EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 13(1)
January 2016: 5–45

A Unit Root in Postwar U.S. Real GDP Still Cannot Be Rejected, and Yes, It Matters

David O. Cushman[1]

LINK TO ABSTRACT

Does real gross domestic product (GDP) have an autoregressive unit root or is it trend stationary, perhaps with very occasional trend breaks? If real GDP has an autoregressive unit root, then at least a portion of each time period's shock to real GDP is permanent. Booms and recessions therefore permanently change the future path of real GDP. But if real GDP is trend stationary, (almost all) shocks to real GDP are transitory, and real GDP (almost always) reverts to an already existing trend.[2]

Here I present evidence that, despite the findings or implications of some recent papers and blog entries, the hypothesis of a unit root in U.S. real GDP after World War II cannot be rejected, once the issues of individual and multiple-test size distortion are dealt with. But are the implied permanent shocks important relative to transitory shocks? They are, according to an unobserved components model and a vector error correction model (VECM) proposed by John Cochrane (1994). Moreover, permanent shocks have strong effects according to impulse response analysis of the Cochrane VECM. Finally, models with unit roots forecast better over the postwar period than trend stationary models. This is particularly true for forecasts of recoveries from seven postwar recessions, and Cochrane's VECM, with its specific identification of permanent and transitory shocks, performs best

1. Westminster College, New Wilmington, PA 16172; University of Saskatchewan, Saskatoon, SK S7N 5A5, Canada.

2. Cochrane (2015b) presented a fine graph illustrating some possibilities: a random walk (pure unit root), a unit root plus a stationary component, and a stationary process.

overall. Thus, by every measure the unit root has "oomph," as Stephen Ziliak and Deirdre McCloskey would say.[3]

Background

In a landmark paper, Charles Nelson and Charles Plosser (1982) argued against the then-prevailing assumption (noted by Stock and Watson 1999) of trend stationarity for U.S. real GDP. Nelson and Plosser showed that the unit root could not be rejected in favor of trend stationarity. Their claim has led to a large number of papers examining the issue with a variety of tests, specifications, data periods, and countries. For U.S. real GDP, James Stock and Mark Watson (1999) judged that the unit root testing literature supported Nelson and Plosser's (1982) conclusion. More recently, Spencer Krane (2011) concluded that the Blue Chip Consensus real GDP forecasts have important permanent components. Sinchan Mitra and Tara Sinclair (2012) stated that "fluctuations in [U.S. real GDP] are primarily due to permanent movements." But other papers since Stock and Watson's (1999) judgment have claimed that the unit root *can* be rejected (see, e.g., Ben-David, Lumsdaine, and Papell 2003; Papell and Prodan 2004; Vougas 2007; Beechey and Österholm 2008; Cook 2008; Shelley and Wallace 2011).

In recent years, the unit root-GDP issue has made lively appearances in in economics blogs. Greg Mankiw (2009a) argued that the likely presence of a unit root cast doubt on the incoming Obama administration's 2009 forecast of strong recovery (Council of Economic Advisers 2009). Brad DeLong (2009) and Paul Krugman (2009) immediately disputed Mankiw's invocation of a unit root as cause for pessimism, arguing that the shocks that generated the 2008–09 recession were transitory, although Krugman allowed that shocks could occasionally be permanent (Cushman 2012; 2013). In commenting on the Mankiw-DeLong-Krugman controversy, Menzie Chinn (2009) ran an econometric test that rejected the unit root. Stephen Gordon (2009), responding in turn to "DeLong-Krugman-Mankiw-Chinn," said: "There's little evidence that we should be operating under the unit-root hypothesis during recessions. Or during expansions, come to that." He provocatively followed this with: "The world would be a better place if…the whole unit root literature never existed."

Cochrane (2012) wrote that "the economy has quite reliably returned to the trend line after recessions," providing a GDP graph with a trend line fitted over 1967–2007. This claim seems to support the trend stationarity hypothesis. But Cochrane has believed for some time that GDP has a unit root (Cochrane 1994;

3. *See* Ziliak and McCloskey 2008; McCloskey and Ziliak 2012.

2015a). The Cochrane (2012) graph may be a dramatic illustration of the blog entry's title ("Just How Bad Is the Economy?") and reflect a belief that shocks causing recessions were mostly transitory prior to 2007. Cochrane's blog entry cited John Taylor (2012), who wrote (citing in turn Bordo and Haubrich 2012) that "deep recessions have always been followed by strong rather than weak recoveries." Similarly, Marcus Nunes (2013) presented a graph that looks like a linear trend stationary model with a break in 2008 from a one-time shock that is permanent and transitory in roughly equal proportions. Nevertheless, although Taylor's presentation and certainly Nunes's are suggestive of the trend stationary hypothesis, Taylor and Nunes did not mention unit roots or stationarity, and so it is not clear what, if anything, they may have believed about the issue. Meanwhile, William Easterly (2012) wrote, facetiously, that the unit root question could be a factor in the 2012 U.S. presidential election, because a unit root in output could be seen as relieving President Obama of blame for the weak recovery. Scott Sumner (2014) pointed out that Mankiw's (2009a; 2009b) unit root-based skepticism about recovery by 2013 had been confirmed.[4]

Quite recently, Roger Farmer (2015) applied a unit root test to U.S. real GDP 1955:2–2014:1. The result was a clear failure to reject the unit root. Farmer concluded: "There is no evidence that the economy is self-correcting."[5] Arnold Kling (2015) used Farmer's result to argue that "the concept of potential GDP can have no objective basis." In contrast, John Taylor (2015) reprinted a graph (from Taylor 2014) showing a linear trend based on the 2001–2007 GDP growth rate that extends through 2024, a trend which he argued could be reached if certain policies are followed. This looks like trend stationarity. Finally, Larry Summers (2015) wrote that for "over 100 recessions [in] industrial countries over the last 50 years…in the vast majority of cases output never returns to previous trends" (referring to Blanchard, Cerutti, and Summers 2015). This implies a unit root.

Thus, seventeen years after Stock and Watson's (1999) judgment about GDP and unit roots, the issue remains controversial.

4. Sumner (2014) was, however, incorrect that the presence of a unit root would imply no "bounceback" after a recession. A variable with a unit root process can also have a stationary component, which means that a recession can have been partly caused by a transitory shock—see Cochrane's (2015b) graph.

5. Farmer used the wrong critical value (I noted this in a comment to the blog entry, which he graciously acknowledged), but correcting the mistake actually strengthens his conclusion. He also ran a KPSS test, which resoundingly rejected the null of trend stationarity.

A summary of what I do and find

I start by analyzing whether the recent unit root rejections in the academic literature (and one in a blog), when combined with earlier rejections, are numerous or strong enough to overturn the 1999 Stock-Watson assessment in favor of unit roots in postwar real GDP. I implement corrections for two problems that cast doubt on the various existing rejections of the unit root in real GDP:

1. Size distortion in individual unit root tests: the tendency of a test to reject at rates exceeding the stated significance level when the null is actually true. The distortion results from the pre-testing for lag order and from the magnitude of the lag parameters, and has seldom been addressed in papers that reject the unit root in real GDP.
2. The multiple-test problem: what to decide when some tests reject but others do not, while maintaining the stated significance level. This has not been addressed in any papers on the GDP/unit root question.

For problem (1), I use bootstrapped distributions to get size-adjusted *p*-values for individual tests. For problem (2), I use the bootstrapped distributions to get several variations of size-adjusted joint *p*-values for various sets of unit root tests. The most comprehensive joint *p*-values show no rejections at the 0.05 level, and hardly any at the 0.10 level. I thus conclude that the unit root null cannot be rejected.

I then turn to the question of the economic importance of the unit root. The question matters because the unit root could be dominated by a transitory component, rendering its statistical significance economically unimportant. I therefore compute indicators of the relative importance of permanent shocks using an unobserved components approach, as in Krane (2011) and Mitra and Sinclair (2012), and the vector error correction approach of Cochrane (1994). The estimates of permanent shocks are economically significant.

Finally, I examine how allowing for unit roots would have affected forecasts of real GDP in the postwar period. Lawrence Christiano and Martin Eichenbaum (1990) and Cochrane (1991a; 1991b) pointed out that a unit root process could be very closely approximated by a stationary process. Nevertheless, I find that specifying a unit root usually does improve forecasts, particularly after recessions. The advantage is most consistent using Cochrane's (1994) VECM that separately identifies both transitory and permanent shocks.

The unit root test literature on postwar U.S. real GDP, and its interpretation

Table 1 lists papers over the last 31 years that have applied unit root tests to postwar U.S. real GDP.[6] The null hypothesis in these papers is that a unit root is present. I am unaware of any papers in peer-reviewed journals focusing on U.S. real GDP for which trend stationarity is the null.[7] This is consistent with a sentiment, attributed to Bennett McCallum, that "it seems strange that anything could be trend stationary" (see McCallum 1991).

