Supplementary Material to Accompany:

"Instrument-Hacking"

Michael Keane, Timothy Neal, and Patrick Vu

Date: July 1, 2026

- **Appendix OA.A: Results Allowing for Overidentified Specifications**
- **Appendix OA.B: Additional Proofs**
- **Appendix OA.C: Non-Monotonicity With Heterogeneous Instruments**
- **Appendix OA.D: Details on Simulating Instrument-Hacking**
- **Appendix OA.E: Results for Different Values of ρ**
- **Appendix OA.F: Results for Very Strong Instruments**
- **Appendix OA.G: Results for Instrument Strengths from JEL Codes**

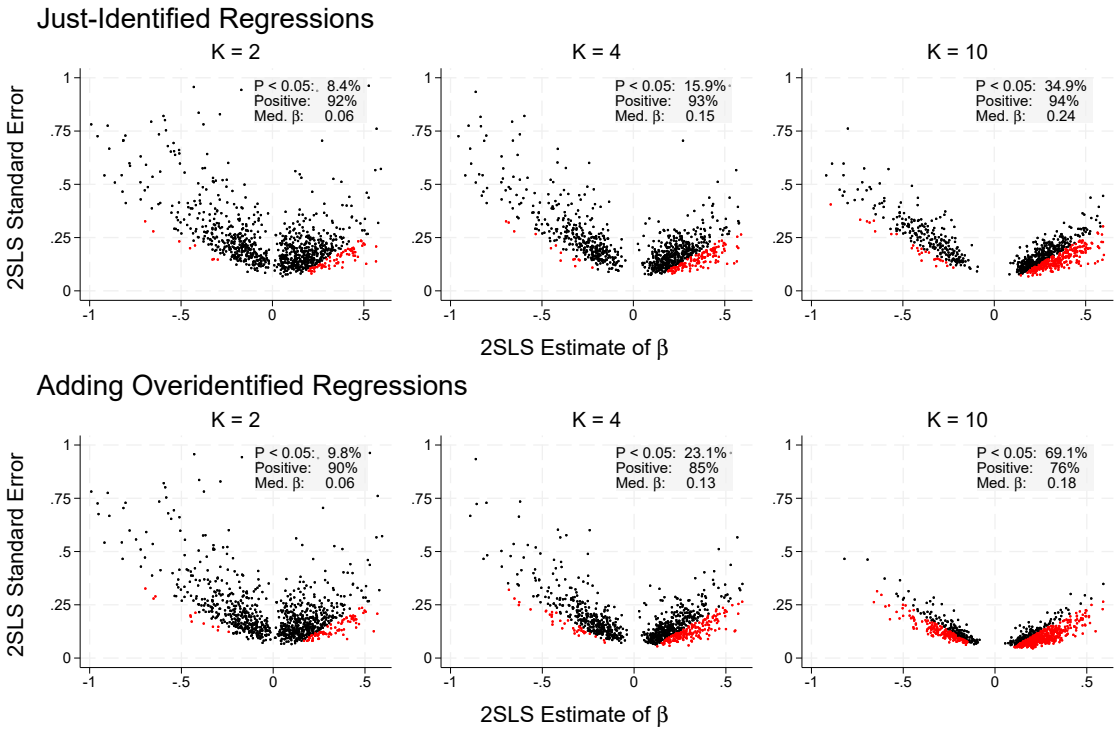
OA.A. **Results Allowing for Overidentified Specifications**

In Section 5 of the main article, we simulated researchers picking from a series of just-identified regressions and reporting the specification that favored their selection strategy. In this appendix, we present results of simulations that allow researchers to consider all possible overidentified combinations of instruments as well. For instance, when two instruments are available a researcher can select from instrument sets (z_1), (z_2), and (z_1,z_2). When three are available, they select from (z_1), (z_2), (z_3), (z_1,z_2), (z_1,z_3), (z_2,z_3), and (z_1,z_2,z_3). The number of potential specifications is $2^K - 1$. This significantly increases the number of specifications to select from, relative to the main results where researchers were restricted to K specifications, but it also allows for the possibility of researchers to utilize more exogenous variation in the regressor to estimate the second-stage effects.

Figure OA.A.1 shows the joint distribution of $\hat{\beta}$ and standard errors using the empirical model in Section 5 of the main article. The top row shows the joint distribution when researchers are restricted to just-identified specifications, as in Table 1 of the main article. The bottom row shows the comparable distribution when overidentified specifications are included within the choice set. When K is small the joint distributions look similar, although size distortion is worse in the overidentified case. It is when K is large that the tradeoff becomes apparent: while allowing overidentified specifications increases size distortion, it reduces median bias somewhat and rejections are less unbalanced around the true null.

Table OA.A.1 shows this more clearly. When K is small size and bias are similar between the two, but as K increases median bias is restrained somewhat while size distortion explodes. Size increases as there are more opportunities to find statistically significant results in each dataset, but rejections are somewhat more balanced since highly significant results are more likely to occur on the negative side of the distribution (relative to just-identified regressions). Median bias remains severe and rejections are unbalanced around the true null. We omit for brevity results for first-stage hacking as they are almost identical, since overidentified specifications rarely have the largest first-stage F statistic. For these reasons, we conclude that the results of the main article are not particularly sensitive to our decision to analyze the case of researchers choosing only from just-identified specifications.

FIGURE OA.A.1. Joint Distribution of $\hat{\beta}$ and Standard Errors Under Second-Stage Hacking



Notes: 2SLS estimates and standard errors when researchers engage in second-stage p-hacking among specifications. The DGP is outlined in Section 5. In all cases, $\beta = 0$, $\rho = 0.5$, and π_k is drawn from a first-stage F distribution of $Gamma(1.66, 18.42)$. Red dots indicate statistically significant estimates at 5% level. '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.

TABLE OA.A.1 – Second-Stage Hacking Allowing Overidentified Regressions

	Only Just-ID Specifications			Include Over-ID Specifications		
K	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias
1	0.040	92%	0.000	0.040	93%	0.000
2	0.084	92%	0.058	0.098	89%	0.056
3	0.122	93%	0.114	0.158	87%	0.103
5	0.194	93%	0.171	0.307	82%	0.144
8	0.293	93%	0.223	0.553	77%	0.175
10	0.349	93%	0.243	0.691	76%	0.182

OA.B. **Additional Proofs**

OA.B.1 Proof of Proposition 3

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 $\text{sign}(\rho)$. As in the proof of Proposition 1, probabilities involving M_K^* are understood conditional on the event that at least one admissible specification exists, since we are concerned with the distribution of reported estimates.

By Lemma OA.B.2, admissible specifications satisfy $M_k > -C/\rho$. On this admissible region, $\hat{\beta}(M) = \frac{M}{C+\rho M}$ is strictly increasing in M. Therefore, it is enough to prove the corresponding result for the selected draw M_K^*. We will show separately that this holds for both lexicographic-threshold preferences and smooth preferences.

Lexicographic Preferences. As shown in the proof of Proposition 2, the first-stage statistic satisfies $F(M_k) \propto (C + \rho M_k)^2$, which is strictly increasing over the admissible region. Hence, for any threshold $c > 0$, there exists a unique $m_c > -C/\rho$ such that $F(M_k) \geqslant c \iff M_k \geqslant m_c$. Also, by Lemma OA.B.1, when $\beta = 0$, $t(M_k) \propto |M_k|(C + \rho M_k)$. Thus the lexicographic-threshold rule is equivalent to maximizing the score

$$s(M_k) = \begin{cases} (0, M_k), & -C/\rho < M_k < m_c, \\ (1, t(M_k)), & M_k \geqslant m_c, \end{cases}$$

where the ordering is lexicographic and inadmissible draws receive score $-\infty$.

