \sum_{l \neq k} \pi_l \sigma_{z_l,z_k} \text{ and } M_k \equiv \frac{1}{N}\sum_{i=1}^{N} \tilde{z}_{ki}u_i$$

and $\sigma_{z_k} \equiv \sqrt{\frac{1}{N}\sum_i \tilde{z}_{ki}^2} = \sqrt{\frac{1}{N}\sum_i (z_{ki} - \bar{z}_k)^2}$ is the standard deviation of instrument z_{ki}; and $\sigma_{z_l,z_k} \equiv \frac{1}{N}\sum_i (z_{li} - \bar{z}_l)(z_{ki} - \bar{z}_k) = \frac{1}{N}\sum_i \tilde{z}_{li}\tilde{z}_{ki}$ is the covariance between z_{li} and z_{ki}.

A.1 Proof of Proposition 1

First, normalize the sign of each instrument so that its first-stage coefficient is positive, implying $\pi > 0$. This is without loss of generality because replacing z_{ki} by $-z_{ki}$ changes the sign of π but leaves the corresponding IV estimand, first-stage F-statistic, and selection rule unchanged. Under sample orthogonality (Assumption 4) and homogeneous instruments (Assumption 5), $C = \pi\sigma_{z_k}^2 > 0$. Second, assume without loss of generality that $\rho > 0$. When $\rho < 0$, the same argument applies with all inequalities reversed, implying that the median bias becomes more negative as K increases, and therefore moves toward the OLS estimand in absolute value.

The proof relies on Lemma OA.B.1 in Online Appendix OA.B, which derives the functional form of the absolute t-ratio as a function of the model parameters and establishes the several properties of the function used throughout the proof.

Let $M_k \equiv \frac{1}{N}\sum_{i=1}^{N} \tilde{z}_{ki}u_i$ denote the raw draw for candidate instrument k. By Assumption 1, M_k has a continuous distribution symmetric about zero. Under first-stage screening, Lemma OA.B.2 implies that a draw is admissible if and only if $M_k > -C/\rho$. For inadmissible draws satisfying $M_k \leqslant -C/\rho$, we set the score equal to $-\infty$. We denote the absolute second-stage t-ratio by $t(M_k)$. On the event that at least one admissible specification exists, we denote the selected draw as $M_K^* \equiv M_{k*}$, where $k^* \in \arg\max_{1 \leqslant k \leqslant K} t(M_k)$. Thus, K counts raw candidate instruments, not admissible instruments. Given our results concern the distribution of reported estimates, probabilities involving M_K^* are understood conditional on the event that at least one admissible specification exists.[29]

Fix $\beta = 0$. First, consider the case $K = 1$. We show that the median of the reported 2SLS estimator is positive, and hence biased toward the OLS estimand. Since M_1 has a continuous distribution symmetric about zero, $\text{Median}(M_1) = 0$. However, the reported draw is observed only when it is admissible. Hence $M_1^* \sim M_1 \mid M_1 > -C/\rho$. Since $-C/\rho < 0$, screening truncates the symmetric distribution of M_1 only on the negative side, implying $\text{Median}(M_1^*) > 0$.

The corresponding reported estimator is

$$\hat{\beta}_1^* = \frac{M_1^*}{C + \rho M_1^*}$$

On the screened region we have $C + \rho M_1^* > 0$, so $\text{sign}(\hat{\beta}_1^*) = \text{sign}(M_1^*)$. Moreover, the 2SLS estimator is strictly increasing in M_1^* on $\big(-\frac{C}{\rho}, \infty\big)$, and strictly increasing transformations preserve medians. Hence, $\text{Median}(\hat{\beta}_1^*) > 0$.

Next, we show the median bias of the reported 2SLS estimator is strictly increasing in the number of candidate instruments, K. Since $\hat{\beta}(M) = M/(C + \rho M)$ is strictly increasing on the screened region, strict increases in the median of M_K^* imply strict increases in the median bias of the reported 2SLS estimator. Hence, our goal is to show that

$$\text{Median}(M_1^*) < \text{Median}(M_2^*) < \cdots < \text{Median}(M_K^*)$$

Let $G_K(x) \equiv \Pr(M_K^* \leqslant x)$. For any $x \geqslant 0$, decompose the cdf as $G_K(x) = (1 - p_K) + p_K G_K^+(x)$, where $p_K \equiv \Pr(M_K^* > 0)$; and $G_K^+(x) \equiv \Pr(M_K^* \leqslant x \mid M_K^* > 0)$. It follows that

$$G_{K+1}(x) - G_K(x) = (p_K - p_{K+1})(1 - G_K^+(x)) + p_{K+1}(G_{K+1}^+(x) - G_K^+(x))$$

We now show that the distribution of M_K^* shifts to the right with K: $G_{K+1}(x) - G_K(x) < 0$ for all $x \geqslant 0$. Since $1 - G_K^+(x) > 0$ and $p_{K+1} > 0$, it suffices to show that both difference terms are negative. The following lemmas establish these claims.

[29] If no admissible specification exists, then no specification is reported under the maintained first-stage sign-screening rule. Thus the unconditional reported estimator is not defined on this event without imposing an additional reporting convention.

Lemma A.1 (Increasing Probability of Positive Winners)**.** $p_K < p_{K+1}$ for all $K \geqslant 1$.

Lemma A.2 (Stochastic Dominance for Positive Winners)**.** $G^+_{K+1}(x) < G^+_K(x)$ for all $x > 0$.

