One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV*

Joshua Angrist Michal Kolesár

December 2022

Abstract

We revisit the finite-sample behavior of single-variable just-identified instrumental variables (just-ID IV) estimators, arguing that in most microeconometric applications, the usual inference strategies are likely reliable. Three widely-cited applications are used to explain why this is so. We then consider pretesting strategies of the form $t_1 > c$, where t_1 is the first-stage t-statistic, and the first-stage sign is given. Although pervasive in empirical practice, pretesting on the first-stage F-statistic exacerbates bias and distorts inference. We show, however, that median bias is both minimized and roughly halved by setting $c = 0$, that is by screening on the sign of the *estimated* first stage. This bias reduction is a free lunch: conventional confidence interval coverage is unchanged by screening on the estimated first-stage sign. To the extent that IV analysts sign-screen already, these results strengthen the case for a sanguine view of the finite-sample behavior of just-ID IV.

*We thank Ahmet Gulek and Luther Yap for expert research assistance. Thanks also go to Tim Armstrong, Isaiah Andrews, Brigham Frandsen, Guido Imbens, Mike Keane, Dave Lee, Whitney Newey, and Steve Pischke for helpful discussions and insightful comments. Kolesár acknowledges support by the Sloan Research Fellowship and by the National Science Foundation Grant SES-22049356. The views expressed here are our own.

1 Introduction

The heavily over-identified two-stage least squares (2SLS) estimates reported in Angrist (1990) and Angrist and Krueger (1991, AK91) sparked a flood of interest in the finite-sample behavior of instrumental variables (IV) estimators. Kicked off by Bekker (1994) and Bound et al. (1995), attention to bias in 2SLS estimates with many weak instruments has since become a staple of applied microeconometrics. The fact that the finite-sample distribution of 2SLS estimates is shifted towards the corresponding ordinary least squares (OLS) probability limit is especially worrying. IV is often motivated by the belief that OLS estimates are compromised by omitted variable bias (OVB). The IV analyst hopes, therefore, that when IV and OLS are close, this signals minimal OVB rather than substantial finite-sample bias in the IV estimates.

2SLS combines multiple instruments in an effort to estimate a single causal effect with acceptable precision. Strikingly, however, Bound et al. (1995) show that in the AK91 specifications interacting quarter of birth dummies with covariates to generate 180 instruments, replacing real quarter of birth dummies with dummies randomly drawn yields 2SLS estimates and standard errors much like those generated by the real thing.[1] But most IV studies featuring a single endogenous variable build on a single underlying instrument, such as a dummy for draft-eligibility in Angrist (1990) and quarter of birth in AK91. Just-identified IV with a single endogenous variable (just-ID IV) offers a simple, transparent estimation strategy in such cases. Our analysis comes in the wake of renewed interest in the finite-sample properties of just-ID IV, an interest reflected in Andrews and Armstrong (2017), Lee et al. (2022), and Keane and Neal (2022), among others.

We argue here that in typical microeconometric applications of just-ID IV, conventional IV estimates and t-tests are compromised little by failures of the usual asymptotic theory. Our analysis builds on the (approximate) finite-sample normality of reduced-form and first-stage estimators (in the argot of classical simultaneous equations models, these are both estimated "reduced forms"). This modeling framework parallels that used in Andrews et al. (2019) and earlier theoretical investigations of weak instrument problems. The normality of reduced-form estimates is justified by conventional asymptotic reasoning, as well as by the local-to-zero asymptotic sequence used in Staiger and Stock (1997) and Stock and Yogo (2005), in which the first stage shrinks to zero at a rate inversely proportional to the square root of the sample size.

Our setup has two free parameters: the correlation between structural and first-stage residuals (henceforth, "endogeneity"), and the population first-stage F-statistic. This fact lends itself to the construction of easily-interpreted rejection contours characterizing conventional second-stage t-tests and confidence interval coverage rates. We see, for example, that for endogeneity less than about 0.76, 95% confidence interval coverage is distorted by no more than 5% for *any* population F. This is explained by the fact that, even as median bias increases when the first stage gets weaker, second-stage precision falls (we focus on median bias because the conventional just-ID IV estimator has no moments). In contrast with the over-identified case, conventional just-ID IV standard errors reflect this,

[1] This "fake instruments" simulation was originally suggested by Alan Krueger. Although not an empirical study, Bekker (1994) is likewise motivated by a heavily over-identified specification in Angrist (1990) that uses 73 draft lottery dummies plus interaction terms as instruments for Vietnam-era veteran status. This application is featured at the end of Bekker's paper, and, originally, in an Amsterdam bar in 1992, where Paul Bekker first confronted Angrist with claims of finite-sample bias.

and confidence intervals widen accordingly. This fact keeps interval coverage high unless endogeneity is extraordinarily high.

What range of values for endogeneity is relevant? Three applications are used to calibrate it: AK91, the Angrist and Evans (1998, AE98) IV estimates using a dummy for samesex sibships as an instrument for family size, and the Angrist and Lavy (1999, AL99) fuzzy regression discontinuity estimates of class size effects on student learning. These studies span a range of OVB scenarios, from modest (for most of the AK91 estimates), to substantial (in AE98, where OLS exceeds IV by about 50%), to dramatic (in AL99, where IV exceeds small, insignificant OLS estimates by an order of magnitude). Yet, the absolute value of estimated endogeneity is no more than 0.47 in these applications, and over 0.4 only for a single specification and sample. Although three examples do not make a theorem, we argue that the features of these studies that limit endogeneity are common to empirical strategies designed to estimate causal effects or to mitigate attenuation bias in models with measurement error.[2]

Evidence on the reliability of conventional just-ID IV inference notwithstanding, IV practitioners have come to see weak instruments as problematic in just-identified as well as over-identified models. Responding to Bound et al. (1995) and Staiger and Stock (1997), analysts now routinely report first-stage t- and F-statistics, a practice that's hard to argue with. Yet, Hall et al. (1996) and others since note that requiring, say, a first-stage F-statistic greater than 10 when instruments are truly weak often does more harm than good. Requiring first-stage estimates to meet a prespecified cutoff amounts to imposition of a pretest that distorts sampling distributions and makes conventional confidence intervals misleading. This is the IV version of the general pretesting problem highlighted by Leeb and Pötscher (2005).[3]

Analyses of pretesting to date apply when the analyst is agnostic as to the sign of the first stage. Andrews and Armstrong (2017) observe, however, that the typical just-ID IV scenario includes a theoretical sign restriction. This leads us to consider pretesting strategies that maintain a sign restriction on the population first stage. Specifically, we examine pretesting rules of the form $t_1 > c$, where t_1 is the first-stage t-statistic and c is a constant chosen by the analyst. This examination leads to a novel theoretical result: the median bias of just-ID IV conditional on $t_1 > c$ is minimized by choosing $c = 0$, that is, by screening on the sign of the estimated first-stage. Moreover, median bias of just-ID IV is roughly halved by requiring the estimated first-stage to have the expected (that is, population) sign.

Surprisingly, pre-screening on the estimated first-stage sign is also shown to be virtually costless: rejection contours for a sign-screened estimator differ little from those obtained without screening. The upshot is that sign-screening mitigates the already-modest median bias of just-ID IV without degrading coverage. To the extent that such screening is a feature of modern empirical work, reported IV estimates reflect the impressively minimal bias characteristic of sign-screened IV. Our theoretical results on the bias-minimizing and bias-mitigating consequences of requiring $t_1 > 0$ therefore strengthen the case for a sanguine view of conventional inference for just-ID IV.

Finally, the theorem establishing bias mitigation from screening on an estimated first stage sign provides an interesting contrast to Andrews and Armstrong (2017), which shows how to use a sign

[2]The spirit of this argument differs from that in Stock and Yogo (2005), which focuses on worst-case rejection rates over all possible endogeneity values. Lee et al. (2022) and Keane and Neal (2022), discussed further in Section 3 below, consider though largely downplay restrictions on endogeneity.

[3]Andrews et al. (2019) survey recent empirical scholarship to demonstrate the empirical relevance of pretesting based on first-sage F in IV applications.

restriction on the *population* first stage to construct a mean unbiased just-ID IV estimator. We show that conditional on the sign of the *estimated* first stage, this estimator, denoted $\hat{\beta}_U$, is no longer unbiased. Rather, $\hat{\beta}_U$ is unbiased by virtue of the fact that it averages two conditional estimators, each biased but in opposite directions. If, as seems likely, empirical sign-screening is endemic in applied microeconomics, this would seem to reduce the appeal of $\hat{\beta}_U$ for empirical practice.

The next section details the just-ID IV setup assuming normally distributed first-stage and reduced-form estimates and derives an expression for endogeneity in terms of OLS OVB. Section 3 reviews the relationship between t-test rejection rates and the parameters that govern the normal model. This section also explains why endogeneity in applied microeconometrics is unlikely to be high enough for conventional IV inference to mislead, and quantifies the length advantage of conventional confidence intervals relative to Anderson-Rubin-based intervals. Section 4 presents our theoretical results on first-stage screening. Section 5 concludes with a discussion of the implications of our results. Proofs and details behind numerical calculations appear in the appendix.

2 Setup

The sample is assumed to consist of n units indexed by i, with data on outcome variable, Y_i, a scalar treatment variable, D_i, a vector of covariates, X_i, and a scalar instrument, Z_i. Population regressions of outcome and treatment on the instrument and covariates define the reduced form and first stage. These are written as follows:

$$Y_i = Z_i\delta + X_i'\psi_1 + u_i, \tag{1}$$

$$D_i = Z_i\pi + X_i'\psi_2 + v_i. \tag{2}$$

The parameter of interest is $\beta = \frac{\delta}{\pi}$, the ratio of the reduced-form and first-stage regression coefficients on Z_i. Provided that the instrument, Z_i, satisfies an exclusion restriction and is relevant (i.e. $\pi \neq 0$), this parameter captures the causal effect of D_i on Y_i. More generally, if treatment effects are heterogeneous and a monotonicity condition holds, β is a weighted average of individual causal effects (Angrist & Imbens, 1995; Imbens & Angrist, 1994). While treatment effect heterogeneity affects the interpretation of β, heterogeneity has no bearing on the behavior of the estimators and inference procedures considered in our just-ID IV setting.

Let $\hat{\delta} = \sum_{i=1}^n \tilde{Z}_i Y_i / \sum_{i=1}^n \tilde{Z}_i^2$ and $\hat{\pi} = \sum_{i=1}^n \tilde{Z}_i D_i / \sum_{i=1}^n \tilde{Z}_i^2$ denote OLS estimates of δ and π, where $\tilde{Z}_i$ is the residual from a regression of Z_i on X_i. Under mild regularity conditions that allow the errors (u_i, v_i) to be non-normal, heteroskedastic, and serially or cluster-dependent, $(\hat{\delta}, \hat{\pi})$ is consistent and asymptotically normal as $n \to \infty$, with an asymptotic covariance matrix that can be consistently estimated. Importantly, this holds regardless of the strength of the instrument. We therefore follow Andrews et al. (2019) and earlier analyses of weak instrument problems by assuming this large-sample approximation holds exactly. Specifically, we assume:

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \pi\beta \\ \pi \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{\hat{\delta}}^2 & \sigma_{\hat{\delta}\hat{\pi}} \\ \sigma_{\hat{\delta}\hat{\pi}} & \sigma_{\hat{\pi}}^2 \end{pmatrix} \right), \tag{3}$$

with a known covariance matrix, Σ. This distributional assumption is implied by the Staiger and Stock (1997) weak-instrument asymptotic sequence (see Andrews et al. (2019, Section 3.2) for additional discussion and references). Finite-sample results under eq. (3) can therefore be seen as asymptotic under the Staiger and Stock (1997) sequence.

