Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 10

11-1-2004

# Size and Power of the RESET Test as Applied to Systems of Equations: A Bootstrap Approach

Ghazi Shukur
*Jönköping and Växjö Universities, Sweden*, ghazi.shukur@ehv.vxu.se

Panagiotis Mantalos
*Lund University, Sweden*, Panagiotis.Mantalos@stat.lu.se

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Shukur, Ghazi and Mantalos, Panagiotis (2004) "Size and Power of the RESET Test as Applied to Systems of Equations: A Bootstrap Approach," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 2 , Article 10.
DOI: 10.22237/jmasm/1099267800

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 370-385
Copyright © 2004 JMASM, Inc.
1538 – 9472/04/$95.00

# Size and Power of the RESET Test as Applied to Systems of Equations: A Bootstrap Approach

Ghazi Shukur
Departments of Economics and Statistics
Jönköping and Växjö Universities, Sweden

Panagiotis Mantalos
Department of Statistics
Lund University, Sweden

The size and power of various generalization of the RESET test for functional misspecification are investigated, using the "Bootsrap critical values", in systems ranging from one to ten equations. The properties of 8 versions of the test are studied using Monte Carlo methods. The results are then compared with another study of Shukur and Edgerton (2002), in which they used the asymptotic critical values instead and found that in general only one version of the tests works well regarding size properties. In our study, when applying the bootstrap critical values, we find that all the tests exhibits correct size even in large systems. The power of the test is low, however, when the number of equations grows and the correlation between the omitted variables and the RESET proxies is small.

Key words: RESET, Systems of Equations, Bootstrap

## Introduction

The RESET test proposed by Ramsey (1969) is a general misspecification test, which is designed to detect both omitted variables and inappropriate functional form. The RESET test is based on the Lagrange Multiplier principle and usually performed using the critical values of the *F*-distribution. While most authors (e.g., Ramsey and Gilbert (1972); Thursby and Schmidt (1977)) have studied the properties of the RESET tests in single equation situations, Shukur and Edgerton (2002), in what follows referred to as SE, examine the small sample properties of various generalization of the RESET test in an environment of equation systems.

The latter authors used Monte Carlo methods to study the properties of eight different versions of the RESET test in systems ranging from one to ten equations. By using the critical values of the *F*-distribution, the authors find that the Rao's *F*-test exhibits the best performance as regards correct size, while, by using the critical values of the $\chi^2$ - distribution, they find that the commonly used LRT (uncorrected for degrees-of-freedom), and LM and Wald tests (both corrected and uncorrected) behave badly even in a single equation situation. SE also find that the power of the test decreases when the number of equations grows and the correlations between the omitted variables and the RESET proxies are small.

Note that by using the critical values of the $\chi^2$ - distribution, the LRT, LM and Wald tests are strictly valid only asymptotically. Therefore, making inferences on the basis of them can be a risky undertaking. Some authors, e.g., Kivit (1986), have used Monte Carlo methods to compare different LM, Wald and LR alternatives for single equation models. When testing for autocorrelation they have shown that the standard *F*-test, which is also only valid asymptotically, is in general more accurate as regards size properties.

However, an effective misspecification test should have correct significance levels under the null hypothesis, irrespective of the values of the regression parameters and other distributional parameters. It should also have

Ghazi Shukur is Professor of statistics and econometrics at two universities in Sweden (Jönköping University and Växjö University). Email: ghazi.shukur@ehv.vxu.se. Panagiotis Mantalos in an Associate Professor in statistics at Lund University, Sweden. Email address: panagiotis.mantalos@stat.lu.se.

reasonable power against the class of alternative specifications under investigation, but low power against other alternatives.

The purpose of this article is to improve the critical values of the test statistics by employing bootstrap technique, so that the size of the test approaches its nominal value. Horowitz (1994) and Mantalos and Shukur (1998) recommended this approach. Given the bootstrap critical values, analyzed here is the size and power of a different generalization of the systemwise RESET test, followed by a comparison with results found by SE.

Model Specification

The regression model investigated is the same model as in SE and consists of *n* linear stochastic equations given by

$$\mathbf{Y}_t = \mathbf{X}_t \mathrm{B} + \varepsilon_t \,, \tag{1}$$

where $\boldsymbol{Y}_t$ and $\varepsilon_t$ are $(1 \times n)$ vectors of endogenous variables and disturbances respectively, $\boldsymbol{X}_t$ is a $(1 \times m)$ matrix of exogenous variables, $\boldsymbol{B}$ is a $(m \times n)$ matrix of parameters, and $t = 1,\ldots,T$. The data matrices $\boldsymbol{Y}$ and $\boldsymbol{X}$ are $(T \times n)$ and $(T \times m)$ respectively. The null hypothesis of correct specification implies that the error terms will be independently and identically distributed conditional on the exogenous variables, and in many cases a normal distribution is also assumed,

$$\varepsilon_t | \mathbf{X}_t \sim N(0,\, \Sigma_{\varepsilon}) \cdot \tag{2}$$

The hypothesis of correct functional form is equivalent to assuming that the disturbances have zero conditional mean, $H_0 : E(\varepsilon_t | \mathbf{X}_t) = 0$.

The class of alternative hypotheses to this null hypothesis is very general; omitted variables and incorrect functional form will obviously be members of the class, but so to will endogeneity of the $\boldsymbol{X}$ variables.

The alternative hypothesis is specified through the following model:

$$\mathbf{Y}_t = \mathbf{X}_t \mathrm{B} + \mathbf{Z}_t \Gamma + \varepsilon_t \cdot \tag{3}$$

$\boldsymbol{Z}$ is in general unknown, and the tests that we will investigate use a proxy $\hat{\boldsymbol{Z}}$. The following regression is estimated instead of (3),

$$\mathbf{Y}_t = \mathbf{X}_t \mathrm{B} + \hat{\mathbf{Z}}_t \Gamma^* + \delta_t \cdot \tag{4}$$

If the null hypothesis is correct, then $\Gamma = \Gamma^* = 0$ whatever the choice of $\hat{\boldsymbol{Z}}$. If the hypothesis is incorrect, then the choice of $\hat{\boldsymbol{Z}}$ will obviously affect the power of any test based on (4). The greater the correlation between $\hat{\boldsymbol{Z}}$ and the non-linear part of the true conditional mean of $\boldsymbol{Y}$, then, in general, the greater the power will be. If we suspect certain variables to have been omitted, then using these variables as $\hat{\boldsymbol{Z}}$ will obviously be most appropriate.