Since $\mathrm{Median}(M_1^*) > 0$ and the distribution is continuous, we have $p_1 > \frac{1}{2}$. Lemma A.1 therefore implies $p_K > \frac{1}{2}$ for all $K \geqslant 2$, which in turn implies that $\mathrm{Median}(M_K^*) > 0$ for all $K \geqslant 2$.

Combining Lemmas A.1 and A.2 gives $G_{K+1}(x) < G_K(x)$ for all $x \geqslant 0$. Evaluating at $x = \mathrm{Median}(M_K^*) \geqslant 0$ yields $G_{K+1}(\mathrm{Median}(M_K^*)) < G_K(\mathrm{Median}(M_K^*)) = \frac{1}{2}$. It follows that $\mathrm{Median}(M_{K+1}^*) > \mathrm{Median}(M_K^*)$, which completes the proof. □

Proof of Lemma A.1: Recall that $M_k \equiv \frac{1}{N}\sum_{i=1}^N \tilde{z}_{ki} u_i$ denotes the raw, unscreened draw. Admissible draws satisfy $M_k > -C/\rho$ (Lemma OA.B.2). For admissible draws, the absolute second-stage score is $|t(M_k)|$, while inadmissible draws are assigned score $-\infty$. Define $t_K^* \equiv \max_{1 \leqslant k \leqslant K} |t(M_k)|$ as the winning score after K draws, and let M_K^* denote the associated winning draw. Since the distribution of M_k is continuous, ties occur with probability zero. Define $p_K = \Pr(M_K^* > 0)$. We will show that $p_{K+1} > p_K$ for every $K \geqslant 1$.

Next, define the replacement event $R_{K+1} \equiv \{t_{K+1} > t_K^*\}$, which occurs whenever the $(K+1)$-st draw replaces the current winner. Then, using the law of total probability, we can write

$$
\begin{aligned}
p_K &= \Pr(M_K^* > 0, R_{K+1} = 1) + \Pr(M_K^* > 0, R_{K+1} = 0),\\
p_{K+1} &= \Pr(M_{K+1}^* > 0, R_{K+1} = 1) + \Pr(M_{K+1}^* > 0, R_{K+1} = 0)\\
&= \Pr(M_{K+1} > 0, R_{K+1} = 1) + \Pr(M_K^* > 0, R_{K+1} = 0).
\end{aligned}
$$

The final equality follows because on the event $\{R_{K+1} = 1\}$, the $(K+1)$-st draw becomes the new winner, so $M_{K+1}^* = M_{K+1}$; while on $\{R_{K+1} = 0\}$, the incumbent winner is unchanged, so $M_{K+1}^* = M_K^*$. Subtracting the two expressions and writing joint probabilities as conditional probabilities gives:

$$
p_{K+1} - p_K = \mathbb{P}(R_{K+1} = 1) \cdot \Big[\mathbb{P}(M_{K+1} > 0 \mid R_{K+1} = 1) - \mathbb{P}(M_K^* > 0 \mid R_{K+1} = 1)\Big]
$$

$$
= \mathbb{E}\Big[\mathbb{P}\big(R_{K+1} = 1 \mid t_K^* = u\big) \cdot \Big(\mathbb{P}(M_{K+1} > 0 \mid R_{K+1} = 1, t_K^* = u) - \mathbb{P}(M_K^* > 0 \mid R_{K+1} = 1, t_K^* = u)\Big)\Big]
$$

where the second equality follows from the law of iterated expectations, and the expectation is taken over the distribution of the winning score u.

We prove $p_{K+1}-p_K > 0$ by showing the expression inside the expectation is weakly positive for every u, and strictly positive on a set of u's with positive probability. First, note that $\mathbb{P}(R_{K+1} = 1 | t^*_K = u) > 0$ for every finite u, because the score function $t(M)$ is unbounded above on the admissible region $\left(-\frac{C}{\rho}, \infty\right)$. Hence, a new draw always has strictly positive probability of exceeding the current winning score.

Thus, it suffices examine the sign of $\Pr(M_{K+1} > 0 \mid R_{K+1} = 1,\, t^*_K = u) - \Pr(M^*_K > 0 \mid R_{K+1} = 1,\, t^*_K = u)$. We will show that this difference is weakly positive for every u, and strictly positive for $u \in [0, t^-_{\max})$, where $t^-_{\max} \equiv \max_{M<0} t(M)$.

First, consider the case $u > t^-_{\max}$, where $t^-_{\max} \equiv \max_{M<0} t(M)$ is finite by Lemma OA.B.1. No negative draw can attain a score above $t^-_{\max}$. Hence, the incumbent winner and any replacement draw must be positive whenever $t^*_K = u > t^-_{\max}$ and $R_{K+1} = 1$. This means that $\Pr(M^*_K > 0 \mid R_{K+1} = 1, t^*_K = u) = \Pr(M_{K+1} > 0 \mid R_{K+1} = 1, t^*_K = u) = 1$, and thus the difference equals zero.