Equation (3) is our only substantive restriction; this assumption allows us to focus on the weak instrument problem, separating this from other finite-same problems, such as the effect of high-leverage observations on the quality of the normal approximation to the distribution of the OLS estimators $(\hat{\delta}, \hat{\pi})$ and the challenge of standard-error estimation with clustered data.[4] With (3) as foundation, we derive finite-sample properties of the IV estimator:

$$\hat{\beta}_{IV} = \frac{\hat{\delta}}{\hat{\pi}}, \tag{4}$$

and the null rejection rate for the corresponding Wald test. The latter is based on the t-statistic centered at the parameter of interest, β, divided by the estimated IV standard error, $\hat{\sigma}_{IV}$:

$$t_W = \frac{\hat{\beta}_{IV} - \beta}{\hat{\sigma}_{IV}}; \qquad \hat{\sigma}^2_{IV} = \frac{\sigma^2_{\hat{\delta}} - 2\sigma_{\hat{\delta}\hat{\pi}}\hat{\beta}_{IV} + \sigma^2_{\hat{\pi}}\hat{\beta}^2_{IV}}{\hat{\pi}^2}, \tag{5}$$

where $\hat{\sigma}^2_{IV}$ estimates the asymptotic variance of $\hat{\beta}_{IV}$ under standard $n \to \infty$ asymptotics. The corresponding theoretical variance is $\sigma^2_{IV} = (\sigma^2_{\hat{\delta}} - 2\sigma_{\hat{\delta}\hat{\pi}}\beta + \sigma^2_{\hat{\pi}}\beta^2)/\pi^2$. In a homoskedastic model with constant causal effects, this simplifies to the familiar formula

$$\sigma^2_{IV} = \frac{\sigma^2_{\epsilon}}{nE[\ddot{Z}^2_i]\pi^2},$$

where $\ddot{Z}_i$ is the residual from the population projection of Z_i on X_i, and σ^2_ϵ is the variance of the residual in the structural equation,

$$Y_i = D_i\beta + X'_i(\psi_1 - \psi_2\beta) + \epsilon_i, \tag{6}$$

that motivates IV estimation in the classic linear set-up (the structural residual is $\epsilon_i = u_i - v_i\beta$).

Given the assumption of a known covariance matrix for the first-stage and reduced-form estimates, both t_W and $\hat{\beta}_{IV}$ depend on the data only through $(\hat{\delta}, \hat{\pi})$. These have distributions determined by the two unknown parameters, π and β. It's illuminating, however, to reparametrize in terms of instrument strength and the degree of endogeneity (a reparametrization adopted in Staiger and Stock (1997) and Lee et al. (2022), among others). The first parameter in this scheme, denoted $E[F]$, is defined as:

$$E[F] = \pi^2/\sigma^2_{\hat{\pi}} + 1.$$

Because $E[F]$ is the expectation of $F = \hat{\pi}^2/\sigma^2_{\hat{\pi}}$, the F-statistic testing $\pi = 0$, it's sometimes called the *population first-stage F-statistic*, a term adopted here. Since π is a scalar, $E[F] = E[t_1]^2 + 1$, where $t_1 = \hat{\pi}/\sigma_{\hat{\pi}}$ is the first-stage t-statistic.

[4] Young (2022) discusses these problems in an IV context.

The second parameter is defined as:

$$\rho = \text{cor}(\hat{\delta} - \hat{\pi}\beta, \hat{\pi}) = \frac{\sigma_{\hat{\pi}}}{\sqrt{\sigma^2_{\hat{\delta}} - 2\beta\sigma_{\hat{\delta}\hat{\pi}} + \sigma^2_{\hat{\pi}}\beta^2}} \times (\sigma_{\hat{\delta}\hat{\pi}}/\sigma^2_{\hat{\pi}} - \beta). \tag{7}$$

With independent heteroskedastic errors, ρ is also given by $\text{cor}(\ddot{Z}_i\epsilon_i, \ddot{Z}_i v_i)$. When, in addition, the errors (u_i, v_i) are homoskedastic, $\rho = \text{cor}(\epsilon_i, v_i)$, where ϵ_i is the structural residual in (6). We therefore refer to ρ as (the degree of) *endogeneity.*[5]

With weak instruments as well as homoskedastic error terms, ρ is proportional to the bias of the OLS estimand. This can be seen by using the first-stage and reduced-form equations to write the OLS slope coefficient, β_{OLS}, as follows:

$$\begin{aligned}\beta_{OLS} &= \frac{E[\ddot{D}_i Y_i]}{E[\ddot{D}_i^2]} = \frac{E[\ddot{Z}_i^2]\pi^2\beta + E[u_i v_i]}{E[\ddot{Z}_i^2]\pi^2 + \sigma_v^2} \\ &= R_p^2\beta + (1 - R_p^2)\frac{E[u_i v_i]}{\sigma_v^2}; \qquad R_p^2 = \frac{E[\ddot{Z}_i^2]\pi^2}{E[\ddot{Z}_i^2]\pi^2 + \sigma_v^2},\end{aligned} \tag{8}$$

where $\ddot{D}_i$ is the residual from a population regression of D_i on X_i, and $\sigma_v^2 = E[v_i^2]$. The weight multiplying β in (8), denoted R_p^2, is the population partial R^2 generated by adding the instrument to the first-stage regression. When the instrument is weak, R_p^2 is close to zero, and (8) is approximately $E[u_i v_i]/\sigma_v^2$. The OLS estimand likewise converges to $E[u_i v_i]/\sigma_v^2$ in the Staiger and Stock (1997) weak-instrument sequence (which takes $\pi \to 0$). This in turn equals $\sigma_{\hat{\delta}\hat{\pi}}/\sigma^2_{\hat{\pi}}$ under homoskedasticity, so the second term on the right-hand side of (7),

$$\sigma_{\hat{\delta}\hat{\pi}}/\sigma^2_{\hat{\pi}} - \beta = \beta_{WOLS} - \beta, \tag{9}$$

is the weak-instrument OVB of OLS (where we've introduced the notation β_{WOLS} for $\sigma_{\hat{\delta}\hat{\pi}}/\sigma^2_{\hat{\pi}}$). Moreover, when $\pi = 0$, it follows from (3) that $\beta_{WOLS} - \beta$ is the median bias of $\hat{\beta}_{IV}$, a result requiring no independence or heteroskedasticity assumptions on the errors in (1) and (2).[6] Thus, ρ also measures endogeneity in the sense that it's proportional to the median bias of the IV estimator when the instrument is irrelevant.

3 Rejection Rates in Theory and Practice

We start by considering t-test rejection rates when the null hypothesis is true. For a two-sided t-test with level α, the rejection rate is the probability that the absolute value of a t-statistic, $|t_W|$, exceeds $z_{1-\alpha/2}$, the $1 - \alpha/2$ quantile of a standard normal distribution. The rejection rate of interest can

[5]This simplification is obtained using the fact that, under homoskedasticity, the variance of v_i is $\sigma_v^2 = \sigma^2_{\hat{\pi}} \cdot nE[\ddot{Z}_i^2]$ and the variance of ϵ_i is $\sigma^2_\epsilon = (\sigma^2_{\hat{\delta}} - 2\beta\sigma_{\hat{\delta}\hat{\pi}} + \sigma^2_{\hat{\pi}}\beta^2) \cdot nE[\ddot{Z}_i^2]$, with $\text{cov}(v_i, \epsilon_i) = (\sigma_{\hat{\delta}\hat{\pi}} - \beta\sigma^2_{\hat{\pi}}) \cdot nE[\ddot{Z}_i^2]$. The homoskedastic formula for the variance of ϵ_i also leads yields the simplification of the formula for σ^2_{IV} noted above.

[6]Assumption (3) implies that we can write reduced form and first stage estimates as $\hat{\delta} = \pi\beta + (\sigma_{\hat{\delta}\hat{\pi}}/\sigma_{\hat{\pi}})\mathcal{Z}_\pi + (\sigma^2_{\hat{\delta}} - \sigma^2_{\hat{\delta}\hat{\pi}}/\sigma^2_{\hat{\pi}})^{1/2}\mathcal{Z}_\delta$ and $\hat{\pi} = \pi + \sigma_{\hat{\pi}}\mathcal{Z}_\pi$, where $\mathcal{Z}_\delta$ and $\mathcal{Z}_\pi$ are independent standard normal variables. When $\pi = 0$, therefore, $\hat{\beta}_{IV} = \frac{1}{\sigma_{\hat{\pi}}}(\sigma^2_{\hat{\delta}} - \sigma^2_{\hat{\delta}\hat{\pi}}/\sigma_{\hat{\pi}})^{1/2}(\mathcal{Z}_\delta/\mathcal{Z}_\pi) + \beta_{WOLS}$, the median of which is β_{WOLS} since $\mathcal{Z}_\delta/\mathcal{Z}_\pi$ has a standard Cauchy distribution with zero median.

therefore be written:

$$R_W = P_{E[F],\rho}(|t_W| > z_{1-\alpha/2}),$$

where $P_{E[F],\rho}$ is the distribution of t_W parameterized by $E[F], \rho$. We evaluate R_W by numerical integration, a computation detailed in Appendix A.3.

Summarizing the behavior of a conventional 5% nominal test, Panel (a) in Figure 1 depicts rejection rates for t_W as a contour plot given ρ and $E[F]$. The figure shows that rejection rates substantially exceed the nominal 5% level only if the instrument is weak (i.e., $E[F]$ is close to 1) *and* endogeneity is high. Stock and Yogo (2005, Section 3.2) define instruments as weak if the usual 5% level t-test rejects a true null more than 10% of the time. The figure shows that, as long as $|\rho| < 0.76$, rejection rates stay below 10% regardless of the strength of the first stage. A stricter standard based on any over-rejection is met as long as $|\rho| < 0.565$ (this cutoff is also noted in Lee et al., 2022). A simple corollary to these observations, substantiated below by showing that worryingly high values of ρ should be seen as unusual, is that the coverage of conventional nominal 95% confidence intervals for $\hat{\beta}_{IV}$ is likely to be satisfactory in most applications.

The modest over-rejection seen in Figure 1 is explained by a signal feature of just-ID IV: the median bias of $\hat{\beta}_{IV}$ rises as the instrument grows weaker, but precision falls apace. The IV standard error reflects this lack of precision well enough that, unless endogeneity is egregious, inference is distorted little. This contrasts with over-identified 2SLS with many weak instruments (as in Bekker (1994) and Bound et al. (1995)), where, bias notwithstanding, the usual standard errors for 2SLS remain small enough for the t-statistic to be misleading.

Our conclusions here also contrast with those drawn in Stock and Yogo (2005) and Lee et al. (2022) regarding the reliability of inference based on a conventional just-ID IV t-statistic. Although Lee et al. (2022) present a similar plot, both studies emphasize worst-case rejection rates over ρ, for a given $E[F]$. As can be seen in our Figure 1, this worst-case rejection rate occurs at $|\rho| = 1$. In the same spirit, Keane and Neal (2022) highlights simulations showing that conventional just-ID IV t-tests can be misleading when endogeneity is very high. Sections 3.1 and 3.2 explain why we are not much concerned with high values of ρ.

Keane and Neal (2022) also observe that, since $\hat{\sigma}_{IV}$ and $\hat{\beta}_{IV}$ tend to be negatively correlated when ρ is positive, most false rejections occur when $\hat{\beta}_{IV} > \beta$. This, they argue, militates so strongly against t_W that conventional Wald tests are to be avoided even with a first-stage F in the thousands. As we see it, this observation, which is related to asymmetry in the power function of t_W, does not make conventional frequentist inference unreliable. The conventional standard for reliability of inference is the accuracy of confidence interval coverage, gauged without conditioning on parameter estimates. Our analysis adheres to this standard.[7]

[7] The Keane and Neal (2022) critique rules out many standard procedures. For instance, consider the conventional $1-\alpha$ confidence interval $[U_{(n)}, U_{(n)}/\alpha^{1/n}]$ for the endpoint of a uniform distribution supported on $[0, \theta]$, where $U_{(n)}$ denotes the largest order statistic in a sample of size n. Here all false rejections occur when the minimum variance unbiased estimate $U_{(n)}/(n+1)$ is below θ (provided $\alpha < 1/\exp(1)$).

3.1 The Anatomy of Endogeneity

We put endogeneity in context using three IV applications. These are the AK91 study that launched the modern weak instruments literature, the AE98 study using a dummy for samesex sibships (of first- and second-born children) as an instrument for family size, and the AL99 fuzzy regression discontinuity estimates of class size effects. The AE98 and AL99 first-stage t-statistics exceed those for AK91 and are arguably out of the zone where an instrument might be considered weak. With a first-stage t-statistic of almost 8, the AK91 quarter-of-birth instrument also seems strong enough. But all three studies can be used to calibrate endogeneity and to document contextual features that constrain it.

Table 1 reports key statistics for specifications drawn from each study (some estimates in the table differ slightly from those in the original). The first row in Panel A shows estimates of the economic returns to schooling in the AK91 sample of men born 1920–29. Here, OLS and IV estimates equal 0.080 and 0.072, respectively. These are close, so endogeneity is small in this case, with an estimated ρ of only 0.043. Schooling returns estimated in the second AK91 sample, consisting of men born 1930–39, exhibit more OVB. In this sample, the IV estimate of 0.105 surprisingly exceeds the OLS estimate of 0.071 (IV estimation of the returns to schooling is usually motivated by a concern that omitted ability controls causes OLS estimates to be too large). Endogeneity is correspondingly larger at $\rho = -0.175$, but still well outside the danger zone depicted in Figure 1.[8]

The AK91, AE98, and AL99 studies span a range of OVB scenarios, from modest in the first AK91 sample, to substantial in AE98 (where OLS magnitudes consistently exceed IV by at least 50%), to dramatic in AL99 (where IV exceeds small, insignificant OLS estimates, mostly by an order of magnitude, and sometimes with a sign flip). Yet, the magnitude of endogeneity exceeds 0.40 in only one specification, that for reading scores in the AL99 discontinuity sample (which consists of classes in schools with enrollment near the cutoff that determines class size). Just-ID IV inference in all three of these studies is therefore unlikely to be compromised by weak instruments.