In indicating the unit root rejections in Table 1, I have corrected two erroneous interpretations by Vougas (2007). He reported that his PT and LP-CC test values rejected the unit root, but he apparently misinterpreted the tests' rejection regions; the two results are not rejections. See Appendix 1 for details. I use the correct interpretation for the two tests because I don't want my assessment of the prevalence of unit root rejections affected by this problem.

In Table 1 there are 33 test results reported, and 12 (36 percent) reject the unit root at the 0.05 level.[8] The table also presents the state of affairs regarding several issues that could affect test outcome or validity:

1. Data period: Tests with more data will, ceteris paribus, have more power.
2. Trend specification: A simple linear trend (under alternative hypothesis of stationarity) has been the most common specification, but some tests have specified trends with breaks to allow for infrequent permanent shocks. Other tests have specified nonlinear trends to allow for gradual evolutions in growth rates. If stationarity is the case and real GDP has a breaking or nonlinear trend, then a test that correctly specifies the trend will tend to have more power, although estimating the extra trend parameters tends to offset the power gain.

6. I do not include the few papers that have tested for a unit root in U.S. per capita real GDP (e.g., Stock and Watson 1986; Cheung and Chinn 1997), only in real GDP itself, the more common approach. The table contains standard journal articles with two exceptions, Chinn (2009) and Farmer (2015), which are blog entries. One might not want to include work found in blogs because blogs are not vetted for their econometrics as refereed journal articles normally are. But in this case I include the blog entries as an illustration of the continued interest in applying unit root tests to GDP, and because Chinn and Farmer are highly respected scholars.

7. Cheung and Chinn (1997) test the null of trend stationarity in postwar *per capita* real GNP. Chinn (2009) and Farmer (2015) apply the KPSS test in their blogs.

8. If we count the two erroneous Vougas (2007) rejections, the proportion rises to 42 percent.

3. The adjustment process: The KSS test, applied three times, features nonlinear adjustment. With traditional, so-called 'linear' adjustment specified in a test, then—under stationarity and in the absence of further shocks—a constant fraction of the deviation from the long-run trend is assumed to be eliminated in each time period. The more recent KSS approach is to specify 'nonlinear' adjustment in unit root tests. With nonlinear adjustment, the fraction of the deviation from trend that is eliminated is assumed *larger* when the variable is far from the trend than when it is close. The motivation to model real GDP this way, as given by Meredith Beechey and Pär Österholm (2008), is that the central bank would be likely to respond disproportionately more strongly to significant inflationary booms and recessions than to small fluctuations. If, under stationarity, adjustment is indeed nonlinear and the test's specification of the nonlinear adjustment is reasonably accurate, then the test could have higher power.
4. Sampling distributions: Christian Murray and Charles Nelson (2000) emphasized that unit root tests were likely to suffer from size distortion: excessive rejections under the null, because of data heterogeneity and data-based lag selection. The lag selection issue is that unit root tests often require an autoregressive lag order choice, which is usually based on the same data as used for the test itself. G. William Schwert (1989) and Glenn Rudebusch (1993) discussed another problem: the effect of specific serial correlation parameter values of relevant sign and/or sufficient magnitude. Murray and Nelson's (2000) response was, in addition to using only postwar data, to employ data-based sampling distributions—bootstrapping—and to include the lag choice decision process in the bootstrapping. Schwert (1989), Rudebusch (1993), and Eric Zivot and Donald Andrews (1992) had also previously used data-based sampling distributions to address size distortion.

In Table 1, the reported unit root rejections tend to be associated with more recent papers having longer data sets (10 of the 12 test rejections are in papers published after 2000), with specifications including breaking or nonlinear trends, and/or with specifications with nonlinear adjustment to trend. However, with just one exception the rejections also come from sampling distributions that are not data-based. So, do the rejections reflect higher power from longer data sets and better trend or adjustment specifications, or do the rejections just reflect size distortion from using non-data-based distributions?

TABLE 1. Various papers and their unit root conclusions

Paper	Unit root test	Time period	Trend	Data-based p-values?	Reject unit root at 0.05 level?
Stulz and Wasserfallen (1985)	ADF	1947:1–1981:4	Linear	No	No
Stock and Watson (1986)	ADF, Zα	1950:1–1984:4	Linear	No	No (2)
Evans (1989)	ADF	1951:1–1985:4	Linear	No	No
Perron (1989)	ADF-break	1947:1–1986:3	1 slope change	No	Yes
Zivot and Andrews (1992)	ZA-B	1947:1–1986:3	1 slope change	Yes	No
Rudebusch (1993)	ADF	1948:3–1988:4	Linear	Yes	No
Duggal et al. (1999)	ADF-exp	1960–1989	Exponential	No	Yes
Murray and Nelson (2000)	ADF, DF-GLS, PP-Z*t*	1947:1–1997:3	Linear	Yes	No (3)
Murray and Nelson (2000)	ZA-B	1947:1–1997:3	1 slope change	Yes	No
Vougas (2007)	ADF, DF-GLS, SP, PT	1947:1–2004:4	Linear	No	Yes (1), No (3)
Vougas (2007)	ADF, SP	1947:1–2004:4	Quadratic	No	Yes (1), No (1)
Vougas (2007)	ZA-C	1947:1–2004:4	1 slope and level change	No	Yes
Vougas (2007)	LP-CC	1947:1–2004:4	2 slope and level changes	No	No
Vougas (2007)	LNV-B	1947:1–2004:4	1 smooth level change	No	Yes
Vougas (2007)	LNV-C	1947:1–2004:4	1 smooth slope and lev. chg.	No	Yes
Cook (2008)	ADF, DF-GLS	1955:1–2004:1	Linear	No	No (2)
Cook (2008)	KSS*	1955:1–2004:1	Linear	No	Yes
Beechey and Osterholm (2008)	ADF, DF-GLS	1947:1–2005:2	Linear	No	No (2)
Beechey and Osterholm (2008)	KSS*	1947:1–2005:2	Linear	No	Yes
Chinn (2009)	ADF, DF-GLS	1967:1–2008:4	Linear	No	Yes (2)
Shelley and Wallace (2011)	KSS*	1947:1–2005:2; –2009:3	Linear	Yes	Yes (1); No (1)
Farmer (2015)	ADF	1955:2-2014:1	Linear	No	No

*Unlike the other unit root tests in the table, the KSS test specifies nonlinear rather than linear adjustment to trend under the trend stationary alternative hypothesis.
Test definitions and sources:
ADF = augmented Dickey-Fuller test with linear or quadratic trend (Ouliaris et al. 1989)
ADF-exp = ADF test with exponential trend (Duggal et al. 1999)
ADF-break = ADF test of residuals of the regression of *y* on a trend with a slope change at a known date (Perron 1989)
DF-GLS = modified ADF test of Elliott et al. (1996) with linear or quadratic trend
KSS = Kapetanios et al. (2003) test with linear trend
LNV-B, LNV-C = Leybourne et al. (1998) test, models B and C
LP-CC = Lumsdaine and Papell (1997) test, model CC
PP-Z*t* = Phillips and Perron (1988) Z(*t*) test with a linear trend
PT = Point optimal test of Elliott et al. (1996) with linear trend
SP = Schmidt and Phillips (1992) tau test with linear trend or quadratic trend
Zα = Phillips (1987) Zα test modified with assumed trend
ZA-B, ZA-C = Zivot and Andrews (1992) models B and C

There is another issue, however, for which the table does not present evidence. Is the rejection by 36 percent of tests, as in Table 1, sufficiently strong evidence against the unit root for an overall rejection? When a null is true and many tests are applied either by one researcher or by many, getting at least one rejection becomes increasingly likely. The overall significance of results such as in Table 1 is thus unknown without further work. The problem is a version of the classic multiple test problem, which is the problem of getting correct significance levels when there are many tests using a single sample (see on this Romano et al. 2010).

Suppose 10 tests of the unit root null are applied to a data set (as in Vougas 2007). If, under the null, the 10 tests are uncorrelated with each other, then we will observe one or more rejections not 5 percent of the time, the nominal level of significance, but 40 percent of the time (from the binomial distribution). And when we get rejections, there will be two or more 21 percent of the time. But the unit root test results are hardly uncorrelated because they are applied to the same data set.

