ELSEVIER

Journal of International Economics 57 (2000) 243–273

Journal of INTERNATIONAL ECONOMICS

www.elsevier.nl/locate/econbase

Long-run PPP may not hold after all

Charles Engel*

Department of Economics, University of Washington and NBER, Box 353330, Seattle, WA 98195-3330, USA

Received 5 September 1997; accepted 14 December 1998

Abstract

Recent tests using long data series find evidence in favor of long-run PPP. These tests may have reached the wrong conclusion. Using artificial data calibrated to nominal exchange rates and disaggregated data on prices, we show that tests on long-run PPP have serious size biases. In the baseline case, unit root and cointegration tests with a nominal size of 5% have true sizes that range from 0.90 to 0.99 in 100-year long data series, even though there is a permanent component that accounts for 42% of the 100-year forecast variance of the real exchange rate. Tests of stationarity are shown to have very low power in the same circumstances. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Cointegration; Purchasing power parity; Unit roots

JEL classification: F30; F40

1. Introduction

Recent work on purchasing power parity (PPP) among high-income countries has found evidence in favor of the hypothesis that real exchange rates converge to their PPP level in the long run.[1] This work reaches a conclusion opposite from earlier work that found real exchange rates have unit roots. The more recent work

*Tel.: +1-206-5436-197; fax: +1-206-6857-477.

E-mail address: cmengel@u.washington.edu (C. Engel)

[1]For example, Frankel (1986), Kim (1990), Abuaf and Jorion (1990), Ardeni and Lubian (1991) and Glen (1992).

0022-1996/00/$ – see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0022-1996(99)00011-2

uses longer sample periods (100 years or more) that imbue the tests with greater power to reject the null hypothesis of a unit root.

What economic theory does long-run PPP test for? The question is whether relative price movements (for example, between traded and non-traded goods) are important in determining equilibrium real exchange rate movements. Suppose we write the log of the real exchange rate, q_t, as the sum of two components:

$$q_t = x_t + y_t. \tag{1}$$

To arrive at this decomposition, define the real exchange rate as the relative price of foreign goods:

$$q_t = s_t + p_t^* - p_t, \tag{2}$$

where p_t is the log of the home country price level, p_t^* is the log of the foreign price level and s_t is the log of the nominal exchange rate. Assume that domestic and foreign prices are weighted averages of traded and non-traded goods:

$$p_t = (1 - \pi)p_t^{\mathrm{T}} + \pi p_t^{\mathrm{N}}$$

and

$$p_t^* = (1 - \lambda)p_t^{\mathrm{T}*} + \lambda p_t^{\mathrm{N}*}.$$

Then, x_t and y_t are defined by:

$$x_t = s_t + p_t^{\mathrm{T}*} - p_t^{\mathrm{T}} \tag{3}$$

and

$$y_t = \lambda(p_t^{\mathrm{N}*} - p_t^{\mathrm{T}*}) - \pi(p_t^{\mathrm{N}} - p_t^{\mathrm{T}}). \tag{4}$$

Almost any theory of international price determination implies that deviations from the law of one price for traded goods (x_t) are stationary.[2] So, if the real exchange rate is non-stationary, it must be because y_t, the relative relative price of non-traded goods, is non-stationary. Under the null in tests of long-run PPP (H_0: q_t has a unit root), the relative price of non-traded goods has a unit root and determines movements of the real exchange rate in the long run. It is this hypothesis that has been rejected by unit root tests of the real exchange rate on long time series.[3]

However, this recent literature may have reached the wrong conclusion. Cochrane (1991), Blough (1992) and Faust (1996) contend that there is always a non-stationary representation for a time series that is arbitrarily close to any

[2]This is true even in models with transportation costs, such as Obstfeld and Taylor (1997).

[3]Unit root tests for y_t using long time series are not possible because there are not reliable long time series on these disaggregated prices.

stationary representation. Let x_t be a stationary random variable, while y_t is a random walk. Cochrane and Blough demonstrate that we cannot distinguish between the case in which y_t has an arbitrarily small innovation variance (in which case, q_t is non-stationary), and the case of a zero innovation variance for y_t (i.e., y_t is constant, in which case q_t is stationary.) A test which rejects a unit root in q_t with frequency χ when y_t is constant (so χ is the power of the test) will reject a unit root in q_t with the same frequency for the case in which y_t has an arbitrarily small innovation variance (so χ is the size of the test.) Any test with large power to reject the unit root null when y_t is constant must have a large size against the null that y_t has a very small innovation variance. This paper explores the practical implications of this line of thought for studies of PPP. This paper is an application of the Cochrane–Blough–Faust econometric theory.

To assess the significance of the potential size bias in tests of PPP, we obtain data for disaggregated components of prices over a relatively short horizon (25 years). From that data, we can indeed identify one component, (x_t), the relative price of traded goods across countries, that should be stationary on theoretical grounds. Theory does not preclude that the other component (y_t), which involves the relative price of traded to non-traded goods, is non-stationary. In our 25-year time series, y_t appears more persistent than x_t, and we cannot reject that it has a unit root. This component has a much smaller innovation variance than x_t. We use parameters estimated from the 25-year time series of dissaggregated prices to simulate the behavior of the real exchange rate in 100-year samples.

We find that the size bias in unit root (and cointegration) tests is large, even when the unit root component, y_t, accounts for a large proportion of the conditional variance of real exchange rate at the 100-year horizon. Since our artificial data is calibrated to match the behavior of actual price data, it appears that unit root tests might routinely reject the null hypothesis even when there is a large permanent component based on the relative price of non-traded goods.

Recently, tests which have a null hypothesis of stationarity have been developed.[4] Researchers frequently have taken the position that if one simultaneously rejects a unit root and fails to reject stationarity, there is strong and mutually reinforcing evidence that the series being tested is stationary.[5] Here we show that under the same circumstances in which the unit root tests have large size biases, the stationarity tests have very low power. One is quite likely to reject a unit root and fail to reject stationarity, even though there is a large unit root component embodied in the series.

Section 2 briefly reviews the econometric issues, focusing on the intuition of why standard unit root and cointegration tests may have large size biases. Section

[4]For example, Kwiatkowski et al. (1992) – henceforth referred to as the KPSS test.

[5]See, for example, Chen and Tran (1994) and Cheung and Chinn (1997); or Fischer and Park (1991) for the converse.

3 proposes a model for the components of the real exchange rate, and estimates the parameters of the model on data from the U.S. and U.K. for the 1970–1995 time period.

In Section 4, we construct artificial 100-year time series using the model estimated in Section 3 (Note that the actual time series for the disaggregated components of the real exchange rate are not available prior to 1970.) Using Monte Carlo simulations, we find the true size of several tests for long-run PPP that have nominal sizes of 5%. We examine the Augmented Dickey–Fuller (ADF) test, and a recent test developed by Perron and Ng (1996) (PN) for unit roots. We also simulate the behavior of the single-equation Error Correction Model (ECM) test, and the Horvath and Watson (1995) (HW) test for no cointegration between the nominal exchange rate and relative price levels.

We acknowledge that the behavior of exchange rates and prices in our 25-year sample may be different than the behavior over the past 100 years as a whole. The variance of nominal exchange rates may be different, and price-setting behavior may have changed, hence altering the persistence of the stationary component of the real exchange rate. There may also be measurement error in x_t and y_t which causes their relative variability to be mismeasured. Hence, we use the time series constructed from the model estimated in Section 3 only as a benchmark. We perform Monte Carlo exercises for a wide variety of parameters. Indeed, one exercise calibrates parameters to actual 100-year data on nominal exchange rates and (non-disaggregated) price indexes.

Section 5 performs parallel Monte Carlo exercises on the KPSS test of stationarity. Here, since our artificial data contain a unit root, we are interested in assessing the power of the KPSS test. Section 6 concludes with a discussion of the implications of the size bias in tests for long-run PPP.

2. Econometric issues

Schwert (1987, 1989), Christiano and Eichenbaum (1990), Cochrane (1991), Blough (1992), Faust (1996) and others have discussed the size bias of unit root tests. The purpose of this section is to give an intuitive synopsis of that literature in terms of a simple example. This review is not comprehensive, but does illustrate the main issues we are concerned with.

Assume the stochastic processes for y_t and x_t are given by:

$$y_{t+1} = y_t + w_{t+1}, \tag{5}$$

$$x_{t+1} = \phi x_t + m_{t+1}, \quad -1 < \phi < 1, \tag{6}$$

where w_{t+1} and m_{t+1} are mean zero, i.i.d., serially uncorrelated but contemporaneously correlated random variables. Then, the univariate ARMA representation for Δq_t $(\equiv q_t - q_{t-1})$ is

$$\Delta q_{t+1} = \phi \Delta q_t + \zeta_{t+1} + \mu \zeta_t, \tag{7}$$

where ζ_t is a mean zero, serially uncorrelated random variable, and

$$\mu = -\frac{1+(1+\phi^2)S^2+(1+\phi)SR-0.5\sqrt{(1-\phi^2)^2S^4+4(1-\phi)^2S^2+4(1-\phi^2)(1-\phi)S^3R}}{1+\phi S^2+(1+\phi)RS}. \tag{8}$$

Here, S is the ratio of the variance of w_t to the variance of m_t, σ_w^2/σ_m^2; and R is the correlation between m_t and w_t, $\sigma_{mw}/\sigma_m\sigma_w$.

From Eq. (8), as S goes to zero (i.e., when the unit root component gets very small), μ goes to -1. When the absolute value of the moving average component is close to unity, the number of lags needed in an AR process to obtain a good approximation for Δq_t would need to be very large. The critical values for the ADF test are constructed under the assumption that under the null we can do a good job representing Δq_t with a relatively low order autoregressive process. Apparently this criteria is not satisfied when S is close to zero, so the size of the ADF test is not correct when applied to test the null of a unit root in a series such as q_t from Eq. (8).

While the Phillips and Perron (1989) test ostensibly handles the case of moving-average components, Schwert (1989) finds in Monte Carlo simulations that when the absolute value of the moving-average coefficient is close to unity, there are large size biases in that test (as well as in the Dickey–Fuller tests.) The problem appears to be a related one. For the Phillips–Perron test, one must estimate the asymptotic variance of the sample mean of the regression error form an ADF regression. Standard methods of estimating this variance (for example, Newey and West, 1987) provide an accurate estimate only when autocorrelations in the regression error fall to zero relatively quickly. But, when μ is close to one in absolute value, the autocorrelations die out very slowly. So, again, when there is a small unit root component to q_t, there is significant size bias in the Phillips–Perron test.

