Testing Overidentifying Restrictions with Many Instruments and Heteroskedasticity*

John C. Chao, Department of Economics, University of Maryland, chao@econ.umd.edu.

Jerry A. Hausman, Department of Economics, MIT, jhausman@mit.edu.

Whitney K. Newey, Department of Economics, MIT, wnewey@mit.edu.

Norman R. Swanson, Department of Economics, Rutgers University, nswanson@econ.rutgers.edu.

Tiemen Woutersen, Department of Economics, Johns Hopkins University, woutersen@jhu.edu.

June 2010

Revised October 2012

Corresponding Author:

Whitney K. Newey

Department of Economics

MIT, E52-262D

Cambridge, MA 02142-1347

Abstract

This paper gives a test of overidentifying restrictions that is robust to many instruments and heteroskedasticity. It is based on a jackknife version of the overidentifying test statistic. Correct asymptotic critical values are derived for this statistic when the number of instruments grows large, at a rate up to the sample size. It is also shown that the test is valid when the number instruments is fixed and there is homoskedasticity. This test improves on recently proposed tests by allowing for heteroskedasticity and by avoiding assumptions on the instrument projection matrix. The distribution theory is based on the heteroskedasticity-robust, many-instrument asymptotics of Chao et al. (2010). In Monte Carlo experiments the rejection frequency of the test is found to be very insensitive to the number of instruments. This paper finds in Monte Carlo studies that the test is more accurate and less sensitive to the number of instruments than the Hausman-Sargan or GMM tests of overidentifying restrictions.

1   Introduction

The Sargan (1958) and Hansen (1982) tests of overidentifying restrictions validity can be sensitive
to the number of restrictions being tested. This paper proposes an alternative test that is robust
to many instruments and to heteroskedasticity. It is based on subtracting out the diagonal terms
in the numerator of the overidentifying test statistic and normalizing appropriately. This test has a
jackknife interpretation, as it is based on the objective function of the JIVE2 estimator of Angrist,
Imbens, and Krueger (1999).

We show that the test has correct rejection frequency as long as the number of instruments
goes to infinity with the sample size at any rate up to the sample size. It is also correct under
homoskedasticity with a fixed number of instruments. In Monte Carlo experiments we find that the
rejection frequency is close to its nominal values in all cases we consider, including heteroskedastic
ones with few overidentifying restrictions. In these ways this test solves the problem of sensitivity
to the number of overidentifying restrictions being tested.

Recently Anatolyev and Gospodinov (2009) and Lee and Okui (2010) have formulated tests that
allow for many instruments but impose homoskedasticity. Our test is valid under their conditions
and also with heteroskedasticity. Also, we do not impose side conditions on the instrument projec-
tion matrix. The asymptotic theory is based on the results of Chao et. al. (2010) and Hausman et
al. (2010), including a central limit theorem that imposes no side conditions on the instrumental
variable projection matrix.

In Section 2 we describe the model and test statistic. In Section 3 we give the asymptotic
theory. Section 4 reports the Monte Carlo results.

2   The Model and Test Statistic

We adopt the same model and notation as in Hausman et al. (2010) and Chao et al. (2010). The
model we consider is given by

$$
\begin{aligned}
\underset{n\times 1}{y} &= \underset{n\times G}{X}\,\underset{G\times 1}{\delta_0} + \underset{n\times 1}{\varepsilon}, \\
X &= \Upsilon + U,
\end{aligned}
$$

where n is the number of observations, G is the number of right-hand side variables, Υ is the
reduced form matrix, and U is the disturbance matrix. For the asymptotic approximations, the

elements of Υ will be implicitly allowed to depend on n, although we suppress dependence of Υ on n, for notational convenience. Estimation of δ_0 will be based on an $n \times K$ matrix, Z, of instrumental variable observations with $rank(Z) = K$. Here we will treat Z and Υ as nonrandom for simplicity though it is possible to do asymptotic theory conditional on these as in Chao et al. (2010). We will assume that $E[\varepsilon] = 0$ and $E[U] = 0$. Further explanation of this framework is provided in Section 3.

This model allows for Υ to be a linear combination of Z (i.e. $\Upsilon = Z\pi$, for some $K \times G$ matrix π). Furthermore, some columns of X may be exogenous, with the corresponding column of U being zero. The model also allows for Z to approximate the reduced form. For example, let X_i', Υ_i', and Z_i' denote the ith row (observation) for X, Υ, and Z, respectively. We could let $\Upsilon_i = f_0(w_i)$ be a vector of unknown functions of a vector w_i of underlying instruments and let $Z_i = (p_{1K}(w_i), ..., p_{KK}(w_i))'$, for approximating functions $p_{kK}(w)$, such as power series or splines. In this case, linear combinations of Z_i may approximate the unknown reduced form.

For estimation of δ we consider the heteroskedasticity-robust version of the Fuller (1977) estimator of Hausman et al. (2010), referred to as HFUL. Other heteroskedasticity and many instrument robust estimators could also be used, such as jackknife instrumental variable (IV) estimators of Angrist, Imbens, and Krueger (1999) or the continuously updated GMM estimator (CUE). We focus on HFUL because of its high efficiency relative to jackknife IV, because it has moments, and because it is computationally simple relative to CUE. To describe HFUL, let

$$P = Z(Z'Z)^{-1}Z',$$

P_{ij} denote the ij^{th} element of P, and $\bar{X} = [y, X]$. Let

$$\tilde{\alpha} \text{ be the smallest eigenvalue of } (\bar{X}'\bar{X})^{-1}(\bar{X}'P\bar{X} - \sum_{i=1}^{n} P_{ii}\bar{X}_i\bar{X}_i').$$

Although the matrix in this expression is not symmetric it has real eigenvalues because $(\bar{X}'\bar{X})^{-1}$ is positive definite and $\bar{X}'P\bar{X} - \sum_{i=1}^{n} P_{ii}\bar{X}_i\bar{X}_i'$ is symmetric. Let

$$\hat{\alpha} = [\tilde{\alpha} - (1 - \tilde{\alpha})/T]/[1 - (1 - \tilde{\alpha})/T].$$

HFUL is given by

$$\hat{\delta} = \left(X'PX - \sum_{i=1}^{n} P_{ii}X_iX_i' - \hat{\alpha}X'X \right)^{-1} \left(X'Py - \sum_{i=1}^{n} P_{ii}X_iy_i - \hat{\alpha}X'y \right).$$

Thus, HFUL can be computed by finding the smallest eigenvalue of a matrix and then using this explicit formula. Motivation for HFUL is further discussed in Hausman et al. (2010).

