

**Measurement Ahead of Theory? The Economic Significance  
of Recent Research in Time Series Modeling**

William J. Polley  
Assistant Professor of Economics  
Foster College of Business Administration  
Bradley University  
1501 W. Bradley Ave.  
Peoria, IL 61625

[wpolley@bradley.edu](mailto:wpolley@bradley.edu)

August 14, 2003  
Revised February 4, 2004

Final Version Published in *Journal of Financial and Economic Practice* Fall 2004  
Personal copy on website with permission of the editor

## Measurement Ahead of Theory? The Economic Significance of Recent Research in Time Series Modeling

### Abstract

Econometric studies of market efficiency can be traced through three generations of models. The first generation, in particular, had problems which were exposed in a number of works by McCloskey. Later generations of models have addressed these some of these critiques. However, a substantial gap between measurement and theory remains and must be closed in order to fully address McCloskey's critiques. A new generation of theoretical models informed by good quantitative measurement is necessary for scientific advancement to continue in this area.

### INTRODUCTION

Econometric models of certain economic or financial relationships have advanced by leaps and bounds in the past 15 to 20 years. Whether this has led to advances in our understanding of these relationships is less certain. For example, consider the purchasing power parity (PPP) hypothesis. Stated in its most general form, absolute PPP requires that at every point in time

$$p = ep^f .$$

The domestic price level is  $p$ . The exchange rate is  $e$ , and the foreign price level is  $p^f$ . Interest in this problem dates back to the classical economists. After the demise of the Bretton Woods system, there was a resurgence of interest that led to a flurry of empirical papers. Many of these rejected the PPP hypothesis; not only in its absolute form, but also in its relative form

$$\% \Delta p = \% \Delta e + \% \Delta p^f .$$

Relative PPP could hold even if absolute PPP failed because of the presence of, for example, transaction costs driving a permanent wedge between prices in different countries in absolute terms.

The repeated tests of PPP in the literature indicate the profession's reluctance to give up on the theory. At its most basic level, PPP theory suggests that in equilibrium there are no opportunities for arbitrage – “no free lunches.” To suggest otherwise would require some explanation. Thus, more testing and more theory would follow as economists searched for some confirmation that free lunches were indeed rare.<sup>1</sup>

Many of the post-Bretton Woods tests of PPP in the late 1970s used ordinary least squares (OLS) to estimate a model like the following

$$\log p_t = \alpha + \beta \log e_t p_t^f + \varepsilon_t . \tag{1}$$

The last term is the error term which is distributed normally,  $N(0, \sigma)$ . The model can also be posed in terms of changes in the log of the price levels for testing relative PPP. PPP requires  $\beta = 1$ . However, if there are long and variable lags to the adjustment process, there is no reason to expect the theory to hold precisely when subjected to this sort of test. Between observations (which may be a month or a year apart, depending on the data set used), the economies may be subjected to many shocks and adjusting to them. The adjustment processes of prices and exchange rates may be linked, but they also may adjust at different rates and be influenced by different factors.

So the theory evolved to consider the short run and long run separately with short run OLS hypothesis tests falling out of favor as the research progressed. By the manner in which we measure these variables, short run PPP may never precisely hold, but there should be a sense in which it holds in the long run. The simplest way to test this is to model the problem as an AR(1) process and test for a unit root. A unit root would mean that the effect of a one-time shock would not die out over time – a failure of *long run* PPP. Such tests of PPP can reveal the “half-life” of the shock. In practice this has been found to be about 3 or 4 years. While this reassures us that deviations are not permanent, one may argue that 3 or 4 years is a rather long time in the currency market. One could see this as still indicating arbitrage opportunities that are not being exploited. Of course, the true adjustment process is likely to be more rapid than is implied by the measurements we are taking. That is, we may look at monthly data, but the variables actually respond to shocks on a daily (or even hourly) basis. As Taylor (2001) points out, this can cause bias in the testing of the PPP hypothesis, causing the half-life to be overstated.

PPP is not the only example of this sort of problem. The unbiased forward rate hypothesis (UFRH) considers whether the forward exchange rate is an unbiased predictor of the future spot rate. The comparisons of interest rates across countries as well as their term structure within a country provide fertile ground for econometric modeling. These problems all generally deal with issues of market efficiency and have also benefited greatly from recent advances in time series modeling. Though the prime focus of this paper is on the PPP question, the topics discussed here apply to other market efficiency questions.

For a number of years, Deirdre McCloskey has been a self-described Cassandra, warning economists of the dangers of putting too much emphasis on statistical significance.<sup>2</sup> Some of her criticism has been aimed at tests of PPP. Given the type of testing that was done at the time, her statements were entirely appropriate. Today, however, the tests have changed. The old criticisms should not be forgotten, but rather examined in a new light.