First, suppose that $m_c \geqslant 0$. Then all threshold-clearing draws lie on the positive branch, where $t(M_k)$ is strictly increasing in M_k by Lemma OA.B.1. Thus, if any draw clears the threshold, the lexicographic rule selects the largest admissible value of M_k, since all non-clearing draws have $M_k < m_c$ and all clearing draws have $M_k \geqslant m_c$. If no draw clears the threshold, the rule maximizes $F(M_k)$, which is also equivalent to maximizing M_k on the admissible region by the proof of Proposition 2. Hence the rule is equivalent to first-stage instrument-hacking, and Proposition 2 immediately gives median bias toward the OLS estimand and strict monotonicity in K.

Next, consider the case $m_c < 0$. As shown in the proof of Proposition 1, when there is a single available instrument ($K = 1$), the median of the selected draw satisfies $\text{Median}(M_1^*) > 0$, and hence $\text{Median}(\hat{\beta}_1^*) > 0$. It therefore remains only to show that

Median(M_K^*) is strictly increasing in K under lexicographic-threshold preferences.

The remainder of the proof follows the same structure as Proposition 1. In particular, the argument there relies on two claims: first, that the probability the selected draw is positive is strictly increasing in K; and second, that conditional on the selected draw being positive, the selected draw shifts to the right as K increases. The second claim follows by the same argument as Lemma A.2, because for $y > 0$ we have $s(y) = (1, t(y))$ and $t(y)$ is strictly increasing on the positive branch by Lemma OA.B.1. Thus, the only part that must be reproved for lexicographic-threshold preferences is

$$p_{K+1} > p_K, \qquad p_K \equiv \Pr(M_K^* > 0)$$

Once this is established, the decomposition argument from Proposition 1 applies unchanged and gives strict monotonicity of Median(M_K^*).

We now prove $p_{K+1} > p_K$. Define the winning score at round K as $s_K^* \equiv \max_{1 \leqslant k \leqslant K} s(M_k)$. Let R_{K+1} denote the event that the $(K+1)$-st draw replaces the winner. As in the proof of Lemma A.1, $p_{K+1} - p_K$ can be written as

$$\begin{aligned}
&\mathbb{E}\Big[\Pr(R_{K+1}=1 \mid s_K^*=u)\Big\{\Pr(M_{K+1}>0 \mid R_{K+1}=1, s_K^*=u) - \Pr(M_K^*>0 \mid R_{K+1}=1, s_K^*=u)\Big\}\Big] \\
&= \mathbb{E}\Big[\Pr(R_{K+1}=1 \mid s_K^*=u)\Big\{\Pr(M_{K+1}>0 \mid R_{K+1}=1, s_K^*=u) - \Pr(M_K^*>0 \mid s_K^*=u)\Big\}\Big] \qquad (1)
\end{aligned}$$

Here the expectation is taken over the distribution of the incumbent score $u = s_K^*$. The second equality follows from the fact that, conditional on $s_K^* = u$, the replacement event depends only on the independent $(K+1)$-st draw, and is therefore conditionally independent of the event $\{M_K^* > 0\}$.

Hence, because $\Pr(R_{K+1} = 1 \mid s_K^* = u) > 0$, it suffices to show that, after replacement at any given u, the replacing draw is weakly more likely to be positive than the incumbent, with strict inequality for some incumbent scores that occur with positive probability.

First, consider the case where the incumbent does not clear the threshold, so $M_K^* < m_c < 0$. Then $\Pr(M_K^* > 0 | s_K^* = u) = 0$. Because $m_c < 0$, any positive $(K+1)$-st draw clears the threshold and, since the incumbent does not, necessarily replaces it. Hence, $\Pr(M_{K+1} > 0 | R_{K+1} = 1, s_K^* = u) > 0$, giving the desired inequality.

Next, consider the case where the incumbent clears the threshold $(M_K^* \geqslant m_c)$, so its score has the form $u = (1, \tau)$, where τ is the realized value of $t(M_K^*)$. For any

second-stage score value τ, define

$$F_+(\tau) \equiv \Pr(t(M) \leqslant \tau \mid M > 0), \qquad F_-(\tau) \equiv \Pr(t(M) \leqslant \tau \mid m_c \leqslant M < 0),$$

and let $f_+(\tau)$ and $f_-(\tau)$ denote the corresponding conditional densities. This is the same branch-specific notation used in Lemma A.1, except that the negative branch is now restricted to threshold-clearing negative draws. Repeating the derivation from that lemma, proving the desired inequality is equivalent to showing that

$$\frac{f_-(\tau)}{1 - F_-(\tau)} > \frac{f_+(\tau)}{1 - F_+(\tau)} \tag{2}$$

whenever $1 - F_-(\tau) > 0$. If $1 - F_-(\tau) = 0$, then no threshold-clearing negative draw can replace an incumbent with score $(1, \tau)$, so the replacing draw is weakly more likely to be positive than the incumbent.[1] Hence, we assume $1 - F_-(\tau) > 0$ for the remainder.

To prove the hazard inequality in (2), we rewrite both ratios in terms of the cdf and density of a single draw of M, denoted by $H(\cdot)$ and $h(\cdot)$. Consider first the right-hand side. Since $t(M)$ is strictly increasing on the positive branch, each τ has a unique positive inverse, denoted $m_+(\tau)$, satisfying $t(m_+(\tau)) = \tau$. By the same arguments used in Lemma A.1, we can write

$$\frac{f_+(\tau)}{1 - F_+(\tau)} = \frac{h(m_+(\tau))}{1 - H(m_+(\tau))} \left| \frac{1}{t'(m_+(\tau))} \right| = \lambda(m_+(\tau)) \left| \frac{1}{t'(m_+(\tau))} \right| \tag{3}$$

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

Now consider the left-hand side of (2). Recall that $1 - F_-(\tau) > 0$, which implies a threshold-clearing negative draw with score above τ. Since $t(M)$ is continuous and $t(M) \to 0$ as $M \uparrow 0$, there exists a negative solution closest to zero; denote it by $m_-(\tau) < 0$, so that $t(m_-(\tau)) = \tau$, with $m_c \leqslant m_-(\tau) < 0$.

By definition,

$$1 - F_-(\tau) = \frac{\int_{\{m \in [m_c, 0):\, t(m) > \tau\}} h(m)\, dm}{\Pr(m_c \leqslant M < 0)}$$

[1] $1 - F_-(\tau) = 0$ when the incumbent's score τ is at least as large as the maximum score attainable by a threshold-clearing negative draw.

and the change-of-variables formula gives

$$f_-(\tau) = \frac{1}{\Pr(m_c \leqslant M < 0)} \sum_{\{m \in [m_c, 0):\, t(m) = \tau\}} \frac{h(m)}{|t'(m)|}$$

Since $m_-(\tau)$ is the negative solution closest to zero and $t(M) \to 0$ as $M \uparrow 0$, any threshold-clearing negative draw with score above τ must satisfy $M \leqslant m_-(\tau)$. Therefore,

$$1 - F_-(\tau) = \frac{\Pr(t(M) > \tau,\ m_c \leqslant M < 0)}{\Pr(m_c \leqslant M < 0)} \leqslant \frac{\Pr(M \leqslant m_-(\tau))}{\Pr(m_c \leqslant M < 0)} = \frac{1 - H(|m_-(\tau)|)}{\Pr(m_c \leqslant M < 0)}$$

where the last equality uses symmetry of M.

Moreover, since $m_-(\tau)$ is one of the roots in the change-of-variables formula for $f_-(\tau)$, and since $h(\cdot)$ is symmetric,

$$f_-(\tau) \geqslant \frac{h(|m_-(\tau)|)}{\Pr(m_c \leqslant M < 0)|t'(m_-(\tau))|}.$$

Combining this with the preceding bound on $1 - F_-(\tau)$ gives

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

It remains to compare this lower bound to the positive-branch ratio in (3). Since $m_-(\tau)$ is threshold-clearing, $-C/\rho < m_c \leqslant m_-(\tau) < 0$, and hence $0 < |m_-(\tau)| < C/\rho$. Thus, Lemma OA.B.1 implies $t(|m_-(\tau)|) > t(m_-(\tau))$. Since $t(M)$ is strictly increasing on the positive branch and $t(m_+(\tau)) = t(m_-(\tau)) = \tau$, it follows that $|m_-(\tau)| > m_+(\tau)$. Strict log-concavity implies that $\lambda(\cdot)$ is strictly increasing, so $\lambda(|m_-(\tau)|) > \lambda(m_+(\tau))$.

