On-Line Appendix for Consistency without Inference: Instrumental Variables in Practical Application

Alwyn Young
London School of Economics; March 2022

Contents:

A: Sensitivity Tests for Tables IV - XVI in the Paper

This appendix presents sensitivity tests for Tables IV through X in the paper. Table IV in the paper reported Type I error rates and power estimates for 2SLS and OLS using Monte Carlos with normal and "actual" errors, the data generating processes described in 9.1-9.3 and 11.1-11.3 in the paper. Table A1 below adds in the results with chi^2 errors (processes 9.4-9.6 described in the paper). Size distortions are somewhat larger with chi^2 errors, but otherwise the patterns are those described in the paper: Type I error rates above nominal level with non-iid errors are not unique to IV; power declines more, both absolutely and proportionately, with non-iid errors in IV than in OLS; IV is a noticeably less efficient estimator with much lower power when errors are uncorrelated (OLS unbiased); and when errors are correlated, precise but biased OLS estimates give rise to huge size distortions.

Table A1: Average Null Rejection Probabilities at the .01 & .05 Levels
(sensitivity test for Table IV in the paper)

	$H_0=\beta_{dgp}$										$H_0=0$									
	all		low		medium		high		all		all		low		medium		high		all	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
	(a) correlated errors (all results)																			
	2SLS								OLS		2SLS								OLS	
iid normal	.029	.077	.011	.049	.036	.082	.039	.101	.718	.782	.461	.590	.578	.694	.285	.440	.519	.636	.579	.682
h normal	.069	.126	.012	.045	.052	.106	.142	.226	.528	.613	.276	.375	.372	.457	.132	.249	.324	.418	.430	.534
h cl normal	.069	.123	.010	.040	.054	.106	.143	.224	.439	.535	.182	.272	.256	.333	.107	.212	.183	.271	.379	.488
iid chi^2	.026	.075	.012	.051	.032	.076	.035	.097	.672	.759	.477	.603	.579	.697	.314	.461	.539	.650	.508	.605
h chi^2	.077	.139	.015	.049	.066	.126	.152	.242	.502	.603	.287	.392	.368	.457	.151	.275	.342	.445	.424	.525
n cl chi^2	.083	.144	.018	.051	.069	.130	.160	.250	.436	.546	.196	.294	.259	.341	.122	.236	.206	.304	.383	.488
iid "actual"	.025	.072	.014	.051	.021	.064	.039	.101	.693	.771	.432	.553	.585	.702	.231	.362	.481	.594	.525	.623
h "actual"	.035	.085	.010	.046	.042	.092	.053	.117	.678	.752	.409	.539	.583	.708	.207	.349	.436	.559	.491	.593
h cl "actual"	.037	.085	.012	.047	.044	.094	.056	.115	.668	.747	.297	.446	.478	.634	.158	.307	.255	.395	.483	.589
	(b) correlated errors (headline results)																			
	2SLS								OLS		2SLS								OLS	
iid normal	.023	.072	.008	.045	.023	.075	.038	.097	.728	.788	.566	.701	.630	.754	.528	.677	.541	.672	.610	.699
h normal	.060	.118	.009	.044	.037	.096	.136	.214	.520	.606	.348	.455	.387	.483	.277	.404	.379	.478	.444	.552
h cl normal	.062	.117	.007	.039	.044	.099	.136	.213	.424	.520	.229	.333	.284	.372	.223	.349	.179	.278	.374	.495
iid chi^2	.020	.066	.010	.047	.018	.066	.033	.086	.686	.766	.576	.711	.615	.751	.551	.694	.560	.689	.544	.629
h chi^2	.070	.133	.012	.047	.054	.125	.144	.228	.492	.591	.353	.463	.386	.496	.294	.409	.380	.485	.428	.539
n cl chi^2	.075	.137	.017	.053	.058	.123	.149	.235	.414	.528	.234	.342	.270	.373	.240	.359	.192	.294	.372	.489
iid "actual"	.021	.070	.007	.043	.023	.076	.031	.088	.698	.780	.537	.661	.633	.761	.458	.582	.521	.641	.560	.649
h "actual"	.038	.089	.009	.045	.057	.116	.049	.106	.691	.752	.518	.660	.604	.764	.473	.612	.479	.604	.550	.630
h cl "actual"	.046	.092	.010	.042	.062	.120	.067	.114	.708	.769	.404	.573	.489	.689	.431	.577	.293	.453	.570	.644

Table A1: Average Null Rejection Probabilities at the .01 & .05 Levels (continued)

	$H_0=\beta_{dgp}$										$H_0=0$									
	all		low		medium		high		all		all		low		medium		high		all	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
(c) uncorrelated errors (all results)																				
	OLS								2SLS		OLS								2SLS	
iid normal	.012	.053	.010	.048	.013	.056	.013	.055	.018	.063	.830	.874	.947	.963	.740	.812	.802	.848	.472	.594
h normal	.068	.140	.013	.055	.048	.121	.143	.245	.054	.107	.648	.727	.850	.895	.542	.646	.552	.640	.295	.394
h cl normal	.078	.155	.017	.064	.056	.132	.161	.268	.053	.103	.570	.665	.749	.816	.500	.612	.460	.567	.200	.290
iid chi^2	.013	.056	.011	.049	.015	.060	.014	.060	.017	.061	.835	.878	.950	.965	.750	.820	.805	.850	.489	.609
h chi^2	.088	.163	.023	.071	.075	.152	.166	.266	.065	.126	.696	.768	.868	.906	.588	.682	.631	.715	.309	.412
n cl chi^2	.097	.177	.028	.082	.080	.162	.183	.286	.067	.125	.623	.712	.778	.840	.548	.652	.543	.643	.214	.311
iid "actual"	.017	.058	.025	.066	.012	.052	.013	.057	.017	.058	.832	.875	.949	.965	.732	.805	.814	.855	.449	.566
h "actual"	.042	.103	.031	.078	.035	.096	.061	.134	.028	.076	.797	.851	.922	.944	.684	.772	.786	.836	.429	.552
h cl "actual"	.056	.123	.027	.075	.040	.102	.101	.193	.026	.071	.740	.815	.893	.920	.638	.753	.689	.772	.337	.468

Notes: As in Table IV in the paper.

Table A2: Ln Truncated OLS Bias & Relative 2SLS to OLS Bias & Mean Squared Error (sensitivity test for Table V in the paper)

	$\lvert \hat{\beta} \rvert < 1000 * \lvert \beta_{dgp} \rvert$						$\lvert \hat{\beta} \rvert < 10 * \lvert \beta_{dgp} \rvert$					
	OLS bias	relative bias				relative mse	OLS bias	relative bias				relative mse
	all	all	low	medium	high	all	all	all	low	medium	high	all
(a) all results												
iid normal	-0.5	-3.4	-4.0	-2.5	-3.8	-0.3	-0.5	-3.4	-4.0	-2.5	-3.8	-0.6
h normal	-0.5	-2.0	-2.8	-1.6	-1.7	1.9	-0.6	-2.3	-3.0	-1.9	-2.0	0.5
h cl normal	-0.5	-1.1	-1.9	-1.3	-0.2	3.3	-0.6	-1.7	-2.4	-1.4	-1.2	1.2
iid chi^2	-0.5	-3.4	-3.8	-2.6	-3.9	-0.4	-0.5	-3.4	-3.8	-2.6	-3.9	-0.7
h chi^2	-0.4	-2.1	-2.7	-1.6	-2.1	1.4	-0.5	-2.3	-3.0	-1.7	-2.3	0.3
n cl chi^2	-0.4	-1.4	-2.0	-1.4	-0.7	2.9	-0.5	-1.6	-2.3	-1.4	-1.2	1.0
iid "actual"	-0.4	-3.3	-3.9	-2.1	-3.8	0.1	-0.4	-3.4	-4.0	-2.3	-3.9	-0.5
h "actual"	-0.4	-3.0	-3.8	-2.0	-3.1	0.4	-0.5	-3.0	-3.9	-2.1	-3.1	-0.3
h cl "actual"	-0.4	-2.1	-3.0	-1.6	-1.8	1.3	-0.5	-2.3	-3.1	-1.8	-2.2	0.3
(b) headline results												
iid normal	-0.7	-3.6	-4.3	-2.8	-3.7	-0.8	-0.7	-3.6	-4.3	-2.8	-3.7	-0.9
h normal	-0.8	-2.1	-3.2	-1.3	-1.8	1.7	-0.8	-2.4	-3.3	-1.7	-2.2	0.2
h cl normal	-0.8	-1.2	-2.2	-1.1	-0.3	3.2	-0.8	-1.8	-2.8	-1.4	-1.1	1.1
iid chi^2	-0.7	-3.3	-3.6	-2.9	-3.5	-0.9	-0.7	-3.4	-3.7	-3.0	-3.5	-1.0
h chi^2	-0.7	-2.2	-3.4	-1.6	-1.8	1.0	-0.7	-2.5	-3.6	-1.7	-2.3	0.0
n cl chi^2	-0.7	-1.3	-2.0	-1.3	-0.6	2.5	-0.7	-1.7	-2.3	-1.5	-1.2	0.9
iid "actual"	-0.5	-3.8	-4.2	-3.3	-3.8	-0.7	-0.5	-3.9	-4.4	-3.4	-3.9	-1.0
h "actual"	-0.6	-3.4	-4.0	-2.7	-3.6	-0.4	-0.6	-3.5	-4.0	-2.7	-3.6	-0.7
h cl "actual"	-0.6	-2.6	-3.2	-2.3	-2.3	0.4	-0.6	-2.5	-3.0	-2.3	-2.3	-0.1

Notes: As in Table V in the paper.

Table A2 above adds chi^2 errors to Table V's analysis in the paper of relative bias and mean squared error with correlated errors. The patterns with chi^2 errors are very much the same: IV's relative bias advantage falls with non-iid errors while IV mse on average becomes greater than that found in OLS. An appendix further below shows that the change in relative bias with non-iid error processes is positively related to maximum leverage.

Tables VI and VII in the paper examined the effectiveness of the Stock & Yogo (2005) size and bias tests using normal and "actual" errors, and in some cases only for the smallest and largest size and bias bounds given by Stock & Yogo. Tables A3 and A4 below extend the analysis to include chi^2 errors and all of the bounds provided by Stock & Yogo. Results for size bounds with chi^2 errors are generally worse than those found with normal errors, with a higher ratio of the fraction of regressions exceeding the desired size bound in H_1 (strong instrument) to the fraction found in H_0 (weak instrument). Chi^2 results with regards to bias are similar to those found with normal errors. Results for intermediate bounds on size and bias lie between the smallest and largest bounds, as might be expected.

Table A3: Fraction of Regressions with Null Rejection Probabilities Greater than Size Bound in Specifications that Don't (H_0) and Do (H_1) Reject the Stock & Yogo Weak Instrument Null (sensitivity test for Table VI in the paper)

	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 IV coefficient covariance estimate, with default F used as Stock and Yogo test statistic								
iid normal	.126	.000 (.022)	.094	.000 (.013)	.067	.000 (.010)	.053	.000 (.009)
iid chi^2	.141	.001 (.022)	.087	.000 (.013)	.062	.000 (.010)	.048	.000 (.009)
iid "actual"	.085	.003 (.028)	.058	.002 (.017)	.036	.002 (.013)	.040	.002 (.011)
(B) cl/robust IV coefficient covariance estimate, with default F used as Stock and Yogo test statistic								
iid normal	.258	.267 (.022)	.106	.025 (.013)	.062	.014 (.010)	.058	.009 (.009)
h normal	.425	.268 (.020)	.201	.125 (.014)	.097	.077 (.012)	.042	.061 (.011)
h cl normal	.415	.449 (.019)	.270	.358 (.014)	.134	.176 (.012)	.050	.083 (.011)
iid chi^2	.216	.276 (.022)	.074	.024 (.013)	.062	.014 (.010)	.053	.008 (.009)
h chi^2	.565	.448 (.019)	.283	.191 (.012)	.141	.134 (.010)	.047	.075 (.009)
h cl chi^2	.574	.602 (.018)	.319	.432 (.012)	.178	.364 (.010)	.096	.217 (.009)
iid "actual"	.251	.269 (.028)	.058	.025 (.017)	.036	.019 (.013)	.036	.011 (.011)
h "actual"	.254	.389 (.026)	.136	.074 (.017)	.108	.057 (.014)	.091	.045 (.012)
h cl "actual"	.316	.385 (.026)	.159	.192 (.017)	.117	.135 (.014)	.094	.099 (.012)
(C) cl/robust IV coefficient covariance estimate, with cl/robust F used as Stock and Yogo test statistic								
iid normal	.247	.270 (.019)	.118	.024 (.011)	.068	.014 (.009)	.063	.009 (.008)
h normal	.394	.247 (.041)	.185	.119 (.027)	.087	.078 (.021)	.045	.062 (.018)
h cl normal	.470	.383 (.101)	.351	.327 (.059)	.159	.176 (.045)	.055	.094 (.037)
iid chi^2	.215	.273 (.017)	.083	.023 (.011)	.069	.014 (.009)	.058	.008 (.008)
h chi^2	.534	.439 (.038)	.262	.183 (.025)	.142	.132 (.019)	.051	.077 (.016)
h cl chi^2	.589	.605 (.077)	.379	.438 (.047)	.277	.372 (.036)	.163	.220 (.031)
iid "actual"	.236	.275 (.022)	.069	.024 (.014)	.041	.018 (.011)	.041	.011 (.009)
h "actual"	.244	.398 (.028)	.139	.072 (.018)	.114	.055 (.014)	.098	.043 (.012)
h cl "actual"	.349	.378 (.060)	.203	.171 (.036)	.153	.120 (.029)	.128	.084 (.025)

Notes: As in Table VI in the paper.

Table A4: Fraction of Regressions with Relative Bias Greater than Bias Bound in Specifications that Don't and Do Reject the Stock & Yogo Weak Instrument Null (sensitivity test for Table VII in the paper) -

	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	.988	.153 (.162)	.902	.091 (.145)	.878	.052 (.106)	.668	.043 (.068)
h normal	.992	.216 (.137)	.998	.396 (.085)	.960	.522 (.042)	.768	.415 (.025)
h cl normal	.995	.869 (.114)	.997	.828 (.072)	.963	.848 (.037)	.833	.762 (.023)
iid chi^2	.993	.140 (.162)	.910	.069 (.145)	.847	.055 (.105)	.676	.065 (.065)
h chi^2	.982	.366 (.105)	.962	.445 (.065)	.864	.502 (.033)	.515	.296 (.021)
h cl chi^2	.976	.819 (.090)	.955	.803 (.055)	.857	.766 (.029)	.562	.579 (.019)
iid "actual"	.971	.139 (.181)	.911	.052 (.156)	.850	.036 (.112)	.705	.040 (.084)
h "actual"	.961	.116 (.178)	.925	.146 (.151)	.784	.136 (.100)	.580	.176 (.067)
h cl "actual"	.966	.671 (.193)	.941	.689 (.143)	.771	.480 (.101)	.589	.402 (.069)
(B) clustered/robust F used as Stock and Yogo test statistic								
iid normal	.991	.174 (.155)	.914	.127 (.130)	.878	.261 (.066)	.655	.248 (.030)
h normal	.984	.649 (.032)	.988	.699 (.017)	.966	.674 (.009)	.546	.528 (.006)
h cl normal	.972	.944 (.034)	.973	.910 (.017)	.982	.880 (.010)	.970	.759 (.007)
iid chi^2	.997	.171 (.151)	.916	.196 (.112)	.825	.386 (.043)	.601	.337 (.018)
h chi^2	.976	.678 (.026)	.974	.656 (.016)	.938	.591 (.009)	.572	.334 (.006)
h cl chi^2	.937	.908 (.027)	.950	.859 (.017)	.951	.771 (.010)	.940	.530 (.006)
iid "actual"	.989	.172 (.157)	.932	.116 (.128)	.864	.225 (.070)	.725	.273 (.033)
h "actual"	.983	.105 (.163)	.957	.122 (.136)	.818	.128 (.088)	.625	.181 (.052)
h cl "actual"	.966	.604 (.252)	.944	.661 (.159)	.788	.466 (.093)	.615	.396 (.054)

Notes: As in Table VII in the paper.

Stock & Yogo (2005) base their theory around Wald and F-statistics calculated with finite sample corrections (pp. 83-84) but p-values based upon the asymptotic chi^2 distribution (pp. 88), so I follow this approach in Table VI in the paper (as noted in the table's notes) and Table A3 above. Table A5 below reports results using the t-distribution with finite sample degrees of freedom corrections to calculate IV p-values and size. As expected, the fraction of regressions with Type I error probabilities greater than the specified levels falls with these corrections (compare to Table A3), but the patterns are identical to those reported in the paper. In particular, with non-iid errors the fraction of regressions with Type I error probabilities greater than the specified level is often higher in H_1 regressions that reject the weak instrument null than it is in H_0 regressions that do not, and is always much greater than the maximum share that would be consistent with the test itself having .05 size.

