OXFORD BULLETIN OF ECONOMICS AND STATISTICS, SPECIAL ISSUE (1999)
0305-9049

A COMPARATIVE STUDY OF UNIT ROOT TESTS WITH PANEL DATA AND A NEW SIMPLE TEST

*G. S. Maddala and Shaowen Wu**

Simplicity, simplicity, simplicity: I say let your affairs be as two or three and not a hundred or thousand.
Simplify, simplify.

(H. D. Thoreau: *Walden*)

I. INTRODUCTION

Since the appearance of the papers by Levin and Lin (1992, 1993), the use of panel data unit root tests has become very popular among empirical researchers with access to a panel data set. It is by now a generally accepted argument that the commonly used unit root tests like the Dickey–Fuller (DF), augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) tests lack power in distinguishing the unit root null from stationary alternatives, and that using panel data unit root tests is one way of increasing the power of unit root tests based on a single time series. See, for example, the arguments in Oh (1996), Wu (1996), MacDonald (1996) and Frankel and Rose (1996), who try to resurrect the purchasing power parity (*PPP*) theory using panel data unit root tests.

Such use of panel unit root tests may not be meaningful because different null hypotheses are being tested in each case. For instance, consider the simplified model

$$\Delta y_{i,t} = \rho_i y_{i,t-1} + \varepsilon_{i,t},\ i = 1, 2, \ldots, N,\ t = 1, 2, \ldots, T.$$

Suppose we are interested in testing $\rho_1 = 0$ vs. $\rho_1 < 0$, we apply a single equation unit root for the first time series. The panel data unit root test tests a different hypothesis:

$$H_0\text{: } \rho_i = 0 \text{ vs. } H_1\text{: } \rho_i < 0, \text{ for } i = 1, 2, \ldots, N.$$

Furthermore, there are now more powerful tests available even in the single equation context. See, for example, Elliott, Rothenberg and Stock (1996) and also Perron and Ng (1996) who suggest modifications of the PP test to increase power.

*An earlier version of this paper was presented at the Econometric Society meetings in New Orleans, January 1997. We would like to thank Anindya Banerjee for helpful comments.

In the following sections, we shall present a review of the Levin–Lin tests, their extension by Im, Pesaran and Shin (1997), which will be referred to as the IPS test, and a simple alternative due to Fisher (1932) which will be referred to as the Fisher test. We then present some results from Monte Carlo experiments comparing these three tests and some evidence from the bootstrap which allows for correlation in the errors. Finally, we consider a test based on the Bonferroni inequality and conduct a Monte Carlo experiment to compare this test with the Fisher test.

II. PANEL DATA UNIT ROOT TESTS

Some early papers on testing for unit roots based on panel data are by Quah (1992, 1994) and Breitung and Mayer (1994). Since these have been superseded by the papers by Levin and Lin (1992, 1993), they are not discussed here.

2.1. The Levin–Lin (LL) Tests

Levin and Lin (1992) conduct an exhaustive study and develop unit root tests for the model:

$$\Delta y_{i,t} = \rho y_{i,t-1} + \alpha_0 + \delta t + \alpha_i + \theta_t + \varepsilon_{i,t},\ i = 1, 2, \ldots, N,\ t = 1, 2, \ldots, T.$$

Thus, the model incorporates a time trend as well as individual and time-specific effects. Initially, they assume that $\varepsilon_{i,t} \sim IID(0, \sigma^2)$ but they state that under serial correlation, with the inclusion of lagged first differences as in the ADF test, the test statistics have the same limiting distributions as mentioned subsequently, provided the number of lagged differences increase with sample size. Levin and Lin consider several subcases of this model. In all cases the limiting distributions are as $N \to \infty$ and $T \to \infty$. Also in all cases, the equation is estimated by *OLS* as a pooled regression model. The submodels are:

Model 1: $\Delta y_{i,t} = \rho y_{i,t-1} + \varepsilon_{i,t} \quad H_0$: $\rho = 0$
Model 2: $\Delta y_{i,t} = \rho y_{i,t-1} + \alpha_0 + \varepsilon_{i,t} \quad H_0$: $\rho = 0$
Model 3: $\Delta y_{i,t} = \rho y_{i,t-1} + \alpha_0 + \delta t + \varepsilon_{i,t} \quad H_0$: $\rho = 0, \delta = 0$
Model 4: $\Delta y_{i,t} = \rho y_{i,t-1} + \theta_t + \varepsilon_{i,t} \quad H_0$: $\rho = 0$
Model 5: $\Delta y_{i,t} = \rho y_{i,t-1} + \alpha_i + \varepsilon_{i,t} \quad H_0$: $\rho = 0, \alpha_i = 0$ for all i
Model 6: $\Delta y_{i,t} = \rho y_{i,t-1} + \alpha_i + \delta_i t + \varepsilon_{i,t} \quad H_0$: $\rho = 0, \delta_i = 0$ for all i.

For models 1–4, they show that

(a) $T\sqrt{N}\hat{\rho} \Rightarrow N(0, 2)$
(b) $t_{\rho=0} \Rightarrow N(0, 1)$.

For model 5, if $\sqrt{N}/T \to 0$, then

(a) $T\sqrt{N}\hat{\rho} + 3\sqrt{N} \Rightarrow N(0,\ 10.2)$

(b) $\sqrt{1.25}t_{\rho=0} + \sqrt{1.875N} \Rightarrow N\left(0, \frac{645}{112}\right)$.

In model 6, both intercept and time trend vary with individuals.

In the empirical applications, Oh (1996) uses only models 1 and 5. Wu (1996) uses the complete model with trend, and individual and time-specific effects but uses the distributions derived for model 5. Papell (1997) uses model 5 with lagged first differences added but computes his own exact finite sample critical values using Monte Carlo methods and finds them 3 to 15 percent higher than those tabulated in Levin and Lin (1992).

Levin and Lin argue that in contrast to the standard distributions of unit root test statistics for a single time series, the panel test statistics have limiting normal distributions. However, the convergence rates are faster as $T \to \infty$ (superconsistency) than as $N \to \infty$.

The paper by Levin and Lin (1993) provides some new results on panel unit root tests. These tests are designed to take care of the problem of heteroscedasticity and autocorrelation. They involve the following steps.

(i) Subtract cross-section averages from the data to eliminate the influence of aggregate effects.

