Consistency without Inference: Instrumental Variables in Practical Application*

Alwyn Young
London School of Economics
This draft: March 2019

Abstract

I use Monte Carlo simulations, the jackknife and multiple forms of the bootstrap to study a comprehensive sample of 1359 instrumental variables regressions in 31 papers published in the journals of the American Economic Association. Monte Carlo simulations based upon published regressions show that non-iid error processes adversely affect the size and power of IV estimates, while increasing the bias of IV relative to OLS, producing a very low ratio of power to size and mean squared error that is almost always larger than biased OLS. Weak instrument pre-tests based upon F-statistics are found to be largely uninformative of both size and bias. In published papers, statistically significant IV results generally depend upon only one or two observations or clusters, excluded instruments often appear to be irrelevant, there is little statistical evidence that OLS is biased, and IV confidence intervals almost always include OLS point estimates.

*I am grateful to David Broadstone, Brian Finley and anonymous referees for valuable suggestions, and to Ruoqi Zhou for excellent research assistance.

I: Introduction

The economics profession is in the midst of a "credibility revolution" (Angrist and Pischke 2010) in which careful research design has become firmly established as a necessary characteristic of applied work. A key element in this revolution has been the use of instruments to identify causal effects free of the potential biases carried by endogenous ordinary least squares regressors. The growing emphasis on research design has not gone hand in hand, however, with equal demands on the quality of inference. Despite the widespread use of Eicker (1963)-Hinkley (1977)-White (1980) robust and clustered covariance estimates, the implications of non-iid error processes for the quality of inference, and their interaction in this regard with regression and research design, has not received the attention it deserves. Heteroskedastic and correlated errors produce test statistics whose dispersion is typically much greater than believed, particularly in highly leveraged regressions, which is the dominant feature of regression design in published papers. This adversely affects inference in both ordinary least squares (OLS) and two stage least squares (hereafter, 2SLS or IV), but more so in the latter, where confidence in results depends upon an assessment of the strength of both first and second stage relations.

In this paper I use Monte Carlos, the jackknife and multiple forms of the bootstrap to study the distribution of coefficients and test statistics in a comprehensive sample of 1359 2SLS regressions in 31 papers published in the journals of the American Economic Association. Subject to some basic rules regarding data and code availability and methods applied, I use all papers produced by a keyword search on the AEA website. I maintain, throughout, the exact specification used by authors and their identifying assumption that the excluded instruments are orthogonal to the second stage residuals. When bootstrapping, jackknifing or generating artificial residuals for Monte Carlos, I draw samples in a fashion consistent with the error dependence within groups of observations and independence across observations implied by authors' standard error calculations. Thus, this paper is not about point estimates or the validity of fundamental assumptions, but rather concerns itself with the quality of inference within the framework laid down by authors themselves.

Monte Carlos, using the regression design found in my sample and artificial error disturbances with a covariance structure matching that observed in 1st and 2nd stage residuals, show how non-iid errors damage the relative quality of inference using 2SLS. Non-iid errors weaken 1st stage relations, raising the relative bias of 2SLS and generating mean squared error that is larger than biased OLS in almost all published papers. Non-iid errors also increase the probability of spuriously large test statistics when the instruments are irrelevant, particularly in highly leveraged regressions and particularly in joint tests of coefficients, i.e. 1st stage F tests. Consequently, while 1st stage relations weaken, 1st stage pre-tests become uninformative, providing little or no protection against 2SLS size distortions or bias. 2SLS standard error estimates become more volatile and tail values of the t-statistic are dominated by unusually low realizations of the standard error rather than deviations of mean effects from the null. In the top third most highly leveraged papers in my sample, the ratio of power to size approaches one, i.e. 2SLS is scarcely able to distinguish between a null of zero and the alternative of the mean effects found in published tables.

Monte Carlos show, however, that the jackknife and (particularly) the bootstrap allow for 2SLS and OLS inference with accurate size and a much higher ratio of power to size than achieved using clustered/robust covariance estimates. Thus, while the bootstrap does not undo the increased bias of 2SLS brought on by non-iid errors, it nevertheless allows for improved inference under these circumstances. Inference using conventional standard errors in 2SLS is based on an estimate of a moment that in finite samples often does not exist (as the coefficient has no finite variance when exactly identified). Not surprisingly, the bootstrap's use of resampling to estimate the percentiles of distributions, which always exist, does much better. While asymptotic theory favours the resampling of the t-statistic, I find that avoiding the finite sample 2SLS standard estimate altogether and focusing on the bootstrap resampling of the coefficient distribution alone provides the best performance, with tail rejection probabilities on IV coefficients that are very close to nominal size in iid, non-iid, low and high leverage settings.

When published results are examined through the lens of the jackknife and bootstrap, a number of weaknesses are revealed. In published papers, about ½ to ⅔ of .01 significant IV results depend upon only one or two outlier observations or clusters and rest upon a finding of unusually small standard errors rather than surprising (under the null) mean effects. First stage relations, when re-examined through the jackknife or bootstrap, cannot reject the null that the excluded instruments are irrelevant in about ¼ to ⅓ of cases, while jackknifed and bootstrapped Durbin (1954) - Wu (1973) - Hausman (1978) tests find little evidence that OLS is substantively biased, despite large proportional and frequent sign differences between OLS and 2SLS point estimates, as 2SLS estimation is found to be so inaccurate that 2SLS confidence intervals almost always include OLS point estimates. In sum, whatever the biases of OLS may be, in practical application with non-iid error processes and highly leveraged regression design, the performance of 2SLS methods deteriorates so much that it is rarely able to identify parameters of interest more accurately or substantively differently than is achieved by OLS. The third of my sample with the lowest maximum observational leverage does better on all metrics, but even here only ½ of papers provide any regressions with results that are significantly distinguishable from OLS.

The concern with the quality of inference in 2SLS raised in this paper is not new. Sargan, in his seminal 1958 paper, raised the issue of efficiency and the possibility of choosing the biased but more accurate OLS estimator, leading later scholars to explore relative efficiency in Monte Carlo settings (e.g. Summers 1965, Feldstein 1974). The current professional emphasis on first stage F-statistics as pre-tests originates in Nelson and Startz (1990a, b), who used examples to show that size distortions can be substantial when the strength of the first stage relationship is weak, and Bound, Jaeger and Baker (1995), who emphasized problems of bias and inconsistency with weak instruments. These papers spurred path-breaking research, such as Staiger and Stock (1997) and Stock and Yogo's (2005) elegant derivation and analysis of weak instrument asymptotic distributions, renewed interest (e.g. Dufour 2003, Baum, Schaffer and Stillman 2007, Chernozhukov and Hansen 2008) in older weak instrument robust methods such as that of Anderson-Rubin (1949), and motivated the use of such techniques in critiques of

selected papers (e.g. Albouy 2012, Bazzi and Clemens 2013). The theoretical and Monte Carlo work that motivates this literature is largely iid based, a notable exception being Olea & Pflueger (2013), who argue that heteroskedastic error processes weaken 1st stage relations and, based upon asymptotic approximations, propose a bias test closely related to the 1st stage clustered/ robust F-statistic. This paper supports Olea & Pflueger's insight that non-iid errors effectively weaken 1st stage relations, brings to the fore earlier concerns regarding the practical relative efficiency of 2SLS, shows that iid-motivated weak instrument pre-tests and weak instrument robust methods perform poorly when misapplied in non-iid settings, and highlights the errors induced by finite sample inference using asymptotically valid clustered/robust covariance estimates in highly leveraged settings, including even the Olea & Pflueger bias test.

The paper proceeds as follows: After a brief review of notation in Section II, Section III describes the rules used to select the sample and its defining characteristics, highlighting the presence of non-iid errors, high leverage and sensitivity to outliers. Section IV presents Monte Carlos patterned on the regression design and error covariance found in my sample, showing how non-iid errors worsen inference of all sorts, but especially degrade the ratio of power to size in IV tests while raising the relative bias of 2SLS estimation. 1st stage pre-tests are found to be largely uninformative, although the Olea & Pflueger bias test effectively separates low and high bias in low leverage over-identified 2SLS regressions. Section V provides a thumbnail review of jackknife and "pairs" and "wild" bootstrap methods. The pairs resampling of the coefficient distribution is found to provide a low cost means of performing multiple 1st and 2nd stage tests with tail rejection probabilities as accurate as other methods and very close to nominal size in tests of IV coefficients in particular. Section VI re-examines the 2SLS regressions in my sample using all of the jackknife and bootstrap methods, finding the results mentioned above, while Section VII concludes. An on-line appendix provides alternative versions of many tables, further comparison of the accuracy of different bootstrap methods, and Monte Carlos of some popular weak instrument robust methods, showing that these, which do not directly address the issue of non-iid errors, are not panaceas and in some cases perform much worse than 2SLS.

All of the results of this research are anonymized. Thus, no information can be provided, in the paper, public use files or private conversation, regarding results for particular papers. Methodological issues matter more than individual results and studies of this sort rely upon the openness and cooperation of current and future authors. For the sake of transparency, I provide complete code (in preparation) that shows how each paper was analysed, but the reader eager to know how a particular paper fared will have to execute this code themselves.

II. Notation and Formulae

It is useful to begin with some notation and basic formulae, to facilitate the discussion which follows. With bold lowercase letters indicating vectors and bold uppercase letters matrices, the data generating process is taken as given by:

$$(1)\quad \mathbf{y} = \mathbf{Y}\beta + \mathbf{X}\boldsymbol{\delta} + \mathbf{u}, \quad \mathbf{Y} = \mathbf{Z}\boldsymbol{\pi} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{v},$$

where **y** is the n x 1 vector of second stage outcomes, **Y** the n x 1 matrix of endogenous regressors, **X** the n x k_X matrix of included exogenous regressors, **Z** the n x k_Z matrix of excluded exogenous regressors (instruments), and **u** and **v** the n x 1 vectors of second and first stage disturbances. The remaining (Greek) letters are parameters, with β representing the parameter of interest. Although in principal there might be more than one endogenous right-hand side variable, i.e. **Y** is generally n x k_Y, in practical work this is exceedingly rare (see below) and this paper focuses on the common case where k_Y equals 1.

The nuisance variables **X** and their associated parameters are of no substantive interest, so I use ˜ to denote the residuals from the projection on **X** and characterize everything in terms of these residuals. For example, with ^ denoting estimated and predicted values, the coefficient estimates for OLS and 2SLS are given by:

$$(2)\quad \hat{\beta}_{ols} = (\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}'\tilde{\mathbf{y}}, \quad \hat{\beta}_{2sls} = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\tilde{\mathbf{Y}}}'\tilde{\mathbf{y}}, \quad \text{where } \hat{\tilde{\mathbf{Y}}} = \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{Y}}$$

Finally, although the formulae are well known, to avoid any confusion it is worth spelling out that in referring to "homoskedastic" or "default" covariance estimates below I mean

$$(3)\quad V(\hat{\beta}_{ols}) = (\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\hat{\sigma}_u^2, \quad V(\hat{\beta}_{2sls}) = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\sigma}_u^2, \quad \& \quad \mathrm{V}(\hat{\boldsymbol{\pi}}) = (\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\hat{\sigma}_v^2,$$

where $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ equal the sum of the (OLS or 2SLS) squared residuals divided by n minus the k right hand side variables, while in the case of "clustered/robust" covariance estimates I mean:

$$(4)\quad V(\hat{\beta}_{ols}) = \frac{c\sum_{\mathbf{i}\in\mathbf{I}} \tilde{\mathbf{Y}}_\mathbf{i}'\hat{\mathbf{u}}_\mathbf{i}\hat{\mathbf{u}}_\mathbf{i}'\tilde{\mathbf{Y}}_\mathbf{i}}{(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^2}, \quad V(\hat{\beta}_{2sls}) = \frac{c\sum_{\mathbf{i}\in\mathbf{I}} \hat{\tilde{\mathbf{Y}}}_\mathbf{i}'\hat{\mathbf{u}}_\mathbf{i}\hat{\mathbf{u}}_\mathbf{i}'\hat{\tilde{\mathbf{Y}}}_\mathbf{i}}{(\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^2}, \quad \& \quad V(\hat{\boldsymbol{\pi}}) = c\sum_{\mathbf{i}\in\mathbf{I}} (\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}_\mathbf{i}'\hat{\mathbf{v}}_\mathbf{i}\hat{\mathbf{v}}_\mathbf{i}'\tilde{\mathbf{Z}}_\mathbf{i}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}$$

where **i** denotes the group of clustered observations (or individual observations when merely robust), **I** the set of all such groupings, $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ the second and first stage residuals, *c* a finite sample adjustment (e.g. n/(n-k) in the robust case), and where I have made use of the fact that the inner-product of **Y** is a scalar.

III. The Sample

My sample is based upon a search on www.aeaweb.org using the keyword "instrument" covering the American Economic Review and the American Economic Journals for Applied Economics, Economic Policy, Microeconomics and Macroeconomics which, at the time of its implementation, yielded papers up through the July 2016 issue of the AER. I then dropped papers that:

(a) did not provide public use data files and Stata do-file code;
(b) did not include instrumental variables regressions;
(c) used non-linear methods or non-standard covariance estimates;
(d) provided incomplete data or non-reproducible regressions.

Public use data files are necessary to perform any analysis, and I had prior experience with Stata and hence could analyse do-files for this programme at relatively low cost. Stata is by far the most popular programme as, among papers that provide data, only five make use of other software. The keyword search brought up a number of papers that deal with instruments of policy, rather than instrumental variables, and these were naturally dropped from the sample.

Conventional linear two stage least squares with either the default or clustered/robust covariance estimate is the overwhelmingly dominant approach, so I dropped the exceedingly rare deviations. This included four papers that used non-linear IV methods, uniquely clustered on two variables or used auto-correlation consistent standard errors, as well as a small handful of

GMM regressions in two papers whose 2SLS regressions are otherwise included in the sample. There is little to be learnt or generalized from a handful of specifications, and clustered/robust linear IV is, almost without exception, the industry practice.

Many papers provide partial data, indicating that users should apply to third parties for confidential data necessary to reproduce the analysis. As the potential delay and likelihood of success in such applications is indeterminate, I dropped these papers from my sample. I took as my sample only IV regressions that appear in tables. Alternative specifications are sometimes discussed in surrounding text, but catching all such references and linking them to the correct code is extremely difficult. By limiting myself to specifications presented in tables, I was able to use coefficients, standard errors and supplementary information like sample sizes and test statistics to identify, interpret and verify the relevant parts of authors' code. Cleaning of the sample based upon the criteria described above produced 1400 2SLS regressions in 32 papers. Only 41 of these, however, contain more than one endogenous right hand side variable. As 41 observations are insufficient to draw meaningful conclusions, I further restricted the analysis to regressions with only one endogenous variable.