Although the consistently moderate levels of endogeneity documented in Table 1 do not add up to a theorem, these applications have features in common with many IV-driven microeconometric investigations of causal effects. Specifically, endogeneity in research on causal effects is often capped by the modest size of the causal effects of interest. To make this point, it's helpful to write ρ as a function of OVB. Using eqs. (8) and (9), we can express ρ under homoskedasticity as:

$$\begin{aligned} \rho &= \frac{\sigma_v}{\sigma_\epsilon}(\beta_{WOLS} - \beta) \\ &= \frac{\sigma_v}{\sigma_\epsilon}\left(\frac{\beta_{OLS} - \beta}{1 - R_p^2}\right) \approx \frac{\sigma_D}{\sigma_Y}(\beta_{OLS} - \beta), \end{aligned} \tag{10}$$

where the approximation $\frac{\sigma_v}{\sigma_\epsilon(1-R_p^2)} \approx \frac{\sigma_D}{\sigma_Y}$ holds if the explanatory power of the observables in both the structural and the first-stage equation is low. We can use this expression to compute ρ by replacing β with $\hat{\beta}_{IV}$. The relevance of this representation of ρ can be seen in the AE98 estimates of the effects of a third child on weeks worked by women aged 21–35 in the 1980 Census. Here, the difference between

[8] Endogeneity confidence intervals are computed by inverting the Anderson-Rubin test, and are therefore robust to weak instruments. See Appendix A.1 for details. In the examples analyzed here, the instruments are not particularly weak, so the bias in estimated endogeneity is negligible and conventional delta-method intervals are similar to those reported in the table.

OLS and the corresponding IV estimate is -3.42. Because the first-stage partial R-square (R_p^2) is close to zero, the term multiplying this, $\frac{\sigma_v}{\sigma_\epsilon}$, is well-approximated by the ratio of the endogenous variable standard deviation to the dependent variable standard deviation, $\frac{\sigma_D}{\sigma_Y}$, a ratio of about 0.022. The product of these two terms gives -0.075, equal to the value of ρ reported in the table for this sample.

Equation (10) suggests a bound on endogeneity motivated by plausible limits to effect size and OVB. In the AK91 scenario, for instance, it seems reasonable to assume that the (causal) economic returns to schooling are no more than double the OLS estimand. Under these restrictions, the descriptive statistics in Table 1, which approximate $\frac{\sigma_v}{\sigma_\epsilon}$ at around 5.2 in this case, suggest $|\rho|$ can be no more than about 0.41. Although substantial, this is still below the 0.565 and 0.76 values beyond which coverage deteriorates. With β bounded below by zero, large magnitudes of ρ require β to far exceed β_{WOLS}. Only when the causal effect of schooling is triple the OLS estimand (so that OLS is *too small* by 0.16) does the endogeneity danger zone become relevant.[9]

Many microeconometric IV applications involve linear probability models in which causal effects are changes in probabilities. This also has implications for endogeneity. The AE98 estimates of the effect of the birth of a third child on female labor force participation in 1980, for example, range from roughly -0.18 for OLS to -0.12 for IV. Labor force participation rates for women with only two children run around 57%. Causal effects might therefore be as large as -0.57, but no larger, since probabilities can't be negative. In this case, $\frac{\sigma_v}{\sigma_\epsilon}$ is about 1 (again, using standard deviations in the data rather than residuals), so $\beta_{OLS} - \beta$ can be no larger than $-0.18 + 0.57 = 0.39$, thereby bounding ρ at this value. This generous bound makes no use of the fact that selection bias is likely to make OLS estimates of family-size effects on female supply too large (in magnitude) rather than too small. Other applications with Bernoulli outcomes admit similar sorts of bounds.

A related argument, appropriate for models with continuous outcomes, shows endogeneity to be constrained by plausible values for causal effects measured in standard deviation units. This line of reasoning is especially apt for education research where standardized effect sizes are widely reported. The influential Tennessee STAR class size experiment analyzed in Krueger (1999), for instance, generated a reduction of 7 students per class, roughly one standard deviation of class size in the AL99 data. The STAR experiment yielded treatment effects of about 0.2σ, an impact typical of education interventions deemed to have been effective. At the same time, education researchers often view effect sizes as large as half a standard deviation in the outcome distribution as rare, if not implausible. Using the fact that $\frac{\sigma_v}{\sigma_\epsilon}$ is about equal to $(1 - R_p^2)$ in the AL99 data, the scenario of a half-standard deviation effect size generated by a one-standard deviation reduction in class size implies $\frac{\sigma_v}{\sigma_\epsilon}\frac{\beta}{1-R_p^2} = -0.5$ on the second line of eq. (10). At the same time, OLS estimates of class size effects in AL99 are mostly zero (as is often found in class size research; see e g., Hanushek (1986)), so the magnitude of endogeneity is capped at 0.51.

Contributing to all three of these empirically-grounded arguments is the fact that endogeneity under homoskedasticity can be split into the difference between two R-squared-like terms:

$$\rho \approx \frac{\sigma_D}{\sigma_Y}(\beta_{OLS} - \beta) = \frac{\sigma_D}{\sigma_Y}\beta_{OLS} - \frac{\sigma_D}{\sigma_Y}\beta.$$

[9] Keane and Neal (2022) consider bounds on ρ motivated by the view that OLS estimates of schooling returns should exceed causal effects. Although this seems defensible, it's worth noting that the literature surveyed by Card (2001) reports many IV estimates in excess of the corresponding OLS estimates, a pattern first highlighted by Lang (1993).

The square of the first term, $(\frac{\sigma_D}{\sigma_Y}\beta_{OLS})^2$, is the variation in the dependent variable accounted for by D_i in an analysis-of-variance for Y_i. In microeconometric applications, this sort of R^2 term is mostly small, as is the causal analog that determines the square of the second term, $(\frac{\sigma_D}{\sigma_Y}\beta)^2$. The small size of these two R^2 terms limits the magnitude of the difference between them. Consistent with this claim to generality, the many IV estimates collated in Chernozhukov and Hansen (2008) likewise show modest endogeneity.

It's noteworthy that the bound of 0.41 derived for the AK91 study depends only on β_{OLS}, standard deviations σ_Y and σ_D, and bounds on the causal effect of interest. The resulting calculations therefore seem likely to be relevant for other empirical strategies estimating returns to schooling. As far as details go, our back-of-the-envelope bounds leverage homoskedasticity and the presumption that observables have little explanatory power. In applications where these restrictions are a stretch, eq. (7) gives a basis for bounds that apply more generally. Specifically, because ρ is monotone decreasing in β, plugging bounds on β into eq. (7) along with estimates of components of the covariance matrix, Σ, bounds ρ. Estimates of reduced-form and first-stage standard errors are typically readily available, while estimates of $\sigma_{\hat{\delta}\hat{\pi}}$ can be obtained from equation eq. (A.1) in Appendix A. For the AK91 study, the more general bound computed using covariance matrix estimates reported in Table 1 leads to a bound on ρ that matches the rough cut. Assuming (as above) that effects of a third child on labor force participation lie between 0 and -0.57, and class size effects lie between 0 and $\frac{\sigma_Y}{2\sigma_D}$, yields upper bounds on $|\rho|$ equal to 0.36, and 0.57, respectively. These numbers are likewise close to the corresponding rough-cut bounds of 0.39 and 0.51, respectively.

3.2 When Measurement Error Motivates IV

In addition to estimating causal effects, a second major arena for microeconometric IV involves models with measurement error. Suppose the regression of interest is $Y_i = D_i^*\beta + X_i'\gamma + \eta_i$, where η_i is a residual uncorrelated with (D_i^*, X_i) by definition. The regressor D_i^* is unobserved; we see only a noisy measure, $D_i = D_i^* + e_i$, where the measurement error, e_i, is assumed to be classical, that is uncorrelated with (D_i^*, X_i, η_i). Replacing D_i^* with D_i yields the structural equation to be instrumented:

$$\begin{aligned} Y_i &= D_i\beta + X_i'\gamma + (\eta_i - e_i\beta) \\ &= D_i\beta + X_i'\gamma + \epsilon_i, \end{aligned}$$

where $\epsilon_i = \eta_i - e_i\beta$ is the structural residual. Given an instrument correlated with D_i^* and uncorrelated with ϵ_i, the coefficients of interest are consistently estimated by IV. The first stage in this scenario can be written as in (2), with first-stage residual, v_i.

To calibrate endogeneity in this model, note first that, given the classical measurement error assumption, $\mathrm{cov}(v_i, \epsilon_i) = -\sigma_e^2\beta$. Under homoskedasticity, endogeneity squared can therefore be written:

$$\rho^2 = \frac{\sigma_e^4\beta^2}{\sigma_v^2\sigma_\epsilon^2} = \frac{\sigma_e^4\beta^2}{\sigma_v^2(\sigma_\eta^2 + \beta^2\sigma_e^2)} \leq \frac{\sigma_e^2}{\sigma_v^2} = \frac{1-r}{1-R_p^2}, \tag{11}$$

where $r = \sigma^2_{\tilde{D}^*}/\sigma^2_{\tilde{D}}$ denotes the reliability (or signal-to-noise ratio) of mismeasured D_i, after partialing out covariates.[10] Although we can't speak to reliability across all fields, labor economists have collected evidence on the reliability of key variables of interest. These include schooling, earnings, hours worked, and hourly wages. Schooling often appears on the right-hand side of wage equations, while earnings, hours, and hourly wages are used in various configurations to estimate labor supply elasticities.

The Angrist and Krueger (1999) summary of reliability estimates suggests $r \approx 0.9$ for schooling and $r \approx 0.8$ for earnings, falling to about $0.65-0.75$ for hours worked and hourly wages. The lower end of this range may be more relevant for wage reliability after partialing out covariates or differencing. With $r = 0.65$ as a reasonably conservative value, we'd need R_p^2 equal to at least 0.4 for ρ to reach 0.76. But $E[F] = \frac{nR_p^2}{1-R_p^2} + 1$, so, at this level of first-stage fit, $E[F]$ is nowhere near the trouble zone for any sample size that's empirically relevant. Of course, reliability can be lower than 0.65. Wealth, for instance, is notoriously hard to measure (Saez & Zucman, 2016; Smith et al., 2022), as is consumption (Bee et al., 2015). But neither wealth nor consumption are seen often in the role of a mismeasured endogenous variable to be instrumented. In any case, provided reliability is reasonably high, microeconometric measurement error can be expected to generate parameter combinations for which conventional IV inference is trouble-free.

3.3 Anderson-Rubin vs. Conventional Confidence Intervals

The Anderson and Rubin (1949, AR) statistic for just-ID IV offers an alternative to conventional asymptotic inference. AR inference is appealing by virtue of the fact that AR test size is undistorted by weak instruments under the Staiger and Stock (1997) sequence. Moreover, in the just-ID IV context, an AR test is optimal among unbiased tests (Moreira, 2009). In our setting, the AR statistic can be written:

$$t_{AR} = \frac{\hat{\delta} - \hat{\pi}\beta}{\sqrt{\sigma_{\hat{\delta}}^2 - 2\sigma_{\hat{\delta}\hat{\pi}}\beta + \sigma_{\hat{\pi}}^2\beta^2}}. \tag{12}$$

This differs from t_W in that it replaces $\hat{\beta}_{IV}$ with the null value of β in the formula for $\hat{\sigma}_{IV}^2$: in the context of the just-ID IV model described by eq. (3), the square of t_{AR} equals the Lagrange multiplier statistic testing whether $\frac{\delta}{\pi}$ equals β.[11] AR tests are also compelling by virtue of the fact that, when testing $\beta = 0$, t_{AR} is the t-statistic for the associated reduced form. It's hard to imagine a convincing case for statistical significance of a just-ID IV estimate when the associated reduced form is statistically indistinguishable from zero.

The AR test can be inverted to yield a confidence set that guarantees undistorted coverage for any values of $E[F]$ and ρ. Why not then default to AR confidence sets? For one thing, AR robustness comes at a cost in precision. AR confidence sets have infinite length when $F \leq z_{1-\alpha/2}^2$ (i.e., less than about 4 for 95% intervals). When $F > z_{1-\alpha/2}^2$, AR intervals are longer than the corresponding

[10] The first equality in (11) follows from the definition of correlation, the middle inequality uses the fact that σ_η^2 must be non-negative, and the last equality uses the definition of partial R^2 in eq. (8).