To describe the overidentification statistic, let $\hat{\varepsilon}_i = y_i - X_i'\hat{\delta}$, $\hat{\varepsilon} = (\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)'$, $\hat{\varepsilon}(2) = (\hat{\varepsilon}_1^2, ..., \hat{\varepsilon}_n^2)'$, and $P(2)$ be the n-dimensional square matrix with ij^{th} component equal to P_{ij}^2. Also, let $\sum_{i \neq j}$ denote the double sum over all i not equal to j. The test statistic is

$$\hat{T} = \frac{\hat{\varepsilon}'P\hat{\varepsilon} - \sum_{i=1}^n P_{ii}\hat{\varepsilon}_i^2}{\sqrt{\hat{V}}} + K, \hat{V} = \frac{\hat{\varepsilon}(2)'P(2)\hat{\varepsilon}(2) - \sum_i P_{ii}^2\hat{\varepsilon}_i^4}{K} = \frac{\sum_{i \neq j} \hat{\varepsilon}_i^2 P_{ij}^2 \hat{\varepsilon}_j^2}{K}$$

Treating $\hat{T}$ as if it is chi-squared with $K - G$ degrees of freedom will be asymptotically correct if $K \longrightarrow \infty$ no faster than n and when K is fixed and ε_i is homoskedastic. Let $q_r(\tau)$ be the τth quantile of the chi-squared distribution with r degrees of freedom. A test with asymptotic rejection frequency α will reject the null hypothesis if

$$\hat{T} \geq q_{K-G}(1 - \alpha).$$

We will show that the test with this critical region has a probability of rejection that converges to α.

To explain the form of this test statistic, note that the numerator is

$$\hat{\varepsilon}'P\hat{\varepsilon} - \sum_{i=1}^n P_{ii}\hat{\varepsilon}_i^2 = \sum_{i \neq j} \hat{\varepsilon}_i P_{ij}\hat{\varepsilon}_j.$$

This object is the numerator of the Sargan (1958) statistic with the own observation terms subtracted out. It has a jackknife form, in the sense that it is the sum of sums where the own observations have been deleted. If $\hat{\delta}$ were chosen to minimize this expression, it would be the JIVE2 estimator of Angrist, Imbens, and Krueger (1999).

One effect of removing the own observations is that $\sum_{i \neq j} \hat{\varepsilon}_i P_{ij}\hat{\varepsilon}_j$ would be mean zero if $\hat{\varepsilon}_i$ were replaced by ε_i In fact, $\sum_{i \neq j} \varepsilon_i P_{ij}\varepsilon_j$ has a martingale difference structure that leads to it being asymptotically normal as $K \longrightarrow \infty$, e.g. as in Lemma A2 of Chao et al. (2010). The denominator incorporates a heteroskedasticity consistent estimator of the variance of $\sum_{i \neq j} \varepsilon_i P_{ij}\varepsilon_j$. By dropping terms that have zero expectation, similarly to Chao et al. (2010), it follows that for $\sigma_i^2 = E[\varepsilon_i^2]$,

$$
\begin{aligned}
E[(\sum_{i \neq j} \varepsilon_i P_{ij}\varepsilon_j)^2] &= E[\sum_{i,j} \sum_{k \notin \{i,j\}} P_{ik}P_{jk}\varepsilon_i\varepsilon_j'\varepsilon_k^2 + \sum_{i \neq j} P_{ij}^2\varepsilon_i^2\varepsilon_j^2] \\
&= E[2\sum_{i \neq j} P_{ij}^2\varepsilon_i^2\varepsilon_j^2] = 2\sum_{i \neq j} P_{ij}^2\sigma_i^2\sigma_j^2,
\end{aligned}
$$

4

Similarly to White (1980) the variances are replaced by squared residuals to obtain $\hat{V}$. Also, 2 is replaced by $1/K$, and K is added to normalize the statistic to be chi-squared with K fixed and ε_i homoskedastic. Unfortunately, it does not appear possible to normalize the statistic to be chi-squared if there is heteroskedasticity when K is fixed.

3 Many Instrument Asymptotics

The asymptotic theory we give combines the many instrument asymptotics of Kunitomo (1980), Morimune (1983), and Bekker (1994) with the many weak instrument asymptotics of Chao and Swanson (2005), as further discussed in Chao et al. (2010). Some regularity conditions are important for this theory. Let $Z_i', \varepsilon_i, U_i'$, and Υ_i' denote the ith row of Z, ε, U, and Υ respectively.

Assumption 1: Z includes among its columns a vector of ones, $rank(Z) = K$, and there is a constant C such that $P_{ii} \leq C < 1$, $(i = 1, ..., n)$, $K \longrightarrow \infty$.

The restriction that $rank(Z) = K$ is a normalization that requires excluding redundant columns from Z. It can be verified in particular cases. For instance, when w_i is a continuously distributed scalar, $Z_i = p^K(w_i)$, and $p_{kK}(w) = w^{k-1}$, it can be shown that $Z'Z$ is nonsingular with probability one for $K < n$.[1] The condition $P_{ii} \leq C < 1$ implies that $K/n \leq C$, because $K/n = \sum_{i=1}^n P_{ii}/n \leq C$.

The next condition specifies that the reduced form Υ_i is a linear combination of a set of variables z_i having certain properties.