The goal of this paper is to examine the recent developments in time series analysis and how they are changing the way that researchers look at these problems. Although the state-of-the-art has progressed considerably since McCloskey first began to raise questions about the appropriateness of the testing procedures applied to these problems, important concerns remain. I will identify some of these concerns using McCloskey’s critiques of econometric practice as a focal point. In doing so, I will follow the research through three generations of models. Finally, I discuss some possibilities for research that would shed new light on these questions. Given the precision with which

we measure high frequency financial market data, our ability to measure these variables still outpaces our ability to draw inference from them.

## MCCLOSKEY'S CRITICISM OF FIRST GENERATION MODELS OF PPP

In a number of books and articles, McCloskey has criticized the way statistical analysis has been conducted in economics. For the purposes of this paper, I will focus primarily on two specific (but related) critiques. The first of these is the misuse of hypothesis testing (including a widespread neglect of the power function). The second questions the methodology of the test of PPP specifically, although the criticism could be leveled at other models in international finance as well.

One of the best expositions of McCloskey's dissatisfaction with standard hypothesis testing is in a 1996 *Journal of Economic Literature* paper co-authored with Stephen Ziliak and aptly named, "The Standard Error of Regressions." In this paper, McCloskey and Ziliak examine all 182 of the full-length papers in the *American Economic Review* from the 1980s which use regression analysis. They ask a number of questions of each paper. The questions are designed to evaluate how well each paper follows the econometric practice McCloskey advocates. On some of the questions such as, "Does the paper discuss the size of the coefficients?" the AER papers score well, with 80.2 percent in the affirmative. On others the outcome is less encouraging. Only 41.2 percent avoided using the word "significance" in ambiguous ways, meaning "statistically significant" in one sentence and "large enough for policy or science" in another. Somewhat fewer, only 32.4 percent, included units and descriptive statistics for all the regression variables.<sup>3</sup> Only 4.4 percent of the papers in the AER during the 1980s that used regression analysis considered the power of the test, that is, the probability of rejecting the null hypothesis when it is false.

McCloskey and Ziliak are not the only ones to point out the shortcomings of econometric practice. Leamer (1983) also asks economists to be more careful with their use of statistics. Obviously, many economists have taken their points seriously. McCloskey and Ziliak (1996) point out some examples of this. One particularly instructive passage is cited.

Here and subsequently, all statements about statistical "significance" should not be taken literally. Besides the usual issue of data mining clouding their interpretation, the "sample" analyzed comes close to covering completely the relevant population. Tests of significance are used here as a metric for discussing the relative fit of different versions of the model. In each case, the actual magnitude of the estimated coefficients is of more interest than their precise "statistical significance." (Griliches 1986, p. 146, quoted in McCloskey and Ziliak 1996, p.106).

This passage addresses not only the fact that statistical significance is not equivalent to significance in a scientific sense, but also the fact that when the "sample" is the population the usual interpretation of statistical significance does not apply.<sup>4</sup> The latter implication is relevant to the PPP issue since technically (at least in the first generation of models) it is not a sampling problem in the usual sense. Elsewhere, McCloskey and

Zecher (1984) point out that those studies examine the entire population (all of the realizations of prices and exchange rates in a time period). Clearly this does pose a problem. If the regression involves the entire population, then you are not “estimating” the coefficient from a sample. With the population, you get the true coefficient – from which you can draw scientific inference.

Best practice in regression analysis is mindful of early researchers such as Neyman and Pearson (1933) and Wald (1939) who cautioned readers that the significance of a result in a scientific sense involves more than what we have come to call “statistical significance.” The Neyman and Pearson (1933) reference contains the oft cited passage asking whether it is more serious to convict the innocent or acquit the guilty. In other words, the reference is to the power of the test, all too often neglected from regression studies.

While some of McCloskey’s works have included more controversial discussions of economic practice, the critique of regression analysis is gaining acceptance. She cites (1996) Nobel laureate Kenneth Arrow, Edward Leamer, and Arthur Goldberger as prominent economists who are in agreement with her. Her argument directly references and is in agreement with the seminal works of Neyman, Pearson, and Wald. The evidence is clearly on her side. Recent practice in applied econometric analysis has too often neglected to consider power, confused the term significance, and ignored the question “How large is large?”

The second criticism cuts to the heart of the PPP question, and given recent research, it deserves to be updated. In McCloskey and Zecher (1984) as well as McCloskey and Ziliak (1996), the subject is a model like equation (1). The initial thrust of the argument is that with the large number of data points typically used in these studies, standard errors can be relatively small. Thus, the estimated  $\beta$  in equation (1) might be close to 1, say 0.99, but have a miniscule standard error, say 0.0001, which would lead the researcher to conclude that PPP fails. Yet, 0.99 might be close enough to 1 for the scientific purposes at hand, such as determining the efficacy of monetary policy or determining whether profitable arbitrage exists.