Perron and Ng (1996) address this size bias in detail. They demonstrate that standard kernel estimators of the asymptotic variance exacerbate the size problem. They offer suggestions of alternative estimates that will reduce the size bias. We will consider their alternative tests.

3. A decomposition of the real exchange rate

The relative price of traded goods, x_t, is likely to be a stationary random variable. If all goods in the traded goods price indexes have the same weights at home and abroad, then changes in x_t occur only because of deviations from the law of one price. Although there is considerable evidence that deviations from the

law of one price can be large and persistent (see, for example, Engel, 1993; Rogers and Jenkins, 1995; Wei and Parsley, 1995), they are almost certainly stationary. Goods arbitrage, even in the presence of transportation costs as in Obstfeld and Taylor (1997), rules out the possibility that these deviations could become unbounded, and thus precludes a unit root in x_t. While Engel (1999) presents evidence that almost all movements in the real exchange rate are attributable to the x_t component, the time series in that study (as in this one) are too short to draw any inference about the importance of each component in the long run.

Permanent shocks to productivity could impart a non-stationary component to the relative price of non-traded to traded goods, thus y_t could have a unit root. Early influential work that emphasized this approach includes Balassa (1964) and Samuelson (1964). Cross-sectional studies of prices show that there can be very large differences in non-traded goods prices across countries.[6]

Engel (1999) uses the decomposition in Eq. (1) to separate real exchange rate changes for the U.S. into their x_t and y_t components. That study uses a variety of price indexes for which data is available on sub-components that can be identified as traded and non-traded goods. Here we pay attention to one measure – the GDP deflator for personal consumption expenditures for the U.S. and the U.K. We choose to examine these series because there are 100-year-long annual series for both countries (used by Rogers, 1995) for the personal consumption deflators and the nominal exchange rate and shorter time series on the disaggregated data.

In this section, we make use of quarterly data on sub-categories of the personal consumption deflator for the years 1970 to 1995.[7] The sub-index for the deflator for personal consumption of commodities in each country is used as the price index for traded goods, and the deflator for personal consumption of services is used as the price index for non-traded goods. This is the assignment used by Engel (1999) and Stockman and Tesar (1995), although clearly these categories are not precise classifications of traded and non-traded goods. We shall attempt to deal with some of the measurement problems in Section 4.

We begin by performing the usual battery of tests for unit roots and cointegration on the 1970–1995 time series.[8] Table 1 summarizes the results. There is ambiguity in the tests for a unit root in q_t. We cannot reject (at the 5% level) the null of a unit root in q_t using the ADF test, but do reject with the Perron–Ng test.

We proceed to test for no cointegration of s_t and $p_t - p_t^*$.[9] We consider two models of cointegration. The first is the single equation Error-Correction Model

[6] See Rogoff (1996) for an illustration using data from the Penn World Tables.

[7] All of the 25-year data was obtained from Datastream. The weights in the price indexes are constant, constructed from the average share of personal consumption expenditures devoted to expenditures on commodities versus services.

[8] The tests are described in detail in Appendix A.

[9] Baillie and Selover (1987), Edison (1987), Taylor (1988), Mark (1990), Patel (1990), Kim (1990), Cheung and Lai (1993) and Edison et al. (1997) are some of the studies that test PPP using cointegration techniques.

Table 1
Unit root tests and cointegration tests on disaggregated data

	Test statistic	5% critical value
Tests on s_t and $p_t - p_t^*$ ($q_t \equiv s_t + p_t^* - p_t$)		
Augmented Dickey–Fuller	−2.80	−2.89
Perron–Ng	−25.75	−14.0
Error-Correction Model	−2.63	−2.81
Horvath–Watson	7.92	10.18
Tests on s_t and $p_t^T - p_t^{T*}$ ($x_t \equiv s_t + p_t^{T*} - p_t^T$)		
Augmented Dickey–Fuller	−2.82	−2.89
Perron–Ng	−32.26	−14.0
Error-Correction Model	−2.65	−2.81
Horvath–Watson	6.39	10.18
Tests on y_t		
Augmented Dickey–Fuller	−1.73	−2.89
Perron–Ng	−2.68	−14.0

(ECM) of Kremers et al. (1992). The single-equation methodology incorporates in the alternative hypothesis the assumption that $p_t - p_t^*$ is weakly exogenous for the cointegration parameters. We impose the null that the cointegrating vector for s_t and $p_t - p_t^*$ is $(1, -1)$ following the procedures in Zivot (1995). The Horvath–Watson (HW) test for no cointegration is a two-equation test that also imposes the cointegrating vector, but does not assume weak exogeneity. Table 1 shows that we fail to reject the null hypothesis of no cointegration at the 5% level using either the ECM or HW test.

Three of our four tests using the 25-year sample fail to reject the null that long-run PPP does not hold. This is not too surprising. The motivation for the studies which use very long series on prices and exchange rates was that tests for unit roots (and cointegration) have little power in time series as short as 25 years. Indeed, we can reject a unit root in the real exchange rate with 100 years of data that is comparable to our shorter series.[10] We note that even in the shorter series, we reject a unit root using the Perron–Ng test.

We also perform unit root tests on the components of the real exchange rate. Table 1 indicates that we fail to reject a unit root at the 5% level for x_t using the ADF test, but again reject the unit root null with the Perron–Ng test. Both the ECM and HW tests fail to reject the null of no cointegration of s_t with $p_t^{\mathrm{T}} - p_t^{\mathrm{T}*}$.

Finally, we fail to reject a unit root in y_t using either the ADF test or the Perron–Ng test. Note that the pattern of findings from the Perron–Ng tests are seemingly inconsistent. While $q_t = x_t + y_t$, we reject the null of a unit root for q_t and x_t, but fail to reject for y_t. We cannot have q_t and x_t stationary, but y_t

[10]The ADF test statistic is −2.94, which is significant at the 5% level. We get a Perron–Ng statistic of −17.23, which is also significant at the 5% level.

non-stationary. Perhaps this is an example of the size bias in unit root tests of q_t that is the subject of this paper.

We have argued that it is likely that x_t is stationary, while y_t may have a unit root. That is the conclusion one could reach from the Perron–Ng tests, though we fail to reject a unit root in either series with 25 years of data using the ADF test. Some further support for this view comes from the estimated degree of persistence of the two variables. For example, the estimate of the coefficient on lagged x_t in a first-order autoregression is 0.8721. For y_t, the corresponding estimate is 0.9668. While the estimates are not statistically significantly different, they are consistent with the view that y_t is more persistent and more likely to have a unit root than x_t.

Further support for that view comes from Fig. 1. This figure plots the variance ratio statistics, as in Cochrane (1988), for q_t, x_t and y_t.[11] So, for x_t (and likewise for q_t and y_t), the figure plots $\mathrm{Var}(x_{t+k}-x_t)/\mathrm{Var}(x_{t+1}-x_t)$ for horizons of $k=1$ to $k=75$. If a series follows a random walk, the (population) variance of the k-difference in that series will be k times the variance of the first difference. For a stationary series, the variance ratio approaches a limit. From Fig. 1 we can see that the variance ratio statistics for y_t rise more steeply than for x_t or q_t, indicating

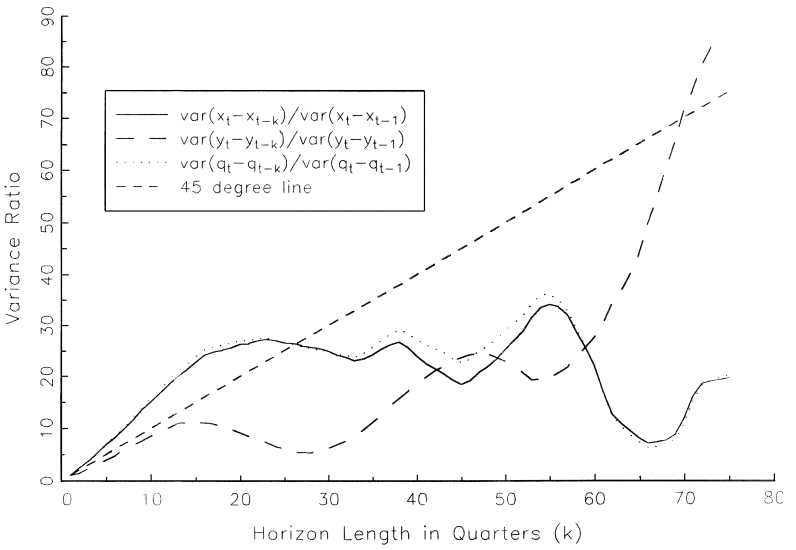


Fig. 1. U.S./U.K. real exchange rate variance component ratios quarterly deflator data from 1970:1 to 1994:4.

[11]These variance ratios were calculated using Cochrane's formula, which corrects for small-sample bias.

more persistence.[12] It is also noteworthy that the variance ratios for x_t and q_t are nearly identical, particularly at the shorter horizons, which is a reflection of the small contribution that y_t makes to movements in q_t over the short and medium run.

Of course, we could never establish that y_t is non-stationary precisely because of near-observational equivalence. For any non-stationary model we choose, there is a stationary one that is arbitrarily close. Our purpose here is to show that if y_t is non-stationary and generated by the model we calibrate below, then unit root tests give misleading answers.

We will calibrate an example model. Simulations from this example model will be used to judge whether standard unit root and cointegration tests can rule out an economically significant permanent component arising from the relative price of non-traded goods. These findings lead us to model y_t as non-stationary and x_t as stationary. We assume y_t follows a simple random walk:[13]

$$y_{t+1} - y_t = au_{t+1}, \tag{9}$$

where u_t is an i.i.d., N(0,1) random variable. This equation determines the movements of the relative prices of non-traded to traded goods, so u_{t+1} incorporates shocks to tastes and technologies that cause this relative price to change permanently.

Define z_t to be the relative price levels, unadjusted for the exchange rate (so, $x_t = s_t - z_t$):

$$z_t \equiv p_t^{\mathrm{T}} - p_t^{\mathrm{T}*}. \tag{10}$$

Note, we will need a model of z_t in our simulations, rather than just y_t and x_t, because we examine tests of no cointegration of $p_t - p_t^*$, which equals $y_t - z_t$, and s_t.