Ramsey (1969) proposed approximating the unknown conditional expectation of $\boldsymbol{Y}$ by using a Taylor expansion around the conditional expectation *under the null hypothesis*, that is $\boldsymbol{X\beta}$ (Ramsey considered a single equation, and $\boldsymbol{\beta}$ was thus a vector). Because the parameters are unknown, this was in turn approximated using $\hat{\mathbf{Y}} = \mathbf{X}\hat{B}$, where $\hat{B}$ was the OLS parameter estimate from the single equation version of (1). This is the RESET test procedure.

Define a systemwise version of the RESET test. Following common terminology of double regression tests, refer to equation (1) as the *primary regression*. The first stage of the RESET test is performed by calculating the least squares' predictions from the primary regression, i.e., $\hat{\mathbf{Y}} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$. These predictions are then used in the following *auxiliary regression*,

$$\mathbf{Y}_t = \mathbf{X}_t \mathrm{B} + \hat{\mathbf{Y}}_t^2 \Gamma_1^* + \hat{\mathbf{Y}}_t^3 \Gamma_2^* \ldots + \hat{\mathbf{Y}}_t^{G+1} \Gamma_G^* + \delta_t \,, \tag{5}$$

where the $(t,\ i)$:*th* elements of the power matrices are given by $[\hat{\mathbf{Y}}^j]_{ti} = \hat{y}_{ti}^{\,j}$. The RESET test is now performed by testing the hypothesis $\Gamma_1^* = \cdots = \Gamma_G^* = 0$.

The practical implementation of the RESET test now depends on two factors. Firstly

it must decided how many power matrices to include in the auxiliary regression (i.e., determine *G*). Secondly, it must be decided which test method to use. We concentrate on the second question, and set *G* = 1 throughout.

Denote by $\hat{\delta}U$ the $(T \times n)$ matrix of estimated residuals from the *unrestricted* regression (5), and by $\hat{\delta}R$ the equivalent matrix of residuals from the *restricted* regression with $H_0'$ imposed. The matrix of cross-products of these residuals will be defined as $\mathbf{S}_U = \hat{\delta}U'\hat{\delta}U$ and $\mathbf{S}_R = \hat{\delta}R'\hat{\delta}R$ respectively. Bewley (1986, Chapter 4) showed that the Wald, Likelihood Ratio and Lagrange Multiplier test statistics are given by

$$W = T(\operatorname{tr}\mathbf{S}_U^{-1}\mathbf{S}_R - n), \tag{6}$$

$$LR = T\ln U, \text{ and} \tag{7}$$

$$LM = T(n - \operatorname{tr}\mathbf{S}_R^{-1}\mathbf{S}_U), \tag{8}$$

where $U = \det\mathbf{S}_R / \det\mathbf{S}_U$. The above statistics are all asymptotically $\chi^2(p)$ distributed under the null hypothesis, where $p = Gn^2$ is the number of restrictions imposed by the null hypothesis. It is well known, however, that this asymptotic result becomes less and less accurate in small samples as the number of equations grows, see for example Laitinen (1978). A simple small sample correction is to replace *T* by $\Delta = T - (m + Gn)$, the degrees of freedom in the equations of the auxiliary regression (4). The *corrected* statistics are thus given by $WC = (\Delta/T)W$, $LRC = (\Delta/T)LR$ and $LMC = (\Delta/T)LM$, which have the same asymptotic distribution as given above.

Another more sophisticated approximation is that given by theorem 8.6.2 in Anderson (1958, p. 321). This uses an Edgeworth expansion, and if we choose the simplest form (which is accurate to the order $T^{-2}$) this corrected LR statistic is given by

$$LRE = \Delta_E \ln U, \tag{9}$$

where $\Delta_E = \Delta + \tfrac{1}{2}[n(G-1) - 1]$. Note that when $G = 1$, the difference between *LRC* and *LRE* is merely that the numerator in the correction is $\Delta$ in the first case and $\Delta - \tfrac{1}{2}$ in the second.

A final approximation is that given by Rao (1973, p. 556), namely

$$RAO = (q/p)(U^{1/s} - 1), \tag{10}$$

where *p* and $\Delta_E$ are defined above, $r = p/2 - 1$, $q = \Delta_E s - r$, and

$$s = \sqrt{\frac{p^2 - 4}{n^2(G^2 + 1) - 5}}. \tag{11}$$

*RAO* is approximately distributed as $F(p,q)$ under the null hypothesis, and reduces to the standard *F* statistic when $n = 1$.

Factors that Affect the Small Sample Properties of the RESET Test

A number of factors obviously can affect the size of the RESET tests, SE have investigated these factors systematically, and we therefore follow their line of investigation. The number of equations (*n*), the sample size (*T*), degrees of freedom ($\Delta$) and the order of the restrictions (*G)* are four such factors. The power of the tests will also be affected by the size and form of $Z_t\,\Gamma$ in (3). In this paper we will also study the consequences of varying *n* and $\Delta$, while *T* is chosen so as to give compatible values of $\Delta$ for different models ($T = \Delta + m + Gn$). We will also mainly concentrate on the case where *G* = 1.

A number of other factors can also affect the properties of the RESET tests, for example the distributions of $\boldsymbol{X_t}$, and $\boldsymbol{\varepsilon_t}$, and the values of $\boldsymbol{B}$. In the rest of this section we will consider these factors in some more detail. In this paper, we consider only stochastic exogenous variables $\boldsymbol{X_t}$ and although SE find that serial dependence in *x* has no practicable

effect on either the size or power of the RESET tests, we will allow autocorrelation in the exogenous variables in our study. The following generating processes are used,

$$x_{tj} = \alpha x_{t-1,j} + \nu_{tj} \ , \quad j = 1, \ldots, \text{m-1}$$
$$\text{and } t = 1, \ldots, \text{T} \qquad (12)$$

where and $\nu_t$ is a multivariate normal white noise process with covariance matrix $\Sigma_\nu$. In our Monte Carlo study we have included a constant term among the exogenous variables, so that (12) has only been applied to the remaining $m - 1$ variables.

The power (but not the size) of the tests will also be affected by $\mathbf{Z}_t \Gamma$ in (3). Intuitively, the power of the test ought to increase with an increase in the omitted portion of the regression. That is to say, an increase in the absolute value of $\Gamma$ should imply an increase in the seriousness of the misspecification caused by using (1) instead of (3). Accordingly, we would expect the power of the RESET test to increase with $\Gamma$. The problem is to decide how large a value of $\Gamma$ is needed to constitute "serious" misspecification.