As shown in Table I below, the final sample consists of 31 papers, 16 appearing in the American Economic Review and 15 in other AEA journals. Of the 1359 IV regressions in these papers, 1087 are exactly identified by one excluded instrument and 272 are overidentified. When equations are overidentified, the number of instruments can be quite large, with an average of 17 excluded instruments (median of 5) in 13 papers. Thus, econometric issues concerning the higher dimensionality of instruments are relevant in a substantial subset of equations and papers. Although instrumental variables regressions are central to the argument in all of these papers, with the keyword "instrument" appearing in either the abstract or the title, the actual number of IV regressions varies greatly, with 5 papers presenting 98 to 286 such regressions, while 9 have only between 2 to 10. In consideration of this and the fact there is a great deal of similarity within papers in regression design, in presenting averages in tables and text below unless

Table I: Characteristics of the Sample

31 papers		1359 2SLS regressions			
journal	# of 2SLS regressions	excluded instruments	covariance estimate	distribution	
16 AER	9 2-10	1087 1	105 default	753	t & F
6 AEJ: A. Econ.	9 11-26	138 2-5	1039 clustered	606	N & chi^2
4 AEJ: E. Policy	8 35-72	134 6-60	215 robust		
5 AEJ: Macro	5 98-286				

Notes: AER = American Economic Review; AEJ = American Economic Journal, Applied Economics, Economic Policy and Macro; t, F, N & chi^2 = t, F, standard normal and chi-squared distributions.

otherwise noted I always take the average *across* papers of the *within* paper average. Thus, each paper carries an equal weight in determining summary results.

Turning to statistical inference, all but one of the papers in my sample use the Eicker (1963)-Hinkley (1977)-White (1980) robust covariance matrix or its multi-observation cluster extension. Different Stata commands make use of different distributions to evaluate the significance of the same 2SLS estimating equation, with the sample divided roughly equally between results evaluated using the t and F (with finite sample covariance corrections) and those evaluated using the normal and chi^2. In directly evaluating authors' results, I use the distributions and methods they chose. For more general comparisons, however, I move everything to a consistent basis, using clustered/robust[1] covariance estimates and the t and F distributions for all 2SLS and OLS results.

Table II shows that non-normality, intra-cluster correlation and heteroskedasticity of the disturbances are important features of the data generating process in my sample. Using Stata's test of normality based upon skewness and kurtosis, I find that in the average paper more than 80% of the OLS regressions which make up the 2SLS point estimates reject the null that the residuals are normal. In equations which cluster, cluster fixed effects are also found to be significant more than 80% of the time. In close to ½ of these regressions the authors' original

[1]I use the robust covariance estimate for the one paper that used the default covariance estimate throughout, and also cluster four regressions that were left unclustered in papers that otherwise clustered all other regressions.

Table II: Tests of Normality, Cluster Correlation and Heteroskedasticity
(average across 31 papers of fraction of regressions rejecting the null)

	Y on **Z**, **X** (1st stage)		**y** on **Z**, **X** (reduced form)		**y** on **Y**, **X** (OLS)	
	.01	.05	.01	.05	.01	.05
normally distributed residuals	.807	.831	.811	.882	.830	.882
no cluster fixed effects	.846	.885	.855	.899	.835	.879
homoscedastic (K) on **Z**, **X** or **Y**, **X**	.739	.806	.635	.698	.657	.727
homoskedastic (W) on **Z**, **X** or **Y**, **X**	.743	.807	.660	.714	.677	.737

Notes: .01/.05 = level of the test. K = Koenker 1981 and W = Wooldridge 2013. Cluster fixed effects only calculated for papers which cluster and where the dependent variable varies within clusters. Where authors weight I use the weights to remove the known heteroskedasticity in the residuals before running the tests.

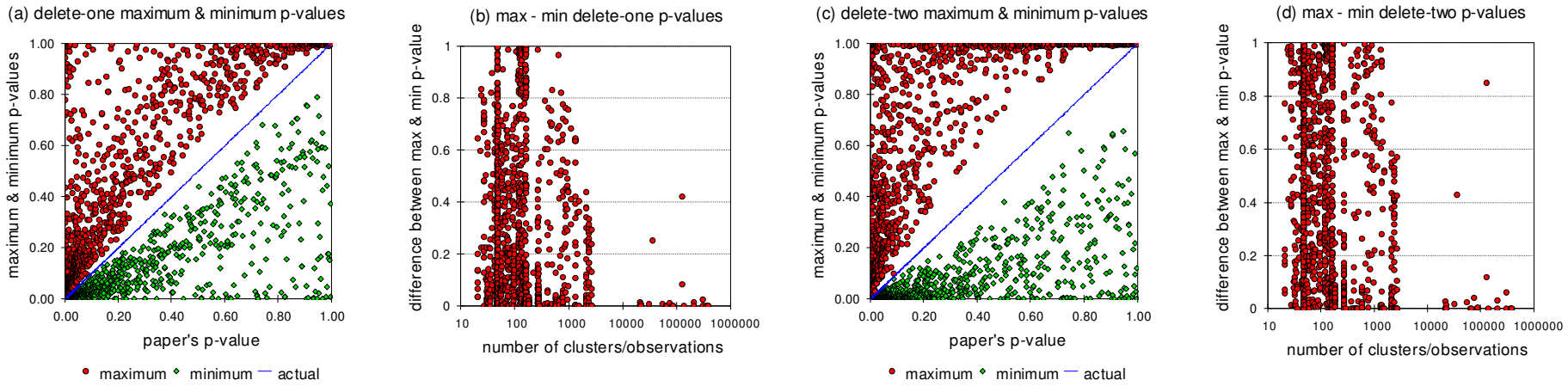
specification includes cluster fixed effects, but where there is smoke there is likely to be fire, i.e. it is unlikely that the cluster correlation of residuals is limited to a simple mean effect; a view apparently shared by the authors, as they cluster standard errors despite including cluster fixed effects. Tests of homoskedasticity involving the regression of squared residuals on the authors' right-hand side variables using the test statistics and distributions suggested by Koenker (1981) and Wooldridge (2013) reject the null between ⅔ and .80 of the time. In Monte Carlos further below I find that while non-normality is of relatively little import, heteroskedasticity and intra-cluster correlation seriously degrade the performance of 2SLS relative to OLS and render existing 1st stage pre-tests uninformative.

The other defining characteristic of published IV results is their extraordinary sensitivity to outliers. Panel a of Figure I below graphs the maximum and minimum p-values that can be found by deleting one cluster or observation in each regression in my sample against the authors' p-value for that instrumented coefficient.[2] With the removal of just one cluster or observation, in the average paper .49 of reported .01 significant 2SLS results can be rendered insignificant at that level, with the average p-value when such changes occur rising to .071. With the deletion of

[2]I use authors' methods to calculate p-values and where authors cluster, I delete clusters, otherwise I delete individual observations. All averages reported in the paragraph above, as elsewhere in the paper, refer to the average across papers of the within paper average measure.

Figure I: Sensitivity of P-Values to Outliers (Instrumented Coefficients)

2SLS



OLS

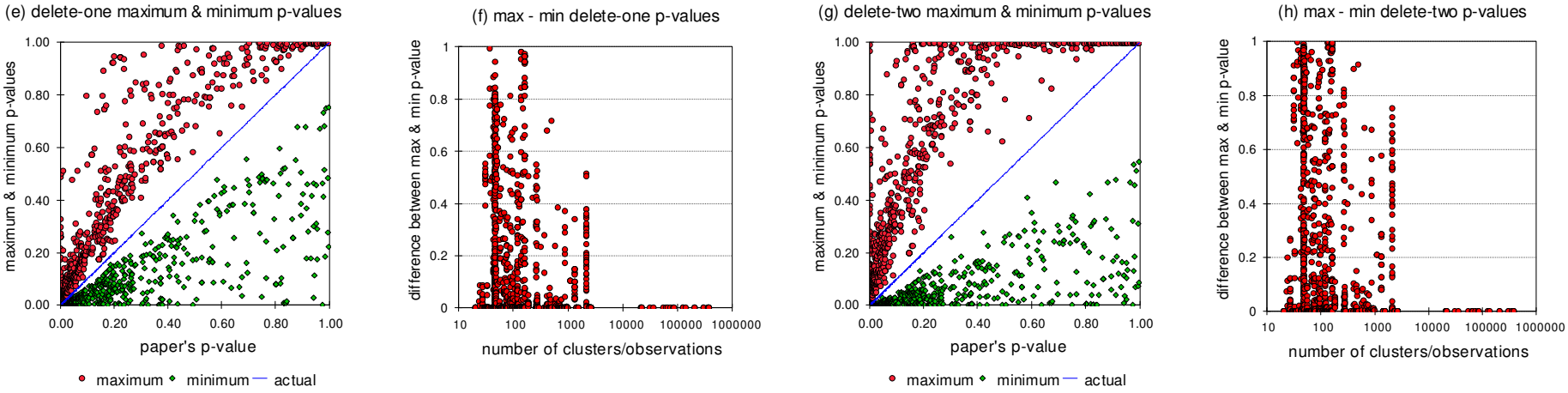
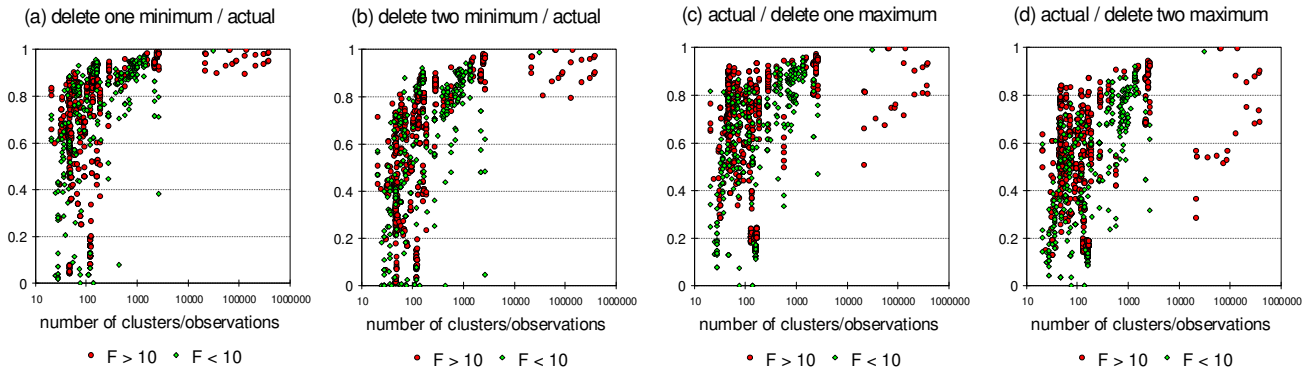


Figure II: Proportional Change of First Stage F with Removal of One or Two Clusters or Observations



two observations (panel c), in the average paper no less[3] than .66 of .01 significant IV results can be rendered insignificant, with the average p-value when such changes occur rising to .149. Conversely, in the average paper .29 and .45 of .01 insignificant IV results can be rendered .01 significant with the removal of one or two clusters or observations, respectively. As panels a and c show, the changes can be extraordinary, with p-values moving from close to 0 to near 1.0, and vice-versa. Not surprisingly, the gap between maximum and minimum delete -one and -two IV p-values is decreasing in the number of clusters or observations, as shown in panels b and d, but very large max-min gaps remain common even with 1000s of clusters and observations. Figure I also reports the sensitivity of the p-values of OLS versions of the authors' estimating equations (panels e through h). Insignificant OLS results are found to be similarly sensitive to outliers, as in the average paper .33 and .47 of .01 insignificant results can be rendered significant with the removal of one or two clusters or observations, respectively. Significant OLS results, however, are more robust, with an average of only .26 and .39 of .01 significant results showing delete-one or -two sensitivity, respectively.

In my sample the F-statistics that authors use to assure readers of the strength of the 1st stage relation are also very sensitive to outliers. Figure II graphs the ratio of the minimum clustered/robust F-statistic found by deleting one or two clusters or observations to the full

[3]"No less" because computation costs prevent me from calculating all possible delete-two combinations. Instead, I delete the cluster/observation with the maximum or minimum delete-one p-value and then calculate the maximum or minimum found by deleting one of the remaining clusters/observations.

sample F (panels a and b) and the ratio of the full sample F to the maximum delete-one or -two F (panels c and d). With the removal of just one or two observations, the average paper F can be lowered to .72 and .58 of its original value, respectively, or increased to the point that the original value is just .69 or .56, respectively, of the new delete-one or -two F. Fs greater than 10 are proportionately somewhat more robust to outliers than Fs less than 10,[4] but, as before, substantial sensitivity can be found even in samples with thousands, if not hundreds of thousands, of observations/clusters.

The delete-one or –two sensitivity of p-values and F-statistics in my sample arises from the concentration of "leverage" in a few clusters and observations. Consider the generic regression of a vector **y** on a matrix of regressors **X**. The change in the estimated coefficient for a particular regressor **x** brought about by the deletion of the vector of observations **i** is given by:

$$(5)\quad \hat{\beta}_{\sim \mathbf{i}} - \hat{\beta} = -\tilde{\mathbf{x}}_{\mathbf{i}}'\boldsymbol{\varepsilon}_{\mathbf{i}} / \tilde{\mathbf{x}}'\tilde{\mathbf{x}}$$

where $\tilde{\mathbf{x}}$ is the vector of residuals of **x** projected on the other regressors, $\tilde{\mathbf{x}}_{\mathbf{i}}$ the **i** elements thereof, and $\boldsymbol{\varepsilon}_{\mathbf{i}}$ the vector of residuals for observations **i** calculated using the delete-**i** coefficient estimates. The default and clustered/robust covariance estimates are of course given by:

$$(6)\ \text{default}: \frac{1}{n-k}\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}};\ \text{clustered/ robust}: \frac{c\sum_{\mathbf{i}} \tilde{\mathbf{x}}_{\mathbf{i}}'\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}'\tilde{\mathbf{x}}_{\mathbf{i}}}{(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^2}$$

Define $\boldsymbol{\varepsilon}_{\mathbf{i}}'\boldsymbol{\varepsilon}_{\mathbf{i}} / \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, $\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}'\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}} / \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ and $\tilde{\mathbf{x}}_{\mathbf{i}}'\tilde{\mathbf{x}}_{\mathbf{i}} / \tilde{\mathbf{x}}'\tilde{\mathbf{x}}$ as the group **i** shares of squared delete-**i** residuals, squared actual residuals, and "coefficient leverage",[5] respectively. Clearly, coefficients, standard errors and t-statistics will be more sensitive to the deletion of individual observations when these shares are uneven, i.e. concentrated in a few observations.

Table III summarizes the maximum coefficient leverage and residual shares found in my sample. In the 1st stage and reduced form projections on the excluded instruments **Z**, in the

[4]The average paper delete one (two) min ratio is .68 (.52) for Fs < 10 and .73 (.60) for Fs > 10, while the average delete-one (two) max ratio is .61 (.46) for Fs < 10 and .71 (.58) for Fs > 10.

[5]So called since "leverage" is typically defined as the diagonal elements of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ formed using all regressors, while the measure described above is the equivalent for the partitioned regression on $\tilde{\mathbf{x}}$.