We posit a simple error-correction model for z_t and s_t, so that these two variables are cointegrated with cointegrating vector $(1, -1)$:

$$s_{t+1} - s_t = -\delta(s_t - z_t) + bu_{t+1} + cv_{t+1}, \tag{11}$$

$$z_{t+1} - z_t = \gamma(s_t - z_t) + d\varepsilon_{t+1} + fv_{t+1} + gu_{t+1}, \tag{12}$$

where v_t and ε_t are also i.i.d., N(0,1) random variables.

[12] At the very long horizons ($k > 75$) all of the statistics drop off significantly, but these statistics should probably be ignored since they are calculated using very few data points. So, they are not included in Fig. 1.

[13] We suppress the intercept terms in the presentation of the model for expositional clarity. Intercept terms were included when the model was estimated (following the most common practice in the PPP literature), although all of them were insignificantly different from zero. The artificial data created for the Monte Carlo exercises in Section 4 does not include intercept terms.

The nominal exchange rate in the model is affected by the u_t shock, as well as a monetary shock, v_t. The monetary shock does not affect the relative price, y_t, but is incorporated in the nominal exchange rate. z_t is a nominal variable whose units are the same as the nominal exchange rate. u_t and v_t affect z_t, as well as ε_t, which is a source of shocks to the PPP relationship in traded goods prices. It might represent shocks to the degree of market segmentation.

Eqs. (11) and (12) imply that the relative price of traded goods, x_t, is stationary and follows an AR(1) process:

$$x_{t+1} = \rho x_t - d\varepsilon_{t+1} + (c - f)v_{t+1} + (b - g)u_{t+1}, \tag{13}$$

where $\rho \equiv 1 - \delta - \gamma$.

The system (9), (11) and (12) is estimated by an iterative GLS procedure.[14] The coefficient estimates are reported in Table 2. Standard errors are constructed from the inverse of the estimated information matrix. Among the coefficients on the random errors, $\hat{c}$ is nearly five times as large as the next largest coefficient. Nominal exchange rates are much more variable than nominal prices. Then, $\hat{d}$, from the equation for z_t is next largest, while $\hat{a}$ and $\hat{g}$ are not much smaller. The other coefficients on the random errors – $\hat{b}$ and $\hat{f}$ – are nearly zero. The implications of these estimates are, first, that nominal prices are much less variable than the nominal exchange rate ($\hat{d}$ and $\hat{a}$ are much smaller than $\hat{c}$.) Second, there is almost no correlation between innovations in the nominal exchange rate and nominal prices ($\hat{b}$ and $\hat{f}$ are near zero.) Third, there is some correlation between shocks to the two terms involving only nominal prices, y_t and z_t ($\hat{g}$ is non-zero.) This latter correlation implies that innovations to the permanent component (y_t) and the transitory component (x_t) of the real exchange rate are correlated in our simulations below.

Also note that the estimate of γ is actually negative, but not significantly different than zero. The fact that it is near zero helps justify our assumption above

Table 2
Estimates of coefficients from Eqs. (9), (11) and (12)

Coefficient	Estimate	Standard error
δ	0.080382	0.03812
γ	−0.003415	0.00849
a	0.005725	0.00038
b	0.001088	0.00509
c	0.050770	0.00361
g	0.006109	0.00113
f	0.000632	0.00113
d	0.011286	0.00085

[14]The six parameters a, b, c, d, f and g are exactly identified from the six parameters of the covariance matrix for u, v and ε.

in the ECM test that prices are weakly exogenous. The persistence of x_t, given by $\hat{\rho} = 0.92303$, is high, but this is lower than the usual measure for q_t. The estimated half-life for x_t is about nine quarters.

None of these results is surprising in light of Engel (1999). That study does not estimate a formal model of exchange rates and prices, but does decompose the mean-squared error of $q_{t+j} - q_t$, for $j = 1,2,\ldots,100$ (using quarterly data), into x_t's share and y_t's share. Generally, over 90% of the mean-squared error of $q_{t+j} - q_t$ is attributable to the MSE of $x_{t+j} - x_t$, even when j is very large. This is compatible with the model of Eqs. (9), (11) and (12), given the coefficient estimates.[15] The model does imply that for j large enough, the variance of $q_{t+j} - q_t$ must be dominated by the variance of $y_{t+j} - y_t$, since y_t is the unit root component. But, 25 years is too short a time span for the y_t component to dominate, because the innovation variance of x_t is much larger than that of y_t. The innovation variance of x_t is $(c-f)^2 + (b-g)^2 + d^2$, which is estimated to be 0.002667, as compared to 0.0000328, which is the estimate of a^2, the innovation variance of y_t. Also note that x_t is highly persistent, so that the effects of large shocks persist over a substantial portion of the 25-year period.

4. Monte Carlo measurements of size of tests for long-run PPP

In this section, we report the results of Monte Carlo exercises to measure the true size of the ADF, PN, ECM and HW tests for long-run PPP. The details of the Monte Carlo experiments are in Appendix A.

The baseline case we consider is conditional on the parameter estimates from the model of Section 3, and assumes that y_t follows a random walk. Each artificial series we create has 400 data points, which corresponds to a 100-year sample (because our parameters are estimated on quarterly data). This sample size roughly corresponds to those in several recent studies of long-run PPP.[16] For each artificial series, we perform all four tests. We create 5000 artificial series, and record in each case whether we reject the unit root null with a 5% test.[17]

For each set of parameters, we also calculate the fraction of the forecast variance at an horizon of 400 quarters, $\mathrm{Var}_t(q_{t+400})$, that should theoretically be attributed to the unit root component. While there is some ambiguity about this decomposition in general because y_t and x_t are correlated, in practice the

[15] For the U.S./U.K. data, the x_t component accounts for over 90% of the mean-squared error of q_t for all horizons out to 60 quarters. Beyond 60 quarters, the share attributable to the x_t component drops as low as 75%, but the statistics at the longer horizons are calculated using very few data points.

[16] For example, Frankel (1986), Edison (1987), Edison and Klovland (1987), Kim (1990), Abuaf and Jorion (1990), Ardeni and Lubian (1991), Glen (1992) and Cheung and Lai (1994).

[17] A constant term but no time trend is included in all of the regressions, following common practice in the literature.

correlation is small enough that the decomposition is not very dependent on how the correlation is treated. We report the calculation for:

$$\frac{\mathrm{Var}_t(y_{t+400})}{\mathrm{Var}_t(x_{t+400}) + \mathrm{Var}_t(y_{t+400})} = \frac{400a^2}{[(1-\rho^{800})/(1-\rho^2)]\,[(b-g)^2 + (c-f)^2 + d^2] + 400a^2}. \quad (14)$$

This is meant as a summary statistic for the seven parameters: a, b, c, d, f, g and ρ. It shows for each set of parameters how important the y_t component is in explaining the real exchange rate at the 100-year horizon. One should not be misled by this summary statistic. As was emphasized at the end of Section 3, the y_t component may account for a very small fraction of the variance of q_t at shorter horizons, but it will become more important at the longer horizons. For example, for the baseline parameters, at the 100-year horizon, y_t accounts for 42.12% of the variance of q_t.[18] But at a 1-year horizon, y_t accounts for 1.51% of the variance of q_t; at 5 years, 3.65%; at 10 years, 6.79%; and at 25 years, 15.39%. The reason for focusing on the 100-year horizon is that we are interested in whether y_t is important in long-run movements of the real exchange rate.

4.1. Alternative approaches

Before moving to the results of the simulations, it is useful to ask why we simulate this model to use in assessing the potential size bias in tests of PPP. Why not, for example, undertake an exercise similar to Schwert (1987)? Schwert estimates univariate difference-stationary ARMA models for several economic series (though not for the real exchange rate). He then asks, if the series were really determined by his ARMA models (hence, were really non-stationary), what would be the size of standard unit root tests applied to such series? He interpolates from the Monte Carlo exercises of Schwert (1989) to assess the size bias.

Note that our objective is different from Schwert's, who is only interested in the amount of the size bias if the variables did indeed follow his estimated ARMA process. We know from Cochrane (1991) and Blough (1992) that any unit root test with good power when there is no unit root will have large size biases when there is a small unit root component. Our main objective is to assess whether the size bias is of practical importance for tests of long-run PPP. We have calibrated our measure of the permanent component to data on traded and non-traded goods prices in order to capture the Balassa–Samuelson effect, and attempt to measure how important that permanent component is at longer horizons.

Even if we did wish to pursue Schwert's approach, there are difficulties in

[18] For reasons discussed below, this probably significantly underestimates the importance of y_t.

applying it in this case. First, it is unlikely we could estimate an ARMA model for the real exchange rate such as Eq. (7). We have seen that the autoregressive parameter for the stationary component of the real exchange rate [ϕ in Eq. (7)] is close to unity. We also find the moving average parameter, μ, is close to -1. The model of Section 3 implies values for these parameters of 0.92 and -0.80, respectively. We have a problem of ''near parameter redundancy''. For example, if $\phi = 1$ and $\mu = -1$, then q_t is really a simple random walk – the parameters ϕ and μ are redundant. Clark (1988) and Christiano and Eichenbaum (1990) demonstrate that in cases such as this, estimates of the ARMA model (7) are very inaccurate. Clark shows, on the basis of Monte Carlo results, that the parameter estimates obtained from the maximum likelihood estimation of (7) are typically biased and have very large variance.

Clark (1988) recommends decomposing the series into permanent and transitory components with a Kalman filter. However, as Christiano and Eichenbaum (1990) emphasize, we cannot obtain such a decomposition without making an assumption about the correlation of the innovations to the permanent and transitory components. Nelson and Plosser (1982) show that the univariate difference-stationary Wold representation for a random variable can be decomposed into permanent and transitory components in an infinite number of ways, depending on that correlation.

Actually, according to the estimates of our model of Section 3, the permanent and transitory components are not very highly correlated. We could not, of course, have estimated that correlation from the aggregate data, but we can infer it from the disaggregated model. Once we know that this correlation is nearly zero, might we then proceed to use only the aggregate data and perform Clark's decomposition which assumes the correlation is zero? Even that approach is problematic. In Section 5 of this paper, we show that the KPSS test for stationarity has very low power for the real exchange rate. But, the KPSS test is in essence a Wald test of the null that the variance of the random walk component in Clark's decomposition is zero. The statement that this test has low power is equivalent to saying that the variance of the unit root component is estimated with a great deal of imprecision. So, we cannot learn much about whether or not there is a significant random walk component to the real exchange rate following this approach.

Still, a disadvantage to our approach is that the disaggregated data we base our model on is only available for 25 years. This is one of the reasons why we extensively investigate the sensitivity of our findings to different parameter values.