Table A5: Stock & Yogo Size Tests with P-Values Calculated using t-Distribution
(sensitivity test for Table VI in the paper)

	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 IV coefficient covariance estimate, with default F used as Stock and Yogo test statistic								
iid normal	.116	.000 (.022)	.075	.000 (.013)	.067	.000 (.010)	.048	.000 (.009)
iid chi^2	.128	.001 (.022)	.075	.000 (.013)	.062	.000 (.010)	.048	.000 (.009)
iid "actual"	.083	.003 (.028)	.055	.002 (.017)	.036	.002 (.013)	.036	.002 (.011)
(B) cl/robust IV coefficient covariance estimate, with default F used as Stock and Yogo test statistic								
iid normal	.209	.216 (.022)	.094	.024 (.013)	.062	.014 (.010)	.053	.001 (.009)
h normal	.400	.234 (.020)	.181	.108 (.014)	.093	.073 (.012)	.039	.059 (.011)
h cl normal	.394	.442 (.019)	.240	.333 (.014)	.116	.165 (.012)	.049	.079 (.011)
iid chi^2	.182	.228 (.022)	.074	.023 (.013)	.062	.014 (.010)	.053	.005 (.009)
h chi^2	.533	.413 (.019)	.253	.175 (.012)	.127	.123 (.010)	.045	.066 (.009)
h cl chi^2	.552	.560 (.018)	.295	.420 (.012)	.168	.351 (.010)	.087	.198 (.009)
iid "actual"	.190	.224 (.028)	.055	.024 (.017)	.036	.018 (.013)	.036	.006 (.011)
h "actual"	.212	.313 (.026)	.120	.065 (.017)	.107	.056 (.014)	.085	.042 (.012)
h cl "actual"	.291	.351 (.026)	.147	.182 (.017)	.111	.132 (.014)	.088	.092 (.012)
(C) cl/robust IV coefficient covariance estimate, with cl/robust F used as Stock and Yogo test statistic								
iid normal	.205	.217 (.019)	.105	.023 (.011)	.068	.014 (.009)	.058	.001 (.008)
h normal	.367	.211 (.041)	.163	.103 (.027)	.085	.074 (.021)	.043	.060 (.018)
h cl normal	.456	.377 (.101)	.324	.303 (.059)	.144	.166 (.045)	.052	.091 (.038)
iid chi^2	.189	.223 (.017)	.083	.022 (.011)	.069	.014 (.009)	.058	.005 (.008)
h chi^2	.501	.403 (.038)	.231	.171 (.025)	.127	.123 (.019)	.046	.068 (.016)
h cl chi^2	.563	.552 (.077)	.357	.431 (.047)	.267	.358 (.036)	.147	.202 (.031)
iid "actual"	.180	.226 (.022)	.065	.023 (.014)	.041	.017 (.011)	.041	.006 (.010)
h "actual"	.209	.317 (.028)	.123	.063 (.018)	.113	.054 (.014)	.092	.040 (.012)
h cl "actual"	.328	.335 (.060)	.192	.162 (.036)	.147	.117 (.029)	.119	.079 (.025)

Notes: As in Table VI in the paper.

Table A6 below divides the results for the Stock & Yogo size test by leverage group, a sensitivity test for Table VI in the paper. With iid error processes and the default covariance estimate used to evaluate F-statistics and calculate IV standard errors, the test, as shown in panel A of the table, does well in all leverage groups, although only the medium leverage group has substantial weak instrument induced size distortions. With clustered/robust covariance estimates used to calculate IV standard errors, results in the medium and high leverage groups are extraordinarily poor, whether or not default (panel B) or clustered/robust (panel C) covariance estimates are used in the calculation of the 1^{st} stage test statistic, as with non-iid errors Type I error probabilities are often as large or greater in the H_1 "strong instrument" group than in the H_0 group that fails to reject the weak instrument null. The test does appear to work better in non-iid settings in low leverage papers (panels B and C), but this is largely a consequence of the fact that size distortions with clustered/robust covariance estimates in these papers are almost always very low for both H_0 and H_1 regressions. In the low leverage cases where rejection probabilities greater than nominal value appear, size distortions in H_1 papers with non-iid errors in panels B and C are occasionally as high as in H_0 regressions and very often above the level consistent with the Stock & Yogo test itself having .05 size.

As noted in the paper, the results for Stock & Yogo's bias test cannot be meaningfully broken down by leverage group. The 134 regressions for which Stock & Yogo provide bias bounds only cover one high leverage paper and 3 low leverage papers, and in the latter almost all observations, but for those from one regression, exceed the bounds.

Table A6: Stock & Yogo Size Tests by Leverage Group
(sensitivity test for Table VI in the paper)

	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 covariance estimate used in 1st stage test statistic and to evaluate coefficient significance								
low								
iid normal	.042	.000 (.011)	.045	.000 (.005)	.062	.000 (.003)	.000	.000 (.003)
iid chi^2	.064	.000 (.011)	.045	.000 (.005)	.000	.000 (.003)	.000	.000 (.003)
iid "actual"	.022	.009 (.011)	.046	.008 (.005)	.000	.008 (.003)	.000	.008 (.003)
medium								
iid normal	.160	.001 (.063)	.106	.001 (.037)	.070	.000 (.029)	.059	.000 (.025)
iid chi^2	.174	.003 (.061)	.097	.001 (.036)	.070	.000 (.028)	.053	.000 (.025)
iid "actual"	.113	.001 (.082)	.063	.000 (.050)	.037	.000 (.038)	.041	.000 (.032)
high								
iid normal	.000	.000 (.006)	.000	.000 (.002)	.000	.000 (.001)	.000	.000 (.001)
iid chi^2	.021	.000 (.006)	.000	.000 (.002)	.000	.000 (.001)	.000	.000 (.001)
iid "actual"	.013	.000 (.010)	.032	.000 (.004)	.053	.000 (.002)	.072	.000 (.002)

Notes: Low, medium, high refer to papers grouped on the basis of average maximum leverage, as in Table II in the paper. Otherwise, as in Table VI in the paper.

Table A6: Stock & Yogo Size Tests by Leverage Group - continued
(sensitivity test for Table VI in the paper)

	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)
(B) default covariance estimate used in 1st stage test statistic, clustered/robust covariance estimate used to evaluate coefficient significance								
low								
iid normal	.100	.033 (.011)	.045	.000 (.005)	.062	.000 (.003)	.000	.000 (.003)
h normal	.061	.046 (.013)	.027	.001 (.007)	.031	.001 (.005)	.034	.001 (.005)
h cl normal	.041	.009 (.012)	.043	.006 (.008)	.022	.001 (.007)	.022	.002 (.006)
iid chi^2	.069	.036 (.011)	.045	.000 (.005)	.063	.000 (.003)	.000	.000 (.003)
h chi^2	.107	.045 (.012)	.028	.005 (.007)	.033	.001 (.005)	.037	.001 (.004)
h cl chi^2	.061	.051 (.012)	.021	.010 (.008)	.022	.001 (.007)	.022	.002 (.006)
iid "actual"	.050	.039 (.011)	.046	.008 (.005)	.000	.008 (.003)	.000	.007 (.003)
h "actual"	.048	.035 (.011)	.045	.004 (.005)	.000	.000 (.003)	.000	.000 (.003)
h cl "actual"	.137	.034 (.012)	.034	.005 (.006)	.000	.000 (.004)	.000	.000 (.004)
medium								
iid normal	.279	.031 (.063)	.119	.001 (.037)	.064	.000 (.029)	.065	.000 (.025)
h normal	.466	.222 (.049)	.217	.127 (.034)	.101	.036 (.029)	.037	.010 (.025)
h cl normal	.439	.271 (.040)	.285	.192 (.029)	.144	.067 (.025)	.051	.019 (.022)
iid chi^2	.236	.015 (.061)	.082	.002 (.036)	.065	.000 (.028)	.059	.000 (.025)
h chi^2	.618	.769 (.049)	.307	.212 (.032)	.143	.120 (.026)	.040	.021 (.023)
h cl chi^2	.633	.515 (.041)	.352	.225 (.028)	.183	.170 (.023)	.093	.065 (.021)
iid "actual"	.269	.021 (.082)	.063	.000 (.050)	.037	.000 (.038)	.035	.000 (.032)
h "actual"	.309	.490 (.072)	.162	.140 (.049)	.126	.102 (.039)	.104	.081 (.033)
h cl "actual"	.377	.497 (.067)	.180	.186 (.045)	.134	.130 (.036)	.106	.083 (.031)
high								
iid normal	.284	.511 (.006)	.000	.053 (.002)	.000	.031 (.001)	.000	.021 (.001)
h normal	.619	.404 (.005)	.275	.188 (.003)	.134	.146 (.003)	.088	.131 (.002)
h cl normal	.640	.794 (.007)	.400	.664 (.005)	.189	.352 (.004)	.075	.177 (.003)
iid chi^2	.242	.540 (.006)	.010	.051 (.002)	.000	.031 (.001)	.000	.019 (.001)
h chi^2	.861	.448 (.004)	.468	.269 (.002)	.335	.211 (.001)	.194	.155 (.001)
h cl chi^2	.857	.928 (.005)	.510	.794 (.003)	.407	.696 (.002)	.263	.447 (.002)
iid "actual"	.298	.518 (.010)	.032	.048 (.004)	.053	.036 (.002)	.072	.021 (.002)
h "actual"	.162	.524 (.010)	.009	.072 (.004)	.012	.058 (.003)	.014	.044 (.002)
h cl "actual"	.199	.509 (.010)	.118	.296 (.005)	.066	.212 (.003)	.060	.163 (.002)

Notes: Low, medium, high refer to papers grouped on the basis of average maximum leverage, as in Table II in the paper. Otherwise, as in Table VI in the paper.

Table A6: Stock & Yogo Size Tests by Leverage Group - continued
(sensitivity test for Table VI in the paper)

	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)
(C) clustered/robust covariance estimate used in 1st stage test statistic and to evaluate coefficient significance								
low								
iid normal	.079	.038 (.011)	.045	.000 (.004)	.061	.000 (.003)	.000	.000 (.003)
h. normal	.043	.054 (.039)	.012	.000 (.025)	.014	.000 (.020)	.015	.000 (.017)
h. cl. normal	.027	.000 (.064)	.021	.001 (.052)	.008	.000 (.045)	.008	.000 (.042)
iid chi^2	.059	.038 (.011)	.046	.000 (.005)	.063	.000 (.003)	.000	.000 (.003)
h. chi^2	.055	.058 (.040)	.011	.006 (.026)	.013	.000 (.020)	.015	.000 (.017)
h. cl. chi^2	.053	.052 (.063)	.013	.010 (.052)	.008	.000 (.046)	.008	.000 (.042)
iid "actual"	.046	.040 (.011)	.046	.008 (.005)	.000	.008 (.003)	.000	.008 (.003)
h "actual"	.035	.038 (.013)	.036	.004 (.006)	.000	.000 (.004)	.000	.000 (.004)
h cl "actual"	.068	.044 (.031)	.018	.005 (.014)	.000	.000 (.010)	.000	.000 (.008)
medium								
iid normal	.284	.055 (.050)	.133	.001 (.030)	.071	.001 (.025)	.070	.001 (.022)
h normal	.427	.192 (.089)	.211	.115 (.051)	.091	.036 (.038)	.033	.011 (.031)
h cl normal	.410	.196 (.116)	.302	.133 (.063)	.141	.049 (.046)	.046	.016 (.038)
iid chi^2	.245	.035 (.046)	.091	.002 (.030)	.072	.000 (.024)	.064	.001 (.021)
h chi^2	.574	.872 (.075)	.284	.218 (.045)	.141	.119 (.033)	.035	.022 (.027)
h cl chi^2	.565	.570 (.095)	.351	.186 (.054)	.224	.135 (.040)	.117	.044 (.033)
iid "actual"	.267	.059 (.062)	.072	.000 (.038)	.042	.000 (.030)	.040	.000 (.026)
h "actual"	.326	.459 (.065)	.180	.127 (.042)	.144	.092 (.034)	.120	.073 (.030)
h cl "actual"	.374	.509 (.076)	.169	.196 (.050)	.128	.134 (.040)	.105	.083 (.035)
high								
iid normal	.198	.510 (.004)	.000	.053 (.001)	.000	.031 (.001)	.000	.021 (.000)
h normal	.650	.355 (.016)	.283	.174 (.011)	.153	.145 (.009)	.118	.131 (.008)
h cl normal	.728	.880 (.114)	.576	.716 (.058)	.266	.401 (.042)	.092	.223 (.035)
iid chi^2	.204	.528 (.003)	.005	.050 (.001)	.000	.031 (.001)	.000	.019 (.000)
h chi^2	.915	.344 (.016)	.475	.237 (.010)	.303	.201 (.007)	.157	.156 (.006)
h cl chi^2	.898	.953 (.069)	.656	.869 (.038)	.548	.759 (.028)	.348	.483 (.023)
iid "actual"	.224	.510 (.006)	.053	.047 (.002)	.086	.036 (.001)	.115	.021 (.001)
h "actual"	.123	.559 (.014)	.031	.073 (.008)	.026	.058 (.005)	.025	.044 (.004)
h cl "actual"	.422	.508 (.066)	.293	.272 (.040)	.219	.194 (.031)	.184	.145 (.027)

Notes: Low, medium, high refer to papers grouped on the basis of average maximum leverage, as in Table II in the paper. Otherwise, as in Table VI in the paper.

Table VIII in the paper examined the effectiveness of the Olea & Pflueger (2013) bias test in overidentified equations (where the finite sample 1st moment exists with normal errors) using normal and "actual" errors. Table A7 presents results including chi^2 errors. As noted in the paper, the test performs somewhat worse with artificial chi^2 errors, with bias levels in the low and medium leverage sample and non-iid errors always exceeding the maximum bound consistent with the test having a Type-I error rate of .05. Table A8 below applies the Olea & Pflueger bias test to the exactly identified equations in my Monte Carlo simulations. As the finite sample IV coefficients in these equations most likely do not have a 1st moment, I evaluate relative truncated bias using coefficients whose absolute value is less than 1000 times the absolute value of the underlying parameter of the data generating process. As noted in the paper, the test functions somewhat worse in this sample than in over-identified equations, as in all leverage groups and with all error processes H_1 regressions now show bias levels that are multiples of the limit consistent with a .05 Type-I error rate.

Table A7: Fraction of Regressions with Relative Bias Greater than Bias Bound in Specifications that Don't and Do Reject the Olea & Pflueger Weak Instrument Null (sensitivity test for Table VIII in the paper)

	bias = .05		bias = .10		bias = .20		bias = ⅓	
	H_0	H_1 (max)	H_0	H_1 (max)	H_0	H_1 (max)	H_0	H_1 (max)
	174 over-identified regressions in 8 low and medium leverage papers							
iid normal	.939	.040 (.249)	.861	.045 (.226)	.815	.033 (.196)	.587	.041 (.146)
h. normal	.907	.240 (.391)	.907	.182 (.266)	.871	.183 (.204)	.650	.194 (.177)
h. cl. normal	.938	.432 (.698)	.894	.258 (.376)	.880	.264 (.242)	.767	.235 (.199)
iid chi^2	.925	.073 (.247)	.864	.031 (.226)	.786	.034 (.196)	.581	.042 (.152)
h chi^2	.903	.355 (.258)	.903	.316 (.170)	.826	.314 (.130)	.460	.207 (.113)
h cl chi^2	.912	.611 (.431)	.876	.465 (.227)	.839	.364 (.145)	.590	.219 (.120)
iid "actual"	.930	.074 (.251)	.876	.001 (.229)	.799	.001 (.199)	.648	.002 (.169)
h "actual"	.877	.093 (.334)	.879	.044 (.242)	.729	.043 (.210)	.538	.062 (.181)
h cl "actual"	.899	.219 (.381)	.886	.135 (.258)	.736	.071 (.211)	.543	.071 (.181)
	52 over-identified regressions in 4 high leverage papers							
iid normal	.000	.197 (.024)	.000	.118 (.012)	.000	.000 (.005)	.000	.000 (.003)
h. normal	.969	.206 (.050)	.878	.207 (.036)	.839	.191 (.026)	.865	.219 (.021)
h. cl. normal	.985	.906 (.908)	.978	.842 (.376)	.968	.847 (.198)	.899	.843 (.147)
iid chi^2	.002	.186 (.020)	.000	.069 (.010)	.000	.021 (.005)	.000	.020 (.003)
h chi^2	.930	.198 (.046)	.855	.196 (.031)	.733	.190 (.021)	.302	.079 (.017)
h cl chi^2	.985	.927 (.611)	.961	.847 (.269)	.922	.855 (.141)	.817	.789 (.103)
iid "actual"	.000	.083 (.023)	.000	.070 (.011)	.000	.064 (.006)	.000	.041 (.004)
h "actual"	.485	.162 (.034)	.197	.074 (.027)	.000	.026 (.017)	.000	.024 (.012)
h cl "actual"	.528	.480 (.246)	.485	.471 (.111)	.326	.365 (.049)	.305	.277 (.037)

Notes: As in Table VIII in the paper.

Table A8: Olea & Pflueger Bias Tests in the Exactly Identified Sample
(sensitivity test for Table VIII in the paper)

	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) 253 regressions in 9 low leverage papers								
iid normal	.777	.162 (.029)	.485	.098 (.017)	.322	.048 (.010)	.318	.033 (.007)
h normal	.888	.141 (.058)	.848	.077 (.050)	.758	.099 (.040)	.627	.083 (.034)
h cl normal	.881	.361 (.095)	.878	.197 (.073)	.858	.107 (.066)	.826	.075 (.062)
iid chi^2	.762	.142 (.029)	.554	.095 (.017)	.339	.042 (.009)	.335	.037 (.007)
h chi^2	.903	.118 (.058)	.845	.045 (.050)	.709	.062 (.041)	.586	.045 (.035)
h cl chi^2	.867	.301 (.088)	.899	.161 (.071)	.835	.125 (.065)	.768	.094 (.062)
iid "actual"	.718	.182 (.029)	.539	.137 (.017)	.334	.080 (.009)	.310	.058 (.007)
h "actual"	.685	.247 (.047)	.602	.167 (.022)	.420	.089 (.012)	.329	.032 (.009)
h cl "actual"	.717	.365 (.107)	.571	.198 (.054)	.548	.136 (.028)	.377	.088 (.021)
(B) 395 regressions in 8 medium leverage papers								
iid normal	.916	.126 (.066)	.776	.154 (.044)	.527	.108 (.027)	.409	.071 (.021)
h normal	.910	.577 (.149)	.834	.231 (.092)	.799	.168 (.064)	.714	.175 (.053)
h cl normal	.914	.781 (.249)	.849	.474 (.132)	.769	.195 (.083)	.717	.187 (.068)
iid chi^2	.903	.189 (.060)	.750	.177 (.040)	.501	.116 (.025)	.346	.072 (.020)
h chi^2	.885	.432 (.114)	.854	.212 (.082)	.781	.142 (.062)	.660	.138 (.053)
h cl chi^2	.925	.776 (.221)	.851	.378 (.116)	.813	.168 (.077)	.701	.150 (.064)
iid "actual"	.923	.262 (.075)	.843	.248 (.056)	.618	.182 (.038)	.399	.103 (.030)
h "actual"	.901	.382 (.084)	.814	.269 (.056)	.588	.218 (.040)	.526	.191 (.033)
h cl "actual"	.845	.501 (.123)	.798	.290 (.064)	.653	.253 (.048)	.570	.230 (.040)
(C) 435 regressions in 9 high leverage papers								
iid normal	.782	.159 (.011)	.377	.078 (.006)	.102	.031 (.003)	.103	.020 (.002)
h normal	.953	.220 (.021)	.913	.184 (.018)	.830	.134 (.015)	.766	.127 (.013)
h cl normal	.983	.991 (.424)	.959	.935 (.220)	.901	.877 (.127)	.855	.814 (.097)
iid chi^2	.733	.168 (.009)	.331	.079 (.005)	.105	.024 (.002)	.075	.012 (.001)
h chi^2	.969	.196 (.021)	.939	.154 (.018)	.889	.124 (.015)	.826	.106 (.013)
h cl chi^2	.965	.956 (.206)	.938	.923 (.115)	.894	.861 (.072)	.830	.800 (.057)
iid "actual"	.744	.124 (.015)	.618	.095 (.009)	.294	.045 (.005)	.189	.028 (.003)
h "actual"	.826	.122 (.021)	.717	.097 (.017)	.582	.080 (.013)	.459	.069 (.011)
h cl "actual"	.896	.558 (.299)	.815	.630 (.104)	.694	.561 (.057)	.543	.418 (.049)

Notes: As in Table VIII in the paper.