Thus far, the argument is a concrete application of her criticism of statistical significance in general. As such, it is classic. It is hard to come up with an economic example that so clearly induced the behavior that she denounces so vehemently.<sup>5</sup> While it still makes a great pedagogical example for teaching future generations of economists and statisticians how *not* to interpret a regression, it is no longer precisely relevant to the way that research into this question is being conducted. Although McCloskey and Ziliak (1996, p. 98) refer to this as the “usual” test of PPP, it was by this time disappearing from the literature.<sup>6</sup> My point is not to refute the argument or simply categorize it as obsolete, but to update it. Improper analysis using newer methods would be just as dangerous as with the old methods. This will be discussed in more detail in the sections of this paper dealing with second and third generation models.

McCloskey’s criticism of PPP testing, however, does not stop at this. McCloskey and Zecher (1984) point out that the success or failure of PPP could be interpreted to imply whether all arbitrage opportunities have been exploited. If prices in different countries diverge, it should induce trade flows and prices to respond. If these divergences in prices persist for months or years – which is in fact precisely the finding, even of the newer models – it implies that profitable arbitrage is not taking place. As

McCloskey and Zecher (1984, p. 131) suggest, “Go thee and prosper.” This is a serious challenge to the PPP literature, then and now.

Of course, as Lipsey (1984) states in a commentary on McCloskey and Zecher (1984), such claims are typically not made by researchers addressing the PPP question. Economists have an aversion to openly claiming that agents are irrational; even though McCloskey and Zecher’s criticism might give one pause. However, present in McCloskey and Zecher (1984) and Lipsey (1984) as well as in many PPP studies of the time is the general notion that there is an adjustment process that takes place over time, and that arbitrage of aggregate indices is difficult, not to mention that there are risks and transport costs. Rational economic actors could very well be arbitraging as much as they can within the period, and also arbitraging over time using financial instruments. (The latter are not present in PPP studies of the period.)

Finally, the criticism of McCloskey and Zecher (1984) illustrates the error of what they characterize as the “Martian approach” to macroeconomics. Despite the amusing name, this is a serious point. What they term the “Martian approach” is an approach to macroeconomics that treats the U.S. as if it were Mars – totally disconnected from the rest of the world – a closed economy. Such an approach implies that the Fed would have the freedom to set prices and interest rates without being constrained by conditions in the rest of the world. By contrast, what they term the “Iowa City approach” is a reference to the premise that a hypothetical monetary authority in Iowa City could not independently set interest rates and prices in Iowa City. Iowa City is inextricably linked with other nearby communities and thus to the U.S. at large. The “Iowa City approach” starts from the premise that the U.S. is to the world economy as Iowa City is to the U.S.

The bottom line of all of this is that if PPP does not hold because price levels diverge and exhibit no consistent endogenous relationship to one another, then the Martian approach is acceptable. We lose little or nothing by using the closed economy macro model. However, McCloskey and Zecher (1984) and McCloskey in several other papers, advocate the Iowa City approach.<sup>7</sup> McCloskey strongly believes that the interconnections between economies simply cannot be ignored and is not charitable to closed economy macroeconomics. On this, there is some agreement with Lipsey (1984).

Fortunately, there is another theme to the paper... It is the important and reasonable one that “it is hard to believe that foreign prices or interest rates did not matter. It is hard to believe that American prices and interest rates are not at all constrained directly by the forces of arbitrage,” and “[purchasing-power parity] is not so great a failure that macroeconomics can go on ignoring the rest of the world.” If that is the point the authors really want to make, even many skeptics about purchasing-power parity could agree. (p. 156)

In the end, the tests of PPP in the style of equation (1) common in the 1970s did not lead to conclusive answers. There are many reasons for this, and they are not at issue here.<sup>8</sup> However, theoretical developments in time series analysis led to investigations of PPP and other financial market questions using more sophisticated methods.

## THE AUTOREGRESSIVE MODEL VERSUS THE STATIC REGRESSION MODEL – THE SECOND GENERATION

Time series modeling advanced considerably with the seminal papers by Dickey and Fuller (1979, 1981) (hereafter DF) which considered the basic AR(1) model

$$y_t = \rho y_{t-1} + \varepsilon_t. \quad (2)$$

The error terms are independent draws from a normal distribution  $N(0, \sigma)$ . The researcher often desires to know if the coefficient,  $\rho$ , is equal to 1 – that is, a unit root. If there is a unit root, then the process is nonstationary, or a random walk. The effects of the random disturbances do not dissipate over time, but persist forever. If a time series is nonstationary, the researcher may want to consider the first differenced version of the original series. Thus, the null hypothesis is typically that  $\rho = 1$ . DF (1979) computed the power of this test of the null hypothesis for different numbers of observations and for true values of  $\rho$  from 0.8 to 1.05. Though the DF test continues to be a popular means to testing whether a time series is stationary, other tests are common in the literature. The Augmented Dickey-Fuller (ADF), and the Phillips-Perron test (Phillips and Perron 1988) generalize the basic DF test. The KPSS test of Kwiatkowski, Phillips, Schmidt, and Shin (1992) tests the null hypothesis of stationarity with the unit root being the alternative. Useful introductions to these methods are Banerjee, Dolado, Galbraith, and Hendry (1993) and Dufrénot and Mignon (2002).