4.2. Simulation results

For the parameters estimated in Section 3, the y_t component accounts for 42.12% of the 100-year variance. Yet, using a nominal size of 5%, all four tests almost always reject a unit root (or, equivalently, no cointegration of s_t and $p_t - p_t^*$.) The true size for the ADF test is 0.8978; for the PN test, 0.9552; for the

ECM test, 0.9434; and for the HW test, 0.9606. So, if the data-generating process in Section 3 produced 100 years of quarterly data, we would almost always conclude from it that long-run PPP held, even though it would not. It is striking that the rejection rate for the Perron–Ng test is so high, since that test is specifically designed to correct for size bias in the Phillips and Perron (1988) test when the difference-stationary series has a large negative moving-average component as the real exchange rate would in our simulated data.

The baseline model may not produce a representative 100-year series for a number of reasons. In the remainder of this section, we consider alternative parameterizations to get an idea of the scope of the size biases in tests for long-run PPP. The first two experiments emphasize how the relative importance of the y_t component may have been underestimated in the model of Section 3.

First, we consider whether the nominal exchange rate variability that we estimate from our 1970–1995 sample is representative of the exchange rate variance over the 100-year sample. The dollar/pound rate over the latter period has been extremely volatile, but over the past 100 years there have been periods in which both the nominal and real exchange rates were much more stable. One way to handle this would be to model switches of regime from low volatility to high volatility states, and examine the consequences for the size of the PPP tests.[19] Here, we undertake the simpler exercise of investigating the consequences of different values of the parameter c from Eq. (11). So, we fix all of the other parameters at their values reported in Table 2, but then conduct Monte Carlo exercises for various values of c. The results are reported in Figs. 2a–d.

Fig. 2a shows the results for the Augmented Dickey–Fuller test. It graphs the true size of the test against the fraction of the 100-year variance accounted for by the unit root component. The baseline case is $c = 0.051$. Most of the values of c that we investigate are smaller, allowing for the effect of more quiescent nominal exchange rates. Fig. 2a is striking – the probability of rejecting a unit root remains very high even when the unit root component accounts for a very large fraction of long-run real exchange rate movements. For example, when y_t accounts for 84.1% of the 100-year variance of q_t, the probability of rejecting a unit root using a 5% test is still 55.1%. Figs. 2b–d show the comparable results for the PN, ECM and HW tests, respectively. The ADF test actually has the smallest size bias of the four tests.

Engel (1999) discusses why our measure of the relative price on non-traded goods may understate the variance of y_t. If there is a large non-traded component in x_t (due perhaps to marketing and distribution costs), then the true variance of changes in y_t may be understated by the measure used here. So, we examine the effect of allowing larger values for the parameter a from Eq. (9). These results are reported in Figs. 3a–d.

[19]Engel and Kim (1999) estimate such a model with 100 years of U.S./U.K. real exchange rate data.

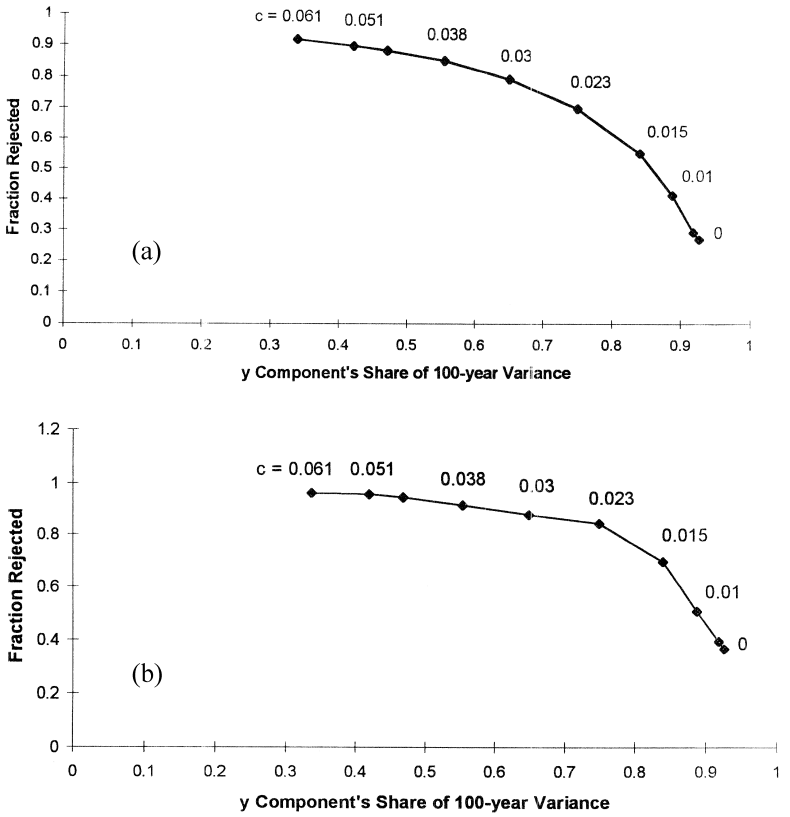
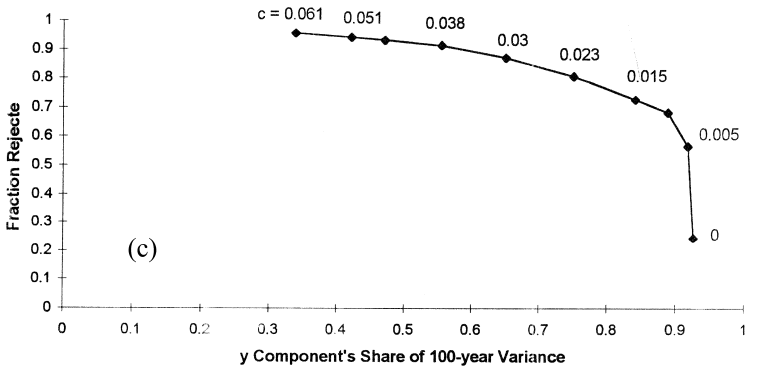


Fig. 2. (a) Augmented Dickey–Fuller test as c varies. (b) Perron–Ng test as c varies. (c) Error-correction model as c varies. (d) Horvath–Watson test as c varies.

As in Fig. 2, there is a trade-off between the fraction of q_t's long-run variance attributed to y_t and the true size of the test. In all cases, the tests appear to have large size biases. For all of the tests, the true size is above 50% even when the y_t component accounts for over 85% of the 100-year variance of the real exchange rate.

When a is set to zero, there is no unit root component in the real exchange rate. Then, the probability of rejection measures the power of the test – the probability of rejecting a unit root when there is none. The Perron–Ng test has the greatest power for this data-generating process, but all four of the tests have impressively high power. The worst of them, the ADF, still has a 95% chance of rejecting the null when the null is false.

Next, we allow for different values of δ from Eq. (11). In our baseline simulations, we set δ equal to 0.077. Figs. 4a–d trace the outcome from Monte



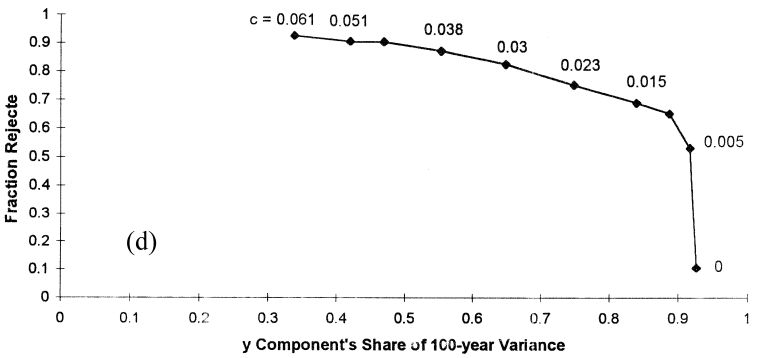


Fig. 2. (*continued*)

Carlo simulations for values of δ ranging from 0.01 to 0.09. For values of δ that are quite small, the tests appear to have less size bias. So, when δ is equal to 0.01, the size of all the tests is around 0.10. However, the reason the tests have smaller size biases in this instance undoubtedly is not that they detect the unit root component, y_t. Certainly, what is occurring is that x_t is very persistent in the case where δ is small. Even if the y_t component were not present, the tests would fail to reject the null of a unit root in x_t (or the null of no cointegration between s_t and z_t.)

Given that we have set γ from Eq. (12) equal to zero (see Appendix A), when δ is also set to zero, both x_t and y_t are random walk processes. The real exchange rate follows a simple random walk. In this case, we find (not surprisingly) that the size of the tests is correct: 0.050 for the ADF; 0.067 for the PN; 0.049 for the ECM; and 0.048 for the HW.

We also consider various values of the parameter d in Eq. (12). However, varying this parameter had little effect on our conclusions about the size of the tests for long-run PPP. Monte Carlo experiments were performed for values of d

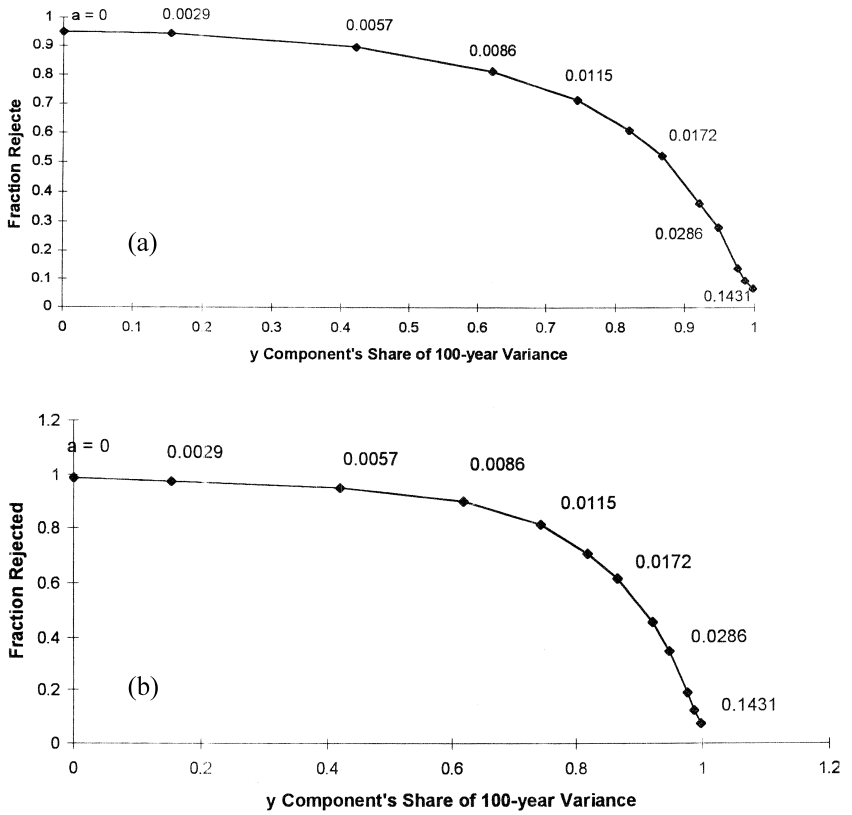


Fig. 3. (a) Augmented Dickey–Fuller test as a varies. (b) Perron–Ng test as a varies. (c) Error-correction model test as a varies. (d) Horvath–Watson test as a varies.

ranging from 0 to 15 times the baseline value, and the true size of all tests was at least 0.89 when the nominal size is set at 0.05.