Table A9 Average Rejection Rates of True Nulls at the .05 Level in 1st Stage Tests
(sensitivity test for Table IX in the paper)

	default			clustered/robust: low leverage			clustered/robust: medium leverage			clustered/robust: high leverage		
	all	$k_Z > 1$ coef	$k_Z > 1$ joint	all	$k_Z > 1$ coef	$k_Z > 1$ joint	all	$k_Z > 1$ coef	$k_Z > 1$ joint	all	$k_Z > 1$ coef	$k_Z > 1$ joint
iid normal	.051	.050	.050	.056	.057	.061	.149	.071	.235	.134	.111	.355
h normal	.404	.253	.463	.062	.061	.070	.132	.053	.149	.281	.156	.481
h cl normal	.595	.355	.652	.066	.064	.068	.133	.054	.144	.308	.199	.500
iid chi^2	.052	.051	.056	.056	.054	.059	.123	.065	.192	.126	.105	.347
h chi^2	.401	.247	.459	.069	.063	.072	.156	.059	.194	.299	.161	.490
h cl chi^2	.594	.346	.653	.080	.066	.074	.160	.061	.195	.341	.199	.515
iid "actual"	.054	.051	.056	.056	.057	.059	.132	.065	.203	.124	.101	.342
h "actual"	.196	.138	.223	.057	.066	.070	.208	.084	.273	.203	.136	.359
h cl "actual"	.372	.232	.390	.061	.074	.075	.211	.083	.276	.226	.136	.397

Notes: As in Table IX in the paper.

Table A10: Average Rejection Rates of True Nulls at the .01 Level in 1st Stage Tests
(sensitivity test for Table IX in the paper)

	default			clustered/robust: low leverage			clustered/robust: medium leverage			clustered/robust: high leverage		
	all	$k_Z > 1$ coef	$k_Z > 1$ joint	all	$k_Z > 1$ coef	$k_Z > 1$ joint	all	$k_Z > 1$ coef	$k_Z > 1$ joint	all	$k_Z > 1$ coef	$k_Z > 1$ joint
iid normal	.010	.010	.010	.013	.012	.013	.075	.020	.133	.062	.045	.273
h normal	.312	.190	.390	.015	.013	.016	.057	.015	.074	.175	.087	.376
h cl normal	.512	.284	.583	.017	.013	.015	.058	.015	.070	.194	.113	.391
iid chi^2	.012	.012	.014	.014	.012	.013	.053	.017	.094	.055	.041	.264
h chi^2	.309	.186	.386	.021	.016	.020	.080	.018	.114	.195	.091	.382
h cl chi^2	.510	.274	.583	.031	.017	.021	.083	.019	.113	.230	.115	.412
iid "actual"	.014	.012	.015	.015	.013	.012	.062	.017	.106	.055	.040	.263
h "actual"	.112	.069	.132	.013	.015	.014	.119	.027	.172	.110	.057	.246
h cl "actual"	.275	.146	.284	.014	.017	.016	.119	.027	.173	.119	.051	.305

Notes: As in Table IX in the paper.

Table IX in the paper reported rejection rates for 1st stage tests using normal and "actual" errors at the .05 level. Tables A9 and A10 above add in chi^2 errors and .01 level results. The pattern of results is much the same as in the paper's discussion of Table IX. Size distortions are very large in medium and high leverage papers and grow with the dimensionality of the test, as evidenced by the comparison of the average rejection rate for tests of individual instruments against that of the joint test of all instruments in papers with overidentified 2SLS regressions.

Table X in the paper reported Monte Carlo estimates of null rejection probabilities of clustered/robust, jackknife and bootstrap methods at the .01 and .05 levels using normal and "actual" errors. Table A11 below adds in results based upon the chi^2 distribution. As elsewhere, the pattern of results are very similar to those reported in the paper: (a) jackknife and bootstrap methods reduce the size distortions of clustered/robust methods while producing a higher ratio of power to size; (b) the bootstrap-c appears to be as accurate as the -t in tests of IV coefficients and is by no means systematically worse in other tests; and (c) no matter which method is used power declines with non-iid error processes. Table A12 reports results broken down by leverage group. As noted in the paper, the improvement in size afforded by the jackknife and bootstrap are concentrated in medium and high leverage papers, while in low leverage papers the alternative methods are as accurate as clustered/robust inference.

Table A11: Sensitivity test for Table X: Improved Finite Sample Inference Using the JackKnife & Bootstrap
(average within paper rejection rates at .01 and .05 levels, 10 Monte Carlos for each of 1309 equations)

	tests of true nulls												tests of false nulls											
	clustered/ robust		jackknife		pairs bootstrap				wild bootstrap				clustered/ robust		jackknife		pairs bootstrap				wild bootstrap			
					c		t		c		t						c		t		c		t	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
IV coefficients (correlated 1st and 2nd stage errors): $H_0 = \beta_{dgp}$ or 0																								
iid normal	.028	.081	.018	.050	.009	.042	.021	.065	.009	.046	.011	.052	.455	.588	.391	.518	.312	.482	.384	.544	.257	.434	.376	.551
h. normal	.069	.126	.024	.061	.011	.048	.025	.063	.015	.051	.016	.058	.263	.364	.202	.284	.181	.270	.182	.270	.156	.245	.218	.323
h cl normal	.070	.124	.023	.048	.009	.041	.025	.059	.013	.049	.015	.055	.190	.273	.127	.186	.102	.177	.121	.184	.100	.174	.137	.228
iid chi^2	.024	.065	.014	.040	.007	.033	.017	.050	.007	.038	.010	.045	.482	.606	.403	.530	.330	.494	.396	.543	.279	.450	.403	.565
h chi^2	.080	.144	.031	.067	.016	.059	.030	.072	.018	.066	.025	.072	.288	.395	.214	.304	.191	.287	.182	.282	.168	.262	.238	.349
h cl chi^2	.075	.141	.027	.058	.012	.051	.028	.067	.017	.072	.022	.075	.189	.288	.132	.200	.103	.189	.107	.187	.099	.183	.148	.250
iid "actual"	.025	.073	.014	.044	.007	.035	.019	.060	.007	.042	.011	.050	.428	.551	.370	.485	.311	.447	.355	.495	.263	.425	.362	.520
h "actual"	.034	.081	.012	.040	.005	.035	.022	.063	.010	.049	.014	.059	.407	.535	.339	.453	.274	.416	.322	.470	.226	.380	.342	.501
h cl "actual"	.035	.083	.014	.039	.004	.032	.024	.064	.009	.045	.015	.057	.293	.444	.226	.350	.157	.294	.228	.375	.139	.303	.273	.424
1st Stage F-tests (correlated errors): $H_0 = \boldsymbol{\pi}_{dgp}$ or $\mathbf{0}$																								
iid normal	.051	.119	.023	.073	.008	.054	.017	.065	.010	.053	.012	.056	.925	.950	.894	.933	.848	.924	.855	.912	.833	.915	.858	.921
h normal	.085	.162	.034	.081	.020	.081	.015	.059	.018	.072	.017	.065	.759	.825	.693	.772	.688	.787	.547	.655	.699	.790	.668	.758
h cl normal	.091	.171	.030	.078	.023	.088	.012	.056	.023	.076	.017	.065	.647	.729	.562	.658	.571	.680	.434	.551	.576	.683	.540	.636
iid chi^2	.038	.099	.019	.054	.006	.038	.016	.053	.008	.046	.009	.050	.928	.955	.901	.940	.855	.927	.856	.913	.841	.918	.870	.929
h chi^2	.101	.183	.044	.095	.027	.088	.024	.067	.028	.075	.027	.079	.778	.848	.712	.793	.701	.812	.553	.666	.738	.818	.713	.800
h cl chi^2	.108	.184	.053	.101	.036	.103	.029	.072	.031	.086	.033	.084	.658	.737	.575	.665	.579	.691	.434	.546	.618	.707	.583	.678
iid "actual"	.040	.105	.018	.054	.009	.041	.015	.051	.008	.042	.009	.044	.880	.924	.837	.897	.795	.879	.806	.873	.791	.882	.816	.889
h "actual"	.081	.160	.029	.075	.011	.056	.017	.064	.017	.067	.016	.066	.857	.910	.778	.855	.738	.853	.718	.820	.751	.854	.754	.856
h cl "actual"	.084	.162	.032	.078	.015	.066	.018	.062	.020	.067	.015	.063	.766	.846	.666	.766	.621	.777	.588	.724	.617	.767	.613	.755

Table A11: Sensitivity test for Table X (continued)

	tests of true nulls												tests of false nulls											
	clustered/ robust		jackknife		pairs bootstrap				wild bootstrap				clustered/ robust		jackknife		pairs bootstrap				wild bootstrap			
					c		t		c		t						c		t		c		t	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
	Hausman tests: $H_0 = (\beta_{iv} = \beta_{ols})$																							
	(uncorrelated errors)												(correlated errors)											
iid normal	.021	.071	.008	.036	.005	.030	.008	.041	.006	.044	.010	.050	.373	.493	.255	.373	.216	.354	.262	.405	.266	.408	.306	.450
h normal	.065	.145	.011	.047	.006	.036	.007	.035	.013	.051	.014	.068	.268	.378	.153	.211	.147	.216	.146	.202	.158	.242	.191	.279
h cl normal	.076	.157	.008	.029	.003	.025	.005	.025	.011	.050	.020	.070	.210	.316	.098	.147	.089	.139	.085	.138	.103	.178	.135	.219
iid chi^2	.017	.060	.007	.033	.004	.031	.007	.035	.007	.041	.010	.051	.344	.470	.239	.347	.210	.338	.236	.363	.247	.397	.276	.421
h chi^2	.083	.158	.018	.054	.010	.047	.011	.046	.016	.068	.027	.087	.285	.383	.150	.210	.142	.215	.140	.202	.165	.250	.198	.293
h cl chi^2	.101	.182	.014	.045	.006	.039	.008	.040	.017	.065	.033	.101	.219	.323	.089	.129	.079	.132	.078	.119	.099	.177	.140	.233
iid "actual"	.023	.073	.006	.030	.003	.024	.007	.040	.024	.050	.013	.052	.375	.484	.266	.364	.221	.346	.258	.377	.211	.318	.314	.441
h "actual"	.039	.100	.007	.032	.003	.028	.007	.041	.021	.051	.021	.068	.335	.454	.211	.315	.186	.297	.205	.323	.171	.272	.260	.387
h cl "actual"	.049	.111	.008	.033	.003	.024	.007	.037	.021	.052	.022	.075	.278	.405	.139	.240	.097	.204	.134	.260	.127	.232	.218	.350

Notes: As in Tsble X in the paper.

Table A12: : Rejection Probabilities of True Nulls by Test and Leverage Group
(sensitivity test for Table X in the paper)

	low leverage						medium leverage						high leverage					
	clust-robust	jack-knife	pairs bootstrap c	pairs bootstrap t	wild bootstrap c	wild bootstrap t	clust-robust	jack-knife	pairs bootstrap c	pairs bootstrap t	wild bootstrap c	wild bootstrap t	clust-robust	jack-knife	pairs bootstrap c	pairs bootstrap t	wild bootstrap c	wild bootstrap t
							IV coefficients (correlated errors): .01 level											
iid normal	.013	.011	.008	.013	.011	.016	.036	.026	.014	.030	.007	.009	.035	.017	.006	.021	.008	.008
h normal	.013	.008	.007	.007	.010	.010	.054	.028	.011	.028	.010	.010	.141	.038	.016	.039	.024	.027
h cl normal	.016	.013	.009	.013	.010	.013	.053	.026	.010	.027	.009	.008	.142	.030	.009	.036	.019	.025
iid chi^2	.009	.007	.004	.008	.006	.007	.031	.016	.011	.021	.009	.011	.033	.019	.006	.022	.005	.012
h chi^2	.014	.016	.009	.013	.015	.014	.069	.039	.022	.038	.019	.026	.157	.038	.018	.039	.020	.035
h cl chi^2	.012	.013	.010	.009	.012	.012	.063	.034	.016	.036	.017	.023	.151	.036	.009	.039	.021	.031
iid "actual"	.012	.010	.007	.012	.011	.015	.021	.012	.008	.016	.006	.010	.044	.019	.005	.029	.005	.008
h "actual"	.010	.007	.006	.007	.006	.009	.043	.013	.006	.032	.019	.029	.050	.015	.003	.028	.004	.004
h cl "actual"	.010	.007	.004	.009	.005	.010	.041	.014	.004	.033	.018	.027	.055	.021	.005	.031	.005	.009
							IV coefficients (correlated errors): .05 level											
iid normal	.058	.053	.048	.058	.049	.060	.084	.051	.051	.064	.049	.052	.102	.045	.025	.074	.039	.044
h normal	.053	.045	.037	.043	.043	.056	.101	.058	.044	.059	.045	.048	.224	.079	.064	.086	.066	.070
h cl normal	.048	.038	.036	.042	.043	.054	.096	.051	.041	.055	.046	.046	.228	.056	.044	.079	.058	.064
iid chi^2	.035	.031	.029	.037	.032	.041	.071	.044	.042	.051	.042	.050	.090	.044	.030	.061	.040	.043
h chi^2	.057	.047	.045	.044	.053	.057	.131	.074	.063	.079	.068	.073	.244	.082	.068	.093	.076	.086
h cl chi^2	.049	.040	.039	.037	.052	.059	.135	.063	.060	.077	.087	.077	.239	.073	.056	.088	.077	.087
iid "actual"	.050	.045	.043	.050	.045	.052	.062	.035	.034	.051	.051	.045	.108	.051	.026	.079	.031	.053
h "actual"	.039	.034	.035	.040	.035	.042	.095	.042	.045	.070	.072	.078	.109	.044	.024	.078	.041	.056
h cl "actual"	.039	.032	.028	.044	.032	.042	.092	.038	.039	.071	.069	.073	.117	.047	.028	.077	.035	.055

Notes: As in Table X in the paper.

Table A12: : Rejection Probabilities of True Nulls by Test and Leverage Group (continued)

	low leverage						medium leverage						high leverage					
	clust-robust	jack-knife	pairs bootstrap		wild bootstrap		clust-robust	jack-knife	pairs bootstrap		wild bootstrap		clust-robust	jack-knife	pairs bootstrap		wild bootstrap	
			c	t	c	t			c	t	c	t			c	t	c	t
							1st Stage F-tests (correlated errors): .01 level											
iid normal	.010	.009	.011	.009	.012	.011	.081	.032	.007	.011	.009	.012	.062	.028	.008	.030	.011	.011
h normal	.016	.011	.016	.006	.011	.010	.054	.023	.018	.007	.014	.012	.187	.066	.025	.033	.029	.030
h cl normal	.018	.017	.020	.009	.012	.011	.050	.021	.016	.006	.017	.011	.206	.053	.032	.022	.039	.029
iid chi^2	.014	.011	.011	.010	.011	.011	.044	.016	.004	.006	.008	.010	.055	.029	.003	.031	.004	.004
h chi^2	.029	.023	.023	.014	.017	.021	.074	.041	.020	.014	.020	.021	.200	.069	.036	.044	.047	.040
h cl chi^2	.026	.019	.018	.013	.019	.017	.088	.057	.034	.022	.023	.039	.209	.084	.054	.051	.051	.042
iid "actual"	.013	.009	.010	.010	.011	.010	.049	.017	.007	.008	.007	.009	.059	.029	.011	.028	.007	.008
h "actual"	.008	.006	.009	.005	.006	.006	.130	.046	.014	.020	.030	.027	.105	.035	.009	.027	.015	.014
h cl "actual"	.010	.005	.012	.004	.006	.004	.130	.051	.017	.018	.031	.032	.113	.038	.015	.032	.022	.008
							1st Stage F-tests (correlated errors): .05 level											
iid normal	.073	.065	.068	.069	.063	.069	.147	.085	.055	.042	.046	.052	.137	.070	.039	.085	.050	.047
h normal	.067	.055	.066	.054	.055	.057	.127	.067	.062	.031	.061	.054	.293	.121	.115	.092	.101	.085
h cl normal	.075	.061	.076	.053	.061	.059	.129	.054	.072	.034	.058	.053	.309	.119	.116	.082	.108	.083
iid chi^2	.052	.052	.047	.052	.051	.053	.118	.045	.037	.030	.044	.055	.127	.066	.029	.077	.043	.043
h chi^2	.065	.060	.065	.048	.059	.055	.168	.097	.083	.053	.061	.079	.315	.129	.116	.100	.103	.103
h cl chi^2	.067	.053	.065	.045	.055	.054	.175	.109	.106	.062	.080	.095	.310	.142	.138	.110	.122	.105
iid "actual"	.050	.043	.046	.044	.048	.046	.130	.050	.039	.033	.039	.046	.136	.070	.038	.076	.040	.041
h "actual"	.053	.048	.051	.042	.049	.052	.219	.097	.066	.064	.089	.084	.208	.080	.051	.086	.063	.061
h cl "actual"	.057	.047	.056	.038	.047	.047	.221	.098	.069	.062	.084	.088	.207	.090	.072	.088	.070	.053

Notes: As in Table X in the paper.