Next, by the derivative formula in Lemma OA.B.1, $|t'(m_+(\tau))| \propto C + 2\rho m_+(\tau)$. Similarly, on the negative branch, $|t'(m_-(\tau))| \propto |-C - 2\rho m_-(\tau)| = C - 2\rho|m_-(\tau)|$, where the right-hand side is positive because $m_-(\tau)$ is the closest-to-zero negative solution and hence lies on the part of the negative branch where $t(M)$ is decreasing in M (Lemma OA.B.1). It follows immediately that $|t'(m_-(\tau))| < |t'(m_+(\tau))|$.

Putting this together implies

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

This is the desired hazard inequality. Combining the case in which the incumbent does not clear the threshold ($M_K^* < m_c < 0$) with the threshold-clearing case ($M_K^* \geqslant m_c$), the bracketed difference in (1) is weakly positive for every incumbent score u. Moreover, the first case occurs with positive probability and gives strict positivity of the bracketed difference. Since $\Pr(R_{K+1} = 1 \mid s_K^* = u) > 0$, it follows that $p_{K+1} > p_K$, completing the proof.

Smooth Preferences. Under smooth preferences, the researcher maximizes $s_k = \lambda F_k + (1-\lambda)|t_k|$, with $\lambda \in (0,1)$. After normalizing by a positive constant, this is equivalent to maximizing $s(M) = \widetilde{\lambda}(C+\rho M)^2 + (1-\widetilde{\lambda})|M|(C+\rho M)$, with $\widetilde{\lambda} \in (0,1)$; as before, inadmissible draws are assigned score $-\infty$.[2]

We first state three properties of this score. First, on the positive branch, when $M > 0$, $s'(M) = 2\widetilde{\lambda}\rho(C+\rho M) + (1-\widetilde{\lambda})(C+2\rho M) > 0$. Second, for any $M \in (0, C/\rho)$, $s(M) - s(-M) > 0$. Third, define $s_0 \equiv s(0) = \widetilde{\lambda}C^2$. Since all positive draws have score at least s_0, any admissible draw with score below s_0 must be negative.

As in the lexicographic case, the case $K = 1$ follows from the proof of Proposition 1. Thus, it remains to show that $\text{Median}(M_K^*)$ is strictly increasing in K.

The proof again follows the decomposition argument from Proposition 1. Since $s(y)$ is strictly increasing for $y > 0$, the same monotone-likelihood-ratio argument as in Lemma A.2 implies that, conditional on the selected draw being positive, the selected draw shifts to the right as K increases. Thus, as in the lexicographic case, proving monotonicity of median bias requires showing that $p_{K+1} > p_K$ under smooth preferences, where $p_K \equiv \Pr(M_K^* > 0)$.

Define the winning score at round K as $s_K^* \equiv \max_{1 \leqslant k \leqslant K} s(M_k)$, and let R_{K+1} denote the event that the $(K+1)$-st draw replaces the incumbent winner. The same argument used above means that $p_{K+1} - p_K$ can be written exactly as in (1). Thus, as before, it suffices to show that the bracketed difference is weakly positive for every

[2]Specifically, under homogeneous instruments, the first-stage statistic and second-stage t-ratio can be written as $F(M) = a(C+\rho M)^2$ and $|t(M)| = b|M|(C+\rho M)$, where $a, b > 0$ are constants common across instruments. Thus $\lambda F(M) + (1-\lambda)|t(M)| = \lambda a(C+\rho M)^2 + (1-\lambda)b|M|(C+\rho M)$. Dividing by the positive constant $\lambda a + (1-\lambda)b$ gives the displayed score with $\widetilde{\lambda} \equiv \lambda a/[\lambda a + (1-\lambda)b] \in (0,1)$.

incumbent score u, and strictly positive for some incumbent scores that occur with positive probability.

First, consider scores below s_0. Since all positive draws have score at least s_0, an incumbent with score $u < s_0$ must be negative. Hence, $\Pr(M_K^* > 0 \mid s_K^* = u) = 0$. But any positive $(K+1)$-st draw has score at least $s_0 > u$, and therefore replaces the incumbent. Hence $\Pr(M_{K+1} > 0 \mid R_{K+1} = 1, s_K^* = u) > 0$ and the bracketed difference in (1) is strictly positive for such scores.

Next, consider scores $u \geqslant s_0$. Define $F_+(u) \equiv \Pr(s(M) \leqslant u \mid M > 0)$ and $F_-(u) \equiv \Pr(s(M) \leqslant u \mid -C/\rho < M < 0)$; and let $f_+(u)$ and $f_-(u)$ denote the corresponding conditional densities. As in Lemma A.1, proving the desired inequality is equivalent to showing the following hazard inequality

$$\frac{f_-(u)}{1 - F_-(u)} > \frac{f_+(u)}{1 - F_+(u)} \tag{4}$$

whenever $1 - F_-(u) > 0$. If $1 - F_-(u) = 0$, then no admissible negative draw can replace an incumbent with score u, so the replacing draw is weakly more likely to be positive than the incumbent. Hence, suppose $1 - F_-(u) > 0$ for the remainder.

The boundary case $u = s_0$ can be ignored because the incumbent score has a continuous distribution. Suppose $u > s_0$. As for the lexicographic proof, let $H(\cdot)$ and $h(\cdot)$ denote the cdf and density of a single draw of M. Since $s(M)$ is strictly increasing on the positive branch, there is a unique positive inverse $m_+(u) > 0$ satisfying $s(m_+(u)) = u$. The same change-of-variables calculation as in Lemma A.1 gives

$$\frac{f_+(u)}{1 - F_+(u)} = r(m_+(u)) \left| \frac{1}{s'(m_+(u))} \right| \tag{5}$$

where $r(x) \equiv h(x)/(1 - H(x))$ denotes the hazard function.

Now consider the negative branch. Since $1 - F_-(u) > 0$, there exists an admissible negative draw with score above u. Since $s(M)$ is continuous and $s(M) \to s_0 < u$ as $M \uparrow 0$, there exists a negative solution closest to zero; denote it by $m_-(u) < 0$, so that $s(m_-(u)) = u$. Since $m_-(u)$ is the negative solution closest to zero and $s(M) \to s_0 < u$ as $M \uparrow 0$, any admissible negative draw with score above u must satisfy $M \leqslant m_-(u)$. Therefore,

$$1 - F_-(u) = \frac{\Pr(s(M) > u,\ -C/\rho < M < 0)}{\Pr(-C/\rho < M < 0)} \leqslant \frac{\Pr(M \leqslant m_-(u))}{\Pr(-C/\rho < M < 0)} = \frac{1 - H(|m_-(u)|)}{\Pr(-C/\rho < M < 0)}$$

Moreover, since $m_-(u)$ is one of the roots in the change-of-variables formula for $f_-(u)$, symmetry of $h(\cdot)$ gives

$$f_-(u) \geqslant \frac{h(|m_-(u)|)}{\Pr(-C/\rho < M < 0)|s'(m_-(u))|}$$

Combining the last two displays yields

$$\frac{f_-(u)}{1-F_-(u)} \geqslant r(|m_-(u)|)\left|\frac{1}{s'(m_-(u))}\right|$$

It remains to compare this lower bound to the positive-branch ratio in (5). First, note that $m_-(u)$ is admissible and negative, so $0 < |m_-(u)| < C/\rho$. The positive-negative score comparison established above implies $s(|m_-(u)|) > s(m_-(u)) = u = s(m_+(u))$. Since $s(M)$ is strictly increasing on the positive branch, it follows that $|m_-(u)| > m_+(u)$. Moreover, strict log-concavity implies that the hazard function $r(\cdot)$ is strictly increasing, giving $r(|m_-(u)|) > r(m_+(u))$.