For a concrete example, suppose the correlation for each test pair (with 10 tests, there are 45 pairs) under the null is 0.80 and the individual test statistics happen to be standard normal. Then the probability of getting one or more rejections falls to 16 percent, but this still exceeds 5 percent. Moreover, if we get any rejections in this case, there will now be two or more about 40 percent of the time. And there will be four rejections (approximately the proportion in Table 1) or more 33 percent of the time. Overall, under the null, observing more than one rejection among 10 tests with a probability significantly exceeding the stated 5 percent significance level is quite likely.[9]

Unit root tests with the above issues addressed

I apply a large number of previously used unit root tests and variations, 42 tests in all. The same estimation period is used for all, so different conclusions from the tests do not reflect different time periods.[10] To account for individual-test size distortion, I employ a bootstrap approach using five data-generating processes (DGPs). To account for the multiple test problem, for each DGP I compute joint *p*-values in four ways for all 42 tests and several subsets. And then I compute overall joint *p*-values of the four joint *p*-values for the various subsets. Finally, because results differ among the DGPs, I try several ways of weighting the DGPs' joint *p*-values to yield an overall *p*-value for each group of tests using each *p*-value method.

9. Under independence, the probabilities can be calculated using the binomial distribution. For the correlated case, the probabilities are estimated with 40,000 replications assuming a multivariate normal distribution for the test statistics.

10. Generally, TSP 5.1 is used for econometric computations in the paper. Exceptions are noted below.

Data set

The data sets in most papers in Table 1 start in 1947:1 (or not too long thereafter) and end shortly before the paper was written, and I could follow this pattern. But Gary Shelley and Frederick Wallace (2011) find no rejections, contrary to Beechey and Österholm (2008), when the data are extended to include the 2008–2009 recession. Farmer's (2015) data also includes the recession, and he reports non-rejection. This could reflect a permanent shock to a generally trend stationary real GDP. Such a break adds a challenge to the rejection of the unit root (supposing it should be rejected) that earlier papers did not face. But I do not want my critique to be influenced by this problem. Therefore, my data set starts in 1947:1 and ends not with the most recent data available to me, but in 2007:3, just before the recession's start in 2007:4.[11]

The unit root tests

I employ all the tests and trend specifications in Table 1, but with a few changes. Regarding trends, I omit one and add one. The omitted trend is the exponential trend of Vijaya Duggal, Cynthia Saltzman, and Lawrence Klein (1999), who included the term $\exp(t^{\Phi})$ as the (only) deterministic term in an ADF test. But the estimated trend curvature and the unit root test result are then dependent on setting the initial date for t equal to 1. An exponential trend without this defect is used by Robert Barro and Xavier Sala-i-Martin (1992), but there is no unit root test with this trend. For postwar U.S. real GDP, however, the Barro and Sala-i-Martin (1992) exponential trend and the quadratic trend are almost identical, so the unit root tests with a quadratic trend approximate the Barro and Sala-i-Martin trend.[12]

The trend I add has two slope changes *without* level changes. This extends the Zivot and Andrews (1992) one-slope-change model B to the two-slope-change case and is therefore like the Robin Lumsdaine and David Papell (1997) model CC but with no level changes. I label the result ZA-LP. Thus, we now have one- *and* two-slope-change tests both with level changes and without them.

The additional adjustments to the Table 1 tests are the following. First, instead of Stock and Watson's (1986) inclusion of an assumed trend in Peter Phillips's (1987) $Z\alpha$ test, I follow Phillips and Pierre Perron's (1988) inclusion of an estimated trend. I label the latter version PP-$Z\alpha$. The second adjustment is to substitute the superior DF-GLS test for the regular ADF test in the case of the quadratic

11. Real GDP is '1 decimal' in chained 2005 dollars, downloaded from FRED on 30 January 2011, in logs.
12. If we drop the random errors and assume constant technology and population growth, the trend in Barro and Sala-i-Martin (1992) becomes $y = a_0 + a_1 t + [1 - \exp(\beta_0 + \beta_1 t)]$.

trend, and the third is to substitute the modified feasible point optimal test (MPT) from Serena Ng and Pierre Perron (2001) for the unmodified version (PT) of Graham Elliott, Thomas Rothenberg, and James Stock (1996).[13] Next, I do not use Perron's (1989) known-breakpoint approach, only the Zivot and Andrews (1992) estimated-breakpoint version, ZA-B. And, because the unit root rejections in Table 1 for the three KSS applications with a simple linear trend suggest relevance for nonlinear adjustment, I add tests with KSS nonlinear adjustment for all other trend specifications.[14]

To adjust for serial correlation, the tests in Table 1 almost always use autoregressive lags. The most common method for lag order choice is the AIC, followed by the BIC and the Ng and Perron (1995) recursive test-down. Steven Cook (2008) simply employs the same lag order every time. Consistent with accounting for many test procedures, I try both the AIC (or the Ng and Perron 2001 modification, MAIC) and BIC for all autoregressive cases. But time is limited so I omit the test-down and fixed-lag approaches. The exceptions in Table 1 to using autoregressive lags are the nonparametric approaches in the $Z\alpha$, PP-$Z\alpha$, and PP-Zt tests and in Vougas's (2007) implementation of the PT test. I use the standard PP-$Z\alpha$ and PP-Zt tests but, because I replace the PT test with the MPT test, I follow Ng and Perron's (2001) use of autoregressive lags for the MPT test. And I use the MAIC for the DF-GLS and MPT tests as in Ng and Perron (2001).

In choosing the lag, the maximum considered is usually 14, from *maxlag* = Int[12(T/100)$^{0.25}$] (Schwert 1989) where T = 243, the sample size. The exceptions are the tests from Peter Schmidt and Peter Phillips (1992). With linear trend, *maxlag* = 14 gives lag order 1, not 2 as in Vougas (2007), and then the unit root is not rejected, unlike in Vougas (2007). Perhaps he used a smaller *maxlag*. With *maxlag* = 12, I get his result. I do not want my critique of unit root rejections to be affected by this minor distinction, and so for the Schmidt-Phillips tests I set *maxlag* = 12.[15]

Bootstrapping and exact *p*-value procedure for the individual tests

I compute individual test *p*-values using a wild bootstrap that adjusts for size distortion from the specific lag coefficients in DGPs and heteroskedasticity from the Great Moderation, viz., the reduction in real GDP variance from 1984 onwards

13. Cushman (2002) is the first I know of to include a quadratic trend in the DF-GLS test.
14. The KSS test first detrends the data with OLS, then applies a Dickey-Fuller-like test with nonlinear adjustment. One can apply this procedure to any form of trend.
15. The difference in lag order and unit root test result from the *maxlag* choice is the same whether I use my data period or his. I am able to exactly replicate his results with his data period and vintage of real GDP data from the February 2005 BEA release, which is consistent with when he likely worked on the paper.

(see Kim and Nelson 1999; Stock and Watson 2002). Plausible DGP equations must first be determined. Under the null they are first-difference models with a constant. In the unit root tests, the AIC, MAIC, and BIC almost always choose one or two lags, so AR(1) and AR(2) first-difference models are my first two DGPs.

Next, I apply "sequential elimination of regressors" (Brüggemann and Lütkepohl 2001) to the autoregressive 14-lag first-difference model (AR(14)). The procedure is similar to general-to-specific modeling in David Hendry and Hans-Martin Krolzig (2001). The full model is estimated, and then the lag with the *t*-ratio closest to zero is dropped. The equation is re-estimated and once again the lag with the *t*-ratio closest to zero is dropped. This continues until all remaining lags have a *t*-ratio that in absolute value exceeds a given value. The result (using either 1.96 or 1.65) has nonzero coefficients at lags 1, 5, and 12. This gives my third DGP, designated as AR(12-3).[16]

A different general-to-specific approach is Ng and Perron's (1995) test-down approach. I apply it to the DF-GLS test with a linear and then a quadratic trend, starting with the AR(14) specification. For both trends, the choice is an AR(12). Thus, an AR(12) first-difference model becomes my next DGP. It is also one of the two DGPs used by Murray and Nelson (2000). Their other was the AR(1).

John Campbell and Greg Mankiw (1987) extensively used first-difference autoregressive, moving-average models (ARMAs) to analyze unit roots in real GDP. More recently, James Morley, Nelson, and Zivot (2003) proposed an ARMA(2,2) for the first differences of U.S. real GDP. The ARMA(2,2) is thus my final DGP.[17]

The wild bootstrap is from Sílvia Gonçalves and Lutz Kilian (2004) and Herman Bierens (2011). Random errors are applied to the estimated DGP equation to recursively build data sets. The error for each time period is a random normal error with a mean of zero and standard deviation equal to the absolute value of the model's error for the given time period.[18] The actual values of real GDP are used to initialize each simulated data set, and 5,000 sets are created for each DGP. The unit root tests are applied to the same sequence of simulated data sets so that the

16. Brüggemann and Lütkepohl (2001) also propose the alternative that in each round one removes the lag that gives the greatest improvement in an information criterion like the AIC or BIC until no more improvement is possible. In the present case, the AIC leads to the same model already determined by the *t*-statistic approach. The BIC additionally eliminates the fifth lag.