[11] Since the moment restrictions in the IV model are linear, Proposition 3 in Newey and West (1987) implies that t_{AR}^2 is also the relevant likelihood ratio statistic. The t_W vs t_{AR} distinction arises solely by virtue of different variance estimators in the denominator, making tests based on these two statistics first-order equivalent. See Anderson and Rubin (1949) for the AR statistic in over-identified models with a fixed number of instruments and Mikusheva and Sun (2022) for an adaptation to models with many weak instruments.

conventional intervals. In particular, Appendix A.2 shows that finite AR intervals can be written:

$$\hat{\beta}_{IV} - r_{AR} \pm \tau_1 \left(\hat{\rho}^2(\tau_1^2 - 1) + 1\right)^{1/2} \times z_{1-\alpha/2}\hat{\sigma}_{IV}, \tag{13}$$

where $\hat{\rho}$ is an endogeneity estimator given in eq. (A.2), $\tau_1 = (1 - z_{1-\alpha/2}^2/F)^{-1/2}$, and $r_{AR} = \hat{\rho}\hat{\sigma}_{IV}F^{1/2}(\tau_1^2 - 1)$. AR therefore recenters the usual interval at $\hat{\beta}_{IV} - r_{AR}$, while adjusting conventional critical value $z_{1-\alpha/2}$ by $\tau_1 \left(\hat{\rho}^2(\tau_1^2 - 1) + 1\right)^{1/2} > 1$.

The AR length penalty is inversely proportional to estimated first-stage strength; this penalty is substantial even for moderately strong instruments. When $F = 16$, for instance, the AR interval adjustment factor ranges from 14.7–31.6%, depending on estimated endogeneity, while the penalty ranges from 8.7–18.2% when $F = 25$. Lee et al. (2022) develop an appealingly simple alternative robust inference strategy, called tF, that adjusts critical values for t_W depending on the value of F. Because the tF adjustment is made presuming worst-case endogeneity, however, the penalty here is even larger: tF intervals are 42% longer than the usual interval when $F = 16$, and 25% longer for $F = 25$.[12]

The bounding arguments illustrated in Sections 3.1 and 3.2 suggest endogeneity is typically too low for conventional intervals to suffer substantial distortion. In such cases, AR and tF intervals may incur a substantial length penalty while mattering little for coverage. AR and tF intervals are most valuable in applications where endogeneity could plausibly be exceptionally high.

4 Bias Under a Good Sign

Having made an empirical case for conventional inference with just-ID IV, we add a novel analytical argument showing remarkable bias improvements from screening IV estimates on the sign of the estimated first stage. This argument builds on the idea that IV identification strategies are most credible when institutional or theoretical foundations explain the first stage. Such foundations usually imply a sign for π. In the AK91 application, for example, the quarter-of-birth first stage arises from the fact that children born later in the year enter school younger, and are therefore constrained by compulsory attendance laws to stay in school longer than those born earlier. The AE98 samesex instrument for family size is predicated on parents' preference for mixed-sex sibships. The AL99 Maimonides Rule instrument for class size is derived from Israeli regulations that determine class size as a function of enrollment. In these and many other applied micro applications, institutions or preferences sign π.

4.1 Sign-Screened Bias and Coverage

We gauge estimator performance under sign restrictions using median bias since the expectation of a just-ID IV estimator is undefined (2SLS moments exists only for over-identified models). Assuming the sign of π is known, the theorem below shows that $c = 0$ minimizes the median bias of $\hat{\beta}_{IV}$ among screening rules of the form $t_1 > c$.

[12] The tF interval adjustment is smaller than the AR adjustment for values of F between $z_{1-\alpha/2}^2$ and 6.8.

Theorem 1. *Consider the model in eq. (3), and suppose $\pi > 0$. The absolute value of the median bias of $\hat{\beta}_{IV}$ conditional on $t_1 > c$, $|\text{median}_{E[F],\rho}(\hat{\beta}_{IV} - \beta \mid t_1 > c)|$, is minimized at $c = 0$.*

Note that empirical sign-screening yields the greatest bias reduction uniformly over *all* parameter values $(E[F], \rho)$. In particular, empirical sign-screening reduces median bias relative to no screening, since the latter sets $c = -\infty$.

For intuition as to why sign-screening is optimal, note first that by virtue of joint normality of first-stage and reduced form estimates in eq. (3), the distribution of $\hat{\beta}_{IV} - \beta$ conditional on t_1 is normal with a mean and median that can be written:

$$E[\hat{\beta}_{IV} - \beta \mid t_1] = (\beta_{WOLS} - \beta)\frac{t_1 - E[t_1]}{t_1}. \tag{14}$$

Suppose $\rho > 0$, so $\beta_{WOLS} - \beta > 0$ is positive. When t_1 is positive, eq. (14) implies that conditional median bias is increasing in t_1. Hence, for any $a > 0$, screening on $t_1 > 0$ is better than screening on $t_1 > a$, since IV estimates in samples with $t_1 \in [0, a]$ are less biased than estimates with $t_1 > a$. To see why screening on $t_1 > 0$ is better than screening on $t_1 > a$ for $a < 0$, note that conditional on negative t_1, eq. (14) implies that IV bias exceeds that of OLS, because $(t_1 - E[t_1])/t_1 > 1$. But, as we show in Theorem 2 below, median bias of IV conditional on $t_1 > 0$ is smaller than OLS bias. Inclusion of samples with negative t_1 therefore increases median bias. The upshot is an optimal screening cutoff of zero. Figure 2 demonstrates this for selected values of $E[F]$ (numerical computation of median bias is described in Appendix A.7). The kinks at zero in the figure reflect the fact that the median of $\hat{\beta}_{IV}$ conditional on t_1 is discontinuous at zero.

While sign-screening is always salutary, pretesting on $t_1 > c$ for $c > 0$ exacerbates median bias relative to no pretesting unless $E[F]$ is exceedingly small. The figure demonstrates this by marking values of the screening cutoff beyond which screening aggravates bias (these values, determined by bias when $c = -\infty$, are 0.86 for $E[F] = 2$, 0.5 for $E[F] = 3.5$ and 0.38 for $E[F] = 5$). The fact that critical values of c are small in this context explains why pretest rules such as $F > 10$ are often counter-productive. Intuitively, large Fs signal realizations in which the in-sample correlation between instruments and structural errors is largest, exacerbating the median bias of $\hat{\beta}_{IV}$.

The fact that sign-screening affords *substantial* bias reduction is established by a theorem characterizing median IV bias scaled by the weak-IV bias of OLS. Rescaling simplifies bias formulas, while the relationship between conditional and unconditional bias stands without this.[13] The theorem below gives a result for worst-case relative bias over ρ, which obtains in the limit as $|\rho| \to 0$ (this is not the same as relative bias when $\rho = 0$; with no endogeneity, both IV and OLS are unbiased, so that relative bias is discontinuous in ρ). The relationship between ρ and relative median bias derived here contrasts with that in Section 3, which shows higher endogeneity leads to worse coverage. This reversal reflects the fact that, although the bias of $\hat{\beta}_{IV}$ increases with endogeneity, OLS OVB increases faster. Worst-case relative bias is characterized by:

Theorem 2. *Consider the model in eq. (3), and suppose that $\pi > 0$. Then, the unconditional relative*

[13] Stock and Yogo (2005) use a similar rescaling, focusing on relative mean bias for 2SLS models with over-identifying restrictions.

median bias of $\hat{\beta}_{IV}$ is given by

$$\sup_{\rho}\left|\frac{\text{median}_{E[F],\rho}(\hat{\beta}_{IV}-\beta)}{\beta_{WOLS}-\beta}\right| = \frac{\phi(E[t_1])}{E[t_1][\Phi(E[t_1])-1/2]+\phi(E[t_1])}. \tag{15}$$

Moreover, if $E[t_1] \geq 0.84$, the relative median bias of $\hat{\beta}_{IV}$ conditional on $\hat{\pi} > 0$ satisfies

$$\sup_{\rho}\left|\frac{\text{median}_{E[F],\rho}(\hat{\beta}_{IV}-\beta \mid \hat{\pi} > 0)}{\beta_{WOLS}-\beta}\right| = \frac{\phi(E[t_1])}{E[t_1]\Phi(E[t_1])+\phi(E[t_1])}. \tag{16}$$

Equivalently, these expressions give the limit of relative unconditional and conditional median bias as $|\rho| \to 0$.

Maintaining the assumptions of the theorem, IV with or without sign-screening has bias below the weak-instrument bias of OLS, since the right-hand sides of both (15) and (16) are in $(0, 1)$ as long as $\pi > 0$. Note also that the ratio of the two bias expressions in the theorem is close to 0.5 for all but the smallest values of $E[F]$. Specifically, the ratio of conditional to unconditional median bias is:

$$1 - \frac{0.5E[t_1]}{E[t_1]\Phi(E[t_1])+\phi(E[t_1])}.$$

This quantity is within 1 percentage point of 0.5 once $E[t_1]$ greater than about 1.5 since the normal cdf is then close to one and the normal density close to zero.

Theorem 2 describes worst-case bias over ρ. Remarkably, however, the bias reduction from sign-screening varies little with the degree of endogeneity. This is documented in Figure 3, which plots relative bias as a function of the population first-stage F, using shading to mark variation in relative bias as a function of ρ (as in the previous figure, this figure plots numerical calculations detailed in Appendix A.7). We see, for example, that for $E[F]$ around 3.5, the ratio of sign-screened to unconditional bias varies between 0.5–0.52, converging quickly to 0.5 thereafter.[14]

The substantial bias reduction generated by sign-screening may seem surprising since wrong-signed first-stage estimates are rare unless $E[F]$ is small. For instance, when $E[F] = 3.5$, the probability of a wrong-signed estimate is $P(t_1 < 0) = \Phi(-E[t_1]) \approx 5.6\%$. Bias gains from sign-screening arise from the fact that, for positive ρ, the distribution of $\hat{\beta}_{IV}$ conditional on $t_1 < 0$ is heavily shifted to the right. Sign-screening therefore discards samples mostly in the far right tail of the IV sampling distribution. In fact, eq. (14) implies that wrong-signed-conditional median IV bias exceeds OLS bias. So screening yields a material improvement in median IV bias even while the events screened out are rare.

While Theorem 2 establishes the bias-mitigation payoff to sign-screening, Hall et al. (1996) and others show that screening on the first-stage F-statistic is a form of pretesting that may degrade inference. The problem here is that, when π is truly zero, large F-statistics overstate first-stage strength, leading to overly optimistic standard errors. And, as noted in the discussion of Figure 2, conditional bias is aggravated by the fact that large Fs signal sample realizations in which the in-sample correlation between instruments and structural errors is largest, exacerbating the bias of $\hat{\beta}_{IV}$.

[14] Richardson (1968) shows that under homoskedasticity, the relative (mean) bias of over-identified 2SLS relative to (weak-instrument) OLS bias is unrelated to ρ.

Consequently, when instruments are truly weak, pretesting can lead to confidence intervals with very poor coverage. It's therefore worth investigating whether empirical sign-screening runs a similar risk. As noted above, wrong-signed first stage estimates are rare when the first stage is nonzero, but pretesting problems are most salient when instruments are indeed weak.

As it turns out, pretesting concerns here are unfounded. By way of evidence on this point, Panel (b) in Figure 1 plots rejection contours for a conventional (second-stage) t-test conditional on $\hat{\pi} > 0$. That is, the figure plots contours for:

$$R_W^c = P_{E[F],\rho}(|t_W| > z_{1-\alpha/2} \mid \hat{\pi} > 0).$$

Comparison of the two panels in Figure 1 suggests sign-screening affects rejection rates little. For instance, the endogeneity threshold required to keep rejections rates below 10% is $|\rho| \leq 0.75$, close to the unconditional value of 0.76 otherwise required for this. This happy finding is explained by the fact that, when the instrument is very weak, sign-screening has two effects. On one hand, screening out wrong-signed first-stage estimates tends to overestimate first-stage strength. At the same time, in contrast to screening on first-stage F, the median bias of $\hat{\beta}_{IV}$ is reduced. These two effects are just about offsetting, so that the conditional rejection contours in panel (b) of Figure 1 are much like the unconditional contours plotted in panel (a).