(ii) Apply the augmented Dickey–Fuller (ADF) test to each individual series and normalize the disturbances. For illustration, we use model 5. The ADF regression

$$\Delta y_{i,t} = \rho_i y_{i,t-1} + \sum_{j=1}^{p_i} \theta_{ij} \Delta y_{i,t-j} + \alpha_i + \varepsilon_{i,t} \tag{1}$$

is equivalent to performing two auxiliary regressions of Δy_{it} and $y_{i,t-1}$ on the remaining variables in equation (1). Let the residuals from these two auxiliary regressions be $\hat{e}_{i,t}$ and $\hat{V}_{i,t-1}$ respectively. Now regress $\hat{e}_{i,t}$ on $\hat{V}_{i,t-1}$:

$$\hat{e}_{i,t} = \rho_i \hat{V}_{i,t-1} + \varepsilon_{i,t}$$

to get $\hat{\rho}_i$ which is equivalent to the *OLS* estimator of ρ_i in (1) directly. Since there is heteroscedasticity in $\varepsilon_{i,t}$, they suggest the following normalization to control it:

$$\hat{\sigma}^2_{e_i} = \frac{1}{T - p_i - 1} \sum_{t=p_i+2}^{T} (\hat{e}_{i,t} - \hat{\rho}_i \hat{V}_{i,t-1})^2$$

$$\tilde{e}_{i,t} = \frac{\hat{e}_{i,t}}{\hat{\sigma}_{e_i}}$$

$$\tilde{V}_{i,t-1} = \frac{\hat{V}_{i,t-1}}{\hat{\sigma}_{e_i}}.$$

Asymptotically, $\tilde{e}_{i,t}$ will be i.i.d. for all individual i.

(iii) Estimate the ratio of long-run to short-run standard deviation for each individual series and then calculate the average ratio for the panel as:

$$\hat{S}_{NT} = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{\sigma}_{y_i}}{\hat{\sigma}_{e_i}}$$

where the long-run variance $\hat{\sigma}^2_{y_i}$ is estimated by

$$\hat{\sigma}^2_{y_i} = \frac{1}{T-1}\sum_{t=2}^{T}\Delta y^2_{i,t} + 2\sum_{L=1}^{\overline{K}} w_{\overline{K}L}\left(\frac{1}{T-1}\sum_{t=L+2}^{T}\Delta y_{i,t}\Delta y_{i,t-L}\right).$$

$\overline{K}$ is the lag truncation parameter and $w_{\overline{K}L}$ is some lag window.

(iv) Compute the panel test statistic. Then consider the following regression:

$$\tilde{e}_{i,t} = \rho\tilde{V}_{i,t-1} + \tilde{\varepsilon}_{i,t}$$

using all i and t. The resulting t-statistic is

$$t_{\rho=0} = \frac{\hat{\rho}}{RSE(\hat{\rho})}$$

where

$$RSE(\hat{\rho}) = \hat{\sigma}_{\varepsilon}\left[\sum_{i=1}^{N}\sum_{t=2+p_i}^{T}\hat{V}^2_{i,t-1}\right]^{-1/2}$$

$$\hat{\sigma}^2_{\varepsilon} = \frac{1}{N\tilde{T}}\sum_{i=1}^{N}\sum_{t=2+p_i}^{T}(\tilde{e}_{i,t} - \hat{\rho}\tilde{V}_{i,t-1})^2$$

$$\tilde{T} = T - p - 1 \text{ and } \overline{p} = \frac{1}{N}\sum_{i=1}^{N}p_i$$

is the average lag length used in the individual ADF regression.

Since the test statistic is not centered at zero, Levin and Lin suggest using the following adjusted t-statistic:

$$t^*_{\rho} = \frac{t_{\rho=0} - N\tilde{T}\hat{S}_{NT}\hat{\sigma}_{\varepsilon}^{-2}RSE(\hat{\rho})\mu^*_{\tilde{T}}}{\sigma^*_{\tilde{T}}}$$

where $\mu^*_{\tilde{T}}$ and $\sigma^*_{\tilde{T}}$ are the mean and the standard deviation adjustment terms which are obtained from Monte Carlo simulation and tabulated in their paper. Under

$$H_0: \rho = 0,\ t^*_{\rho} \Rightarrow N(0, 1).$$

In the simulations reported later we used this procedure.

The major limitation of the Levin–Lin tests is that ρ is the same for all observations. Thus, if we denote by ρ_i the value of ρ for the ith cross-section unit then the Levin–Lin test specifies the null H_0 and alternative H_1 as:

$$H_0: \rho_1 = \rho_2 = ... = \rho_N = \rho = 0$$
$$H_1: \rho_1 = \rho_2 = ... = \rho_N = \rho < 0.$$

The null makes sense under some circumstances, but the alternative is too strong to be held in any interesting empirical cases. For example, in testing the convergence hypothesis in growth models, one can formulate the null as implying that none of the economies under study converges and thus $\rho = 0$ for all countries. But it does not make any sense to assume that all the countries will converge at the same rate if they do converge.

2.2. *The Im–Pesaran–Shin (IPS) Test (1997)*

IPS relax the assumption that $\rho_1 = \rho_2 = ... = \rho_N$ under H_1. The basic idea of the test is very simple. Take model 5 in Levin and Lin and substitute ρ_i for ρ. Essentially what we have is a model with a linear trend for each of the N cross-section units. Thus, instead of pooling the data, we use separate unit root tests for the N cross-section units. Consider the t-test for each cross-section unit based on T observations. Let $t_{i,T}$ $(i = 1, 2, \ldots, N)$ denote the t-statistics for testing unit roots, and let $E(t_{i,T}) = \mu$ and $V(t_{i,T}) = \sigma^2$. Then

$$\sqrt{N}\frac{(\bar{t}_{N,T} - \mu)}{\sigma} \Rightarrow N(0, 1), \text{ where } \bar{t}_{N,T} = \frac{1}{N}\sum_{i=1}^{N} t_{i,T}.$$

The problem is computing μ and σ^2. This they do by Monte Carlo methods and tabulate them for ready reference (Tables 3 and 4 of their paper).

Although IPS talk of their test as a generalization of the LL tests, the important thing to note is that the IPS test is a way of combining the evidence on the unit root hypothesis from the N unit root tests performed on the N cross-section units. Note that implicit in the test is the assumption that T is the same for all cross-section units and hence $E(t_{i,T})$ and $V(t_{i,T})$ are the same for all i. Thus, we are considering only balanced panel data. In practice, if unbalanced data are used, more simulations have to be carried out to get critical values.