Assumption 2: $\Upsilon_i = S_n z_i/\sqrt{n}$ where $S_n = \tilde{S} \operatorname{diag}(\mu_{1n}, ..., \mu_{Gn})$ and $\tilde{S}$ is nonsingular. Also, for each j either $\mu_{jn} = \sqrt{n}$ or $\mu_{jn}/\sqrt{n} \longrightarrow 0$, $\mu_n = \min_{1 \leq j \leq G} \mu_{jn} \longrightarrow \infty$, and $\sqrt{K}/\mu_n^2 \longrightarrow 0$. Also, there is $C > 0$ such that $\|\sum_{i=1}^n z_i z_i'/n\| \leq C$ and $\lambda_{\min}(\sum_{i=1}^n z_i z_i'/n) \geq 1/C$, for n sufficiently large.

This condition is similar to Assumption 2 of Hansen, Hausman, and Newey (2008). It accommodates linear models where included instruments (e.g. a constant) have fixed reduced form coefficients and excluded instruments with coefficients that can shrink as the sample size grows, as

[1]The observations $w_1, ..., w_n$ are distinct with probability one and therefore, by $K < n$, cannot all be roots of a Kth degree polynomial. It follows that for any nonzero a there must be some i with $a'Z_i = a'p^K(w_i) \neq 0$, implying that $a'Z'Za > 0$.

further discussed in Hausman et al. (2010). The μ_n^2 can be thought of as a version of the concentration parameter, determining the convergence rate of estimators of δ_0, just as the concentration parameter does in other settings. For $\mu_n^2 = n$, the convergence rate will be $\sqrt{n}$, where Assumptions 1 and 2 permit K to grow as fast as the sample size, corresponding to a many instrument asymptotic approximation as in Kunitomo (1980), Morimune (1983), and Bekker (1994). For μ_n^2 growing slower than n, the convergence rate will be slower than $1/\sqrt{n}$, corresponding to the many weak asymptotics of Chao and Swanson (2005).

Assumption 3: There is a constant $C > 0$ such that $(\varepsilon_1, U_1), ..., (\varepsilon_n, U_n)$ are independent, with $E[\varepsilon_i] = 0$, $E[U_i] = 0$, $E[\varepsilon_i^2] < C$, $E[\|U_i\|^2] \leq C$, $Var((\varepsilon_i, U_i')') = diag(\Omega_i^*, 0)$, and $\lambda_{\min}(\sum_{i=1}^n \Omega_i^*/n) \geq 1/C$.

This assumption requires second conditional moments of disturbances to be bounded. It also imposes uniform nonsingularity of the variance of the reduced form disturbances, that is useful in the consistency proof.

Assumption 4: There is a π_{K_n} such that $\sum_{i=1}^n \|z_i - \pi_{K_n} Z_i\|^2 /n \longrightarrow 0$.

This condition and $P_{ii} \leq C < 1$ will imply that for a large enough sample

$$
\begin{aligned}
\sum_{i \neq j} P_{ij} \Upsilon_i \Upsilon_j'/n &= \Upsilon' P \Upsilon/n - \sum_{i=1}^n P_{ii} \Upsilon_i \Upsilon_i'/n = \sum_{i=1}^n (1 - P_{ii}) \Upsilon_i \Upsilon_i'/n - \Upsilon'(I - P)\Upsilon/n \\
&= \sum_{i=1}^n (1 - P_{ii}) \Upsilon_i \Upsilon_i'/n + o(1) \geq (1 - C) \sum_{i=1}^n \Upsilon_i \Upsilon_i'/n,
\end{aligned}
$$

so that the structural parameters are identified asymptotically. Also, Assumption 4 is not very restrictive because flexibility is allowed in the specification of Υ_i. If we simply make Υ_i the expectation of X_i given the instrumental variables then Assumption 4 holds automatically.

Assumption 5: There is a constant, $C > 0$, such that with probability one, $\sum_{i=1}^n \|z_i\|^4 /n^2 \longrightarrow 0$, $E[\varepsilon_i^4] \leq C$ and $E[\|U_i\|^4] \leq C$.

It simplifies the asymptotic theory to assume that certain objects converge and to allow for two cases of growth rates of K relative to μ_n^2. These conditions could be relaxed at the expense of further notation and detail, as in Chao et al.. Let $\sigma_i^2 = E[\varepsilon_i^2]$, $\gamma_n = \sum_{i=1}^n E[U_i \varepsilon_i]/ \sum_{i=1}^n \sigma_i^2$, $\tilde{U} = U - \varepsilon \gamma_n'$, having i^{th} row $\tilde{U}_i'$; and let $\tilde{\Omega}_i = E[\tilde{U}_i \tilde{U}_i']$.

Assumption 6: $\mu_n S_n^{-1} \longrightarrow S_0$ and either I) $K/\mu_n^2 \longrightarrow \alpha$ for finite α or; II) $K/\mu_n^2 \longrightarrow \infty$. Also, each of the following exists:

$$H_P = \lim_{n \to \infty} \sum_{i=1}^{n} (1 - P_{ii}) z_i z_i'/n, \Sigma_P = \lim_{n \to \infty} \sum_{i=1}^{n} (1 - P_{ii})^2 z_i z_i' \sigma_i^2/n,$$

$$\Psi = \lim_{n \to \infty} \sum_{i \neq j} P_{ij}^2 \left(\sigma_i^2 E[\tilde{U}_j \tilde{U}_j'] + E[\tilde{U}_i \varepsilon_i] E[\varepsilon_j \tilde{U}_j'] \right)/K.$$

The first result shows that the chi-square approximation is asymptotically correct when K grows with n.

THEOREM 1: *If Assumptions 1-6 are satisfied then* $\Pr(\hat{T} \geq q_{K-G}(1 - \alpha)) \longrightarrow \alpha$.

The next result shows asymptotic validity of the chi-squared approximation when K is fixed.

THEOREM 2: *If* $E[\varepsilon_i^2] = \sigma^2$, *$K$ is fixed,* $Z'Z/n \longrightarrow Q$ *nonsingular,* $Z'\Upsilon/n \longrightarrow D$ *with* $rank(D) = G$, $E[\varepsilon_i^4] \leq C$, $\|\Upsilon_i\| \leq C$, *and Assumption 3 is satisfied, then* $\Pr(\hat{T} \geq q_{K-G}(1-\alpha)) \longrightarrow \alpha$.