4.2 Theoretical Sign Restrictions Only

Andrews and Armstrong (2017) introduce an estimator that's unbiased given a sign restriction on the population first-stage coefficient, rather than the sign of estimated $\hat{\pi}$. This unbiased estimator is:

$$\hat{\beta}_U \equiv \hat{\tau}(\hat{\delta} - \beta_{WOLS}\hat{\pi}) + \beta_{WOLS} = t_1\mu(t_1)\hat{\beta}_{IV} + (1 - t_1\mu(t_1))\beta_{WOLS}, \tag{17}$$

where $\mu(x) = \frac{1-\Phi(x)}{\phi(x)}$ is the Mills' ratio of a standard normal random variable (ϕ and Φ denote the standard normal density and cdf, respectively). Estimator $\hat{\beta}_U$ is unbiased because, under normality of first-stage estimates and given $\pi > 0$,

$$E\left[\frac{\mu(t_1)}{\sigma_{\hat{\pi}}}\right] = \frac{1}{\pi},$$

that is, $\hat{\tau} \equiv \frac{\mu(t_1)}{\sigma_{\hat{\pi}}}$ is an unbiased estimator of the reciprocal of π. Moreover, $\hat{\delta} - \beta_{WOLS}\hat{\pi}$ and $\hat{\tau}$ are uncorrelated, since β_{WOLS} is the slope in the regression of the estimated reduced form on the estimated first stage. Unbiasedness then follows from the fact that $E[\hat{\delta} - \beta_{WOLS}\hat{\pi}] = (\beta - \beta_{WOLS})\pi$.

$\hat{\beta}_U$ has an interesting and counter-intuitive shrinkage interpretation when $t_1 > 0$. Observe that

$$0 \leq 1 - t_1\mu(t_1) \leq \frac{1}{t_1^2} \tag{18}$$

when $t_1 > 0$ (this is implied by a Mill's ratio inequality given in Feller (1968, p. 175)). Thus, when the first stage is right-signed, the weights $t_1\mu(t_1)$ in eq. (17) lie between 0 and 1, and $\hat{\beta}_U$ shrinks the conventional IV estimate towards OLS.

The shrinkage interpretation of $\hat{\beta}_U$ seems surprising since shrinkage *towards* OLS increases bias. This fact is reconciled with the unbiasedness of $\hat{\beta}_U$ by the following theorem:

Theorem 3. *Consider the model in (3), and suppose that $\pi > 0$. Then, the mean bias of $\hat{\beta}_U$ conditional on $t_1 > 0$ can be written:*

$$E[\hat{\beta}_U - \beta \mid t_1 > 0] = \frac{0.5e^{-0.5E[t_1]^2}}{\Phi(E[t_1])}(\beta_{WOLS} - \beta),$$

while, conditional on $t_1 < 0$, relative mean bias is:

$$E[\hat{\beta}_U - \beta \mid t_1 < 0] = -\frac{0.5e^{-0.5E[t_1]^2}}{1 - \Phi(E[t_1])}(\beta_{WOLS} - \beta).$$

Note that the denominators of these expressions equal the probability t_1 is positive and negative, respectively. $\hat{\beta}_U$ is therefore unbiased because it averages conditional positive bias when $t_1 > 0$ and conditional negative bias when $t_1 < 0$.

An analyst who is prepared to sign the population first stage seems unlikely to ignore the sign of the estimated first stage. Yet, when it comes to $\hat{\beta}_U$, empirical sign-screening results in more bias not less. This would seem to strip $\hat{\beta}_U$ of its appeal. And use of median rather than mean bias to measure performance does not ameliorate this: Appendix A.7 shows that the conditional median bias of $\hat{\beta}_{IV}$ is always less than that of $\hat{\beta}_U$, and at least 50% smaller once $E[t_1] \geq 1$.[15]

5 Summary and Conclusions

Assuming reduced-form and first-stage estimates are approximately normally distributed, null rejection rates for conventional t-tests in just-ID IV models are distorted little unless endogeneity is exceptionally high. A corollary to this fact is good coverage of conventional confidence intervals. Three widely-cited applications, two of which exhibit considerable OLS OVB, are characterized by moderate endogeneity and consequently fall well inside the low-distortion just-ID IV comfort zone. We've argued that these three examples should be seen as representative rather than idiosyncratic: the structure of much applied micro research naturally bounds endogeneity.

We also develop a new theoretical argument alleviating bias concerns in just-ID IV. As Andrews and Armstrong (2017) note, the most convincing applications of just-ID IV restrict the sign of the first stage. Unlike Andrews and Armstrong (2017), however, we impose the same sign restriction on the estimated as well as the theoretical first stage. In contrast to screening on first-stage F, which may do more harm than good, empirical sign-screening roughly halves the median bias of the IV estimator without degrading coverage. Since most analysts likely impose an estimated first-stage sign screen as a matter of course, the bias reduction sign-conditioning engenders should be reflected in published empirical work.

[15] Andrews and Armstrong (2017) shows numerically that the unconditional median bias of $\hat{\beta}_U$ is smaller than that of $\hat{\beta}_{IV}$ when $E[F]$ is small, while this bias ranking reverses for larger $E[F]$. Andrews and Armstrong (2017) notes also that the median absolute deviation of $\hat{\beta}_U$ is always smaller than that of $\hat{\beta}_{IV}$. Our numerical calculations indicate that, conditional on the estimated first stage sign, this no longer holds for all parameter values.

What practical lesson should we draw from this? In the context of the AK91, AE98, and AL99 studies, first-stage sign screening adds no action items to the empirical agenda. The first-stage estimates in these applications are robustly right-signed. In applications with weaker instruments than these, an empirical strategy that begins by examining the first-stage sign would seem to have no downside. Claims for credible causal evidence requires more than this, however. In AK91, for instance, the quarter-of-birth story holds water because schooling can be seen to move sharply up and down with quarter of birth as predicted by compulsory attendance laws, across 30 birth cohorts in three data sets, and because graduate degree completion, which should be changed little by compulsory attendance, moves little with quarter of birth. This sort of coherence contributes as much or more than statistical significance to first stage strength.

References

Anderson, T. W., & Rubin, H. (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. https://doi.org/10.1214/aoms/1177730090

Andrews, I., & Armstrong, T. B. (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, *8*(2), 479–503. https://doi.org/10.3982/QE700

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*(1), 727–753. https://doi.org/10.1146/annurev-economics-080218-025643

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, *80*(3), 313–336. https://www.jstor.org/stable/10.2307/2006669

Angrist, J. D., & Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, *88*(3), 450–477. https://www.jstor.org/stable/116844

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, *90*(430), 431–442. https://doi.org/10.1080/01621459.1995.10476535

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (pp. 1277–1366). Elsevier. https://doi.org/10.1016/S1573-4463(99)03004-7

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575. https://doi.org/10.1162/003355399556061

Bee, A., Meyer, B. D., & Sullivan, J. X. (2015). The validity of consumption data: Are the consumer expenditure interview and diary surveys informative? In C. D. Carroll, T. F. Crossley, & J.

Sabelhaus (Eds.), *Improving the measurement of consumer expenditures* (pp. 204–240). The University of Chicago Press.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, *62*(3), 657–681. https://doi.org/10.2307/2951662

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450. https://doi.org/10.1080/01621459.1995.10476536

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*(5), 1127–1160. https://doi.org/10.1111/1468-0262.00237

Chernozhukov, V., & Hansen, C. (2008). The reduced form: A simple approach to inference with weak instruments. *Economics Letters*, *100*(1), 68–71. https://doi.org/10.1016/j.econlet.2007.11.012

Feller, W. (1968). *An introduction to probability theory and its application* (3rd ed., Vol. 1). Wiley.

Hall, A. R., Rudebusch, G. D., & Wilcox, D. W. (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review*, *37*(2), 283. https://doi.org/10.2307/2527324

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, *24*(3), 1141–1177. https://www.jstor.org/stable/2725865

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475. https://doi.org/10.2307/2951620

Keane, M., & Neal, T. (2022). *A practical guide to weak instruments* (Working Paper). SSRN. https://doi.org/10.2139/ssrn.3846841

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497–532. https://doi.org/10.1162/003355399556052

Lang, K. (1993). *Ability bias, discount rate bias, and the return to education* [Unpublished manuscript, Boston University].

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. (2022). Valid t-ratio inference for IV. *American Economic Review*, *112*(10), 3260–3290. https://doi.org/10.1257/aer.20211063

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21–59. https://doi.org/10.1017/S0266466605050036

Mikusheva, A., & Sun, L. (2022). Inference with many weak instruments. *The Review of Economic Studies*, *89*(5), 2663–2686. https://doi.org/10.1093/restud/rdab097

Moreira, M. J. (2009). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics*, *152*(2), 131–140. https://doi.org/10.1016/j.jeconom.2009.01.012

Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, *28*(3), 777–787. https://doi.org/10.2307/2526578

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics - Simulation and Computation*, *9*(4), 389–419. https://doi.org/10.1080/03610918008812164

Richardson, D. H. (1968). The exact distribution of a structural coefficient estimator. *Journal of the American Statistical Association*, *63*(324), 1214–1226. https://doi.org/10.1080/01621459.1968.10480921

Saez, E., & Zucman, G. (2016). Wealth inequality in the United States since 1913: Evidence from capitalized income tax data. *The Quarterly Journal of Economics*, *131*(2), 519–578. https://doi.org/10.1093/qje/qjw004

Smith, M., Zidar, O., & Zwick, E. (2022). Top wealth in America: New estimates under heterogeneous returns. *The Quarterly Journal of Economics*. https://doi.org/10.1093/qje/qjac033

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557–586. https://doi.org/10.2307/2171753

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge University Press. https://doi.org/10.1017/CBO9780511614491.006

Young, A. (2022). Consistency without inference: Instrumental variables in practical application. *European Economic Review*, *147*. https://doi.org/10.1016/j.euroecorev.2022.104112

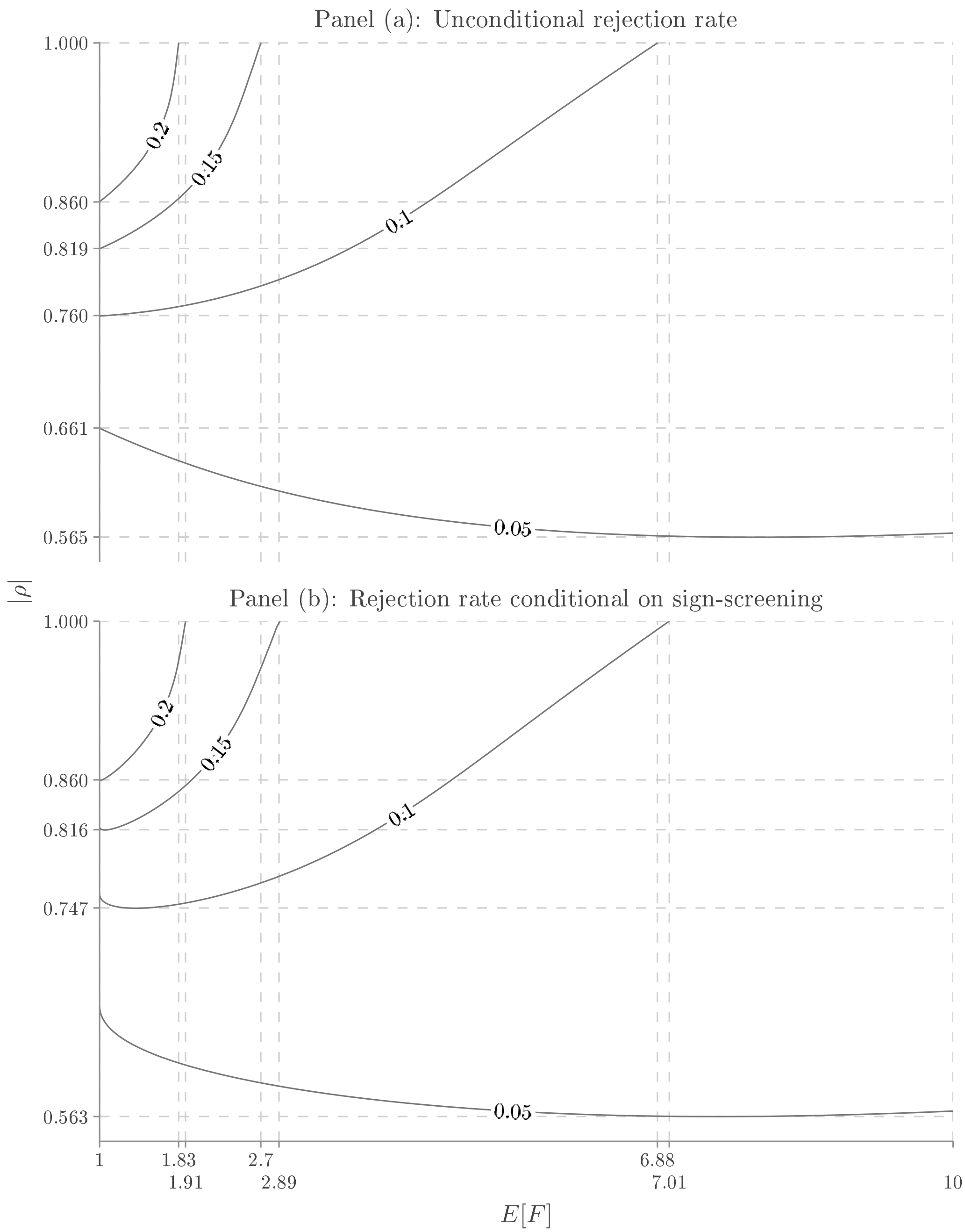

Figure 1: Contour plot of the rejection rate of conventional t-test with nominal level $\alpha = 0.05$ as function of $E[F]$ and ρ. Panel (a) plots the unconditional rejection rate R_W. Panel (b) plots the rejection rate R_W^c conditional on $\hat{\pi} > 0$. See Appendix A.3 for computational details.