Applied to PPP and other market efficiency questions, the researcher asks whether the variable of interest (e.g. difference in the logs of the real exchange rates between two countries) is stationary. Stationarity is evidence of long run reversion to equilibrium (e.g. long run PPP). This is a step in the right direction because the static PPP regression such as that in equation (1) is likely to be plagued with serial correlation of the residuals. The PPP problem has obvious dynamic implications as discussed in the previous section. Perhaps the problem could never be properly specified in a model such as (1) even with corrections for serial correlation and other biases. Perhaps because economic agents are arbitraging across time as well as across countries, the convergence to long run equilibrium after a random shock should not be expected to be immediate.

The unit root test of the AR(1) model is simple and specific. Given a process constructed from independent random shocks, determine if the shocks are permanent or transitory. If they are transitory, estimate how long it takes for the effect of a single shock to dissipate (half-life).

Studies of PPP recently have turned up evidence both supporting and rejecting the unit root hypothesis. However, it should be noted that failure to reject does not automatically mean irrationality on the part of financial market participants or a “Martian” model of macroeconomics. The power of the DF test is low when there are few observations. As an example of the variation of results found in the literature, Frankel and Rose (1996) were able to reject the unit root hypothesis, but Engel (2000) was not. In the former, the authors use panel data, which they acknowledge gives them a better chance of finding convergence than with a single time series because of the low power of the test. In the latter paper, the author uses 25 years of quarterly data (100 observations) and even this may not be enough to overcome the limited power of the test.

The analysis of PPP from the perspective of time series analysis begins to address the most obvious criticism of model (1), specifically that this is a dynamic problem requiring explicitly dynamic tools. Also, recall McCloskey's argument that the static regression model of PPP is not a sampling problem – the researcher has the whole population of data. The AR(1) model overcomes this criticism since the *theory* underlying the model is more consistent with this being a sampling problem. Model (1) is too rigid given that we know (or at least have good reason to strongly believe) that the true process is more complicated. Model (1) imposes a rather confining behavioral assumption, that arbitrage produces a stable link between prices in different countries and any deviation from that is a random error. We only get to see one realization of prices for each period in each country, but that is in fact the entire population. The regression and the hypothesis test will determine whether we should, on the basis of the errors, declare that the assumed relationship does not exist. There is no obvious *economic* interpretation of the errors themselves, and because of the possibility of serial correlation, the independence assumption is certainly suspect. In short, model (1) is plagued with multiple problems.

Model (2) imposes a different, less confining, behavioral assumption, that arbitrage pushes the accumulated deviation from PPP towards zero at a certain rate, but at the same time, the economy is being hit by persistent random shocks, each of which is an independent random draw from a distribution of possible shocks. This is an important difference. The random shocks can be interpreted as combinations of news, policy, or demand shifts which may reasonably be considered independent. In fact, it is the shocks themselves that are the sample. Assuming an AR(1) framework, the regression will estimate the speed of convergence from this realization of shocks. The unit root test will report whether we can distinguish the realized path from a random walk on the basis of this particular sample of shocks. The static regression model (1) had to deal with serial correlation of the residuals. The time series model (2) assumes serial correlation in the data generating process itself, which is much more upfront given the belief that adjustment to a long run equilibrium after a shock takes time. Thus, the AR model (2) overcomes at least some of the problems associated with model (1).

What of McCloskey's repeated attacks on the PPP tests using static regression analysis and the misinterpretation of "significance" in those tests? The specific model may no longer be relevant, but the spirit of the criticism is as relevant as ever, and for the same reason. Whereas model (1) lends itself to a simple reject/do not reject decision based on the estimate of  $\beta$ , model (2) lends itself to the same sort of decision based on the estimate of  $\rho$ . Obviously if we found  $\rho$  to be less than 1, rejected the null hypothesis, rejoiced in the finding of long run purchasing power parity, and found nothing more interesting to say, then we could just change the equation, change a few words, and McCloskey's criticism would be just as biting. The good news is that recent research does not stop here.