Now consider $u \in [0, t^-_{\max})$. For the first term in the difference, Bayes' rule gives

$$\mathbb{P}(M_{K+1} > 0 \mid R_{K+1} = 1,\, t^*_K = u)$$

$$\begin{aligned}
&= \frac{\mathbb{P}(R_{K+1} = 1 \mid M_{K+1} > 0,\, t^*_K = u)\, \mathbb{P}(M_{K+1} > 0 \mid t^*_K = u)}{\sum_{s\in\{-1,1\}} \mathbb{P}(R_{K+1} = 1 \mid \operatorname{sign}(M_{K+1}) = s,\, t^*_K = u)\, \mathbb{P}(\operatorname{sign}(M_{K+1}) = s \mid t^*_K = u)} \\
&= \frac{\frac{1}{2}\left[1 - F_+(u)\right]}{\frac{1}{2}\left[1 - F_+(u)\right] + \frac{1}{2}\left[1 - F_-(u)\right]} = \frac{1 - F_+(u)}{1 - F_+(u) + 1 - F_-(u)} \qquad (3)
\end{aligned}$$

where $F_+(u) \equiv \mathbb{P}(t(M) \leqslant u \mid M > 0)$ and $F_-(u) \equiv \mathbb{P}(t(M) \leqslant u \mid M < 0)$, with inadmissible draws assigned score $-\infty$. The second last equality uses independence of M_{K+1} and t^*_K, and symmetry of the raw distribution of M_{K+1}, which implies $\mathbb{P}(M_{K+1} > 0 \mid t^*_K = u) = \mathbb{P}(M_{K+1} > 0) = \frac{1}{2}$, and similarly for negative draws. Finally, writing the expression in terms of $F_+(u)$ and $F_-(u)$ follows because conditional on $t^*_K = u$, the replacement event $R_{K+1} = 1$ is equivalent to $t(M_{K+1}) > u$, where inadmissible draws have score $-\infty$.

For the second term in the difference, note that $\mathbb{P}(M^*_K > 0 \mid R_{K+1} = 1,\, t^*_K = u) = \mathbb{P}(M^*_K > 0 \mid t^*_K = u)$, because conditional on $t^*_K = u$, the replacement event $R_{K+1} = 1$ depends only on the independent draw M_{K+1}. Hence,

$$\mathbb{P}(M^*_K > 0 \mid t^*_K = u) = \frac{\mathbb{P}(M^*_K > 0,\, t^*_K = u)}{\mathbb{P}(t^*_K = u)}$$

Since t_K^* has a continuous distribution, interpret the conditional probability through shrinking intervals around u. For small $\varepsilon > 0$,

$$\mathbb{P}\big(M_K^* > 0,\, t_K^* \in [u, u+\varepsilon)\big) = K \cdot \tfrac{1}{2}\big[F_+(u+\varepsilon) - F_+(u)\big]F(u)^{K-1} + o(\varepsilon)$$
$$\mathbb{P}\big(t_K^* \in [u, u+\varepsilon)\big) = K \cdot \tfrac{1}{2}\big[F_+(u+\varepsilon) - F_+(u) + F_-(u+\varepsilon) - F_-(u)\big]F(u)^{K-1} + o(\varepsilon)$$

where $F(u) \equiv \mathbb{P}(t(M) \leqslant u)$. The first expression follows because exactly one of the K draws must fall in $[u, u+\varepsilon)$ on the positive branch, while the remaining $K-1$ draws must have scores below u; the factor $\frac{1}{2}$ follows from symmetry of the underlying distribution of M. Dividing the two expressions and taking $\varepsilon \downarrow 0$ gives

$$\mathbb{P}(M_K^* > 0 \mid t_K^* = u) = \frac{f_+(u)}{f_+(u) + f_-(u)} \tag{4}$$

Combining (3) and (4), and then rearranging, it suffices to prove

$$\frac{f_-(u)}{1 - F_-(u)} > \frac{f_+(u)}{1 - F_+(u)} \tag{5}$$

To show this inequality holds, we first rewrite both ratios in terms of the cdf and density of a single draw of M, which we denote by $H(\cdot)$ and $h(\cdot)$, respectively.

Consider first the right-hand side. Since $t(M)$ is strictly increasing on $(0, \infty)$ (Lemma OA.B.1), each $u \in [0, t_{\max}^-)$ has a unique positive inverse, which we denote by $m^+(u)$. Hence,

$$1 - F_+(u) = \mathbb{P}(t(M) > u \mid M > 0) = \mathbb{P}(M > m^+(u) \mid M > 0) = 2\big[1 - H(m^+(u))\big]$$

For the numerator, the change-of-variable formula gives $f_+(u) = \frac{2\,h(m^+(u))}{|t'(m^+(u))|}$. Therefore,

$$\frac{f_+(u)}{1 - F_+(u)} = \frac{h(m^+(u))}{1 - H(m^+(u))}\left|\frac{1}{t'(m^+(u))}\right| = \lambda(m^+(u))\left|\frac{1}{t'(m^+(u))}\right|$$

where $\lambda(M) \equiv \frac{h(M)}{1-H(M)}$ denotes the hazard function.

Next, consider the left-hand side of (5). Fix the score $u \in (0, t_{\max}^-)$, and let $m_-(u) < 0$ denote the negative solution closest to zero satisfying $t(m_-(u)) = u$. Since $m_-(u)$ is the negative solution closest to zero and $t(M) \to 0$ as $M \uparrow 0$, any negative draw with score above u must satisfy $M \leqslant m_{-}(u)$. Therefore,

$$1 - F_{-}(u) = \Pr(t(M) > u \mid M < 0) \leqslant \Pr(M \leqslant m_{-}(u) \mid M < 0) = 2\big[1 - H(|m_{-}(u)|)\big]$$

where the final equality uses symmetry of M.