Table A12: : Rejection Probabilities of True Nulls by Test and Leverage Group (continued)

	low leverage						medium leverage						high leverage					
	clust-robust	jack-knife	pairs bootstrap		wild bootstrap		clust-robust	jack-knife	pairs bootstrap		wild bootstrap		clust-robust	jack-knife	pairs bootstrap		wild bootstrap	
			c	t	c	t			c	t	c	t			c	t	c	t
	Hausman tests (uncorrelated errors): .01 level																	
iid normal	.016	.010	.007	.009	.010	.012	.016	.005	.005	.006	.004	.009	.030	.009	.003	.011	.005	.008
h normal	.011	.005	.006	.004	.009	.011	.043	.004	.001	.002	.011	.012	.143	.025	.010	.016	.019	.020
h cl normal	.020	.007	.007	.006	.007	.017	.052	.004	.001	.002	.011	.018	.155	.014	.002	.007	.015	.024
iid chi^2	.014	.007	.006	.006	.005	.012	.012	.002	.001	.002	.006	.013	.026	.012	.005	.014	.010	.007
h chi^2	.021	.008	.006	.003	.010	.015	.080	.013	.008	.010	.017	.030	.149	.032	.016	.020	.022	.036
h cl chi^2	.037	.010	.007	.007	.012	.024	.071	.007	.004	.004	.017	.029	.193	.024	.008	.014	.021	.048
iid "actual"	.008	.005	.004	.005	.033	.007	.028	.003	.002	.002	.034	.024	.031	.012	.004	.013	.004	.008
h "actual"	.020	.007	.006	.005	.027	.011	.049	.003	.002	.003	.034	.041	.047	.011	.002	.013	.002	.009
h cl "actual"	.018	.002	.001	.001	.027	.011	.065	.005	.007	.008	.033	.047	.064	.016	.002	.010	.003	.008
	Hausman tests (uncorrelated errors): .05 level																	
iid normal	.068	.050	.045	.050	.053	.062	.053	.016	.018	.023	.037	.044	.091	.041	.026	.049	.041	.043
h normal	.080	.044	.039	.037	.046	.073	.109	.023	.018	.016	.041	.051	.247	.074	.052	.051	.066	.080
h cl normal	.083	.033	.030	.031	.042	.070	.118	.019	.019	.013	.041	.057	.269	.036	.026	.030	.068	.082
iid chi^2	.049	.032	.031	.037	.039	.051	.057	.026	.026	.023	.041	.052	.075	.040	.035	.046	.043	.049
h chi^2	.072	.045	.044	.033	.050	.067	.149	.043	.037	.034	.072	.096	.253	.073	.061	.070	.081	.097
h cl chi^2	.099	.049	.046	.038	.060	.090	.138	.031	.021	.022	.063	.084	.310	.055	.048	.060	.072	.128
iid "actual"	.042	.032	.029	.033	.058	.041	.074	.014	.018	.029	.079	.067	.102	.043	.024	.059	.014	.050
h "actual"	.065	.041	.043	.050	.067	.059	.119	.018	.020	.024	.069	.091	.116	.036	.022	.050	.017	.054
h cl "actual"	.069	.034	.028	.035	.058	.061	.132	.028	.026	.032	.073	.106	.132	.037	.017	.044	.025	.058

Notes: As in Table X in the paper.

Because of the high computational cost of calculating jackknife and bootstrap p-values, Table X in the paper (and Tables A11 and A12 above) estimated null rejection probabilities using only 10 simulations for each of 1309 equations in 30 papers. To address the question of whether this leads to inaccurate estimates, Tables A13 reports clustered/robust results using 10 and 1000 simulations per equation. As can be seen, 10 and 1000 results are very similar. Table X aims to measure average rejection rates across 30 papers, not the average rejection rate in any given equation, and in this regard, as noted in the paper, 10 simulations per equation appear to yield reasonably accurate estimates of the average and relative performance of the different methods.

Table A13: Clustered/Robust Rejection Rates at the .01 and .05 Levels
(sensitivity test for Table X in the paper, 10 vs 1000 Monte Carlos per equation)

	IV coefficients				1st Stage F-tests				Hausman tests			
	$H_0 = \beta_{dgp}$		$H_0 = 0$		$H_0 = \boldsymbol{\pi}_{dgp}$		$H_0 = \mathbf{0}$		uncorrelated errors		correlated errors	
	10	1000	10	1000	10	1000	10	1000	10	1000	10	1000
	.01 level											
iid normal	.028	.029	.455	.461	.051	.050	.925	.924	.021	.020	.373	.374
h. normal	.069	.069	.263	.276	.085	.082	.759	.759	.066	.065	.268	.270
h cl normal	.070	.069	.190	.182	.091	.090	.647	.642	.076	.076	.210	.205
iid chi^2	.024	.026	.482	.477	.038	.041	.928	.927	.017	.018	.344	.349
h. chi^2	.080	.077	.288	.287	.101	.099	.778	.775	.083	.084	.285	.289
h cl chi^2	.075	.083	.189	.196	.108	.115	.658	.662	.101	.103	.219	.224
iid "actual"	.025	.025	.428	.432	.040	.044	.880	.887	.023	.019	.375	.374
h "actual"	.034	.035	.407	.409	.081	.080	.857	.860	.039	.038	.335	.334
h cl "actual"	.035	.037	.293	.297	.084	.084	.766	.761	.049	.046	.278	.279
	.05 level											
iid normal	.081	.077	.588	.590	.119	.113	.950	.953	.071	.069	.493	.490
h. normal	.126	.126	.364	.375	.162	.158	.825	.825	.145	.139	.378	.374
h cl normal	.124	.123	.273	.272	.171	.169	.729	.720	.157	.155	.316	.312
iid chi^2	.065	.075	.606	.603	.099	.102	.955	.953	.060	.067	.470	.472
h. chi^2	.144	.139	.395	.392	.183	.174	.848	.847	.158	.158	.383	.391
h cl chi^2	.141	.144	.288	.294	.184	.194	.737	.742	.182	.182	.323	.329
iid "actual"	.073	.072	.551	.552	.105	.104	.924	.927	.073	.068	.484	.487
h "actual"	.081	.085	.535	.539	.160	.156	.910	.910	.100	.098	.454	.455
h cl "actual"	.083	.085	.444	.446	.162	.166	.846	.845	.111	.110	.405	.407

Notes: 10 and 1000 = number of Monte Carlos per equation used in calculation of average rejection rates. Otherwise, as in Table X in the paper.

Table A14: Significance of 2SLS Coefficients (sensitivity test for Table XI)
(average across papers of the fraction of coefficients rejecting the null of 0)

	headline results		all results							
			all		low		medium		high	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
authors' methods	.522	.788	.365	.558	.543	.719	.215	.400	.336	.555
clustered/robust	.463	.768	.339	.531	.524	.716	.173	.347	.322	.531
jackknife	.382	.537	.250	.401	.467	.674	.095	.235	.187	.293
pairs bootstrap - c	.243	.520	.160	.340	.346	.600	.074	.168	.062	.252
pairs bootstrap - t	.308	.599	.247	.453	.444	.692	.088	.289	.210	.378
wild bootstrap - c	.231	.444	.115	.337	.219	.603	.092	.246	.035	.163
wild bootstrap - t	.512	.719	.346	.535	.600	.768	.231	.414	.208	.425

Notes: As in Table XI.

Table A15: Frequency with which IV Confidence Intervals contain OLS Point Estimates
(sensitivity test for Table XIII)

	headline results		all results							
			all		low		medium		high	
	.99	.95	.99	.95	.99	.95	.99	.95	.99	.95
clustered/robust	.831	.673	.870	.750	.820	.706	.951	.830	.840	.713
jackknife	.862	.801	.902	.825	.801	.727	.973	.915	.930	.833
pairs bootstrap - c	.895	.790	.934	.852	.849	.753	.972	.925	.981	.877
pairs bootstrap - t	.895	.769	.902	.779	.825	.697	.984	.890	.897	.750
wild bootstrap - c	.847	.759	.916	.801	.836	.733	.920	.771	.990	.899
wild bootstrap - t	.858	.664	.887	.719	.768	.622	.940	.787	.952	.748

Notes: As in Table XIII.

In Section VI's analysis of the sample aggregate information is given for all and headline results, but (for reasons of space) detail by leverage group is only given for headline results. Tables A14-A18 reverse this, providing detail for all results by leverage groups. The patterns by leverage group are the same as those found for headline results reported in the paper with, in particular, the greatest differences between cl/robust and jackknife and bootstrap significance rates appearing in medium and high leverage papers.

Table A16: Rejection Rates in Hausman Tests (tests of OLS bias)
(sensitivity test for Table XIV)

	headline results		all results							
			all		low		medium		high	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
clustered/robust	.309	.445	.232	.382	.290	.441	.228	.358	.177	.348
jackknife	.188	.254	.135	.227	.252	.344	.071	.162	.083	.174
pairs bootstrap - c	.138	.249	.098	.200	.190	.310	.066	.154	.037	.136
pairs bootstrap - t	.110	.300	.110	.243	.233	.349	.065	.176	.031	.205
wild bootstrap - c	.187	.319	.129	.247	.203	.313	.147	.278	.036	.149
wild bootstrap - t	.237	.470	.175	.328	.283	.421	.209	.341	.034	.221

Notes: As in Table XIV.

Table A17: Identification in the First-Stage (sensitivity test for Table XV)
(rejection rates in tests of instrument irrelevance)

	headline results		all results							
			all		low		medium		high	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
clustered/robust	1.00	1.00	.858	.929	.913	.966	.802	.853	.858	.969
jackknife	.835	.945	.718	.827	.903	.948	.630	.728	.621	.805
pairs bootstrap - c	.781	.967	.661	.874	.869	.977	.526	.768	.588	.878
pairs bootstrap - t	.755	.877	.638	.773	.859	.923	.571	.704	.484	.693
wild bootstrap - c	.794	.967	.704	.886	.892	.961	.585	.727	.636	.971
wild bootstrap - t	.783	.952	.660	.856	.879	.942	.604	.749	.497	.876

Notes: As in Table XV.

Table A18: Does 2SLS Provide Information that is Strongly Statistically Different from OLS?
(average fraction of 2SLS regressions rejecting $\boldsymbol{\pi} = 0$ & $\beta_{ols} \in CI_{2sls}$ or β_{ols} unbiased)
(sensitivity test for Table XVI)

	(i) at .01 level					(ii) at .05 level				
	headline	all results				headline	all results			
	results	all	low	med	high	results	all	low	med	high
cl/robust	.309	.234	.285	.209	.210	.445	.378	.439	.329	.366
jackknife	.188	.130	.239	.071	.081	.271	.228	.341	.159	.184
pairs boot - c	.138	.097	.190	.064	.037	.221	.183	.310	.152	.086
pairs boot - t	.138	.127	.220	.066	.093	.355	.277	.353	.203	.273
wild boot - c	.187	.116	.205	.133	.009	.319	.249	.315	.274	.158
wild boot - t	.287	.177	.276	.199	.058	.502	.353	.433	.322	.305

Notes: As in Table XVI.

B: Selection of Headline Results

As noted in the paper, at the request of reviewers I separate out headline results in the discussion and analysis. The text of my paper gives the criteria used to define a headline result. Table B below reports the location of headline results in each paper, along with notes indicating how they were identified. Papers are identified by the initials of the last names of the authors and the year of publication (see appendix L below for the full citations), followed by an equals sign and the number of headline results. The location of headline results in tables is then identified by a number indicating the table followed by a parentheses where the row (if needed) and column of headline results are listed, separated by "/" marks. To illustrate: 2(3/4) means table 2, IV coefficients in columns 3 and 4; 5(B14) means table 5, IV coefficients in panel B row 1 column 4.

Table B: Selection of Headline Results	
paper table(rowcol)	notes
A2011 = 8 2(3/4)	Abstract/intro mention black and white poverty, black-white income disparities, black incomes, within white inequality. Repeated in first sentence of conclusion. Table 2 covers all of these under column title main results.
A2012 = 1 3(A1)	Critique of another paper: use first replication regression, which has strongest 1st stage.
ACS2014 = 1 4(6)	Cols 4 & 6 very close and given equal weight in text and match # reported in abstract/intro. Col 6 used to construct estimates included lagged effects reported in text and abstract/intro.
ADH2013 = 1 3(6)	Table 3, col. 6 noted as preferred specification. Abstract/intro/conclusion discuss other significant effects, but these come much later in presentation within paper and intermingled with insignificant results.
AJRY2008 = 2 5(4), 6(4)	Both instruments given equal weight in introduction. Col 4 is baseline specification, given more discussion in terms of 1st stage and coef.
AZ2011 = 3 6(2/4/6)	Columns with full controls given emphasis in text and match reported results in introduction. Table 7 covers other measures of quality of governance, but rule of law singled out in this and other sections.
BC2010 = 7 3(A22/B32/C22/D12), 4(14), 5(B14), 6(F35)	Abstract: divorce, intermarriage, fewer children, for some living outside ethnic enclave; introduction: lower marriage, ever married, higher divorce, spouse fluent, more educated and earns more, marriage outside ethnicity and nationality, fewer children, living outside ethnic enclaves; conclusion: divorced, marrying US native, more educated and higher earning spouse, fewer children, for some living outside ethnic enclaves. Common to at least two of the above: divorce, intermarriage, spouse more educated and higher earning, fewer children, for some outside ethnic enclave. Intermarriage - spouse has same country of birth seems to summarize best the four measures; fertility - text indicates women's results more easy to interpret as fertility; living outside enclave - second measure deemed more accurate at top p. 183.
BC2013 = 3 1(1), 3(1/3)	Critique of other papers: use regressions that replicate original results for equations with a single instrumented variable.
BD2006 = 1 5(13)	Result mentioned in intro, other results in table are specification checks and with caveats.
BHW2011 = 1 1(7)	Non-textile results highlighted in abstract/introduction/conclusion. This instrument highlighted as primary specification in introduction (p. 94).
BL2010 = 1 4(A2)	All instruments, only point estimate for that table summarized in text (p. 139), panels A & B (with more controls) very similar, insignificant results on movement to autocracy qualified in conclusion and given less emphasis in introduction/abstract/conclusion.
BL2012 = 1 3(2)	Only coefficent estimate for that table discussed in text, remainder described as specification checks. [Alternative: Cols 5 & 8, but 8 involves multiple instrumented coefficients - my paper only examines single instrumented as multiple is rare, see text of my paper - and both have low 1st stage F - since yield same coef with higher s.e., seen by authors as specification checks].
C2015 = 4 2(B2/4/6/7)	No particular outcome mentioned in introduction, no abstract. Multiple outcomes, text discusses rape, larceny, motor vehicle theft & aggravated assault.
CFLW2012 = 1 3(6)	Highlighted as preferred specification in text.
CLGJ2010 = 1 5(C4)	Women's results considered more reliable than men's (text), overall infant mortality result noted in abstract.
CS2013 = 3 3(A1/2) & 5(A3)	Property values, income, population, employment, poverty rates effects mentioned in abstract. First 3 repeated in introduction and conclusion. Population effects (table 5) repeated in conclusion. All other results in these tables compared in text to those in panel A.

Table L: Selection of Headline Results (continued)	
paper table(rowcol)	notes
D2011 = 2 4(7/8)	Employment and hours/wages highlighted in abstract/introduction. Hours/wages only OLS, employment IV also. Remaining results described as mechanisms. Point estimate on female employment/participation repeated in introduction/text/conclusion. These two columns highlighted as preferred specification in text. Male results on participation in table 5 qualified in text.
D2015 = 1 8(1)	Coefficients rise (OLS & IV) with additional covariates and discussion in text focuses on lowest OLS value (col 1) and possibility that it overstates. This column given precedence in discussion and indicated to be preferred specification in introduction.
DMW2011 = 1 4I(C4)	Highlighted as preferred estimate in conclusion, in range reported in introduction. [Alternative: Table 2, reg # 2, col 4 reported as preferred specification in text, but not highlighted in conclusion].
GK2010 = 3 3(last 3 rows of 2)	6, 12 & 18 month horizons highlighted in introduction and conclusion.
H2014 = 1 5(6)	Productivity result highlighted in introduction/conclusion. Tables 4 & 5 are main results, text indicates dissatisfaction with 1st stage until get to last column of table 5. Remaining tables described as testing robustness of results.
HG2010 = 2 7(H/J)	Described in text as preferred IV specifications. Then repeated in first column of table 8 which is then used to summarize results in conclusion. Post-college results in table 8 not as significant.
J2015 = 2 3(3) & 7(3)	IV results highlighted in abstract and introduction. Asymmetries explored in table 9 and discussed in introduction and conclusion, but is last table in paper and hence seems less central. Table 4 is a sub-category of 3, tables 5 & 6 IV insignificant and not featured in abstract/introduction. Cols. 4 of tables (different specification) described as addressing some concerns, but in some cases results opposite to headline results or insignificant.
K2014 = 1 4(2)	Closest match to number reported in abstract. Cols 2 and 3 noted in text to have higher 1st stage F due to fact more important in these sub-samples. Col 1 insignificant. [Alternative: column 1, because full sample].
LMB2013 = 2 6(6), 7(6)	Identified as preferred specification in text. Outcomes highlighted in intro,
MVW2014 = 1 4(2)	Point estimate quoted in introduction, singled out in text. [Alternative: per patent estimate, col 4, but not quoted in intro].
O2006 = 3 4(12/42/82)	Returns to schooling mentioned in both introduction and conclusion. Introduction also mentions other outcomes, but not in conclusion and not reviewed in text until last. Table 4 is table that delivers summary result (mentioned in introduction and conclusion) of 10-14%, compares 3 countries in text.
SW2011 = 1 1(6)	Agrees with point estimates summarized in introduction. In text, col. 4 quickly dismissed in favour of col. 5, col. 5 then described as "naive".
T2008 = 1 8(A2)	Abstract/introduction emphasis on HIV positive purchasing condoms and number purchased. Because only use eqns with one instrumented coef in this study [see text of my paper], exclude results Table 7. Also, Table 8 separates estimates out by HIV status, which is what is emphasized in introduction. Table 8 - HIV positive, purchasing condoms, is closest to emphasis in abstract/introduction.
Y2014 = 1 2(D2)	Considers defense spending as best instrument & use of capacity utilization controls important, value of -.750 used in later discussion. (Specification with -.750 at bottom of table is a summary effect, including effects of lags which are not instrumented).