Another set of simulations fixes the parameters at the baseline set of parameters, but allows the sample size to vary. As is typical in PPP research, all four tests have fairly low rejection rates with short sample sizes. For example, at 100 quarters, the rejection rates for the ADF, PN, ECM, and HW tests were 0.2606, 0.3958, 0.2548, and 0.2142, respectively. The size bias is smaller at shorter samples. In a sense, the tests have smaller size bias for the ''wrong'' reason. We claim x_t is stationary but highly persistent and has a large innovation variance, and y_t is a random walk with a very small innovation variance. Then, in short samples, the behavior of q_t is dominated by x_t because its innovation variance is so large relative to y_t's. In short samples, unit root tests of x_t have low power because x_t is so persistent. So, in short time series, we do not reject a unit root in q_t because it looks like x_t and x_t

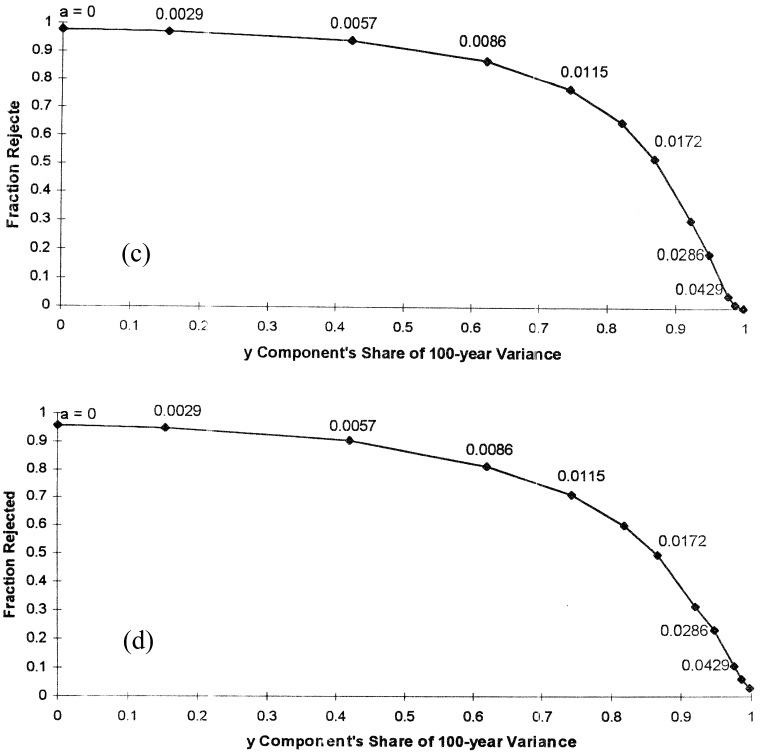


Fig. 3. (*continued*)

looks like a non-stationary process. The size bias is small in short samples not because the tests detect the true random-walk component, y_t, but because the stationary component, x_t, looks non-stationary.

At horizons longer than 100 years, the tests still have large size biases. Two recent studies of PPP use time series longer than 100 years. Lothian and Taylor (1996) use 200 years of wholesale price index data, while Froot et al. (1995) use 700 years of commodity price data. Table 3 reports results from simulations on the baseline model, but allowing the sample size to vary from 50 years (200 quarters) to 200 years (800 quarters). Holding the maximum lag length constant, the size bias is actually larger the larger the sample size in this range. When we allow the maximum lag length in the tests to increase, the size bias begins to diminish. But, as Table 3 indicates, the size bias is still extremely large for 200 years of artificial data, even when the maximum lag length is set at 48 quarters.[20]

[20]Faust (1996) shows that the size of tests whose null is that the variable follows an arbitrary unit root process must go to one as the sample size increases. Here, we are examining the size of tests with a specific null hypothesis.

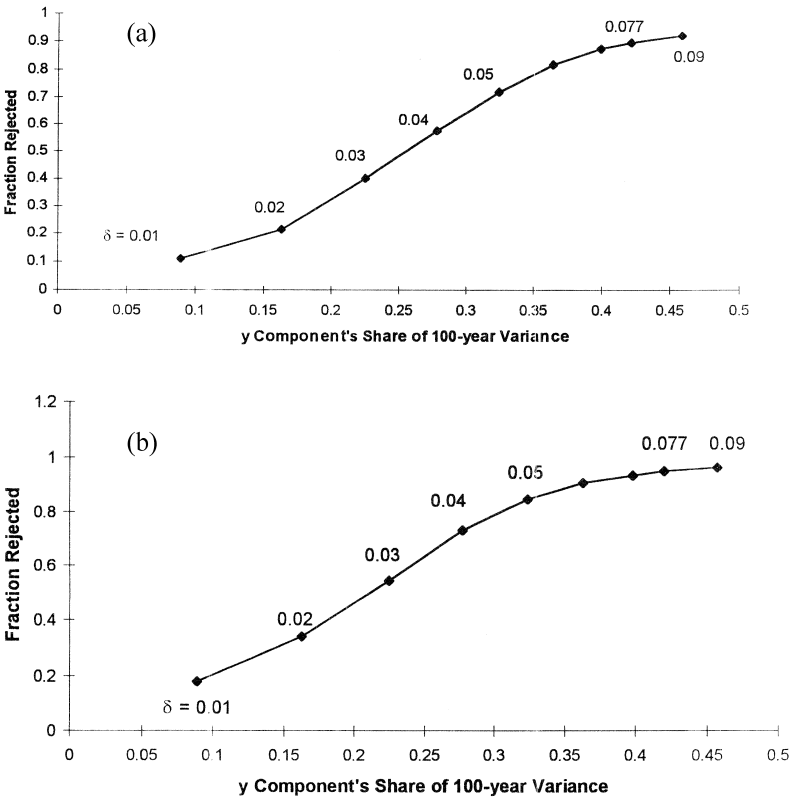


Fig. 4. (a) Augmented Dickey–Fuller test as δ varies. (b) Perron–Ng test as δ varies. (c) Error-correction model test as δ varies. (d) Horvath–Watson test as δ varies.

We do not try to replicate the sample length of Froot, Kim and Rogoff. But, note that the data used by these authors are constructed from a limited number of goods. The price data generally do not incorporate the non-traded goods prices that, according to our model, account for the non-stationary component of the real exchange rate. So, while they convincingly reject a unit root, their tests are more properly considered tests for a unit root in the x_t component.

Finally, we consider some simulations in which the parameters of the model in (9), (11) and (12) are calibrated to a 100-year data sample. This data, from Rogers (1995), consists of the nominal dollar/pound exchange rate, and the personal consumption deflator in the U.S. and U.K. from 1892 to 1992. The description of the data in Rogers and in the original source, Mitchell (1988), is sparse. The exchange rate is described as a period average, and the prices appear to be end-of-period data. All of the data is annual. In this 100-year data, we noted above that we reject the null of a unit root in the real exchange rate at the 5% level using the ADF and PN tests.

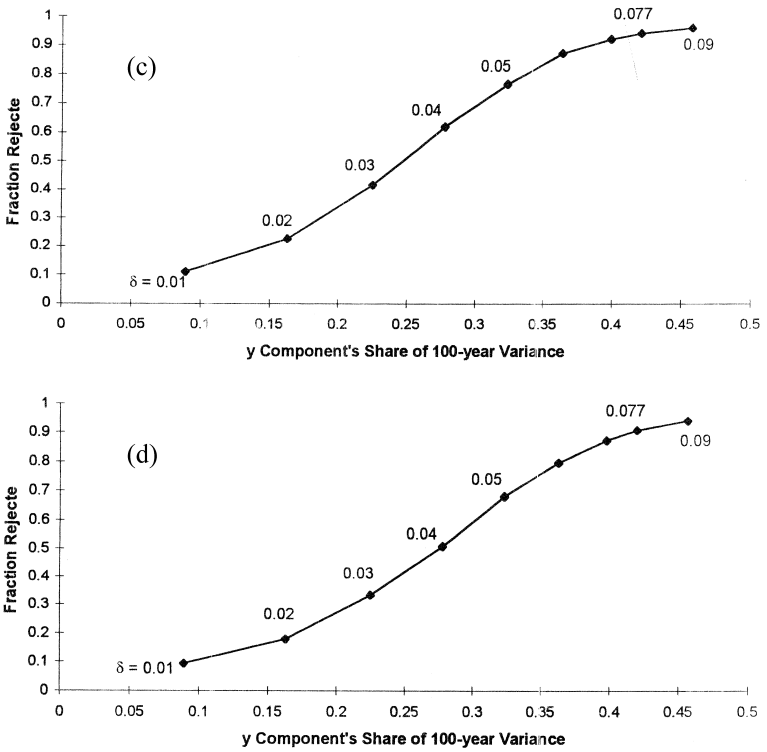


Fig. 4. (*continued*)

Note that this data is not disaggregated data. It cannot be used to construct the x_t and y_t components of the real exchange rate. We choose parameter values in the model to match three moments in the data – the first-order autocorrelation coefficient of the real exchange rate, and the variances of the annual difference in the relative price levels and the 3-year difference in the nominal exchange rate.[21] Appendix A describes how we create artificial series whose moments match those from the true data.

