State of the Art Unit Root Tests and Purchasing Power Parity

Claude Lopez[a] Christian J. Murray[b] David H. Papell[b]

[a]University of Cincinnati, *[b]University of Houston*

April 2004

Abstract

Although the question of whether Purchasing Power Parity (PPP) holds in the long run has been extensively studied, the answer is still controversial. Some of the strongest evidence is provided by Taylor (2002), who concludes that long-run PPP held over the twentieth century. We argue that this conclusion is quite sensitive to the use of sub-optimal lag selection in unit root tests. Using superior lag selection methods, we find that long run PPP held for the real exchange rates of only 9 out of the 16 industrialized countries in Taylor's sample with the U.S. dollar as the base currency.

Correspondence:

Claude Lopez, Department of Economics, University of Cincinnati
tel: (513) 556 2346, fax: (513) 556 2669 email: claude.lopez@uc.edu

Chris Murray, Department of Economics, University of Houston
tel: (713) 743 3835, fax: (713) 743 3798, email: cmurray@mail.uh.edu

David Papell, Department of Economics, University of Houston
tel: (713) 743 3807, fax: (713) 743 3798, email: dpapell@mail.uh.edu

1. Introduction

The question of whether Purchasing Power Parity (PPP) holds in the long run has been the subject of voluminous research. Some of the strongest evidence of long-run PPP is provided by Taylor (2002). An important contribution of Taylor's work is to construct real exchange rate data for over 100 years for 20 countries. For the first time, it is possible to investigate PPP using long-horizon data with approximately the same set of advanced countries as is commonly used in studies with post-1973 floating real exchange rate data. After performing the Augmented Dickey-Fuller (ADF) test and the more powerful DF-GLS test of Elliott, Rothenberg, and Stock (1996), Taylor concludes that "PPP has held in the long run over the twentieth century for my sample of twenty countries." The evidence is sufficient to lead him to write: "If PPP holds in the long run, it is no longer productive to devote further attention to the stationarity question."

In this paper, we take issue with Taylor's conclusion that PPP has held over the twentieth century. More precisely, we show that strong evidence of PPP can be found for only 9 out of 16 industrialized countries with the U.S. dollar as the base currency. The reason for the difference lies in the seemingly arcane topic of lag selection for augmenting unit root regressions in the presence of serial correlation. Taylor employs a method that selects very short lag lengths, mostly zero. Research by Hall (1994) and Ng and Perron (1995, 2001) shows that the use of too short lag lengths lowers power for ADF tests and makes DF-GLS tests oversized. They recommend a general-to-specific procedure for ADF tests and a modified Akaike information criterion for DF-GLS tests. These procedures have become standard practice for conducting unit root tests.

We perform unit root tests using standard lag selection techniques. Whether we use the ADF or the DF-GLS test, the result is the same. The unit root null hypothesis can be rejected in favor of either level or trend stationarity for 9 out of 16 countries at the 5 percent level. We do not see how finding evidence of PPP for one more than half of the counties justifies describing the twentieth century as "A Century of Purchasing Power Parity."

The rest of this paper is organized as follows. Section 2 discusses the data, unit root tests, and issues regarding lag selection. Section 3 performs an exercise which demonstrates that any desired conclusion about the existence of long run PPP can be reached if one chooses the lag to rig the results. Section 4 summarizes and offers concluding remarks.

2. The Data and Unit Root Tests

Taylor (2002) collects nominal exchange rate and price level data through 1996 for 20 countries, each for over 100 years, yielding 19 US dollar denominated real exchange rates. The price levels are consumer price deflators or, if not available, GDP deflators. We extend Taylor's data through 1998, and omit Argentina, Brazil, and Mexico, in order to focus solely on developed countries. This leaves us with 16 dollar denominated real exchange rates: Australia, Belgium, Canada, Denmark, Finland, France, Germany, Italy, Japan, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom. Taylor conducts both ADF and DF-GLS tests, with and without deterministic time trends.[1]

The ADF test, without a deterministic time trend, runs the following regression:

$$q_t = c + \alpha q_{t-1} + \sum_{i=1}^{k} \psi_i \Delta q_{t-i} + u_t \tag{1}$$

where q_t is the natural logarithm of the real exchange rate. This regression includes *k* lagged first differences to account for serial correlation.