Table C1: Increase in ln Relative 2SLS to OLS Relative Bias and Maximum Leverage
(each cell, enclosed in a box, represents a separate regression)

		increase in relative bias from iid errors to heteroskedastic errors			increase in relative bias from iid errors to clustered & heteroskedastic errors		
		$\|\hat{\beta}\|<1000*\|\beta_{dgp}\|$	$\|\hat{\beta}\|<100*\|\beta_{dgp}\|$	$\|\hat{\beta}\|<10*\|\beta_{dgp}\|$	$\|\hat{\beta}\|<1000*\|\beta_{dgp}\|$	$\|\hat{\beta}\|<100*\|\beta_{dgp}\|$	$\|\hat{\beta}\|<10*\|\beta_{dgp}\|$
normal errors	β	4.0	3.8	3.8	5.6	4.8	4.3
	s.e.	(1.5)	(1.3)	(1.2)	(1.3)	(1.1)	(1.1)
	p-v	.012	.019	.012	.001	.002	.002
chi^2 errors	β	3.6	3.7	3.7	5.5	5.5	5.2
	s.e.	(1.3)	(1.2)	(1.0)	(1.4)	(1.1)	(1.1)
	p-v	.013	.010	.006	.007	.002	.001
"actual" errors	β	2.2	2.2	2.3	2.9	2.7	2.9
	s.e.	(0.8)	(0.7)	(0.7)	(1.4)	(1.3)	(1.3)
	p-v	.011	.009	.014	.037	.029	.029

Notes: Each cell represents a separate regression of the increase in ln 2SLS to OLS relative bias on maximum leverage and a constant term using paper averages (30observations). β & s.e. = coefficient and heteroskedasticity robust standard error for maximum leverage, p-v = resampling bootstrap-t p-value calculated using 1000 bootstrap draws.

C: Maximum Leverage and Increases in Relative Bias

As noted in the paper's discussion of Table V, although the increase in relative 2SLS to OLS bias with non-iid error processes by broad leverage group (low, medium, high) is not monotonic, the two variables are positively and significantly related at the paper level. To show this, Table C1 regresses the increase in ln relative 2SLS to OLS bias found in moving from iid to heteroskedastic or clustered & heteroskedastic errors on maximum leverage, separately examining results using normal, chi^2 and "actual" error processes and different levels of truncation in calculating relative bias. Regressions are done at the paper level using paper averages. Reported standard errors are heteroskedasticity robust and p-values are calculated using the bootstrap-t. All of the relations are positive and significant at the .05 level or less.

D: Using Wild Bootstrap Data Generating Methods to Approximate the Characteristics of a Data Generating Process

In the paper and elsewhere in this on-line appendix I use transformations of jackknifed residuals to approximate the distribution of results produced by the data generating process underlying the actual data of my sample. Such simulations are identified by the moniker "actual" in the relevant tables. In this appendix I present two approaches to approximating the results produced by an underlying data generating process using wild bootstrap transformations of estimated residuals and apply them to the artificial data generating processes 9.1 - 9.6 described in the paper, whose true characteristics can be determined by simulation. In the first I use standard estimated residuals and in the second jackknifed delete-**i** residuals. I find that wild transformations of estimated residuals do a poor job of replicating the pattern of results produced by an underlying data generating process, perhaps because estimated residuals are shrunken toward zero in high leverage observations. In contrast, wild transformations based on jackknifed residuals approximate some of the results produced by an underlying data generating process.

The two methods examined in simulations below are:

(1) **Wild bootstrap**. Given a set of *base* data, estimate the 2SLS equation system:

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

and then produce artificial data

(1.2) $\mathbf{y}^{w} = \mathbf{Y}^{w}\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \mathbf{u}, \quad \mathbf{Y}^{w} = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \mathbf{v}$,

where $(\mathbf{u},\mathbf{v})$ are transformations of the estimated residual pairs $(\hat{\mathbf{u}}, c_1\hat{\mathbf{v}})$, where $c_1=(n/(n-k_z-k_x))^{1/2}$ is an adjustment for the reduction in variance brought about by OLS fitting. The transformations vary by the assumption regarding the underlying data generating process:

(1.3a) iid - the residual pairs are multiplied by a 50/50 iid draw from ±1 at the observation level and randomly shuffled across observations;

(1.3b) heteroskedastic - the residual pairs are multiplied by a 50/50 iid draw from ±1 at the observation level, but not shuffled;

(1.3c) heteroskedastic & clustered - the residual pairs are multiplied by a 50/50 iid draw from ±1 at the cluster level and not shuffled.

(2) **Wild bootstrap with the jackknife**. Given a set of *base* data, estimate the 2SLS coefficients

$$(2.1)\quad \mathbf{y} = \mathbf{Y}\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \hat{\mathbf{u}}, \quad \mathbf{Y} = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \hat{\mathbf{v}},$$

estimate delete-**i** residuals based upon delete-**i** coefficient estimates

$$(2.2)\quad \breve{\mathbf{u}}_{\mathbf{i}} = \mathbf{y}_{\mathbf{i}} - \mathbf{Y}_{\mathbf{i}}\hat{\beta}_{iv\sim\mathbf{i}} + \mathbf{X}_{\mathbf{i}}\hat{\boldsymbol{\delta}}_{\sim\mathbf{i}} \text{ and } \breve{\mathbf{v}}_{\mathbf{i}} = \mathbf{Y}_{\mathbf{i}} - \mathbf{Z}_{\mathbf{i}}\hat{\boldsymbol{\pi}}_{\sim\mathbf{i}} + \mathbf{X}_{\mathbf{i}}\hat{\boldsymbol{\gamma}}_{\sim\mathbf{i}},$$

where ~**i** indicates coefficient estimates excluding cluster **i** (or an individual observation when the regression is not clustered) and **i** the variables for cluster **i**, and then produce artificial data

$$(2.3)\quad \mathbf{y}^{w_{jk}} = \mathbf{Y}^{w_{jk}}\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \mathbf{u}, \quad \mathbf{Y}^{w_{jk}} = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \mathbf{v},$$

where $(\mathbf{u}, \mathbf{v})$ are transformations of the estimated delete-**i** residuals pairs $(\breve{\mathbf{u}}, \breve{\mathbf{v}})$ using the processes described in (1.3a) - (1.3c). Where the regressions include cluster fixed effects, the delete-**i** residuals are estimated using cluster demeaned variables, so the delete-**i** residuals have a zero cluster mean. Delete-**i** residuals are estimated at the cluster level when the regression is clustered, regardless of whether the subsequent transformations (1.3a) - (1.3c) are clustered or not, so as to use a consistent set of residuals across the different transformations.

The above methods each describe a data generating process which tries to replicate the data generating process underlying the *base* data. To avoid confusion, I shall refer to the data generating process of the *base* data as *dgp*, and the two data generating processes described above as *wild* and *wild jk*. I refer to the underlying IV parameter value of each data generating process as β. For *dgp*, based as it is upon simulations 9.1-9.6 described in the paper, this is the $\hat{\beta}_{iv}$ of the papers' data. In contrast, the β of *wild* and *wild jk* in the simulations below will be the $\hat{\beta}_{iv}$ of each *base* data draw from *dgp*.

Table D1 reports rejection rates in clustered/robust tests of the instrumented coefficient for the true data generating processes (*dgp*) 9.1-9.6 that produce the *base* data, and for the wild bootstrap data generating processes *wild* and *wild jk*. "Type I error rate" is the probability that the null that the parameter value equals the β of each process is rejected, while "power" is the rejection probability of the incorrect null of zero effects. I use 1000 draws from *dgp* to calculate

Table D1: Type I Error Rates and Power using Wild Bootstrap Data Generating Methods vs Actual Characteristics for the Artificial Data Generating Processes Described in the Paper (average across papers of within paper averages)

	low leverage papers			medium leverage papers			high leverage papers		
	dgp	*wild*	*wild jk*	*dgp*	*wild*	*wild jk*	*dgp*	*wild*	*wild jk*
	Type I error rate: IV rejection rate of true nulls (.01 level)								
iid normal	.011	.012	.012	.036	.036	.024	.039	.040	.041
h normal	.012	.031	.016	.052	.056	.035	.142	.135	.132
h cl normal	.010	.047	.016	.054	.061	.033	.143	.105	.132
iid chi^2	.012	.014	.013	.032	.031	.021	.035	.037	.038
h chi^2	.015	.031	.015	.066	.055	.036	.152	.118	.130
h cl chi^2	.018	.047	.017	.069	.056	.034	.160	.099	.130
	Type I error rate: IV rejection rate of true nulls (.05 level)								
iid normal	.049	.050	.049	.082	.082	.069	.101	.102	.103
h normal	.045	.070	.054	.106	.107	.087	.226	.216	.215
h cl normal	.040	.085	.051	.106	.111	.084	.224	.178	.208
iid chi^2	.051	.052	.052	.076	.076	.066	.097	.099	.099
h chi^2	.049	.070	.053	.126	.105	.090	.242	.199	.216
h cl chi^2	.051	.086	.053	.130	.105	.086	.250	.168	.209
	power: IV rejection rate of the incorrect null of zero effects (.01 level)								
iid normal	.578	.590	.582	.285	.356	.286	.519	.546	.512
h normal	.372	.396	.395	.132	.222	.162	.324	.379	.363
h cl normal	.256	.294	.286	.107	.197	.145	.183	.286	.234
iid chi^2	.579	.596	.587	.314	.377	.304	.539	.558	.525
h chi^2	.368	.403	.400	.151	.243	.186	.342	.413	.376
h cl chi^2	.259	.281	.274	.122	.206	.148	.206	.298	.240
	power: IV rejection rate of the incorrect null of zero effects (.05 level)								
iid normal	.694	.693	.686	.440	.494	.416	.636	.659	.626
h normal	.457	.481	.478	.249	.347	.282	.418	.481	.459
h cl normal	.333	.381	.369	.212	.320	.258	.271	.387	.325
iid chi^2	.698	.696	.689	.461	.512	.436	.650	.666	.633
h chi^2	.457	.489	.485	.275	.372	.307	.445	.519	.477
h cl chi^2	.341	.372	.362	.236	.328	.263	.304	.405	.335

Notes: (1) *dgp* = cl/robust rejection rates for the data generating processes listed in the left-most column, as determined by 1000 simulations per equation; (2) *wild* and *wild jk* = cl/robust rejection rates as determined by simulated distributions using 1000 transformations of residuals for 10 draws from *dgp*, with transformations 1.3a in the text used for iid *dgp*, 1.3b for heteroskedastic *dgp*, and 1.3c for heteroskedastic and clustered *dgp*.

its true rejection probabilities, while for the wild bootstrap methods I use 1000 wild transformations for each of 10 *base* data draws from *dgp*. Reported numbers are the average of within paper averages. Since comparisons in the paper are often based upon leverage, I divide the sample papers into the low, medium and high leverage groups described in the paper.

As can be seen in Table D1, *wild* does exceptionally poorly. In moving from iid to heteroskedastic and then heteroskedastic and clustered errors, it indicates large Type I error rates in low leverage papers (which is not actually a characteristic of *dgp*), a distinctly non-monotonic relationship in high leverage papers (which again is not a characteristic of *dgp*), and substantially understates the decline in power found in *dgp* in medium and high leverage papers. In contrast, *wild jk* does a much better job of approximating the patterns of Type I error rates and power found in *dgp,* although it does not fully capture the degree to which Type I errors rise and power falls with heteroskedastic and clustered errors in medium and high leverage papers.

Table D2 reports average ln relative OLS to IV truncated relative bias and mean squared error, as well as ln absolute OLS bias, calculated across realized coefficients whose absolute value is less than 1000 times the absolute value of the parameter of the data generating process. As can be seen in the table, *wild* once again does poorly, as both relative bias and relative mean squared error do not rise nearly as fast as in *dgp* with a movement from iid to heteroskedastic and clustered errors, especially in high leverage papers. In contrast, *wild jk* provides a much closer approximation of the movements in relative IV to OLS bias and mean squared error that arise with heteroskedastic and clustered errors at different levels of leverage. Both methods tend to overstate slightly the ln proportional bias of OLS itself, with *wild* doing somewhat better on this metric. This is not a measure, however, that I emphasize much in the paper, beyond noting that it changes little in moving from iid to heteroskedastic errors, which seems to be true in all of the simulations.

At the request of readers, I include simulations using the data generating process produced by *wild jk* in the paper. As shown in the tables above, it approximates IV Type I error rates and power and the relative bias and mse of IV and OLS, which are the results discussed in Section IV of the paper. That said, jackknifed residuals are most certainly not the actual errors of a data generating process and it must be borne in mind that it simply is not possible to extract

Table D2: OLS Bias and Relative Truncated Bias and Mean Squared Error using Wild Bootstrap Data Generating Methods vs Actual Characteristics for the Artificial Data Generating Processes Described in the Paper (average across papers of within paper averages)

	low leverage papers			medium leverage papers			high leverage papers		
	dgp	*wild*	*wild jk*	*dgp*	*wild*	*wild jk*	*dgp*	*wild*	*wild jk*
				ln absolute value of IV to OLS bias					
iid normal	-4.0	-4.0	-4.0	-2.5	-2.5	-2.3	-3.8	-3.8	-3.7
h normal	-2.8	-3.0	-3.1	-1.6	-1.7	-1.4	-1.7	-2.1	-1.7
h cl normal	-1.9	-2.3	-2.2	-1.3	-1.6	-1.3	-0.2	-1.6	-0.5
iid chi^2	-3.8	-3.9	-4.0	-2.6	-2.5	-2.3	-3.9	-3.8	-3.8
h chi^2	-2.7	-3.0	-3.0	-1.6	-1.9	-1.6	-2.1	-2.2	-1.9
h cl chi^2	-2.0	-2.3	-2.2	-1.4	-1.7	-1.4	-0.7	-1.6	-0.6
				ln IV to OLS mean squared error					
iid normal	-0.8	-0.9	-1.0	0.5	0.7	1.1	-0.6	-0.4	-0.2
h normal	1.3	0.6	0.7	2.1	1.7	2.3	2.3	1.4	1.9
h cl normal	2.9	1.9	2.2	2.3	1.8	2.5	4.8	1.7	3.1
iid chi^2	-0.8	-0.8	-0.9	0.5	0.7	1.1	-0.7	-0.5	-0.4
h chi^2	1.1	0.5	0.6	1.8	1.6	2.2	1.5	1.1	1.6
h cl chi^2	2.8	1.7	2.0	2.0	1.7	2.4	3.8	1.6	3.0
				ln OLS bias					
iid normal	-0.6	-0.5	-0.5	-0.3	-0.4	-0.3	-0.6	-0.6	-0.5
h normal	-0.6	-0.3	-0.3	-0.3	-0.3	-0.2	-0.6	-0.5	-0.4
h cl normal	-0.7	-0.3	-0.3	-0.3	-0.4	-0.3	-0.6	-0.4	-0.4
iid chi^2	-0.6	-0.6	-0.5	-0.3	-0.4	-0.3	-0.6	-0.5	-0.5
h chi^2	-0.6	-0.3	-0.3	-0.2	-0.4	-0.3	-0.4	-0.5	-0.4
h cl chi^2	-0.6	-0.3	-0.2	-0.2	-0.4	-0.3	-0.4	-0.4	-0.3

Note: Values calculated based upon truncated central .99 of the coefficient distributions. Bias and mse around the parameter β of the data generating process. Relative bias = ln(|IV bias|/|OLS bias|), relative mse = ln(IV mse/OLS mse), and OLS bias = ln(|OLS bias/β|).

the true residuals from a single realization of *base* data or uncover from these the distribution of results produced by the *dgp* that produced that *base* data. Were such miracles possible, standard errors would not be needed.

E: Comparing Tests of OLS Bias using Monte Carlos

This appendix compares simulation results for two forms of the Durbin (1954) - Wu (1973) - Hausman (1978) tests of OLS bias. The first test is based upon the "artificial regression" suggested by Hausman (1978), wherein the residuals of the 1st stage regression are entered into an OLS version of the 2nd stage regression and, using the notation of the paper, we test of the significance of θ in:

(E1) $\mathbf{y} = \mathbf{Y}\beta + \mathbf{X}\boldsymbol{\delta} + \hat{\mathbf{v}}\theta + \mathbf{u}, \quad \text{where} \quad \hat{\mathbf{v}} = \mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\pi}} - \mathbf{X}\hat{\boldsymbol{\gamma}}.$

The second is based upon the "vector of contrasts", i.e. the difference between the IV and OLS coefficients on **Y** in the second stage regression, using the test statistic:

(E2) $\dfrac{(\hat{\beta}_{iv} - \hat{\beta}_{ols})^2}{\mathrm{V}(\hat{\beta}_{iv}) - \mathrm{V}(\hat{\beta}_{ols})},$

where $\mathrm{V}(\hat{\beta}_{iv})$ and $\mathrm{V}(\hat{\beta}_{ols})$ are estimates of the variance of the two coefficients. (E1) can easily be adapted to a non-iid environment with the use of a clustered/robust variance estimate for θ. However, while with the same n-k finite sample adjustment the default or homoskedastic variance estimate for $\hat{\beta}_{iv}$ is always greater than that for $\hat{\beta}_{ols}$, this is not always the case with clustered/robust variance estimates. Consequently, it is not possible to use non-iid adjustments in tests of the form of (E2), and this leads to large size distortions in the conventional test and comparatively weaker power when jackknife and bootstrap corrections are applied.

Table E1 below presents the relevant simulations. The simulations use the error processes described in 9.1 - 9.6 in the paper, there are 10 simulations per data generating process per equation, and the table reports the average across papers of the average within paper rejection rate. "Correlated errors", based upon the covariance structure of errors found in the residuals of the 2SLS systems of the samples (see (9) in the paper), produce OLS bias and are used in tests of power. "Uncorrelated errors", where the off-diagonal elements of the covariance matrix are set to zero, generate a true null where OLS is unbiased, and are used to estimate Type I error rates.