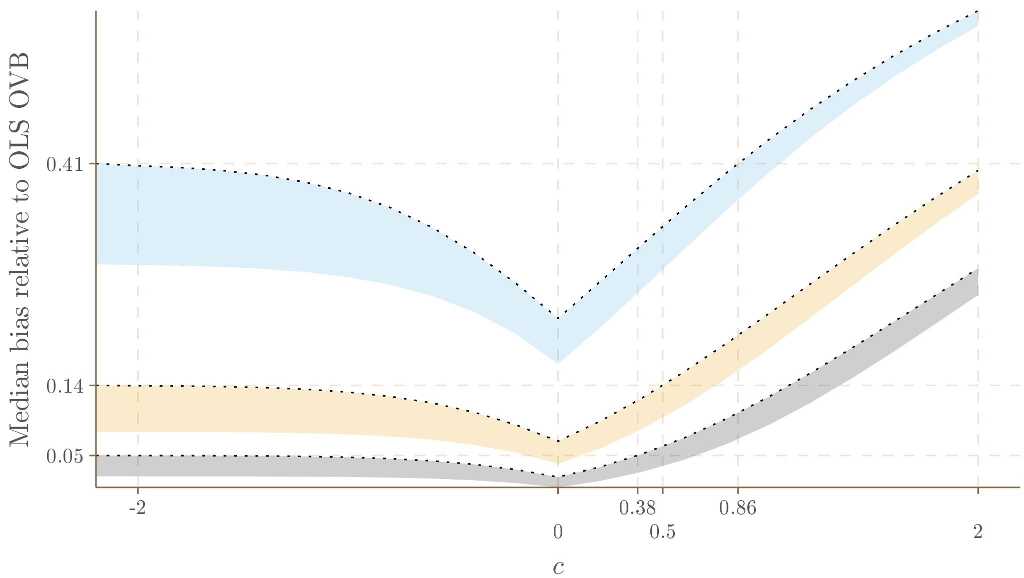

Figure 2: Relative median bias of $\hat{\beta}_{IV}$ conditional on $t_1 > c$. The figure plots median$(\hat{\beta}_{IV} - \beta \mid t_1 > c)/|\beta_{WOLS} - \beta|$ as a function of test cutoff c for select values of $E[F]$. Shaded areas cover the range of variation in relative bias over possible values of ρ. Blue: $E[F] = 2$, orange: $E[F] = 3.5$, gray: $E[F] = 5$. Dotted lines denote limiting relative bias as $\rho \to 0$. For each value of $E[F]$, horizontal gridlines mark the value of this relative bias when $c = -\infty$, and vertical gridlines mark the cutoff value for which screening on $t_1 > c$ increases bias relative to no screening. See Appendix A.7 for computational details.

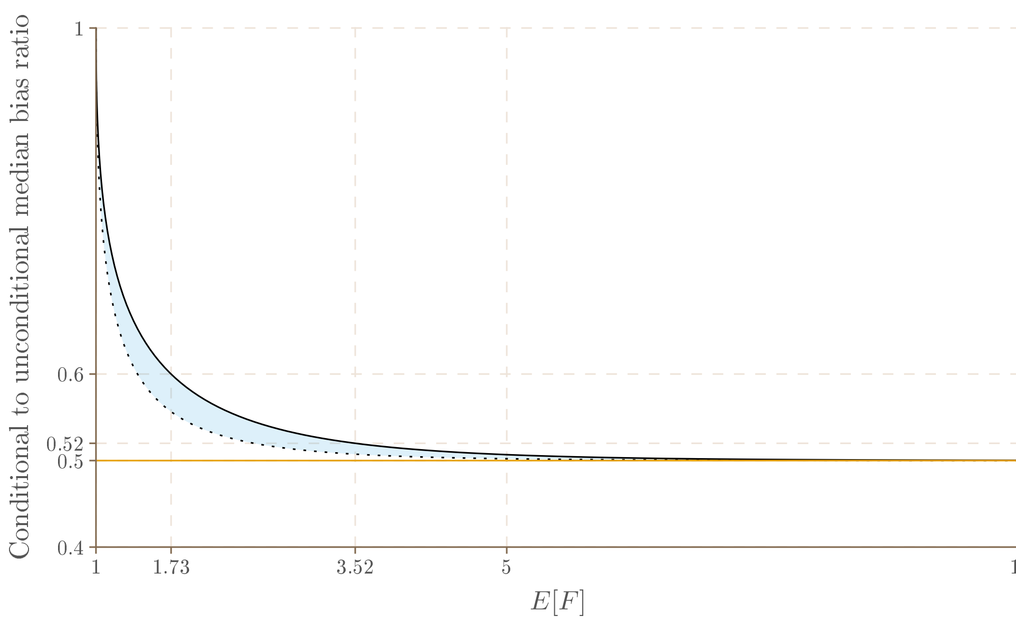

Figure 3: Median bias of $\hat{\beta}_{IV}$ conditional on $t_1 > 0$ relative to unconditional median bias. The solid line plots the bias ratio when $\rho = 1$; the dotted line denotes the limit of the bias ratio as $\rho \to 0$. The blue shaded area covers the range of the bias ratio over possible values of ρ. The solid orange line marks a reference line at 0.5. See Appendix A.7 for computational details.

Table 1: Estimates and Endogeneity in Three IV Applications.

				Estimates				
	Outcome	Treatment	Instrument	OLS	$\hat{\pi}$	$\hat{\delta}$	IV	$\hat{\rho}$
Sample	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			A. AK91					
Men born 1920–29	Log weekly wage	Years of education	Born in Q1	0.080	−0.122	−0.009	0.072	0.043
($n = 247,199$)	[0.65]	[3.36]	[0.43]	(0.0004)	(0.016)	(0.003)	(0.023)	($-0.210, 0.292$)
Men born 1930–39				0.071	−0.106	−0.011	0.105	−0.175
($n = 329,509$)	[0.68]	[3.28]	[0.43]	(0.0004)	(0.013)	(0.003)	(0.025)	($-0.407, 0.068$)
			B. AE98					
1980 Census, mothers aged 18–35	Worked for pay	More than 2 kids	Samesex	−0.176	0.061	−0.007	−0.117	−0.058
($n = 394,840$)	[0.50]	[0.49]	[0.50]	(0.002)	(0.001)	(0.002)	(0.025)	($-0.105, -0.010$)
	Weeks worked			−8.978	0.061	−0.340	−5.559	−0.075
	[22.29]	[0.49]	[0.50]	(0.071)	(0.001)	(0.069)	(1.118)	($-0.122, -0.027$)
1990 Census, mothers aged 18–35	Worked for pay			−0.164	0.062	−0.006	−0.092	−0.069
($n = 380,007$)	[0.48]	[0.48]	[0.50]	(0.002)	(0.002)	(0.002)	(0.025)	($-0.116, -0.022$)
	Weeks worked			−8.649	0.062	−0.341	−5.462	−0.067
	[22.87]	[0.48]	[0.50]	(0.077)	(0.002)	(0.073)	(1.156)	($-0.114, -0.020$)
			C. AL99					
5th grade, full sample	Reading	Class size	Maimonides' rule	0.009	0.477	−0.125	−0.263	0.357
($n = 2,019$)	[7.68]	[6.55]	[6.11]	(0.034)	(0.041)	(0.042)	(0.094)	($0.207, 0.499$)
	Math			0.036	0.477	−0.126	−0.264	0.315
	[9.60]	[6.54]	[6.10]	(0.045)	(0.041)	(0.056)	(0.123)	($0.160, 0.462$)
5th grade, discontinuity sample	Reading			−0.070	0.481	−0.197	−0.410	0.469
($n = 471$)	[8.18]	[7.42]	[7.50]	(0.050)	(0.057)	(0.050)	(0.118)	($0.279, 0.639$)
	Math			0.090	0.481	−0.089	−0.185	0.322
	[10.20]	[7.42]	[7.50]	(0.070)	(0.057)	(0.072)	(0.155)	($0.107, 0.521$)

Notes: This table reports IV and OLS estimates replicating the AK91, AE98, and AL99 studies discussed in the text. For each study, the table reports IV and OLS estimates from multiple samples, as well as the corresponding first-stage estimates, $\hat{\pi}$, reduced-form estimates, $\hat{\delta}$, and estimates of the endogeneity parameter, $\hat{\rho}$. Standard errors appear in parentheses. These are robust for AK91 and AE98, and clustered on school for AL99. Confidence intervals for ρ, reported below parameter estimates, are computed as described in Appendix A.1. Standard deviations for the outcome, treatment, and instrument are reported in columns 1–3 in brackets.

Appendix A Derivations and Proofs

The appendix uses the notation $\tilde{\beta}_{IV} = (\hat{\beta}_{IV} - \beta)/|\beta_{WOLS} - \beta|$ and $\tilde{\beta}_U = (\hat{\beta}_{IV} - \beta)/|\beta_{WOLS} - \beta|$ to denote the IV and the unbiased estimator, after centering and scaling by the weak-IV OVB of OLS. Also, we let $\omega = \rho/\sqrt{1-\rho^2}$.

A.1 Estimating ρ

We estimate ρ (defined in eq. (7)) using first-stage and IV estimates, and the associated first-stage, reduced-form and IV standard errors. To see how this works, rewrite eq. (5) as:

$$\sigma_{\hat{\delta}\hat{\pi}} = \frac{\sigma_{\hat{\pi}}^2 \hat{\beta}_{IV}^2 - \hat{\pi}^2 \hat{\sigma}_{IV}^2 + \sigma_{\hat{\delta}}^2}{2\hat{\beta}_{IV}}. \tag{A.1}$$

With this in hand for $\sigma_{\hat{\delta}\hat{\pi}}$, endogeneity can be computed as the sample analog of eq. (7), replacing β with $\hat{\beta}_{IV}$. The resulting endogeneity estimator is:

$$\hat{\rho} = \frac{\sigma_{\hat{\pi}}}{|\hat{\pi}|\hat{\sigma}_{IV}} \times (\sigma_{\hat{\delta}\hat{\pi}}/\sigma_{\hat{\pi}}^2 - \hat{\beta}_{IV}). \tag{A.2}$$

Under the normal model in eq. (3), this estimator depends on the data only through $\hat{\beta}_{IV}$, with the derivative given by $\partial\hat{\rho}/\partial\hat{\beta}_{IV} = \hat{\sigma}_{IV}^{-1} \cdot (\hat{\rho}^2 - 1)/|t_1|$. Hence, the delta-method standard error for $\hat{\rho}$ is simply $(1-\hat{\rho}^2)/|t_1|$.

Paralleling concerns with finite-sample coverage of the usual confidence interval for β, we might worry that confidence intervals for ρ based on delta-method standard errors suffer from undercoverage if endogeneity is high and the instruments are weak. We therefore compute confidence sets for ρ by inverting the AR statistic. Specifically, denote the AR confidence set by $[\beta_\ell, \beta_u]$ when this is finite. Since ρ is monotone decreasing in β, this leads to a confidence set for ρ that can be written $[\varrho(\beta_u), \varrho(\beta_\ell)]$, where $\varrho(\beta) = \frac{\sigma_{\hat{\pi}}}{\sqrt{\sigma_{\hat{\delta}}^2 - 2\beta\sigma_{\hat{\delta}\hat{\pi}} + \sigma_{\hat{\pi}}^2\beta^2}} \times (\sigma_{\hat{\delta}\hat{\pi}}/\sigma_{\hat{\pi}}^2 - \beta)$. When the AR confidence set takes the form $(-\infty, \beta_\ell] \cup [\beta_u, \infty)$, the confidence set for ρ is $[-1, \varrho(\beta_u)] \cup [\varrho(\beta_\ell), 1]$.