Because $m_-(u)$ is the negative solution closest to zero, it is the right endpoint of the negative values of M whose score exceeds u. Hence, as M increases through $m_-(u)$, the score crosses the level u from above to below. This implies $s'(m_-(u)) < 0$. Therefore

$$\begin{aligned}
s'(m_+(u)) - |s'(m_-(u))| &= s'(m_+(u)) + s'(m_-(u)) \\
&= \left[2\widetilde{\lambda}\rho(C+\rho m_+(u)) + (1-\widetilde{\lambda})(C+2\rho m_+(u))\right] \\
&\quad + \left[2\widetilde{\lambda}\rho(C+\rho m_-(u)) - (1-\widetilde{\lambda})(C+2\rho m_-(u))\right] \\
&= 2\widetilde{\lambda}\rho\{2C+\rho(m_+(u)+m_-(u))\} + (1-\widetilde{\lambda})2\rho\{m_+(u)-m_-(u)\} \\
&= 2\widetilde{\lambda}\rho\{2C+\rho(m_+(u)-|m_-(u)|)\} + 2(1-\widetilde{\lambda})\rho\{m_+(u)+|m_-(u)|\} > 0
\end{aligned}$$

The final inequality follows because $0 < |m_-(u)| < C/\rho$ implies $2C + \rho(m_+(u) - |m_-(u)|) > 0$, and $m_+(u) + |m_-(u)| > 0$.

Putting these facts together gives

$$\frac{f_-(u)}{1-F_-(u)} \geqslant r(|m_-(u)|)\left|\frac{1}{s'(m_-(u))}\right| > r(m_+(u))\left|\frac{1}{s'(m_+(u))}\right| = \frac{f_+(u)}{1-F_+(u)}.$$

This is the desired hazard inequality. Therefore, the bracketed difference in (1) is weakly positive for every incumbent score u. Moreover, the case $u < s_0$ occurs

with positive probability and gives strict positivity of the bracketed difference. Since $\Pr(R_{K+1} = 1 \mid s_K^* = u) > 0$, it follows that $p_{K+1} > p_K$, which completes the proof. □

Lemma OA.B.1 (t-Ratio Properties)**.** Consider model (1) with $\beta = 0$ and Assumptions 1–5. Let $\rho, \pi > 0$. Then the absolute t-ratio of a just-identified specification satisfies the following properties:

1. (Closed-Form Expressions) The function can be expressed as

$$t(M) \propto |M|(C + \rho M) \tag{6}$$

   For $M > -C/\rho$ and $M \neq 0$, its first and second derivatives satisfy

$$t'(M) \propto \operatorname{sign}(M)(C + 2\rho M) \tag{7}$$
$$t''(M) \propto \operatorname{sign}(M) \tag{8}$$

2. (Unique Negative Maximum) There exists a unique $M_{\max}^- \in (-C/\rho, 0)$ such that $t'(M_{\max}^-) = 0$ and
$$t(M_{\max}^-) = \max_{M \in (-C/\rho, 0)} t(M)$$

3. (Positive Dominance) For all $M \in (0, C/\rho)$,
$$t(M) > t(-M)$$

Proof of Lemma OA.B.1, Part 1. The absolute t-ratio is given by

$$t(M) = \left| \frac{\hat{\beta}(M)}{\sqrt{\operatorname{Var}(\hat{\beta}(M))}} \right| \tag{9}$$

We first derive a closed-form expression for the variance term. Let γ_k and π_k denote the reduced-form and first-stage coefficients for instrument k, and let $(\hat{\gamma}_k, \hat{\pi}_k)$ denote the corresponding estimators. Hence, the true effect β is given by $g(\gamma_k, \pi_k) = \frac{\gamma_k}{\pi_k}$.

By the Delta method, the asymptotic variance of the 2SLS estimator based on

instrument k is

$$\nabla g(\gamma_k, \pi_k)^T \Sigma_k \nabla g(\gamma_k, \pi_k) = \frac{1}{\pi_k^2}\sigma_{\gamma,k}^2 - 2\frac{\gamma_k}{\pi_k^3}\sigma_{\gamma\pi,k} + \frac{\gamma_k^2}{\pi_k^4}\sigma_{\pi,k}^2 = \frac{\sigma_{\pi,k}^2\beta^2 - 2\sigma_{\gamma\pi,k}\beta + \sigma_{\gamma,k}^2}{\pi_k^2} \tag{10}$$

where

$$\Sigma_k = \begin{pmatrix} \sigma_{\gamma,k}^2 & \sigma_{\gamma\pi,k} \\ \sigma_{\gamma\pi,k} & \sigma_{\pi,k}^2 \end{pmatrix}$$

denotes the covariance block corresponding to $(\hat{\gamma}_k, \hat{\pi}_k)$ within the full covariance matrix Σ from Assumption 2. The second equality uses $\beta = \gamma_k/\pi_k$. The conventional variance estimator replaces (γ_k, π_k) with $(\hat{\gamma}_k, \hat{\pi}_k)$.

Next, we express the covariance terms as functions of the model parameters.

Since $\beta = 0$, the structural equation reduces to $y_i = u_i$. Instrument validity therefore implies

$$\gamma_k = \frac{\mathrm{Cov}(z_k, y)}{\mathrm{Var}(z_k)} = \frac{\mathrm{Cov}(z_k, u)}{\mathrm{Var}(z_k)} = 0$$

Since z_{ki} is binary, the reduced-form OLS estimator is equal to the difference in sample means. Since $\gamma_k = 0$ and $y_i = u_i$,

$$\hat{\gamma}_k - \gamma_k = \hat{\gamma}_k = \frac{1}{N_1}\sum_{\{i:z_{ki}=1\}} u_i - \frac{1}{N_0}\sum_{\{i:z_{ki}=0\}} u_i \equiv \bar{u}_1 - \bar{u}_0$$

Similarly, since z_{ki} is binary, the first-stage OLS estimator is the difference in conditional sample means. Using $x_i = \delta + \sum_{\ell=1}^{\bar{K}} \pi_\ell z_{\ell i} + \rho u_i$ gives

$$\hat{\pi}_k = \pi_k + \sum_{\ell \neq k} \pi_\ell(\bar{z}_{\ell,1} - \bar{z}_{\ell,0}) + \rho(\bar{u}_1 - \bar{u}_0)$$

where $\bar{z}_{\ell,1} \equiv \frac{1}{N_1}\sum_{\{i:z_{ki}=1\}} z_{\ell i}$ and $\bar{z}_{\ell,0} \equiv \frac{1}{N_0}\sum_{\{i:z_{ki}=0\}} z_{\ell i}$. Moreover, for $\ell \neq k$,

$$\bar{z}_{\ell,1} - \bar{z}_{\ell,0} = \frac{\widehat{\mathrm{Cov}}(z_k, z_\ell)}{\widehat{\mathrm{Var}}(z_k)} = 0$$

by Assumption 4. Hence $\hat{\pi}_k - \pi_k = \rho(\bar{u}_1 - \bar{u}_0)$.