Table E1: Tests of OLS Bias
(average within paper rejection rates, 10 Monte Carlo for each of 1309 equations)

	Type I error rates (uncorrelated errors)						power (correlated errors)					
	conventional	jack-knife	pairs bootstrap		wild bootstrap		conventional	jack-knife	pairs bootstrap		wild bootstrap	
			c	t	c	t			c	t	c	t
(a) artificial regression: test of θ *in* $\mathbf{y} = \mathbf{Y}\beta + \mathbf{X}\boldsymbol{\delta} + \hat{\mathbf{v}}\theta + \mathbf{u}$ (.01 level)												
iid normal	.021	.008	.005	.008	.006	.010	.373	.255	.216	.262	.266	.306
h normal	.065	.011	.006	.007	.013	.014	.268	.153	.147	.146	.158	.191
cl h normal	.076	.008	.003	.005	.011	.020	.210	.098	.089	.085	.103	.135
iid chi^2	.017	.007	.004	.007	.007	.010	.344	.239	.210	.236	.247	.276
h chi^2	.083	.018	.010	.011	.016	.027	.285	.150	.142	.140	.165	.198
h & cl chi^2	.101	.014	.006	.008	.017	.033	.219	.089	.079	.078	.099	.140
(a) artificial regression: test of θ *in* $\mathbf{y} = \mathbf{Y}\beta + \mathbf{X}\boldsymbol{\delta} + \hat{\mathbf{v}}\theta + \mathbf{u}$ (.05 level)												
iid normal	.071	.036	.030	.041	.044	.050	.493	.373	.354	.405	.408	.450
h normal	.145	.047	.036	.035	.051	.068	.378	.211	.216	.202	.242	.279
cl h normal	.157	.029	.025	.025	.050	.070	.316	.147	.139	.138	.178	.219
iid chi^2	.060	.033	.031	.035	.041	.051	.470	.347	.338	.363	.397	.421
h chi^2	.158	.054	.047	.046	.068	.087	.383	.210	.215	.202	.250	.293
h & cl chi^2	.182	.045	.039	.040	.065	.101	.323	.129	.132	.119	.177	.233
(b) vector of contrasts: test based upon $(\hat{\beta}_{iv} - \hat{\beta}_{ols})^2 / [\mathrm{V}(\hat{\beta}_{iv}) - \mathrm{V}(\hat{\beta}_{ols})]$ (.01 level)												
iid normal	.005	.008	.004	.012	.006	.011	.283	.241	.187	.248	.250	.309
h normal	.238	.012	.006	.010	.011	.015	.429	.148	.134	.144	.150	.188
cl h normal	.434	.005	.003	.005	.010	.016	.546	.081	.070	.077	.091	.134
iid chi^2	.007	.006	.003	.009	.006	.013	.288	.232	.186	.233	.238	.287
h chi^2	.248	.016	.007	.013	.012	.024	.429	.152	.132	.149	.151	.188
cl h chi^2	.420	.009	.004	.008	.011	.027	.539	.081	.070	.076	.089	.129
(b) vector of contrasts: test based upon $(\hat{\beta}_{iv} - \hat{\beta}_{ols})^2 / [\mathrm{V}(\hat{\beta}_{iv}) - \mathrm{V}(\hat{\beta}_{ols})]$ (.05 level)												
iid normal	.038	.035	.027	.051	.042	.052	.421	.349	.311	.398	.397	.451
h normal	.341	.041	.032	.043	.050	.065	.552	.199	.189	.204	.231	.282
cl h normal	.531	.022	.018	.025	.045	.063	.647	.119	.116	.119	.160	.206
iid chi^2	.037	.030	.026	.039	.039	.052	.417	.338	.310	.368	.380	.434
h chi^2	.354	.049	.041	.049	.056	.082	.550	.209	.203	.213	.235	.287
cl h chi^2	.527	.035	.031	.034	.057	.089	.649	.118	.115	.117	.161	.220

Notes: As in Table X in the paper. (a) calculated using cl/robust covariance estimates for both the conventional test and the bootstrap; (b) calculated using default/homoskedastic covariance estimates for both methods.

In the tests based upon "artificial regressions", clustered/robust covariance estimates and associated degrees of freedom are used to evaluate the significance of θ in (E1), whereas in the tests based upon the vector of contrasts default/homoskedastic covariance estimates and the chi squared distribution are used to compute and evaluate (E2). Bootstrap-t covariance estimates follow those used in each conventional test. As shown in the table, with non-iid errors the test based upon the vector of contrasts produces very large size distortions in the conventional test and weaker power in the jackknife and bootstrap tests. Moreover, in the actual analysis of the sample using the jackknife and bootstrap, the artificial regression produces higher rejection rates, i.e. results that are more favourable to the sample (see results reported further below). For these reasons, I report results based upon the artificial regression in Section VI of the paper.

Table F1: Wild Bootstrap Methods (null not imposed)

	estimated residuals	jackknifed residuals
preliminary estimation	$\mathbf{y} = \mathbf{Y}\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \hat{\mathbf{u}}$ $\mathbf{Y} = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \hat{\mathbf{v}}$	$\mathbf{y} = \mathbf{Y}\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \hat{\mathbf{u}}$ $\mathbf{Y} = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \hat{\mathbf{v}}$
adjustment of residuals	$\tilde{\mathbf{v}} = \sqrt{n/(n-k_Z-k_X)} * \hat{\mathbf{v}}$	$\breve{\mathbf{u}}_\mathbf{i} = \mathbf{y}_\mathbf{i} - \mathbf{Y}_\mathbf{i}\hat{\beta}_{iv\sim\mathbf{i}} + \mathbf{X}_\mathbf{i}\hat{\boldsymbol{\delta}}_{\sim\mathbf{i}}$ $\breve{\mathbf{v}}_\mathbf{i} = \mathbf{Y}_\mathbf{i} - \mathbf{Z}_\mathbf{i}\hat{\boldsymbol{\pi}}_{\sim\mathbf{i}} + \mathbf{X}_\mathbf{i}\hat{\boldsymbol{\gamma}}_{\sim\mathbf{i}}$
data generating process	$\mathbf{y}^w = \mathbf{Y}^w\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \boldsymbol{\eta} \circ \hat{\mathbf{u}}$ $\mathbf{Y}^w = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \boldsymbol{\eta} \circ \tilde{\mathbf{v}}$	$\mathbf{y}^w = \mathbf{Y}^w\hat{\beta}_{iv} + \mathbf{X}\hat{\boldsymbol{\delta}} + \boldsymbol{\eta} \circ \breve{\mathbf{u}}$ $\mathbf{Y}^w = \mathbf{Z}\hat{\boldsymbol{\pi}} + \mathbf{X}\hat{\boldsymbol{\gamma}} + \boldsymbol{\eta} \circ \breve{\mathbf{v}}$

Notes: ○ denotes Hadamard product. $\boldsymbol{\eta}$ is composed of observation or cluster level iid draws of a transformation variable, as described in the text. k_Z and k_X denote the number of regressors in **Z** and **X**.

F: Comparing Wild-Bootstrap Methods using Monte Carlos

This section describes various forms of the wild bootstrap and examines their relative performance in Monte Carlos. The methods which impose the null, yielding the most accurate size and following what is considered to be "best practice", are used in the paper.

Table F1 begins by detailing the methods I follow in implementing wild bootstrap tests where the null is not imposed on the data generating process. Following the customary estimation of 2SLS coefficients, the residuals are modified. In the case where "estimated residuals" are used, the modification is a small adjustment of 1st stage residuals for the reduction in variance brought about by OLS fitting.[1] Where "jackknifed residuals" are used, the estimated residuals are replaced with the delete-**i** residuals. The modified residuals are then Hadamard multiplied by a transformation vector $\boldsymbol{\eta}$ which involves iid observation or cluster level[2] draws of a random variable, and added to the estimated 2SLS predicted values to generate $\mathbf{y}^w$ and $\mathbf{Y}^w$.

[1] There is no theoretical justification for modifying 2nd stage residuals in this manner, so they are left as is.

[2] In all simulations or tests reported in the paper and this appendix, I implement the wild bootstrap using transformations that follow authors' covariance estimates, i.e. clustered where they cluster and at the observation level where they do not. I do this even in simulations with non-clustered iid or heteroskedastic error processes, as this allows us to see how the methods used in the tests of the actual sample would perform if the authors' clustering were uncalled for.

Table F2: Wild Bootstrap Methods (null imposed)

	tests of IV coefficients	tests of IV coefficients (RER)
preliminary estimation	$\mathbf{y}-\mathbf{Y}\beta=\mathbf{X}\hat{\boldsymbol{\delta}}+\hat{\mathbf{u}}$ $\mathbf{Y}=\mathbf{Z}\hat{\boldsymbol{\pi}}+\mathbf{X}\hat{\boldsymbol{\gamma}}+\hat{\mathbf{v}}$	$\mathbf{y}-\mathbf{Y}\beta=\mathbf{X}\hat{\boldsymbol{\delta}}+\hat{\mathbf{u}}$ $\mathbf{Y}=\mathbf{Z}\hat{\boldsymbol{\pi}}+\mathbf{X}\hat{\boldsymbol{\gamma}}+\hat{\mathbf{u}}\hat{\alpha}+\hat{\mathbf{v}}$
adjustment of residuals	$\tilde{\mathbf{u}}=\sqrt{n/(n-k_X)}*\hat{\mathbf{u}}$ $\tilde{\mathbf{v}}=\sqrt{n/(n-k_Z-k_X)}*\hat{\mathbf{v}}$	$\tilde{\mathbf{u}}=\sqrt{n/(n-k_X)}*\hat{\mathbf{u}}$ $\tilde{\mathbf{v}}=\sqrt{n/(n-k_Z-k_X)}*(\hat{\mathbf{u}}\hat{\alpha}+\hat{\mathbf{v}})$
data generating process	$\mathbf{y}^w=\mathbf{Y}^w\beta+\mathbf{X}\hat{\boldsymbol{\delta}}+\boldsymbol{\eta}\circ\tilde{\mathbf{u}}$ $\mathbf{Y}^w=\mathbf{Z}\hat{\boldsymbol{\pi}}+\mathbf{X}\hat{\boldsymbol{\gamma}}+\boldsymbol{\eta}\circ\tilde{\mathbf{v}}$	$\mathbf{y}^w=\mathbf{Y}^w\beta+\mathbf{X}\hat{\boldsymbol{\delta}}+\boldsymbol{\eta}\circ\tilde{\mathbf{u}}$ $\mathbf{Y}^w=\mathbf{Z}\hat{\boldsymbol{\pi}}+\mathbf{X}\hat{\boldsymbol{\gamma}}+\boldsymbol{\eta}\circ\tilde{\mathbf{v}}$
	tests of 1st stage coefficients	tests of OLS bias
preliminary estimation	$\mathbf{Y}-\mathbf{Z}\boldsymbol{\pi}=\mathbf{X}\hat{\boldsymbol{\gamma}}+\hat{\mathbf{v}}$	$\mathbf{y}=\mathbf{Y}\hat{\beta}_{ols}+\mathbf{X}\hat{\boldsymbol{\delta}}+\hat{\mathbf{u}}$ $\mathbf{Y}=\mathbf{Z}\hat{\boldsymbol{\pi}}+\mathbf{X}\hat{\boldsymbol{\gamma}}+\hat{\mathbf{v}}$
adjustment of residuals	$\tilde{\mathbf{v}}=\sqrt{n/(n-k_X)}*\hat{\mathbf{v}}$	$\tilde{\mathbf{u}}=\sqrt{n/(n-k_X-1)}*\hat{\mathbf{u}}$ $\tilde{\mathbf{v}}=\sqrt{n/(n-k_Z-k_X)}*\hat{\mathbf{v}}$
data generating process	$\mathbf{Y}^w=\mathbf{Z}\boldsymbol{\pi}+\mathbf{X}\hat{\boldsymbol{\gamma}}+\boldsymbol{\eta}\circ\tilde{\mathbf{v}}$	$\mathbf{y}^w=\mathbf{Y}^w\hat{\beta}_{ols}+\mathbf{X}\hat{\boldsymbol{\delta}}+\boldsymbol{\eta}_2\circ\tilde{\mathbf{u}}$ $\mathbf{Y}^w=\mathbf{Z}\hat{\boldsymbol{\pi}}+\mathbf{X}\hat{\boldsymbol{\gamma}}+\boldsymbol{\eta}_1\circ\tilde{\mathbf{v}}$

Notes: Unless otherwise noted, as in Table F1. RER = restricted efficient residuals.

Following Davidson-Flachaire's (2008) analysis of the wild bootstrap, I consider symmetric transformations where η_i takes on the values [1,-1] with a 50/50 probability, and asymmetric transformations where it takes on the values [(1-√5)/2, (1+√5)/2] with probabilities [(√5+1)/2√5, (√5-1)/2√5]. On each draw of $\boldsymbol{\eta}$, the 2SLS coefficients $\hat{\beta}_{iv}$ and $\hat{\boldsymbol{\pi}}$ and their respective variance estimates can be estimated, allowing implementation of the bootstrap-c and -t, as described in the paper.

An alternative wild bootstrap approach involves imposing the null, as described in Table F2. In this case, preliminary estimation imposes the restriction implied by the null. Aside from the simple imposition of the null, there is also the "wild restricted efficient residual bootstrap"

(Davidson & McKinnon 2010), which uses the 2nd stage OLS residuals to try to get more efficient estimates of 1st stage relations when the instruments may be weak. Since the null varies according to what is being tested, a separate data generating process is used for tests of IV and 1st stage coefficients. The table also presents a wild bootstrap data generating process for tests of OLS bias. As the null is that OLS is unbiased, preliminary estimation uses OLS for both 1st and 2nd stage coefficients. In the case of this test, I will consider two versions of the test: (i) where the transformations on the 1st and 2nd stage residuals are the same, $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$; and (ii) where the transformations are independent. Version (ii) looks to see whether power can be increased by strengthening the null (that the residuals are uncorrelated and OLS is unbiased) to include the assumption that the residuals are actually completely independent (which is not a necessary implication of lack of correlation when errors are non-normal). In the case of each method described in Table F2, on each draw of $\mathbf{y}^w$ and $\mathbf{Y}^w$ one estimates the coefficents (and associated variance estimates) relevant to the null being tested, i.e. $\hat{\beta}_{iv}$ for tests of the IV coefficient, $\hat{\boldsymbol{\pi}}$ for 1st stage coefficients, and, for tests of OLS bias, either the vector of contrasts $\hat{\beta}_{iv}$ - $\hat{\beta}_{ols}$ or coefficient $\hat{\theta}$ on the 1st stage residuals in the artificial 2nd stage OLS regression (Appendix E).

Table F3 below presents Monte Carlo estimates of Type I error rates, comparing methods that impose the true null to those that do not. I use the six data generating processes described in the paper[3] and run 10 Monte Carlo simulations per equation (with 1000 wild bootstrap draws with symmetric transformations used to construct a p-value for each test), i.e. 13090 Monte Carlo p-values per data generating process. Reported is the average across papers of the within paper average rejection rate of true nulls (i.e. the parameter values of the underlying data generating processes). When the null is not imposed and estimated residuals used, wild bootstrap rejection probabilities are grossly larger than nominal value and, in the case of the -c, actually worse than simply cl/robust conventional techniques in tests of 1st stage coefficients

[3]As fewer computer resources were available to me towards the end of this project, I did not run (and hence do not report) the comparisons reported in the table for the data generating process based upon "actual" errors.

Table F3: Wild Bootstrap Inference With and Without Imposing the Null
(average within paper rejection rates of true nulls, 10 Monte Carlo simulations per equation)

	IV coefficients (β_{iv})							
	c		t		c		t	
	.01	.05	.01	.05	.01	.05	.01	.05
	estimated residuals				jackknifed residuals			
iid normal	.036	.091	.033	.080	.016	.053	.028	.069
h normal	.079	.133	.060	.104	.026	.060	.040	.081
cl h normal	.069	.118	.059	.094	.017	.042	.042	.077
iid chi^2	.028	.079	.030	.070	.012	.039	.025	.057
h chi^2	.089	.141	.069	.114	.029	.061	.047	.087
cl h chi^2	.081	.133	.063	.109	.025	.060	.045	.085
	null imposed				null imposed (RER)			
iid normal	.008	.047	.015	.055	.009	.046	.011	.052
h normal	.011	.047	.024	.070	.015	.051	.016	.058
cl h normal	.011	.047	.025	.068	.013	.049	.015	.055
iid chi^2	.006	.037	.014	.049	.007	.038	.010	.045
h chi^2	.017	.060	.037	.087	.018	.066	.025	.072
cl h chi^2	.016	.063	.030	.089	.017	.072	.022	.075

	1st stage F-tests ($\boldsymbol{\pi}$)											
	estimated residuals				jackknifed residuals				null imposed			
	c		t		c		t		c		t	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
iid normal	.075	.141	.033	.087	.039	.082	.022	.063	.010	.053	.012	.056
h normal	.141	.200	.051	.095	.065	.106	.031	.069	.018	.072	.017	.065
cl h normal	.145	.209	.056	.099	.067	.111	.029	.069	.023	.076	.017	.065
iid chi^2	.065	.122	.029	.072	.031	.066	.022	.056	.008	.046	.009	.050
h chi^2	.160	.216	.065	.115	.079	.128	.040	.083	.028	.075	.027	.079
cl h chi^2	.164	.220	.074	.118	.086	.126	.048	.087	.031	.086	.033	.084

Notes: Reported figures are the average across 30 papers of the within paper average rejection rate. c/t = bootstrap-c or bootstrap-t tests using symmetric transformations $\boldsymbol{\eta}$ as described in text accompanying Table F2. .01/.05 = nominal size of the test. iid normal & chi2, heteroskedastic (h) and clustered (cl) denote the data generating process for the Monte Carlo disturbances, as described in 9.1 - 9.6 in the paper. All simulations with correlated 1st and 2nd stage residuals. RER = restricted efficient residuals.

(compare with Table X in the paper). Use of jackknifed residuals improves on these results, moving rejection rates closer to nominal level, but imposing the null in most cases does even better. There are simply very large advantages to knowing the underlying parameter of the data generating process, as is the case when looking for the distribution of a test statistic under a

Table F4: Wild Bootstrap Inference using Symmetric & Asymmetric Transformations
(average within paper rejection rates of true nulls, 10 Monte Carlo simulations per equation)

	symmetric				asymmetric			
	c		t		c		t	
	.01	.05	.01	.05	.01	.05	.01	.05
IV coefficients (null imposed, correlated errors)								
iid normal	.008	.047	.015	.055	.006	.046	.015	.058
h normal	.011	.047	.024	.070	.006	.041	.027	.071
cl h normal	.011	.047	.025	.068	.008	.044	.033	.074
iid chi^2	.006	.037	.014	.049	.004	.035	.016	.050
h chi^2	.017	.060	.037	.087	.011	.054	.038	.089
cl h chi^2	.016	.063	.030	.089	.012	.056	.034	.086
IV coefficients (null imposed, restricted efficient residual, correlated errors)								
iid normal	.009	.046	.011	.052	.007	.046	.012	.052
h normal	.015	.051	.016	.058	.009	.043	.016	.058
cl h normal	.013	.049	.015	.055	.012	.048	.020	.056
iid chi^2	.007	.038	.010	.045	.005	.037	.012	.042
h chi^2	.018	.066	.025	.072	.012	.056	.025	.072
cl h chi^2	.017	.072	.022	.075	.013	.061	.025	.072
1^{st} stage F-tests (null imposed, correlated errors)								
iid normal	.010	.053	.012	.056	.008	.047	.009	.054
h normal	.018	.072	.017	.065	.007	.047	.011	.052
cl h normal	.023	.076	.017	.065	.007	.050	.010	.051
iid chi^2	.008	.046	.009	.050	.006	.043	.007	.047
h chi^2	.028	.075	.027	.079	.011	.057	.019	.066
cl h chi^2	.031	.086	.033	.084	.013	.067	.026	.071

Notes: Symmetric and asymmetric transformations refer to the wild bootstrap draws for $\boldsymbol{\eta}$. Otherwise, as in Table F3.

particular null. Among wild bootstrap methods that impose the null in the evaluation of the significance of instrumented coefficients, those using restricted efficient residuals do appear to produce Type I error rates that are generally somewhat closer to nominal value. Table F4 compares size with symmetric and asymmetric transformations $\boldsymbol{\eta}$ in wild bootstrap methods that impose the null. For IV coefficients inference with asymmetric transformations is sometimes more and sometimes less accurate. In 1^{st} stage tests, asymmetric transformations mostly result in lower rejection rates across all types of data generating processes. This brings Type I error rates closer to or further away from nominal level depending upon whether they are initially above or below it, but does not systematically improve accuracy.