For the numerator, the change-of-variables formula for the density of $t(M)$ on the negative branch includes a contribution from the root $m_{-}(u)$. Hence, using the symmetry of $h(\cdot)$, we get

$$f_{-}(u) \geqslant \frac{2h(|m_{-}(u)|)}{|t'(m_{-}(u))|}$$

Combining the preceding two displays gives

$$\frac{f_{-}(u)}{1 - F_{-}(u)} \geqslant \frac{\frac{2h(|m_{-}(u)|)}{|t'(m_{-}(u))|}}{2[1 - H(|m_{-}(u)|)]} = \lambda(|m_{-}(u)|)\left|\frac{1}{t'(m_{-}(u))}\right|$$

where $\lambda(M) \equiv h(M)/(1 - H(M))$ is the hazard function.

By Lemma OA.B.1, $t(|m_{-}(u)|) > t(m_{-}(u)) = u = t(m^{+}(u))$. Since $t(M)$ is strictly increasing on the positive branch, it follows that $|m_{-}(u)| > m^{+}(u)$. Strict log-concavity implies that the hazard function $\lambda(\cdot)$ is strictly increasing, which implies $\lambda(|m_{-}(u)|) > \lambda(m^{+}(u))$.

Finally, Lemma OA.B.1 gives $t'(M) \propto \operatorname{sign}(M)(C + 2\rho M)$. Since $t(M)$ is strictly concave on $(-C/\rho, 0)$, and satisfies $t(M) \to 0$ as $M \uparrow 0$, the negative root closest to zero satisfies $t'(m_{-}(u)) < 0$ (Lemma OA.B.1). Hence

$$|t'(m_{-}(u))| = -t'(m_{-}(u)) \propto C - 2\rho|m_{-}(u)| < C < C + 2\rho m^{+}(u) \propto |t'(m^{+}(u))|$$

Combining the hazard comparison and the derivative comparison proves the desired inequality in (5) for every $u \in (0, t^{-}_{\max})$,

$$\frac{f_{-}(u)}{1 - F_{-}(u)} \geqslant \lambda(|m_{-}(u)|)\left|\frac{1}{t'(m_{-}(u))}\right| > \lambda(m^{+}(u))\left|\frac{1}{t'(m^{+}(u))}\right| = \frac{f_{+}(u)}{1 - F_{+}(u)}$$

Thus, we have shown that the expression inside the expectation in (3) is weakly positive for every u, and strictly positive for $u \in (0, t^{-}_{\max})$. Since this interval has positive probability and $\Pr(R_{K+1} = 1 \mid t^{*}_{K} = u) > 0$, it follows that $p_{K+1} - p_K > 0$, which completes the proof. $\square$

Proof of Lemma A.2: Let $h(\cdot)$ denote the density of a single draw M_k. Lemma OA.B.3 establishes that $M_1, \ldots, M_K$ are i.i.d. with a continuous distribution.

For $y > 0$, the event $\{M_K^* \in dy,\ M_K^* > 0\}$ occurs if exactly one candidate draw falls in the infinitesimal interval $(y, y + dy)$ and each of the remaining $K - 1$ draws has absolute t-ratio no larger than $t(y)$. Since draws are i.i.d.,

$$\Pr(M_K^* \in dy,\ M_K^* > 0) = K h(y) \Pr(t(M_k) \leqslant t(y))^{K-1} dy$$

Recall $G_K^+(x) \equiv \Pr(M_K^* \leqslant x \mid M_K^* > 0)$. For $x > 0$,

$$G_K^+(x) = \frac{\Pr(M_K^* \leqslant x,\ M_K^* > 0)}{\Pr(M_K^* > 0)} = \frac{\int_0^x h(y) \Pr(t(M_k) \leqslant t(y))^{K-1} dy}{\int_0^\infty h(y) \Pr(t(M_k) \leqslant t(y))^{K-1} dy}$$

Let $g_K^+(y)$ denote the density corresponding to $G_K^+(y)$. Then

$$g_K^+(y) = \frac{h(y) \Pr(t(M_k) \leqslant t(y))^{K-1}}{\int_0^\infty h(s) \Pr(t(M_k) \leqslant t(s))^{K-1} ds}$$

and

$$\frac{g_{K+1}^+(y)}{g_K^+(y)} = \Pr(t(M_k) \leqslant t(y)) \cdot \left(\frac{\int_0^\infty h(s) \Pr(t(M_k) \leqslant t(s))^{K-1} ds}{\int_0^\infty h(s) \Pr(t(M_k) \leqslant t(s))^{K} ds} \right)$$

The second factor is a constant that does not depend on y. By Lemma OA.B.1, the absolute t-ratio can be written as $t(y) \propto |y|(C + \rho y)$. Thus, when $y > 0$, we have $t(y) \propto y(C + \rho y)$, which is strictly increasing in y.

Since M_k is continuously distributed, $\Pr(t(M_k) \leqslant t(y))$ is also strictly increasing in $y > 0$. Thus, $g_{K+1}^+(y)/g_K^+(y)$ is strictly increasing in y and $g_{K+1}^+(\cdot)$ dominates $g_K^+(\cdot)$ in the monotone likelihood-ratio order. This implies first-order stochastic dominance, so that $G_{K+1}^+(x) < G_K^+(x)$ for all $x > 0$, as desired. □

A.2 Proof of Proposition 2

As in the proof of Proposition 1, normalize the sign of each instrument so that its first-stage coefficient is positive. Under homogeneous instruments, this allows us to write $\pi > 0$. Also normalize $\rho > 0$; when $\rho < 0$, the same argument applies after multiplying all bias statements by $\mathrm{sign}(\rho)$.