A.2 AR confidence sets for β

The AR confidence set consists of all points β_0 that are not rejected by the AR test. These points must therefore satisfy the inequality $z_{1-\alpha/2}^2 \geq \frac{(\hat{\delta}-\hat{\pi}\beta_0)^2}{\sigma_{\hat{\delta}}^2 - 2\sigma_{\hat{\delta}\hat{\pi}}\beta_0 + \sigma_{\hat{\pi}}^2\beta_0^2}$. Letting $\Delta_0 = \beta_0 - \hat{\beta}_{IV}$, and using eqs. (A.2) and (5), we can write this inequality as

$$(F - z_{1-\alpha/2}^2)\Delta_0^2 + 2z_{1-\alpha/2}^2 \hat{\rho} F^{1/2} \hat{\sigma}_{IV} \Delta_0 - z_{1-\alpha/2}^2 F \hat{\sigma}_{IV}^2 \leq 0.$$

Solving this quadratic inequality, we obtain that when $F > z_{1-\alpha/2}^2$, the confidence interval for $\beta - \hat{\beta}_{IV}$ has endpoints given by

$$-(\tau_1^2 - 1)\hat{\rho}F^{1/2}\hat{\sigma}_{IV} \pm \tau_1 z_{1-\alpha/2}\hat{\sigma}_{IV}\sqrt{(\tau_1^2-1)\hat{\rho}^2 + 1}.$$

It follows that the confidence interval for β is given by eq. (13).

A.3 t-Test Rejection Rates

This section writes the rejection probabilities of the t-test as an integral indexed by $(E[F], \rho)$. Stock and Yogo (2005) use Monte Carlo methods to compute unconditional rejection probabilities in a similar setup. The calculation described here is much faster. More importantly, it allows us to easily compute both unconditional rejection rates and rejection rates conditional on sign-screening.

Using eq. (7), and the fact that $\beta_{WOLS} - \beta$ and ρ have the same sign, we may write t_{AR} as

$$t_{AR} = \frac{(\hat{\delta} - \hat{\pi}\beta)|\rho|}{\sigma_{\hat{\pi}}|\beta_{WOLS} - \beta|}. \tag{A.3}$$

Consequently,

$$\tilde{\beta}_{IV} = \frac{\hat{\delta} - \beta\hat{\pi}}{\sigma_{\hat{\pi}} t_1 |\beta_{WOLS} - \beta|} = \frac{t_{AR}}{|\rho| t_1}. \tag{A.4}$$

Thus,

$$t_W = \frac{\operatorname{sign}(t_1) t_{AR}}{\sqrt{\frac{\sigma^2_{\hat{\delta}}/\sigma^2_{\hat{\pi}} - 2\beta_{WOLS}\beta + \beta^2}{(\beta_{WOLS} - \beta)^2}\rho^2 + \frac{t^2_{AR}}{t^2_1} - 2\rho\frac{t_{AR}}{t_1}}} = \frac{\operatorname{sign}(t_1) t_{AR}}{\sqrt{1 + t^2_{AR}/t^2_1 - 2\rho t_{AR}/t_1}} \tag{A.5}$$

where the first equality uses eq. (A.4) and the definition of β_{WOLS}, and the second equality uses eq. (7). This expression for t_W implies that conditional on t_1, the rejection region $\{|t_W| \geq z_{1-\alpha/2}\}$ is quadratic in t_{AR}. Solving this quadratic inequality implies that the rejection region is given by

$$t_{AR} \in \begin{cases} \emptyset & \text{if } t_1^2 \leq (1-\rho^2) z^2_{1-\alpha/2}, \\ [a_1, a_2] & \text{if } (1-\rho^2) z^2_{1-\alpha/2} \leq t_1^2 \leq z^2_{1-\alpha/2}. \\ (-\infty, a_2) \cup (a_1, \infty) & \text{if } t_1^2 \geq z^2_{1-\alpha/2} \end{cases}$$

where

$$a_1 = \frac{\rho z^2_{1-\alpha/2} t_1 - |t_1| z_{1-\alpha/2}\sqrt{t_1^2 - (1-\rho^2) z^2_{1-\alpha/2}}}{z^2_{1-\alpha/2} - t_1^2},$$

$$a_2 = \frac{\rho z^2_{1-\alpha/2} t_1 + |t_1| z_{1-\alpha/2}\sqrt{t_1^2 - (1-\rho^2) z^2_{1-\alpha/2}}}{z^2_{1-\alpha/2} - t_1^2}.$$

Note that $\operatorname{cor}(t_{AR}, t_1) = \rho$, so that

$$P(t_{AR} \leq x \mid t_1) = \Phi((x - \rho(t_1 - E[t_1]))/\sqrt{1-\rho^2}). \tag{A.6}$$

Thus, conditional on t_1, the rejection probability is given by

$$\begin{aligned} P(|t_W| \geq z_{1-\alpha/2} \mid t_1) &= (P(t_{AR} \leq a_2 \mid t_1) - P(t_{AR} \leq a_1 \mid t_1))\,\mathbb{I}\{t_1^2 \geq z^2_{1-\alpha/2}(1-\rho^2)\} \\ &\quad + \mathbb{I}\{t_1^2 \geq z^2_{1-\alpha/2}\} \\ &= f(t_1; E[t_1], \rho)\,\mathbb{I}\{t_1^2 \geq (1-\rho^2) z^2_{1-\alpha/2}\} + \mathbb{I}\{t_1^2 \geq z^2_{1-\alpha/2}\}, \end{aligned} \tag{A.7}$$

where

$$f(t_1; E[t_1], \rho) = \Phi\left(\frac{a_2 - \rho(t_1 - E[t_1])}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{a_1 - \rho(t_1 - E[t_1])}{\sqrt{1-\rho^2}}\right).$$

Since $t_1 \sim \mathcal{N}(E[t_1], 1)$, the rejection probability conditional on $t_1 \geq c$ is therefore given by

$$P(|t_W| \geq z_{1-\alpha/2} \mid t_1 \geq c) = \frac{\int_c^\infty (\mathbb{I}\{\frac{t_1^2}{1-\rho^2} \geq z_{1-\alpha/2}\} f(t_1; E[t_1], \rho) + \mathbb{I}\{t_1^2 \geq z_{1-\alpha/2}^2\})\phi(t_1 - E[t_1])dt_1}{\Phi(E[t_1] - c)}.$$

The unconditional rejection probability R_W obtains by setting $c = -\infty$. The rejection probability conditional on sign screening, R_W^c, obtains by setting $c = 0$. The rejection contours in Figure 1 evaluate the above expression as a function of $(\rho, E[t_1])$ by numerical integration.

A.4 Proof of Theorem 1

The distribution of $\tilde{\beta}_{IV}$ conditional on t_1 can be written as

$$P_{\omega, E[t_1]}(\tilde{\beta}_{IV} \leq x \mid t_1) = \begin{cases} P(t_{AR} \leq |\rho| t_1 x \mid t_1) = \Phi(\omega[E[t_1] - (1 - \text{sign}(\omega)x)t_1]) & \text{if } t_1 \geq 0, \\ P(t_{AR} \geq |\rho| t_1 x \mid t_1) = \Phi(-\omega[E[t_1] - (1 - \text{sign}(\omega)x)t_1]) & \text{if } t_1 < 0, \end{cases} \tag{A.8}$$

where the first equality uses eq. (A.4), and the second equality follows from eq. (A.6). Observe that since $P_{-\omega, E[t_1]}(\tilde{\beta}_{IV} \leq x \mid t_1) = 1 - P_{\omega, E[t_1]}(\tilde{\beta}_{IV} \leq -x \mid t_1)$, the distribution is symmetric in ω. It therefore suffices to consider $\omega > 0$.

Let $p_c(x) = P_{\omega, E[t_1]}(\tilde{\beta}_{IV} \leq x \mid t_1 > c)$ denote the distribution of $\tilde{\beta}_{IV}$ conditional on $t_1 > c$ (this shorthand notation ignores the dependence on ω and $E[t_1]$). By eq. (A.8),

$$p_c(x) = \begin{cases} \frac{1}{\Phi(E[t_1]-c)} \int_{c-E[t_1]}^\infty \Phi\left(y(z)\right) \phi(z) dz & \text{if } c \geq 0, \\ \frac{1}{\Phi(E[t_1]-c)} \left[\int_{c-E[t_1]}^{-E[t_1]} \Phi(-y(z))\phi(z)dz + \int_{-E[t_1]}^\infty \Phi(y(z))\phi(z)dz\right] & \text{if } c < 0, \end{cases}$$

where $y(z) = \omega[xE[t_1] - (1-x)z]$. Observe that for all c, the conditional median of $\tilde{\beta}_{IV}$, denoted $m_c = m_c(\omega)$, is smaller than 1. In particular, for $c \geq 0$,

$$p_c(1) = \frac{\Phi\left(\omega E[t_1]\right)}{\Phi(E[t_1] - c)} \int_{c-E[t_1]}^\infty \phi(z) dz = \Phi(\omega E[t_1]) \geq 1/2, \tag{A.9}$$

while if $c < 0$,

$$\begin{aligned} p_c(1) &= \frac{1}{\Phi(E[t_1] - c)} \left[\Phi(E[t_1] - c) - \Phi(E[t_1]) + \Phi(\omega E[t_1])(-\Phi(E[t_1] - c) + 2\Phi(E[t_1]))\right] \\ &\geq \frac{1}{\Phi(E[t_1] - c)} \left[\Phi(E[t_1] - c) - \Phi(E[t_1]) + \frac{1}{2}(2\Phi(E[t_1]) - \Phi(E[t_1] - c))\right] = \frac{1}{2}. \end{aligned}$$

We now show that the m_c is minimized at $c = 0$. If $c \geq 0$, by Leibniz rule,

$$\frac{\partial p_c(x)}{\partial c} = \frac{\phi(E[t_1] - c)}{\Phi(E[t_1] - c)} \left[p_c(x) - \Phi\left(\omega[E[t_1] - (1-x)c]\right)\right],$$

which is negative for $x \leq 1$, since it follows from eq. (A.8) that $p_c(x) \leq \frac{1}{\Phi(E[t_1]-c)} \int_{c-E[t_1]}^{\infty} \Phi(\omega[E[t_1] - (1-x)c])\phi(z)dz = \Phi\left(\omega[E[t_1] - (1-x)c]\right)$. Therefore, m_c is decreasing for positive c.

Now consider $c < 0$. From eq. (A.8), it follows that

$$p_c(x) = \frac{1}{\Phi(E[t_1]-c)} \left[\int_{c-E[t_1]}^{-E[t_1]} \Phi(\omega[-E[t_1] + (1-x)(z+E[t_1])])\phi(z)dz + \Phi(E[t_1])p_0(x) \right].$$

Suppose $p_0(m_c) < 1/2$. Then it follows from the preceding display that for $x \leq 1$,

$$\frac{1}{2} < \frac{1}{\Phi(E[t_1]-c)} \left[\int_{c-E[t_1]}^{-E[t_1]} \Phi(\omega[-E[t_1] + (1-x)(z+E[t_1])])\phi(z)dz + \Phi(E[t_1])\frac{1}{2} \right]$$
$$\leq \frac{1}{\Phi(E[t_1]-c)} \left[\frac{1}{2} \int_{c-E[t_1]}^{-E[t_1]} \phi(z)dz + \Phi(E[t_1])\frac{1}{2} \right] = \frac{1}{2},$$

where the second inequality uses the fact that $\Phi(\omega[-E[t_1] + (1-x)(z+E[t_1])]) \leq \Phi(-E[t_1]\omega) \leq \Phi(0)$ over the range of integration. Hence, $p_0(m_c) \geq 1/2$, which implies that $m_0 \leq m_c$. By the proof of Theorem 2 $m_0 \geq 0$. Thus, setting $c = 0$ minimizes $|m_c|$ for all c, as claimed.