Tables F5 and F6 below explore the impact of different residual transformations on Type I error rates and power in tests of OLS bias (described in Appendix E). In these tables, Type I error rates report the the probability of rejecting the null that OLS is unbiased when 1st and 2nd stage errors are uncorrelated, while power reports rejection rates when they are correlated (see description of simulations in 9.1-9.6 and associated text in paper). Once again, there is no indication that asymmetric transformations allow for systematically more accurate Type I error rates, even in the case of skewed chi^2 error processes. Using independent transformations $\boldsymbol{\eta}$ on the 1st and 2nd stage errors in most instances and on average improves power. The test based upon the artificial regression also appears to be systematically more powerful than that based upon the vector of contrasts, as already noted in Appendix E earlier.

In results reported in the paper itself I impose the null, as this appears to be essential for accurate wild bootstrap inference. For the Monte Carlos (Table X), when estimating Type I error rates I impose the null that the parameter value equals that of the data generating process and when estimating power I impose the null that the parameter value equals zero. For the analysis of the sample itself, I impose the null that the parameter value equals zero. For tests of IV coefficients, in both Monte Carlos and the analysis of the sample, I report results using restricted efficient residuals. In tests of OLS bias, I use the the Hausman test based upon the artificial regression (in preference to the test based upon the vector of contrasts) and independent transformations, as both of these allow greater power. For both Monte Carlos and the analysis of the sample, I use symmetric transformations, as asymmetric transformations neither provide obvious advantages in Monte Carlos nor produce results that are systematically more favourable to the sample. In an appendix further below, I report wild bootstrap results for the sample using **all** methods and tests described in this appendix that impose the null. The range of results varies very little from the subset reported in Section VI of the paper itself.

Table F5: Type I Error Rates & Power in Tests of OLS Bias based on Artificial Regressions (average within paper rejection rates, 10 Monte Carlo simulations per equation)

	$\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$				$\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ independent			
	c		t		c		t	
	.01	.05	.01	.05	.01	.05	.01	.05
	Type I error rates (uncorrelated 1st and 2nd stage errors), symmetric transformations							
iid normal	.006	.042	.010	.051	.006	.044	.010	.050
h normal	.019	.062	.019	.073	.013	.051	.014	.068
cl h normal	.012	.044	.021	.070	.011	.050	.020	.070
iid chi^2	.006	.039	.012	.048	.007	.041	.010	.051
h chi^2	.020	.063	.032	.086	.016	.068	.027	.089
cl h chi^2	.014	.053	.038	.092	.017	.065	.033	.101
	Type I error rates (uncorrelated 1st and 2nd stage errors), asymmetric transformations							
iid normal	.004	.032	.013	.058	.004	.042	.009	.047
h normal	.005	.032	.029	.087	.006	.046	.014	.063
cl h normal	.005	.031	.030	.080	.006	.045	.016	.064
iid chi^2	.002	.026	.014	.054	.004	.038	.009	.045
h chi^2	.005	.037	.041	.099	.009	.054	.028	.084
cl h chi^2	.006	.037	.046	.108	.008	.058	.034	.099
	power (correlated 1st and 2nd stage errors), symmetric transformations							
iid normal	.223	.371	.263	.413	.266	.408	.306	.450
h normal	.153	.225	.185	.276	.158	.242	.191	.279
cl h normal	.084	.145	.113	.195	.103	.178	.135	.219
iid chi^2	.236	.384	.275	.420	.247	.397	.276	.421
h chi^2	.161	.241	.210	.299	.165	.250	.198	.293
cl h chi^2	.088	.158	.141	.227	.099	.177	.140	.233
	power (correlated 1st and 2nd stage errors), asymmetric transformations							
iid normal	.166	.306	.294	.441	.231	.391	.290	.447
h normal	.122	.178	.201	.291	.136	.225	.181	.283
cl h normal	.069	.114	.137	.220	.089	.163	.128	.218
iid chi^2	.158	.294	.284	.423	.207	.375	.254	.414
h chi^2	.123	.190	.217	.304	.135	.229	.188	.290
cl h chi^2	.071	.116	.152	.241	.082	.162	.132	.225

Notes: Type I error rates using uncorrelated 1st and 2nd stage errors; power using correlated 1st and 2nd stage errors. Symmetric and asymmetric transformations refer to the wild bootstrap draws for $\boldsymbol{\eta}$. Otherwise, as in Table F3.

Table F6: Type I Error Rates & Power in Tests of OLS Bias based on the Vector of Contrasts (average within paper rejection rates, 10 Monte Carlo simulations per equation)

	$\eta_1 = \eta_2$				η_1 and η_2 independent			
	c		t		c		t	
	.01	.05	.01	.05	.01	.05	.01	.05
	Type I error rates (uncorrelated 1st and 2nd stage errors), symmetric transformations							
iid normal	.005	.041	.010	.053	.006	.042	.011	.052
h normal	.018	.062	.022	.073	.011	.050	.015	.065
cl h normal	.014	.045	.018	.061	.010	.045	.016	.063
iid chi^2	.006	.037	.013	.050	.006	.039	.013	.052
h chi^2	.020	.060	.026	.078	.012	.056	.024	.082
cl h chi^2	.015	.051	.027	.078	.011	.057	.027	.089
	Type I error rates (uncorrelated 1st and 2nd stage errors), asymmetric transformations							
iid normal	.004	.030	.006	.043	.005	.041	.009	.051
h normal	.005	.031	.021	.068	.005	.042	.020	.074
cl h normal	.005	.029	.022	.063	.006	.042	.019	.072
iid chi^2	.002	.026	.006	.039	.004	.039	.009	.050
h chi^2	.005	.034	.024	.073	.007	.054	.021	.087
cl h chi^2	.006	.035	.031	.084	.007	.053	.029	.097
	power (correlated 1st and 2nd stage errors), symmetric transformations							
iid normal	.214	.356	.271	.424	.250	.397	.309	.451
h normal	.149	.214	.182	.270	.150	.231	.188	.282
cl h normal	.080	.134	.108	.182	.091	.160	.134	.206
iid chi^2	.224	.369	.287	.429	.238	.380	.287	.434
h chi^2	.156	.232	.193	.282	.151	.235	.188	.287
cl h chi^2	.085	.146	.120	.208	.089	.161	.129	.220
	power (correlated 1st and 2nd stage errors), asymmetric transformations							
iid normal	.159	.296	.218	.365	.215	.385	.276	.436
h normal	.120	.174	.173	.256	.127	.217	.181	.284
cl h normal	.067	.110	.120	.189	.081	.153	.136	.213
iid chi^2	.154	.283	.207	.355	.195	.367	.253	.419
h chi^2	.119	.182	.184	.269	.126	.218	.177	.285
cl h chi^2	.069	.109	.126	.202	.077	.153	.128	.222

Notes: As in Table F5.

G: Comparing Symmetric & Asymmetric Tests using Monte Carlos

As noted in the paper, Hall (1992) argues that size in bootstrapped symmetric tests converges more rapidly to nominal value than in asymmetric tests because symmetric tests minimize the influence of skewness. Symmetric tests calculate p-values using the fraction of bootstrapped results that exceed the absolute value of the t-statistic or coefficient deviation from the null, while equal-tailed asymmetric tests calculate the bootstrapped p-value as 2 times the minimum of the fraction of results that are either greater or less than the actual value of the t-statistic or coefficient deviation. Wald based F-statistics are by construction positive and hence not (sensibly) amenable to asymmetric tests.

Table G1 below confirms the finite sample validity of Hall's asymptotic result using the Monte Carlos described earlier.[4] For the pairs bootstrap, in 36 different comparisons of rejection rates for tests of true nulls for IV coefficients (.01 & .05 levels for the nine data generating processes given in the table for the boot-c and boot-t), Type I error rates are closer to nominal value using an asymmetric test only 8 times (with an average improvement of .005) and further from nominal value 28 times (with an average increased deviation of .086), while in 36 comparisons of Hausman tests Type I error rates are closer to nominal value using an asymmetric test 7 times (with an average improvement of .003) and further 29 times (with an average increased deviation of .014). For the wild bootstrap using symmetric transformations, in 36 different comparisons of Type I error rates for tests of IV coefficients asymmetric tests are closer to nominal value 11 times (with an average improvement of .001) and further from nominal value 25 times (with an average increased deviation of .006), while in 36 comparisons for Hausman tests they are closer 19 times (.0006 improvement) and further 17 times (.0008 worse deviation from nominal level). For the wild bootstrap using asymmetric transformations, in 24 comparisons of test of IV coefficients Type I error rates using asymmetric tests are closer

[4]Wild bootstrap tests of IV coefficients are those using the null imposed with restricted efficient residuals, while wild bootstrap tests of OLS bias use independent transformations ($\boldsymbol{\eta}$) of residuals, both as described earlier in Appendix F.

Table G1: Type I Bootstrap Error Rates in Symmetric & Asymmetric Tests
(average within paper rejection rates, 10 Monte Carlo simulations for each of 1309 equations)

	symmetric tests						asymmetric equal-tailed tests					
	pairs bootstrap		symmetric wild bootstrap		asymmetric wild bootstrap		pairs bootstrap		symmetric wild bootstrap		asymmetric wild bootstrap	
	c	t	c	t	c	t	c	t	c	t	c	t
	IV coefficients (correlated errors): .01 level											
iid normal	.009	.021	.009	.011	.007	.012	.017	.056	.008	.009	.006	.027
h normal	.011	.025	.015	.016	.009	.016	.064	.132	.020	.021	.009	.064
h cl normal	.009	.025	.013	.015	.012	.020	.059	.173	.018	.015	.013	.058
iid chi2	.007	.017	.007	.010	.005	.012	.013	.057	.006	.010	.003	.028
h.chi2	.016	.030	.018	.025	.012	.025	.071	.127	.026	.026	.012	.072
h cl chi2	.012	.028	.017	.022	.013	.025	.067	.178	.027	.028	.016	.085
iid "actual"	.007	.019	.007	.011	NA	NA	.012	.065	.008	.010	NA	NA
h "actual"	.005	.022	.010	.014	NA	NA	.015	.089	.013	.015	NA	NA
h cl "actual"	.004	.024	.009	.015	NA	NA	.020	.120	.015	.014	NA	NA
	IV coefficients (correlated errors): .05 level											
iid normal	.042	.065	.046	.052	.046	.052	.053	.113	.047	.050	.037	.080
h normal	.048	.063	.051	.058	.043	.058	.122	.211	.059	.062	.040	.132
h cl normal	.041	.059	.049	.055	.048	.056	.120	.260	.062	.059	.046	.121
iid chi2	.033	.050	.038	.045	.037	.042	.048	.111	.038	.046	.026	.081
h.chi2	.059	.072	.066	.072	.056	.072	.137	.204	.076	.080	.048	.143
h cl chi2	.051	.067	.072	.075	.061	.072	.136	.266	.085	.085	.057	.160
iid "actual"	.035	.060	.042	.050	NA	NA	.054	.132	.045	.047	NA	NA
h "actual"	.035	.063	.049	.059	NA	NA	.060	.164	.055	.058	NA	NA
h cl "actual"	.032	.064	.045	.057	NA	NA	.068	.196	.062	.068	NA	NA

Notes: At end of table below.

to nominal value 3 times (with an average improvement of .003) and further from nominal value 21 times (with an average increased deviation of .029), while in Hausman tests they are closer 2 times (.003 improvement) and worse 22 times (.017 increased deviation). Thus, in finite samples symmetric tests are seen to have rejection rates that are systematically closer to nominal value.

Table G1: Type I Bootstrap Error Rates in Symmetric & Asymmetric Tests (continued)

	symmetric tests						asymmetric equal-tailed tests					
	pairs bootstrap		symmetric wild bootstrap		asymmetric wild bootstrap		pairs bootstrap		symmetric wild bootstrap		asymmetric wild bootstrap	
	c	t	c	t	c	t	c	t	c	t	c	t
Hausman Tests of OLS Bias (uncorrelated errors): .01 level												
iid normal	.005	.008	.006	.010	.004	.009	.008	.006	.007	.010	.004	.022
h normal	.006	.007	.013	.014	.006	.014	.041	.006	.012	.014	.005	.035
h cl normal	.003	.005	.011	.020	.006	.016	.040	.005	.011	.019	.005	.035
iid chi2	.004	.007	.007	.010	.004	.009	.006	.006	.006	.010	.002	.024
h chi2	.010	.011	.016	.027	.009	.028	.048	.010	.016	.029	.007	.053
h cl chi2	.006	.008	.017	.033	.008	.034	.047	.008	.017	.035	.008	.057
iid "actual"	.003	.007	.024	.013	NA	NA	.000	.005	.025	.013	NA	NA
h "actual"	.003	.007	.021	.021	NA	NA	.000	.007	.021	.019	NA	NA
h cl "actual"	.003	.007	.021	.022	NA	NA	.001	.004	.021	.022	NA	NA
Hausman Tests of OLS Bias (uncorrelated errors): .05 level												
iid normal	.030	.041	.044	.050	.042	.047	.041	.033	.043	.050	.034	.070
h normal	.036	.035	.051	.068	.046	.063	.100	.028	.050	.067	.031	.109
h cl normal	.025	.025	.050	.070	.045	.064	.087	.020	.049	.069	.033	.105
iid chi2	.031	.035	.041	.051	.038	.045	.035	.027	.041	.051	.030	.074
h chi2	.047	.046	.068	.087	.054	.084	.106	.038	.067	.088	.040	.127
h cl chi2	.039	.040	.065	.101	.058	.099	.099	.029	.063	.101	.047	.131
iid "actual"	.024	.040	.050	.052	NA	NA	.001	.032	.050	.051	NA	NA
h "actual"	.028	.041	.051	.068	NA	NA	.002	.032	.049	.068	NA	NA
h cl "actual"	.024	.037	.052	.075	NA	NA	.003	.030	.053	.077	NA	NA

Notes: Symmetric and asymmetric in the context of the wild bootstrap refer to the residual transformations, as described earlier in Appendix F. Symmetric versus asymmetric equal-tailed in the context of tests refer to use of the absolute value of coefficients and t-statistics versus the actual value of the coefficients and t-statistics, as described in the text above. NA = not available, due to limitations on computer resources towards the end of this project these simulations were not performed. Reported figures are the average across 30 papers of the within paper average rejection rate.

Table H1: Type I Error Rates of the BCA Bootstrap Compared to other Methods
(average within paper rejection rates, 10 Monte Carlo simulations for each of 1309 equations)

	clustered /robust		bca bootstrap		pairs bootstrap symmetric tests				pairs bootstrap asymmetric equal tailed tests			
					c		t		c		t	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
	IV coefficients (correlated errors)											
iid normal	.028	.081	.025	.068	.009	.042	.021	.065	.017	.052	.056	.113
h normal	.069	.126	.055	.128	.011	.048	.025	.063	.064	.122	.132	.211
h cl normal	.070	.124	.072	.155	.009	.041	.025	.059	.059	.120	.173	.260
iid chi2	.024	.065	.025	.068	.007	.033	.017	.050	.013	.048	.057	.111
h chi2	.080	.144	.061	.126	.016	.059	.030	.072	.071	.137	.127	.204
h cl chi2	.075	.141	.082	.162	.012	.051	.028	.067	.067	.136	.178	.266
iid "actual"	.025	.073	.029	.078	.007	.035	.019	.060	.012	.054	.065	.132
h "actual"	.034	.081	.034	.091	.005	.035	.022	.063	.015	.060	.089	.164
h cl "actual"	.035	.083	.039	.099	.004	.032	.024	.064	.020	.068	.120	.196
	Hausman Tests of OLS Bias based on Artificial Regressions (uncorrelated errors)											
iid normal	.021	.071	.020	.061	.005	.030	.008	.041	.008	.041	.006	.033
h normal	.065	.145	.052	.122	.006	.036	.007	.035	.041	.100	.006	.028
h cl normal	.076	.157	.067	.142	.003	.025	.005	.025	.040	.087	.005	.020
iid chi2	.017	.060	.024	.066	.004	.031	.007	.035	.006	.035	.006	.027
h chi2	.083	.158	.052	.127	.010	.047	.011	.046	.048	.106	.010	.038
h cl chi2	.101	.182	.064	.155	.006	.039	.008	.040	.047	.099	.008	.029
iid "actual"	.023	.073	.023	.077	.003	.024	.007	.040	.000	.001	.005	.032
h "actual"	.039	.100	.034	.095	.003	.028	.007	.041	.000	.002	.007	.032
h cl "actual"	.049	.111	.043	.102	.003	.024	.007	.037	.001	.003	.004	.030

Notes: Symmetric and asymmetric in this context refer to tests using the absolute value of the t-statistic and equal tailed tests using the percentiles of the t-statistic, respectively, as described earlier in Appendix G. .01/.05 = level.

H: Monte Carlos for the Bias Corrected and Accelerated Bootstrap

As noted in a footnote in the paper, the bias corrected and accelerated (BCA) bootstrap is another refinement of the pairs resampling bootstrap. By correcting for skewness, it asymptotically provides $O(n^{-1})$ convergence to nominal size, as opposed to the $O(n^{-½})$ achieved by the bootstrap-c in asymmetric equal tailed tests. The convergence rate of the bootstrap-t in asymmetric equal tailed tests is also $O(n^{-1})$, but the bootstrap-t is not transformation respecting, so the BCA method in theory provides a means of attaining $O(n^{-1})$ performance with a transformation respecting asymmetric test (Hall 1992, Efron & Tibshirani 1994).