The more powerful DF-GLS test runs the following auxiliary regression:

$$q_t^{\mu} = \alpha q_{t-1}^{\mu} + \sum_{i=1}^{k} \psi_i \Delta q_{t-i}^{\mu} + u_t \tag{2}$$

where q_t^{μ} is the GLS demeaned real exchange rate. That is, $q_t^{\mu} = q_t - \tilde{\beta} z_t$, where $z_t = 1$, $\tilde{\beta} = \left(\sum \tilde{z}_t^2\right)^{-1} \sum \tilde{z}_t \tilde{q}_t$, $\tilde{q}_t = (q_1, (q_2 - \alpha q_1), ..., (q_T - \alpha q_{T-1}))'$, $\tilde{z}_t = (1, (1-\alpha), ..., (1-\alpha))'$, $\alpha = 1 + c/T$, and $c = -7$.

Taylor uses a Lagrange Multiplier (LM) criterion to choose the number of lags in the unit root regressions, and in most cases the selected lag is zero. For the real exchange rates that we consider in this paper, Taylor finds that with the ADF test, 6 of 16 unit root tests are rejected at the 5% level, while with the DF-GLS test, 10 of the 16 unit root tests are rejected at the 5% level. The increased rejection rate with the DF-GLS test is

[1] The difference between our results and Taylor's is not caused by extending the data from 1996 to 1998. Taylor also tests for PPP using a "world" basket as the base currency.

attributed to the higher power of the test. Much research has been devoted to the topic of lag selection in unit root tests, and to our knowledge, the LM criterion has not been studied in this context.[2] Also, there are well documented problems with unit root tests when the chosen lag is too small. Specifically, as shown by Hall (1994) and Ng and Perron (1995), the ADF test suffers from low power when the lag length is too small. Also, Ng and Perron (2001) demonstrate that the DF-GLS test suffers from size distortions when the lag is too small. Since the former problems leads to too few rejections, and the latter problems leads to too many, lag selection alone may be responsible for the difference in rejections rate that Taylor finds between ADF and DF-GLS tests.

There exist lag selection procedures whose properties are well understood, and which offer a better combination of size and power. Specifically, the existing literature suggests using the general-to-specific (GS) procedure of Hall (1994) for ADF tests and the Modified Akaike information criterion (MAIC) of Ng and Perron (2001) for DF-GLS tests. Accordingly, we report ADF and DF-GLS tests in Table 1 and Table 2 respectively. Every series ends in 1998 but not all have the same starting date. The sample period for each series is indicated in Tables 1 and 2. For completeness, we consider LM, GS, and MAIC lag selection for each test, although we will focus on GS for ADF test, and MAIC for DF-GLS test.

With GS lag selection, we find 8 of 16 rejections at the 5% level using ADF tests; 2 more than Taylor finds for the same 16 countries with data ending in 1996. With MAIC lag selection, we find 9 of 16 rejections at the 5% level with DF-GLS test, 1 less than Taylor with the same data ending in 1996. Thus, using proper lag selection procedures, the difference in rejection rates between the ADF and DF-GLS test has almost been eliminated.[3]

Taylor also reports rejections of the unit root null based on ADF and DF-GLS tests with the inclusion of a deterministic time trend. While rejection of the unit root null in this case is not evidence of PPP in the strict sense of a mean reverting real exchange rate,

[2] See Hall (1994), Ng and Perron (1995), and Ng and Perron (2001).

[3] The 8 rejections in Table 1 are not a proper subset of the 9 rejections in Table 3. In total, there are 11 countries that reject the unit root null in Tables 1 and/or 3. We do not report 11 rejections since this testing strategy would lead our test to be oversized.