17. The exact maximum likelihood option in TSP 5.1 is used to estimate all ARMAs in this paper.

18. There is no evidence for non-normality once a simple model of heteroskedasticity is allowed for. Divide the residuals from each DGP's estimated regression into two parts defined by one of Stock and Watson's (2002) estimated variance break points of 1983:2. The Jarque-Bera normality test gives a result nowhere near rejection for either time period for any of the five DGPs. Note that the wild bootstrap does not impose a specific breakpoint, consistent with uncertainty over the exact date and allowing that other patterns of heteroskedasticity are present.

joint *p*-value computations can account for the simulated correlations among the test statistics. The two-break tests (LP-CC, ZA-LP, and their KSS versions) take such a long time to compute that I use only the first 2,500 data sets for them. Each test chooses its lag order in every replication, just as with the actual data. This gives Papell (1997) and Murray and Nelson's (2000) "exact *p*-values."

Results for individual tests

The total number of tests is 42. Table 2 presents the results for 21 tests. The 21 tests omitted from the table generally consist of the same tests but with the lag order determined by the alternative information criterion. The exception is the PP-$Z\alpha$ test as the alternative to the PP-Zt test. The alternative lag-choice criterion (usually the BIC) almost always gives *p*-values less significant (thus, the omitted tests are not more favorable towards rejecting the unit root). Table 2 also gives *p*-values based on published critical values that are either asymptotic or adjusted only for sample size. The comparison of exact p-values with those from published critical values shows the size distortion from using such critical values.

Where one of my tests repeats one in Table 1, my test statistic is similar in value to that in the cited paper. Therefore, differences in my conclusions probably don't reflect the difference in sample periods. In Table 2, the bootstrapped *p*-values usually show lower statistical significance than do the *p*-values from published critical values. For example, the number of 0.05 rejections from the ten Vougas (2007) tests falls from six to one, two, or three, depending on the DGP.

Overall, my size-adjusted results show a rejection rate far less than the 36 percent rate in Table 1, at most 5 of 21 tests (24 percent). The importance of accounting for size distortion in individual tests is thus confirmed. Nevertheless, rejections do remain after bootstrapping. Are the remaining rejections sufficiently strong or numerous for an overall unit root rejection? Joint *p*-values provide an answer.

TABLE 2. Individual tests: statistics, and tabular and bootstrapped *p*-values

							Data generating process				
Test	Criterion	Trend	Group	Test statistic	Lag order	Non-boot *p*-value	AR(1)	AR(2)	AR(12-3)	AR(12)	ARMA
ADF	AIC	Linear	A	−2.79	1	0.20	0.34	0.33	0.43	0.33	0.37
DF-GLS	MAIC	Linear	A	−2.08	1	> 0.10	0.24	0.23	0.34	0.31	0.23
SP	AIC	Linear	A	−3.14	2	***0.040***	0.055	0.054	0.14	0.092	0.10
MPT	MAIC	Linear	A	10.16	1	> 0.10	0.24	0.23	0.31	0.29	0.19
DF-GLS	MAIC	Quadratic	A	−3.72	1	***0.029***	***0.017***	***0.013***	0.072	***0.040***	***0.031***
SP	AIC	Quadratic	A	−4.12	2	***< 0.010***	***0.024***	***0.023***	0.089	***0.050***	0.072
ZA-C	BIC	1 slope & level change	A	−5.68	2	***< 0.010***	0.064	0.065	0.24	0.37	0.44
LNV-B	AIC	1 smooth level change	A	−5.27	2	***< 0.010***	***0.037***	***0.034***	0.11	0.078	0.14
LNV-C	AIC	1 smooth slope & level change	A	−5.38	2	***0.014***	0.067	0.060	0.15	0.11	0.19
LP-CC	BIC	2 slope & level changes	A	−6.20	2	> 0.100	0.39	0.37	0.70	0.82	0.88
KSS	AIC	Linear	B	−4.11	2	***< 0.010***	0.057	0.057	0.068	0.054	0.12
PP-Z*t*	n.a.	Linear	C	−2.14	14	0.52	0.57	0.56	0.80	0.79	0.70
ZA-B	AIC	1 slope change	C	−4.72	2	***0.022***	0.10	0.11	0.20	0.15	0.21
ZA-LP	AIC	2 slope changes	D	−5.66	2		0.24	0.24	0.37	0.28	0.43
KSS	AIC	Quadratic	D	−4.91	2		***0.033***	***0.032***	***0.037***	***0.028***	0.071
KSS-ZA-B	AIC	1 slope change	D	−5.41	2		0.063	0.067	0.056	***0.044***	0.13
KSS-ZA-C	AIC	1 slope & level change	D	−7.14	2		0.48	0.51	0.37	0.38	0.49
KSS-LNV-B	AIC	1 smooth level change	D	−4.56	2		0.15	0.14	0.13	0.099	0.2
KSS-LNV-C	AIC	1 smooth slope & level change	D	−5.13	2		0.12	0.12	0.084	0.062	0.15
KSS-ZA-LP	AIC	2 slope changes	D	−6.52	2		0.39	0.40	0.26	0.26	0.46
KSS-LP-CC	AIC	2 slope & level changes	D	−9.07	3		0.72	0.76	0.56	0.59	0.70

Notes: The MPT test was computed using Gauss code from Serena Ng. *P*-values of 0.100 or less are given to 3 decimal places, and values of 0.050 or less are in boldface italics. The PP-Z*t* lag order is the order in the nonparametric correction. Non-bootstrapped p-value sources: ADF, PP-Z*t*: MacKinnon (1994); DF-GLS (linear), MPT: Elliott et al. (1996, Table I); DF-GLS (quad): Harvey et al. (2011, n.3); KSS: Kapetanios et al. (2003, Table 1); LNV: Leybourne et al. (1998, Table 1); LP-CC: Lumsdaine and Papell (1997, Table 3); SP: Schmidt and Phillips (1992, Tables 1A and 1B); ZA-B, ZA-C: Zivot and Andrews (1992, Tables 3A and 4A).

Joint *p*-values

I first compute joint *p*-values for the complete set of 42 tests. Is this an excessive number? If we knew some of the tests had little power, then it would be good to omit them. Failure to reject when using a joint *p*-value would not be very persuasive if the joint *p*-value was partly computed from useless tests. One might also argue for the omission of tests whose results are very highly correlated with other tests. Such tests would add little new evidence. To avoid cherry picking, the decision to omit a test would have to be done based on information exogenous to the data and before running the test and seeing the result. In any event, all the test types, trends, adjustments, and lag selection techniques in my 42 tests have been judged potentially useful and powerful by various authors in the past. Also, my joint *p*-value procedures take account of the correlations among the tests. I see no sound basis to omit any of the tests.

Nevertheless, as a sensitivity check I do look at the following subsets of tests composed of the groups identified in Table 1: A, A+B, A+B+C, and A+B+C+D.[19] Subset A contains the tests (with my modifications) that Vougas (2007) used. Subset A+B adds the KSS nonlinear adjustment test with linear trend and thus contains all tests applied to the real GDP-unit root question in the last 10 years (with my modifications). Subset A+B+C contains all the trends, adjustment specifications, and test types ever appearing in the real GDP-unit root literature (with my modifications), with one lag choice method for each. Finally, subset A+B+C+D adds the two-slope-change test and the nonlinear adjustment, KSS-type tests for all the trends beyond the simple linear one, with one lag choice method for each.

The literature contains several approaches to joint *p*-value computation, but it does not tell us which is best. To see whether my results are robust to different methods, I try four. Each involves a comparison of how unusual the actual bootstrapped *p*-values for the tests of interest are compared with the joint distribution of their simulated *p*-values for a given DGP. For a given set of tests, the joint distribution of *p*-values reflects the correlations among the test statistics. I briefly describe the four joint methods here. Details are given in Appendix 2.

The first method is a bootstrapped version of the "improved Bonferroni procedure" of R. J. Simes (1986), which gives a joint *p*-value based on the actual

19. Of course, even my list of 42 tests omits many possible tests. There are other lag specification methods and an endless array of possible trend specifications, and no doubt additional unit root test approaches yet to be invented. I am implicitly assuming that as these possibilities increase in number, they become increasingly redundant and not remarkably more powerful.

p-values adjusted according to their number and rank in terms of size. The next two methods compute how unusual, relative to the bootstrapped distribution, are the actual number of individual rejections, i.e., the number of small-enough *p*-values (Cushman 2008). The two methods are differentiated by their use of the 0.05 or 0.10 level to determine rejection. The fourth method determines how unusual relative to the bootstrapped distribution are the specific *p*-values themselves (Cushman and Michael 2011). There are several other techniques that I do not use (Ayat and Burridge 2000; Harvey et al. 2009; Hanck 2012) because they only apply to two tests.