Combining the previous two expressions gives

$$\sqrt{N}\begin{pmatrix}\hat{\gamma}_k - \gamma_k \\ \hat{\pi}_k - \pi_k\end{pmatrix} = \sqrt{N}(\bar{u}_1 - \bar{u}_0)\begin{pmatrix}1 \\ \rho\end{pmatrix}$$

Under the present DGP, the covariance block corresponding to instrument k therefore takes the form

$$\begin{aligned}\Sigma_k &= \begin{pmatrix}\sigma^2_{\gamma,k} & \sigma_{\gamma\pi,k} \\ \sigma_{\gamma\pi,k} & \sigma^2_{\pi,k}\end{pmatrix} \\ &= \text{Var}\left(\sqrt{N}(\bar{u}_1 - \bar{u}_0)\begin{pmatrix}1 \\ \rho\end{pmatrix}\right) \\ &= N\sigma_u^2\left(\frac{1}{N_1} + \frac{1}{N_0}\right)\begin{pmatrix}1 \\ \rho\end{pmatrix}\begin{pmatrix}1 & \rho\end{pmatrix} \\ &= \frac{1}{\sigma_z^2}\begin{pmatrix}1 & \rho \\ \rho & \rho^2\end{pmatrix} \qquad (11)\end{aligned}$$

where the final equality uses the normalization $\sigma_u^2 = 1$ and $\sigma_z^2 = N_1 N_0/N^2$.[3]

Substituting these coefficients into the plug-in variance estimator corresponding to equation (10) gives

$$\begin{aligned}\text{Var}\left(\hat{\beta}(M)\right) &= \frac{1}{N}\frac{\sigma^2_{\pi,k}\hat{\beta}(M)^2 - 2\sigma_{\gamma\pi,k}\hat{\beta}(M) + \sigma^2_{\gamma,k}}{\hat{\pi}_k^2} \\ &= \frac{1}{N\sigma_z^2}\frac{\rho^2\hat{\beta}(M)^2 - 2\rho\hat{\beta}(M) + 1}{\hat{\pi}_k^2} \\ &= \frac{\sigma_z^2}{N}\frac{(\rho\hat{\beta}(M) - 1)^2}{(C + \rho M)^2} \\ &= \frac{\sigma_z^2}{N}\frac{C^2}{(C + \rho M)^4}\end{aligned}$$

where the third line uses $\hat{\pi}_k = \frac{\sum_i \tilde{z}_{ki}\tilde{x}_i}{\sum_i \tilde{z}_{ki}^2} = \frac{C+\rho M}{\sigma_z^2}$ and the last line uses $\hat{\beta}(M) = \frac{M}{C+\rho M}$.

[3]Normalizing $\sigma_u^2 = 1$ is without loss of generality, because only relative scaling matters for the t-ratio.

Hence, the absolute t-ratio in (9) can be written as

$$t(M) = \left| \frac{\hat{\beta}(M)}{\sqrt{\text{Var}\left(\hat{\beta}(M)\right)}} \right| = \left| \frac{M/(C+\rho M)}{\sqrt{\frac{\sigma_z^2}{N}\frac{C^2}{(C+\rho M)^4}}} \right| = \sqrt{\frac{N}{\sigma_z^2}} \left| \frac{M(C+\rho M)}{C} \right|$$

By Lemma OA.B.2, first-stage screening implies $M > -C/\rho \iff C + \rho M > 0$. Therefore, $|M(C+\rho M)| = |M|(C+\rho M)$ and hence $t(M) \propto |M|(C+\rho M)$.

Differentiating separately on the positive and negative branches gives the (proportional) first and second derivative expressions. □

Proof of Lemma OA.B.1, Part 2. Consider the negative branch

$$-\frac{C}{\rho} < M < 0.$$

Using the proportional absolute-t function from equation (6),

$$g(M) = |M|(C+\rho M).$$

On this branch, $|M| = -M$, and therefore

$$g(M) = (-M)(C+\rho M) = -MC - \rho M^2.$$

Hence the maximizer satisfies

$$g'(M^-_{\max}) = -C - 2\rho M^-_{\max} = 0 \quad \iff \quad M^-_{\max} = -\frac{C}{2\rho}.$$

Given $\rho, C > 0$,

$$-\frac{C}{\rho} < -\frac{C}{2\rho} < 0.$$

Moreover, $M^-_{\max}$ is the unique maximizer because

$$g''(M) = -2\rho < 0,$$

so $g(\cdot)$ is strictly concave over this branch.

Finally, since $t(\cdot)$ is proportional to $g(\cdot)$ with a positive constant that is invariant to M, $M^-_{\max}$ is also the unique maximizer of $t(\cdot)$ on the negative branch. □

Proof of Lemma OA.B.1, Part 3. Fix $M \in \left(0, \frac{C}{\rho}\right)$. Then using the proportional function in equation 6, we have

$$g(M) - g(-M) = M(C + \rho M) - M(C - \rho M) = 2\rho M^2 > 0$$

The proportionality constant for $g(\cdot)$ is positive and invariant to M, and it therefore also follows that $t(M) - t(-M) > 0$, which is what we wanted to show. □

Lemma OA.B.2 (Screening Region)**.** Let $\rho, \pi_k > 0$. Under Assumptions 3 and 4, we have that $\text{sign}(C_k) = \text{sign}(\pi_k) = 1$ and $M_k > -\frac{C_k}{\rho}$.

Proof. Under sample orthogonality (Assumption 4), $C_k = \pi_k \sigma^2_{z_k}$. Since $\sigma^2_{z_k} > 0$ for any nondegenerate instrument and $\pi_k > 0$ by hypothesis, $\text{sign}(C_k) = \text{sign}(\pi_k) = 1$.

Next, observe that

$$\hat{\pi}_k = \frac{\frac{1}{N}\sum_{i=1}^N \widetilde{z}_{ki} x_i}{\sigma^2_{z_k}} = \frac{C_k + \rho M_k}{\sigma^2_{z_k}}$$

Since $\sigma^2_{z_k} > 0$, $\text{sign}(\hat{\pi}_k) = \text{sign}(C_k + \rho M_k)$.

Under first-stage screening (Assumption 3), the reported instrument satisfies $\text{sign}(\hat{\pi}_k) = \text{sign}(\pi_k) = 1$. Hence, $C_k + \rho M_k > 0 \iff M_k > -\frac{C_k}{\rho}$, as required. □

Lemma OA.B.3 (I.I.D. Candidate Error Components)**.** Under Assumptions 1–5, the candidate error components

$$M_k \equiv \frac{1}{N}\sum_{i=1}^N \widetilde{z}_{ki} u_i, \qquad k = 1, \ldots, K,$$

are i.i.d. with a continuous distribution.

Proof. Homogeneity of instruments implies that candidate instruments have a common assignment share, $q_k = q$, and are symmetrically distributed across k (Assumption 5). Hence each M_k has the same marginal distribution.

Now fix $k \neq \ell$. Since

$$M_k = \frac{1}{N}\sum_{i=1}^N \widetilde{z}_{ki} u_i, \qquad M_\ell = \frac{1}{N}\sum_{i=1}^N \widetilde{z}_{\ell i} u_i,$$

we have

$$\mathrm{Cov}(M_k, M_\ell) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \widetilde{z}_{ki} \widetilde{z}_{\ell j} \,\mathrm{Cov}(u_i, u_j)$$

Structural errors are independent across observations (Assumption 1), so $\mathrm{Cov}(u_i, u_j) = 0$ for $i \neq j$. Hence

$$\mathrm{Cov}(M_k, M_\ell) = \frac{\sigma_u^2}{N^2} \sum_{i=1}^{N} \widetilde{z}_{ki} \widetilde{z}_{\ell i}.$$

By sample orthogonality of instruments (Assumption 4), we have $\sum_{i=1}^{N} \widetilde{z}_{ki} \widetilde{z}_{\ell i} = 0$, and therefore

$$\mathrm{Cov}(M_k, M_\ell) = 0$$

Assumption 2 imposes a joint Gaussian approximation on the first-stage estimation errors. To relate this approximation to $(M_1, \ldots, M_K)$, note that the first-stage coefficient from regressing x_i on $(1, z_{ki})$ is

$$\hat{\pi}_k = \frac{\frac{1}{N} \sum_{i=1}^{N} \widetilde{z}_{ki} \widetilde{x}_i}{\sigma_{z_k}^2}.$$

The numerator can be written as

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \widetilde{z}_{ki} \widetilde{x}_i &= \frac{1}{N} \sum_{i=1}^{N} \widetilde{z}_{ki} \left(\sum_{\ell=1}^{K} \pi_\ell \widetilde{z}_{\ell i} + \rho u_i \right) \\
&= \sum_{\ell=1}^{K} \pi_\ell \frac{1}{N} \sum_{i=1}^{N} \widetilde{z}_{ki} \widetilde{z}_{\ell i} + \rho \frac{1}{N} \sum_{i=1}^{N} \widetilde{z}_{ki} u_i \\
&= \pi_k \sigma_{z_k}^2 + \rho M_k.
\end{aligned}$$

where the third equality uses sample orthogonality together with the definitions of $\sigma_{z_k}^2$ and M_k.