one can ascribe a Balassa-Samuelson interpretation to a stationary real exchange rate around a time trend.[4] A sensible strategy when testing for unit roots in persistent data, when some of the series are clearly trending, is to employ a two-step procedure. First, test the unit root null against the alternative of level stationarity. If the null is rejected, conclude that the series is level stationary, or in our case, that long run PPP holds. If the null cannot be rejected against level stationarity, test the unit root null against the alternative of trend stationarity. If the null is rejected, conclude that the series is trend stationary. If the null cannot be rejected, conclude that the series is I(1), and that long run PPP fails to hold. This testing strategy, which Taylor essentially follows, has intuitive appeal when testing for unit roots in real exchange rate data, since some series are clearly trending. Suppose that the true data generating process is level stationarity. In this case the unit root test described by equation (1) or (2) is appropriate. A time trend is an extraneous regressor, and its inclusion reduces power. However, if the true data generating process were trend stationarity, failing to include a time trend also results in a reduction in power of the test. In addition, this loss or power from excluding a time trend when it should be present is more severe than the reduction in power associated with including a time trend when it is extraneous; see West (1987). Finally, if the series is I(1), then both steps of this procedure should lead to a failure to reject the unit root null.

Using this two-step procedure, Taylor finds 3 additional 5% rejections for the ADF test, and 1 additional 5% rejection for the DF-GLS test. He cites these extra rejections from including a time trend as an important ingredient to his conclusion that PPP holds for over a century. However, it appears that it is the use of sub-optimal lag selection which leads Taylor to find so many rejections in the presence of linear time trend, not the (trend) stationarity of the data. When one uses proper lag selection methods, there is only one additional unit root rejection for the ADF test and none for the DF-GLS test.

We report the results of both tests, again using all 3 methods of lag section for completeness. The ADF unit root test with a time trend runs the following regression:

$$q_t = c + bt + \alpha q_{t-1} + \sum_{i=1}^{k} \psi_i \Delta q_{t-i} + u_t \,,$$

[4] See Papell and Prodan (2003) for further discussion.

the results of which are reported in Table 3. For the ADF test with GS lag selection the 10% rejection for the U.S./Australia exchange rate is strengthened to a 5% rejection.

Table 4 reports the results of the DF-GLS test, also allowing for a time trend under the alternative hypothesis, which runs the following auxiliary regression:

$$q_t^\tau = \alpha q_{t-1}^\tau + \sum_{i=1}^{k} \psi_i \Delta q_{t-i}^\tau + u_t$$

where q_t^τ is the GLS detrended real exchange rate. For the DF-GLS test with MAIC lag selection, there are no additional unit root rejections.

Thus, using state of the art unit root tests and lag selection, 9 of 16 countries are consistent with PPP, while the remaining 7 are consistent with the failure of PPP to hold in the long run. It is interesting to note that the number of rejections at the 5 percent level is the same for the ADF and the DF-GLS test – use of the more powerful DF-GLS test does not make any difference in this case.

3. Lag Selection for the Insane

We perform the following exercise to provide further evidence that lag selection plays an important role in determining the existence, or lack thereof, of long run PPP. Suppose the lag length were chosen to minimize the p-value of the unit root statistic. That is, one considers all possible values of *k*, and chooses the value which maximizes the evidence against the unit root null. Equivalently, this procedure maximizes the evidence that long run PPP holds. Admittedly, choosing *k* in this manner would lead to unit root tests with absurd size and power properties, and we certainly do not advocate using this method of lag selection in practice. We employ it here to make the point that lag selection for unit root tests is not merely a theoretical curiosity but can have dramatic implications for empirical findings.

The results from this exercise are reported in Table 5. We conduct ADF and DF-GLS tests, both with and without time trends, by choosing the lag length to maximize the evidence the long run PPP holds.[5] We focus only on the results of unit root tests without time trends, as there are no substantive differences in rejection frequencies of the unit

[5] We only report the decision to reject or not at the 5% level. P-values of the unit root t-statistic and chosen lag lengths are omitted for brevity.

root null by including a time trend.[6] We find 10/16 and 13/16 findings of long run PPP for the ADF and DF-GLS respectively. These results are stronger than what we find in Section 2, where we employ standard lag selection procedures. With the DF-GLS test in particular, this method of lag selection would provide compelling evidence that long run PPP holds for most of the real exchange rates in Taylor's (2002) data set.