This test should have power against some forms of misspecification. Under misspecification, $\hat{V}$ will still be bounded and bounded away from zero. Also, for $\bar{\varepsilon}_i = E[y_i - X_i' plim(\hat{\delta})]$, the normalized numerator $\sum_{i \neq j} P_{ij} \hat{\varepsilon}_i \hat{\varepsilon}_j/\sqrt{K}$ will be centered at

$$\left(\bar{\varepsilon}' P \bar{\varepsilon} - \sum_i P_{ii} \bar{\varepsilon}_i^2 \right)/\sqrt{K}.$$

Assuming a linear combination of Z approximates $\bar{\varepsilon}$ this is close to

$$\sum_i \bar{\varepsilon}_i^2 (1 - P_{ii})/\sqrt{K}$$

This will increase at rate $n/\sqrt{K}$ by P_{ii} bounded away from one.

$\hat{T}$ provides a specification check for many instrument estimator $\hat{\delta}$. Note however that it may not be optimal as a test of the null hypothesis that $E[\bar{\varepsilon}_i] = 0$. The magnitude of the test statistic under the alternative grows more quickly when K grows more slowly. Thus, for higher power it would be better to use fewer instruments.

4   Monte Carlo Experiments

In this Monte Carlo simulation, we provide evidence concerning the finite sample behavior of $\hat{T}$. We consider two designs, the first of which is the same as in Hausman et al. (2010). The model for this design is

$$y_i = \delta_{10} + \delta_{20}x_{2i} + \varepsilon_i, x_{2i} = \pi z_{1i} + U_{2i} \tag{1}$$

where $z_{i1} \sim N(0,1)$ and $U_{2i} \sim N(0,1)$. The ith instrument observation is

$$Z_i' = (1, z_{1i}, z_{1i}^2, z_{1i}^3, z_{1i}^4, z_{1i}D_{i1}, ..., z_{1i}D_{i,K-5}), \tag{2}$$

where $D_{ik} \in \{0,1\}$, $\Pr(D_{ik} = 1) = 1/2$, and D_{ik} is independent of z_{i1}. Thus, the instruments consist of powers of a standard normal up to the fourth power plus interactions with dummy variables. Only z_{i1} affects the reduced form.

The structural disturbance, ε, is allowed to be heteroskedastic, being given by

$$\varepsilon = \rho U_2 + \sqrt{\frac{1-\rho^2}{\phi^2 + (0.86)^4}}(\phi v_1 + 0.86v_2), v_1 \sim N(0, z_1^2), v_2 \sim N(0, (0.86)^2),$$

where v_1 and v_2 are independent of U_2. This is a design that will lead to LIML being inconsistent with many instruments, as discussed in Hausman et al. (2010).

We report properties of the overidentifying test statistic $\hat{T}$ considered in this paper, the homoskedasticity based Sargan test statistic $\hat{\varepsilon}'P\hat{\varepsilon}/\hat{\varepsilon}'\hat{\varepsilon}$, and the Hansen GMM overidentifying test statistic that uses a heteroskedasticity consistent weighting matrix. Throughout we use HFUL as the estimator of δ. We consider $n = 800$ and $\rho = 0.3$ throughout, and let the number of instrumental variables be $K = 10, 30, 50$. Experiments with $n = 1600$ produced very similar results to those reported here. We choose π so that the concentration parameter is $n\pi^2 = 8$ and 32. We also choose ϕ so that the R-squared for the regression of ε^2 on the instruments is 0 or .2. Experiments with the R-squared for ε^2 is .1 gave similar results.

Tables 1 and 2 report rejection frequencies for nominal 5 percent and 1 percent tests. Table 1 is a homoskedastic case and Table 2 is heteroskedastic. We find that the actual rejection frequencies for the test $\hat{T}$ we propose to be close to their nominal values throughout these tables, including with and without heteroskedasticity and with few or many instrumental variables. We also find that the GMM overidentifying statistic is more sensitive to the number of overidentifying restrictions but in this design the Jstat is not very sensitive.

To see if these results were sensitive to the design we also tried a design where the heteroskedasticity was more closely related to the instruments. Note than in the previous design most of the instruments are uncorrelated with z_1 and that the heteroskedasticity depend entirely on z_1. The alternative design we considered was the same as above except that

$$\varepsilon = (\rho U_2 + v_2)(1 + \frac{K-5}{2})|z_1|, U_2 \sim N(0,1), v_2 \sim N(0, \sqrt{.91}).$$

This design has stronger heteroskedasticity that increases in strength with the number of instruments. Table 3 reports the results of this experiment. Here we find that, presumably due to the heteroskedasticity, the actual rejection frequencies for $Jstat$ are far from their nominal values. Also, the rejection frequencies for the GMM test are even more sensitive to the number of instruments than in the previous design. Remarkably, the test statistic proposed here continues to have rejection frequencies very close to nominal values.

To compare the power of the statistics we considered a homoskedastic design that allows for simple instrument misspecification. This design has a model, reduced form disturbance, and instrumental variables as in equations (1) and (2), but the structural disturbance is

$$\varepsilon_i = \rho_u U_{2i} + \rho_z z_{1i} + \rho_\eta \eta_i,$$

where η_i is standard normal and independent of (z_{1i}, U_{2i}) and $\rho_u^2 + \rho_z^2 + \rho_\eta^2 = 1$. We set the correlation ρ_u between structural and reduced form disturbances to be $\rho_u = .3$. We also allow the correlation ρ_z between structural disturbance and the instrument to vary between 0 and 1. The total number of instruments is taken to be $K = 10$. Also, we choose the concentration parameter $n\pi^2 = 8$ for sample sizes of $n = 400$ and $n = 800$.

For $n = 400$ the rejection frequencies under the null hypothesis are .47, .050, .041 respectively for $\hat{T}$, $Jstat$, and GMM. For $n = 800$ they are .052, .053, .049 respectively. The power curves (rejection frequencies) for the three tests as a function of ρ_z are plotted in Figures 1 and 2 for 400 and 800 observations respectively. The ranking of the power curves is similar to the ranking of sizes. Also the power curves are quite close. We do find that the power of $\hat{T}$ is quite similar to that of the other statistics. Thus, there seems to be little cost in terms of power under homoskedasticity to using a test statistic that is valid under heteroskedaticity with many instruments.