Hence

$$\hat{\pi}_k = \pi_k + \rho \frac{M_k}{\sigma_{z_k}^2} \iff M_k = \frac{\sigma_{z_k}^2}{\rho} (\hat{\pi}_k - \pi_k).$$

Thus each M_k is a scalar multiple of the first-stage estimation error $\hat{\pi}_k - \pi_k$. Since Assumption 2 imposes a joint Gaussian approximation on the vector of first-stage estimation errors, it follows that $(M_1, \ldots, M_K)$ is jointly Gaussian.

For jointly Gaussian random variables, zero covariance implies independence. Thus $M_1, \ldots, M_K$ are mutually independent. Since they also share a common marginal

distribution, they are i.i.d.

Finally, each M_k has strictly positive variance under Assumption 2. Hence, each marginal Gaussian distribution is nondegenerate and therefore continuous. □

OA.C. **Non-Monotonicity Example With Heterogeneous Instruments**

OA.C.1 Setup

Consider a setting with many weak instruments and a small fraction of strong instruments. Specifically, suppose that for $k = 1, \ldots, \bar{K}$,

$$\pi_k = \begin{cases} 0.05, & \text{with probability } 0.90, \\ 0.15, & \text{with probability } 0.10. \end{cases}$$

Let $q_k = 0.5$ for all k, so that $\text{Var}(z_k) = q_k(1 - q_k) = 0.25$, and assume instruments are mutually independent, so that $\text{Cov}(z_k, z_\ell) = 0$ for all $k \neq \ell$.

Consider the model

$$x_i = \sum_{k=1}^{\bar{K}} \pi_k z_{ki} + \rho u_i, \qquad u_i \sim N(0,1), \quad \rho = 0.5, \quad N = 5,000.$$

We rescale the model so that $\text{Var}(x_i) = 1$, which is without loss of generality since only relative signal strength matters for the behavior of the IV estimator.

Under this normalization, the true first-stage R^2 from a regression of x_i on instrument z_{ki} alone is

$$R_k^2 = \frac{\pi_k^2 \,\text{Var}(z_k)}{\text{Var}(x_i)} = \frac{\pi_k^2}{4},$$

and hence the corresponding true first-stage F-statistic is

$$F_k = N \frac{R_k^2}{1 - R_k^2} = N \frac{\pi_k^2/4}{1 - \pi_k^2/4}.$$

With $N = 5,000$, this implies that weak instruments with $\pi_k = 0.05$ correspond to $F_k = 3.13$, while strong instruments with $\pi_k = 0.15$ correspond to $F_k = 28.28$.

OA.C.2 Simulation

This section illustrates how the simulation works by expressing the model in terms of key variables: M_k and C_k.

For each candidate instrument k, define $M_k \equiv \frac{1}{N}\sum_{i=1}^{N} \widetilde{z}_{ki} u_i$, where $\widetilde{z}_{ki} = z_{ki} - \bar{z}_k$. This is the sample covariance between instrument k and the structural error. Conditional on the instrument vector z_k, and because we impose sample orthogonality across instruments,

$$M_k \mid z_k \sim N\left(0, \frac{1}{N^2}\sum_{i=1}^{N} \widetilde{z}_{ki}^2\right) = N\left(0, \frac{q_k(1-q_k)}{N}\right) = N\left(0, \frac{1}{4N}\right).$$

The first equality follows from the definition of M_k and the assumption that $u_i \sim N(0,1)$. The second uses $\frac{1}{N}\sum_{i=1}^{N} \widetilde{z}_{ki}^2 = q_k(1-q_k)$, and the third uses $q_k = 0.5$ for all k.

Next, define $C_k \equiv \pi_k \sigma_{z_k}^2 + \sum_{\ell \neq k} \pi_\ell \sigma_{z_\ell, z_k} = \frac{\pi_k}{4}$, where the last equality uses the orthogonality assumption. It follows that

$$C_k = \begin{cases} 0.0125, & \text{with probability } 0.90, \\ 0.0375, & \text{with probability } 0.10. \end{cases}$$

Note that all instruments have the same assignment probability $q_k = 0.5$ and that the distribution of M_k does not depend on the realized value of π_k. Hence M_k is independent of C_k. Moreover, by Lemma OA.B.3, the random variables $\{M_k\}_{k=1}^{K}$ are i.i.d. Gaussian with variance $1/(4N)$.

The just-identified 2SLS estimator using instrument k satisfies

$$\hat{\beta}_k - \beta = \frac{M_k}{C_k + \rho M_k}.$$

Finally, note that

$$\widehat{\mathrm{Cov}}(z_k, x) = C_k + \rho M_k,$$

and the first-stage F-statistic is a monotone function of $\widehat{\mathrm{Cov}}(z_k, x)^2$. Hence, selecting the instrument that maximizes F_k is equivalent to selecting the instrument that maximizes $|C_k + \rho M_k|$.

In accordance with Assumption 3, we retain only instruments satisfying $C_k + \rho M_k > 0$, which is equivalent to requiring that the estimated first-stage coefficient has the same sign as the true first-stage coefficient.

Thus, we can simulate first-stage hacking by drawing (C_k, M_k) for $k = 1, \ldots, K$ and selecting the instrument that maximizes F_k. The Figure in the main text presents

the results from this simulation.

OA.D. **Details on Simulating Instrument-Hacking**

This Appendix provides details on the simulations in Section 5. Restating the DGP:

$$y_i = \beta x_i + u_i, \qquad x_i = \sum_{k=1}^{\bar{K}} \pi_k z_{ki} + \rho u_i, \qquad i = 1, 2, ..., N \tag{12}$$

Parameter values are stated in Section 5. The main text describes the selection model we estimate to randomly draw true first-stage F-statistics, F_k. To simulate the DGP above, we need to convert these first-stage F-statistics to first-stage coefficients, π_k.

To do this, note that the population first-stage strength associated with using z_{ki} as the single instrument can be written as follows:

$$F_k = \frac{N\pi_k^2\sigma_z^2}{\sum_{\ell\neq k}^{\bar{K}} \pi_\ell^2\sigma_z^2 + \rho^2\sigma_u^2} = \frac{N\pi_k^2\sigma_z^2}{\sum_{\ell=1}^{\bar{K}} \pi_\ell^2\sigma_z^2 + \rho^2\sigma_u^2 - \pi_k^2\sigma_z^2} = \frac{N\pi_k^2\sigma_z^2}{1 - \pi_k^2\sigma_z^2} \tag{13}$$

where the last equality uses the normalization $\text{Var}(x_i) = \text{Var}\left(\sum_{k=1}^{\bar{K}} \pi_k z_{ki} + \rho u_i\right) = \sum_{k=1}^{\bar{K}} \pi_k^2\sigma_z^2 + \rho^2\sigma_u^2 = 1$. The previous display shows F_k is a strictly increasing function of π_k^2. Thus, under the normalization that the sign of the first-stage coefficient is positive, $\pi_k \geqslant 0$, each F_k corresponds to a unique value of π_k.

To ensure that the restriction holds, we adopt the following procedure:

1. Set $\bar{K}$ to be one plus the maximum number of instruments available to the researcher that we want to analyse, in this case 10, so we set $\bar{K} = 11$.