Conversely, we next consider a lag selection procedure which maximizes the p-value of the unit root statistic. In contrast to our previous exercise, we consider all possible values of *k*, and choose the value which provides the greatest evidence *against* long run PPP. These results are reported in Table 6, again for ADF and DF-GLS tests, with and without time trends. The results are even more striking in this case. We find a failure to reject the null that long run PPP does not hold for 15/16 and 14/16 cases for the ADF and DF-GLS tests respectively. These results would suggest that long run PPP did not hold over the 20th century between the U.S. and any other country in our sample except for Finland and possibly Italy.

The message from this exercise is clear. Different lag selection methods lead to drastically different conclusions regarding the existence of long run PPP. By choosing the lag length to rig the outcome, basically any desired conclusion can be reached. Therefore, it is important to employ lag selection procedures which lead to unit root tests with good size and power properties, so that the outcomes of unit root tests are credible.

4. Conclusion

Taylor (2002) constructs long-horizon US dollar denominated real exchange rate data for 16 developed and 3 developing countries from nominal exchange rates and consumer price deflators. This data allows researchers, for the first time, to study long-horizon real exchange rates for developed countries with comparable coverage and construction to the data that are commonly used to investigate post-1973 real exchange rates. He concludes that, using more powerful DF-GLS tests, long-run PPP has held over the twentieth century.

We find his conclusion too strong. When one performs unit root tests with superior methods of lag selection, the unit root null hypothesis is rejected in favor of long run PPP

[6] Again, rejecting the unit root null when a time trend is included is not equivalent to finding long PPP in the strict sense of a mean reverting real exchange rate.

at the 5% level for only 9 of 16 real exchange rates. To us, this evidence is *not* sufficient to conclude that long run PPP has held over the last century or that devoting further attention to the issue of stationarity would be unproductive.

References

Elliott, G., Rothenberg, T., and Stock, J.H., 1996, “Efficient Tests for an Autoregressive Unit Root,” *Econometrica* 64:813-836.

Hall, A., 1994, “Testing for a Unit Root in Time Series with Pretest Data-Based Model Selection,” *Journal of Business and Economic Statistics* 12: 461-70.

Ng, S. and P. Perron, 1995, “Unit Root Test in ARMA Models with Data Dependent Methods for the Selection of the Truncation Lag,” *Journal of the American Statistical Association* 90: 268-281.

Ng. S. and P. Perron, 2001 “Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power.” *Econometrica* 69: 1519-1554.

Papell, D.H., and R. Prodan, 2003, “Long Run Purchasing Power Parity: Cassel or Balassa-Samuelson?” working paper, University of Houston.

Taylor, A.M., 2002, “A Century of Purchasing Power Parity,” *The Review of Economics and Statistics* 84: 139-150.

West, K.D., 1987, “A Note on the Power of Least Squares Tests for a Unit Root,” *Economics Letters* 24: 249-252.

Table 1. ADF Unit Root Tests Without Time Trends

$$q_t = c + \alpha q_{t-1} + \sum_{i=1}^{k} \psi_i \Delta q_{t-i} + u_t$$

Country	Sample	ADF_{LM}	k_{LM}	ADF_{GS}	k_{GS}	ADF_{MAIC}	k_{MAIC}
Australia	1870-1998	-2.26	0	-2.62*	1	-2.26	0
Belgium	1880-1998	-3.32**	0	-4.15***	1	-2.92**	3
Canada	1870-1998	-1.62	0	-1.62	0	-1.62	0
Denmark	1880-1998	-2.27	0	-1.24	6	-1.24	6
Finland	1881-1998	-4.58***	0	-6.02***	1	-4.58***	0
France	1880-1998	-3.55***	1	-3.55***	1	-2.53	6
Germany	1880-1998	-2.95**	1	-2.95**	1	-2.39	2
Italy	1880-1998	-3.33**	0	-4.28***	2	-3.33**	0
Japan	1885-1998	-0.37	0	-1.02	1	-0.74	2
Netherlands	1870-1998	-2.14	0	-2.79*	1	-2.46	2
Norway	1870-1998	-2.58*	0	-3.67***	1	-2.23	5
Portugal	1890-1998	-2.69*	0	-2.25	5	-1.99	4
Spain	1880-1998	-2.43	0	-3.24**	1	-2.43	3
Sweden	1880-1998	-2.95**	0	-3.72***	1	-3.09**	2
Switzerland	1892-1998	-2.18	1	-1.50	2	-1.50	2
UK	1870-1998	-3.20**	0	-2.61*	4	-3.20**	0

*, **, and *** denote significance at the 10%, 5%, and 1% level respectively.