Table III: Largest Shares of Coefficient Leverage & Squared Residuals

	coefficient leverage		residuals: **Y** on **Z** (1st stage)		residuals: **y** on **Z** (reduced form)		residuals: **y** on **Y** (OLS)		delete-one/two sensitivity of < .01 p-values	
	$\frac{\tilde{\mathbf{Z}}_i'\tilde{\mathbf{Z}}_i}{\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}}$	$\frac{\tilde{\mathbf{Y}}_i'\tilde{\mathbf{Y}}_i}{\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}}$	$\frac{\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i}{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}$	$\frac{\hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_i}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}$	$\frac{\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i}{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}$	$\frac{\hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_i}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}$	$\frac{\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i}{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}$	$\frac{\hat{\boldsymbol{\varepsilon}}_i'\hat{\boldsymbol{\varepsilon}}_i}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}$	2SLS	OLS
	(a) all papers (average of 31 paper averages)									
one cl/obs	.17	.13	.14	.12	.12	.10	.12	.10	.49	.26
two cl/obs	.26	.20	.21	.19	.18	.17	.18	.17	.66	.39
	(b) low leverage (10 papers)									
one cl/obs	.04	.04	.05	.05	.05	.05	.05	.05	.27	.05
two cl/obs	.07	.07	.08	.08	.09	.08	.09	.08	.41	.09
	(c) medium leverage (11 papers)									
one cl/obs	.14	.13	.16	.12	.15	.14	.15	.14	.53	.28
two cl/obs	.26	.21	.25	.21	.24	.22	.24	.22	.67	.41
	(d) high leverage (10 papers)									
one cl/obs	.33	.23	.21	.18	.14	.12	.14	.12	.72	.51
two cl/obs	.46	.32	.31	.27	.21	.19	.21	.19	.95	.73

Notes: Reported figures, as elsewhere, are the average across papers of the within paper average measure; cl/obs = clusters or observations, depending upon whether the regression is clustered or not; delete-one/two sensitivity reports the share of .01 significant p-values that are sensitive to the deletion of one/two observations; four papers have no .01 significant p-values and are not included therein; low, medium and high divide the sample based upon the share of the largest cl/obs in **Z** leverage.

average paper the largest one or two clusters or observations on average account for .17 and .26 of total coefficient leverage, while the delete-**i** or estimated residual shares range from .10 to .14 and .17 to .19 for the largest one and two clusters/observations, respectively. With large outliers, in both regressors and residuals, the coefficients and covariance matrices of the 1st and 2nd stage are heavily influenced by one or two clusters or observations. Maximum leverage in the OLS projection on Y is somewhat smaller, but residuals are similarly large. The sample is, however, very heterogeneous, so I divide it into thirds based upon the average share of the largest cluster/observation in **Z** (instrument) coefficient leverage in each paper. While the average share of the largest cluster/observation is just under .04 in the 10 papers with the smallest maximum

leverage, it is .33 in the 10 papers with the largest maximum leverage. Maximum coefficient leverage in the OLS regression and the share of the largest one or two residual groupings move more or less in tandem with the shares of maximum **Z** leverage, reflecting perhaps the influence of **Z** on the endogenous regressor **Y** and the correlation of residuals with extreme values of the regressors noted earlier in Table II. As expected, delete-one and -two sensitivity varies systematically with maximum leverage as well.

Sensitivity to a change in the sample and accuracy of inference given the sample are related problems. When leverage is concentrated in a few observations, the clustered/robust covariance estimate places much of its weight on relatively few residuals, whose observed variance is reduced by the strong response of coefficient estimates to their realizations. This tends to produce downward biased standard error estimates with a volatility greater than indicated by the nominal degrees of freedom typically used to evaluate test statistics. The dispersion of test statistics increases when, in addition, heteroskedasticity is correlated with extreme values of the regressors, as the standard error estimate becomes heavily dependent upon the realizations of a few highly volatile residuals. Regressor correlated heteroskedasticity, however, arises very naturally. Random heterogeneity in the effects of righthand side variables, for example, mechanically generates heteroskedasticity that is correlated with those variables. It is precisely this correlation of heteroskedasticity with regressors that clustered/robust covariance estimates are supposed to correct for,[6] but, as shown below, in finite samples these asymptotically valid methods produce highly volatile covariance estimates and, consequently, test statistics with underappreciated thick tail probabilities.

IV. Monte Carlos: 2SLS and OLS in iid & non-iid Settings

In this section I explore the relative characteristics of 2SLS and OLS in iid and non-iid settings using Monte Carlos based on the practical regressions that appear in my sample. I begin

[6]It is easily seen, in (6) for example, that when leverage is uncorrelated with residuals, or leverage is even, so that the correlation is perforce zero, the robust covariance estimate reduces to the default estimate.

by estimating the coefficients and residuals of the 1st and 2nd stage equations using 2SLS and then calculating the Cholesky decomposition of the covariance matrix of the residuals:[7]

$$(7)\quad \mathbf{y} = \mathbf{Y}\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \hat{\mathbf{u}}, \quad \mathbf{Y} = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \hat{\mathbf{v}}, \quad \& \quad \mathbf{CC}' = \mathbf{V} = \frac{1}{n-k}\begin{bmatrix} \hat{\mathbf{u}}'\hat{\mathbf{u}} & \hat{\mathbf{u}}'\hat{\mathbf{v}} \\ \hat{\mathbf{v}}'\hat{\mathbf{u}} & \hat{\mathbf{v}}'\hat{\mathbf{v}} \end{bmatrix}.$$

I then generate independent random variables $\boldsymbol{\varepsilon_1}$ & $\boldsymbol{\varepsilon_2}$ drawn from a standardized distribution (i.e. demeaned and divided by its standard deviation), and artificial values of **y** and **Y** using the data generating process:

$$(8)\quad \mathbf{y} = \mathbf{Y}\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \mathbf{u}, \quad \mathbf{Y} = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \mathbf{v}, \quad \text{where} \quad [\mathbf{u},\mathbf{v}] = \mathbf{C}[\boldsymbol{\varepsilon_1},\boldsymbol{\varepsilon_2}].$$

I use six different data generating processes for the observation specific values (ε_i) of $\boldsymbol{\varepsilon_1}$ & $\boldsymbol{\varepsilon_2}$:

8.1. iid standard normal
8.2. iid standardized chi^2
8.3. heteroskedastic standard normal, where $\varepsilon_i = h_i\eta_i$, η ~ iid standard normal
8.4. heteroskedastic standardized chi^2, where $\varepsilon_i = h_i\eta_i$, η ~ iid standardized chi^2
8.5. heteroskedastic clustered standard normal, where $\varepsilon_i = h_i(\eta_i + \eta_c)/2^{½}$, η ~ iid standard normal
8.6. heteroskedastic clustered standardized chi^2, where $\varepsilon_i = h_i(\eta_i + \eta_c)/2^{½}$, η ~ iid standardized chi^2

A standardized chi^2 distribution ranges from -.7 to infinity, i.e. is decidedly skewed and non-normal. To produce heteroskedastic residuals, I use **h** equal to the sample standardized value of the first element **z** in **Z**. Heteroskedastic effects of this kind naturally arise when there is heterogeneity in the effects of **z** on **Y** and **Y** on **y**.[8] In modelling unaccounted for intracluster correlation, there is little point in using simple cluster random effects, as more than half of clustered regressions have cluster fixed effects. Instead, I model the cluster effect as representing iid cluster level draws in the heterogeneity of the impact of **z** on **Y** and **Y** on **y**, with the independent cluster (η_c) and observation specific (η_i) components carrying equal weight. By sample standardizing **z** and dividing by √2 when combining cluster and observation components I ensure that the covariance matrix of the disturbances remains unchanged and equal to the sample estimate **V** across the six data generating processes. In comparing 2SLS and OLS, it will

[7] I multiply the estimated covariance matrix by n/(n-k) to correct for the reduction in 1st stage residuals brought about by OLS fitting. There is no particular justification for or against multiplying the asymptotically valid 2nd stage residuals by anything, so, for simplicity, I multiply the entire covariance matrix by this adjustment.

[8] For example, let $\tilde{Y}_i = \tilde{z}_i(\pi + \pi_i) = \tilde{z}_i\pi + \tilde{z}_i\pi_i$ and $\tilde{y}_i = \tilde{Y}_i(\beta + \beta_i) = \tilde{Y}_i\beta + \tilde{z}_i(\pi + \pi_i)\beta_i$, where π_i and β_i are mean zero random variables that are independent of **z**.

be useful to consider cases where OLS is unbiased. To this end, I also run simulations in which I set the off-diagonal elements of **V** in (7) to 0. Such simulations are noted below as having "uncorrelated errors", as opposed to the "correlated errors" of the baseline analysis.

(a) 2SLS vs OLS: Inference and Accuracy

Table IV below begins by reporting the size and power of IV and OLS methods. Size is measured by calculating rejection rates at the .01 level of the null that the underlying parameter equals the $\hat{\beta}_{iv}$ used in the data generating process (8), while power tests the null that it equals 0.[9] In the frequentist world of starred results reported in economics journals only size should matter, but in the quasi-Bayesian way in which these results seem to be evaluated by the profession, power is also relevant. I run 1000 Monte Carlo simulations for each of the six data generating processes for each of the 1359 equations and, as usual, report cross paper averages of within paper average rejection rates. The first four columns of the table compare rejection rates using the default and clustered/robust IV covariance estimates when 1st and 2nd stage errors are correlated. These establish that, whatever the flaws of clustered/robust covariance calculations, their use is clearly preferred to the default covariance estimate, which produces slightly better size when errors are iid and gross size distortions when they are not. To be sure, the biases that produce large size distortions translate into greater power, so a tendency to over-reject, a weakness when the null is true, becomes a positive feature when the null is false. However, from the perspective of updating prior beliefs, it is the ratio of power to size that matters, and in this respect the default covariance matrix also performs very poorly in non-iid settings. To save space, in the presentation below I focus on results using clustered/robust covariance estimates.

Table IV reveals three interesting patterns. First, while IV has larger size distortions than OLS when errors are iid, in these simulations it actually provides for more accurate inference under true nulls when the disturbances are heteroskedastic and/or correlated within clusters. The

[9]In this paper I focus on the sensitivity and characteristics of results at the .01 level, as these are the types of results, as judged by the number of stars attached, that readers probably find most convincing. The on-line appendix presents versions of all tables at the .05 level. The patterns and issues that arise are very much the same.

Table IV: Average Rejection Rates at the .01 Level
(1000 Monte Carlo simulations for each of 1359 equations)

	correlated errors				uncorrelated errors			
	IV default		IV cluster/robust		IV cluster/robust		OLS cluster/robust	
	size	power	size	power	size	power	size	power
iid normal	.015	.437	.028	.458	.018	.469	.013	.835
iid chi^2	.016	.445	.026	.475	.017	.487	.013	.840
h. normal	.265	.529	.067	.278	.052	.296	.067	.658
h. chi^2	.266	.535	.075	.290	.064	.311	.085	.705
h. & cl. normal	.465	.611	.067	.176	.052	.194	.077	.572
h. & cl. chi^2	.467	.609	.080	.190	.065	.208	.095	.621

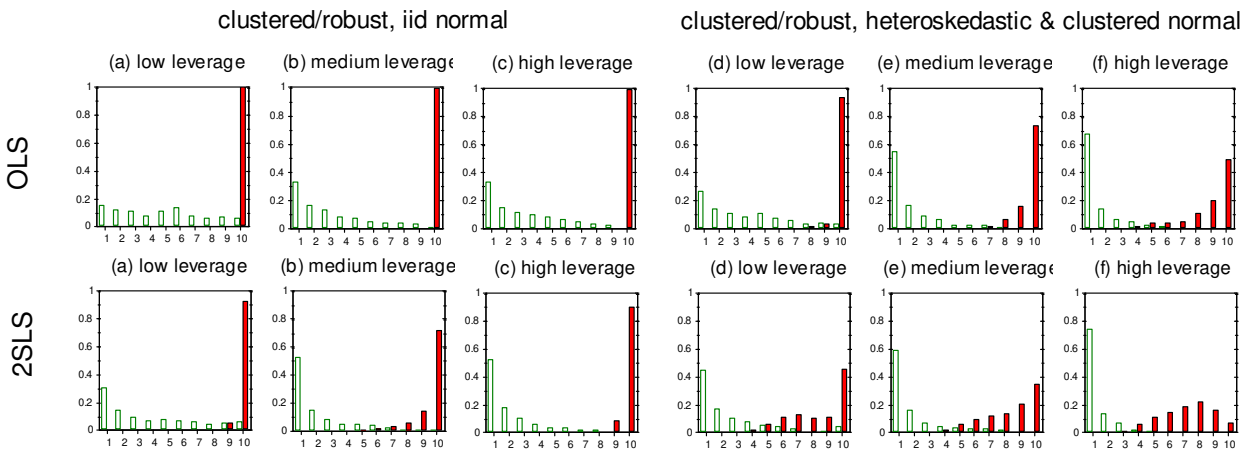
	IV cluster/robust (correlated errors)						OLS cluster/robust (uncorrelated errors)		
	low leverage		medium leverage		high leverage		low	medium	high
	size	power	size	power	size	power	size	size	size
iid normal	.009	.569	.034	.303	.040	.517	.010	.015	.013
iid chi^2	.011	.569	.031	.331	.035	.540	.009	.014	.014
h. normal	.011	.371	.051	.150	.142	.326	.013	.046	.143
h. chi^2	.012	.367	.063	.170	.152	.344	.021	.072	.164
h. & cl. normal	.010	.244	.051	.109	.142	.182	.017	.054	.164
h. & cl. chi^2	.018	.246	.064	.123	.159	.206	.028	.077	.182

Notes: Correlated and uncorrelated errors, here and elsewhere in Monte Carlos, refers to the relation between 1st and 2nd stage residuals, not to cross-correlations within clusters; default & cluster/robust refer to the covariance estimate and associated degrees of freedom used in the test; size and power as described in the text; h. and cl. refer to heteroskedastic and clustered data generating processes as described in 9.1-9.6 and associated text; low, medium and high leverage divide the sample based upon maximum **Z** leverage (as in Table III earlier).

theoretical and Monte Carlo discussion of weak instruments and their potential impact on size has focused on iid-settings. As one moves away from this ideal, however, the dominant problem becomes that inference of any sort, whether IV or OLS, may have large size biases, despite the use of clustered/robust covariance estimates. Second, the lower panel of Table IV shows that while size distortions for both OLS and IV can be very large in high leverage settings, in the low leverage sample, particularly for IV, they are very close to nominal size. Clustered/robust covariance estimates correct for non-iid error processes in finite samples, but only when maximal leverage is not too large.[10] Third, Table IV shows that the power of 2SLS, already substantially

[10]One might be tempted to conclude that these covariance estimates work well in low maximum leverage samples simply because they equal the default estimate when leverage is evenly distributed. This conclusion is

Figure III: Shares of Deciles of Coefficients & Standard Errors at .01 Tail of Squared t-statistic



Notes: Standard errors in hollow bars; absolute coefficient deviations in solid bars; clustered/robust = covariance estimate used.

lower than OLS in an iid world, declines more rapidly with non-iid errors. In the high leverage sample with clustered and heteroskedastic errors, IV power is only slightly above size, so from a Bayesian perspective a statistically significant result does little to shift beliefs from an effect of 0 in favour of the coefficient estimates reported by authors.