Table 1: $n = 800$, $\mathcal{R}^2_{\varepsilon^2|z_1^2} = 0$

μ^2	K	5% Jack	Jstat	GMM	1% Jack	Jstat	GMM
8	10	5.25	5.12	4.61	0.87	1.06	0.71
8	30	5.25	5.38	3.99	0.99	0.90	0.57
8	50	4.88	4.62	3.06	0.85	0.85	0.44
32	10	5.02	5.04	4.70	1.11	0.99	0.92
32	30	4.71	4.76	3.67	0.88	0.94	0.41
32	50	5.11	4.93	2.97	0.91	0.76	0.36

***Results based on 10,000 simulations.

Table 2: $n = 800$; $\mathcal{R}^2_{\varepsilon^2|z_1^2} = .2$

μ^2	K	.05 Jack	Jstat	GMM	.01 Jack	Jstat	GMM
8	10	5.68	5.79	5.34	1.11	1.20	0.91
8	30	5.40	4.98	4.09	0.85	0.89	0.49
8	50	4.79	4.79	3.05	0.86	0.85	0.32
32	10	5.05	5.15	4.69	0.97	1.00	0.71
32	30	4.65	4.84	3.77	0.93	0.89	0.59
32	50	4.80	4.80	3.12	0.83	0.62	0.30

***Results based on 10,000 simulations.

Table 3: $n = 800$;

μ^2	K	HFUL .05	JSTAT .05	GMM .05	HFUL .01	JSTAT .01	GMM .01
8	10	5.46	79.35	10.45	1.16	65.18	2.68
8	30	5.66	98.63	6.15	0.98	96.08	0.91
8	50	5.04	99.93	3.47	0.71	99.70	0.35
32	10	5.30	78.93	10.55	0.85	64.76	2.56
32	30	5.48	98.85	5.69	0.86	96.63	0.99
32	50	5.50	99.94	4.01	0.88	99.96	0.81

***Results based on 10,000 simulations.

5   Appendix A - Proofs of Theorems

We will define a number of abbreviations as well as some notation and. C denotes a generic positive constant that may be different in different uses and let M, CS, and T denote the Markov inequality, the Cauchy-Schwartz inequality, and the Triangle inequality respectively. Also, for random variables W_i, Y_i, and η_i, let $\bar{w}_i = E[W_i]$, $\tilde{W}_i = W_i - \bar{w}_i$, $\bar{y}_i = E[Y_i]$, $\tilde{Y}_i = Y_i - \bar{y}_i$, $\bar{\eta}_i = E[\eta_i]$, $\tilde{\eta}_i = \eta_i - \bar{\eta}_i$, $\bar{y} = (\bar{y}_1, ..., \bar{y}_n)'$, $\bar{w} = (\bar{w}_1, ..., \bar{w}_n)'$,

$$
\begin{aligned}
\bar{\mu}_W &= \max_{1 \leq i \leq n} |\bar{w}_i|, \ \bar{\mu}_Y = \max_{1 \leq i \leq n} |\bar{y}_i|, \ \bar{\mu}_\eta = \max_{1 \leq i \leq n} |\bar{\eta}_i|, \\
\bar{\sigma}_W^2 &= \max_{i \leq n} Var[W_i], \bar{\sigma}_Y^2 = \max_{i \leq n} Var[Y_i], \ \bar{\sigma}_\eta^2 = \max_{i \leq n} Var[\eta_i];
\end{aligned}
$$

The following Lemmas are special cases of results in Chao et al. (2010) but are given here for exposition:

LEMMA A1: *Suppose that the following conditions hold: i) P is a symmetric, idempotent matrix with $rank(P) = K$, $P_{ii} \leq C < 1$; ii) $(W_{1n}, U_1, \varepsilon_1)$, ..., $(W_{nn}, U_n, \varepsilon_n)$ are independent and $D_n = \sum_{i=1}^n E[W_{in}W_{in}']$ satisfies $\|D_n\| \leq C$; iii) $E[W_{in}'] = 0$, $E[U_i] = 0$, $E[\varepsilon_i] = 0$ and there exists a constant C such that $E[\|U_i\|^4] \leq C$, $E[\varepsilon_i^4] \leq C$; iv) $\sum_{i=1}^n E\left[\|W_{in}\|^4\right] \longrightarrow 0$; v) $K \longrightarrow \infty$ as $n \to \infty$. Then for*

$$
\bar{\Sigma}_n \overset{def}{=} \sum_{i \neq j} P_{ij}^2 \left(E[U_i U_i']E[\varepsilon_j^2] + E[U_i \varepsilon_i]E[\varepsilon_j U_j'] \right) / K
$$

and any sequences c_{1n} and c_{2n} depending on Z conformable vectors with $\|c_{1n}\| \leq C$, $\|c_{2n}\| \leq C$, $\Xi_n = c_{1n}'D_n c_{1n} + c_{2n}'\bar{\Sigma}_n c_{2n} > 1/C$, it follows that

$$
Y_n = \Xi_n^{-1/2}(c_{1n}'\sum_{i=1}^n W_{in} + c_{2n}'\sum_{i \neq j} U_i P_{ij}\varepsilon_j/\sqrt{K}) \overset{d}{\longrightarrow} N(0,1).
$$

Proof: This is Lemma A2 of Chao et al. (2010) when Z and Υ are not random. Q.E.D.

LEMMA A2: *If Assumptions 1-3 are satisfied then*

$$
S_n^{-1}\sum_{i \neq j} X_i P_{ij} X_j' S_n^{-1'} = O_p(1), S_n^{-1}\sum_{i \neq j} X_i P_{ij}\varepsilon_j = O_p(1 + \sqrt{K/r_n}).
$$

Proof: The second conclusion holds by Lemma A5 of Chao et al. (2010), and by that same result,

$$
S_n^{-1}\sum_{i \neq j} X_i P_{ij} X_j' S_n^{-1'} = \sum_{i \neq j} z_i P_{ij} z_j'/n + o_p(1).
$$

We also have

$$\sum_{i\neq j} z_i P_{ij} z_j'/n = z'Pz/n - \sum_i P_{ii} z_i z_i/n$$

and both $z'Pz/n \leq z'z/n$ and $\sum_i P_{ii} z_i z_i'/n \leq z'z/n$ are bounded, giving the first conclusion. Q.E.D.