Table 2. ADF Unit Root Tests With Time Trends

$$q_t = c + bt + \alpha q_{t-1} + \sum_{i=1}^{k} \psi_i \Delta q_{t-i} + u_t$$

Country	Sample	ADF_{LM}	k_{LM}	ADF_{GS}	k_{GS}	ADF_{MAIC}	k_{MAIC}
Australia	1870-1998	-3.22*	0	-3.66**	1	-3.22*	0
Belgium	1880-1998	-3.90**	0	-5.11***	1	-3.90**	0
Canada	1870-1998	-2.98	0	-2.98	0	-2.98	0
Denmark	1880-1998	-2.90	0	-1.97	6	-1.97	6
Finland	1881-1998	-4.69***	0	-6.22***	1	-4.69***	0
France	1880-1998	-4.16***	1	-4.16***	1	-1.67	8
Germany	1880-1998	-3.32*	1	-3.32*	1	-2.75	2
Italy	1880-1998	-3.33*	0	-4.27***	2	-3.33*	0
Japan	1885-1998	-2.07	0	-1.98	7	-2.32	2
Netherlands	1870-1998	-2.40	0	-3.19*	1	-2.40	0
Norway	1870-1998	-2.71	0	-3.95**	1	-2.64	5
Portugal	1890-1998	-2.60	0	-2.15	5	-1.72	4
Spain	1880-1998	-2.37	0	-3.23*	1	-2.34	3
Sweden	1880-1998	-3.40*	0	-4.52***	1	-3.40*	0
Switzerland	1892-1998	-3.63**	1	-2.78	2	-2.78	2
UK	1870-1998	-3.38*	0	-2.74	4	-3.38*	0

*, **, and *** denote significance at the 10%, 5%, and 1% level respectively.

Table 3. DF-GLS Unit Root Tests without Time Trends

$$q_t^{\mu} = \alpha q_{t-1}^{\mu} + \sum_{i=1}^{k} \psi_i \Delta q_{t-i}^{\mu} + u_t$$

Country	Sample	*DF-GLS*$_{LM}$	k_{LM}	*DF-GLS*$_{GS}$	k_{GS}	*DF-GLS*$_{MAIC}$	k_{MAIC}
Australia	1870-1998	-2.29**	0	-2.65***	1	-2.29**	0
Belgium	1880-1998	-2.84***	0	-3.57***	1	-2.38**	3
Canada	1870-1998	-1.40	1	-1.29	0	-1.29	0
Denmark	1880-1998	-2.24**	0	-1.20	6	-1.20	6
Finland	1881-1998	-4.49***	0	-5.85***	1	-4.49***	0
France	1880-1998	-2.24**	1	-1.34	4	-1.07	6
Germany	1880-1998	-2.52**	1	-2.52**	1	-1.97**	2
Italy	1880-1998	-3.35***	0	-4.29***	2	-3.35***	0
Japan	1885-1998	0.25	0	-0.08	1	-0.08	1
Netherlands	1870-1998	-1.84*	0	-2.51**	1	-2.20**	2
Norway	1870-1998	-1.61*	0	-2.49**	1	-1.31	5
Portugal	1890-1998	-1.88*	0	-1.52	5	-1.29	6
Spain	1880-1998	-2.13**	0	-2.08**	3	-2.08**	3
Sweden	1880-1998	-2.35**	0	-2.36**	2	-2.36**	2
Switzerland	1892-1998	-1.47	1	-0.76	2	-0.76	2
UK	1870-1998	-2.81***	0	-2.26**	4	-2.26**	4

*, **, and *** denote significance at the 10%, 5%, and 1% level respectively.