A.5 Proof of Theorem 2

The proof begins by characterizing the distribution of $\tilde{\beta}_{IV}$ conditional on $t_1 > 0$. The previous proof establishes that the conditional median for this distribution is less than 1, so it suffices to consider $p_0(x)$ for $x \leq 1$. By the mean value theorem, for some $\tilde{\omega} = \tilde{\omega}(x, \omega) \in [0, \omega]$,

$$p_0(x) = \Phi(0) + \frac{\omega}{\Phi(E[t_1])} \int_{-E[t_1]}^{\infty} ((x-1)z + xE[t_1])\phi\left(\tilde{\omega}((x-1)z + xE[t_1])\right)\phi(z)dz$$
$$= \frac{1}{2} + \frac{\omega}{\tilde{\omega}^2(1-x)\Phi(E[t_1])} \int_{-\infty}^{\tilde{\omega}E[t_1]} y\phi\left(y\right)\phi(a+by)dy,$$

where the second line uses the change of variables $y = \tilde{\omega}xE[t_1] - \tilde{\omega}(1-x)z$, and we let $a = xE[t_1]/(1-x)$, $b = -\frac{1}{\tilde{\omega}(1-x)}$. By line 111 of Table 1 in Owen (1980),

$$\int x\phi(x)\phi(a+bx) = \frac{\phi(a/t)}{t^2}\left[-\phi(tx+ab/t) - \frac{ab}{t}\Phi(tx+ab/t) \right], \quad t = \sqrt{1+b^2}. \tag{A.10}$$

Applying this result to the preceding display then yields

$$p_0(x) = \frac{1}{2} + \frac{\omega}{\tilde{\omega}^2(1-x)\Phi(E[t_1])}\frac{\phi(a/t)}{1+b^2}\left[-\phi\left(t\tilde{\omega}E[t_1] + ab/t\right) - \frac{ab}{\sqrt{1+b^2}}\Phi(t\tilde{\omega}E[t_1] + ab/t) \right]$$
$$= \frac{1}{2} + \frac{\omega}{\Phi(E[t_1])}\frac{\phi(a/t)(1-x)}{\tilde{\omega}^2(1-x)^2+1}\left[\frac{x}{|1-x|}\frac{E[t_1]}{\tilde{g}(x,\tilde{\omega})}\Phi(E[t_1]g(x,\tilde{\omega})) - \phi\left(E[t_1]g(x,\tilde{\omega})\right) \right],$$

where $g(x, \tilde{\omega}) = \frac{\tilde{\omega}^2|1-x| + \text{sign}(1-x)}{\sqrt{\tilde{\omega}^2(1-x)^2+1}}$, and $\tilde{g}(x, \tilde{\omega}) = \sqrt{\tilde{\omega}^2(1-x)^2+1}$. When evaluated at the conditional median, m_0, the expression in square brackets must equal zero by definition of the median. Therefore,

$m_0 > 0$, and since we also know from eq. (A.9) that $m_0 < 1$, the conditional median must satisfy

$$m_0 = \frac{1}{\frac{E[t_1]}{\tilde{g}(m_0,\tilde{\omega}(m_0,\omega))}\frac{\Phi(E[t_1]g(m_0,\tilde{\omega}(m_0,\omega)))}{\phi(E[t_1]g(m_0,\tilde{\omega}(m_0,\omega)))}+1}, \tag{A.11}$$

We have

$$\frac{E[t_1]}{\tilde{g}}\frac{\Phi\left(E[t_1]g\right)}{\phi(E[t_1]g)} \geq \frac{E[t_1]}{\tilde{g}}\frac{\Phi\left(E[t_1]\tilde{g}\right)}{\phi(E[t_1]\tilde{g})}, \quad \text{and} \quad \frac{E[t_1]}{\tilde{g}}\frac{\Phi\left(E[t_1]\tilde{g}\right)}{\phi(E[t_1]\tilde{g})} \geq E[t_1]\frac{\Phi\left(E[t_1]\right)}{\phi(E[t_1])} \quad \text{if } E[t_1] \geq 0.84.$$

Here the first inequality follows because $\Phi(x)/\phi(x)$ is increasing in x, and $g \geq \tilde{g}$, and the second inequality follows because $\frac{\Phi(x)}{x\phi(x)}$ is increasing for $x \geq 0.84$, and $\tilde{g} \geq 1$. Therefore,

$$m_0 \leq \frac{\phi(E[t_1])}{E[t_1]\Phi(E[t_1]) + \phi(E[t_1])} = \lim_{\omega\downarrow 0} m_0(\omega),$$

where the equality follows since the right-hand side of eq. (A.11) converges to $\frac{\phi(E[t_1])}{E[t_1]\Phi(E[t_1])+\phi(E[t_1])}$ as $\omega \to 0$.

We now prove the claims concerning the unconditional distribution of $\tilde{\beta}_{IV}$. By arguments in the proof of Theorem 1, the median is smaller than 1, so it suffices to consider $p_{-\infty}(x)$ for $x \leq 1$. By arguments as in the conditional case,

$$\begin{aligned}
p_{-\infty}(x) &= \frac{1}{2} + \frac{\omega}{\tilde{\omega}^2(1-x)}\left[\int_{-\infty}^{\tilde{\omega}E[t_1]} y\phi\left(y\right)\phi(a+by)dy - \int_{\tilde{\omega}E[t_1]}^{\infty} y\phi\left(y\right)\phi(a+by)dy\right] \\
&= \frac{1}{2} + \frac{\omega}{\tilde{\omega}^2(1-x)}\frac{\phi(a/t)}{1+b^2}\left[-2\phi(t\tilde{\omega}E[t_1]+ab/t) - 2\frac{ab}{t}\Phi(t\tilde{\omega}E[t_1]+ab/t) + \frac{ab}{t}\right] \\
&= \frac{1}{2} + \frac{\omega}{\tilde{\omega}^2(1-x)}\frac{\phi(a/t)}{1+b^2}\left[-2\phi(E[t_1]g(\tilde{\omega},x)) + 2\frac{x}{1-x}\frac{E[t_1]}{\tilde{g}}\Phi(E[t_1]g(\tilde{\omega},x)) - \frac{x}{1-x}\frac{E[t_1]}{\tilde{g}}\right].
\end{aligned}$$

Here the first line follows by the mean value theorem, where $\tilde{\omega} = \tilde{\omega}(x,\omega) \in [0,\omega]$, the second line uses eq. (A.10), and the last line follows by algebra. When evaluated at $x = m_{-\infty}$, the expression in square brackets must equal zero by definition of the median. Therefore, $m_{-\infty} > 0$, and it must satisfy

$$m_{-\infty} = \frac{1}{\frac{E[t_1]}{\tilde{g}}\frac{\Phi(E[t_1]g)-1/2}{\phi(E[t_1]g)}+1} \tag{A.12}$$

Now,

$$\frac{E[t_1]}{\tilde{g}}\frac{\Phi(E[t_1]g)-1/2}{\phi(E[t_1]g)} \geq \frac{E[t_1]}{\tilde{g}}\frac{\Phi(E[t_1]\tilde{g})-1/2}{\phi(E[t_1]\tilde{g})} \geq E[t_1]\frac{\Phi(E[t_1])-1/2}{\phi(E[t_1])}.$$

Here the first inequality follows because $\Phi(x)/\phi(x)$ is increasing in x, and $g \geq \tilde{g}$, and the second inequality follows because $\frac{\Phi(x)-1/2}{x\phi(x)}$ is increasing for $x > 0$. As a result,

$$m_{-\infty} \leq \frac{\phi(E[t_1])}{E[t_1](\Phi(E[t_1]) - 1/2 + \phi(E[t_1]))} = \lim_{\omega\downarrow 0} m_{-\infty}(\omega),$$

where the equality follows since the right-hand side of eq. (A.12) converges to $\frac{\phi(E[t_1])}{E[t_1](\Phi(E[t_1])-1/2+\phi(E[t_1]))}$

as $\omega \to 0$.

A.6 Proof of Theorem 3

We may write

$$\tilde{\beta}_U = t_1\mu(t_1)\tilde{\beta}_{IV} + (1 - t_1\mu(t_1))\,\mathrm{sign}(\rho) = \mu(t_1)\frac{t_{AR}}{|\rho|} + (1 - t_1\mu(t_1))\,\mathrm{sign}(\rho) \tag{A.13}$$

where the first equality follows from eq. (17), and the fact that $\beta_{WOLS} - \beta$ and ρ have the same sign, and the second equality applies eq. (A.4).

Since $E[t_{AR} \mid t_1] = \rho(t_1 - E[t_1])$, the relative bias conditional on t_1 is given by

$$E[\tilde{\beta}_U \mid t_1] = \mathrm{sign}(\rho)\,[1 - E[t_1]\mu(t_1)]\,.$$

By arguments analogous to those in the proof of Lemma 2.1 in Andrews and Armstrong (2017), we have

$$\begin{aligned}
E[t_1]E[\mu(t_1) \mid t_1 > 0] &= \frac{E[t_1]}{\Phi(E[t_1])}\int_{t=0}^{\infty}\frac{1-\Phi(t)}{\phi(t)}\phi(t - E[t_1])dt = \frac{e^{-\frac{E[t_1]^2}{2}}}{\Phi(E[t_1])}\int_{t=0}^{\infty}(1-\Phi(t))E[t_1]e^{E[t_1]t}dt \\
&= \frac{e^{-E[t_1]^2/2}}{\Phi(E[t_1])}\left\{\left[e^{E[t_1]t}(1 - \Phi(E[t_1]))\right]_{t=0}^{\infty} + \int_{t=0}^{\infty}\phi(t)e^{E[t_1]t}dt\right\} \\
&= \frac{1}{\Phi(E[t_1])}\left[-\frac{1}{2}e^{-E[t_1]^2/2} + \int_{t=0}^{\infty}\phi(t - E[t_1])dt\right] = -\frac{1}{2}\frac{e^{-E[t_1]^2/2}}{\Phi(E[t_1])} + 1,
\end{aligned}$$

where the first line uses the definition of the Mills' ratio, the second line uses integration by parts, and the third follows by completing the square. It therefore follows that

$$\frac{E[\hat{\beta}_U - \beta \mid t_1 > 0]}{\beta_{WOLS} - \beta} = \frac{1}{2}\frac{e^{-E[t_1]^2/2}}{\Phi(E[t_1])}.$$

The second claim follows by an analogous argument.

A.7 Median Bias Comparisons

To evaluate the relative median bias of $\hat{\beta}_{IV}$ as a function of both $E[F]$ and ρ conditional on $t_1 \geq c$, we first evaluate its distribution

$$P(\tilde{\beta}_{IV} \leq x \mid t_1 \geq c; \rho, E[t_1]) = \frac{1}{\Phi(E[t_1] - c)}\int_{c-E[t_1]}^{\infty} f_{IV}(z; x, E[t_1], \omega)\phi(z)dz \tag{A.14}$$

by numerical integration. Here we use the formula $f_{IV}(z; x, \rho, E[t_1]) = \Phi(\omega[\mathrm{sign}(z + E[t_1])E[t_1] - (1 - x)|z + E[t_1]|])$ from eq. (A.8) for the cdf conditional on $z = t_1 - E[t_1]$. We then numerically solve for the median. The shaded regions in Figure 2 correspond to the range of the absolute value of the relative median bias as ρ varies between -1 and 1. Similarly, the shaded regions in Figure 3 show how the median bias conditional on $t_1 \geq 0$ relative to the unconditional median bias (that sets $c = -\infty$) varies with ρ.

To compare the relative median bias to that of $\hat{\beta}_U$, it suffices to consider $\rho > 0$, since the distributions of $\hat{\beta}_U$ and $\hat{\beta}_{IV}$ are symmetric in ρ. By eq. (A.13), it follows that for $t_1 > 0$,

$$P(\tilde{\beta}_U \leq x \mid t_1; \omega) = P\left(\tilde{\beta}_{IV} \leq x - (1-x)\frac{(1 - t_1\mu(t_1))}{t_1\mu(t_1)} \;\Big|\; t_1; \omega\right),$$

which for $x < 1$ is smaller than $P(\tilde{\beta}_{IV} \leq x \mid t_1; \omega)$. Since the median of $\tilde{\beta}_{IV}$ conditional on $t_1 > 0$ is smaller than 1, it follows that the conditional median bias of $\tilde{\beta}_{IV}$ is always smaller than that of $\tilde{\beta}_U$.

To compare the relative magnitudes of the median biases, we compute the relative median bias of $\tilde{\beta}_U$ analogously to that of $\tilde{\beta}_{IV}$, except we replace f_{IV} in eq. (A.14) with $f_U(z; x, E[t_1], \rho) = \Phi(\omega[E[t_1] - (1 - \operatorname{sign}(\omega)x)/\mu(E[t_1] + z)])$ (it follows from eqs. (A.6) and (A.13) that this is the cdf $\tilde{\beta}_U$ conditional on $z = t_1 - E[t_1]$). We then compute the ratio $\operatorname{median}_{E[t_1],\rho}(\tilde{\beta}_U \mid t_1 > 0)/\operatorname{median}_{E[t_1],\rho}(\tilde{\beta}_{IV} \mid t_1 > 0)$ of the median biases on a fine grid of values of $(\rho, E[t_1])$. This ratio is greater than 2 if $E[F] \geq 2$, and greater than 3 if $E[F] \geq 3$, regardless of the value of ρ. Likewise, comparison of the ratio of the conditional and unconditional median IV bias, $\operatorname{median}_{E[t_1],\rho}(\tilde{\beta}_{IV} \mid t_1 > 0)/\operatorname{median}_{E[t_1],\rho}(\tilde{\beta}_{IV}) = \operatorname{median}_{E[t_1],\rho}(\hat{\beta}_{IV} - \beta \mid t_1 > 0)/\operatorname{median}_{E[t_1],\rho}(\hat{\beta}_{IV} - \beta)$ shows that the ratio lies between 0.5 and 0.525 for $E[t_1] \geq 1.5$, regardless of the value of ρ.