I should provide evidence that the joint procedures I employ do work, in the sense of having good size and power properties. Existing evidence on the joint procedure properties is meager. Simes (1986) shows that his improved Bonferroni procedure has good size properties and better power than the standard Bonferroni method in the case of 10 tests, particularly when the test statistics are correlated. There is not any evidence concerning the counting methods I used in Cushman (2008). Nils Michael and I (2011) found good size properties for our approach as applied to six tests, but we did not examine power.

Given this shortage of evidence and the centrality of the joint approaches to my critique, I conducted a Monte Carlo examination of size and power of the unbootstrapped and bootstrapped Simes procedure, the 0.10 and 0.05 counting methods, and the Cushman-Michael procedure. The details are in Appendix 3. All the approaches have good size and power. The unbootstrapped and bootstrapped Simes *p*-values are very highly correlated ($r \geq 0.990$) for every data-generating process, and so including just one is reasonable.

Table 3 gives the results for the four joint procedures for the various subsets of tests using the various DGPs. If one includes all 42 tests and uses the 0.05 level for rejection, the unit root is not rejected for any combination of DGP and joint *p*-value approach. But what if the significance level is relaxed to 0.10 or certain smaller test groups are considered more appropriate? Then unit root rejection remains possible. It depends on how we resolve the varying conclusions from the joint approaches, and on which DGPs are more plausible.

Let us first consider the issue of varying joint *p*-value conclusions. We are back to a multiple test problem. Given my conclusion that the four joint approaches are all reasonable, then a joint approach can be applied to the joint tests. I therefore apply the unbootstrapped Simes procedure to the joint tests.[20] The resulting *p*-values are labeled "Simes Joint" in Table 3. Only one is significant at the

20. I am not applying any of the bootstrapped joint approaches to the present problem because of the immense amount of time it would require.

0.05 level, for the subset A+B using the AR(2) DGP. However, several more are significant at the 0.10 level, for subsets A and A+B using the AR(1) or AR(2) DGP.

TABLE 3. Joint *p*-values

		Data generating process					
Subset	Method	AR(1)	AR(2)	AR(12-3)	ARMA	AR(12)	Mean
A (10 tests)	Simes	0.088	0.091	0.23	0.24	0.17	0.16
	Count 0.05	0.060	0.057	1.00	0.26	0.13	0.13
	Count 0.10	***0.028***	***0.026***	0.26	0.25	0.084	0.13
	CM	0.051	***0.033***	0.19	0.17	0.12	0.11
	Simes Joint	0.080	0.066	0.35	0.26	0.17	0.19
A+B (11 tests)	Simes	0.084	0.086	0.21	0.23	0.16	0.15
	Count 0.05	0.067	0.065	1.00	0.29	0.14	0.14
	Count 0.10	***0.021***	***0.018***	0.16	0.28	0.055	0.11
	CM	***0.041***	***0.023***	0.17	0.14	0.10	0.094
	Simes Joint	0.082	***0.046***	0.28	0.29	0.16	0.17
A+B+C (13 tests)	Simes	0.098	0.099	0.25	0.25	0.18	0.18
	Count 0.05	0.079	0.083	1.00	0.33	0.16	0.17
	Count 0.10	***0.028***	***0.028***	0.19	0.31	0.076	0.13
	CM	0.051	***0.034***	0.22	0.15	0.15	0.12
	Simes Joint	0.098	0.068	0.33	0.35	0.19	0.21
A+B+C+D (21 tests)	Simes	0.12	0.12	0.23	0.28	0.16	0.18
	Count 0.05	0.088	0.092	0.42	0.43	0.092	0.23
	Count 0.10	***0.031***	***0.028***	0.11	0.32	***0.027***	0.10
	CM	0.067	0.054	0.12	0.15	0.068	0.092
	Simes Joint	0.12	0.11	0.24	0.43	0.11	0.20
All (42 tests)	Simes	0.13	0.14	0.33	0.49	0.27	0.27
	Count 0.05	0.095	0.10	0.51	0.53	0.21	0.29
	Count 0.10	0.050	0.039	0.24	0.50	0.16	0.20
	CM	0.095	0.082	0.20	0.31	0.26	0.19
	Simes Joint	0.13	0.13	0.44	0.53	0.27	0.30

Notes: A *p*-value of 1.00 means there were no rejections at the 0.05 level for the Count 0.05 test. In such cases I omit the test in computing the average *p*-value. CM is the method of Cushman and Michael (2011). *P*-value significance is highlighted as in Table 2.

This brings us to the question of resolving the different results from different DGPs. We could assume that each is equally plausible. If so, multiplying each joint *p*-value by 1/5 and summing (i.e., computing the mean) generates an overall *p*-value for a given test group.[21] The final column of Table 3 gives the results. There are no rejections at the 0.05 level, but there are two marginal rejections at the 0.10 level.

21. This reflects the Law of Total Probability. Each individual *p*-value is the probability of a test statistic at least as extreme as the one observed given the DGP. Thus, the overall probability of a test statistic at least

But are the DGPs equally plausible? The data can be used to generate AIC and BIC weights using methods from Bruce Hansen (2007). Using AIC weights overwhelmingly favors the AR(12-3) model with a weight of 0.997. Using BIC weights almost equally favors the AR(12-3) model, with a weight of 0.932; the AR(1) is second at 0.067. With either the AIC or BIC, the weighted-mean *p*-values are essentially the *p*-values for the AR(12-3), and there are no rejections at even the 0.10 level.

The economic importance of the permanent shocks

Given the failure to reject the unit root, from here on I assume real GDP has a unit root. I now present several estimates of the unit root's economic importance—addressing Ziliak and McCloskey's (2008) desire for "oomph" analysis.

An unobserved components model

The first oomph estimate is from an unobserved components model. Krane (2011) also used such a model. The model gives an estimate of the relative magnitude of permanent shock variance to transitory shock variance. I use the local level model with constant drift in Koopman et al. (2000) and found in their STAMP software (version 6.21), which I use for the estimations. Let y be real GDP. The model is (with variables in logs):

$$y_t = \mu_t + \varepsilon_t \ , \ \varepsilon \sim NID\left(0, \ \sigma_\varepsilon^2\right) \tag{1a}$$

$$\mu_t = d + \mu_{t-1} + \eta_t, \ \ \eta \sim NID\left(0, \ \sigma_\eta^2\right) \tag{1b}$$

where ε is the transitory shock and η the permanent shock. The current, long-run level of y is determined by μ_t, and the long-run rate of growth is d. The key parameters are the transitory variance σ_ε^2 and permanent variance σ_η^2. To identify the two variances, Koopman et al. (2000) assume zero contemporaneous correlation between the errors. Serial correlation can be specified with the substitution of $y_t - \sum_{j=1}^{k} \varrho_j y_{t-j}$ for y_t in equations (1a) and (1b). I use lag order $k = 3$, consistent with

as extreme as the one observed is the sum of the probabilities conditional on the DGP where each is first multiplied by its probability.

two lags in first differences as is sometimes suggested by the information criteria applied to the unit root tests. The substitution is equivalent to assuming that e and η are autoregressively correlated with identical lag coefficients.

An estimate of the relative importance of the permanent shocks is s_η/s_e, permanent to transitory shock standard deviation. The value for the 1947–2007 data is 1.17. This magnitude is, however, quite misleading because of significant small-sample bias. The estimation procedure has a hard time distinguishing partial persistence (ϱ_j) from permanent persistence (η), and it tends to interpret partial persistence as permanent. To compensate, I have computed bias-adjusted estimates. Details are given in Appendix 4. The mean-bias-adjusted ratio is 0.42. However, the estimate is subject to very large uncertainty (see Appendix 4).

Evidence from a vector error correction model

Cochrane (1994) proposed a two-variable vector error correction model (VECM) to identify the permanent and transitory components of real GDP. Nondurable goods consumption plus services consumption gives households' permanent consumption. According to the permanent income hypothesis, permanent consumption is a random walk and is a constant fraction of permanent income on average. Thus, income (real GDP) and permanent consumption are cointegrated. Changes in permanent consumption reflect consumers' views of permanent shocks to GDP. Deviations of GDP from the cointegration vector reflect transitory shocks to GDP. With two first-difference lags as employed by Cochrane (1994), the VECM model is:

$$\Delta c_t = a_c + e_c\left(c_{t-1} - y_{t-1}\right) + \sum_{i=1}^{2}\left(b_{c,1,i}\,\Delta c_{t-i} + b_{c,2,i}\,\Delta c_{t-i}\right) + \upsilon_{c,t} \tag{2a}$$

$$\Delta y_t = a_y + e_y\left(c_{t-1} - y_{t-1}\right) + \sum_{i=1}^{2}\left(b_{y,1,i}\,\Delta c_{t-i} + b_{y,2,i}\,\Delta c_{t-i}\right) + \upsilon_{y,t} \tag{2b}$$

where c is permanent consumption and y is real GDP. With permanent consumption a random walk, $e_c = 0$. The constant ratio of permanent consumption to income gives the unitary coefficient on y_{t-1} in the cointegration vector.[22] Actually, consumption in equation (2a) is not a pure random walk because of the presence of the first-difference lag terms. Nevertheless, the $e_c = 0$ constraint ensures that the permanent shocks can be solely associated with consumption.[23]

22. The value of the ratio is contained in the constant of equation (2b), which also includes the long-run growth rate.

23. The constraint is accepted at the 0.05 level when the VECM is estimated over the 1947–2007 period and for three subperiods defined by consumption break dates of 1969:4 and 1991:4 found by Stock and Watson (2002).