Figure III examines the roles played by coefficient and standard error estimates at the tail ends of the distributions of OLS and 2SLS t-statistics. For each of the 1000 Monte Carlo simulations in each of 1359 equations I calculate the deciles of the standard error estimates and the absolute deviation of the coefficient estimate from the true null. I then graph the density of these deciles when the squared t-statistic associated with the test of the null is in the .01 tail of its distribution. Panels (a)-(c) provide information for clustered/robust inference in low, medium and high leverage samples with iid normal errors, while panels (d)-(f) illustrate results with heteroskedastic clustered normal errors. The OLS and 2SLS figures are based upon simulations with uncorrelated and correlated 1st and 2nd stage errors, respectively. By focusing on ideal environments for each method, tests of true nulls, and actual .01 tail values, the figures remove confounding influences, allowing a better understanding of the factors behind size distortions and diminishing power. Alternative scenarios are described below and in the on-line appendix.

---

incorrect. In the low leverage sample the average ln difference between cl/robust and default standard error estimates is .78 with heteroskedastic normal errors and 1.64 with heteroskedastic clustered normal errors.

Beginning with OLS, Figure III shows how leverage and non-iid errors interact to systematically increase size and reduce relative power. In the close-to-ideal low leverage environment of panel (a), extreme coefficient values completely dominate tail t-values. With medium and high leverage, panels (b)-(c), standard errors become somewhat more volatile, so that small standard errors play a bigger role in rejections. This is precisely the pattern one sees if one, for example, simulates the distribution of the ratio of a chi^2 squared coefficient estimate to an n-k chi^2 standard error estimate, and then lowers n-k. To the degree that this increased standard error volatility is not recognized by degrees of freedom, size is greater than nominal value. The shift to regressor-correlated heteroskedasticity in panels (d)-(f) greatly increases the volatility of standard errors, especially in higher leverage samples, and their role in tail values, at the expense of extreme coefficient realizations. As the degrees of freedom typically used to evaluate these distributions don't change between (a)-(c) and (d)-(f), size distortions rise. Power relative to size also falls, as the overall distribution of the test statistic is dominated by the dispersal of the denominator.[11]

From the perspective of this paper, however, the most interesting aspect of Figure III is the contrast between OLS and 2SLS results. As can be seen in the OLS panels, the frequency with which large coefficient deviations appear in the tail realizations of t-statistics declines as the role of small standard errors rises. In the case of 2SLS, however, this relationship is qualitatively different, as large coefficient deviations play a smaller role in every panel and, in the extreme, virtually vanish altogether in high leverage regressions with heteroskedastic errors. The distribution of the standard error, and not the distribution of coefficients, utterly dominates the tail realizations of 2SLS t-statistics.[12] This result is confirmed when I use the different forms of

[11]The greater volatility of coefficient estimates, as accounted for in larger standard errors, also plays a role as a given null shift in the numerator has less of an effect on any given t-statistic.

[12]The result in panel (f) for 2SLS, with the modal concentration of coefficient deviations below the largest decile, is clearly suggestive of strong correlations between standard error estimates and coefficient deviations. The on-line appendix shows that 2SLS retains this characteristic when the 1st and 2nd stage disturbances are uncorrelated. With correlated errors, the .01 tail of the t- test that the biased OLS coefficient estimate equals the true null is populated less by large coefficient deviations, but the density remains monotonic and (with non-iid errors) large coefficient deviations are much more important than in 2SLS.

the bootstrap to analyse the actual regressions in my sample further below. Published IV coefficients are generally found to be significant using conventional clustered/robust inference not because the coefficient estimate has an unusually extreme value (given the null), but because the standard error estimate is surprisingly small given the data generating process. This is suggestive of either spurious rejections or a very low ratio of power to size.[13]

Table V below reports Monte Carlo estimates of the average ln relative 2SLS to OLS truncated variance, bias and mean squared error (MSE) around the parameter value of the data generating process. With normal disturbances only the first $k_{\mathbf{Z}}–k_{\mathbf{Y}}$ finite sample moments of 2SLS estimates exist (Kinal 1980). Consequently, in these simulations moments do not exist for most of my sample, which is only exactly identified. However, the percentiles of a distribution always exist and one can always legitimately estimate percentile truncated moments. In the table I estimate moments after removing the largest and smallest ½ of one percent of outcomes of both 2SLS and OLS, i.e. moments calculated across the central 99 percentiles of the distribution of coefficients, providing some insight into how non-iid errors affect their relative properties.

As shown in the table, non-iid disturbances have a distinctly adverse effect on the relative performance of 2SLS. With iid normal errors, the average truncated relative bias of 2SLS coefficient estimates is ln -3.4 lower than OLS when OLS is biased, while MSE is ln -.7 lower. In contrast, with heteroskedastic & clustered normal errors 2SLS has an average bias that is only ln -1.3 lower, and a MSE that is ln 2.3 times greater and, in fact, is on average greater in 27 out of 31 papers. With heteroskedastic and clustered errors 1st stage predicted values are more heavily influenced by the realization of individual errors that are correlated with the 2nd stage. This weakens the bias advantage of 2SLS that in iid settings offsets its greater relative variance,

[13]Figure III standardizes by looking at the .01 tail of each distribution. The on-line appendix reports the deciles that appear when results are rejected at the .01 nominal level in size and power tests. As would be expected, size distortions move further into each distribution, picking up less extreme values of both coefficients and standard errors, without changing the fundamental result concerning OLS vs 2SLS inference. Also as would be expected, power tests, where the coefficient deviation from a false null of 0 is located in the deciles of the coefficient deviations from the true parameter value, increase the role of extreme coefficient deviations, but this effect is very weak in situations where 2SLS has low power.

Table V: Ln Relative 2SLS to OLS Truncated Absolute Bias, Variance & MSE
(1000 Monte Carlo simulations for each of 1359 equations)

	correlated 1st & 2nd stage errors			uncorrelated 1st & 2nd stage errors		
	bias	variance	mse	bias	variance	mse
iid normal	-3.4	3.4	-.70	1.5	3.2	3.2
iid chi^2	-3.7	2.8	-.75	1.4	3.2	3.2
h. normal	-2.2	3.3	.89	1.7	3.1	3.1
h. chi^2	-2.4	2.7	.68	1.4	3.0	3.0
h. & cl. normal	-1.3	3.9	2.3	2.2	3.8	3.8
h. & cl. chi^2	-1.4	3.5	2.1	1.9	3.7	3.7

	correlated 1st & 2nd stage errors								
	bias			variance			mse		
	low	medium	high	low	medium	high	low	medium	high
iid normal	-3.9	-2.6	-3.6	4.5	3.3	2.5	-1.3	-0.1	-0.7
iid chi^2	-4.6	-2.6	-3.9	3.9	2.5	2.0	-1.3	-0.2	-0.8
h. normal	-3.1	-1.7	-1.9	4.1	3.2	2.6	0.2	1.1	1.3
h. chi^2	-3.2	-2.0	-2.0	3.5	2.6	2.1	0.2	1.0	0.9
h. & cl. normal	-1.9	-1.4	-0.7	4.7	3.3	3.9	2.0	1.7	3.3
h. & cl. chi^2	-1.9	-1.5	-0.8	4.3	2.9	3.4	1.9	1.6	2.8

Notes: Estimates calculated by removing the largest and smallest ½ of one percentile of IV & OLS coefficient outcomes. Low, medium and high refer to groups of papers based on maximum **Z** leverage (Table III earlier).

resulting in substantially greater relative MSE. When errors are uncorrelated, the estimated truncated bias of OLS is trivially small, as is that of 2SLS, so all that matters is the relative inefficiency or variance of 2SLS, which is quite large, as shown in the upper right-hand panel of the table. In sum, as is well known, when OLS is unbiased 2SLS is an inferior estimator of the parameter of interest because of its greater variance. In the presence of clustered and heteroskedastic errors it is also an inferior estimator, in most papers in my sample, even when OLS is biased, as, for a given variance and correlation of 1st and 2nd errors much of the advantage of 2SLS in terms of bias disappears.[14] Leverage does not play a particularly dominant role in this process, as can be seen by the fact that the middle leverage group experiences the smallest

[14]Calculations using the central 95 or 90 percentiles of each distribution, in the on-line appendix, show a similar pattern. With more of the tails of the distributions removed, the relative variance of 2SLS to OLS declines, but there is still a marked worsening of relative bias in the presence of non-iid heteroskedastic and clustered errors, resulting in higher truncated MSE than OLS in more than 70 percent of papers.

increase in relative bias in the table. Leverage does, however, impact the ability of tests to provide assurances on the limits of this bias, as shown shortly below.

(b) First stage Pre-Tests and F-tests

Following the influential work of Nelson and Startz (1990a,b) and Bound, Jaeger and Baker (1995), which identified the problems of size, bias and inconsistency associated with a weak 1st stage relation, all of the papers in my sample try to assure the reader that the excluded instruments are relevant and their relationship with the right-hand side endogenous variable strong. Twenty-two papers explicitly report 1st stage F statistics in at least some tables, with the remainder using coefficients, standard errors, p-values and graphs to make their case. The reporting of 1st stage F-statistics is, in particular, motivated by Staiger and Stock's (1997) derivation of the weak instrument asymptotic distribution of the 2SLS estimator in an iid world and, on the basis of this, Stock and Yogo's (2005) development of weak instrument pre-tests using the first stage F-statistic to guarantee no more than a .05 probability that 2SLS has size under the null or proportional bias relative to OLS greater than specified levels. In this section I show that in non-iid settings these tests are largely uninformative, but cluster/robust modifications work somewhat better, provided maximal leverage is low.

Tables VI and VII apply Stock and Yogo's weak instrument pre-tests to each of the 1000 draws for each IV regression from each of the six data generating processes described earlier. I divide regressions based upon whether or not they reject the weak instrument null (H_0) in favour of the strong instrument alternative (H_1) and report the fraction of regressions so classified which, based upon the entire Monte Carlo distribution, have size or bias greater than the indicated bound.[15] I also report (in parentheses) the maximum fraction of H_1 observations violating the bounds that would be consistent with the test having its theoretical nominal size of

[15]In other words, each individual data draw is classified into H_0 or H_1 based upon its 1st stage F statistic, but the size or bias characteristics of all 1000 data draws for a particular regression specification are evaluated using their combined distribution. Generally in this paper I calculate size using the finite sample t and F distributions. However, Stock & Yogo (2005) base their theory around Wald and F-statistics with finite sample corrections (pp. 83-84) but p-values calculated using the asymptotic chi^2 distribution (pp. 88), so I follow this approach in the table. Results using the finite sample t-distribution are similar, and are presented in the on-line appendix.

Table VI: Fraction of Regressions with Size Greater than Indicated in Specifications that Don't (H_0) and Do (H_1) Reject the Stock & Yogo Weak Instrument Null
(1000 simulations for each error process in 1327 IV regressions)

	maximum acceptable size for a nominal .05 test							
	.10		.15		.20		.25	
	H_0	H_1 (max)	H_0	H_1 (max)	H_0	H_1 (max)	H_0	H_1 (max)
(A) default F used as Stock and Yogo test statistic								
default cov								
iid normal	.115	.000 (.021)	.078	.000 (.012)	.062	.000 (.010)	.048	.000 (.009)
iid chi^2	.117	.000 (.021)	.078	.000 (.013)	.061	.000 (.010)	.053	.000 (.009)
cl/robust cov								
iid normal	.221	.268 (.021)	.097	.024 (.012)	.062	.014 (.010)	.053	.005 (.009)
iid chi^2	.203	.281 (.021)	.074	.021 (.013)	.066	.014 (.010)	.048	.000 (.009)
h normal	.438	.250 (.019)	.215	.122 (.013)	.096	.076 (.011)	.040	.058 (.010)
h chi^2	.557	.427 (.018)	.299	.182 (.012)	.149	.130 (.010)	.053	.081 (.009)
h cl normal	.425	.433 (.018)	.262	.343 (.013)	.122	.173 (.011)	.043	.072 (.010)
h cl chi^2	.559	.526 (.017)	.314	.405 (.012)	.178	.351 (.010)	.093	.237 (.009)
(B) clustered/robust F used as Stock and Yogo test statistic								
cl/robust cov								
iid normal	.212	.269 (.018)	.109	.024 (.011)	.068	.013 (.009)	.057	.005 (.008)
iid chi^2	.199	.278 (.017)	.085	.020 (.011)	.074	.013 (.009)	.053	.000 (.008)
h normal	.403	.223 (.039)	.190	.116 (.025)	.090	.075 (.020)	.046	.058 (.017)
h chi^2	.525	.416 (.037)	.271	.173 (.024)	.148	.127 (.018)	.061	.081 (.015)
h cl normal	.453	.386 (.108)	.327	.328 (.063)	.146	.181 (.048)	.042	.087 (.041)
h cl chi^2	.531	.540 (.082)	.334	.440 (.050)	.252	.377 (.040)	.159	.253 (.034)

Notes: Sample is restricted to regressions for which Stock & Yogo (2005) provide critical values; default and cl/robust cov = using these covariance matrices to calculate t-statistics and p-values; max = maximum share of the sample that rejects H_0 in favour of H_1 with size greater than indicated bound consistent with the test itself having size .05 (see text and accompanying footnote); error processes as describe earlier above. Size estimates based upon 1000 Monte Carlo simulations per error process per IV regression.

no greater than .05.[16] With critical values dependent upon the number of instruments and endogenous regressors, Stock and Yogo provide size critical values for 1327 of the 1359 regressions in my sample, but bias critical values for only 180 of the over-identified regressions,

[16]Let N_0 and N_1 denote the known number of regressions classified under H_0 and H_1, respectively, and W_0, W_1, S_0 and S_1 the unknown number of regressions with weak and strong instruments in each group, with $W_1 = \alpha(W_0+W_1)$ and $S_0 = (1-p)(S_0+S_1)$, where $\alpha \leq .05$ and p denote size and power. Then $W_1/N_1 = (\alpha/(1-\alpha))(N_0-S_0)/N_1$, which, for given N_0 & N_1, is maximized when p = 1 and $\alpha = .05$, with $W_1/N_1 = (1/19)(N_0/N_1)$. The relative number of regressions in the N_0 and N_1 groups for each test can be calculated by inverting this equation.

where the first moment can be taken as existing.[17]

Table VI begins by using the default covariance estimate to evaluate both the F-statistic and coefficient significance when the data generating process is consistent with Stock and Yogo's iid-based theory.[18] In this context, the test performs remarkably well. Only a miniscule fraction (ranging from .00009 to .00037) of the regressions which reject the weak instrument null H_0 in favour of the strong alternative H_1 have a size greater than the desired bound. Outside of this ideal environment, however, the test rapidly becomes uninformative. When the clustered/robust covariance estimate is used to evaluate coefficient significance with iid disturbances the test still provides good insurance against large size distortions at the .05 nominal level, but for lower levels of size the biases introduced by the clustered/robust approach dominate any issues associated with instrument strength. When in addition non-iid error processes are introduced, the test appears to provide no information at any level, as the fraction of regressions with size greater than the specified level in H_1 regressions is generally equal to or greater than that found in H_0 and always much larger than the maximum share consistent with the Stock & Yogo test itself having a nominal size of .05. Use of the clustered/robust 1st stage F-statistic as the test-statistic, an ad-hoc adjustment of Stock and Yogo's iid-based theory generally implemented by users,[19] provides no improvement whatsoever. Stock and Yogo's bias test, as shown in Table VII, performs only slightly better. In non-iid settings the fraction of regressions with IV to OLS relative bias greater than the specified amount in H_1 is generally lower than in the H_0 sample, but, at levels often reaching .90, much too high to either be consistent with the test having a .05 Type-I error rate or provide much comfort to users. Results

[17]Staiger and Stock (1997) showed that for iid disturbances of any sort the asymptotic weak instrument distribution follows the finite sample normal distribution, which has a first moment when the regression is over-identified. Consequently, I evaluate these tests using the actual bias, as estimated by the full Monte Carlo distribution, rather than the truncated bias used in the analysis of all regressions earlier above.