In the case of serial correlation, IPS propose using the ADF t-test for individual series. However, $E(t_{i,T})$ and $V(t_{i,T})$ will vary as the lag length included in the ADF regression varies. They tabulate $E(t_{i,T})$ and $V(t_{i,T})$ for different lag lengths. In practice, however, to make use of their tables, we are restricted implicitly to using the same lag length for all the ADF regressions for individual series.

IPS also suggest an *LR*-bar test based on likelihood ratio statistics, but we

shall concentrate our discussion on their *t*-bar test. The same arguments apply to the *LR*-bar test.

2.3. *Fisher's (p_λ) Test*

It should be noted that the IPS test is for testing the significance of the results from N independent tests of a hypothesis. There is a large amount of literature on this issue dating back to Tippett (1931) and Fisher (1932). This problem has been studied under the title 'Meta Analysis' and the different tests are reviewed in Hedges and Olkin (1985, Chapter 3). All these procedures depend on different ways of combining the observed significance levels (p-values) from the different tests. If the test statistics are continuous, the significance levels π_i ($i = 1, 2, \ldots, N$) are independent uniform (0, 1) variables, and $-2\log_e \pi_i$ has a χ^2 distribution with two degrees of freedom. Using the additive property of the χ^2 variables, we get $\lambda = -2\sum_{i=1}^{N} \log_e \pi_i$ has a χ^2 distribution with $2N$ degrees of freedom. This is the test suggested by Fisher (1932). Pearson suggested a slight modification of this and the Fisher test goes under the name of p_λ test. It is discussed in Rao (1952, p. 44) and by Maddala (1977, p. 47) but there have not been many econometric applications of this test.

Tippett suggested using the distribution of the smallest of the p-values, π_i. There have been several other suggestions about the p-value combinations. Becker (1977) lists 16 of them, but no p-value combination is most powerful. However, the Fisher test based on the sum of the log-p-values has been widely recommended. In this paper, we shall use the Fisher test. (We tried the Tippett test as well but it was not as powerful as the Fisher test, and hence the results are not reported.)

The advantage of this test is that it does not require a balanced panel as in the case of the IPS test. Also, one can use different lag lengths in the individual ADF regression. Another advantage of the Fisher test is that it can also be carried out for any unit root test derived. The disadvantage is that the p-values have to be derived by Monte Carlo simulation. The IPS test is easy to use because there are ready tables available in the paper for $E(t_{i,T})$ and $V(t_{i,T})$. However, these are valid only for the ADF test.

III. A COMPARISON OF THE DIFFERENT TESTS

Some broad comments on the merits and demerits of the three tests would be useful in interpreting the results to be presented later. Hence we shall go through these first.

(1) The LL tests test a very restrictive hypothesis that is rarely of practical interest.

(2) The IPS test is claimed to be a generalization of the LL tests. However, it is better viewed as a way of combining the evidence of several independent unit root tests.

(3) Im–Pesaran–Shin present a power comparison of the LL and IPS tests and argue that the IPS test is more powerful than the LL test. However, strictly speaking, the power comparison is not valid. Although the null hypothesis is the same in the two tests, the alternative hypothesis is different. The LL tests are based on homogeneity of the autoregressive parameter (although there is heterogeneity in the error variances and the serial correlation structure of the errors). Thus the tests are based on pooled regressions. The IPS test, on the other hand, is based on heterogeneity of the autoregressive parameter. As argued earlier, the test amounts to a combination of different independent tests. There is no pooling of data involved as in the LL tests. In the following sections, we shall also present power comparisons with the LL test but it should be borne in mind that the LL test will necessarily come out worse because the LL test has to use the panel estimation method which is not valid if there is no need for pooling.

(4) The Fisher test and the IPS test *are* directly comparable. The aim of both tests is a combination of the significance of different *independent tests*. The Fisher test is non-parametric; whatever test statistic we use for testing for a unit root for each sample, we can get the p-values π_i and then $-2\sum \log_e \pi_i \sim \chi^2$ with $2N$ d.f., where N is the number of separate samples. The IPS test, on the other hand, is parametric. The distribution of the t-bar statistic involves the mean and variance of the t-statistics used. IPS compute this for the ADF test statistic for different values of the number of lags used and different sample sizes. However, these tables are valid *only* if the ADF test is used for the unit root tests. Also, if the length of the time series for the different samples is different, there is a problem using the tables prepared by IPS. The Fisher test does not have any such limitations. It can be used with any unit root test and even if the ADF test is used, the choice of the lag length for each sample can be separately determined. Also, there is no restriction of the sample sizes for different samples (they can vary according to availability of the data).

(5) The Fisher test is an exact test. The IPS test is an asymptotic test. Note that this does not lead to a huge difference in finite sample results, since the adjustment terms in the IPS test are derived from simulations while the p-values in the Fisher test are also derived from simulations. However, the asymptotic validity of the tests depends on different conditions. For the IPS test the asymptotic results depend on N going to infinity while for the Fisher test they depend on T going to infinity.

(6) The crucial element that distinguishes the two tests is that the Fisher test is based on combining the *significance levels* of the different tests, and the IPS test is based on combining the *test statistics*. Which is better is the question. We conducted Monte Carlo studies with these issues in mind.

(7) Both the Fisher test and IPS test are based on combining independent tests. So if there is contemporaneous correlation, then there are correlations

among the individual test statistics. Both tests will need modifications in this case.

IV. DESIGN OF THE MONTE CARLO STUDIES

In this section, we present details on the designs of Monte Carlo simulations. Basically, there are two experiments. The first one studies the size and power performances of the three unit root tests under the general setup of the null and the alternative hypotheses, i.e., H_0: $\rho_i = \rho = 0$ for all i, and H_1: $\rho_i = \rho < 0$ for all i. In the second experiment, we change the alternative to H_1: $\rho_i = 0$ for some i, and $\rho_i < 0$ for the other i. We do so due to the following consideration. In reality, some series in a panel under study might be stationary while the others are non-stationary. We expect the null to be rejected for such panels. But due to the masking problems, power might be relatively low for panel unit root tests. Thus we want to investigate the powers of the three panel unit root tests in this circumstance. Details of these two designs are as follows.

4.1. Basic Monte Carlo Simulation Design

In this experiment, we use the following data-generating process (*DGP*) for a dynamic panel containing group and time-specific effects. The coefficient on the lagged dependent variable is denoted by ϕ_i in the *DGP*. Thus,

$$\text{Drift model: } \Delta y_{i,t} = -\phi_i \mu_i + \phi_i y_{i,t-1} + u_{i,t}$$

$$\text{Trend model: } \Delta y_{i,t} = \mu_i - \phi_i \mu_i t + \phi_i y_{i,t-1} + u_{i,t}$$

where $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, and $y_{i,0}$ is random.