I use the Cochrane VECM model to estimate the relative importance of permanent shocks. But what about possible structural change? Perron (1989) and Perron and Tatsuma Wada (2009) argued that a permanent change in real GDP's growth rate occurred in 1973. Stock and Watson (2002) found a break in services consumption in 1969 and in nondurables consumption in 1991. There is also the Great Moderation (Kim and Nelson 1999). A VECM that is estimated for the entire data period but ignores such changes might give misleading results. Unfortunately, the dates are not definite and standard VAR/VECM estimation procedures do not allow for variance breaks. Therefore, I employ an alternative method, which is to estimate rolling VECMs. This approach gets estimates that gradually adapt to changing conditions. The rolling estimation periods are 15 years long.[24]

I use impulse response (IR) analysis to measure the importance of the unit root. Suppose some unexpected event, a shock, occurs. An impulse response is the deviation of a variable's subsequent path from the path it would have followed if no shock had occurred. I wish to compare the effects of permanent shocks with those of transitory shocks. To do so unambiguously, we need a structural model with orthogonal permanent and transitory shocks. But the VECM is a reduced form model and does not reveal the orthogonal shocks. (The structural vector autoregression model can have unlagged terms for c and y on the right-hand sides of the equations.) To identify the structural model and shocks, I use the Choleski recursive approach. I assume that shocks to y contemporaneously affect only y, not c, while shocks to c contemporaneously affect both c and y. This follows from the permanent income model's implication that consumption alone contains the single unit root process. If we were to reverse the Choleski ordering, orthogonal income shocks would have permanent effects, contradicting the model. Under the Choleski ordering I use, permanent shocks in the structural model are the v_c terms

24. The first rolling period starts in 1947:4 (allowing three periods for the lags) and ends in 1962:3, the second starts in 1948:1 and ends in 1962:4, and so on, until the last period starts in 1992:4 and ends in 2007:3. As in Cochrane (1994), permanent consumption is measured by the sum of real personal consumption expenditures on nondurable goods and on services (seasonally adjusted). My data is from the BEA, vintage March 27, 2014. Nondurable goods consumption is series DNDGRA3Q086SBEA, an index number series with base year 2009 that starts in 1947. I convert it to 2009 dollars using series PCNDGC96, which is a real dollar series starting in 1999 and perfectly correlated with its index number version over the 1999–2013 period). Services is series PCESVC96, an index number series also starting in 1947 with base year 2009 and converted to 2009 dollars using real dollar series DSERRA3Q086SBEA that also begins in 1999 and is perfectly correlated with its index number version. The series are converted to logs. Real GDP is series GDPC1 of the same vintage. The work in the present section was done more recently than the unit root test work, and I also use the newer vintage because it gives a real consumption series with no linking required to account for base year changes, unlike the data available only a few years earlier.

in the VECM consumption equation (2a).[25] The impulse responses are estimated with JMulti 4.24.

In the IR analysis, the magnitudes of the initiating shocks are the estimated standard deviations of the structural shocks over the estimation period (they are thus of 'typical' size). With η the permanent shock and e the transitory shock in the structural model, the initiating shocks for the impulse responses are therefore s_η and s_e. Across all 181 rolling models, s_η has a mean of 0.0041 and s_e has a mean of 0.0067. There is noticeable variation in both, illustrated below. The relative importance of the permanent shock, s_η/s_e, has a mean across the 181 rolling periods of 0.64 (much more than the estimate of 0.42 from the unobserved components model applied to the full data period).[26]

Important though the structural shock magnitudes are, IR analysis shows that their relative magnitudes are not very good indicators of subsequent real GDP responses. In what follows, the impact quarter is 0, and the forecast horizons are from 1 to 20 quarters ahead. To indicate relative importance, I report the sizes of the real GDP responses relative to the size of the initial transitory shock. In Figures 1a–1f, solid red solid lines depict permanent shocks or their GDP responses, and blue dashed lines depict transitory shocks or their GDP responses. The horizontal lines give the means across the 181 rolling periods. The dates on the horizontal axes give the starting period of the rolling VECM.

Figure 1a gives the actual shocks. A decline in shock magnitude in the Great Moderation is easily seen. Figures 1b–1f give GDP responses. At horizon 0 (the impact period), the response to the permanent shock averages half the value of the transitory shock. At horizon 1, the response to the smaller permanent shock has already caught up to that from the transitory shock. At horizon 4, the response to the permanent shock is on average 1.7 times the response to the transitory shock and 2.25 times the size of its initiating permanent shock. And at horizon 8, the permanent shock response is over 3 times the response to the transitory shock, and still 2 times the size of its initiating shock. During the Great Moderation, the responses at horizons 4 and 8 to transitory shocks have become relatively more important, but still not equal to those from permanent shocks. Note also that, at longer horizons, the growing relative importance of the permanent shock is increasingly a consequence of the dying away of the effects of the transitory shock.

25. If we don't impose the permanent income model, then we can relax its error correction zero restriction on consumption adjustment and try both Choleski variable orderings. As a check, I compared the results of these alternative specifications with those of my main model using the full 1947:1 to 2007:3 period. The only significant difference is that, with income first in the ordering, it takes two years instead of less than one for consumption shocks to dominate income shocks in their effects on real GDP. The responses to income shocks continue to largely die away, just taking a bit longer.

26. Note that $0.64 \neq 0.0041/0.0067$: the mean of the ratios is not equal to the ratio of the means.

By horizon 20, the responses are close to their long-run values, with the permanent shock response 1.5 times its initiating shock in all but the early rolling VECMs, and the transitory shock response nearly back to zero. From these results, it would be hard to argue that permanent shocks are unimportant.[27]

Figure 1a. Initial shocks (horizon 0)

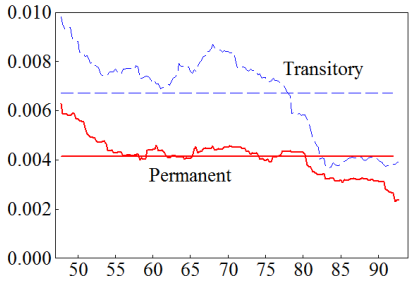


Figure 1b. Impulse responses at horizon 0

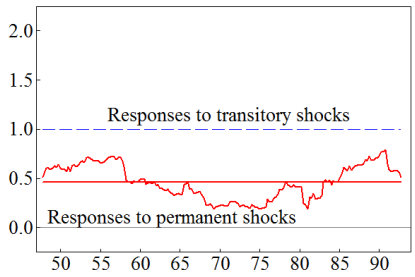


Figure 1c. Impulse responses at horizon 1

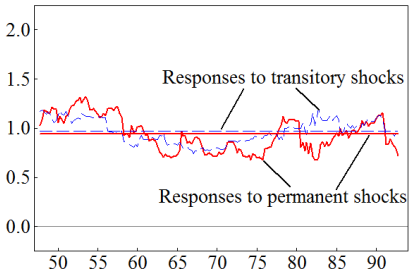


Figure 1d. Impulse responses at horizon 4

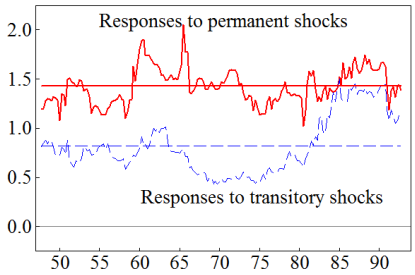


Figure 1e. Impulse responses at horizon 8

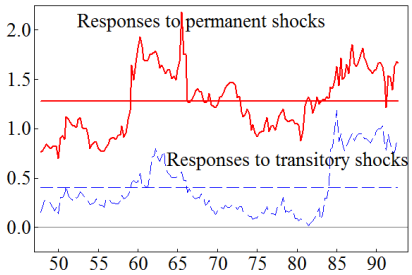
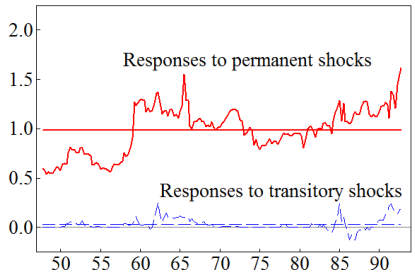


Figure 1f. Impulse responses at horizon 20



27. To focus on oomph and to minimize graph complexity, I do not include confidence limits in the impulse response figures. Here is a summary of the results from 2,000 replications of an Efron bootstrap for each of the 181 rolling VECMs. Except with some rolling estimation periods for impact period 0, the lower 95 percent limit for GDP response to the permanent shock is always well above zero. The lower 95 percent limit for GDP's response to the transitory shock is always above zero until horizon 4 for 33 rolling periods, for all 181 rolling periods at horizon 10 and longer. Thus, the responses to permanent shocks are always statistically significant (with a few exceptions in the impact period, and the responses to transitory shocks are statistically significant for 3 quarters after which point they are often not, consistent with the responses nearing their long-run equilibrium value of zero.