[18]As the number of papers with any results classified in H_1 varies substantially as one moves down the columns or across the rows of the table, here, and in Tables VII & VIII, I depart from the practice elsewhere of reporting averages across papers of within paper averages, and simply weight each simulation regression equally.

[19]Eleven of the papers in my sample that report F-statistics make direct reference to the work of Stock and his co-authors. All of these report clustered/robust measures, although two report default F-statistics as well.

Table VII: Fraction of Regressions with Relative Bias Greater than Indicated in Specifications that Don't and Do Reject the Stock & Yogo Weak Instrument Null (1000 simulations for each error process in 180 over-identified IV regressions)

	maximum acceptable relative bias							
	.05		.10		.20		.30	
	H_0	H_1 (max)	H_0	H_1 (max)	H_0	H_1 (max)	H_0	H_1 (max)
(A) default F used as Stock and Yogo test statistic								
iid normal	.980	.110 (.070)	.913	.054 (.065)	.797	.045 (.053)	.651	.026 (.038)
iid chi^2	.961	.098 (.070)	.906	.029 (.064)	.872	.035 (.052)	.640	.024 (.037)
h normal	.999	.280 (.063)	.987	.273 (.046)	.952	.326 (.026)	.761	.292 (.017)
h chi^2	.974	.278 (.053)	.968	.281 (.037)	.789	.302 (.022)	.506	.202 (.014)
h cl normal	1.00	.909 (.055)	.988	.911 (.040)	.957	.875 (.023)	.801	.800 (.015)
h cl chi^2	.977	.916 (.046)	.965	.887 (.032)	.785	.836 (.019)	.527	.728 (.013)
(B) clustered/robust F used as Stock and Yogo test statistic								
iid normal	.985	.116 (.068)	.923	.074 (.060)	.803	.151 (.037)	.633	.161 (.020)
iid chi^2	.966	.110 (.067)	.916	.091 (.055)	.849	.252 (.026)	.568	.212 (.012)
h normal	.992	.536 (.022)	.984	.517 (.012)	.921	.484 (.007)	.665	.383 (.005)
h chi^2	.968	.507 (.019)	.974	.474 (.012)	.887	.387 (.007)	.636	.234 (.005)
h cl normal	.988	.917 (.063)	.995	.905 (.041)	.965	.862 (.031)	.965	.716 (.027)
h cl chi^2	.986	.901 (.054)	.971	.875 (.040)	.956	.745 (.030)	.934	.568 (.026)

Notes: as in Table VI above.

for both tables broken down by paper leverage (in the on-line appendix) do not find these tests to be informative in low, medium or high leverage sub-samples either.[20] The misapplication of Stock & Yogo's iid based test in non-iid settings does not yield useful results.

Olea and Pflueger (2013), noting that the widespread application of Stock & Yogo's test in non-iid settings is not justified by theory, undertake the challenging task of extending the method to allow for non-iid errors, deriving critical values for the null hypothesis that the IV Nagar bias is smaller than a "worst-case" benchmark. The Nagar bias is the bias of an approximating distribution based on a third-order Taylor series expansion of the asymptotic distribution, while the worst-case benchmark equals the OLS bias in the case of iid errors. The

[20]In low leverage papers, both H_0 and H_1 regressions generally have minimal size distortions, but the H_1 sample often exceeds the desired size bound more frequently than the H_0 sample, indicating that the critical values of the test are not particularly discerning. In the bias test for the low leverage sample, regressions in H_1 exceed the bias bound less frequently than those in H_0, but more often than is consistent with the test having a .05 Type-I error probability.

Table VIII: Fraction of Regressions with Relative Bias Greater than Indicated in Specifications that Don't and Do Reject the Olea & Pflueger Weak Instrument Null (1000 simulations for each error process in 272 over-identified IV regressions)

	maximum acceptable relative bias							
	.05		.10		.20		⅓	
	H_0	H_1 (max)	H_0	H_1 (max)	H_0	H_1 (max)	H_0	H_1 (max)
(A) 54 over-identified regressions in 3 low leverage papers								
iid normal	.215	.080 (.005)	.303	.000 (.003)	.470	.000 (.002)	.000	.000 (.002)
iid chi^2	.217	.080 (.005)	.298	.020 (.003)	.480	.019 (.002)	.000	.000 (.002)
h normal	.455	.158 (.046)	.346	.069 (.021)	.131	.000 (.009)	.000	.000 (.006)
h chi^2	.304	.186 (.046)	.262	.026 (.021)	.000	.000 (.009)	.000	.000 (.006)
h cl normal	.983	.161 (.405)	.988	.153 (.388)	.950	.000 (.377)	.930	.001 (.371)
h cl chi^2	.981	.164 (.414)	.969	.153 (.382)	.931	.152 (.368)	.911	.012 (.359)
(B) 166 over-identified regressions in 6 medium leverage papers								
iid normal	.940	.045 (.321)	.870	.039 (.285)	.748	.039 (.239)	.576	.048 (.171)
iid chi^2	.932	.011 (.318)	.863	.000 (.286)	.807	.012 (.237)	.555	.022 (.178)
h normal	.936	.361 (.543)	.921	.257 (.339)	.908	.226 (.250)	.686	.236 (.212)
h chi^2	.919	.467 (.321)	.928	.379 (.200)	.748	.309 (.148)	.439	.216 (.127)
h cl normal	.972	.904 (1.45)	.908	.254 (.537)	.914	.269 (.307)	.786	.264 (.241)
h cl chi^2	.900	.718 (.619)	.890	.522 (.279)	.764	.366 (.167)	.555	.251 (.135)
(C) 52 over-identified regressions in 4 high leverage papers								
iid normal	.000	.253 (.024)	.000	.165 (.012)	.000	.105 (.005)	.000	.061 (.003)
iid chi^2	.002	.186 (.020)	.000	.092 (.010)	.000	.042 (.005)	.000	.041 (.003)
h normal	.962	.293 (.049)	.940	.297 (.036)	.898	.277 (.026)	.908	.307 (.021)
h chi^2	.933	.232 (.046)	.851	.199 (.031)	.843	.254 (.021)	.246	.124 (.017)
h cl normal	.970	.831 (.856)	.979	.845 (.355)	.962	.869 (.191)	.843	.853 (.141)
h cl chi^2	1.00	1.00 (.657)	.988	.942 (.285)	.967	.872 (.153)	.880	.774 (.112)

Notes: as in Table VI above.

test statistic is related to the clustered/robust 1st stage F-statistic, but the calculation of sample dependent degrees of freedom for the test is computationally costly and impractical for the many simulations underlying the table which follows. Olea and Pflueger note, however, that conservative degrees of freedom can be estimated using only the eigenvalues of the clustered/ robust 1st stage F-statistic, and I make use of this approach along with the table of critical values they provide. These conservative degrees of freedom should lower the probability of a Type-I error, i.e. classifying as H_1 a regression with a relative bias greater than the desired level, below the .05 size of the test.

Table VIII above applies Olea & Pflueger's test to the Monte Carlo sample. As before, I divide regressions by whether or not they reject the weak instrument null and report the fraction of regressions in each group where the relative bias of IV to OLS, as estimated from the Monte Carlo distribution, exceeds the acceptable bound. In fairness, this bias bound is not the object of the test, which concerns asymptotic bias relative to a worst case IV-approximation benchmark, but I would argue it is the object of interest to users, who use 2SLS in order to avoid OLS bias. I divide the analysis of the 272 over-identified equations in my sample by the average level of leverage in each paper, as earlier classified in Table III. As shown in the table, bias levels in regressions which reject H_0 in favour of H_1 are generally very much lower in low leverage papers, although they sometimes exceed the maximum bound consistent with the test having no more than a .05 probability of Type-I error. The bias of H_1 results rises, however, in medium and high leverage papers and, in the latter, in the case of heteroskedastic and clustered errors becomes virtually indistinguishable from regressions which cannot reject H_0 for all bias bounds. In the on-line appendix I show that the Olea & Pflueger critical values result in lowers levels of bias than the iid based Stock & Yogo test in H_1 groups in all leverage sub-samples with non-iid errors. Olea & Pflueger also provide critical values for exactly identified equations, as the Nagar bias always exists even if the first moment does not. Applying these and comparing relative 2SLS to OLS bias in the truncated central 99 percentiles of their distributions, in the on-line appendix I find results similar to those of Table VIII: in the high leverage sample relative bias often differs very little between the H_0 and H_1 groups, but the test is fairly discerning in the low leverage sample, with substantially lower bias levels in the H_1 group, although generally greater than is consistent with the test providing a maximum .05 probability of a Type-I error.

Table IX reports Monte Carlo estimates of size in 1st stage F-tests using default and clustered/robust covariance estimates. As expected, estimated size with the default covariance estimate is close to its theoretical value with iid disturbances, but explodes with non-iid errors. Clustered/robust covariance estimates provide better results, especially in low leverage papers, but size distortions are very high in medium and high leverage papers, particularly in over-

Table IX: Average Rejection Rates of True Nulls at the .01 Level in 1st Stage Tests
(1000 Monte Carlo simulations for each of 1359 equations)

	default			clustered/robust								
				low leverage			medium leverage			high leverage		
	all	$k_Z > 1$		all	$k_Z > 1$		all	$k_Z > 1$		all	$k_Z > 1$	
		coef	joint		coef	joint		coef	joint		coef	joint
iid normal	.010	.010	.010	.011	.010	.011	.071	.020	.115	.061	.045	.276
iid chi^2	.012	.012	.015	.012	.010	.009	.051	.017	.083	.055	.041	.267
h. normal	.310	.184	.382	.013	.010	.013	.052	.015	.061	.170	.088	.380
h. chi^2	.312	.182	.384	.020	.012	.016	.076	.019	.099	.195	.091	.385
h. & cl. normal	.526	.334	.613	.013	.012	.014	.055	.015	.063	.195	.111	.386
h. & cl. chi^2	.525	.323	.614	.028	.013	.022	.077	.020	.094	.225	.110	.400

Notes: all = average across all equations in all papers; $k_Z > 1$ = average across 3 low, 6 medium and 4 high leverage papers in equations with more than 1 excluded instrument; coef = test of individual coefficients; joint = joint test of all excluded instruments.

identified equations. Part of this has to do with the way size distortions increase when more than one coefficient is tested, which the table shows by comparing the average size of coefficient level (t) tests of the excluded instruments in over-identified equations with the much higher rejection rates found in the joint (F) tests of these instruments. Intuition for this result can be found by considering that the familiar F-statistic actually equals 1/k times the maximum squared t-statistic that can be found by searching over all possible linear combinations **w** of the estimated coefficients, that is

$$(9)\quad \frac{\hat{\boldsymbol{\beta}}'\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\beta}}}{k} = \frac{1}{k}\operatorname*{Max}_{\mathbf{w}}\frac{(\mathbf{w}'\hat{\boldsymbol{\beta}})^2}{\mathbf{w}'\hat{\mathbf{V}}\mathbf{w}}$$

When errors are iid and $\hat{\mathbf{V}} = \mathbf{V}\hat{\sigma}^2/\sigma^2$, this search is actually limited. Employing the transformations $\tilde{\mathbf{w}} = \mathbf{V}^{½}\mathbf{w}$ and $\tilde{\boldsymbol{\beta}} = \mathbf{V}^{-½}\hat{\boldsymbol{\beta}}$, plus the normalization $\tilde{\mathbf{w}}'\tilde{\mathbf{w}} = 1$, one sees that:

$$(10)\quad \frac{1}{k}\operatorname*{Max}_{\mathbf{w}}\frac{(\mathbf{w}'\hat{\boldsymbol{\beta}})^2}{\mathbf{w}'\hat{\mathbf{V}}\mathbf{w}} = \frac{n-k}{k}\operatorname*{Max}_{\tilde{\mathbf{w}}}\frac{(\tilde{\mathbf{w}}'\tilde{\boldsymbol{\beta}})^2}{(n-k)\tilde{\mathbf{w}}'\mathbf{V}^{-½}\hat{\mathbf{V}}\mathbf{V}^{-½}\tilde{\mathbf{w}}} = \frac{n-k}{k}\,\frac{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}{(n-k)\hat{\sigma}^2/\sigma^2}$$

The last equality follows because the denominator reduces to $(n-k)\hat{\sigma}^2/\sigma^2$, a chi^2 variable with n-k degrees of freedom, no matter what $\tilde{\mathbf{w}}$ such that $\tilde{\mathbf{w}}'\tilde{\mathbf{w}} = 1$. In this case one can separately maximize the numerator across $\tilde{\mathbf{w}}$ and find that the maximand equals $\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}$, which is an independent chi^2 variable with k degrees of freedom. Consequently, the entire expression is

distributed $F_{k,n-k}$. However, when errors are no longer iid and in finite samples $\hat{\mathbf{V}}$ is no simple scalar multiple of the true covariance $\mathbf{V}$, the denominator takes on different values as $\tilde{\mathbf{w}}$ varies. In the test of a single coefficient, the clustered/robust covariance estimate may have bias and a volatility greater than nominal degrees of freedom, but a joint test involves a search across all possible combinations of this bias and volatility to generate maximal test statistics, producing tail probabilities that are even more distorted away from iid-based nominal values.

In the asymptotic world that forms the foundation of Olea & Pflueger's results clustered/robust covariance estimates should allow for exact inference. As shown by Table IX, in the finite sample highly-leveraged world of published papers this is far from the case. Problems of inaccurate covariance estimation are compounded in higher dimensional tests, making large clustered/robust 1st stage Fs much more likely than suggested by asymptotic theory. This probably renders the Olea/Pflueger less informative than it might otherwise be.