The error term $u_{i,t}$ contains a time-specific effect θ_t and random component $\varepsilon_{i,t}$:

$$u_{i,t} = \theta_t + \varepsilon_{i,t}$$

where $\theta_t = 0.9\theta_{t-1} + \omega_t$, $\omega_t \sim N(0, 1)$ and $\varepsilon_{i,t} = \lambda_i \varepsilon_{i,t-1} + e_{i,t}$, λ_i's are randomly generated on $U[0.2, 0.4]$ and different for each i where U denotes the uniform distribution. We assume $e_{i,t}$ to be jointly normal distributed with

$$E(e_{i,t}) = 0,\ E(e_{i,t}, e_{j,s}) = \begin{cases} \sigma_{ij} & \text{for } t = s \\ 0 & \text{for } t \neq s. \end{cases}$$

If we let Σ denote $(\sigma_{ij})_{i,j=1}^{N}$, then non-zero terms on the off-diagonal terms in Σ represents the existence of cross-correlations. Here we randomly generate some positive definite matrices for Σ.

In the simulation, μ_i is generated randomly from $N(0, 1)$ and then fixed

for each model. The randomness of $y_{i,0}$ is achieved by generating extra y's and discarding some initial observations.

This experiment is very similar to IPS's experiments. The biggest difference is that we allow contemporaneous correlation in $\varepsilon_{i,t}$.

After the data have been generated, we apply the LL, IPS and Fisher tests. For the Fisher test, we apply the ADF(p) test for each individual series. The following two models are estimated.

$$\text{Drift model: } \Delta y_{i,t} = \alpha_i + \rho_i y_{i,t-1} + \sum_{j=1}^{p} \gamma_{ij} \Delta y_{i,t-j} + \text{residual}$$

$$\text{Trend model: } \Delta y_{i,t} = \alpha_i + \delta_i t + \rho_i y_{i,t} + \sum_{j=1}^{p} \gamma_{ij} \Delta y_{i,t-j} + \text{residual}$$

where $p = 0, 1, 2$ are used and p-values using the Dickey–Fuller t-distributions generated by 100,000 simulations for the corresponding ADF t-test statistics are computed. Consequently, $p_\lambda = (-2\sum \log_e \pi_i)$ is calculated. For the LL and IPS tests, we follow the procedures described in their papers.

In each experiment, we consider the cases of $N = 10, 25, 50, 100$ and $T = 10, 25, 50, 100$. A total of 2000 trials are used in computing the empirical size and power of the tests. In analyzing the size, ϕ_i is set to be 0, and in calculating the power, ϕ_i is set to be -0.1 for all i. The results are presented in Tables 1 and 2.

Table 1 reports the simulation results of the model with constant term only. Overall, the performance of the LL test is the worst but it should be noted that as argued earlier, the power comparison between the LL test and the other two tests is not valid. The IPS and the Fisher tests are directly comparable because they are tests of the same hypothesis. When reading Table 1, it should be kept in mind that the power comparisons are not appropriate because there are significant size distortions. However, we have not computed size-adjusted power because some broad conclusions can be drawn from the figures presented, without looking at the size-adjusted power.

The major conclusion that one can draw is that for high values of T (50 or 100) and values of $N = 50$ and 100 the Fisher test dominates the IPS test in the sense that the Fisher test has smaller size distortions and comparable power. The same conclusions follow from Table 2, the case with trend. We shall see that this conclusion about the dominance of the Fisher test is reinforced in the results presented in Table 3 later.

Another observation in Tables 1 and 2 is the effect of selection of order of the ADF regressions. We generate the error terms according to an $AR(1)$ model thus using ADF(1) is appropriate. From the results we can see the danger of under-selecting the order of ADF regression which is also pointed out by IPS. Meanwhile, over-selecting the order of ADF regression, i.e.,

TABLE 1
Size and Power of the Unit Root Tests (Constant Term Only)

			ADF(0)		*ADF(1)*		*ADF(2)*	
N	*T*		*Size*	*Power*	*Size*	*Power*	*Size*	*Power*
25	25	IPS	0.0065	0.0095	0.1060	0.4235	0.0590	0.2535
		Fisher	0.0060	0.0005	0.0440	0.1910	0.0250	0.0875
		LL	0.0230	0.0245	0.1585	0.4760	0.0880	0.2575
	50	IPS	0.0025	0.0650	0.0880	0.9000	0.0645	0.8085
		Fisher	0.0075	0.0175	0.0845	0.7575	0.0665	0.6300
		LL	0.0165	0.0205	0.1415	0.7660	0.1030	0.5550
	100	IPS	0.0010	0.7725	0.1050	1.0000	0.0920	0.9995
		Fisher	0.0015	0.1090	0.0230	0.9975	0.0175	0.9925
		LL	0.0220	0.0265	0.1440	0.9905	0.1300	0.9525
50	25	IPS	0.0000	0.0060	0.1040	0.6520	0.0400	0.3705
		Fisher	0.0040	0.0025	0.1070	0.4945	0.0575	0.2695
		LL	0.0115	0.0170	0.1805	0.6950	0.0840	0.3555

	50	IPS	0.0015	0.0690	0.1035	0.9920	0.0715	0.9620
		Fisher	0.0025	0.0040	0.0730	0.9355	0.0510	0.8250
		LL	0.0080	0.0105	0.1550	0.9400	0.1030	0.7865
	100	IPS	0.0000	0.9880	0.0905	1.0000	0.0745	1.0000
		Fisher	0.0005	0.6970	0.0545	1.0000	0.0410	1.0000
		LL	0.0025	0.0800	0.1410	1.0000	0.1130	0.9990
100	25	IPS	0.0000	0.0030	0.0985	0.8615	0.0235	0.5415
		Fisher	0.0000	0.0005	0.1315	0.7080	0.0495	0.4040
		LL	0.0040	0.0070	0.1950	0.8760	0.0660	0.5085
	50	IPS	0.0000	0.0830	0.0945	1.0000	0.0545	0.9995
		Fisher	0.0005	0.0025	0.0740	0.9970	0.0415	0.9780
		LL	0.0010	0.0075	0.1770	0.9990	0.0990	0.9495
	100	IPS	0.0000	1.0000	0.0875	1.0000	0.0645	1.0000
		Fisher	0.0000	0.9010	0.0460	1.0000	0.0345	1.0000
		LL	0.0005	0.0790	0.1575	1.0000	0.1225	1.0000