LEMMA A3: *If $\hat{\delta} \longrightarrow \delta$, $E[\|X_i\|^2] \leq C$, $E[\varepsilon_i^4] \leq C$, $\varepsilon_1, ..., \varepsilon_n$ are mutually independent, and either $K \longrightarrow \infty$ or $\max_{i \leq n} P_{ii} \longrightarrow 0$ then*

$$\frac{\sum_{i\neq j} P_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2}{K} - \frac{\sum_{i\neq j} P_{ij}^2 \sigma_i^2 \sigma_j^2}{K} \stackrel{p}{\longrightarrow} 0.$$

Proof: Hence by $\hat{\delta} \stackrel{p}{\longrightarrow} \delta$ we have $\left\|\hat{\delta} - \delta\right\|^2 \leq \left\|\hat{\delta} - \delta\right\|$ with probability approaching one (w.p.a.1). Hence w.p.a.1, for $d_i = 3(1 + \|X_i\|^2)$,

$$\left|\hat{\varepsilon}_i^2 - \varepsilon_i^2\right| \leq 2\|X_i\| \left\|\hat{\delta} - \delta\right\| + \|X_i\|^2 \left\|\hat{\delta} - \delta\right\|^2 \leq d_i \left\|\hat{\delta} - \delta\right\|.$$

Also by $\sum_{i,j} P_{ij}^2 = \sum_i P_{ii} = K$,

$$E[\sum_{i\neq j} P_{ij}^2 d_i d_j]/K \leq C \sum_{i\neq j} P_{ij}^2/K \leq C, E[\sum_{i\neq j} P_{ij}^2 \varepsilon_i^2 d_j]/K \leq C.$$

Then by M,

$$\sum_{i\neq j} P_{ij}^2 d_i d_j/K = O_p(1), \sum_{i\neq j} P_{ij}^2 \varepsilon_i^2 d_j/K = O_p(1).$$

Therefore, for $\hat{V}_n = \sum_{i\neq j} P_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2/K, \tilde{V}_n = \sum_{i\neq j} P_{ij}^2 \varepsilon_i^2 \varepsilon_j^2/K$ we have

$$\begin{aligned}
\left|\hat{V}_n - \tilde{V}_n\right| &\leq \sum_{i\neq j} P_{ij}^2 \left|\hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2 - \varepsilon_i^2 \varepsilon_j^2\right|/K \\
&\leq \left\|\hat{\delta} - \delta\right\|^2 \sum_{i\neq j} P_{ij}^2 d_i d_j/K + 2\left\|\hat{\delta} - \delta\right\| \sum_{i\neq j} P_{ij}^2 \varepsilon_i^2 d_j/K \stackrel{p}{\longrightarrow} 0.
\end{aligned}$$

Let $V_n = \sum_{i\neq j} P_{ij}^2 \sigma_i^2 \sigma_j^2/K$ and $v_i = \varepsilon_i^2 - \sigma_i^2$ Note that by $P_{ij} = P_{ji}$,

$$\sum_{i\neq j} P_{ij}^2 \varepsilon_i^2 \varepsilon_j^2 - \sum_{i\neq j} P_{ij}^2 \sigma_i^2 \sigma_j^2 = 2\sum_{i\neq j} P_{ij}^2 v_i \sigma_j^2 + \sum_{i\neq j} P_{ij}^2 v_i v_j.$$

Note that $E[v_i^2] \leq E[\varepsilon_i^4] \leq C$, so we have

$$\begin{aligned}
E[(\sum_{i\neq j} P_{ij}^2 v_i \sigma_j^2/K)^2] &= K^{-2} \sum_i \sum_{j\neq i} \sum_{k\neq i} P_{ij}^2 P_{ik}^2 E[v_i^2] \sigma_j^2 \sigma_k^2 \\
&\leq CK^{-2} \sum_i \sum_j P_{ij}^2 \sum_k P_{ik}^2 = CK^{-2} \sum_i P_{ii}^2 \\
&\leq CK^{-1} \max_{i \leq n} P_{ii} \sum_i P_{ii}/K \leq CK^{-1} \max_{i \leq n} P_{ii} \longrightarrow 0.
\end{aligned}$$

Also, by CS, $\max_{i,j \le n} P_{ij}^2 \le \max_{i \le n} P_{ii}^2$, so that

$$
\begin{aligned}
E[(\sum_{i \ne j} P_{ij}^2 v_i v_j / K)^2] &= 2K^{-2} \sum_{i \ne j} P_{ij}^4 E[v_i^2] E[v_j^2] \le CK^{-2} \sum_{i,j} P_{ij}^4 \\
&\le CK^{-1} \max_{i \le n} P_{ii}^2 \sum_{i,j} P_{ij}^2 / K = CK^{-1} \max_{i \le n} P_{ii}^2 \longrightarrow 0.
\end{aligned}
$$

Then by T and M we have $\tilde{V}_n - V_n \overset{p}{\longrightarrow} 0$. The conclusion then follows by T. Q.E.D.