SE found that a good measure of misspecification is given by the relative increase of goodness-of-fit, achieved by going from the incorrect model under the null (1) to the correct model under the alternative (3), i.e.,

$$R_D^2 = \frac{R_1^2 - R_0^2}{1 - R_0^2} \qquad (13)$$

where $R_0^2$ and $R_1^2$ are the theoretical $R^2$ measures from the null and alternative models respectively. The reasoning behind this choice of misspecification measure, and the relationships that exist between goodness-of-fit and the other parameters of the model, are explained in the Appendix of their paper. An advantage of using $R_D^2$ as a measure of misspecification is that it is bounded between zero (no misspecification) and one (a perfect alternative).

The power of the test will also depend on the joint distribution of the included and omitted variables. If this distribution is joint normal, then the regression of the omitted variables on the included variables is exactly linear, and no loss of fit will occur through the exclusion of the omitted variables.

The RESET test will have zero power in such circumstances, even though the parameter estimates will be biased, unless the omitted variable is also uncorrelated with the included variables. If the omitted variables are non-normal, then their conditional means can be non-linear in the included variables, and the RESET test can have power. The strength of the power, however, might depend on the correlation between the omitted variable (Z) and the proxy variables ($\hat{Y}^j$) used in the auxiliary regression (4). In this paper, and as in SE, we concentrate on an omitted variable which is the square of one of the (normally distributed) included variables.

Bootstrap-hypothesis testing, critical values.

Two aspects are of primary importance when the properties of a test procedure are investigated. Firstly, determine if the *actual size* of the test (i.e., the probability of rejecting the null when true) is close to the *nominal size* (used to calculate the critical values). Given that actual size is a reasonable approximation to the nominal size, then investigate the *actual power* of the test (i.e., the probability of rejecting the null when false) for a number of different alternative hypotheses. When comparing different tests, therefore, those in which (a) actual size lies close to the nominal size and, given that (a) holds, (b) have greatest power are preferred. In most cases, however, the distributions of the test statistics used are known only asymptotically.

As a result, the tests do not have the correct size and inferential comparisons and judgments based on them might be misleading. However, by using bootstrap technique it is possible to improve the critical values so that the true size of the test approaches its nominal value.

In the regression model (1), the null hypothesis of correct specification implies that the error term $\varepsilon_t$ will be independently and identically distributed, conditional on the exogenous variables. The most convenient way

to apply bootstrap, here, is to resample the $\varepsilon_t$. Since the errors are not observable and the usual solution is then to use the calculated Least Squares (OLS) residuals instead. A direct residual resampling gives:

$$\mathbf{Y}_{t^*}^* = \mathbf{X}_t \hat{B}_{ols} + \varepsilon_t^*, \qquad (1a)$$

where $\varepsilon_t^*$ are i.i.d observations $\varepsilon_1^*, \ldots, \varepsilon_T^*$, drawn from the empirical distribution ($\hat{F}_\varepsilon$) of the LS residuals. This method is called the *bootstrap based on residuals*, abbreviated as RB, proposed by Efron (1979). Note that, in what follows, all bootstrap statistics will marked by an asterisk $^{(*)}$. An important assumption for the RB is that $\varepsilon_t$ are i.i.d, but even if this assumption holds, the empirical distribution $\hat{F}_\varepsilon$ is not based on exactly i.i.d data, namely observed residuals $\hat{\varepsilon}_t$. Therefore the following adjustments are necessary.

First, subtract the sample mean of the OLS residuals from the residuals: ($\hat{\varepsilon}_i - \bar{\varepsilon}$)

where $\bar{\varepsilon} = T^{-1}\sum_{i=1}^{T} \hat{\varepsilon}_i$ i = 1, ... , T. Thus, $E_*(\varepsilon_t^*) = 0$ for all t. And $E_*(\hat{\mathbf{B}}^*{}_{\mathrm{OLS}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E_*(Y^*) = \hat{\mathbf{B}}_{\mathrm{OLS}}$, where $\hat{\mathbf{B}}^*_{\mathrm{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\ \mathbf{Y}^*$, and $\mathrm{Var}_*\hat{\mathbf{B}}^*_{\mathrm{OLS}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$,

where

$$\hat{\sigma}^2 = \mathrm{Var}_*(\varepsilon_t^*) = \mathrm{T}^{-1}\sum_{i=1}^{T} (\hat{\varepsilon}_t - \bar{\varepsilon})^2, \quad \mathrm{i} = 1, 2, \ldots, \mathrm{T}.$$

This bootstrap procedure produces consistent variance but is downward biased (Efron, 1982). To remove this negative bias, Efron (1982) suggested the bootstrap data to be drawn from the empirical distribution $\hat{F}_\varepsilon$ putting mass 1/T to the adjusted OLS residuals $(\hat{\varepsilon}_i - \bar{\varepsilon})/\sqrt{1-[m/T]}$, i = 1,...,T. This is called *the adjusted residual resampling* ARR.

The basic principle is to draw a number of bootstrap samples from the model under the null hypothesis. The bootstrap test statistic ($T_s^*$) can then be calculated by repeating this step $N_b$ number of times. Then, take the (1-α):*th* quintile of the bootstrap distribution of $T_s^*$ and get the α - level "bootstrap critical values" ($c_{t\alpha}^*$). Generally, the bootstrap procedure is summarized by the following steps:

(1) Estimate the test statistic as previously described, which is called ($T_s$).

(2) Use *the adjusted residual resampling* ARR, $(\hat{\varepsilon}_i - \bar{\varepsilon})/\sqrt{1-[m/T]}$ i = 1,...,T to draw i.i.d. data $\varepsilon_1^*, \ldots, \varepsilon_T^*$ and define: $\mathbf{Y}_{t^*}^* = \mathbf{X}_t \hat{B}_{ols} + \varepsilon_t^*$.

Then, calculate the test statistic ($T_s^*$) as described, i.e., by applying the RESET test procedure to the (1a) model. Repeating this step $N_b$ number of times and taking the (1-α):*th* quintile of the bootstrap distribution of $T_s^*$, we obtain the α - level "bootstrap critical values" ($c_{t\alpha}^*$), and finally, we then reject Ho if $T_s \geq c_{t\alpha}^*$. This is our bootstrap test approach to investigate the size and power of the various generalization of the systemwise RESET test.

Monte Carlo Experiment

In a Monte Carlo study, the estimated size is estimated by simply observing how many times the null is rejected in repeated samples under conditions where the null is true. By varying factors such as described in the previous section, a succession of estimated sizes under different conditions is obtained. In general, the closer an estimated size is to the nominal size the better the test. Most of the factors discussed earlier either have very small effect, or have no effect at all on the estimated size of the tests. To show the effect of the remaining factors on the performances of the tests, the estimated sizes of the tests are displayed in the tables.