When there is a single available instrument ($K = 1$), the median of the 2SLS estimator is biased in the direction of the OLS estimand. This follows from the same argument in Proposition 1, even though β need not equal zero in this case. Under the sign normalizations stated above, this implies that the median bias is positive at $K = 1$. It therefore remains only to show that median bias is monotonically increasing in K under first-stage instrument selection.

To begin, we derive an expression for the first-stage F-statistic. The first stage corresponding to instrument k is the OLS regression: $\tilde{x}_i = \pi \tilde{z}_{ki} + e_{ki}$ where $e_{ki} = \rho u_i + \pi \sum_{\ell \neq k} \tilde{z}_{\ell i}$. Given this is a just-identified specification, the first-stage F-statistic is equal to the square of the t-statistic on the estimate for π. Hence, to derive the first-stage F-statistic function, $F(\cdot)$, first see that:

$$\hat{\pi}_k = \frac{\frac{1}{N}\sum_i \tilde{z}_{ki}\tilde{x}_i}{\frac{1}{N}\sum_i \tilde{z}_{ki}^2} = \frac{C + \rho M_k}{\sigma_{z_k}^2} \qquad \text{and} \qquad \hat{\sigma}_{\hat{\pi}}^2 = \frac{\sigma_e^2}{\sum_i \tilde{z}_{ki}^2} = \frac{\sigma_e^2}{N\sigma_{z_k}^2}.$$

Under homogeneous instruments (Assumption 5) and sample orthogonality (Assumption 4), both $\sigma_{z_k}^2$ and σ_e^2 are constant across k.[30] Hence, the denominator of the first-stage t-statistic is common across candidate instruments, which implies:

$$\implies F(M_k) = \left(\frac{\hat{\pi}_k}{\sqrt{\hat{\sigma}_{\hat{\pi}}^2}}\right)^2 = \frac{N}{\sigma_{z_k}^2 \sigma_e^2} \cdot (C + \rho M_k)^2 \propto (C + \rho M_k)^2 \qquad (6)$$

Thus, maximizing the first-stage F-statistic, $F(M_k)$, is equivalent to maximizing $(C + \rho M_k)^2$. Note that this is a convex quadratic function minimized at $M = -\frac{C}{\rho}$ with $F\big(-\frac{C}{\rho}\big) = 0$.

By Lemma OA.B.2, admissible specifications satisfy $M_k > -\frac{C}{\rho}$. Let $a \equiv -C/\rho$ denote the admissibility cutoff. Importantly, over the admissible region, $F(M)$ is strictly increasing in M. Therefore, first-stage instrument-hacking to maximize the first-stage F-statistic is equivalent to selecting the admissible instrument with the largest value of M_k.

Let $H(\cdot)$ denote the cdf of a single raw draw M_k. Lemma OA.B.3 establishes that $M_1, \ldots, M_K$ are i.i.d. across k. Conditional on at least one admissible specification being available, the selected draw satisfies $M_K^* = \max_{1 \leqslant k \leqslant K} M_k$ conditional on $\max_{1 \leqslant k \leqslant K} M_k > a$.

[30] In particular, if all instruments have the same variance and are mutually orthogonal, then $\text{Var}(e_{ki}) = \rho^2 \text{Var}(u_i) + \pi^2 \sum_{\ell \neq k} \text{Var}(\tilde{z}_{\ell i})$ is identical for every k, because the summation contains the same number of identically distributed instruments, differing only in the omitted index.

Hence, for $x > a$, the distribution of the selected draw conditional on reporting is

$$G_K(x) \equiv \Pr(M_K^* \leqslant x) = \Pr\left(\max_{1\leqslant k\leqslant K} M_k \leqslant x \;\middle|\; \max_{1\leqslant k\leqslant K} M_k > a \right) = \frac{H(x)^K - H(a)^K}{1 - H(a)^K}$$

Let $m_K \equiv \text{Median}(M_K^*)$. Since the distribution is continuous, m_K satisfies $\frac{H(m_K)^K - H(a)^K}{1-H(a)^K} = \frac{1}{2} \iff H(m_K) = \left[\frac{1+H(a)^K}{2}\right]^{1/K}$. Since $H(a) \in (0,1)$, the right-hand side is the generalized mean of the two numbers 1 and $H(a)$, with equal weights and order K. Generalized means are strictly increasing in their order whenever the two numbers are not equal. Therefore $H(m_K)$ is strictly increasing in K. Since $H(\cdot)$ is strictly increasing, it follows that m_K is strictly increasing in K.

Finally, since bias is strictly increasing in M_k over the admissible region[31], the median bias of the selected 2SLS estimator is strictly increasing in K, which is what we wanted to show. □

A.3 Proof of Proposition 4

As in Proposition 1, without loss of generality, normalize $\rho > 0$ and orient each instrument so that $\pi_k > 0$. Under first-stage screening, Lemma OA.B.2 implies that $M_k > -C_k/\rho$ for admissibility. As in the proof of Proposition 1, define $k^* = \arg\max_{1\leqslant k\leqslant K} |t_k|$, where inadmissible realizations are assigned score $-\infty$. Denote reported statistics as $\hat{\beta}_K^* = \hat{\beta}_{k^*}$, $M_K^* = M_{k^*}$, and $t_K^* = t_{k^*}$.