2. Draw F_k for $k = 1, 2, ..., \bar{K} - 1$ and convert each to π_k using (13).

3. Normalizing $\text{Var}(x_i) = 1$ implies $\sum_{k=1}^{\bar{K}} \pi_k^2 = \frac{1-\rho^2\sigma_u^2}{\sigma_z^2}$. Thus, we set $\pi_{\bar{K}}^2 = \left(\frac{1-\rho^2\sigma_u^2}{\sigma_z^2} - \sum_{k=1}^{\bar{K}-1} \pi_k^2\right)$ to ensure that $\text{Var}(x_i) = 1$.[4]

4. Simulate a replication of (12).

[4]The only potential problem with this procedure is if $\sum_{k=1}^{\bar{K}-1} \pi_k^2 > \frac{1-\rho^2\sigma_u^2}{\sigma_z^2}$. When that occurs, it is impossible to set $\pi_{k=\bar{K}}$ equal to any number to ensure that the variance of x_i is 1. Nevertheless, with the parameter values we set for our DGP this has not occurred. If it were to for a potential scenario, one would have to reduce $\bar{K}$ to be able to achieve a distribution of first-stage F-statistics that are sufficiently strong to match the empirical data.

5. Run $\bar{K} - 1$ just-identified IV regressions using each instrument, z_{ki}, and collect statistics (e.g. first-stage F-statistic, estimated treatment effect etc.).

6. Repeat steps 2-5 100,000 times.

Once the DGP has been simulated many times and statistics collected, we simulate first- and second-stage instrument hacking as described in main paper.

OA.E. **Results with Different Values for ρ**

This appendix extends the results of Section 5 of the main article, which used empirically derived variable instrument strengths to simulate instrument-hacking by a researcher. While empirical papers often report the strength of the instrument, they rarely report estimates of the strength of the endogeneity $\hat{\rho}$. Accordingly, the results in the main article assumed a moderate degree of endogeneity $\rho = 0.5$, and found severe bias under several strategies of selecting instruments. In this section, we consider the performance of the t-test test under two other cases: where endogeneity is either weak, with $\rho = 0.2$, or severe, with $\rho = 0.8$.

Table OA.E.1 outlines the size, proportion of rejections towards the biased OLS estimate, and median bias for second-stage instrument hacking under $\rho = 0.2$, 0.5, and 0.8. For $\rho = 0.2$, we see a fairly minor reduction in median bias and size across all cases where $K > 1$. Importantly, rejections are more balanced at roughly 75% in the same direction as the biased OLS estimate. This is because the standard error of the 2SLS regression is always minimized when $\hat{\beta}_{2SLS} = \hat{\beta}_{OLS}$, and when ρ is small $\hat{\beta}_{OLS}$ is closer to zero. When $\hat{\beta}_{OLS}$ is closer to zero, as it is here with $\beta = 0$ and $\rho = 0.2$, the distribution of standard errors become closer to symmetric around the true null. This leads to more balanced rejections, but even at empirically relevant levels of instrument strengths even $\rho = 0.2$ is not small enough to obtain close to balanced rejections. In summary, less meaningful endogeneity reduces size distortion and improves the balance of rejections. What is surprising is that the magnitude of median bias is only slightly smaller. Accordingly, we conclude that the impact of instrument-hacking on the integrity of the 2SLS estimator is quantitatively important even with mild endogeneity.

For $\rho = 0.8$, we observe larger size distortion under all K. This is expected as size increases with ρ even without instrument hacking. Even fewer rejections occur when

$\hat{\beta}_{2SLS} < 0$, which is due to the power asymmetry problem of IV estimators. Lastly, and most importantly, we see very similar median bias to the $\rho = 0.5$ case. The reason for this is that at $\rho = 0.5$ rejections are already close to 100% in the same direction as $\hat{\beta}_{OLS}$, which determines the severity of the added median bias that instrument-hacking contributes to the estimator. Increasing ρ further makes rejections slightly more unbalanced, and with slightly more median bias.

TABLE OA.E.1 – Simulation Results for Second-Stage Hacking by K and p

	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
K	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias
1	0.031	74%	0.001	0.043	93%	0.002	0.061	99%	0.002
2	0.062	74%	0.040	0.084	93%	0.062	0.119	98%	0.074
3	0.090	74%	0.086	0.123	93%	0.116	0.172	98%	0.128
5	0.145	74%	0.145	0.196	93%	0.177	0.269	98%	0.185
8	0.221	74%	0.192	0.293	94%	0.225	0.394	99%	0.230
10	0.268	74%	0.213	0.353	94%	0.246	0.466	99%	0.250

Notes: This table presents rejection rates and median bias from the DGP outlined in Section 5 of the main article for second-stage instrument hacking under different values for ρ. Pos Rej. refers to the proportion of rejections that feature a $\hat{\beta} > 0$.

Table OA.E.2 presents results for first-stage instrument hacking. Here, we see a clearer linear relationship for size and median bias with values of ρ. Reducing ρ to 0.2 more than halves median bias, while increasing ρ to 0.8 increases median bias by about 60%. Bias is only meaningful for moderate or severe degrees of endogeneity, in contrast to second-stage hacking.

OA.F. **Results for Very Strong Instruments**

This appendix considers a hypothetical distribution of instrument strengths characterized by very strong instruments. True first-stage F-statistics are drawn from $Gamma(10, 14)$, resulting in a median first-stage sample F of 137. 99% of draws are above a first-stage F of 50. Figure OA.F.1 shows the joint distribution of $\hat{\beta}$ and standard errors of 2SLS when instruments are drawn from this distribution under different selection strategies. When $K = 1$, we see that the correlation between the $\hat{\beta}_{2SLS}$ and $SE(\hat{\beta}_{2SLS})$ remains clear. Amazingly, the graph still illustrates severe distortions in the 2SLS estimator under second-stage hacking as K increases. Size distortion increases as is typical for any estimator under p-hacking, but the rejections are still

TABLE OA.E.2 – Simulation Results for First-Stage Hacking by K and p

	$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.8$		
K	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias
1	0.031	74%	0.001	0.043	93%	0.002	0.061	99%	0.002
2	0.040	73%	0.008	0.054	92%	0.020	0.078	98%	0.031
3	0.043	73%	0.010	0.059	92%	0.025	0.086	98%	0.040
5	0.046	72%	0.012	0.064	92%	0.030	0.097	98%	0.049
8	0.048	73%	0.013	0.069	93%	0.033	0.107	98%	0.054
10	0.049	73%	0.014	0.071	93%	0.035	0.112	98%	0.056

Notes: This table presents rejection rates and median bias from the DGP outlined in Section 5 of the main article for first-stage instrument hacking under different values for ρ. Pos Rej. refers to the proportion of rejections that feature a $\hat{\beta} > 0$.

highly unbalanced under the true null with approx 71% in the direction of the OLS bias. This is less severe than in the empirical distribution of instrument strengths, where it was approx 92%. Nevertheless, it remains meaningful.