In the context of the static regression model (1), McCloskey implores researchers to ask how significant  $\beta$  is for science or for policy. Indeed, the significance of  $\beta$  for science or for policy (hereafter referred to as "economic significance") is frequently hard to determine and always dependent on the question at hand. At this point, one is reminded of Wald (1939 p. 302) who says, "The statistician who wants to test certain hypotheses must first determine the relative importance of all possible errors, which will

depend on the special purposes of his investigation” (quoted in McCloskey and Ziliak (1996 p. 98)). If  $\beta = 0.7$  in model (1), for example, it might mean that policymakers should take an “Iowa City approach” to monetary policy (if for policy intervention purposes 0.7 is “close” to 1), or it could be a sign that arbitrage opportunities exist (if for arbitrage purposes 0.7 is not “close” to 1). Of course, we also need to specify a metric against which either of these policy/scientific claims could be evaluated – and that may very well be the most difficult task of all.

On the other hand, the economic significance of  $\rho$  is more evident. The half-life of the deviation is

$$h = \frac{\left( \ln \frac{1}{2} \right)}{\ln \rho}.$$

For example,  $\rho = 0.8$  translates to a half-life of just over 3 periods. This is quantitative science. It should lead the researcher to ask why it takes this long for the shock to dissipate. How do transaction costs, other financial markets, trade barriers, etc. affect the length of the half-life? How does the frequency of the data affect it? These represent just some of the questions that could be asked.

Unfortunately, as noted earlier in this section, the evidence from this model has been mixed. Some researchers have been unable to reject the unit root null. The low power of the test notwithstanding, this remained a stumbling block. An updated version of McCloskey’s critique might argue that the debate over unit roots versus near unit roots is just as bad as the debate over the value of  $\beta$  in model (1).<sup>9</sup> However, the fact that the power of the tests used (particularly the DF test) was low against the unit root null hypothesis was widely known and proved important for pushing researchers to develop better tests and better models.

## NONLINEAR MODELS – THE THIRD GENERATION

The second generation of econometric models of PPP represented a dramatic improvement over standard static regressions. The static model was simply not suited to the true dynamic problem. As the profession later realized, it was an inappropriate model for the task. Even worse, it lent itself to the abuses McCloskey would come to criticize.

The autoregressive model acknowledged the difference between the short and long run in a way that static regressions could not. This led to something of a split in the literature based on the finding (or lack of finding) unit roots. Making an analogy to the world of athletics, Taylor (2001) refers to the opposing strands of literature as “teams,” and summarizes the schism thus:

Differences between the two teams center on whether this [AR(1)] model holds, or in what form, for the contemporary period. One team may be termed the “whittling down half-lives” team; for them, the half-lives of deviations are small and reasonable, and price gaps are stationary; but even here, half-lives tend to be measured in years. On the other side is the “whittling up half-lives” team; for them, half-lives are much longer than seems reasonable, and could even be

infinite; price gaps might follow a random walk. Considerable ink has been spilled on both sides of this debate, and still no broad agreement seems at hand. The sense of conundrum is only furthered by the realization that, for the most part, studies have been run on essentially the same model and the same data, deriving different conclusions from slight changes in samples, pooling and panel techniques, stationarity tests, and the like. (p. 474)

Taylor (2001) follows this up by aligning himself with the “whittling down half-lives” strand of literature. There is an obvious similarity to the way the previous generation’s literature was divided between those researchers who found  $\beta = 1$  and those who did not. Eliminate the words “stationarity tests” from the statement and it would be difficult to tell if the last half of that paragraph was written about model (1) or model (2).

The primary contribution of Taylor (2001) was to demonstrate that two pervasive problems can affect inference in the canonical AR(1) model: temporal aggregation (sampling the data at lower frequencies than the frequency of the true data generating process) and nonlinearity. He shows that both problems can bias the estimate of the half-life upward. Combining these problems leads to drastic reductions in the power of the DF test for a unit root. Since the DF test has low power to begin with, especially with a short span of data, this is a real concern. Consider a span of data equaling 25 times the half-life of the deviations. If the true process converges at a certain rate per day (high frequency) with a half-life of 30 days, but the data is sampled only monthly (low frequency) as opposed to daily, the power of the test drops from 0.65 to 0.21 (Taylor 2001 p. 488). This drop in power is from temporal aggregation alone. Misspecification resulting from modeling the process as linear when the true process is nonlinear decreases power even further.

Recent literature has been working to address the nonlinearity issue. Today’s tests of PPP and other market efficiency questions utilize cointegration techniques, error correction models, and threshold autoregressive models (TARs). TARs have spawned a growing literature and produced interesting results.