As regards the *estimated power functions* of the tests, these have mainly been compared graphically. This has proved to be quite adequate, since those tests that give reasonable results as regard size usually differed very little regarding power.

The Monte Carlo experiment was performed by generating data according to (1), (2) and (12), estimating the auxiliary regression (5) and then calculating the test statistics, $T_s$, defined above.

Because the number of regressors in the auxiliary regression (5) is (m + n), we draw i.i.d. data $\varepsilon_1^*,\ldots,\varepsilon_T^*$ from the empirical distribution $\hat{F}_\varepsilon$ putting mass 1/t to the adjusted (LS) residuals $(\hat{\varepsilon}_i - \bar{\varepsilon}) / \sqrt{1-[(m+n)/T]}$, i = 1,...,T.

The bootstrap procedure described in the previous section is followed to obtain the α - level "bootstrap critical values" ($c_{t\alpha}^*$). The α = 0.05 level, for example, is the $T_{sN_b\,96}^*$ of the order test statistic: $T_{sN_b\,1}^* \le T_{sN_b\,2}^* \le \ldots \le T_{sN_b\,100}^*$.

A final consideration is the significance levels to be used when judging the properties of the tests. Theoretically, it is possible to construct the empirical distributions of the test statistics, and to compare these with the theoretical asymptotic results. In this study, the tests of the null hypothesis were carried out using nominal significance levels ($\pi_0$) of 1%, 5%, 10% and 20%. Hence, for the 1%, 5%, 10%, and 20% levels, the "bootstrap critical values" $c_{t\alpha}^* = T_{sN_b\,99}^*$, $c_{t\alpha}^* = T_{sN_b\,95}^*$, $c_{t\alpha}^* = T_{sN_b\,90}^*$ and $c_{t\alpha}^* = T_{sN_b\,80}^*$ were chosen, respectively. Then, reject Ho if $T_s \ge c_{t\alpha}^*$.

An approximate 95% confidence interval for the actual size ($\pi$) can be given as

$$\hat{\pi} \pm 2\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}}, \qquad (13)$$

where $\hat{\pi}$ is the estimated size and $N$ is the number of replications.

However, because the main interest is in the behavior of the distributions in the tails, only results using the conventional 5% significance level have been analyzed. A summary of the design can be found in Table 1 and 2, and in Table 3 approximate 95% confidence intervals for the actual size, based on (13) are presented. Letting the number of replications per model is 10,000, which by (13) seems to be sufficient when estimating size. Note again that SE's Monte Carlo design is followed, and a summary of the relationships between the various factors can be found in their article (in their Appendix).

Regarding the $N_b$ number of the bootstrap samples used to estimate bootstrap critical value, Horowitz (1994) used the value of $N_b$ = 100. However, it follows from Hall (1986) that the error in the size of a test using the "bootstrap critical values" is independent on the number of the bootstrap sample used to estimate $c_{t\alpha}^*$. $N_b$ = 500 in the current study. Increasing the number of the bootstrap samples beyond 500 has little effect on the results of the experiment and takes longer time.

The primary interest lies in the analysis of systemwise tests, and thus the number of equations to be estimated is of central importance. As the number of equations grows, the computation time becomes longer. A system with ten equations was selected as the largest model when studying the size of the tests. This represents a fairly large consumption model of the type that is used in, for example, agricultural economics. Medium size models are represented by five- and seven-equation systems, and two- and three-equation systems are typical of the small models used when separability is imposed.

Another important factor that affects the performance of tests is the number of observations. The number of degrees of freedom, Δ, was held constant between models of different sizes, because this allows a fair comparison. If the number of observations, *T*, were held constant then tests in models with a large number of equations would automatically perform more poorly, simply due to the reduced degrees of freedom (a new predictor is included for each equation in the system).

Table 1. Values of Factors Held Constant for Different Models

| Factor | Symbol | Design |
|---|---|---|
| Constant term | | 1 (size) or 0 (power) |
| Number of *X* variables (excl constant) | $n + 1$ | number of equations + 1 |
| Mean of *X* variables | $\boldsymbol{\mu}_X$ | $\mathbf{0}$ |
| Parameters of *X* variables | $\boldsymbol{B}$ | $\boldsymbol{E}$ |
| Distribution of *X* variables | | Normal |
| Covariance of *X* variables | $\boldsymbol{\Sigma}_X$ | $(1-\rho_X)\boldsymbol{I} + \rho_X\boldsymbol{E}$ |
| Properties of *X* in repeated samples | | Stochastic |
| Distribution of error terms | | Normal |
| Covariance of error terms | $\boldsymbol{\Sigma}_\varepsilon$ | $\sigma^2\boldsymbol{I}$ |

Table 2a. Values of Factors that Vary for Different Models - Size Calculations

| Factor | Symbol | Design | | | |
|---|---|---|---|---|---|
| Number of equations | $n$ | 1 | 2 | 3, 5, 7 | 10 |
| Degrees of freedom | $\Delta$ | 15, 25, 45, 75 | | | |
| Nominal size | $\pi_0$ | 1%, 5%, 10%, 20% | | | |
| Goodness-of-fit in null | $R_0^2$ | .1, .3, .5, .7, .9 | .3, .5, .7 | .3, .7 | .3 .7 |
| AR parameter for $X$ | $\alpha$ | 0, .5, .9 | 0, .5 | | |
| Correlation $(X_i,X_j)$ | $\rho_X$ | 0, .5, .9 | 0, .5 | | |

Table 2b. Values of Factors that Vary for Different Models - Power Calculations

| Factor | Symbol | Design |
|---|---|---|
| Number of equations | $n$ | 1, 2, 3, 5, 7, 10 |
| Degrees of freedom | $\Delta$ | 15, 25, 45, 75 |
| Nominal size | $\pi_0$ | 1%, 5%, 10%, 20% |
| Goodness-of-fit in null | $R_0^2$ | .3, .5, .7 |
| Relative difference in $R^2$ | $R_D^2$ | 0, .1, .2, .3, .4, .5, .6, .7, .8, .9 |
| AR parameter for $X$ | $\alpha$ | 0, .5 |
| Correlation $(\eta,z)$ | $\rho_{\eta z}$ | .1, .3, .5, .7, .9 |

$z$ is the omitted variable (the square of $x_1$) and $\eta$ is the square of the conditional expected value of $y$.