TABLE 2
Size and Power of the Unit Root Tests (Time Trend)

			ADF(0)		*ADF(1)*		*ADF(2)*	
N	*T*		*Size*	*Power*	*Size*	*Power*	*Size*	*Power*
25	25	IPS	0.0010	0.0005	0.1145	0.2000	0.0450	0.0900
		Fisher	0.0010	0.0000	0.1060	0.1805	0.0535	0.0765
		LL	0.0015	0.0020	0.1090	0.2285	0.0150	0.0320
	50	IPS	0.0000	0.0015	0.1055	0.4895	0.0655	0.3340
		Fisher	0.0000	0.0000	0.0940	0.3670	0.0525	0.2305
		LL	0.0000	0.0030	0.1180	0.5635	0.0365	0.2430
	100	IPS	0.0000	0.0190	0.1015	0.9895	0.0780	0.9620
		Fisher	0.0000	0.0040	0.0435	0.9210	0.0335	0.8390
		LL	0.0000	0.0130	0.1250	0.9665	0.0650	0.8420
50	25	IPS	0.0000	0.0000	0.1320	0.2635	0.0365	0.0790
		Fisher	0.0000	0.0000	0.1340	0.2295	0.0465	0.0780
		LL	0.0000	0.0005	0.1135	0.2850	0.0065	0.0195
	50	IPS	0.0000	0.0000	0.1155	0.6920	0.0705	0.4775
		Fisher	0.0000	0.0000	0.0705	0.4560	0.0385	0.2625
		LL	0.0000	0.0010	0.1165	0.7420	0.0235	0.3170

	100	IPS	0.0000	0.0275	0.0975	1.0000	0.0720	0.9990
		Fisher	0.0000	0.0005	0.0420	0.9970	0.0285	0.9830
		LL	0.0000	0.0100	0.1210	0.9990	0.0530	0.9710
100	25	IPS	0.0000	0.0000	0.1265	0.3305	0.0185	0.0700
		Fisher	0.0000	0.0000	0.1570	0.2850	0.0360	0.0785
		LL	0.0000	0.0000	0.0925	0.3335	0.0010	0.0085
	50	IPS	0.0000	0.0000	0.1105	0.9015	0.0470	0.6930
		Fisher	0.0000	0.0000	0.0645	0.7100	0.0260	0.4215
		LL	0.0000	0.0000	0.1155	0.9220	0.0130	0.4525
	100	IPS	0.0000	0.0240	0.0900	1.0000	0.0645	1.0000
		Fisher	0.0000	0.0005	0.0355	1.0000	0.0265	0.9995
		LL	0.0000	0.0030	0.1255	1.0000	0.0440	0.9990

Note: Empirical size is reported when the nominal size is 5%. Power is reported when ϕ is set to be -0.1.

$p = 2$, seems to alleviate the size distortions for all three tests, the price is the decline in power, as has been observed often.

4.2. Simulation with a Mixture of Stationary and Non-stationary Alternatives

In this experiment we generate the data according to the *DGP* described in the last section except that we set $\phi_i = 0$ for some of the individual series and $\phi_i < 0$ for the others as our maintained hypothesis. We study the case of $N = 25$ and $T = 50$. First, we consider the case where there is only one stationary process in the group, and we then increase the number of stationary processes (k) to 2, 4, 8, 10 and 12. For the cases of $k = 1, 2, 4$, we set $\phi_i = -0.2$, while for cases of $k = 8, 10, 12$, ϕ_i is generated from $U[-0.3, -0.1]$ and is different for each series. The powers of the three tests are reported in Table 3. The general result is that the Fisher test has the highest power in all cases. The more the number of stationary processes included, the stronger the relative advantage. Thus if only part of the panel is stationary, the Fisher test is the most likely one to point it out.

Broadly speaking, our conclusions are as follows:

1. In general, when there is no cross-sectional correlation in the errors, the IPS test is slightly more powerful than the Fisher test, in the sense that the IPS test has higher power when the two have the same size. Both tests are more powerful than the LL test.
2. As for the issues of heteroscedasticity and serial correlation in the errors, all the tests can take care of these problems. But when the errors in the different samples (or cross-section units) are cross-correlated (as would often be the case in empirical work) none of the tests can handle this problem well. However, the Monte Carlo evidence suggests that this problem is less severe with the Fisher test than with the LL or the IPS test. More specifically, when T is large but N is not very large, the size distortion with the Fisher test is small. But for

TABLE 3
*Power Comparison: Design 2 ($N = 25$, $T = 50$)**

Number of Stationary Units	*IPS*	*Fisher*	*LL*
1	0.0785	0.1085	0.0885
2	0.1270	0.1875	0.1040
4	0.1835	0.2490	0.1350
8	0.4685	0.5900	0.2545
10	0.6165	0.6840	0.3240
12	0.7385	0.8145	0.4045

* In the Monte Carlo simulation design, for the cases of stationary units included is 1, 2, 4, 8, we set ϕ_i to -0.2. In the cases of 10 and 12, the ϕ_i are generated at $U[-0.3, -0.1]$.

medium values of T and large N, the size distortion of the Fisher test is of the same level as that of the IPS test.

3. For the cases that we include a mixture of stationary and non-stationary series in the group as an alternative hypothesis, the Fisher test is the best because it has the highest power in distinguishing the null and the alternative.

Overall, we can conclude from these results that the Fisher test is better than the IPS and LL tests.

V. CORRELATED ERRORS AND THE BOOTSTRAP ALTERNATIVE

Recall that the properties of all three tests are based on the assumption that the error terms are not cross-correlated. When this assumption is violated the derived distributions for these test statistics are no longer valid. They all suffer from nuisance parameter problems. More specifically, if there is cross-correlation in the data, as noted earlier, the distributions of the test statistics are not the same as before and are not known. For the t-bar test in IPS, the t-statistics are correlated and hence the t-bar statistic does not have the stated variance in its (asymptotic) normal distribution, and for the Fisher test, too, we do not have independent tests and hence the p_λ criterion does not have the χ^2 distribution. Im, Pesaran and Shin (1997) consider a special case of correlated errors. They assume that the cross-correlations are caused by common time-specific effects, i.e. in the model $\Delta y_{i,t} = \rho_i y_{i,t-1} + u_{i,t}$ we have $u_{i,t} = \theta_t + \varepsilon_{i,t}$. They suggest eliminating θ_t by subtracting out the mean $\overline{y} = (1/N)\sum_{i=1}^{N} y_{i,t}$ from $y_{i,t}$ before applying the unit root tests and the t-bar test. In many practical applications the cross-correlation is not likely to be of this simple form. In fact they suggest that if

$$u_{i,t} = r_i \theta_t + \varepsilon_{i,t}$$

this procedure of demeaning would not work.