(c) Weak Instrument Robust Methods

The argument that non-iid error processes effectively weaken 1st stage relations might lead practitioners to mistakenly conclude that well known weak instrument robust inference methods provide easy solutions to the problems described above. In the on-line appendix I use the same Monte Carlos to examine the performance of three of these, the Anderson-Rubin (1949) reduced-form method, limited information maximum likelihood (LIML) and Fuller's-k (1977), relative to 2SLS. In iid settings the Anderson-Rubin method provides exact size no matter what the strength of instruments, while LIML provides nearly exact size and Fuller's-k better bias than 2SLS in the presence of weak instruments (Stock & Yogo 2005). In non-iid Monte Carlos, because size distortions are as much of a problem in OLS as in 2SLS but grow with the dimensionality of the test, I find that the Anderson-Rubin approach provides no improvements in exactly identified equations while delivering much larger size distortions in over-identified cases. LIML provides modestly improved size that nevertheless remains well above nominal value, at the cost of a substantial increase in truncated variance (as the LIML point estimate has no moments). Fuller's-k has modestly larger size distortions than 2SLS and suffers larger increases

in bias from non-iid errors, but nevertheless retains some advantage in bias and MSE in non-iid simulations. None of these methods provides anything approaching a solution to the size distortions and increased bias of IV estimation brought on by non-iid errors.

V. Improved Finite Sample Inference Using the JackKnife and Bootstrap

This section shows that in non-iid settings the jackknife and the bootstrap provide improved finite sample inference, with smaller size distortions and a higher ratio of power to size than found using standard clustered/robust covariance estimates and their associated degrees of freedom. These methods are often evaluated based upon their asymptotic properties, but their practical usefulness lies in their superior finite sample performance, which is often quite unrelated to asymptotic results. I begin with a brief description of the methods and then use Monte Carlos to establish their finite sample benefits.

(a) **The JackKnife**

The jackknife variance estimate based on the full sample ($\hat{\boldsymbol{\beta}}$) and m delete-**i** ($\hat{\boldsymbol{\beta}}_{\sim \mathbf{i}}$) coefficient values is given by:

$$(11) \quad \frac{m-1}{m}\sum_{\mathbf{i}}(\hat{\boldsymbol{\beta}}_{\sim \mathbf{i}}-\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{\sim \mathbf{i}}-\hat{\boldsymbol{\beta}})' \;=\; \frac{m-1}{m}\sum_{\mathbf{i}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{\mathbf{i}}'\boldsymbol{\varepsilon}_{\mathbf{i}}\boldsymbol{\varepsilon}_{\mathbf{i}}'\mathbf{X}_{\mathbf{i}}(\mathbf{X}'\mathbf{X})^{-1}$$

where, for expositional purposes, in the second expression I have substituted using the formula for the delete-**i** change in coefficient estimates in the OLS regression on variables **X**. Hinkley (1977) showed that the jackknife variance estimate is asymptotically robust to arbitrary heteroskedasticity, and as such was given credit by MacKinnon and White (1985) as an early developer of clustered/robust methods. The jackknife, however, has largely been superseded by the bootstrap which is considered superior, and indeed in simulations below I find this to be the case. I provide the jackknife in response to referees who have asked whether finite sample corrections of clustered/robust covariance estimates do not allow for better inference than that given by the bootstrap. With $\mathbf{H_{ii}}$ denoting the portion of the hat matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ associated with observations **i**, these finite sample corrections adjust for the observation specific reduction

in error variance brought on by leverage by substituting corrected residuals $(\mathbf{I}-\mathbf{H}_{ii})^{-1/2}\hat{\boldsymbol{\varepsilon}}_i$ or $(\mathbf{I}-\mathbf{H}_{ii})^{-1}\hat{\boldsymbol{\varepsilon}}_i$ for $\hat{\boldsymbol{\varepsilon}}_i$ in the clustered/robust calculation. Unfortunately, for much of my sample the matrices $\mathbf{I}$ - $\mathbf{H}_{ii}$ are singular, as the regressions contain cluster fixed effects, and these corrections cannot be applied. However, when $\mathbf{I}$ - $\mathbf{H}_{ii}$ is not singular, it is easily shown that the delete-**i** residuals $\boldsymbol{\varepsilon}_i$ equal $(\mathbf{I}-\mathbf{H}_{ii})^{-1}\hat{\boldsymbol{\varepsilon}}_i$ and consequently, with the exception of the inconsequential (m-1)/m, in OLS settings the jackknife variance estimate (11) *actually is* the clustered/robust estimate with the $(\mathbf{I}-\mathbf{H}_{ii})^{-1}$ correction of residuals. Although all forms of the clustered/robust covariance estimate are asymptotically identical, MacKinnon and White (1985) showed that these corrections yield the most accurate inference in finite samples. By using delete-**i** residuals the jackknife provides improvement in avoiding the bias brought on by the reduction of the size of residuals in highly leveraged (and consequently weighted) observations in clustered/robust covariance estimates, but it fails to account for the fact that high leverage places disproportionate weight on a small number of residuals, producing test statistics with a more dispersed distribution than anticipated from the standard clustered/robust degrees of freedom, and hence remains inferior to the bootstrap.[21]

(b) The Bootstrap

I use two forms of the bootstrap, the non-parametric "pairs" resampling of the data and the parametric "wild" bootstrap transformation of residuals. Conventional econometrics uses assumptions and asymptotic theorems to infer the distribution of a statistic f calculated from a sample with empirical distribution F_1 drawn from an infinite parent population with distribution F_0, which can be described as $f(F_1|F_0)$. In contrast, the resampling bootstrap estimates the distribution of $f(F_1|F_0)$ by drawing random samples F_2 from the population distribution F_1 and observing the distribution of $f(F_2|F_1)$ (Hall 1992). If f is a smooth function of the sample, then asymptotically the bootstrapped distribution converges to the true distribution (Lehmann and

[21]Both of these problems, of course, disappear asymptotically. I should note that there is a related jackknife variance formula that uses the jackknife pseudo-values, but its standard error estimates are always smaller than those of (11) (producing larger size distortions) and its calculation is not as closely linked to the finite sample clustered/ robust corrections just described.

Romano 2005), as, intuitively, the outcomes observed when sampling F_2 from an infinite sample F_1 approach those arrived at from sampling F_1 from the actual population F_0.

The resampling bootstrap described above is fully nonparametric, as the only assumption made is that the sample can be divided into groups that are independent draws from the data generating function of the population distribution.[22] From a regression perspective, however, the samples are "pairs" of dependent outcomes and regressors and, as such, the estimated distribution of the test statistic is that with both stochastic residuals and regressors. The "wild" bootstrap imposes parametric structure and uses transformations of the residuals to mimic a more traditional resampling of stochastic residuals alone. For example, in the regression model $Y_i = \mathbf{z}_i'\boldsymbol{\beta}_\mathbf{z} + \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + v_i$, to estimate the distribution of coefficients and test statistics under the null that $\boldsymbol{\beta}_\mathbf{z} = \mathbf{0}$ one begins by estimating the restricted equation $Y_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_\mathbf{x} + \hat{v}_i$, generating artificial data $Y_i^{wild} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_\mathbf{x} + \eta_i\hat{v}_i$, where η_i is a 50/50 iid[23] draw from the pair (-1,1), and then running Y_i^{wild} on $\mathbf{z}_\mathrm{i}$ and $\mathbf{x}_\mathrm{i}$. The initial estimation of the parametric data generating process can involve imposing the null, as just done, or not, and the transformations η_i can be symmetric or asymmetric. In Monte Carlo studies I find, as reported in the on-line appendix, that a failure to impose the null results in very large size distortions, while asymmetric transformations provide no advantages, even when the data generating process for the residuals v_i is decidedly asymmetric. Because a separate null has to be imposed on the wild data generating process for each separate test, use of the wild bootstrap is computationally costly and complex. I provide results using this method, however, because it is familiar to many users. Full details on how I impose the null on the data-generating process for each separate wild bootstrap test, and on how this improves the accuracy of inference using the method, are provided in the on-line appendix.

For both the pairs resampling and wild transformations bootstraps I draw inferences using two methods, one based upon the distribution of bootstrapped test statistics (the bootstrap-t) and

[22]Thus, in implementing the method, I follow the assumptions implicit in the authors' covariance calculation methods: resampling clusters where they cluster and resampling observations where they do not.

[23]In the case of clustered data, η_i is drawn and applied at the cluster level.

another based upon the distribution of bootstrapped coefficients (the bootstrap-c). To illustrate with the case of the resampling bootstrap, one can test whether the coefficient estimates $\boldsymbol{\beta}_1$ in the sample F_1 is different from $\mathbf{0}$ by looking at the distribution of the Wald-statistics for the test that the coefficient estimates $\boldsymbol{\beta}_2$ in the sample F_2 drawn from F_1 are different from $\boldsymbol{\beta}_1$ (the known parameter value for the data generating process), computing the probability

$$(12)\quad (\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1)'\mathbf{V}(\boldsymbol{\beta}_2^i)^{-1}(\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1) \;>\; (\boldsymbol{\beta}_1 - \mathbf{0})'\mathbf{V}(\boldsymbol{\beta}_1)^{-1}(\boldsymbol{\beta}_1 - \mathbf{0})$$

where $\boldsymbol{\beta}_1$ is the vector of coefficients estimated using the original sample F_1, $\boldsymbol{\beta}_2^i$ the vector of coefficients estimated in the i^{th} draw of sample F_2 from F_1, and $\mathbf{V}(\boldsymbol{\beta}_1)$ and $\mathbf{V}(\boldsymbol{\beta}_2^i)$ their respective clustered/robust covariance estimates. In the case of a single coefficient, this reduces to calculating the distribution of the squared t-statistic, i.e. the probability:

$$(12)'\quad [(\beta_2^i - \beta_1)/\hat{\sigma}(\beta_2^i)]^2 \;>\; [(\beta_1 - 0)/\hat{\sigma}(\beta_1)]^2$$

where $\hat{\sigma}$ is the estimated standard error of the coefficient. This method, which requires calculating an iteration by iteration covariance or standard error estimate, is the bootstrap-t. Alternatively, one can use the distribution of the bootstrapped coefficients as the common covariance estimate, calculating the probability:

$$(13)\quad (\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1)'\mathbf{V}(F(\boldsymbol{\beta}_2^i))^{-1}(\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1) \;>\; (\boldsymbol{\beta}_1 - \mathbf{0})'\mathbf{V}(F(\boldsymbol{\beta}_2^i))^{-1}(\boldsymbol{\beta}_1 - \mathbf{0})$$

where $\mathbf{V}(F(\boldsymbol{\beta}_2^i))$ denotes the covariance matrix of $\boldsymbol{\beta}_2^i$ calculated using the bootstrapped distribution of the coefficients. In the case of an individual coefficient, the common variance in the denominator on both sides can be cancelled and the method reduces to calculating the probability:

$$(13)'\quad (\beta_2^i - \beta_1)^2 \;>\; (\beta_1 - 0)^2$$

which is simply the tail probability of the squared coefficient deviation from the null hypothesis. This method is the bootstrap-c.

From the point of view of asymptotic theory, the bootstrap-t may be superior, but in practical application it has its weaknesses. Hall (1992) showed that while coverage error in a one-sided hypothesis test of a single coefficient of the resampling bootstrap-t converges to zero at a rate $O(n^{-1})$, the coverage error of the bootstrap-c converges at a rate of only $O(n^{-½})$, i.e. no

better than the standard root-N convergence of asymptotic normal approximations. The intuition for this, as presented by Hall, lies in the fact that the bootstrap-t estimates an asymptotically pivotal distribution, one that does not depend upon unknowns, while the bootstrap-c estimates an asymptotically non-pivotal distribution, one that depends upon the estimated variance. As the sample expands to infinity, the bootstrap-c continues to make estimates of this parameter, which results in greater error and slower convergence of rejection probabilities to nominal value. This argument, however, as recognized by Hall himself (1992, p. 167), rests upon covariance estimates being sufficiently accurate so that the distribution of the test statistic is actually pivotal. Hall's concern is particularly relevant in the context of using clustered/robust covariance estimates in highly leveraged finite samples, and all the more so in the case of 2SLS, where the variance estimate, in all of the exactly identified equations in my sample, is an estimate of a finite sample moment that does not even exist. I find, as shown below, that the bootstrap-c performs better than the bootstrap-t in tests of IV coefficients, and is by no means very much worse in individual and joint tests of OLS coefficients either.

"Publication bias" argues in favour of using the bootstrap-c, or at least comparing it to the -t, in a study such as this. As shown in the previous section, with non-iid disturbances conventionally significant clustered/robust IV results arise most often because standard errors are unusually small rather than coefficient estimates unusually large (under the null). If so, then there should be a large discrepancy between significance rates calculated using the bootstrap-c and those calculated using the bootstrap-t. This is the pattern I find in the next section. Significant published IV results do not have unusually large coefficient values under the null. They do, however, have unusually small standard errors, and hence appear systematically more significant when analyzed using the bootstrap-t. While the bootstrap-c and -t have roughly similar size and power in Monte Carlos, as shown shortly below, they provide dramatically different assessments of published IV results. This is precisely what one would expect to find if published results are selected on the basis of statistical significance and if that statistical significance rests heavily upon unusually small draws from the distribution of standard errors.

(c) **Monte Carlos**

Table X below presents a Monte Carlo analysis of the different methods using the six data generating processes described in the previous section. As calculation of the jackknife and bootstrap (with 1000 bootstrap draws in each case) is very costly, I only evaluate 10 realizations of each data generating process for each of 1359 equations. With 13590 p-values per data generating process, this still allows evaluation of average size and power, albeit without the accuracy of the 1359000 iterations used earlier above. For the sake of comparison, I also report size and power for clustered/robust methods using the same 13590 realizations of data. I study the distribution of p-values for instrumented 2SLS coefficients and 1st stage F-tests when 1st and 2nd stage errors are correlated and for OLS versions of the estimating equation when they are uncorrelated,[24] with size involving a test of a null equal to the parameter value underlying the data generating process and power the test of a null of zero. For Durbin-Wu-Hausman tests of the bias of OLS coefficient estimates, the data generating processes for size and power involve uncorrelated and correlated errors, respectively, as these are the circumstances in which the null of no OLS bias is true or false. As the IV clustered/robust covariance estimate is not always larger than that of OLS, in these tests I use the default covariance estimates for both 2SLS and OLS so that the desired inequality is assured.

Several patterns are readily apparent in the table, which reports rejection rates at the .01 level.[25] First, in the presence of non-iid errors the jackknife and all forms of the bootstrap provide average size much closer to nominal value than clustered/robust estimates, while raising the ratio of power to size. Thus, whether one's interest is in frequentist inference or Bayesian updating these methods are on average superior to clustered/robust methods in finite sample non-iid settings. As might be expected, tables broken down by leverage group, in the on-line appendix, find their advantages over clustered/robust methods lie mainly in medium and high

[24]The correlation of 1st and 2nd stage errors does not, of course, matter for 1st stage F-tests, so I simply select the data realizations with correlated errors, as this corresponds to the case where 1st stage tests are relevant.

[25]Rejection rates at the .05 level show similar patterns, and are reported in the on-line appendix.