Table 3. Approximate 95% Confidence Intervals for Actual Size

| $\pi_0$ ∴ $N$ | 2000 | 10000 |
|---|---|---|
| 1% | ±0.44 | ±0.20 |
| 5% | ±0.97 | ±0.44 |
| 10% | ±1.34 | ±0.60 |
| 20% | ±1.79 | ±0.80 |

We have investigated samples typical for annual and quarterly consumption models, using degrees of freedom 15, 25, 45 and 75. This is equivalent to sample sizes of between 20 and 110 observations.

Various values of $R_0^2$ were chosen to represent different explanatory powers under the null with a greater variation in small models. The distribution of the exogenous variables was varied to account for a typical property of economic time series, i.e., that they are trended and/or autocorrelated. SE find that trending had no effect at all on the RESET tests, and it is therefore not considered here. The calculations were performed using GAUSS 3.2, and the results from different models were analysed using MSExcel 4.0.

When calculating the power functions of the tests we used different values of $R_D^2$ to indicate different degrees of misspecification in the model. Different values of $\rho_{\eta z}$ were used to illustrate different strengths in the relationship between the omitted variable and the proxy variable used in the auxiliary regression.

Analysis of the Size of the RESET Tests.

In this section, results are presented of the Monte Carlo experiment concerning the size of the RESET tests. When using the "bootstrap critical values", our primary results reveal that the LM and Wald tests get results identical to their corrected correspondents (i.e., LMC, and WC).

All the LR tests (including the RAO) lead to identical results. Moreover, for a single equation, we find that all the eight test methods yield the same results. Noticeable effects on the estimated size were not found, however, by varying the number of equations, degrees of freedom, autocorrelation in the exogenous variables, the collinearity between the exogenous variables, or the goodness-of-fit under the null hypothesis. These results agree with the results obtained by SE regarding the Rao test only.

The results from the two articles are now compared to show the differences between our findings. Our results are shown in Table 4, were the same goodness-of-fit ($R_0^2 = 0.7$) was used, multicollinearity ($\rho_X = 0.5$), and autocorrelation ($\alpha = 0.0$) in $X$ as in Table SE 4. Note that changing the factors we have held constant in these tables (i.e., goodness-of-fit, multicollinearity and autocorrelation in $X$) would not change the conclusions in any way. Some important results regarding the different variants of the RESET test are presented in Table SE 4. They found that the number of equations in the system (n) and the degrees of freedom ($\Delta$) have noticeable effect on the performances of the tests.

They also found that the RAO test was superior to all the other alternatives, with only one result (out of 30) lying slightly outside the 95% confidence interval, whereas the WALD and LRT tests performed extremely poorly.

When we use the "bootstrap critical values", the results show that all tests perform well, i.e. the superiority of the Rao test to the other tests disappears. The WALD/Wald-C tests perform slightly badly in small samples and large systems. The Rao/LR and LM tests are shown to perform satisfactorily in all situations. Note that in our study, i.e. when we use the "bootstrap critical values", all the tests have identical results for single equation models.

Analysis of the Power of the RESET tests

In this section, the most interesting results of our Monte Carlo experiment regarding the power of the various versions of the RESET test are discussed. The power of different versions of the RESET test was analyzed, using the "bootstrap critical values", in systems ranging from one to ten equations. The power function was estimated by calculating the rejection frequencies in 2,000 replications using different values of the relative differences in goodness-of-fit, $R_D^2$.

Even if a correctly given size is not sufficient to ensure the good performance of a test, it is a prerequisite. SE only present power results for the Rao test, since this test is shown to be superior in all situations. In our study, regarding the size, all tests perform well even in large systems of equations. To compare how the different test methods perform, consider the following power results:

Table 4. Estimated Size for the Alternative RESET Tests at 5% Nominal Size.

| | No. of Equations ($n$) | | | | | |
|---|---|---|---|---|---|---|
| | RAO = LRE = LRT = LRT-C | | | | | |
| $\Delta$ | 1 | 2 | 3 | 5 | 7 | 10 |
| 15 | 0.049 | 0.050 | 0.051 | 0.054 | 0.051 | 0.046 |
| 25 | 0.050 | 0.051 | 0.050 | 0.054 | 0.046 | 0.049 |
| 45 | 0.050 | 0.050 | 0.053 | 0.054 | 0.051 | 0.048 |
| 75 | 0.054 | 0.049 | 0.053 | 0.050 | 0.052 | 0.046 |

Wald = Wald-C

| $\Delta$ | 1 | 2 | 3 | 5 | 7 | 10 |
|---|---|---|---|---|---|---|
| 15 | 0.049 | 0.049 | 0.049 | 0.054 | 0.048 | 0.044 |
| 25 | 0.050 | 0.054 | 0.050 | 0.053 | 0.046 | 0.048 |
| 45 | 0.050 | 0.050 | 0.053 | 0.052 | 0.050 | 0.048 |
| 75 | 0.054 | 0.049 | 0.053 | 0.050 | 0.052 | 0.047 |

LM = LM-C

| $\Delta$ | 1 | 2 | 3 | 5 | 7 | 10 |
|---|---|---|---|---|---|---|
| 15 | 0.049 | 0.049 | 0.051 | 0.054 | 0.051 | 0.051 |
| 25 | 0.050 | 0.053 | 0.051 | 0.054 | 0.048 | 0.050 |
| 45 | 0.050 | 0.051 | 0.052 | 0.053 | 0.052 | 0.049 |
| 75 | 0.054 | 0.049 | 0.053 | 0.050 | 0.053 | 0.048 |

In this table $R_0^2$ = 0.7, $\rho_X$ = 0.5 and $\alpha$ = 0.0. The shading indicates bad performance as defined earlier in Table 3, i.e., when the results lie outside the approximate 95% confidence interval for actual size.