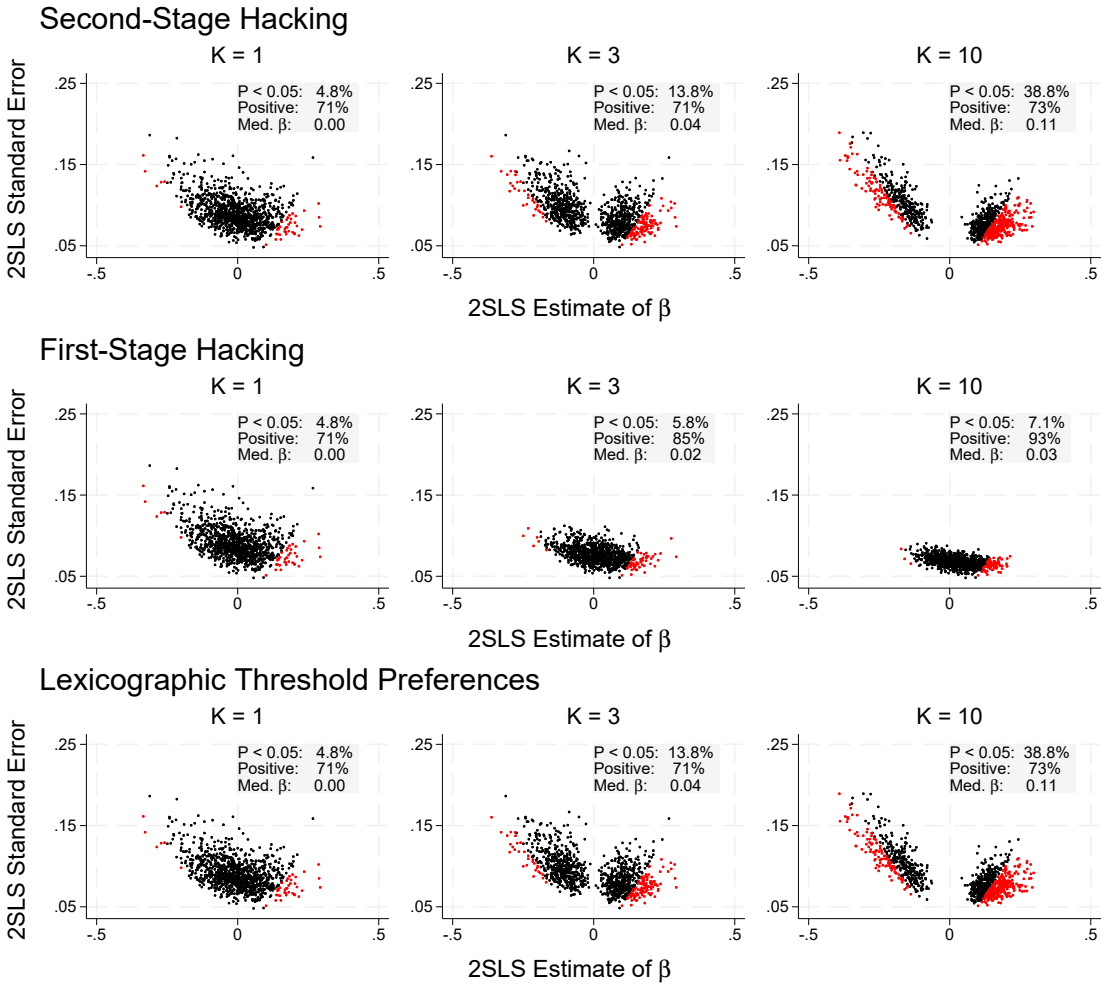
Meanwhile, median bias slightly more than halves than in the empirical distribution of instrument strengths. Again, the magnitude remains meaningful. The results for first-stage hacking with very strong instruments are very similar to the results in the main article. Size distortion and median bias are fairly minor. Rejections around the true null, however, are severely unbalanced and become worse as K increases. Accordingly, the main conclusions we draw in the main article are not specific to weak or moderately weak instruments, and persist even in an ideal case of many very strong instruments available to researchers.

OA.G. **Results for Instrument Strengths Drawn from JEL Codes**

To form the distribution of true F-statistics in Section 5 of the main article, we collected empirical data from IV papers in all JEL codes and used them to estimate a distribution of instrument strengths. Here, we consider the performance of 2SLS under first and second-stage hacking when the distribution of instrument strengths is estimated from empirical papers restricted to particular JEL codes. The goal of this exercise is to determine whether the performance of the 2SLS estimator under selection is sensitive to the particular type of economics papers we use to form the strength of the instrument set.

Table OA.G.1 present results for size, median bias, and the proportion of rejections

FIGURE OA.F.1. Instrument-Hacking with Instrument Strengths Drawn from Data (Very Strong Instruments): Joint Distribution of $\hat{\beta}$ and Standard Errors



Notes: 2SLS estimates and standard errors when researcher p-hack among just-identified specifications from 100,000 replications. In all cases, $\beta = 0$, $\rho = 0.5$, and π_k is drawn from a first-stage F distribution of $Gamma(10, 14)$ (for more details see text). Red dots indicate statistically significant estimates at 5% level. 'Conditional Second-Stage Hacking' refers to minimizing the second-stage p-value among all specifications with a first-stage F-statistic greater than 10 (otherwise maximizing the first-stage F). '$P \leqslant 0.05$' refers to the size (rejection rate) of the estimator, "Positive" refers to the proportion of rejections are 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.

occur in the positive direction (i.e. towards the biased OLS estimate) for second-stage hacking as K increases. Overall, we can see that the results change remarkably little. This is due to the fact that the JEL codes have similar empirical distributions of

first-stage Fs in the data. Table OA.G.2 present the results for first-stage instrument hacking. As before, the results change little across the JEL restrictions.

TABLE OA.G.1 – Simulation Results for Second-Stage Hacking Under Different JEL codes

K	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias
	JEL = All Papers			JEL = J			JEL = D		
1	0.043	93%	0.002	0.042	92%	0.002	0.043	93%	0.001
2	0.084	93%	0.062	0.082	93%	0.064	0.084	93%	0.060
3	0.123	93%	0.116	0.121	93%	0.117	0.123	93%	0.116
5	0.196	93%	0.177	0.195	93%	0.177	0.196	93%	0.177
8	0.293	94%	0.225	0.293	93%	0.225	0.294	94%	0.225
10	0.353	94%	0.246	0.351	93%	0.246	0.354	94%	0.246
	JEL = F			JEL = L			JEL = O		
1	0.042	93%	0.001	0.043	94%	0.002	0.042	94%	0.002
2	0.083	93%	0.057	0.083	94%	0.066	0.082	94%	0.062
3	0.123	93%	0.110	0.123	94%	0.124	0.121	94%	0.117
5	0.197	93%	0.173	0.194	94%	0.186	0.195	94%	0.180
8	0.296	93%	0.220	0.294	95%	0.236	0.293	94%	0.229
10	0.354	94%	0.240	0.353	95%	0.258	0.353	94%	0.250

Notes: This table presents rejection rates and median bias from the DGP outlined in Section 5 using 100,000 replications. Pos Rej. refers to the proportion of rejections that feature a $\hat{\beta} > 0$.

TABLE OA.G.2 – Simulation Results for First-Stage Hacking Under Different JEL codes

K	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias	Size	Pos. Rej.	Med. Bias
	JEL = All Papers			JEL = J			JEL = D		
1	0.043	93%	0.002	0.042	92%	0.002	0.043	93%	0.001
2	0.054	92%	0.020	0.052	92%	0.017	0.054	93%	0.021
3	0.059	92%	0.025	0.057	91%	0.023	0.060	94%	0.029
5	0.064	92%	0.030	0.062	91%	0.027	0.067	94%	0.035
8	0.069	93%	0.033	0.067	92%	0.030	0.072	94%	0.039
10	0.071	93%	0.035	0.067	92%	0.031	0.075	94%	0.041
	JEL = F			JEL = L			JEL = O		
1	0.042	93%	0.001	0.043	94%	0.002	0.042	94%	0.002
2	0.054	93%	0.020	0.055	94%	0.023	0.054	93%	0.021
3	0.060	93%	0.026	0.062	94%	0.032	0.060	94%	0.028
5	0.067	94%	0.033	0.068	94%	0.038	0.067	94%	0.034
8	0.070	94%	0.036	0.075	95%	0.043	0.073	95%	0.039
10	0.072	94%	0.037	0.077	95%	0.044	0.074	95%	0.041

Notes: This table presents rejection rates and median bias from the DGP outlined in Section 5 using 100,000 replications. Pos Rej. refers to the proportion of rejections that feature a $\hat{\beta} > 0$.