Table H1 above applies the BCA method to the Monte Carlos described in the paper and compares results to those found using conventional symmetric clustered/robust tests and the pairs bootstrap c & t in symmetric and asymmetric tests. As shown, in finite sample tests of IV coefficients the BCA method actually performs worse than asymmetric bootstrap-c methods or even conventional symmetric clustered/robust tests (which are also asymptotically $O(n^{-1})$), although it does perform better than the asymmetric bootstrap-t test. It is, however, completely dominated by symmetric bootstrap tests, both -c and -t, which provide much more accurate Type I error rates in tests of IV coefficients. In the Hausman test of OLS bias, the BCA method has size distortions that are somewhat less than the conventional clustered/robust test, but clearly worse than the bootstrap-t in symmetric and asymmetric tests, particularly at the .01 level. In sum, the BCA method does not appear to provide accurate inference in finite samples. It also does not provide reliable improvements over the bootstrap-c and -t in asymmetric tests and is very much dominated by these two methods when they are used in symmetric tests.

Table I1: Significance of OLS Coefficients
(supplement for Table XI)

	all results		headline results							
			all		low		medium		high	
	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05
clustered/robust	.543	.638	.615	.654	.750	.750	.496	.496	.600	.717
jackknife	.466	.589	.529	.633	.750	.750	.350	.496	.486	.652
pairs bootstrap - c	.458	.592	.545	.633	.750	.750	.350	.496	.536	.652
pairs bootstrap - t	.472	.599	.490	.633	.650	.750	.363	.496	.457	.652
wild bootstrap - c	.474	.618	.516	.638	.750	.750	.363	.496	.436	.669
wild bootstrap - t	.455	.608	.516	.605	.750	.750	.363	.496	.436	.569

Notes: Unless otherwise noted, as in Table XI in the paper. Wild bootstrap methods impose the null.

I: OLS Significance Rates in the Sample

Section VI of the paper analyzes the sample's results using jackknife and bootstrap methods. In the discussion of the large differences between bootstrap-c and -t significance rates for instrumented coefficients in Table XI, I note that no such differences exist when these techniques are applied to OLS versions of the estimating equations. Table I1 above reports rejection rates for OLS estimates of the (otherwise) instrumented coefficient in authors' 2^{nd} stage regressions, and shows that this is by and large the case. The only instance where a large difference between -c and -t methods arises is in the pairs bootstrap analysis of headline results at the .01 level, and even here the difference is proportionately much smaller than the comparable difference for IV versions in Table XI and *in the opposite direction* (with -t methods showing lower rather than higher rates of significance). In the paper I argue that the discrepancy between -c and -t results reflects publication bias which selects in favour of spuriously significant IV t-statistics which, as the comparison between -c and -t methods shows, are characterized by unusually large t-statistics rather than unusually large coefficient estimates under the null. No such difference exists in tests of OLS coefficients, which do not form the basis for the publication decision.

J: Alternative Wild Bootstrap & OLS Bias Results for the Sample

In the paper I analyse the sample using wild bootstrap methods with symmetric transformations in symmetric two-sided tests with the null imposed and, in the case of IV coefficients, following the recommendation of Davidson & MacKinnon (2010), restricted efficient residuals. Monte Carlo simulations show that wild bootstrap tests with the null imposed have decidedly more accurate Type I error rates than those without, but other choices I have made are based upon smaller advantages (Appendices E, F & G above). In Table J1 I compare the results reported in the paper (in bold) with those found using the wild bootstrap with asymmetric transformations ($\boldsymbol{\eta}$ in Appendix F), asymmetric equal tailed tests, and tests of IV coefficients that simply impose the null (without restricted efficient residuals). As can be seen, using asymmetric transformations generally produces lower significance rates in the 1st stage F-test than are reported in the paper. In the Hausman test, using independent transformations of the residuals ($\eta_1 \neq \eta_2$), as done in the paper, produces higher rejection rates than using the same transformations in the 1st and 2nd stage ($\eta_1 = \eta_2$). In the context of symmetric Hausman tests, asymmetric transformations do not produce higher rejection rates than those reported in the paper. Asymmetric equal tailed Hausman tests produce the same or lower rejection rates as those reported in the paper, except in the case of those with asymmetric transformations for asymmetric bootstrap-t tests which, as can be seen in Table G1 earlier, have sizeable size distortions in simulation. In tests of IV coefficients, wild bootstrap tests that simply impose the null without restricted efficient residuals produce somewhat higher -c rejection rates and slightly lower -t rejection rates. I reported restricted efficient residual results in the paper for fear that wild bootstrap users, who appear to be convinced that these are the best, would reject the results out of hand if I did not use this method. Otherwise, asymmetric transformations in symmetric tests produce lower rejection rates, as do most asymmetric equal tailed tests. The only exception is once again asymmetric equal tailed bootstrap-t tests with asymmetric transformations which again, as can be seen in Table G1 earlier, appear to have substantial size distortions.

Table J1: Wild Bootstrap Inference in the Sample with the Null Imposed
in Symmetric & Asymmetric Transformations & Tests
(average within paper rejection rates by level of the test)

test:	symmetric two-sided				asymmetric equal tailed			
transformation:	symmetric		asymmetric		symmetric		asymmetric	
level:	.01	.05	.01	.05	.01	.05	.01	.05
				all results				
IV coefficients:								
bootstrap - c (RER)	**.115**	**.337**	.106	.322	.204	.383	.141	.304
bootstrap - t (RER)	**.346**	**.535**	.340	.506	.322	.508	.439	.581
bootstrap - c	.153	.377	.147	.364	.229	.440	.163	.332
bootstrap - t	.343	.533	.324	.503	.329	.512	.445	.591
Hausman test:								
bootstrap - c ($\eta_1=\eta_2$)	.085	.236	.060	.142	.075	.178	.059	.109
bootstrap - t ($\eta_1=\eta_2$)	.156	.323	.167	.346	.122	.278	.184	.333
bootstrap - c ($\eta_1\neq\eta_2$)	**.129**	**.247**	.112	.242	.123	.247	.086	.215
bootstrap - t ($\eta_1\neq\eta_2$)	**.175**	**.328**	.171	.335	.168	.333	.241	.427
1st stage:								
bootstrap - c	**.704**	**.886**	.557	.823	NA	NA	NA	NA
bootstrap - t	**.660**	**.856**	.638	.847	NA	NA	NA	NA
				headline results				
IV coefficients:								
bootstrap - c (RER)	**.231**	**.444**	.194	.467	.334	.544	.205	.453
bootstrap - t (RER)	**.512**	**.719**	.459	.677	.492	.682	.596	.774
bootstrap - c	.235	.508	.231	.560	.387	.654	.258	.484
bootstrap - t	.512	.702	.454	.677	.503	.682	.567	.774
Hausman test:								
bootstrap - c ($\eta_1=\eta_2$)	.153	.252	.067	.186	.081	.261	.033	.108
bootstrap - t ($\eta_1=\eta_2$)	.170	.404	.220	.412	.159	.358	.201	.404
bootstrap - c ($\eta_1\neq\eta_2$)	**.187**	**.319**	.153	.323	.187	.352	.067	.308
bootstrap - t ($\eta_1\neq\eta_2$)	**.237**	**.470**	.220	.428	.237	.470	.280	.498
1st stage:								
bootstrap - c	**.794**	**.967**	.724	.917	NA	NA	NA	NA
bootstrap - t	**.783**	**.952**	.758	.971	NA	NA	NA	NA

Notes: RER = restricted efficient residuals. Figures in bold are those reported in the paper. All methods with the null imposed. NA – not applicable, as the 1st stage F-test is often a joint test of multiple coefficients where the test statistic is, by construction, positive. $\eta_1=\eta_2$ vs $\eta_1\neq\eta_2$: whether the transformations for the wild residuals are the same for both the 1st and 2nd stage or independent, as discussed in Appendix F above.

Table J2: Rejection Rates in Tests of OLS Bias in the Sample
(average within paper rejection rates by level of the test)

	all results		headline results		all results		headline results	
	.01	.05	.01	.05	.01	.05	.01	.05
	artificial regression: test of θ in $\mathbf{y} = \mathbf{Y}\beta + \mathbf{X\delta} + \hat{\mathbf{v}}\theta + \mathbf{u}$				vector of contrasts: test based upon $(\hat{\beta}_{iv} - \hat{\beta}_{ols})^2 / [\mathrm{V}(\hat{\beta}_{iv}) - \mathrm{V}(\hat{\beta}_{ols})]$			
clustered/robust	**.232**	**.382**	**.309**	**.445**	.252	.382	.318	.464
jackknife	**.135**	**.227**	**.188**	**.254**	.116	.199	.138	.221
pairs bootstrap - c	**.098**	**.200**	**.138**	**.249**	.079	.183	.138	.238
pairs bootstrap - t	**.110**	**.243**	**.110**	**.300**	.109	.257	.148	.261
wild bootstrap - c	**.129**	**.247**	**.187**	**.319**	.113	.239	.183	.319
wild bootstrap - t	**.175**	**.328**	**.237**	**.470**	.178	.358	.253	.443

Notes: Figures in bold are those reported in the paper. Symmetric two-sided tests in all cases.

Table J2 reports alternative results for tests of OLS bias in the sample. In Table XIV in the paper I report results based upon the significance of the coefficient on the 1st stage residuals entered into an artificial 2nd stage OLS regression using clustered/robust covariance estimates. As noted in Appendix E above, an alternative test based upon the vector of contrasts, i.e. the differences between 2nd stage IV and OLS coefficients, in non-iid error environments has large size distortions in the conventional test and less power when evaluated using the jackknife or bootstrap. Table J2 shows that in the analysis of the sample the vector of contrasts generally provides higher rejection rates in the conventional test (an average of .012 higher in 4 comparisons between the first rows of the left and right panels of the table) and lower rejection rates in the jackknife and bootstrap versions of the tests (an average of .008 lower in 20 comparisons between the bottom five rows of the left and right panels in the table). Since I emphasize the jackknife and bootstrap results in the paper, and the conventional vector of contrasts test has large size distortions with non-iid errors (Appendix E above), I report results based on the artificial regression in the paper.

Table K1: Leverage, Heteroskedasticity and Differences in IV P-Values
(alternative p-values - cl/robust p-value regressed on leverage & homoskedasticity p-value)

		jackknife	pairs boot-t	pairs boot-c	wild boot-t	wild boot-c
max leverage	β	.217	.059	.154	.112	.213
	s.e.	(.053)	(.045)	(.049)	(.033)	(.062)
	p-v	.000	.319	.016	.016	.021
max lev x homoskedasticity p-value	β	-3.97	-.829	-2.04	-.356	-.273
	s.e.	(1.46)	(.423)	(.941)	(.274)	(1.03)
	p-v	.229	.210	.218	.276	.850
homoskedasticity p-value	β	.866	.203	.442	.006	.076
	s.e.	(.291)	(.096)	(.190)	(.067)	(.212)
	p-v	.231	.253	.216	.933	.799
constant	β	.023	.006	.015	-.013	.004
	s.e.	(.013)	(.008)	(.011)	(.008)	(.013)
	p-v	.164	.459	.248	.085	.769
R^2		.368	.147	.192	.227	.213

Notes: Each column represents a separate regression. Each observation is a paper average, so there are 30 observations in each regression. β & s.e. = coefficient and heteroskedasticity robust standard error, p-v = bootstrap-t p-value calculated using 1000 bootstrap draws. Max lev = maximum instrument leverage share of single observation or cluster (paper level average), as in Table II in the paper. Homoskedasticity p-value = Koenker (1981) p-value on test that residuals are homoskedastic, as in Table III in the paper. Results using Wooldridge (2013) p-value are almost identical.

K: Leverage, Heteroskedasticity and Differences in IV P-Values

Table K1 above regresses the difference between the jackknife and bootstrap p-values and the conventional clustered/robust p-values for the sample regressions (Section VI in the paper) on maximum leverage, the p-value on the test of homoskedasticity, and the interaction between the two. Observations are paper averages, so there are 30 observations in each column's regression. The maximum leverage share of the largest cluster or observation is always positively associated with p-value differences, and this effect is larger when the average p-value on the test that the 1st stage residuals are homoskedastic is low. These results are consistent with the Monte Carlo simulations presented in the paper which indicated that clustered/robust p-values have larger size distortions when leverage is high and the residuals are heteroskedastic. However, although many of the coefficients in the table are deemed to be statistically significant when evaluated using heteroskedasticity robust standard errors, only the coefficients on maximum leverage are found to be significant when evaluated using the bootstrap-t, as reported in the table. The average homoskedasticity p-value is close to zero in ¾ of the papers, so the bootstrap resampling finds that the results are heavily sensitive to a few observations and not statistically significant. These results were described in Section VI in the paper.

L: Papers in the Instrumental Variables Sample

Acconcia, Antonio, Giancarlo Corsetti, and Saverio Simonelli. 2014. "Mafia and Public Spending: Evidence on the Fiscal Multiplier from a Quasi-Experiment." *American Economic Review*, 104(7): 2185–2209.

Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared. 2008. "Income and Democracy." *American Economic Review*, 98 (3): 808–842.

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

Alesina, Alberto, and Ekaterina Zhuravskaya. 2011. "Segregation and the Quality of Government in a Cross Section of Countries." *American Economic Review*, 101 (5): 1872-1911.

Ananat, Elizabeth Oltmans. 2011. "The Wrong Side(s) of the Tracks: The Causal Effects of Racial Segregation on Urban Poverty and Inequality." *American Economic Journal: Applied Economics*, 3 (2): 34–66.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." *American Economic Review*, 103 (6): 2121–2168.

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.

Becker, Sascha O., Erik Hornung, and Ludger Woessmann. 2011. "Education and Catch-up in the Industrial Revolution." *American Economic Journal: Macroeconomics*, 3 (3): 92–126.

Bedard, Kelly, and Olivier Deschênes. 2006. "The Long-Term Impact of Military Service on Health: Evidence from World War II and Korean War Veterans." *American Economic Review*, 96 (1): 176-194.

Bleakley, Hoyt, and Aimee Chin. 2010. "Age at Arrival, English Proficiency, and Social Assimilation Among US Immigrants." *American Economic Journal: Applied Economics*, 2 (1): 165–192.

Brown, Kristine M., and Ron A. Laschever. 2012. "When They're Sixty-Four: Peer Effects and the Timing of Retirement." *American Economic Journal: Applied Economics*, 4(3): 90–115.

Burke, Paul J., and Andrew Leigh. 2010. "Do Output Contractions Trigger Democratic Change?" *American Economic Journal: Macroeconomics*, 2 (4): 124–157

Chalfin, Aaron. 2015. "The Long-Run Effect of Mexican Immigration on Crime in US Cities: Evidence from Variation in Mexican Fertility Rates." *American Economic Review: Papers & Proceedings*, 105(5): 220–225.

Chodorow-Reich, Gabriel, Laura Feiveson, Zachary Liscow, and William Gui Woolston. 2012. "Does State Fiscal Relief During Recessions Increase Employment? Evidence from the American Recovery and Reinvestment Act." *American Economic Journal: Economic Policy*, 4(3): 118–145.

Chou, Shin-Yi, Jin-Tan Liu, Michael Grossman, and Ted Joyce. 2010. "Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan." *American Economic Journal: Applied Economics*, 2 (1): 33–61.

Collins, William J., and Katharine L. Shester. 2013. "Slum Clearance and Urban Renewal in the United States." *American Economic Journal: Applied Economics*, 5(1): 239–273.

Decarolis, Francesco. 2015. "Medicare Part D: Are Insurers Gaming the Low Income Subsidy Design." *American Economic Review*, 105 (4): 1547–1580.

Dinkelman, Taryn. 2011. "The Effects of Rural Electrification on Employment: New Evidence from South Africa." *American Economic Review*, 101 (7): 3078-3108.

Draca, Mirko, Stephen Machin and Robert Witt. 2011. "Panic on the Streets of London: Police, Crime, and the July 2005 Terror Attacks." *American Economic Review*, 101 (5): 2157-2181.

Guryan, Jonathan, and Melissa S. Kearney. 2010. "Is Lottery Gambling Addictive?" *American Economic Journal: Economic Policy*, 2 (3): 90–110

Hornung, Erik. 2014. "Immigration and the Diffusion of Technology: The Huguenot Diaspora in Prussia." *American Economic Review*, 104(1): 84–122.

Hunt, Jennifer, and Marjolaine Gauthier-Loiselle. 2010. "How Much Does Immigration Boost Innovation?" *American Economic Journal: Macroeconomics*, 2 (2): 31–56.

James, Alexander. 2015. "US State Fiscal Policy and Natural Resources." *American Economic Journal: Economic Policy*, 7(3): 238–257.

Kraay, Aart. 2014. "Government Spending Multipliers in Developing Countries: Evidence from Lending by Official Creditors." *American Economic Journal: Macroeconomics*, 6(4): 170–208.

Lipscomb, Molly, A. Mushfiq Mobarak, and Tania Barham. 2013. "Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil." *American Economic Journal: Applied Economics*, 5(2): 200–231.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. "German Jewish Émigrés and US Invention." *American Economic Review*, 104(10): 3222–3255.

Oreopoulos, Philip. 2006. "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter." *American Economic Review*, 96 (1): 152-175.

Saiz, Albert, and Susan Wachter. 2011. "Immigration and the Neighborhood." *American Economic Journal: Economic Policy*, 3 (2): 169–188.

Thornton, Rebecca L. 2008. "The Demand for, and Impact of, Learning HIV Status." *American Economic Review*, 98 (5): 1829-1863.

Young, Alwyn. 2014. "Structural Transformation, the Mismeasurement of Productivity Growth, and the Cost Disease of Services." *American Economic Review*, 104 (11): 3635–3667.

Bibliography[5]

Davidson, Russell and Emmanuel Flachaire. 2008. "The wild bootstrap, tamed at last." *Journal of Econometrics* 146 (): 162-169.

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

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

Efron, Bradley and Robert J. Tibshirani. An Introduction to the Bootstrap. New York: Chapman and Hall/CRC, 1994.

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.

Koenker, Roger. 1981. "A Note on Studentizing a Test for Heteroskedasticity." Journal of Econometrics 17: 107-112.

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

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.

Wooldridge, J. M. 2013. Introductory Econometrics: A Modern Approach. 5th ed. Mason, OH: South-Western.

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

[5]Sources cited in this appendix.