Table SE 4. Estimated Size for the Alternative RESET Tests at 5% Nominal Size.

| | No. of Equations ($n$) | | | | | | No. of Equations ($n$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RAO | | | | | | LRE | | | | | |
| $\Delta$ | 1 | 2 | 3 | 5 | 7 | 10 | 1 | 2 | 3 | 5 | 7 | 10 |
| 15 | .047 | .048 | .047 | .050 | .049 | .058 | .047 | .048 | .048 | .062 | .078 | .182 |
| 25 | .049 | .047 | .053 | .048 | .051 | .048 | .049 | .047 | .054 | .051 | .060 | .078 |
| 45 | .051 | .051 | .052 | .049 | .049 | .048 | .051 | .051 | .053 | .050 | .053 | .055 |
| 75 | .049 | .050 | .050 | .054 | .053 | .054 | .049 | .050 | .050 | .054 | .054 | .056 |

| | LRT-C | | | | | | Wald-C | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta$ | 1 | 2 | 3 | 5 | 7 | 10 | 1 | 2 | 3 | 5 | 7 | 10 |
| 15 | .051 | .056 | .058 | .082 | .110 | .249 | .069 | .120 | .193 | .504 | .841 | .998 |
| 25 | .052 | .051 | .060 | .061 | .075 | .103 | .062 | .085 | .132 | .279 | .559 | .921 |
| 45 | .052 | .053 | .057 | .054 | .059 | .065 | .057 | .074 | .096 | .162 | .291 | .602 |
| 75 | .050 | .051 | .052 | .058 | .058 | .062 | .054 | .062 | .074 | .109 | .167 | .339 |

| | LRT | | | | | | Wald | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta$ | 1 | 2 | 3 | 5 | 7 | 10 | 1 | 2 | 3 | 5 | 7 | 10 |
| 15 | .086 | .164 | .298 | .756 | .985 | 1.00 | .101 | .238 | .457 | .925 | .999 | 1.00 |
| 25 | .072 | .107 | .186 | .468 | .842 | .999 | .081 | .150 | .293 | .708 | .972 | 1.00 |
| 45 | .062 | .083 | .116 | .254 | .500 | .906 | .067 | .102 | .165 | .410 | .760 | .993 |
| 75 | .058 | .069 | .086 | .150 | .284 | .627 | .060 | .079 | .113 | .234 | .469 | .872 |

| | LM | | | | | | LM-C | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta$ | 1 | 2 | 3 | 5 | 7 | 10 | 1 | 2 | 3 | 5 | 7 | 10 |
| 15 | .071 | .087 | .112 | .285 | .608 | .970 | .037 | .012 | .003 | 0.00 | 0.00 | 0.00 |
| 25 | .063 | .069 | .090 | .162 | .353 | .763 | .042 | .025 | .011 | .001 | 0.00 | 0.00 |
| 45 | .058 | .062 | .073 | .105 | .187 | .419 | .047 | .036 | .026 | .008 | .001 | 0.00 |
| 75 | .054 | .058 | .062 | .083 | .118 | .241 | .048 | .042 | .035 | .021 | .009 | .002 |

*Source* : Shukur & Edgerton (2002, Table 4). In this table, $R_0^2$ = 0.7, $\rho_x$ = 0.5 and $\alpha$ = 0.0. The shading indicates bad performance, i.e., when the results lie outside the approximate 95% confidence interval for actual size.

The primary results reveal that, for single equation, the power functions for all the tests methods are identical. Moreover, in systems with more than one equation, we find that all the LR tests (uncorrected and corrected including the Rao's *F*-test) have identical results, and that the corrected and uncorrected Wald have identical results, and the same for the LM and corrected LM tests. This means that in single equation, the eight tests reduces to one and that we can present results from any one of them. In systems with more than one equation, the results differ between the three test groups (Wald, LR & LM).

The factors that affect the power of the RESET tests differ from those that affect the size. Although the number of equations (*n*), and degrees of freedom (Δ) had only a slight effect on the estimated size, they have a considerable effect on the power. As in the case of the size, changes in the autocorrelation between the exogenous variables ($\alpha$), and the goodness-of-fit in the null ($R_0^2$) did not produce any noticeable effects on the power function of the tests, and will not be shown in the diagrams.

The power of the RESET test did, as expected, depend on the degree of misspecification ($R_D^2$) and the correlation between the proxy in the auxiliary regression and the omitted variable ($\rho_{\eta z}$). The greater the misspecification, and the better the RESET proxy mirrors the omitted variable, the greater the power of the tests.

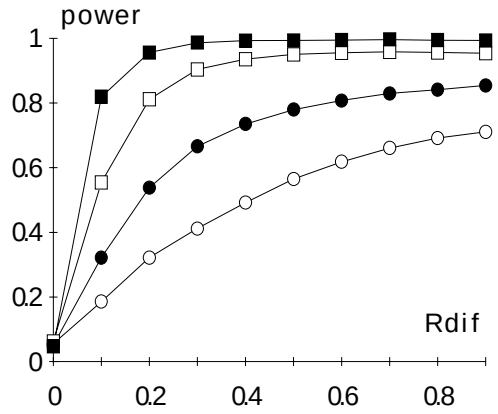
In Figure 1, the power functions of the three test methods are shown at a nominal size of 5% for different degrees of freedom (Δ) and for systems with different numbers of equations (*n*). The autocorrelation in the exogenous variables ($\alpha = 0$) is fixed, the goodness-of-fit in the null ($R_0^2 = 0.7$) and the correlation between the included and omitted variables ($\rho_{\eta z} = 0.5$). The power functions have also been calculated at other values, but because the patterns obtained are essentially the same they are excluded to save space.

It can be seen from the diagrams in Figure 1 that the power functions satisfy the expected properties of increasing with Δ and $R_D^2$ (which is denoted *Rdif* in the figure). The rate at which the power approaches one is heavily dependent on the values of Δ and *n*, however. It is quite clear that the Wald tests exhibits the best power among the others, especially in large samples (when n = 10). The LR tests (or the Rao test) is next best, while the LM test comes in third place. Note that in SE only results for the Rao test have been presented, which are very similar to our results for the LR tests groups, which we refer to as "Rao" in what follows.

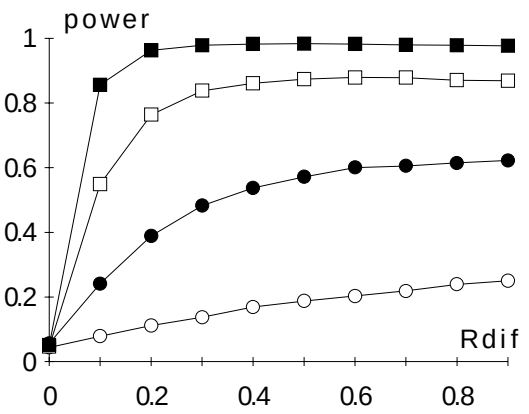
A closer examination of the diagrams shows that in small samples the power functions decrease as *n* increases, while in large samples, i.e., when Δ = 75, it can be seen that the power functions increase as *n* increases. The reason for this is that when *n* increases, the number of proxy variables that are included in the auxiliary regression also increases. Because each of these proxies is correlated with the omitted variable, their combined effect will tend to be greater when *n* increases (to hold this effect under control, the *multiple* correlation between the omitted variable and all of the proxy variables would have to be held constant) will obviously influence the power functions. Note also how, in small samples, the power functions become flatter as the number of equations increases, i.e., the tests become worse and worse, in particular the LM test. For large values of *n* and low degrees of freedom there is, little difference between the estimated size and estimated power.