There are several variations of the TAR model. Consider the version presented in Taylor (2001):

$$x_t = \begin{cases} c + \rho(x_{t-1} - c) + \varepsilon_t & \text{if } x_{t-1} > c \\ x_{t-1} + \varepsilon_t & \text{if } c \geq x_{t-1} \geq -c \\ -c + \rho(x_{t-1} + c) + \varepsilon_t & \text{if } -c > x_{t-1} \end{cases} \quad (3)$$

Model (3) has a “band of inaction” which Taylor (2001) compares to the “gold points” of the classical gold standard. In this region, between  $-c$  and  $c$ , the process follows a random walk. Outside the band, the process follows an AR(1) reverting back to the boundary of the band. In this case, the process has a band that is centered around zero, and the convergence rate is the same on either side of the band. In general, this need not be the case. Samanta and Zadeh (2001) illustrate a TAR with an asymmetric adjustment process. The similarity of this structure to the gold points of the classical gold standard suggests that this model may be well suited to arbitrage problems with transaction costs or other frictions. Certainly this is a prime motivation for using a model of this type. When transaction costs prevent arbitrage, the deviations from PPP (or interest parity, etc)

follow a random walk. Outside the band, when gains from arbitrage exceed the transaction costs, the process reverts to equilibrium. In fact, a generalization of the following form is also possible:

$$x_t = \begin{cases} c + \rho_1(x_{t-1} - c) + \varepsilon_t & \text{if } x_{t-1} > c \\ \rho_2 x_{t-1} + \varepsilon_t & \text{if } c \geq x_{t-1} \geq -c \\ -c + \rho_3(x_{t-1} + c) + \varepsilon_t & \text{if } -c > x_{t-1} \end{cases} \quad (4)$$

This generalization not only allows for asymmetric adjustment, but for reversion to equilibrium within the band perhaps at a different rate. The standard AR(1) model is simply a special case of (4). Thus, proper use of this technique requires testing (4) against the alternative of a standard AR(1) model. If the hypothesis of linearity is rejected, then it is appropriate to use a model such as (4). If the true data generating process is characterized by (4), then the power of the test for a unit root ( $\rho_1 = \rho_2 = \rho_3 = 1$ ) in model (4) should be better than the power of the test in the standard AR(1) model, as Taylor (2001) illustrates.

For an application to interest rate adjustment, see Enders and Granger (1998). Tsay (1989) is an early exposition of the method. Dufrénot and Mignon (2002) present a clear treatment of a variety of TAR models. They demonstrate the power of the tests, and show the application to finance and economics.

## MEASUREMENT AHEAD OF THEORY?

In an important paper, Edward Prescott (1986) made the claim that theoretical advances in quantitative macroeconomics (real business cycle or dynamic stochastic general equilibrium theory) had pushed beyond the profession's current ability to appropriately measure variables of interest. Prescott's concluding statement summarizes the argument.

The match between theory and observation is excellent, but far from perfect. The key deviation is that the empirical labor elasticity of output is less than predicted by theory. An important part of this deviation could very well disappear if the economic variables were measured more in conformity with theory. That is why I argue that theory is now ahead of business cycle measurement and theory should be used to obtain better measures of the key economic time series. Even with better measurement there will likely be significant deviations from theory which can direct subsequent theoretical research. This feedback between theory and measurement is the way that mature, quantitative sciences advance. (p. 21)

Whether Prescott's words still ring true about the state of business cycle theory over 15 years after he made them is certainly beyond the scope of this paper. The final sentence of the quote, however, is directly relevant to the matter at hand, and is consistent with the spirit of McCloskey's critiques.

In matters of financial economics (market efficiency questions such as PPP, UFRH, interest rate arbitrage, and others), I propose that the opposite of Prescott's

statement is true. The third generation of models designed to test market efficiency hypotheses can utilize the wealth of data generated and gathered each day in the financial markets as well as the goods markets. Perhaps the only variables which are not measured with great frequency are the price levels themselves. However, these are also typically less volatile. Exchange rates, forward rates, and interest rates are measured at very high frequencies. Modern financial practice involves complicated strategies to arbitrage in these markets *very* quickly. Any relevant information tends to be quickly reflected in the price of assets. Prices are reported in, for all practical purposes, continuous time as even the small investor can obtain real-time quotes on his or her computer. Our ability to measure, record, and communicate the prices of even the most esoteric derivative securities is remarkable. The newest econometric models in this area are, in fact, measurement tools. They measure, among other things, the speed of convergence and the width of the “band of inaction.” These are variables of scientific interest.

Theory in this area has been slower to progress. What is missing from the literature is an equilibrium economic model that links the goods market to the financial markets in such a way that it can attempt to match the data and provide feedback to the measurement process. This is critical. Measurement without theory is uninformative. Suppose we find the half-life of PPP deviations to be on the order of 3 or 4 years, which may seem to be long. However, McCloskey wants us to always ask, “How large is large?” Are different processes at work in financial markets and goods markets enough to slow convergence? Does price stickiness in goods markets contrast enough with efficient arbitrage in financial markets that prices in these markets wander away from equilibrium relationships for that long?