These results are similar to those of Cochrane (1994). For instance, the magnitudes of the absolute and relative sizes of shocks and their responses at horizons 0, 4, and 8 are almost the same. Differences likely reflect that Cochrane (1994) did not impose $e_c = 0$ when computing impulse responses, and that he did not compute rolling VECMs. Because he did not impose $e_c = 0$, a small part of Cochrane's GDP shock is permanent and never dies away. The rolling approach then reveals that transitory shocks become relatively more persistent during the Great Moderation (GDP's relative responses to transitory shocks become larger at horizons 4 and 8).

Forecasting contests

The measures of unit root oomph presented so far are computed from within-sample data. Let's turn to out-of-sample forecasts. I compare the forecasts of 6 models. The first two are the just-analyzed bivariate VECM and a univariate ARMA(2,2) in first differences. These specify both permanent and temporary shocks.[28] They are therefore unit root models. The second two models are a univariate AR(2) in first differences and a bivariate VAR(2) in first differences using real GDP and permanent consumption. These two models (called AR-dif and VAR-dif below) specify all shocks as having permanent effects, although they do allow transitory dynamics. These, too, are unit root models. The final two models are a univariate trend stationary AR(3) model in levels and a bivariate trend stationary VAR(3) model in levels (with real GDP and permanent consumption). In these two models (called AR-tr and VAR-tr below), all shocks are transitory.

The first forecasting experiment applies the rolling regression approach to the six forecasting models to make forecasts at horizons of $h = 1$ to 20. The number of h-period-ahead forecasts for each model and horizon is $181 - h$. My measure of forecast accuracy is the square root of the mean squared forecast error, i.e., the standard deviation of the forecast errors where the mean in the standard deviation formula is set to zero.[29] Results for a few key horizons are in Figures 2a and 2b. Each cluster of bars shows the forecast error standard deviations for the six models for one forecast horizon. The vertical axis measures the values of the forecast standard deviations. Figure 2a is computed using all forecasts through 2007:3. Fig-

28. I have already discussed the way the VECM specifies permanent and transitory shocks. Tsay (2005) shows how various ARMA models are reduced forms of unobserved component models with permanent and transitory shocks.

29. Mean square forecast errors, not standard deviations as I am using, are often employed to reflect the assumption that large errors are disproportionately more costly than small errors. And they assume the disproportionalities are specifically quadratic. The reader can make adjustments if desired. I employ standard deviations so that the differences among them correspond to easy-to-interpret growth rates, not squared growth rates.

ure 2b shows forecasts from 2007:4 through 2013:4 (for the first time in the paper). The longer-horizon forecasts are dominated by the inability to forecast the Great Recession.

In Figure 2a, the unit root models almost always outperform the linear trend stationary models (in 27 of 32 paired comparisons—four unit root models with two trend stationary models for four horizons). The VECM model always is best. But the margins of victory for the unit root models are not usually large in relative terms. The unit root models win less often in Figure 2b (23 of 32 comparisons), but the VECM model is still best in every case.[30]

Figure 2a. Forecast error standard deviations: earliest through 2007:3

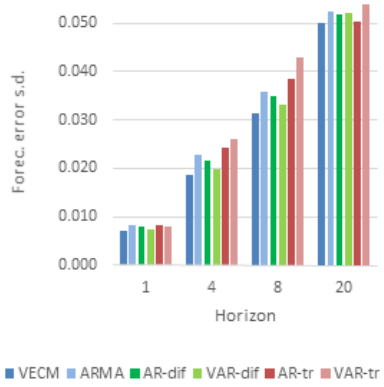
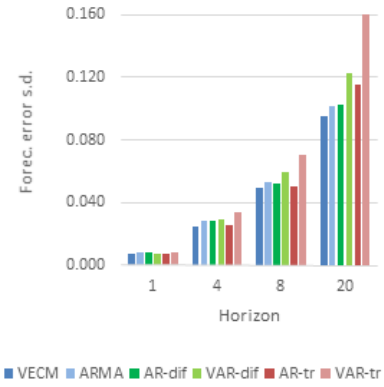


Figure 2b. Forecast error standard deviations: 2007:4–2013:4



Tables 4a and 4b give the underlying numerical values for Figures 2a and 2b, with the addition of summary measures in the form of the mean forecast standard errors for horizons 1–8 and 9–20. The previous conclusions are supported, with the additional revelation that, for the means, every unit root model beats the trend stationary models for forecasts through 2007:3.

Interest in unit root models and forecasting has often centered on whether or not recoveries from recessions would be robust, involving rebounds. Thus, for the second forecasting contest I use actual postwar recessions. Suppose, in what turns out to be the final quarter of the seven postwar recessions with sufficient data preceding them, a hypothetical econometrician uses the above six models to forecast the recovery. The models are estimated using the 60 quarters of data ending in the last quarter prior to the peak that defines the beginning of recession

30. If the VECM models are estimated without the error correction constraint, the results are almost identical. In addition, the reader may wonder about the statistical significance of the results in Figures 2a–2b. At this point, my main aim is oomph, but see Appendix 5 for some simulation results that address statistical significance.

according to the BEA. This is to follow a number of economists who, as previously mentioned, implicitly or explicitly drew trend lines based on pre-recession data.[31]

TABLE 4a. Forecast standard errors, from first rolling model through 2007:3

	Unit root models				Trend stationary models	
Horizon	VECM	ARMA	AR-dif	VAR-dif	AR-tr	VAR-tr
1	0.0072	0.0084	0.0080	0.0075	0.0083	0.0081
4	0.019	0.023	0.022	0.020	0.024	0.026
8	0.031	0.036	0.035	0.033	0.039	0.043
20	0.0501	0.052	0.052	0.052	0.0503	0.054
1–8	0.020	0.023	0.023	0.021	0.025	0.027
9–20	0.043	0.0466	0.0457	0.045	0.0468	0.052

TABLE 4b. Forecast standard errors, 2007:3–2013:4

	Unit root models				Trend stationary models	
Horizon	VECM	ARMA	AR-dif	VAR-dif	AR-tr	VAR-tr
1	0.0070	0.0078	0.0078	0.0075	0.0075	0.0080
4	0.025	0.029	0.029	0.029	0.026	0.034
8	0.049	0.053	0.052	0.060	0.051	0.070
20	0.095	0.101	0.103	0.123	0.115	0.160
1–8	0.028	0.031	0.031	0.033	0.029	0.038
9–20	0.119	0.128	0.126	0.162	0.134	0.202

Note: I sometimes increase decimal accuracy beyond three places to allow distinction of slight differences in magnitude, but this is not to imply that such differences are necessarily of statistical or economic significance.

The results are presented graphically in Figures 3a–3g. The figures include 20 forecast periods after the end of the recession and six quarters prior. In two cases, the 1969:4–1970:4 and 1980:1–1980:3 recessions, the next recession starts before 20 quarters have elapsed. In these cases, vertical lines indicate the middle of the peak quarter that defines the beginning of the next recession. All values are detrended with a simple OLS linear trend from the same 60-quarter forecast period used for the forecasting models. The horizontal zero line corresponds to this trend.

31. My results will not be precisely what the hypothetical econometrician would have obtained in real time because I am using 2014 vintage data, not the data available at the time of the forecasts. In addition, several of the modeling techniques had not been invented at the time of the earlier recessions. Finally, since real-time econometricians would not know the ending period of the recession until sometime later, my approach gives the trend stationary models an advantage in that, if there is going to be a return to a former trend line, it should be most accurately measured starting at the end of the recession.