Because SE only focus on the Rao test, and to facilitate comparison between the two papers, we will also present results for the Rao test. In Figure 2, the effect is shown of different values of $\rho_{\eta z}$ (rho in the figures) on the power function of the *RAO* test with 45 degrees of freedom, for systems with one, three, seven and ten equations. The power functions are shown at a nominal size of 5%, the autocorrelation in the exogenous variables ($\alpha = 0$) are fixed, and the goodness-of-fit in the null ($R_0^2 = 0.7$). The effect of the correlations between the proxies and the omitted variables is noticeable, and plays an important role on how quickly the power reaches the value of one. The effect of this factor is more dramatic in large systems, but again this is in part due to the usage of simple instead of multiple correlations.

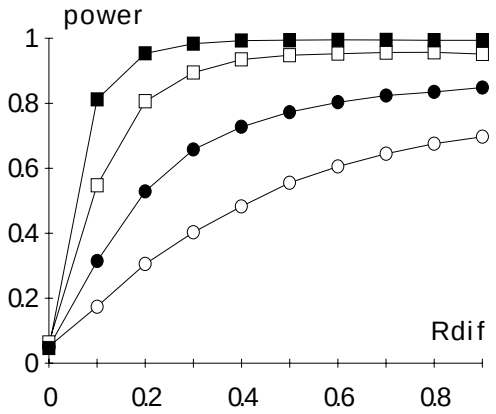
Figure 1 : The Power Function of the Wald, Rao and LM Tests for Three and Ten equations, Using the Bootstrap Critical Values.



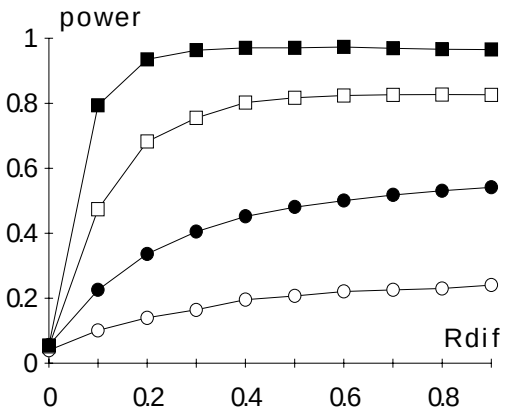

Wald test, 3 eq's :



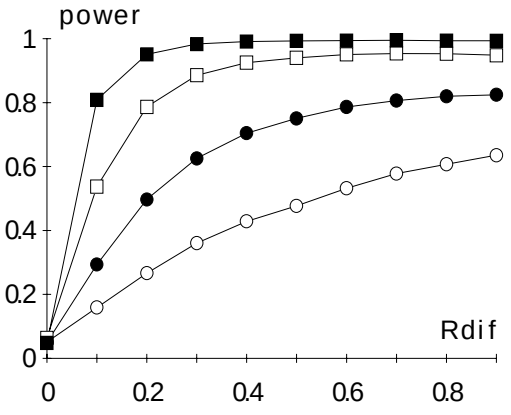

Wald test, 10 eq's :



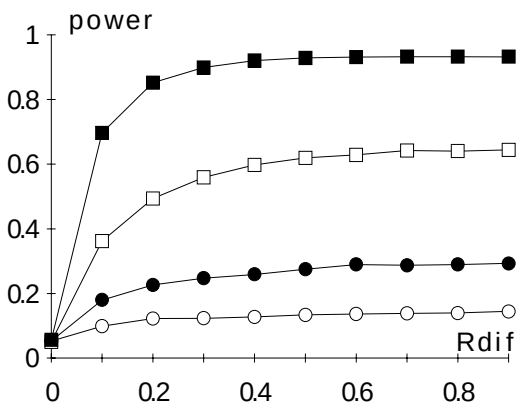

Rao test, 3 eq's :




Rao test, 10 eq's :




LM test, 3 eq's :




LM test, 10 eq's :

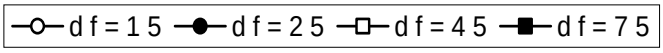
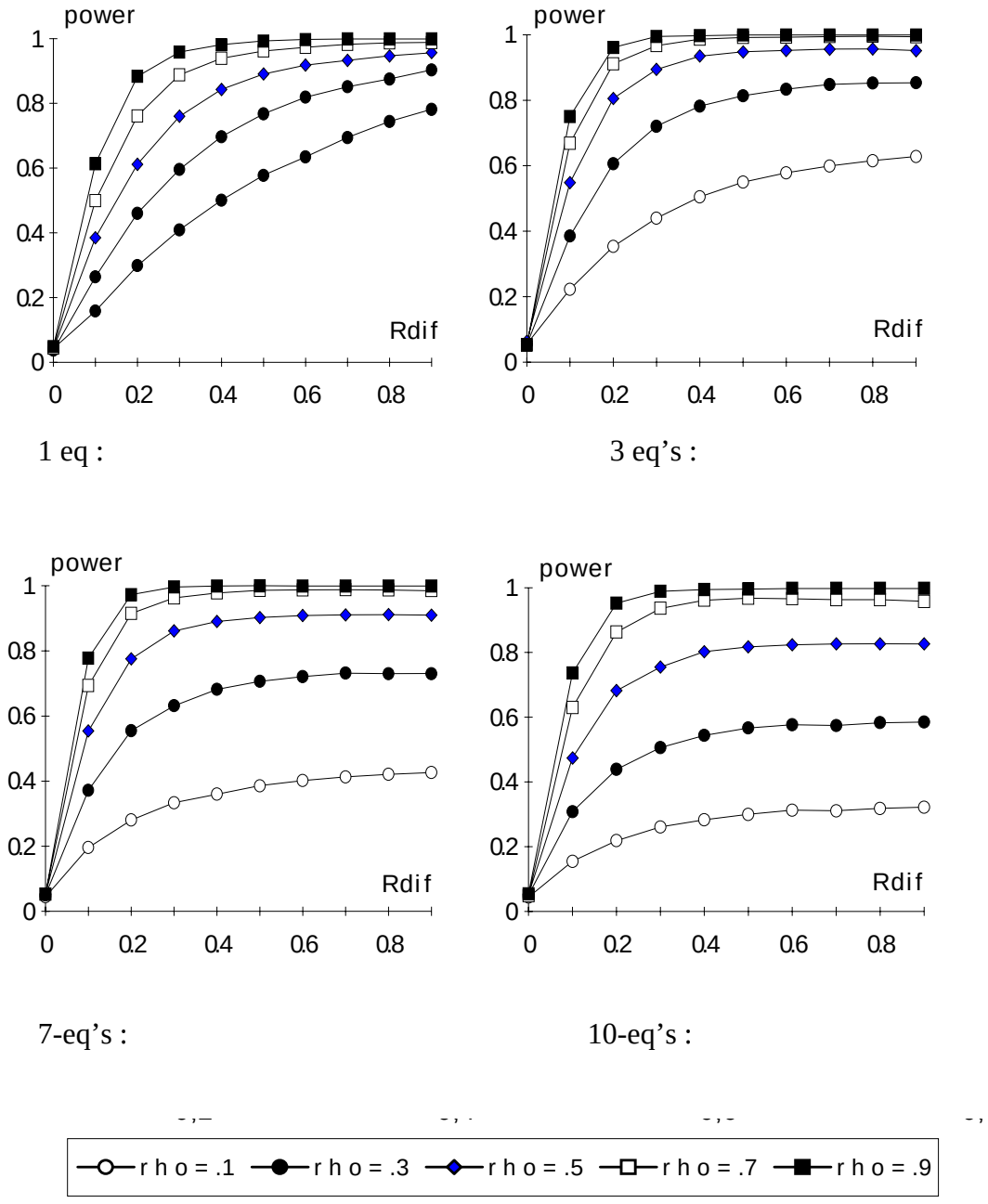