PROOF OF THEOREM 1: Note that

$$
\begin{aligned}
\frac{\sum_{i \ne j}^n \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\sqrt{K}} &= \sum_{i \ne j} \left[\varepsilon_i - X_i'(\hat{\delta} - \delta) \right] P_{ij} \left[\varepsilon_j - X_j'(\hat{\delta} - \delta) \right] / \sqrt{K} \\
&= \frac{\sum_{i \ne j} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{K}} + (\hat{\delta} - \delta)' S_n \left[S_n^{-1} \sum_{i \ne j} X_i P_{ij} X_j' S_n^{-1\prime} \right] S_n'(\hat{\delta} - \delta) / \sqrt{K} \\
&\quad + 2(\hat{\delta} - \delta)' S_n \left[S_n^{-1} \sum_{i \ne j} X_i P_{ij} \varepsilon_j \right] / \sqrt{K}.
\end{aligned}
$$

If $K/\mu_n^2 \longrightarrow \alpha < \infty$ (case I of Assumption 6) then by Theorem 2 of Hausman et al. (2010) we have $S_n'(\hat{\delta} - \delta) = O_p(1)$. Then by Lemma A2 we have

$$
\frac{\sum_{i \ne j}^n \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\sqrt{K}} = \frac{\sum_{i \ne j} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{K}} + o_p(1). \tag{3}
$$

If $K/\mu_n^2 \longrightarrow \infty$ (case II of Assumption 6) then by Theorem 2 of Hausman et al. (2010), $(\mu_n/\sqrt{K}) S_n'(\hat{\delta} - \delta_0) = O_p(1)$, so that by $\sqrt{K}/\mu_n^2 \longrightarrow 0$,

$$
(\hat{\delta} - \delta)' S_n \left[S_n^{-1} \sum_{i \ne j} X_i P_{ij} X_j' S_n^{-1\prime} \right] S_n'(\hat{\delta} - \delta) / \sqrt{K} = O_p(1) \left(K/\mu_n^2 \right) / \sqrt{K} = o_p(1),
$$

$$
\begin{aligned}
(\hat{\delta} - \delta)' S_n \left[S_n^{-1} \sum_{i \ne j} X_i P_{ij} \varepsilon_j \right] / \sqrt{K} &= O_p(1)(\sqrt{K}/\mu_n) O_p(1 + \sqrt{K}/\mu_n) / \sqrt{K} \\
&= O_p(1/\mu_n + \sqrt{K}/\mu_n^2) = o_p(1).
\end{aligned}
$$

Therefore, eq. (3) is also satisfied when $K/\mu_n^2 \longrightarrow \infty$.

Next, note that $\sigma_i^2 \ge C$ by Assumption 3 and $P_{ii} \le C < 1$ by Assumption 1, so that

$$
V_n = \frac{\sum_{i \ne j} \sigma_i^2 P_{ij}^2 \sigma_j^2}{K} > C \left(\frac{\sum_{i,j} P_{ij}^2}{K} - \frac{\sum_i P_{ii}^2}{K} \right) = C \frac{\sum_i P_{ii}(1 - P_{ii})}{K} > C > 0.
$$

Also, $E[\varepsilon_i^4] \leq C$ and as shown above, $E[\sum_{i \neq j}(\varepsilon_i P_{ij}\varepsilon_j)^2] = 2KV_n$. Now apply Lemma A1 with $W_{in} = 0$, $c_{1n} = 0$, and $c_{2n} = 1$. It follows by the conclusion of Lemma A1 that

$$\frac{\sum_{i \neq j} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{2KV_n}} \xrightarrow{d} N(0,1).$$

Next, by Theorem 1 of Hausman et al. (2010) we have $\hat{\delta} \xrightarrow{p} \delta$, so that by Lemma A3, $\hat{V}_n - V_n \xrightarrow{p} 0$. Then by V_n bounded and bounded away from zero, $\sqrt{V_n/\hat{V}_n} \xrightarrow{p} 1$. Therefore by the Slutzky theorem,

$$\frac{\sum_{i \neq j}^{n} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\sqrt{2K\hat{V}_n}} = \frac{\sum_{i \neq j}^{n} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{2K\hat{V}_n}} + \frac{o_p(1)}{\sqrt{2\hat{V}_n}} = \sqrt{\frac{V_n}{\hat{V}_n}} \frac{\sum_{i \neq j}^{n} \varepsilon_i P_{ij} \varepsilon_j}{\sqrt{2KV_n}} + o_p(1) \xrightarrow{d} N(0,1).$$

Next, note that $\hat{T} \geq q_{K-G}(1-\alpha)$ if and only if

$$\frac{\sum_{i \neq j}^{n} \hat{\varepsilon}_i P_{ij} \hat{\varepsilon}_j}{\sqrt{2K\hat{V}_n}} \geq \frac{q_{K-G}(1-\alpha) - K}{\sqrt{2K}}.$$

It is known that as $K \longrightarrow \infty$, $[q_{K-G}(1-\alpha) - (K-G)]/\sqrt{2(K-G)} \longrightarrow q(1-\alpha)$, where $q(1-\alpha)$ is the $1-\alpha$ quantile of the standard normal distribution. Also, we have

$$= \sqrt{\frac{K-G}{K}} \left(\frac{q_{K-G}(1-\alpha) - (K-G)}{\sqrt{2(K-G)}} \right) - \frac{G}{\sqrt{2K}} \longrightarrow q(1-\alpha).$$

The conclusion now follows. Q.E.D.

PROOF OF THEOREM 2: It follows in the usual way from the conditions that

$$\sqrt{n} \left(\hat{\delta} - \delta_0 \right) \xrightarrow{d} N(0, \sigma^2(D'Q^{-1}D)^{-1}).$$

In addition, it is straightforward to show that $Z'Z/n \longrightarrow Q$ nonsingular implies that $\max_{i \leq n} P_{ii} \longrightarrow 0$; e.g. see McFadden (1982). Furthermore, note for $d_i = 3(1 + \|X_i\|^2)$ from the proof of Lemma A3 that

$$E[\sum_i P_{ii}d_i] \leq \sum_i P_{ii}E[d_i] \leq C,$$

so $\sum_i P_{ii}d_i = O_p(1)$. Then similarly to the proof of Lemma A3, by $P_{ii} \geq 0$,

$$\left| \sum_i P_{ii}(\hat{\varepsilon}_i^2 - \varepsilon_i^2) \right| \leq \sum_i P_{ii} \left| \hat{\varepsilon}_i^2 - \varepsilon_i^2 \right| \leq \sum_i P_{ii}d_i \left\| \hat{\delta} - \delta \right\| = O_p(1)o_p(1) \xrightarrow{p} 0.$$