The derivation of the second-stage absolute t-ratio is identical to Lemma OA.B.1, except that the constants are now instrument-specific. Thus,

$$t_k(M_k) = \sqrt{\frac{N}{\sigma_{z_k}^2}} \, \frac{|M_k|(C_k + \rho M_k)}{C_k} \tag{7}$$

First, we show that median bias is weakly positive for any K. Fix a realization of the structural errors $u = (u_1, \ldots, u_N)$ and consider its reflection $-u = (-u_1, \ldots, -u_N)$. Define $M_k(u) = \frac{1}{N}\sum_{i=1}^N \tilde{z}_{ki} u_i$. Since $M_k(\cdot)$ is linear in u, it follows that $M_k(-u) = -M_k(u)$ for every instrument k. Hence, the realizations u and $-u$ generate the same magnitudes $|M_k|$, but opposite signs.

We now prove the sign claim using a sign-reversal argument. By Lemma OA.B.1, replacing a negative realization of M_k with a positive realization of the same magnitude strictly increases the absolute t-ratio whenever both realizations are admissible.

[31] Specifically, $\frac{\partial}{\partial M_k}\left[\hat{\beta}(M_k) - \beta\right] = \frac{C}{(C+\rho M_k)^2} > 0$.

Consider a realization u together with its reflected realization $-u$. If $M_K^*(u) \geqslant 0$, then at least one of u and $-u$ satisfies $M_K^* \geqslant 0$. Otherwise, suppose $M_K^*(u) < 0$, and let the winning instrument under u be j. Since j maximizes the absolute t-ratio under u, we have $|t_j(u)| \geqslant |t_\ell(u)|$ for all ℓ. Under the reflected realization $-u$, instrument j becomes positive while preserving the same magnitude $|M_j|$, and therefore its score strictly increases. Conversely, any instrument that is negative under $-u$ was positive under u, and therefore its score weakly decreases after reflection, with strict inequality whenever $M_\ell \neq 0$. Hence no instrument with $M_\ell(-u) < 0$ can maximize the absolute t-ratio under the reflected realization, except on a probability zero tie event under the continuity assumptions. Therefore, $M_K^*(-u) \geqslant 0$ almost surely.

Thus, for every realization u, at least one of u and $-u$ satisfies $M_K^* \geqslant 0$. Since $u \overset{d}{=} -u$ by Assumption 1, the two realizations are equally likely under the sign-reversal pairing. Therefore, $\Pr(M_K^* \geqslant 0) \geqslant \frac{1}{2}$. Since $\operatorname{sign}(\hat{\beta}_K^*) = \operatorname{sign}(M_K^*)$ under first-stage screening, we have $\Pr(\hat{\beta}_K^* \geqslant 0) \geqslant 1/2$, and hence $\operatorname{Median}(\hat{\beta}_K^*) \geqslant 0$.

We next derive a finite upper bound on the absolute t-ratio attainable on the negative branch. By Lemma OA.B.1, the negative branch is uniquely maximized at $|M_k| = C_k/(2\rho)$. Substituting this into (7) gives

$$t_{k,\max}^- = \sqrt{\frac{N}{\sigma_{z_k}^2}}\,\frac{C_k}{4\rho} = \frac{\sqrt{N}\,\sqrt{\lambda_k}}{4\rho} \leqslant \frac{\sqrt{N}\,\sqrt{\bar{\lambda}}}{4\rho} \equiv t_{\max}^-$$

where the second equality uses $C_k = \pi_k \sigma_{z_k}^2$ and $\lambda_k = \pi_k^2 \sigma_{z_k}^2$; and the inequality uses $\lambda_k \leqslant \bar{\lambda}$. Hence no admissible negative realization can produce an absolute t-ratio exceeding $t_{\max}^-$. This is illustrated in the left panel of Figure A1.

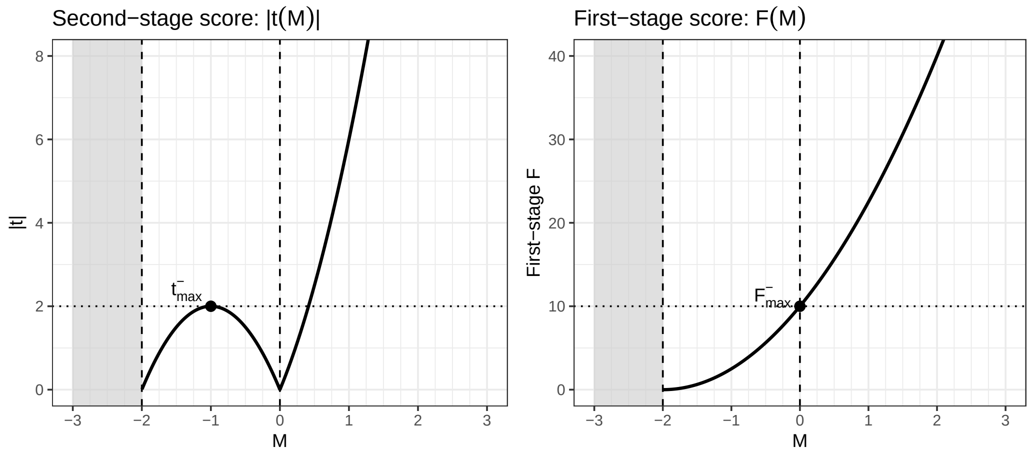
Now consider a positive realization with $b = \hat{\beta}_k$ satisfying $|t_k| \geqslant u$. Solving $\hat{\beta}_k = M_k/(C_k + \rho M_k)$ for M_k gives $M_k = bC_k/(1-\rho b)$. Substituting into (7) yields

$$|t_k| = \sqrt{N}\,\sqrt{\lambda_k}\,\frac{b}{(1-\rho b)^2} \leqslant \bar{B}\frac{b}{(1-\rho b)^2} \equiv \phi(b)$$

where $\bar{B} \equiv \sqrt{N}\,\sqrt{\bar{\lambda}}$. The inequality uses the bound $\lambda_k \leqslant \bar{\lambda}$.