A concrete example of the problem that can arise when there is a gap between theory and measurement is in Samanta and Zadeh (2001). In a TAR model of the UFRH, they find evidence of an asymmetric adjustment process. However, they conclude, “If it is known *a priori* that the positive phase is more persistent than the negative phase, it creates an opportunity for the speculators in the currency markets to use this information to increase their profits. It also provides a different but important input in the intervention and innovation policies for the central bank” (p.33). In response to the first sentence, it is not at all clear that speculators could increase their profits from knowing that the adjustment process is asymmetric, any more than it was clear that speculators could profit from finding that  $\beta = 1$  in the first generation models. We do not know why the process is asymmetric. A theoretical model is the place to explore that idea, and it has not yet been done. Furthermore, asymmetry aside, why are speculators not already responding to the fact that deviations on either side of the band tend to persist? Go thee and prosper! Yet another problem is that the data used in that paper is monthly, whereas speculators would likely operate using daily data. Temporal aggregation remains a serious problem for both theory and measurement. The second sentence echoes McCloskey and Zecher’s (1984) discussion of the Mars vs. Iowa City approaches to monetary policy. Is the asymmetry, let alone the fact that deviations persist at all, something that central banks can use to their advantage to manipulate either domestic variables or exchange rates in the short run? Are the half-lives in question long enough to imply arbitrage opportunities are still available or that monetary policy is more or less effective? These are precisely the questions raised by McCloskey in response to the first

generation of PPP models. We still do not know the answers, but the third generation of models will undoubtedly help to answer them, if theory can catch up with measurement.

One positive step in this direction was made by Obstfeld and Rogoff (2001). They present some useful modeling strategies and apply them to some of the common puzzles of international macroeconomics and finance. They claim that transport costs and non-traded goods play an important role. They cite recent research into the nonlinear models attempting to, in Taylor's (2001) terminology, "whittle down" half-lives. Yet even in this work which surveys and adds to the state-of-the-art in this field, there is still no dynamic theoretical model that quantitatively replicates the long half-lives. Such a model would certainly have to include monopoly pricing power as well as financial variables.

Another positive step was taken by Backus, Kehoe, and Kydland (1992) who develop a quantitative model to explain international business cycles and the comovements of components of GDP across countries. However, since their model measures components of GDP, it is a low frequency model and therefore would not ultimately resolve the PPP puzzle. To be sure, a theoretical model that proposes to add to the discourse on PPP will need to accommodate high frequency financial market data as well as low frequency goods market data as exchange rates are determined by the interaction of agents acting on high and low frequency fluctuations.

## CONCLUSION

Research into issues of market efficiency, most notably the famous PPP question, has progressed through three generations of models. These issues present some of the most pressing unsolved puzzles in macroeconomics and finance. Each puzzle has some similarities and differences to the others, suggesting that there may be common factors (see Obstfeld and Rogoff (2001)). However, our ability to gather data on the variables of interest – empirical estimates of the width of the "band of inaction" for example – exceeds our ability to explain them using theoretical, quantitative, equilibrium models.

The importance of this task is underscored by McCloskey's critiques of econometric practice. At the heart of McCloskey's critique of the first generation econometric models of the PPP question was the lack of theory informing the hypothesis testing. Quite simply, McCloskey wants the researcher to ask, "How large is large?" Are deviations from PPP a sign that profitable arbitrage opportunities exist or that the Fed can influence domestic prices without affecting the exchange rate? The standard against which the estimate is compared may very well be different depending on the question.

Applied econometric research into market efficiency questions is improving the way we measure variables that can inform the next generation of theoretical models. For example, half-lives of deviations from PPP of the width of the "band of inaction" should be the kinds of variables included in theoretical models. Then the researcher can ask the model if the estimates of these variables imply whether, for example, central banks can successfully intervene in currency markets. Econometric models in which we simply accept or reject unit roots cannot by themselves answer these questions. Dynamic theoretical models and simulations that can replicate observed data are essential to determining the economic significance of statistical findings. This kind of quantitative

science is possible given recent advances in measurement through applied econometrics, but it must be used to inform the theory in order for science to advance.

#### ENDNOTES

1. For more background on the PPP question, two excellent surveys are Rogoff (1996) and Froot and Rogoff (1995).
2. McCloskey refers to herself as a Cassandra in her column in the *Eastern Economic Journal* (Summer 1999).
3. This is particularly distressing since knowing the means and standard deviations of the variables under study is essential for quantitative science.
4. Of course, in citing this passage, McCloskey and Ziliak (1996) are quick to point out that it is unclear (and not explained by Griliches) why statistical significance is a relevant metric for comparing different versions of the model.
5. In my experience teaching this to undergraduate students of international economics, it makes a very convincing argument.
6. In fact, in her 1996 work, *The Vices of Economists -- The Virtues of the Bourgeoisie*, this particular example of the argument does not appear.
7. This argument also appears in McCloskey (1994) and in an extended and very readable form in McCloskey (1985).
8. For more examples of 1970s vintage PPP tests, see the special issue of the *Journal of International Economics* from May of 1978 (Dornbusch and Jaffee 1978).
9. McCloskey (1996) mentions the unit root literature only briefly.