Table X: Improved Finite Sample Inference Using the JackKnife & Bootstrap
(average within paper rejection rates at .01 level, 10 Monte Carlo simulations per equation)

	size						power					
	clust-robust	jack-knife	pairs bootstrap		wild bootstrap		clust-robust	jack-knife	pairs bootstrap		wild bootstrap	
			c	t	c	t			c	t	c	t
IV coefficients (correlated errors)												
iid normal	.032	.015	.012	.022	.007	.011	.460	.397	.328	.394	.276	.405
iid chi2	.025	.016	.009	.016	.008	.012	.477	.409	.329	.396	.287	.409
h. normal	.071	.026	.014	.025	.014	.016	.283	.215	.199	.187	.169	.230
h. chi2	.079	.030	.013	.026	.017	.022	.288	.211	.202	.175	.178	.247
h. cl. normal	.069	.018	.010	.026	.015	.014	.183	.120	.100	.106	.097	.138
h. cl. chi2	.088	.035	.013	.032	.023	.028	.168	.108	.090	.091	.092	.133
OLS coefficients (uncorrelated errors)												
iid normal	.008	.007	.007	.007	.006	.006	.831	.822	.825	.821	.816	.824
iid chi2	.012	.009	.011	.008	.011	.014	.846	.828	.826	.815	.816	.825
h. normal	.056	.020	.020	.011	.009	.011	.652	.592	.615	.519	.615	.592
h. chi2	.075	.031	.021	.029	.020	.023	.710	.636	.637	.555	.648	.635
h. cl. normal	.066	.024	.026	.017	.019	.016	.570	.479	.518	.410	.516	.482
h. cl. chi2	.084	.031	.027	.032	.027	.021	.620	.515	.553	.453	.549	.525
1st Stage F-tests (correlated errors)												
iid normal	.046	.017	.006	.012	.009	.005	.934	.902	.861	.866	.832	.868
iid chi2	.045	.019	.011	.013	.009	.011	.941	.911	.866	.867	.854	.883
h. normal	.078	.030	.019	.017	.014	.015	.774	.699	.701	.578	.716	.692
h. chi2	.114	.056	.042	.031	.033	.038	.782	.717	.712	.570	.740	.722
h. cl. normal	.073	.027	.013	.012	.019	.011	.636	.554	.564	.424	.571	.529
h. cl. chi2	.126	.065	.047	.035	.047	.045	.635	.567	.556	.423	.595	.571
Durbin-Wu-Hausman tests												
	(uncorrelated errors)						(correlated errors)					
iid normal	.006	.010	.005	.011	.010	.012	.298	.253	.206	.270	.231	.294
iid chi2	.004	.009	.003	.010	.008	.012	.312	.251	.205	.257	.249	.314
h. normal	.239	.015	.009	.011	.018	.025	.461	.165	.148	.156	.165	.195
h. chi2	.266	.012	.005	.008	.021	.035	.461	.161	.146	.153	.170	.218
h. cl. normal	.466	.008	.003	.008	.010	.016	.595	.082	.073	.078	.078	.108
h. cl. chi2	.446	.015	.004	.010	.015	.036	.563	.086	.073	.076	.086	.128

Notes: Reported figures are the average across 31 papers of the within paper average rejection rate. iid normal & chi^2, heteroskedastic (h.) and clustered (cl.) denote the data generating process for the disturbances, as described earlier, and correlated and uncorrelated refer to the correlation between 1st and 2nd stage errors. Durbin-Wu-Hausman tests performed using default covariance estimates with uncorrelated errors for size and correlated errors for power. Bootstrap-t methods use clustered/robust covariance estimates for IV, OLS and 1st stage coefficients, and default covariance estimates for Durbin-Wu-Hausman tests.

leverage papers, but in low leverage papers they also prove to be equal or better than clustered/robust methods, which are fairly accurate in that setting. Second, as already noted earlier, the bootstrap-c is more accurate than the -t in tests of IV coefficients and is by no means

systematically much worse in other tests. These patterns are repeated when the results are broken down by leverage group in the on-line appendix. The pairs bootstrap-c, in particular, provides for size very near nominal value in tests of IV coefficients in all leverage groups. Whatever the advantages of the bootstrap-t in asymptotic theory, they do not play out very well in the finite sample environment of the typical AEA IV paper.

Third, Table X also shows that the non-parametric resampling pairs bootstrap provides inference, across all tests, similar to that given by the wild bootstrap transformation of residuals, and in fact (using the -c) is more accurate in the case IV coefficients. This, despite the fact that the data generating process is strictly parametric with non-stochastic regressors, and I only report results using the most accurate wild bootstrap methods that impose the null and, in the case of IV coefficients, use Davidson-Mackinnon's (2010) method of "restricted efficient residuals" (see the analysis in the on-line appendix). While a different set of wild bootstrap draws, imposing a separate null, is necessary for each size or power test described in the table, the pairs bootstrap allows the calculation of all of these tests, plus confidence intervals if so desired, in one set of resampling draws. As such, it provides a relatively low cost solution to issues of inference in 2SLS.[26] Computational costs are even lower if one avoids the iteration by iteration calculation of standard errors of the pairs bootstrap-t by implementing the simple pairs bootstrap-c, which has similar size and often higher power (Table X).

VI: Consistency without Inference

Table XI reports the statistical significance of the coefficients of instrumented right-hand side variables in my sample papers using conventional, jackknife and bootstrap techniques. As

[26]A caveat is that I use symmetric tests, as described earlier in (12) and (13). Hall (1992) shows that, because they minimize the influence of skewness, symmetric tests converge to nominal size at twice the rate of asymmetric equal tailed tests. This result turns out to be very relevant in finite samples, as in Monte Carlos (reported in the on-line appendix) I find asymmetric equal-tailed tests are generally less accurate than symmetric tests. However, the disadvantages of asymmetric tests are much more pronounced in the case of the pairs bootstrap. So, if one is interested in using asymmetric tests the wild bootstrap is almost certainly a better choice. Although I use symmetric tests to analyse the sample below, the on-line appendix reports a full set of results for the wild bootstrap, using symmetric and asymmetric tests and transformations, and these results are also summarized in a footnote below.

Table XI: Significance of Coefficients on Instrumented Variables at the .01 Level by Leverage
(average across papers of the fraction of coefficients rejecting the null of 0)

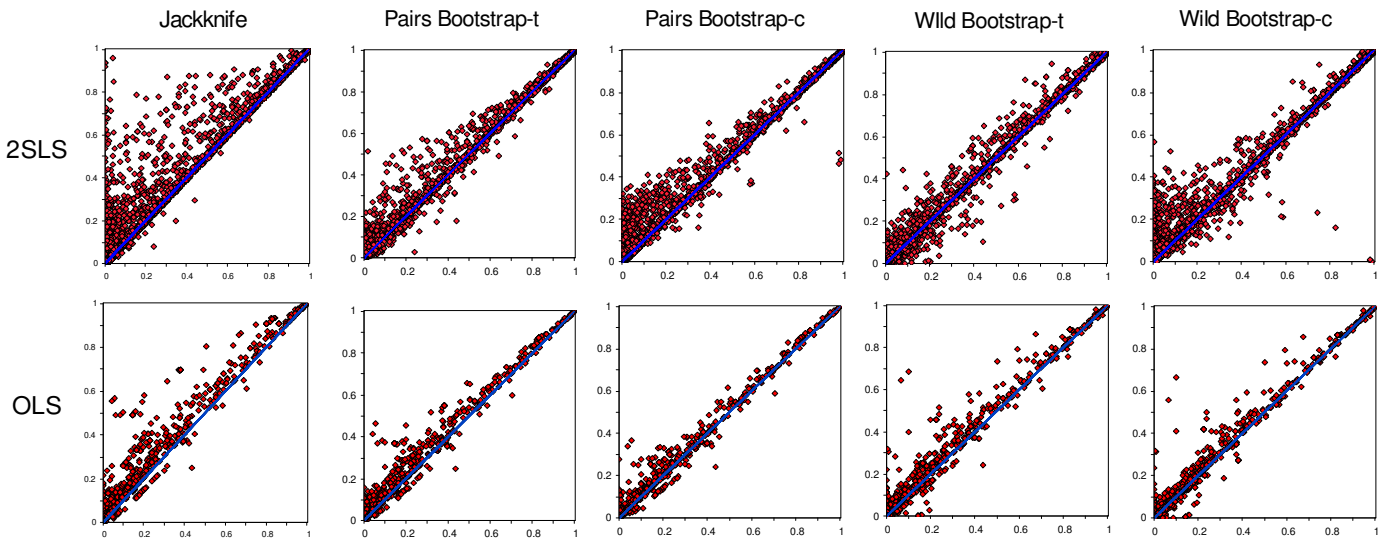
	two-stage least squares				ordinary least squares			
	all	low	medium	high	all	low	medium	high
authors' methods	.364	.532	.236	.336				
clustered/robust	.339	.512	.198	.322	.558	.777	.398	.514
jackknife	.253	.471	.114	.187	.484	.749	.330	.388
pairs bootstrap - t	.253	.444	.115	.215	.495	.741	.331	.431
pairs bootstrap - c	.167	.349	.096	.062	.480	.757	.376	.319
wild bootstrap - t	.345	.598	.241	.208	.473	.767	.345	.320
wild bootstrap - c	.121	.223	.106	.035	.491	.801	.325	.363

Notes: low/medium/high refer to leverage groups, as described in Table III; bootstrap-t implemented using the clustered/robust covariance estimate; iv wild bootstrap using restricted efficient residuals; all bootstrap p-values evaluated using 2000 bootstrap draws; OLS results are for the OLS version of the IV equation

shown, using authors' covariance calculation methods and chosen distribution (normal or t), in the average paper .364 of instrumented coefficients are statistically significant at the .01 level.[27] As authors use diverse methods, in the second row I move things to a consistent framework by using the clustered/ robust covariance matrix and the finite sample t-distribution throughout. This lowers significance rates slightly, mostly because the t-distribution has thicker tails than the normal distribution used by authors in almost half of the regressions in the first row. Use of the jackknife and bootstrap to evaluate significance has a much bigger impact. With the jackknife and nonparametric resampling pairs bootstrap-t, the average fraction of coefficients found to be significant at the .01 level falls to .75 of the level seen in the second row. The reduction is greater (to .49 of the clustered/robust level) with the non-parametric pairs bootstrap-c, and even greater (to .36) with the parametric wild bootstrap-c. These differences are most pronounced in medium and high leverage papers. The wild bootstrap-t finds statistical significance on average in about 1.02 as many regressions as clustered/robust methods, but even this method only registers ⅔ as many significant results in high leverage papers. OLS results for the same estimating equations, as shown in the table, are both more robust and more consistent, with the

[27]The on-line appendix presents .05 level results for all tests reviewed below, as well as results using alternative forms of the wild bootstrap, which are summarized at the end of this section.

Figure IV: Jackknife, Bootstrap & Clustered/Robust P-Values



Notes: X-axes = clustered/robust p-values, Y-axes = jackknife or bootstrap p-values.

jackknife and all forms of the bootstrap registering significance at a rate between .85 and .89 of that found using the t-distribution and clustered/robust covariance estimates. All of this difference, such as it is, is concentrated in medium and high leverage papers.

Figure IV graphs the 2SLS and OLS p-values found using jackknife and bootstrap methods against the corresponding p-values found using clustered/robust covariance estimates and the t-distribution. As shown, for 2SLS the disagreements found using the alternative variance estimate of the jackknife or the coefficient distribution of the bootstrap-c are very large, while those found using the t-statistic distribution in the bootstrap-t tend to be fairly small. When a clustered/robust .01 significant IV coefficient is found to be insignificant at the same level by the jackknife, in the average paper the average p-value rises to .089. Similarly, where disagreement on significance exists, the average p-value rises to .068 with the pairs bootstrap-c and .076 with the wild bootstrap-c. In contrast, where such disagreement arises with the pairs bootstrap-t the average p-value only rises to .031, while the increase in the case of the wild bootstrap-t is merely to .029. Differences in the case of OLS regressions, using all methods, are generally small, as can be seen in the figure.

The very large proportional differences in significance rates found using bootstrap-t and

-c methods are not easily dismissed as reflecting power differences, as the differences in proportional power found in the Monte Carlos earlier above are comparatively small. The results are, however, consistent with the tendency of 2SLS methods (as shown in those same Monte Carlos) to find significance not when point estimates are extreme given the null, but rather when standard errors are unusually small given the sample size and residual variance. The bootstrap-c, whether based upon non-parametric pairs resampling or parametric wild transformations of residuals, indicates that published coefficient estimates are often not unusual under the null of zero average effects. The corresponding bootstrap-t measures, which differ only by considering the standard error, find much larger significance, indicating that published standard error estimates are surprisingly small given the characteristics of the sample. This pattern is confirmed by the jackknife which substitutes its own standard error estimate and finds a substantial increase in p-values. These are the types of results one would expect when "publication bias" selectively picks out the tail outcomes of a method with a low ratio of power to size whose extreme test statistics are completely dominated by the realizations of a highly volatile standard error estimate. It is notable that there is no systematic gap between -t and -c bootstrap results when applied to OLS versions of the estimating equations, which are often not reported and don't form the basis of the publication decision, and where extreme coefficient values feature as prominently as standard error estimates in the tail realizations of test statistics.

Table XII highlights the extraordinary uncertainty surrounding 2SLS estimates. As shown, the conventional clustered/robust .99 2SLS confidence interval contains the OLS point estimate in .863 of the regressions of the typical paper. Jackknife and bootstrapped confidence intervals raise this proportion further, particularly in high leverage papers, where it reaches .98 and .99 using the bootstrap-c.[28] These results reflect the variability of 2SLS estimates, and are

[28]I calculate the jackknife confidence interval by multiplying the jackknife standard error by the critical values of the t-distribution used to evaluate the regression, the bootstrap-c confidence interval from the percentiles of the absolute-value of the coefficient deviation from the null, and the bootstrap-t confidence interval by multiplying the clustered/robust standard error of the original sample by the percentiles of the absolute-value of the distribution of the t-statistic associated with the coefficient deviation from the null.

Table XII: Point Estimates (β) and .99 Confidence Intervals (CI) by Leverage Group

	$\beta_{ols} \in CI_{2sls}$				$\beta_{2sls} \in CI_{ols}$			
	all	low	medium	high	all	low	medium	high
clustered/robust	.863	.806	.937	.840	.332	.180	.389	.422
jackknife	.894	.787	.960	.930	.367	.202	.402	.493
pairs bootstrap - t	.896	.812	.970	.897	.356	.179	.396	.491
pairs bootstrap - c	.926	.832	.960	.981	.365	.199	.388	.506
wild bootstrap - t	.879	.751	.930	.952	.373	.205	.406	.503
wild bootstrap - c	.907	.819	.912	.990	.364	.208	.411	.470

Notes: CI_x refers to .99 confidence interval of coefficients estimated by method x. Otherwise, as in Table XI.

not a consequence of a close similarity between OLS and 2SLS point estimates. First, as shown in the table, OLS confidence intervals contain the 2SLS point estimate only about ⅓ of the time. Second, it is worth noting that in the average paper .17 of 2SLS coefficient estimates are of the *opposite sign* of the OLS estimate for the same equation, while the absolute difference of the 2SLS and OLS point estimates is greater than 0.5 times the absolute value of the OLS point estimate in .74 of regressions and greater than 5.0 times that value in .20 of regressions. 2SLS and OLS point estimates differ substantively, very substantively, in many cases, but statistically the IV estimator rarely rejects the OLS value.