Figure 2 : The Power Function of Various Alternatives of the Rao Test with 45 df, Using the Bootstrap Critical Values.




Note also how the power functions become flatter for small $\rho_{\eta z}$ as the number of equations increases. For high values of *n* and low $\rho_{\eta z}$ there is very little difference between the estimated size and the estimated power. Note that, regarding the Rao test, our results are almost identical to those obtained by SE.

## Conclusion

The *size and power* of systemwise generalisations of Ramsey's RESET test was examined for misspecification errors by using "bootstrap critical values" ($c^*_{t\alpha}$). Shukur and Edgerton (2002) (SE) studied the same properties of the test, but they used the asymptotic critical values instead. The purpose of this paper is to show the ability of the bootstrap technique to produce critical values that might be much more accurate than the asymptotic ones.

We followed the same principle as in SE to construct Wald, Lagrange Multiplier and Likelihood Ratio tests that are applicable to auxiliary regression systems. Various degrees-of-freedom corrections have been investigated, in particular the commonly used simple replacement of the number of observations ($T$) by the degrees-of-freedom ($\Delta$) and, for the LR test, the Edgeworth correction developed by Anderson (1958). We also studied the properties of the systemwise $F$-test approximation proposed by Rao (1973).

The investigation has been carried out using Monte Carlo simulations. A large number of models were investigated, where the number of equations, degrees of freedom, error variance and stochastic properties of the exogenous variables have been varied. For each model, we performed 10,000 replications and studied four different nominal sizes. The power properties have been investigated using 2,000 replications per model, where in addition to the properties mentioned above the degree of misspecification (measured as the relative difference in the explanatory power between the null and true models) and the correlation between the omitted and included variables have also varied.

The analysis reveals that, in single equations, all test method are identical regarding the estimated size and power, while in systems with many equations the eight tests reduce to three groups, namely Wald, LR (or Rao), and LM. Although SE found that the Rao's $F$-test is the best and that the uncorrected LR test and both the corrected and uncorrected Wald and *LM* tests are shown to perform *extremely* badly in all situations, our analysis reveals that, in almost all cases, the performance of all the tests are satisfactorily.

The factors that affect the power of the RESET tests differ from those that affect the size. While the number of equations ($n$), and degrees of freedom ($\Delta$) had only a slight effect on the estimated size, they have a considerable effect on the power. As in the case of the size, changes in the autocorrelation between the exogenous variables ($\alpha$), and the goodness-of-fit in the null ($R_0^2$) did not produce any noticeable effects on the power function of the tests. The power of the RESET test did, as expected, depend on the degree of misspecification ($R_D^2$) and the correlation between the proxy in the auxiliary regression and the omitted variable ($\rho_{\eta z}$). The greater the misspecification, and the better the RESET proxy mirrors the omitted variable, the greater the power of the tests.

As regards the power, the Wald test has been shown to perform somewhat better than the others especially in small samples and large systems, but the differences between the alternative RESET tests are minimal. The Rao test performs well in our study as well as in that of SE, i.e., when using the asymptotic critical values and the "bootstrap critical values", which reinforces our picture of good performance in both cases. Generally, the power functions become flatter for small $\rho_{\eta z}$ as the number of equations increases. For high values of $n$ and low $\rho_{\eta z}$ there is indeed very little difference between the estimated size and the estimated power.

## References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, (2ed). New York, Wiley.

Bewley, R. (1986). Allocation Models: Specification Estimation and Applications. Cambridge Mass. Ballinger.

Edgerton, D. L., & Shukur, G. (1999). Testing Autocorrelation in a System Perspective. *Econometric Reviews*, *18*, 343-386.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. SIAM: Philadel.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, *14*, 1453-1462.

Horowitz, J. L. (1994). Bootstrap-based critical values for the information matrix test. *Journal of Econometric*, *6*1, 395-411.

Kiviet, J. (1986). On the Rigour of Some Specification Test for Modelling Dynamic Relationships. *Review of Economic Studies*, *53*, 241-262.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics, 21*, 255-285.

Mantalos, Panagiotis, & Shukur, G. (1998). Size and Power of the Error Correction Model (ECM) of Cointegration Tests. A Bootstrap Approach. *Oxford Bulletin of Economics and Statistics*, *60*, 249-255.

Ramsey, J. B. (1969). Test for Specification error in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society, Series B, 31*, 350-371.

Rao, C. R. (1973). Linear Statistical *Inference and Its Applications*, ($2^{nd}$ ed.). New York: Wiley.

Shukur, G., & Edgerton, D. L. (2002). The Small Sample Properties of the RESET Test as Applied to Systems of Equations. *Journal of Statistical Computation and Simulation, 72* (12), 909-924, 2002.

Thursby, J. G., & Schmidt, P. (1977): Some Properties of Tests for Specification Error in a Linear Regression Model. *Journal of American Statistical Association*, *72*, 635-641.