#### REFERENCES

- Backus, D., Kehoe, P., and Kydland, F. 1992. International Real Business Cycles. *Journal of Political Economy* 100 (4), 745-775.
- Banerjee, A., Dolado, J., Galbraith, J., and Hendry, D. 1993. *Co-integration, Error Correction, and the Econometric Analysis of Non-stationary Data*. Oxford: Oxford UP.
- Dickey, D. and Fuller, W. 1979. Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association* 74(366), 427-431.
- , 1981. Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica* 49(4), 1057-1072.
- Dornbusch, R. and Jaffee, D. 1978. Purchasing Power Parity and Exchange Rate Problems: Introduction. *Journal of International Economics* 8, 157-161.
- Dufrénot, G. and Mignon, V. 2002. *Recent Developments in Nonlinear Cointegration with Applications to Macroeconomics and Finance*. Boston: Kluwer.

Enders, W. and Granger, C. 1998. Unit Root Tests and Asymmetric Adjustment With an Example Using the Term Structure of Interest Rates. *Journal of Business and Economic Statistics* 16(3), 304-311.

Engel, C. 2000. Long-Run PPP May Not Hold After All. *Journal of International Economics* 51(2), 243-273.

Frankel, J. and Rose, A. 1996. A Panel Project on Purchasing Power Parity: Mean Reversion Within and Between Countries. *Journal of International Economics* 40, 209-222.

Froot, K. and Rogoff, K. 1995. Perspectives on PPP and Long-Run Real Exchange Rates, in G. Grossman and K. Rogoff, eds. *Handbook of International Economics*, vol. 3. Amsterdam: North-Holland.

Griliches, Z. 1986. Productivity, R&D, and Basic Research at the Firm Level in the 1970s. *American Economic Review* 76 (1), 141-154.

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

Leamer, E. 1983. Let's Take the Con Out of Econometrics. *American Economic Review* 73(1), 31-43.

Lipsey, R. 1984. Comment on The Success of Purchasing Power Parity, in M. Bordo and A. Schwartz eds. *A Retrospective on the Classical Gold Standard, 1821-1931*. Chicago: U of Chicago Press.

McCloskey, D. 1985. *The Rhetoric of Economics*. Madison: U of Wisconsin Press.

----- . 1994. *Knowledge and Persuasion in Economics*. Cambridge: Cambridge UP.

----- . 1996. *The Vices of Economists – The Virtues of the Bourgeoisie*. Amsterdam: Amsterdam UP.

----- . 1999. Cassandra's Open Letter to Her Economist Colleagues. *Eastern Economic Journal* 25 (Summer), 357-363.

McCloskey, D. and Zecher, J. 1984. The Success of Purchasing Power Parity, in M. Bordo and A. Schwartz eds. *A Retrospective on the Classical Gold Standard, 1821-1931*. Chicago: U of Chicago Press.

McCloskey, D. and Ziliak, S. 1996. The Standard Error of Regressions. *Journal of Economic Literature* 34, 97-114.

- Neyman, J. and Pearson, E. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character* 231, 289-337.
- Obstfeld, M. and Rogoff, K. 2001. The Six Major Puzzles in International Macroeconomics: Is There a Common Cause, in *NBER Macroeconomics Annual vol. 15*. Cambridge, MA: MIT Press.
- Phillips, P. and Perron, P. 1988. Testing for a Unit Root in Time Series Regression. *Biometrika* 75 (2), 335-346.
- Prescott, E. 1986. Theory Ahead of Business Cycle Measurement. *Federal Reserve Bank of Minneapolis Quarterly Review* 10(4), 9-22.
- Rogoff, K. 1996. The Purchasing Power Parity Puzzle. *Journal of Economic Literature* 34, 647-668.
- Samanta, S. and Zadeh, A. 2001. Foreign Exchange Rates, Asymmetric Adjustment and Threshold Co-integration: Empirical Evidence from Canada. *Journal of Economics* 27(2), 19-35.
- Taylor, A. 2001. Potential Pitfalls for the Purchasing-Power-Parity Puzzle? Sampling and Specification Biases in Mean-Reversion Tests of the Law of One Price. *Econometrica* 69(2), 473-498.
- Tsay, R. 1989. Testing and Modeling Threshold Autoregressive Processes. *Journal of the American Statistical Association* 84(405), 231-240.
- Wald, A. 1939. Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *Annals of Mathematical Statistics* 10 (4), 299-326.