Table 4 shows the sets of parameter values we use to perform our analysis of the size of the long-run PPP tests. As it turns out, the value of the parameter c is about the same across all of the sets, but there is quite a bit of variation in the other parameter values. As the table shows, the unit root component, y_t, accounts

[21]During the periods of ''fixed'' nominal exchange rates, the occasional devaluations yield distant outliers in the 1-year changes in exchange rates. The variance of the 1-year change in these periods is quite large – generally larger than for floating rate periods. Taking the variance of 3-year changes gives a more reasonable picture of the amount of exchange rate volatility. Of course, as we mention above, it would be best to model these jumps explicitly.

Table 3
Size of tests as sample size and maximum lag length vary[a]

Sample size	Maximum lags			
	12	24	36	48
Size of Augmented Dickey–Fuller test				
200	0.5524	0.4304	0.3094	0.2376
400	0.8978	0.7470	0.5706	0.4366
600	0.9700	0.8832	0.7252	0.5732
800	0.9892	0.9266	0.8114	0.6724
Size of Perron–Ng test				
200	0.6726	0.6504	0.5684	0.4064
400	0.9552	0.8718	0.7830	0.7028
600	0.9860	0.9336	0.8512	0.7720
800	0.9932	0.9638	0.8840	0.7980
Sample size	Maximum lags			
	8	16	24	32
Size of single-equation ECM test				
200	0.5920	0.5126	0.3914	0.3174
400	0.9434	0.8486	0.7262	0.5984
600	0.9892	0.9526	0.8758	0.7714
800	0.9966	0.9788	0.9324	0.8450
Size of Horvath–Watson test				
200	0.4884	0.4006	0.2990	0.2370
400	0.9066	0.7868	0.6528	0.5042
600	0.9802	0.9300	0.8208	0.6954
800	0.9942	0.9678	0.9088	0.8108

[a] Numbers reported are rejection probabilities. ''Sample size'' refers to the length of the sample in quarters.

for as little as 0.9% of the 100-year variance of q_t to as much as 89.4%, depending on the particular parameter values. The true size of the tests is nonetheless fairly consistent across all the sets of parameter values and all of the tests. As shown in Table 4, the size ranges only from 0.299 to 0.486. So, for all parameter values, there is considerable size bias in all four tests.

Engel and Kim (1999) estimate an unobserved components model for the U.S./U.K. real exchange rate based on producer price indexes over approximately this same 100-year time span. They allow the transitory component to switch among three regimes – ones of low, medium and high variance. They find the permanent component can be adequately described by a single regime. This model is meant to capture the changing volatility in nominal exchange rates over the past 100 years. Monte Carlo tests performed on data generated from this model yield size values for the unit root and cointegration tests which are very close to those reported in Table 4.

Table 4
Size of long-run PPP tests with parameters calibrated to long-run data

	Specification							
	1	2	3	4	5	6	7	8
a[a]	0.0010	0.0030	0.0060	0.0090	0.0120	0.0150	0.0180	0.0210
d[a]	0.0189	0.0193	0.0195	0.0193	0.0186	0.0174	0.0155	0.0126
c[a]	0.0415	0.0415	0.0417	0.0420	0.0427	0.0438	0.0456	0.0484
δ[a]	0.0240	0.0244	0.0257	0.0281	0.0320	0.0382	0.0476	0.0609
var 1[b]	0.0092	0.0773	0.2587	0.4591	0.6289	0.7550	0.8405	0.8941
ADF[c]	0.3078	0.2986	0.3120	0.3152	0.3358	0.3264	0.3478	0.3566
PN[c]	0.4486	0.4638	0.4438	0.4546	0.4452	0.4584	0.4600	0.4318
ECM[c]	0.3886	0.3810	0.3954	0.3726	0.3738	0.3626	0.3432	0.3200
HW[c]	0.3116	0.3002	0.3164	0.2984	0.3096	0.3040	0.3040	0.3020
KPSS[d]	0.0062	0.0046	0.0054	0.0128	0.0138	0.0204	0.0350	0.0502

[a] The letters *a, d, c* and δ refer to parameter values from Eqs. (9), (11) and (12).

[b] ''var 1'' refers to the fraction of the conditional variance of the real exchange rate at a 100-year horizon accounted for by the unit root component for each set of parameter values.

[c] The numbers reported in the rows ADF, PN, ECM and HW are the true size for a test with a nominal size of 5% for the Augmented Dickey–Fuller, Perron–Ng, Error-Correction Model and Horvath–Watson tests, respectively.

[d] The numbers reported in the KPSS row are the power of a KPSS test for stationarity with a size of 5%.

5. Tests of stationarity

We have presented evidence that standard unit root tests have substantial size biases, in the sense that they reject the null hypothesis of a unit root even when a substantial random walk component is present. An alternative approach is to test the null hypothesis that the series is stationary. This is the motivation of Kwiatkowski et al. (1992). That test, in essence, is a test of the null hypothesis that y_t has a zero variance.[22]

We consider, however, whether the KPSS test has much power to reject the null of stationarity in our case. As in Section 4, we perform Monte Carlo simulations of the real exchange rate based on the model of Section 3. However, since the null is stationarity and the artificial series really have a unit root, when we tabulate the fraction of times the null is rejected we are calculating the power of the KPSS test. The details of the Monte Carlo simulation are in Appendix A.

Unfortunately, the KPSS test has very low power to detect the unit root component. For the baseline model, we reject the null only 8.6% of the time. Figs.

[22] We focus on the KPSS test rather than the test of Leybourne and McCabe (1994) because the KPSS test is more widely used in the literature. Cheung and Chinn (1997) report that the Leybourne–McCabe test has considerable size bias in small samples. The size bias is much greater than for the KPSS test. Since we focus on the power of the stationarity test, it is essential that the tests we examine have nearly the correct size.

5a–c show the power of the test for various values of the parameters c, a and δ [from the model of Eqs. (9), (11) and (12)]. Similar results emerge for various values of the parameter d. In addition, we have constructed artificial series whose

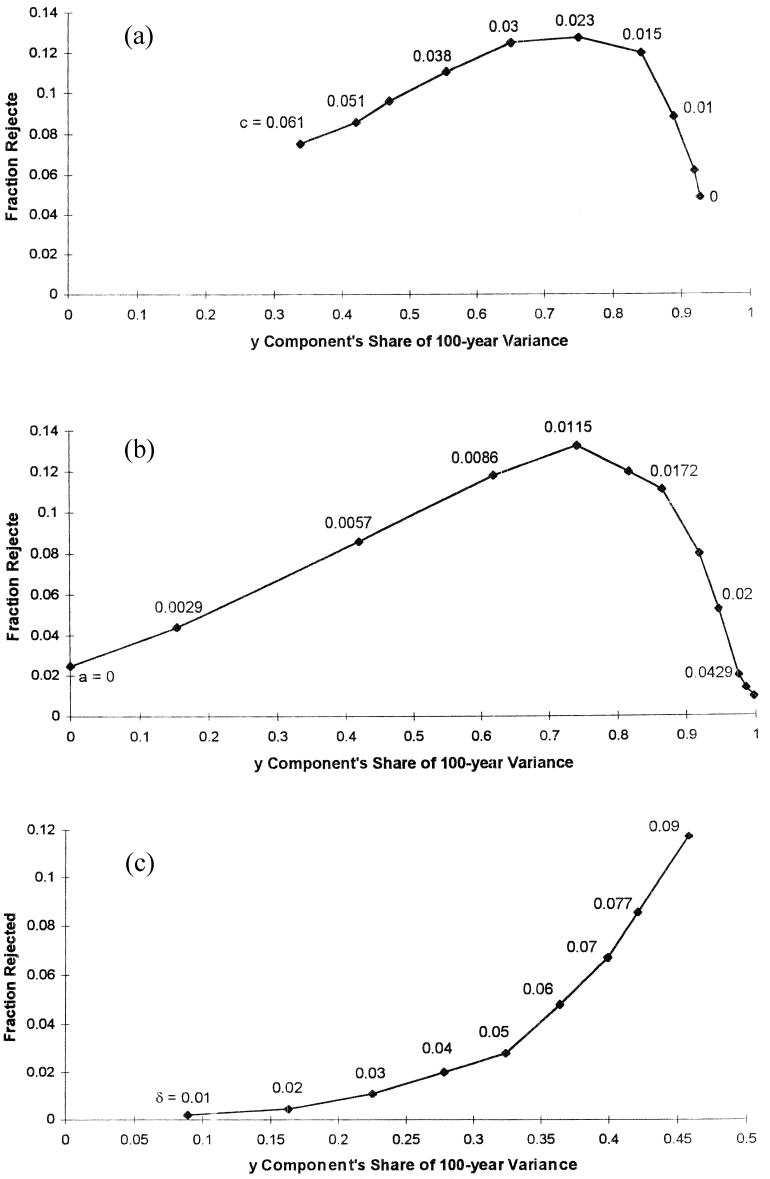


Fig. 5. (a) KPSS test as c varies. (b) KPSS test as a varies. (c) KPSS test as δ varies.

moments match those of our 100-year data (as at the end of Section 4.) These results are reported in the last line of Table 4.

In the best case among all of the parameterizations, the KPSS test rejects the null only 13.3% of the time. In most cases, the power was considerably lower, even though in some parameterizations the unit root component accounts for a large share of the long-run variance.

It is interesting that the KPSS test has such low power. One might be tempted to conclude that since the ADF, PN, ECM and HW tests reject a unit root, and the KPSS test fails to reject stationarity, that there is mutually reinforcing evidence that there is no unit root. But, in fact, all of the tests lead to the wrong inference – there is size bias in the ADF, PN, ECM and HW tests, and the KPSS test has low power.[23]

We have noted that with our 100-year series of U.S./U.K. real exchange rates, we reject a unit root using the ADF and PN tests. Applying the KPSS test to that data, we fail to reject the null hypothesis of stationarity. We calculate the KPSS statistic to be 0.0788, while the 5% critical value is 0.463. The results of this section and the previous one undermine our confidence that these tests rule out a substantial, economically important permanent component to the real exchange rate.

6. Conclusions

We have found that there can be large size biases in tests for long-run PPP. There may be a significant unit root component that is not detected by these tests. We associate that component with the y_t term above, which represents the element corresponding to the relative price of non-traded goods in the real exchange rate.

In this study we do not address the recent panel tests for PPP, such as Frankel and Rose (1996) and Wei and Parsley (1995).[24] It seems likely that these tests would suffer from similar size problems as the ones addressed here. However, note that O'Connell (1998) and Engel et al. (1997) argue that these panel studies have size biases stemming from an entirely different source – failure to control adequately for cross-sectional correlation.