O'Connell (1998) assumes cross-sectional dependence of the form

$$\Omega = \begin{bmatrix} 1 & \omega & \dots & \omega \\ \omega & 1 & \dots & \omega \\ . & . & . & . \\ \omega & \omega & \dots & 1 \end{bmatrix} \quad \omega < 1$$

but $E(u_{i,t}, u_{j,s}) = 0 \; \forall i, j$ if $s \neq t$.

This sort of covariance matrix would arise with random time effects as in random effects models, i.e. $u_{i,t} = \theta_t + \varepsilon_{i,t}$ where θ_t are i.i.d., $\varepsilon_{i,t}$ are i.i.d., θ_t and $\varepsilon_{i,t}$ are mutually independent.

One way out of this problem of cross-correlated errors is to use the bootstrap method to get the empirical distributions of the test statistics to make inferences. In what follows, we shall illustrate the bootstrap method by using an empirical example of testing convergence in real GDPs of 17

European countries. Please refer to Maddala and Wu (1999) for motivation of the problem and detailed discussion of the data. We shall discuss the bootstrap method briefly and then present the results. The bootstrap method for univariate time series is well developed. Readers are referred to Li and Maddala (1996) for a good introduction. Meanwhile the bootstrap method for panel data is in its infancy. We shall discuss our proposed method in detail.

The problem here is to generate the bootstrap distributions of the test statistics. The bootstrap data were generated using the sampling scheme S_3 described in Li and Maddala (1996). Generally speaking, if the null hypothesis is $H_0: \beta = \beta_0$, the sampling scheme S_3 suggests generating bootstrap sample y^* as $y^* = x\beta_0 + \varepsilon^*$ where ε^* is the bootstrap sample from $\varepsilon^0 = y - x\beta_0$. Since we have panel data here, we should also take care of special problems arising from the serial correlation. So we get the bootstrap sample of the error term ε^0 from

$$\Delta y_{i,t} = \eta_i \Delta y_{i,t-1} + \varepsilon^0_{i,t} \tag{2}$$

since under the null hypothesis $y_{i,t}$ has a unit root. Since there are cross-correlations among $\varepsilon^0_{i,t}$, we cannot resample $\varepsilon^0_{i,t}$ directly. We propose resampling $\varepsilon^0_{i,t}$ with the cross-section index fixed, i.e. instead of resampling $\varepsilon^0_{i,t}$, we resample $\varepsilon^0_t = [\varepsilon^0_{1,t}, \varepsilon^0_{2,t}, \ldots, \varepsilon^0_{N,t}]'$ to get ε^*_t. In this way, we can preserve the cross-correlation structure of the error term. Then the bootstrap sample y^* is generated as

$$y^*_{i,t} = y^*_{i,t-1} + u^*_{i,t} \text{ with } y^*_{i,0} = 0$$

$$u^*_{i,t} = \hat{\eta}_i u^*_{i,t-1} + \varepsilon^*_{i,t} \text{ with } u^*_{i,0} = \sum_{j=0}^{m} \hat{\eta}_i^j \varepsilon^*_{-j}$$

where $\hat{\eta}_i$'s are from estimation results of (2). The method to get the initial value of $u^*_{i,0}$ is suggested by Rayner (1990). Here ε^*_{-j}'s are drawn as an independent bootstrap sample and m is selected as 30. A total of 2000 replications are used to generate the empirical distributions.

We can get the critical values at the 5 percent level based on these empirical distributions. For the sample of the 17 European countries under study, these critical values are -0.518, 42.61 and -1.157 for the IPS, Fisher and LL tests respectively. Now, the test statistics for the same sample are -0.950, 48.912 and -1.078. In contrast to the non-rejection results based on the usual critical values (derived from standard normal for the IPS and LL tests and χ^2 for the Fisher test), we can now strongly reject the unit root null for the data set by using the critical values at the 5 percent level for the IPS and the Fisher tests, although for the LL test we fail marginally to reject the null.

In general, the conclusions from the bootstrap experiments are: In the case of cross-correlated errors there are substantial size distortions in using

the conventional test statistics for the IPS, Fisher and LL tests, although the size distortions are less serious for the Fisher test than for the IPS test. Using the bootstrap method results in a decrease of these size distortions, although it does not eliminate them. The Fisher test does better than the IPS test overall using the bootstrap method.

The size distortions with the Levin–Lin test in the case of cross-correlated errors have also been discussed in O'Connell (1998). He talks of dramatic size distortions, the actual size being 50 percent when the nominal size is 5 percent. The size distortions that we observed with the panel data unit root tests (including the LL test), under cross-correlations, were not as dramatic as those noted by O'Connell. It is possible that his results are a consequence of the equi-correlational error structure he assumed.

VI. TESTS BASED ON BONFERRONI INEQUALITY

For the correlated errors case we argued that, for both the Fisher test statistic and the IPS test statistic, the distribution involves nuisance parameters, but bootstrap methods could be used to make inferences. There is, however, one test statistic that is equally applicable in both the uncorrelated and correlated cases. This is the statistic based on the Bonferroni inequality (see Alt, 1982) and discussed in Dufour and Torres (1996). These tests have also been discussed earlier in Savin (1984) and Dufour (1990).

The idea behind this test is to break up the hypothesis H_0: $\rho_i = 0$ for all i, $i = 1, 2, \ldots, N$ into a set of sub-hypotheses H_{0i}: $\rho_i = 0$ and noting that H_0 is wrong if and only if any of its components H_{0i} is wrong. Suppose we choose the significance level π_i for the ith test. Then if we follow the rule that we reject H_0 if at least one of the sub-hypotheses H_{0i} is rejected at significance level π_i for H_{0i}, then the Bonferroni inequality says that the significance level π for H_0 is given by

$$\pi \leqslant \sum_{i=1}^{N} \pi_i.$$

One simple rule that Dufour and Torres suggest is to take $\pi_i = \pi/N$, unless there is an *a priori* compelling reason that some tests ought to be rejected at lower (or higher) significance level than the others.

We shall now compare this test with the Fisher test for both the cases of independent tests and correlated tests. Since the tables of significance values for unit root tests are readily available at the 1 percent significance level, we chose $N = 5$ in our investigation, so that $\pi_i = 0.01$ and $\pi = 0.05$. Note that the case of small N is favorable to the test based on the Bonferroni inequality (see p. 299 of Alt, 1982).