Since $\phi(\cdot)$ is continuous, strictly increasing, satisfies $\phi(b) \to 0$ as $b \downarrow 0$, and satisfies $\phi(b) \to \infty$ as $b \uparrow 1/\rho$, it defines a bijection from $(0, 1/\rho)$ onto $(0, \infty)$. Hence its inverse is well-defined on $(0, \infty)$, and in particular on $(t_{\max}^-, \infty)$. Define $g(u) \equiv \phi^{-1}(u)$ for

Figure A1. Intuition for Propositions 4 and 5



Notes: This figure assumes $\rho > 0$ and $\beta = 0$. The shaded gray area is the inadmissible, screened region.

$u > t^-_{\max}$. Hence, $g : (t^-_{\max}, \infty) \to (0, 1/\rho)$, so $g(u) > 0$ for all $u > t^-_{\max}$. Now $\phi(b) \geqslant |t_k|$ and $|t_k| \geqslant u$ together imply $\phi(b) \geqslant u$. Since ϕ is strictly increasing, applying the inverse function yields $b \geqslant g(u)$.

Hence, any positive realization with $|t_k| \geqslant u$ must also satisfy $\hat{\beta}_k \geqslant g(u)$. Now suppose $\text{Median}(|t^*_K|) > t^-_{\max}$. By the definition of the median, $|t^*_K| \geqslant \text{Median}(|t^*_K|)$ with probability at least one half. Since no negative realization can attain an absolute t-ratio exceeding $t^-_{\max}$, every realization satisfying $|t^*_K| > t^-_{\max}$ must be positive. Therefore, the probability of the event that both $|t^*_K| \geqslant \text{Median}(|t^*_K|)$ and $\hat{\beta}^*_K > 0$ hold is at least one half. Applying the previous implication with $u = \text{Median}(|t^*_K|)$ gives $\hat{\beta}^*_K \geqslant g(\text{Median}(|t^*_K|))$ with probability at least one half. Hence,

$$\text{Median}(\hat{\beta}^*_K) \geqslant g(\text{Median}(|t^*_K|)) > 0$$

Finally, under second-stage instrument-hacking, $|t^*_K| = \max_{1 \leqslant k \leqslant K} |t_k|$. Adding an additional candidate instrument weakly increases this maximum for every realization. Therefore $|t^*_K|$ is weakly increasing in K in the first-order stochastic order, implying that $\text{Median}(|t^*_K|)$ is weakly increasing in K. □

A.4 Proof of Proposition 5

As in Proposition 1, without loss of generality, normalize $\rho > 0$ and orient each instrument so that $\pi_k > 0$. Under first-stage screening, Lemma OA.B.2 implies that $M_k > -C_k/\rho$ for admissibility. Define $k^* = \arg\max_{1 \leqslant k \leqslant K} F_k$, where inadmissible realizations are assigned score $-\infty$. Denote reported statistics as $\hat{\beta}^*_K = \hat{\beta}_{k*}$, $M^*_K = M_{k*}$, and $F^*_K = F_{k*}$.

For instrument k, the estimated first-stage coefficient is

$$\hat{\pi}_k = \frac{\frac{1}{N}\sum_{i=1}^N \widetilde{z}_{ki}\widetilde{x}_i}{\frac{1}{N}\sum_{i=1}^N \widetilde{z}_{ki}^2} = \frac{C_k + \rho M_k}{\sigma^2_{z_k}} \qquad \text{and} \qquad \hat{\sigma}^2_{\hat{\pi}_k} = \frac{\sigma^2_{e,k}}{N\sigma^2_{z_k}}$$

In a just-identified specification, the first-stage F-statistic is equal to the square of the t-statistic. Hence,

$$F_k(M_k) = \frac{N}{\sigma^2_{z_k}\sigma^2_{e,k}}(C_k + \rho M_k)^2.$$

On the admissible region $M_k > -C_k/\rho$, $F_k(M_k)$ is strictly increasing in M_k.

First, we show that median bias is weakly positive for any K. Fix a realization of the structural errors $u = (u_1, \dots, u_N)$ and consider its reflection $-u = (-u_1, \dots, -u_N)$. Define $M_k(u) = \frac{1}{N}\sum_{i=1}^N \widetilde{z}_{ki}u_i$. Since $M_k(\cdot)$ is linear in u, it follows that $M_k(-u) = -M_k(u)$ for every instrument k. Hence, the realizations u and $-u$ generate the same magnitudes $|M_k|$, but opposite signs.

We now prove the sign claim using a sign-reversal argument. Since $F_k(M_k)$ is strictly increasing over the admissible region $M_k > -C_k/\rho$, and equal to $-\infty$ outside this region, replacing a negative realization of M_k with a positive realization of the same magnitude strictly increases the first-stage score.