The conclusion in this paper that we cannot rule out an economically significant permanent component to the real exchange rate is consistent with other evidence recently uncovered. Mark (1995) and Chinn and Meese (1995) find that they can forecast the nominal exchange rate at long horizons by predicting that it returns to a target level – but that target level is not the PPP value. Mark and Choi (1997)

[23] DeJong et al. (1992) note that both unit root tests and stationarity tests are likely to have low power. We focus on a somewhat different issue – that stationarity tests may have low power and unit root tests may have large size biases in the same circumstances.

[24] For a more complete list of citations of recent panel studies of PPP, see Engel et al. (1997).

explicitly allow for the target component of the real exchange rate to move over time according to various models of long-run real exchange rate determination. They find that models in which long-run PPP holds are significantly outperformed, in terms of out-of-sample forecasting power, by models that allow the long-run real exchange rate to vary over time. Note that Eq. (12) can be interpreted as a model for the deviation of q_t from y_t (since $x_t = q_t - y_t$), so that our model for the real exchange rate is also one in which the real exchange rate returns to a target level that changes over time.

The cross-section dispersion of aggregate price levels (corrected for nominal exchange rates) that we see across countries can be easily understood in terms of a unit root component to the real exchange rate arising from a Balassa–Samuelson effect. Rogoff (1996) finds enormous variation in price levels across countries using data from Summers and Heston (1991). In some instances, richer countries have price levels that are an order of magnitude higher than the small countries. Consider the following experiment: the cross-sectional variance in the log of relative prices of consumption deflators in the Summers–Heston data is 0.2686. If we fit an AR(1) model to our 25 years of U.S./U.K. real exchange rates, the variance of the quarterly innovations is 0.00263. Suppose each real exchange rate relative to the U.S. from the Summers–Heston panel follows an identical AR(1) with innovation variance of 0.00263. What value of the quarterly autocorrelation coefficient would produce a cross-section variance of 0.2686? The answer is 0.995 (that is, $0.00263/(1 - 0.995^2) = 0.2686$). So, while it is possible that the Summers–Heston data was generated by stationary real exchange rate series, it is more plausible that there is a unit root component in some of the real exchange rates.

The tests for long-run PPP on long data sets that reject a unit root are probably correctly telling us that there is a stationary component to the real exchange rate. So, they confirm that the x_t component – the relative price of traded goods – does indeed converge. It appears, however, that they cannot rule out a non-stationary component.

Acknowledgements

I thank Eric Zivot, Serena Ng, Charles Nelson, Anthony Rodrigues, Jaewoo Lee, and Frank Diebold for helpful discussions. I also thank two referees, and the co-editor, Gregor Smith, for many helpful comments. Participants in seminars at UCLA, UCSD, Rochester, Irvine, USC and Washington made useful remarks. Mike Hendrickson provided excellent research assistance and knowledgeable input. Some of the work for this paper was completed while I was a Visiting Scholar at the Federal Reserve Bank of Kansas City and at the International Monetary Fund. Neither the Federal Reserve System nor the IMF necessarily shares the views expressed in this paper. I also acknowledge assistance from the

National Science Foundation, NSF grant #SBR-932078 to the National Bureau of Economic Research.

Appendix A

The Monte Carlo tests of Section 4 (and the tests on the actual data in Section 3) examine long-run PPP using four different tests.

First is the Augmented Dickey–Fuller (ADF) test. Here we estimate the equation

$$q_t = \alpha + \beta q_{t-1} + \xi_1 \Delta q_{t-1} + \xi_2 \Delta q_{t-2} + \cdots + \xi_j \Delta q_{t-j} + \eta_t. \tag{A.1}$$

The lag length, j, was chosen by an iterative data-based procedure, as recommended by Ng and Perron (1995). We start with a maximum number of lags (12) and test for the significance of ξ_{12}. If it is significantly different from zero, then $j = 12$. Otherwise, we drop the 12th lag, reestimate the regression, and proceed until ξ_j is significantly different from zero. The null hypothesis is $\beta = 1$.

In the test performed in Section 3, we end up choosing a lag length of six. For each iteration of the Monte Carlo procedure of Section 4, we do the iterative procedure to choose the lag length. The number of lags actually chosen varied from zero to 12. Approximately 30–40% of the time (depending on the parameters used to generate the artificial data) a lag length of zero was chosen.

The Perron and Ng (1996) unit root test is a modified version of the Phillips and Perron (1988) Z_α statistic. We first run the regression:

$$q_t = \kappa + \alpha \cdot q_{t-1} + \eta_t. \tag{A.2}$$

Construct Z_α as

$$Z_\alpha = T(\hat{\alpha} - 1) - (s^2 - s_\eta^2)\left(2T^{-2}\sum_{t=1}^{T-1}(q_t - \bar{q})^2\right)^{-1},$$

where T is the number of observations. Formulas for s^2 and s_η^2 are given below.

The modified Phillips–Perron statistic is given by:

$$MZ_\alpha = Z_\alpha + (T/2)(\hat{\alpha} - 1)^2.$$

We use

$$s_\eta^2 = T^{-1}\sum_{t=1}^{T}\hat{\eta}_t^2,$$

and the s_{AR}^2 measure for s^2 recommended by Perron and Ng. s_{AR}^2 is given by:

$$s_{\mathrm{AR}}^2 = s_{ek}^2/(1 - \hat{b}(1))^2,$$

where

$$s_{ek}^2 = T^{-1} \sum_{t=k+1}^{T} \hat{e}_{tk}^2 \text{ and } \hat{b}(1) = \sum_{j=1}^{k} \hat{b}_j,$$

with $\hat{b}_j$ and $\{\hat{e}_{tk}\}$ obtained from the regression:

$$\Delta q_t = b_0 q_{t-1} + \sum_{j=1}^{k} b_j \Delta q_{t-j} + e_{tk}.$$

The lag length, k, is chosen in each iteration of the Monte Carlo by the same iterative procedure described for choosing the lag length in the ADF test above (with a maximum lag length of 12).

The first cointegration test derives from the single-equation Error-Correction Model (ECM) test proposed by Kremers et al. (1992). We estimate the equation

$$\begin{aligned}\Delta s_t = \alpha - \beta q_{t-1} + \theta_0 \Delta(p_t - p_t^*) + \theta_1 \Delta(p_{t-1} - p_{t-1}^*) + \theta_2 \Delta(p_{t-2} - p_{t-2}^*) \\ + \cdots + \theta_j \Delta(p_{t-j} - p_{t-j}^*) + \xi_1 \Delta s_{t-1} + \xi_2 \Delta s_{t-2} + \cdots + \xi_j \Delta s_{t-j} \\ + \eta_t. \qquad \text{(A.3)}\end{aligned}$$

Note that the number of lags of $\Delta(p_t - p_t^*)$ and Δs_t are constrained to be equal. The lag length is chosen by the same type of data-based procedure as in the ADF test. Here, in each iteration, we test the joint null $\theta_j = 0$, $\xi_j = 0$. Again, when we do the Monte Carlo simulations, we use this iterative procedure to choose the lag length for each set of artificial data.

The ECM test is valid in this case if $p_t - p_t^*$ is weakly exogenous for the cointegration parameters. In general, the ECM method allows for estimation of the cointegrating vector, although here we have imposed that it is $(1, -1)$.

The null hypothesis in this case is $\beta = 0$. Following Zivot (1995) and Hansen (1995), the test statistic for the ECM test depends on the asymptotic covariance matrix of the sample means of η_t and $\Delta s_t - \alpha - \beta q_{t-1}$. This matrix is calculated by the method suggested by Andrews and Monahan (1992), using a Bartlett kernel, with the selection rule for the order of the kernel weight function chosen as in Andrews (1991).[25] The critical values are presented in Hansen (1995).

In each iteration of the Monte Carlo, then, we compute the long-run covariance matrix and use it to compare the test statistic to the critical value from the Hansen table.

The second cointegration test is based on the procedure suggested by Horvath and Watson (1995). Here we estimate the system of equations given by:

[25] See Hamilton (1994, pp. 280–285) for a summary of these methods.

$$\begin{aligned}\Delta s_t &= \alpha_1 - \beta_1 q_{t-1} + \theta_{11}\Delta(p_{t-1} - p^*_{t-1}) + \theta_{12}\Delta(p_{t-2} - p^*_{t-2}) \\ &\quad + \cdots + \theta_{1j}\Delta(p_{t-j} - p^*_{t-j}) + \xi_{11}\Delta s_{t-1} + \xi_{12}\Delta s_{t-2} \\ &\quad + \cdots + \xi_{1j}\Delta s_{t-j} + \eta_{1t}, \\ \Delta(p_t - p^*_t) &= \alpha_2 - \beta_2 q_{t-1} + \theta_{21}\Delta(p_{t-1} - p^*_{t-1}) + \theta_{22}\Delta(p_{t-2} - p^*_{t-2}) \\ &\quad + \cdots + \theta_{2j}\Delta(p_{t-j} - p^*_{t-j}) + \xi_{21}\Delta s_{t-1} + \xi_{22}\Delta s_{t-2} \\ &\quad + \cdots + \xi_{2j}\Delta s_{t-j} + \eta_{2t.} \end{aligned} \tag{A.4}$$

Note that we impose that the cointegrating vector is $(1, -1)$. We also impose that the lag length on $\Delta(p_t - p^*_t)$ and Δs_t are the same for both variables in both regressions. We choose the lag length again by an iterative data-based criterion. Here, in each iteration, to choose the lag length we test the joint null that $\theta_{1j} = 0$, $\xi_{1j} = 0$, $\theta_{2j} = 0$, $\xi_{2j} = 0$.

The null hypothesis of long-run PPP is $\beta_1 = 0$ and $\beta_2 = 0$. We compare the test statistic to the critical values reported in Horvath and Watson.

We construct 5000 replications of data series with 400 observations each. In each case, 450 data points were actually computed, and the first 50 were dropped to avoid any bias from start-up values. The start-up values for all of the variables are zero, which is equal to their unconditional mean given that there are no intercept terms included in the simulations. The error terms u_t, v_t and ε_t from Eqs. (9), (11) and (12) are assumed to be N(0,1), and were created using the ''rndn'' command in Gauss version 3.01. For each of these 5000 artificial series, we perform the four tests.