We consider four cases of *DGP*. The basic *DGP* is

$$\Delta y_{i,t} = \alpha_i + \rho_i y_{i,t-1} + \varepsilon_{i,t}$$

$$\alpha_i \sim U[0, 1]$$

$$\rho_i = \begin{cases} -0.1 & \text{for power} \\ 0 & \text{for size} \end{cases}$$

where assumptions on $\varepsilon_{i,t}$ are different in each case as follows:

Case 1: No serial correlation and contemporaneous correlation

$$\varepsilon_{i,t} \sim N(0, \sigma_i^2), \text{ where } \sigma_i^2 \sim U[0.5, 1.5].$$

Case 2: There is serial correlation, but no contemporaneous correlation

$$\varepsilon_{i,t} = \lambda_i \varepsilon_{i,t-1} + e_{i,t}, \text{ where } \lambda_i \sim U[0.2, 0.4]$$

and

$$e_{i,t} \sim N(0, \sigma_i^2), \text{ with } \sigma_i^2 \sim U[0.5, 1.5].$$

Case 3: There is contemporaneous correlation but no serial correlation

$$\varepsilon_t = [\varepsilon_{1,t}, \varepsilon_{2,t}, \ldots, \varepsilon_{N,t}]' \sim N(0, \Sigma),$$
$$\text{where } \Sigma = \Psi'\Psi \text{ with} \Psi_{N\times N} \sim U[-1, 1]$$

Case 4: There are both serial correlation and contemporaneous correlation

$$\varepsilon_{i,t} = \lambda_i \varepsilon_{i,t-1} + e_{i,t}, \text{ where } \lambda_i \sim U[0.2, 0.4]$$

and

$$e_t = [e_{1,t}, e_{2,t}, \ldots, e_{N,t}]' \sim N(0, \Sigma),$$
$$\text{where } \Sigma = \Psi'\Psi \text{ with } \Psi_{N\times N} \sim U[-1, 1].$$

Table 4 presents the results from the Monte Carlo studies based on 4000 simulations. The results clearly suggest that the power of the Dufour–Torres (DT) test is substantially lower than that of the Fisher test, although there is a slight size distortion for the Fisher test in the case of correlated tests.

VII. PANEL DATA TESTS WITH STATIONARITY AS NULL

Although the use of the Fisher test has been outlined in the previous sections with reference to the ADF test, to compare it with the IPS test, it can be used with *any* of the efficient unit root tests suggested in the literature, such as the Elliott *et al*. (1996) test or the Perron-Ng (1996) test. It can also be used for tests of stationarity as the null, like the test by Kwiatkowski *et al*. (1992) or Leybourne and McCabe (1994). We suggest the same procedure of combining the p-values to get a χ^2 test statistic and using the bootstrap method for obtaining the critical values, to account for the correlations among the test statistics for the individual cross-section units.

TABLE 4
Size and Power Comparison of the Fisher Test with the Dufour–Torres (DT) Test

Case No.	*Test*	*Size*	*Power*
Case 1	Fisher	0.0505	0.1180
$T = 25$	DT	0.0460	0.0693
Case 1	Fisher	0.0473	0.3142
$T = 50$	DT	0.0467	0.1232
Case 2	Fisher	0.0560	0.1135
$T = 25$	DT	0.0565	0.0760
Case 2	Fisher	0.0480	0.2802
$T = 50$	DT	0.0493	0.1185
Case 3	Fisher	0.0785	0.1630
$T = 25$	DT	0.0480	0.0683
Case 3	Fisher	0.0750	0.3448
$T = 50$	DT	0.0432	0.1215
Case 4	Fisher	0.0858	0.1530
$T = 25$	DT	0.0578	0.0767
Case 4	Fisher	0.0825	0.3073
$T = 50$	DT	0.0462	0.1143

VIII. THE CASE OF PANEL COINTEGRATION TESTS

Again the methodology we suggest is applicable to panel cointegration tests, whether they are tests using no cointegration as null, or cointegration as null. (There are several tests in each category – see Chapter 6 of Maddala and Kim (1998) for a review.) The procedure we suggest is simple and universally applicable. There is no need for a separate theory for each type of test. The only problem is the correlation among the test statistics for the different cross-section units, for which we suggest using the bootstrap method to get critical values for the χ^2 test. This is the procedure used by Wu (1998, Chapter 5) for deriving panel data cointegration tests to test the PPP (purchasing power parity) hypothesis. There have been several panel data cointegration tests suggested in the literature. See Pedroni (1995, 1997), Kao (1999) and McCoskey and Kao (1998a). They all allow for heterogeneity in the cointegrating coefficients. But the null and alternatives imply that either *all* the relationships are cointegrated or *all* the relationships are not cointegrated. There is no allowance for some relationships to be cointegrated and others not (the Fisher test allows for this – the p-values can be different).

Pedroni (1997) suggests seven test statistics. There is a lot of complicated algebra in the derivations which depend on the construction of LR covar-

iance matrices of the errors, whose small sample properties are known to be questionable. His tests are for no cointegration as null.

Kao (1999) suggests tests for no cointegration as the null and McCoskey and Kao (1998a) suggest tests for the null of cointegration. The methodology for the derivation of the test statistics in the two cases is different.

We have outlined briefly the deficiencies of these alternative tests for cointegration in panel data. A detailed discussion of these tests is beyond the scope of our paper. This can be found in McCoskey and Kao (1998b) who present a detailed Monte Carlo study of the tests by Kao, Pedroni and McCoskey and Kao. Limit theory for non-stationary panel data can be found in Phillips and Moon (1999).

IX. CONCLUSIONS

The paper compares the Levin–Lin and Im–Pesaran–Shin (IPS) panel data unit root tests with the Fisher test which was suggested over 60 years ago by R. A. Fisher and has a celebrated history in the statistical literature. The main conclusion of the paper is that the Fisher test is simple and straightforward to use and is a better test than the LL and IPS tests. Some other conservative tests (applicable in the case of correlated tests) based on the Bonferroni bounds have also been found to be inferior to the Fisher test. Our arguments also apply to tests using stationarity as null and to panel cointegration tests testing the null of no cointegration as well as testing the null of cointegration. The essential problem there is again one of combining the evidence from several tests. Also, the same problems of correlated tests have to be addressed, using bootstrap methods.