Also, we have

$$
\begin{aligned}
E[(\sum_i P_{ii}\varepsilon_i^2 - K\sigma^2)^2] &= E[(\sum_i P_{ii}\{\varepsilon_i^2 - \sigma^2\})^2] = \sum_i P_{ii}^2 Var(\varepsilon_i^4) \\
&\leq C \max_{i \leq n} P_{ii} \sum_i P_{ii} \longrightarrow 0.
\end{aligned}
$$

Then by the Markov and Triangle inequalities,

$$
\sum_i P_{ii}\hat{\varepsilon}_i^2 \xrightarrow{p} K\sigma^2.
$$

Also, since (as just shown) $\sum_i P_{ii}^2 \longrightarrow 0$ it follows by Lemma A3 and $\sigma_i^2 = \sigma^2$ that

$$
\hat{V}_n - \sigma^4 = \frac{\sum_{i\neq j} P_{ij}^2 \hat{\varepsilon}_i^2 \hat{\varepsilon}_j^2}{K} - \sigma^4 \frac{\sum_{i\neq j} P_{ij}^2}{K} - \sigma^4 \frac{\sum_i P_{ii}^2}{K} = o_p(1) + o(1) \xrightarrow{p} 0.
$$

Therefore,

$$
\hat{T} = \frac{\sigma^2}{\sqrt{\hat{V}}}\frac{\hat{\varepsilon}'P\hat{\varepsilon}}{\sigma^2} + K - \frac{\sum_{i=1}^n P_{ii}\hat{\varepsilon}_i^2}{\sqrt{\hat{V}}} = [1 + o_p(1)]\frac{\hat{\varepsilon}'P\hat{\varepsilon}}{\sigma^2} + o_p(1).
$$

It follows by standard arguments that $\hat{\varepsilon}'P\hat{\varepsilon}/\sigma^2 \xrightarrow{d} \chi^2(K - G)$, so the conclusion follows by the Slutsky Lemma. Q.E.D.

6 References

Anatolyev, S. and Gospodinov, N. (2009). Specification testing in models with many instruments, Econometric Theory 27, 427-441.

Angrist, J.D., G.W. Imbens, and A. Krueger (1999) Jackknife instrumental variables estimation. Journal of Applied Econometrics 14, 57-67.

Bekker, P.A. (1994) Alternative approximations to the distributions of instrumental variable estimators. Econometrica 62, 657-681.

Chao, J.C. and N.R. Swanson (2005) Consistent estimation with a large number of weak instruments. Econometrica 73, 1673-1692.

Chao, J.C., N.R. Swanson, J.A. Hausman, W.K. Newey, and T. Woutersen.(2010): Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments, working paper.

Fuller, W.A. (1977) Some properties of a modification of the limited information estimator," Econometrica 45, 939-954.

Hansen, L.P. (1982): Large sample properties of generalized method of moments estimators, Econometrica 50, 1029-1054.

Hansen, C., J.A. Hausman, and W.K. Newey (2008) Estimation with many instrumental variables. Journal of Business and Economic Statistics 26, 398-422..

Hausman, J.A., W.K. Newey, T.M. Woutersen, J. Chao, and N.R. Swanson (2010) IV estimation with heteroskedasticity and many instruments. Working Paper, MIT.

Kunitomo, N. (1980) Asymptotic expansions of distributions of estimators in a linear functional relationship and simultaneous equations. Journal of the American Statistical Association 75, 693-700.

Lee, Y. and R. Okui (2010) A specification test for instrumental variables regression with many instrument, working paper.

McFadden, D. (1982) Large sample properties of least squares, Lecture notes, MIT.

Morimune, K. (1983) Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. Econometrica 51, 821-841.

Sargan, J.D. (1958), The estimation of economic relationships using instrumental variables, Econometrica 26, 393–415.

White, H. (1980) A heteroskedasticity-consistent variance matrix estimator and a direct test for heteroskedasticity, Econometrica 48, 817-38.
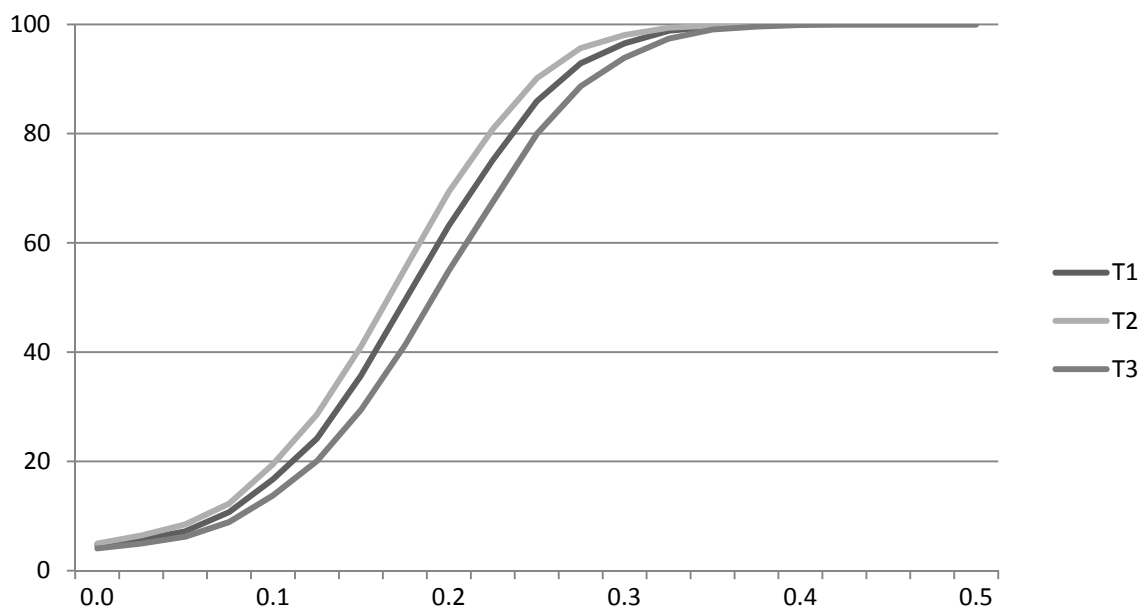
Figure 1, Power of Tests, n = 400

Figure 2, Power of Tests, n = 800