The motivation for using 2SLS stems from the concern that the correlation of endogenous regressors with the error term will produce substantially biased and inconsistent estimates of parameters of interest. Table XIII shows that there is actually limited statistical evidence of this in my sample. I report the Durbin - Wu - Hausman test based upon the Wald statistic formed by the difference between the 2SLS and OLS coefficient estimates. The conventional estimate on average rejects the null in about ¼ of equations. As shown earlier in Monte Carlos, this method has very large size distortions. Not surprisingly, jackknife and bootstrap methods result in much lower average rejection rates, which range between .08 and .16 for the full sample and reach a minimum of .01 (i.e. no larger than nominal size) in the bootstrap-c analysis of high leverage papers. Bootstrap-t methods show higher rejection rates because, once again, the IV variance estimate in the sample itself is found to be surprisingly small producing an unusually large

Table XIII: Rejection Rates in Durbin-Wu-Hausman Tests
at .01 Level by Leverage Group (tests of OLS bias)

	all	low	medium	high
conventional/default	.259	.375	.155	.256
variance	.122	.257	.060	.056
jackknife	.117	.230	.056	.070
pairs bootstrap - t	.087	.201	.053	.010
pairs bootstrap - c	.162	.299	.148	.041
wild bootstrap - t	.082	.183	.056	.010
wild bootstrap - c				

Notes: Default variance estimate used in the conventional and bootstrap-t tests, as the clustered/ robust 2SLS variance estimate is not always greater than the OLS counterpart. Conventional and jackknife test statistics evaluated using the chi^2 distribution. Otherwise, as in Table XI above.

Table XIV: Identification in the First-Stage by Leverage Group
(rejection rates at .01 level in tests of instrument irrelevance)

	all	low	medium	high
clustered/robust	.863	.926	.809	.858
jackknife	.727	.916	.652	.621
pairs bootstrap - t	.652	.880	.593	.489
pairs bootstrap - c	.687	.903	.550	.623
wild bootstrap - t	.672	.903	.621	.497
wild bootstrap – c	.714	.914	.605	.636

Notes: as in Table XI.

Durbin-Wu-Hausman Wald statistic. As already noted, OLS and 2SLS coefficients are often of different sign and the difference in point estimates is, proportionately, often large. However, given the inaccuracy of 2SLS estimation, and the great width of 2SLS confidence intervals, IV estimates in the sample itself provide little guidance as to the magnitude of OLS bias. In the overwhelming majority of regressions reported in published papers, the data actually provide no compelling evidence that use of OLS methods produces substantively biased estimates at all.

Table XIV asks whether 2SLS equations are even identified by testing the null that all first stage coefficients on the excluded exogenous variables are zero. Using the conventional test with the clustered/robust covariance estimate, an average of .863 of first stage regressions in the typical paper reject the null of a rank zero first stage relation at the .01 level. This share falls to

Table XV: Consistency without Inference: 2SLS in Practical Application
(1359 coefficients in 31 papers)

	Durbin-Wu-Hausman & instrument relevance				DWH, instrument relevance, & $\beta_{ols} \notin CI_{2sls}$				DWH, instrument relevance, $\beta_{2sls} \neq 0$			
	all	low	med	high	all	low	med	high	all	low	med	high
(a) average fraction of 2SLS regressions meeting specified criteria at .01 level												
cl/robust	.257	.370	.154	.256	.122	.184	.050	.140	.190	.308	.095	.177
jackknife	.115	.247	.058	.045	.094	.203	.038	.045	.093	.195	.048	.041
pairs boot - t	.106	.220	.053	.051	.074	.175	.028	.022	.075	.177	.043	.006
pairs boot - c	.086	.201	.052	.010	.070	.166	.038	.010	.069	.155	.045	.010
wild boot - t	.146	.275	.138	.028	.101	.235	.061	.009	.122	.248	.109	.009
wild boot - c	.075	.183	.045	.001	.074	.181	.044	.001	.063	.151	.039	.000
(b) number of papers with no 2SLS regressions meeting specified criteria												
cl/robust	12	3	4	5	17	4	7	6	13	3	5	5
jackknife	18	4	7	7	20	5	8	7	20	5	8	7
pairs boot - t	18	4	8	6	21	5	9	7	20	4	8	8
pairs boot - c	20	5	7	8	23	6	9	8	21	5	8	8
wild boot - t	12	2	3	7	17	4	5	8	13	2	3	8
wild boot - c	20	4	7	9	20	4	7	9	22	6	7	9

Notes: Durbin-Wu-Hausman (DWH) = rejecting the null that OLS is unbiased; instrument relevance = rejecting the null that 1st stage coefficients on excluded instruments all equal 0; $\beta_{ols} \notin CI_{2sls}$ = OLS point estimate not in 2SLS .99 confidence interval; $\beta_{2sls} \neq 0$ = rejecting the null that the coefficient on the instrumented right hand side variable equals 0.

between .652 and .727 using bootstrap and jackknife techniques. Once again, differences are most pronounced in medium and high leverage papers, where bootstrap and jackknife rejection rates range from a low of ½ to just under ⅔ of the regressions in the average paper.

Table XV brings the preceding results together. As noted earlier, using authors' methods, results, highlighted by multiple "stars", encourage readers to conclude that 2SLS methods have revealed something about the world. In Table XV I consider alternative criteria for evaluating published results. A good starting point seems to be to require that the Durbin-Wu-Hausman test indicate that there is a statistically significant OLS bias, as the use of inefficient 2SLS when OLS is not substantively biased is a catastrophic error (as shown by the mean squared error comparisons in the Monte Carlos), and, moreover, that one can reject the null hypothesis that the

model is utterly unidentified with all of the first stage coefficients equal to 0, as in this case "identification" is achieved through an undesirable finite sample correlation between the instruments and the error term. Using conventional clustered/robust methods, about ¼ of published results meet these criteria at the .01 level, while only between .075 and .146 can provide these assurances using the much less biased jackknife and bootstrapped tests. Surprisingly, as shown in the bottom panel of the table, depending upon the technique used to evaluate results between ⅓ and ⅔ of published papers have no 2SLS regressions *at all* that reject these important baseline nulls at the .01 level.

With these basic prerequisites for credibility in place, one can then ask whether 2SLS estimates have conveyed new information. Sargan (1958) argued that, given their inefficiency, 2SLS estimates should only be considered if they rule out the OLS point estimate. As Table XV shows, using clustered/robust estimates only .122 of regressions can make this additional claim at the .01 level, and only between .070 and .101 do so using jackknife or bootstrap methods. Even using relatively favourable clustered/robust or wild bootstrap-t methods, more than ½ of the papers in the sample have no regressions at all which meet basic criteria for credibility while producing results that are statistically distinguishable from OLS at the .01 level. Putting aside comparison with OLS, an alternative approach, following the DWH and identification pre-tests, is to ask whether the 2SLS coefficient p-value rejects the null of zero effects, suggesting that, aside from finding that OLS is biased, we have uncovered a meaningful causal relationship. Here the conventional clustered/robust test does somewhat better, finding a significant result in .190 of regressions, while jackknife and bootstrap results vary between .063 and .122. Focusing on the pairs and wild bootstrap-c results, which find significance based upon surprisingly large measured effects rather than unusually small sample standard errors, only about .07 of regressions in the typical paper are both strongly credible and significantly different from the OLS point estimate or zero, while ⅔ of papers have no such results at all. High leverage papers do exceptionally poorly, with an average of between .000 and .010 of results meeting these

criteria, while low leverage papers do vastly better, with between .151 and .181 of all regressions meeting these bootstrap-c based standards.[29]

VII. Conclusion

Contemporary IV practice involves the screening of reported results on the basis of the 1st stage F-statistic, as, beyond argumentation in favour of the exogeneity of instruments, the acceptance of findings rests on evidence of a strong first stage relationship. The results in this paper suggest that this approach is not helpful, and possibly pernicious. Table XVI below reports the Monte Carlo probability of an F greater than 10 appearing in tests of true nulls in my sample. In an ideal iid normal world, using the appropriate default covariance estimate, the probability of an F greater than 10 arising when the instruments are completely irrelevant is a 1 in 1000 event, i.e. a rare occurrence, whether leverage is low, medium or high. This is why an F greater than 10 in that setting suggests the presence of a strong non-zero 1st stage relationship. In a heteroskedastic world, even when clustered/robust covariance estimates are used, this probability is on average between 40 and 60 in 1000, and rises above 100 in 1000 in high leverage papers. Consequently, it is not surprising that in non-iid settings conventional Fs provide none of the bounds and protection on size and bias suggested by asymptotic iid or even non-iid critical values. In a world in which economists experiment with plausible instruments in the privacy of their offices, publicly reported results could easily be filled with instruments

[29]As shown in the on-line appendix, use of alternative forms of the wild bootstrap (asymmetric transformations, asymmetric equal tailed tests or without restricted efficient residuals) whose performance in Monte Carlos is reasonably similar to the wild bootstrap methods used above (symmetric transformations in symmetric tests with restricted efficient residuals for IV coefficients), yields systematically lower rejection rates in 1st stage F-tests and Durbin-Wu-Hausman tests and, inconsistently between c and t versions, higher or lower rejection rates for IV coefficients. Otherwise, results at the .05 level (in the on-line appendix), are somewhat more favourable to the sample, as might be expected. Focusing on the specific methods where the sample does most poorly in the analysis above, Durbin-Wu-Hausman bootstrap-c rejection rates rise from about .08 at the .01 level to .20 at the .05 level, 1st stage bootstrap-t rejection rates rise from ⅔ to about .8, bootstrap-c proportional reductions in the number of significant IV coefficients rise from .36 or .49 to .65, and bootstrap-c estimates of the fraction of IV confidence intervals which include the OLS point estimate fall from .91 or .93 at the .99 level to .79 or .84 at the .95 level. The fraction of regressions meeting the criteria specified in Table XV with the bootstrap-c rises from about .07 at the .01 level to between .14 and .19 at the .05 level. Differences with clustered/robust methods remain concentrated in high and medium leverage papers.

Table XVI: Probability of 1st Stage F > 10 when the Null is True
(1000 Monte Carlo simulations for each of 1359 equations)

	default covariance estimate				clustered/robust covariance estimate			
	all	low	medium	high	all	low	medium	high
iid normal	.001	.001	.001	.001	.011	.002	.009	.022
iid chi^2	.002	.001	.002	.003	.010	.003	.007	.020
h. normal	.215	.194	.121	.341	.039	.002	.015	.102
h. chi^2	.217	.195	.126	.341	.051	.006	.025	.125
h. & cl. normal	.433	.473	.239	.606	.045	.002	.017	.120
h. & cl. chi^2	.431	.475	.243	.592	.061	.010	.028	.148

Notes: Average across papers of average within paper rejection rates; low, medium and high divide the sample into thirds, based upon average leverage, as in Table III earlier.

which, while legitimately exogenous in the population, are nevertheless irrelevant or very nearly so, with the strong reported F being the result of an unfortunate finite sample correlation with the endogenous disturbances, producing undesirably biased estimates. The widespread and growing use of test statistics with underappreciated fat tails to gain credibility is less than ideal.

Economists use 2SLS methods because they wish to gain a more accurate estimate of parameters of interest than provided by biased OLS. In this regard, explicit consideration of the degree to which 2SLS results are distinguishable from OLS seems natural, a point raised early on by Sargan in his seminal (1958) paper. In the analysis of the sample, above, I find that 2SLS rarely rejects the OLS point estimate or is able to provide strong statistical evidence against OLS being unbiased, despite the fact that 2SLS point estimates are often of the opposite sign or substantially different magnitude. This is virtually always true in high leverage papers, but is even true in the low leverage sample, where ½ of papers provide no regressions that can reject a null of zero OLS bias or exclude OLS point estimates at the .01 level. These results do not indicate that OLS point estimates are actually unbiased, but merely that 2SLS is sufficiently inefficient that, despite yielding substantially different point estimates, it does not often provide meaningfully different information. Learning about the world may simply be harder than suggested by simple dichotomies between good and bad research design.

BIBLIOGRAPHY[30]

Albouy, David Y. 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment." *American Economic Review*, 102 (6): 3059-3076.

Anderson, T.W. and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46-63.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3-30.

Baum, Christopher F., Mark E. Schaffer and Steven Stillman. 2007. "Enhanced routines for instrumental variables/generalized method of moments estimation and testing." *Stata Journal* 7 (4): 465-506.

Bazzi, Samuel, and Michael A. Clemens. 2013. "Blunt Instruments: Avoiding Common Pitfalls in Identifying the Causes of Economic Growth." *American Economic Journal: Macroeconomics*, 5(2): 152–186.

Chernozhukov, Victor and Christian Hansen. 2008. "The Reduced Form: A Simple Approach to Inference with Weak Instruments." *Economics Letters* 100 (1): 68-71.

Davidson, Russell and James G. MacKinnon. 2010. "Wild Bootstrap Tests for IV Regression." *Journal of Business & Economic Statistics* 28 (1): 128-144.

Dufour, Jean-Marie. 2003. "Identification, Weak-Instruments, and Statistical Inference in Econometrics." *Canadian Journal of Economics* 36 (4): 767-808.

Durbin, J. 1954. "Errors in Variables." *Review of the International Statistical Institute* 22: 23-32.

Eicker. F. 1963. "Asymptotic normality and consistency of the least squares estimators for families of linear regressions." *Annals of Mathematical Statistics* 34 (2): 447-456.

Fuller, Wayne A. 1977. "Some Properties of a Modification of the Limited Information Estimator." *Econometrica* 45 (4): 939-953.

Hall, Peter. 1992. The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag, 1992.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (6): 1251-1271.

Hinkley, David V. 1977. "Jackknifing in Unbalanced Situations." *Technometrics* 19 (3): 285-292.

Kinal, Terrence W. 1980. "The Existence of Moments of k-Class Estimators." *Econometrica* 48 (1): 241-249.

Lehmann, E.L and Joseph P. Romano. 2005. Testing Statistical Hypotheses. Third edition. New York: Springer Science + Business Media, 2005.

[30]Sources cited in this paper. See on-line appendix for list of papers in the sample.

MacKinnon, James G. and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics* 29 (): 305-325.

Nelson, Charles R. and Richard Startz. 1990a. "The Distribution of the Instrumental Variable Estimator and Its t Ratio When the Instrument Is a Poor One." *Journal of Business* 63 (1): S125–S140.

Nelson, Charles R. and Richard Startz. 1990b. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator." *Econometrica* 58 (4): 967–976.

Olea, Jose Luis Montiel and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." Journal of Business and Economic Statistics 31 (3): 358-369.

Sargan, J.D. 1958. "The Estimation of Economic Relationships using Instrumental Variables." *Econometrica* 26 (3): 393-415.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557-586.

Stock, James H. and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In Andrews, Donald W.K. and James H. Stock, eds, Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. New York: Cambridge University Press.

Summers, Robert. 1965. "A Capital Intensive Approach to the Small Sample Properties of Various Simultaneous Equation Estimators." *Econometrica* 33 (1): 1-41.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817-838.

Wu, De-Min. 1973. "Alternative Tests of Independence between Stochastic Regressors and Disturbances." *Econometrica* 41 (4): 733-750.