One major problem with the literature on unit root and cointegration tests in panel data is that there is an urge to generalize the tests used in univariate data to panel data under assumptions that are not likely to be meaningful in practice. There is more concentration on technical details and less on the questions being answered. This makes them not very useful in practice. For instance, the tests are almost all tests for the hypothesis that *all* series are stationary vs. *all* series are non-stationary, or that *all* series are cointegrated vs. that *none* is cointegrated. This is almost always a hypothesis of questionable value to test.

It is very important to bear in mind Thoreau's admonition quoted at the beginning of our paper.

The Ohio State University
SUNY at Buffalo

Date of Receipt of Final Manuscript: July 1999

REFERENCES

Alt, F. B. (1982). 'Bonferroni Inequalities and Intervals', in Johnson, N. L. and Kotz, S. (eds.) *Encyclopedia of Statistical Sciences*, Vol. I, 294-301, Wiley, New York.

Becker, B. J. (1997). '*P*-Values: Combination', in Kotz, S. and Read, C. B. (eds.) *Encyclopedia of Statistical Sciences: Update*, Vol. I, 448-453, Wiley, New York.

Breitung, J. and Mayer, W. (1994). 'Testing for Unit Roots in Panel Data: Are Wages on Different Bargaining Levels Cointegrated?', *Applied Economics*, 26, 353-361.

Cavanagh, C. L., Elliott, G. and Stock, J. H. (1995). 'Inference in Models with Nearly Nonstationary Regressors', *Econometric Theory*, 11, 1131-1147.

Dufour, J. M. (1990). 'Exact Tests and Confidence Sets in Linear Regressions with Autocorrelated Errors', *Econometrica*, 58, 479-494.

Dufour, J. M. and Torres, O. (1996). 'Union-Intersection and Sample-Split Methods in Econometrics with Application to MA and SURE Models', Paper presented at the summer meetings of the Econometric Society, Iowa City.

Elliott, G., Rothenberg, T. J. and Stock, J. H. (1996). 'Efficient Tests for an Autoregressive Unit Root', *Econometrica,* 64, 813-836.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh, 4th Edition.

Frankel, J. A. and Rose, A. K. (1996). 'A Panel Project on Purchasing Power Parity: Mean Reversion Within and Between Countries', *Journal of International Economics*, 40, 209-224.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta Analysis*, Academic Press, New York.

Im, K. S., Pesaran, M. H. and Shin, Y. (1997). 'Testing for Unit Roots in Heterogeneous Panels', Mimeo, Department of Applied Economics, University of Cambridge.

Kao, C. (1999). 'Spurious Regression and Residual-Based Tests for Cointegration in Panel Data', *Journal of Econometrics*, 90, 1-44.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. and Shin, Y. (1992). 'Testing for the Null of Stationarity Against the Alternative of a Unit Root', *Journal of Econometrics*, 54, 159-178.

Levin, A. and Lin, C. F. (1992). 'Unit Root Test in Panel Data: Asymptotic and Finite Sample Properties', University of California at San Diego, Discussion Paper No. 92-93.

Levin, A. and Lin, C. F. (1993). 'Unit Root Test in Panel Data: New Results', University of California at San Diego, Discussion Paper No. 93-56.

Leybourne, S. J. and McCabe, B. P. M. (1994). 'A Consistent Test for a Unit Root', *Journal of Business and Economic Statistics*, 12, 157-166.

Li, H. and Maddala, G. S. (1996). 'Bootstrapping Time Series Models' (with discussion), *Econometric Reviews*, 15, 115-195.

McCoskey, S. and Kao, C. (1998a). 'A Residual-Based Test for the Null of Cointegration in Panel Data', *Econometric Reviews*, 17, 57-84.

McCoskey, S. and Kao, C. (1998b). 'A Monte Carlo Comparison of Tests for Cointegration in Panel Data', Mimeo, Center for Policy Research, Syracuse University, February.

MacDonald, R. (1996). 'Panel Unit Root Tests and Real Exchange Rates', *Economics Letters*, 50, 7-11.
Maddala, G. S. (1977). *Econometrics*, McGraw-Hill, New York.
Maddala, G. S. and Wu, S. (1999). 'Cross-country Growth Regressions: Problems of Heterogeneity, Stability and Interpretation', forthcoming in *Applied Economics*.
Maddala, G. S. and Kim, I. M. (1998). *Unit Roots, Cointegration and Structural Change*, Cambridge University Press, Cambridge.
O'Connell, P. (1998). 'The Overvaluation of Purchasing Power Parity', *Journal of International Economics*, 44, 1-19.
Oh, K. Y. (1996). 'Purchasing Power Parity and Unit Root Tests using Panel Data', *Journal of International Money and Finance*, 15, 405-418.
Papell, D. H. (1997). 'Searching for Stationarity: Purchasing Power Parity under the Current Float', *Journal of International Economics,* 43, 313-332.
Pedroni, P. (1995). 'Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the *PPP* Hypothesis', Indiana University Working Paper in Economics No. 95-013.
Pedroni, P. (1997). 'Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the *PPP* Hypothesis: New Results', Indiana University Working Paper in Economics.
Perron, P. and Ng, S. (1996). 'Useful Modifications to some Unit Root Tests with Dependent Errors and their Local Asymptotic Properties', *Review of Economic Studies,* 63, 435-465.
Phillips, P. C. B. and Moon, H. (1999). 'Linear Regression Limit Theory for Nonstationary Panel Data', *Econometrica*, 67, 1057–1111.
Quah, D. (1992). 'International Patterns of Growth I: Persistence in Cross-Country Disparities', LSE Working Paper.
Quah, D. (1994). 'Exploiting Cross-Section Variation for Unit Root Inference in Dynamic Data', *Economics Letters*, 44, 9-19.
Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*, Wiley, New York.
Rayner, R. K. (1990). 'Bootstrapping P-values and Power in the First-Order Autoregression: A Monte Carlo Investigation', *Journal of Business & Economic Statistics*, 8, 251-263.
Savin, N. E. (1984). 'Multiple Hypothesis Testing', *Handbook of Econometrics*, eds. Griliches, Z. and Intriligator, M. D., Chapter 14, North-Holland, Amsterdam.
Tippett, L. H. C. (1931). *The Methods of Statistics*, Williams & Norgate, London, 1st Edition.
Wu, S. (1998). 'Nonstationary Panel Data Analysis', Thesis, Ohio State University.
Wu, Y. (1996). 'Are Real Exchange Rates Stationary? Evidence from a Panel Data Test', *Journal of Money, Credit and Banking*, 28, 54-63.