Table 4. DF-GLS Unit Root Tests with Time Trends

$$q_t^\tau = \alpha q_{t-1}^\tau + \sum_{i=1}^{k} \psi_i \Delta q_{t-i}^\tau + u_t$$

Country	Sample	$DF\text{-}GLS_{LM}$	k_{LM}	$DF\text{-}GLS_{GS}$	k_{GS}	$DF\text{-}GLS_{MAIC}$	k_{MAIC}
Australia	1870-1998	-2.77*	0	-3.17**	1	-2.77*	0
Belgium	1880-1998	-3.93***	0	-5.14***	1	-3.93***	0
Canada	1870-1998	-1.48	0	-1.48	0	-1.48	0
Denmark	1880-1998	-2.79*	0	-1.86	6	-1.86	6
Finland	1881-1998	-4.72***	0	-6.27***	1	-4.72***	0
France	1880-1998	-4.04***	1	-4.04***	1	-1.60	8
Germany	1880-1998	-3.35**	1	-3.35**	1	-2.78*	2
Italy	1880-1998	-3.35**	0	-4.30***	2	-3.35**	0
Japan	1885-1998	-1.92	0	-1.98	7	-2.28	2
Netherlands	1870-1998	-2.43	0	-3.23**	1	-2.43	0
Norway	1870-1998	-2.58	0	-3.81***	1	-2.60	5
Portugal	1890-1998	-2.49	0	-2.15	5	-1.71	4
Spain	1880-1998	-2.37	0	-3.22**	1	-2.36	3
Sweden	1880-1998	-3.44**	0	-4.55***	1	-3.44**	0
Switzerland	1892-1998	-3.64***	1	-2.78*	2	-2.78*	2
UK	1870-1998	-2.99*	0	-2.45	4	-2.99*	0

*, **, and *** denote significance at the 10%, 5%, and 1% level respectively

Table 5. Choosing *k* to Maximize the Evidence that PPP Holds

	Outcome of Unit Root Test at the 5% Significance Level			
	ADF Test		**DF-GLS Test**	
Country	**w/o Time Trend**	**w/ Time Trend**	**w/o Time Trend**	**w/ Time Trend**
Australia	FTR	Reject	Reject	Reject
Belgium	Reject	Reject	Reject	Reject
Canada	FTR	FTR	FTR	FTR
Denmark	FTR	FTR	Reject	Reject
Finland	Reject	Reject	Reject	Reject
France	Reject	Reject	Reject	Reject
Germany	Reject	FTR	Reject	Reject
Italy	Reject	Reject	Reject	Reject
Japan	FTR	FTR	FTR	FTR
Netherlands	FTR	FTR	Reject	Reject
Norway	Reject	Reject	Reject	Reject
Portugal	Reject	FTR	Reject	Reject
Spain	Reject	FTR	Reject	Reject
Sweden	Reject	Reject	Reject	Reject
Switzerland	FTR	Reject	FTR	Reject
UK	Reject	Reject	Reject	Reject
Long Run PPP	**10/16**	**9/16**	**13/16**	**14/16**

The outcome of the test, rejection or failure to reject, is determined using 5% critical values.

Table 6. Choosing *k* to Minimize the Evidence that PPP Holds

	Outcome of Unit Root Test at the 5% Significance Level			
	ADF Test		**DF-GLS Test**	
Country	**w/o Time Trend**	**w/ Time Trend**	**w/o Time Trend**	**w/ Time Trend**
Australia	FTR	FTR	FTR	FTR
Belgium	FTR	FTR	FTR	Reject
Canada	FTR	FTR	FTR	FTR
Denmark	FTR	FTR	FTR	FTR
Finland	Reject	FTR	Reject	FTR
France	FTR	FTR	FTR	FTR
Germany	FTR	FTR	FTR	FTR
Italy	FTR	FTR	Reject	FTR
Japan	FTR	FTR	FTR	FTR
Netherlands	FTR	FTR	FTR	FTR
Norway	FTR	FTR	FTR	FTR
Portugal	FTR	FTR	FTR	FTR
Spain	FTR	FTR	FTR	FTR
Sweden	FTR	FTR	FTR	Reject
Switzerland	FTR	FTR	FTR	FTR
UK	FTR	FTR	FTR	FTR
No Long Run PPP	**15/16**	**16/16**	**14/16**	**14/16**

The outcome of the test, rejection or failure to reject, is determined using 5% critical values.