One issue arises in the Horvath–Watson cointegration tests. The estimated value of γ from Eq. (12) is actually negative, but not significantly different from zero. If we construct our artificial data using this negative value, there would be some problems of interpretation with the HW test. Suppose, for example, that $-\gamma = \delta$. Then x_t, and therefore q_t, has a unit root even if y_t were zero. But the HW test would conclude that s_t and $p_t - p^*_t$ (which equals z_t when y_t is zero) are cointegrated, because it would test the joint hypothesis that $\gamma = 0$ and $\delta = 0$. So, in our simulations, we set γ equal to zero, and set our measure of δ equal to the estimated value of $\delta + \gamma$. This leaves the persistence of the x_t component unchanged.

At the end of Section 4, simulations are based on the 100-year data sample. We search for combinations of the four parameters – a, c, d and δ – that produce 100-year artificial series whose moments match those of the data. (The other parameters are kept equal to their values estimated in Section 3.) To construct the 100-year series for the nominal exchange rate, 400 data points were generated and every four were averaged together. This is meant to replicate the construction of the actual data, which is an annual average. 400 data points are produced for the prices as well, and every fourth number is used since the actual price data are

end-of-year data. The Monte Carlo experiments used to choose these values of a, c, d and δ employ 5000 replications of each series.

We use a hill-climbing technique to find these parameters. We begin with a guess at a set of parameter values, construct 5000 artificial time series, and for each replication calculate the first-order autocorrelation coefficient of the real exchange rate, and the variances of the annual difference in the relative price levels and the 3-year difference in the nominal exchange. We calculate the average value of those statistics over the 5000 replications. We adjust the parameters, and construct 5000 new artificial time series, until the average over the 5000 replications of those three statistics reproduce their values in the actual 100-year data.

To construct the KPSS test, we calculate $r_t = q_t - \bar{q}$, where $\bar{q}$ is the sample average of q_t. Then, define $S_t = \Sigma_{i=1}^{t} r_i$. The KPSS statistic is calculated as:

$$\frac{\sum_{t=1}^{T} S_t^2}{[T^2 \hat{\sigma}^2]},$$

where $\hat{\sigma}^2$ is an estimate of the asymptotic variance of the sample mean of r_t. We calculate this variance as we did in calculating the asymptotic variance of η_t for the HW statistic described above. Again, each Monte Carlo experiment involved construction of 5000 artificial data series of length 400.

References

Abuaf, N., Jorion, P., 1990. Purchasing power parity in the long run. Journal of Finance 45, 157–174.

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

Andrews, D.W.K., Monahan, J.C., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. Econometrica 60, 953–966.

Ardeni, P.G., Lubian, D., 1991. Is there trend reversion in purchasing power parity? European Economic Review 35, 1035–1055.

Baillie, R.T., Selover, D.D., 1987. Cointegration and models of exchange rate determination. International Journal of Forecasting 3, 43–51.

Balassa, B., 1964. The purchasing power parity doctrine: a reappraisal. Journal of Political Economy 72, 584–596.

Blough, S.R., 1992. The relationship between power and level for generic unit root tests in finite samples. Journal of Applied Econometrics 7, 295–308.

Chen, B., Tran, K.C., 1994. Are we sure that the real exchange rate follows a random walk? A reexamination. International Economic Journal 8, 33–44.

Cheung, Y.-W., Chinn, M.D., 1997. Further investigation of the uncertain unit root in GNP. Journal of Business and Economic Statistics 15, 68–73.

Cheung, Y.-W., Lai, K., 1993. Long-run purchasing power parity during the recent float. Journal of International Economics 34, 181–192.

Cheung, Y.-W., Lai, K., 1994. Mean reversion in real exchange rates. Economics Letters 46, 251–256.

Chinn, M.D., Meese, R.A., 1995. Banking on currency forecasts: how predictable is change in money? Journal of International Economics 38, 161–178.
Christiano, L.J., Eichenbaum, M., 1990. Unit roots in real GNP: do we know and do we care? Carnegie–Rochester Conference Series on Public Policy 32, 7–61.
Clark, P.K., 1988. Nearly redundant parameters and measures of persistence in economic time series. Journal of Economic Dynamics and Control 12, 447–461.
Cochrane, J.H., 1988. How big is the random walk in GNP? Journal of Political Economy 96, 893–920.
Cochrane, J.H., 1991. A critique of the application of unit root tests. Journal of Economic Dynamics and Control 15, 275–284.
DeJong, D.N., Nankervis, J.C., Savin, N.E., Whiteman, C.H., 1992. Integration versus trend stationarity in time series. Econometrica 60, 423–433.
Edison, H.J., 1987. Purchasing power parity in the long run: a test of the dollar/pound exchange rate (1890–1978). Journal of Money, Credit and Banking 19, 376–387.
Edison, H.J., Klovland, J.T., 1987. A quantitative reassessment of the purchasing power parity hypothesis: some evidence on Norway and the United States. Journal of Applied Econometrics 2, 309–334.
Edison, H.J., Gagnon, J.E., Melick, W.R., 1997. Understanding the empirical literature on purchasing power parity: the post-Bretton Woods era. Journal of International Money and Finance 16, 1–17.
Engel, C., 1993. Real exchange rates and relative prices: an empirical investigation. Journal of Monetary Economics 32, 35–50.
Engel, C., 1999. Accounting for U.S. real exchange rate changes. Journal of Political Economy, forthcoming.
Engel, C., Hendrickson, M.K., Rogers, J.H., 1997. Intra-national, intra-continental and intra-planetary PPP. Journal of the Japanese and International Economies 11, 480–501.
Engel, C., Kim, C.-J., 1999. The long-run U.S./U.K. real exchange rate. Journal of Money, Credit and Banking, forthcoming.
Faust, J., 1996. Near observational equivalence and theoretical size problems with unit root tests. Econometric Theory 12, 724–731.
Fischer, E.O'N., Park, J.Y., 1991. Testing purchasing power parity under the null hypothesis of cointegration. Economic Journal 101, 1476–1484.
Frankel, J.A., 1986. International capital mobility and crowding out in the U.S. economy: imperfect integration of financial markets or goods markets. In: Hafer, R.W. (Ed.), How Open is the U.S. Economy, Lexington Books, Lexington.
Frankel, J.A., Rose, A., 1996. A panel project on purchasing power parity: mean reversion within countries and between countries. Journal of International Economics 40, 209–224.
Froot, K.A., Kim, M.C., Rogoff, K., 1995. The law of one price over 700 years. National Bureau of Economic Research, working paper No. 5132.
Glen, J.D., 1992. Real exchange rates in the short, medium, and long run. Journal of International Economics 33, 147–166.
Hansen, B.E., 1995. Rethinking the univariate approach to unit root testing: using covariates to increase power. Econometric Theory 11, 1148–1171.
Horvath, M.T., Watson, M., 1995. Testing for cointegration when some of the cointegrating vectors are known. Econometric Theory 11, 984–1014.
Kim, Y., 1990. Purchasing power parity in the long run: a cointegration approach. Journal of Money, Credit and Banking 22, 491–503.
Kremers, J.J.M., Ericsson, N.R., Dolado, J.J., 1992. The power of cointegration tests. Oxford Bulletin of Economics and Statistics 54, 325–348.
Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? Journal of Econometrics 54, 159–178.
Leybourne, S.J., McCabe, B.P.M., 1994. A consistent test for a unit root. Journal of Business and Economic Statistics 12, 157–166.

Lothian, J.R., Taylor, M.P., 1996. Real exchange rate behavior: the recent float from the perspective of the past two centuries. Journal of Political Economy 104, 488–509.
Mark, N.C., 1990. Real and nominal exchange rates in the long run. Journal of International Economics 28, 115–136.
Mark, N.C., 1995. Exchange rates and fundamentals: evidence on long-horizon predictability. American Economic Review 85, 201–218.
Mark, N.C., Choi, D.-Y., 1997. Real exchange-rate prediction over long horizons. Journal of International Economics 43, 29–60.
Mitchell, B.R., 1988. British Historical Statistics, Cambridge University Press, Cambridge.
Nelson, C.R., Plosser, C.I., 1982. Trends and random walks in macroeconomic time series: some evidence and implications. Journal of Monetary Economics 10, 139–162.
Newey, W.K., West, K.D., 1987. A simple positive semi-definite, heteroskdeasticity and autocorrleation consistent covariance matrix. Econometrica 55, 703–708.
Ng, S., Perron, P., 1995. Unit root tests in ARMA models with data dependent methods for the truncation lag. Journal of the American Statistical Association 90, 268–281.
Obstfeld, M., Taylor, A., 1997. Nonlinear aspects of goods-market arbitrage and adjustment: Heckscher's commodity points revisited. Journal of the Japanese and International Economies 11, 441–478.
O'Connell, P., 1998. The overvaluation of purchasing power parity. Journal of International Economics 44, 1–19.
Patel, J., 1990. Purchasing power parity as a long-run relation. Journal of Applied Econometrics 5, 367–379.
Perron, P., Ng, S., 1996. Useful modifications to some unit root tests with dependent errors and their local asymptotic properties. Review of Economic Studies 63, 435–463.
Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.
Rogers, J.H., 1995. Real shocks and real exchange rates in really long-term data. Board of Governors of the Federal Reserve.
Rogers, J.H., Jenkins, M.A., 1995. Haircuts or hysteresis: sources of movements in real exchange rates. Journal of International Economics 38, 339–360.
Rogoff, K., 1996. The purchasing power parity puzzle. Journal of Economic Literature 34, 647–668.
Samuelson, P.A., 1964. Theoretical notes on trade problems. Review of Economics and Statistics 46, 145–154.
Schwert, G.W., 1987. Effects of model specification on tests for unit roots in macroeconomic data. Journal of Monetary Economics 20, 73–103.
Schwert, G.W., 1989. Tests for unit roots: a Monte Carlo investigation. Journal of Business and Economics Statistics 7, 147–159.
Stockman, A.C., Tesar, L.L., 1995. Tastes and technology in a two-country model of the business cycle: explaining international comovements. American Economic Review 85, 168–185.
Summers, R., Heston, A., 1991. The Penn World Table (Mark 5): an expanded set of international comparisons, 1950–88. Quarterly Journal of Economics 106, 327–368.
Taylor, M.P., 1988. An empirical examination of long run purchasing power parity using cointegration techniques. Applied Economics 20, 1369–1381.
Wei, S.-J., Parsley, D.C., 1995. Purchasing power disparity during the floating rate period: exchange rate volatility, trade barriers and other culprits. National Bureau of Economic Research, working paper No. 5032.
Zivot, E., 1995. The power of single equation tests for cointegration when the cointegrating vector is known. Department of Economics, University of Washington.