Consider a realization u and its reflected realization $-u$. If $M^*_K(u) \geqslant 0$, then at least one of u and $-u$ satisfies $M^*_K \geqslant 0$. Otherwise, suppose $M^*_K(u) < 0$, and let j denote the winning instrument under u. Since j maximizes the first-stage F-statistic, $F_j(u) \geqslant F_\ell(u)$ for all ℓ. Under the reflected realization $-u$, instrument j becomes positive while preserving the same magnitude $|M_j|$, and therefore its first-stage score strictly increases. Conversely, any instrument that is negative under $-u$ was positive under u, and therefore its score weakly decreases after reflection, with strict inequality whenever $M_\ell \neq 0$. Hence no instrument with $M_\ell(-u) < 0$ can maximize the first-stage score under the reflected realization, except on a probability zero tie event under the continuity assumptions. Therefore, $M^*_K(-u) \geqslant 0$ almost surely.

Thus, for every realization u, at least one of u and $-u$ satisfies $M^*_K \geqslant 0$. Since $u \overset{d}{=} -u$ by Assumption 1, the two realizations are equally likely under the sign-reversal pairing. Therefore, $\Pr(M^*_K \geqslant 0) \geqslant \frac{1}{2}$. Now see that $\hat{\beta}_k - \beta = M_k/(C_k + \rho M_k)$ and $C_k + \rho M_k > 0$ (Lemma OA.B.2) imply that $\text{sign}(\hat{\beta}^*_K - \beta) = \text{sign}(M^*_K)$. Hence

$\Pr(\hat{\beta}_K^* - \beta \geqslant 0) \geqslant 1/2$, and so $\text{Median}(\hat{\beta}_K^* - \beta) \geqslant 0$.

We next derive a finite upper bound on the first-stage F-statistic attainable on the negative branch. On the admissible negative branch, $-C_k/\rho < M_k < 0$, we have

$$0 < C_k + \rho M_k < C_k \quad \Longrightarrow \quad F_k(M_k) < \frac{N}{\sigma_{z_k}^2 \sigma_{e,k}^2} C_k^2 = \frac{N\lambda_k}{\sigma_{e,k}^2} \leqslant \frac{N\bar{\lambda}}{\underline{\sigma}_e^2} \equiv F_{\max}^-$$

where the equality uses $C_k = \pi_k \sigma_{z_k}^2$ and $\lambda_k = \pi_k^2 \sigma_{z_k}^2$, while the inequality uses $\lambda_k \leqslant \bar{\lambda}$ and $\sigma_{e,k}^2 \geqslant \underline{\sigma}_e^2 > 0$, where $\underline{\sigma}_e^2 \equiv \min_{1 \leqslant k \leqslant \bar{K}} \sigma_{e,k}^2$. Hence no admissible negative realization can produce a first-stage F-statistic exceeding $F_{\max}^-$. This is illustrated in the right panel of Figure A1.

Now consider a positive realization with $b = \hat{\beta}_k - \beta$ satisfying $F_k \geqslant u$. Solving $\hat{\beta}_k - \beta = \frac{M_k}{C_k + \rho M_k}$ for M_k gives $M_k = \frac{bC_k}{1-\rho b}$. Substituting into the expression for $F_k(M_k)$ yields

$$F_k = \frac{N}{\sigma_{z_k}^2 \sigma_{e,k}^2} \frac{C_k^2}{(1-\rho b)^2} \leqslant \frac{F_{\max}^-}{(1-\rho b)^2} \equiv \phi(b)$$

Since $\phi(\cdot)$ is continuous, strictly increasing, and satisfies $\phi(b) \to \infty$ as b approaches $1/\rho$, it defines a bijection from $(0, 1/\rho)$ onto $(F_{\max}^-, \infty)$. Define its inverse $h(u) \equiv \phi^{-1}(u)$. Now $\phi(b) \geqslant F_k$ and $F_k \geqslant u$ together imply $\phi(b) \geqslant u$. Since ϕ is strictly increasing, applying the inverse function yields $b \geqslant h(u)$. Hence, for any $u > F_{\max}^-$ and $F_k \geqslant u$, we have $\hat{\beta}_k - \beta \geqslant h(u) > 0$, where $h(u) > 0$ because $h : (F_{\max}^-, \infty) \to (0, 1/\rho)$.

Suppose $\text{Median}(F_K^*) > F_{\max}^-$. By the definition of the median, $F_K^* \geqslant \text{Median}(F_K^*)$ with probability at least one half. Since no negative realization can attain a first-stage F-statistic exceeding $F_{\max}^-$, every realization satisfying $F_K^* > F_{\max}^-$ must be positive. Therefore, the probability of the event that both $F_k^* \geqslant \text{Median}(F_K^*)$ and $\hat{\beta}_K^* - \beta > 0$ hold is at least one half. Applying the previous implication with $u = \text{Median}(F_K^*)$ gives $\hat{\beta}_K^* - \beta \geqslant h(\text{Median}(F_K^*))$ with probability at least one half. Hence,

$$\text{Median}(\hat{\beta}_K^* - \beta) \geqslant h(\text{Median}(F_K^*)) > 0.$$

Finally, under first-stage instrument-hacking, $F_K^* = \max_{1 \leqslant k \leqslant K} F_k$. Adding an additional candidate instrument weakly increases this maximum for every realization. Therefore, F_K^* first-order stochastically increases with K. Hence, $\text{Median}(F_K^*)$ is weakly increasing in K. $\square$