Fig. 3a. After the 1969:4–1970:4 recession

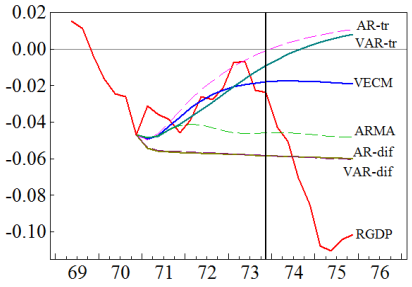


Fig. 3b. After the 1973:4–1975:1 recession

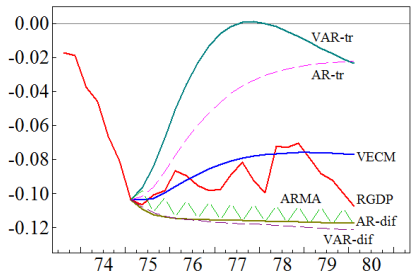


Fig. 3c. After the 1980:1–1980:3 recession

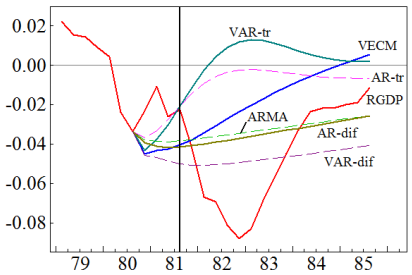


Fig. 3d. After the 1981:3–1982:4 recession

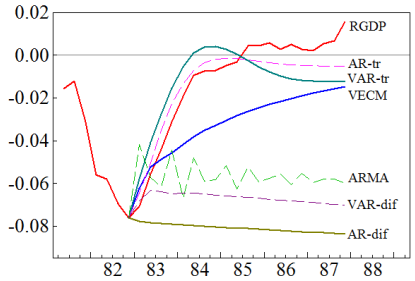


Fig. 3e. After the 1990:3–1991:1 recession

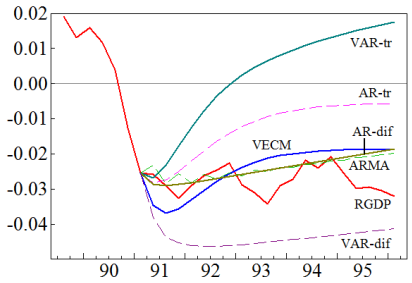


Fig. 3f. After the 2001:1–2001:4 recession

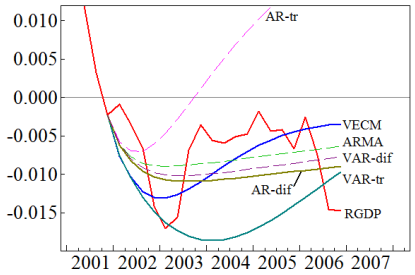
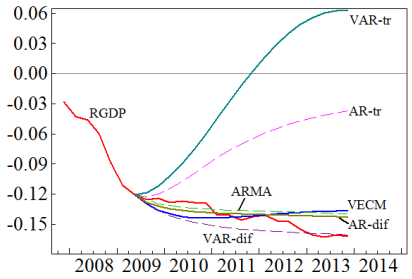


Fig. 3g. After the 2007:4–2009:2 recession



After recessions 1 (1969:4–1970:4), 3 (1980:1–1980:3), and 4 (1981:3–1982:4), real GDP more or less returns to the simple prior trend given by the zero line. However, in contrast to some beliefs previously noted, real GDP *often* has no tendency to return to a prior trend; see recessions 2 (1973:4–1975:1), 5 (1990:3–1991:1), and 7 (2007:4–2009:2). And after recession 6 (2001:1–2001:4), GDP's path follows an ambiguous path. It thus appears that the relative importance of the permanent and transitory shocks associated with recessions has varied.

Which forecasting models manage these diverse situations best? Consider the two linear trend stationary models (AR-tr and VAR-tr) versus the two first difference models (AR-dif and VAR-dif). The linear trend stationary models do better when GDP heads back to the prior trend. This is not surprising as these models always assume reversion to trend. The first difference models do better when there is no significant return to the former trend, again unsurprising as these models assume shocks are permanent. The first difference models somewhat outperform the trend stationary models in the ambiguous case of recession 6. But what about the ARMA and VECM models? The graphs suggest that the ARMA model is a sort of compromise, consistent with its forecasts estimating any recent shock to be a typical blend of permanence and transience. The VECM provides a better compromise, doing reasonably well after every recession, which can be attributed to its explicit, and apparently often successful, identification of recent permanent versus transitory shocks. Thus, unless you know what kind of shock caused the recession, the VECM seems to be the best choice.[32]

The impressions from looking at Figure 3 are sharpened by examining the forecast error standard deviations. I compute them for the six models and seven recessions using the same 20 post-recession quarters as in Figure 3, except for the two recessions where another recession intervenes in less than 20 quarters. In those cases, the final forecast error quarter is that of the peak defining the beginning of the next recession. The results are in Figure 4.[33] For example, as it appeared from the graphs, after recessions 2, 5, 6, and 7, the trend stationary models generally do much worse than the models specifying unit roots. The trend stationary models do well after recessions 1, 3, and 4. The VECM model, although frequently not the best, always does fairly well with no huge mistakes as sometimes occur with the other models.

32. Regarding the VECM's success in forecasting no recovery from the last recession, Cochrane (2015c) remarked: "It's interesting that consumers seem to have seen the doldrums coming that all the professional forecasters did not."

33. The forecast error bars from the linear trend stationary models for the final recession are so large that I have truncated the height of the graph to maintain reasonably visible differences among the other forecast error bars.

I obtain overall conclusions from the Figure 4 results by computing averages of each model's forecast error standard deviations across the recessions. They are in Table 5. The first row gives the arithmetic means of the seven forecast error standard deviations for each model. The second row gives the means with forecast error standard deviations weighted by the number of forecast periods used to compute them (fewer for the two recessions whose recoveries were cut short by another recession). The mean standard deviations for the AR-tr and VAR-tr models are strongly influenced by very large forecast error standard deviations for the final recession. To see what difference the last recession makes, the third and fourth rows omit the last recession.

Figure 4. Forecast error standard deviations after recessions

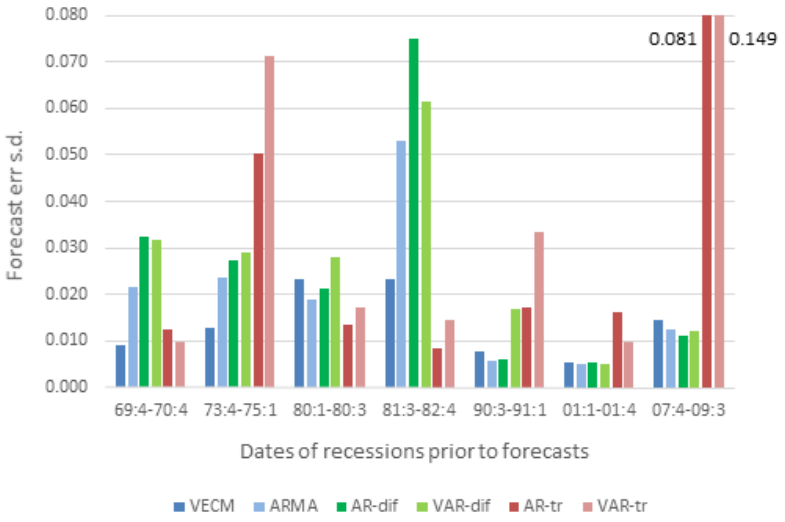


TABLE 5. Average forecast error standard deviations across recessions

	Unit root models				Trend stationary models	
Type of average	VECM	ARMA	AR-dif	VAR-dif	AR-tr	VAR-tr
Mean	0.014	0.020	0.025	0.026	0.028	0.044
Weighted mean	0.015	0.024	0.031	0.031	0.038	0.060
Mean without last recession	0.014	0.021	0.028	0.029	0.020	0.026
Weighted mean without last recession	0.016	0.027	0.036	0.036	0.027	0.036

The first two rows of Table 5 support the conclusion from Figures 3a–3g and Figure 4 that the unit root models allowing for both permanent and transitory shocks, VECM and ARMA, do best on average. In fact, the VECM does much better than any of the other models. After the VECM and ARMA, the unit root

models explicitly specifying only permanent shocks, AR-dif and VAR-dif, are next best. The linear trend models are worst. Accordingly, concluding that GDP is trend stationary, or that unit roots don't matter, is likely to lead to the worst forecasts, on average. Meanwhile, if the 2007:4–2009:2 recession is omitted, perhaps because one believes it was an extraordinary event, the tidy pattern just described is somewhat muted. In rows 3 and 4, the ARMA is now only the approximate equal of the AR-tr. The VECM model, however, continues to be substantially better than the other models.

The above results suggest that:

1. Models specifying only one kind of shock only do well when that kind of shock dominates. The modest advantage on average of the AR-dif and VAR-dif models over the AR-tr and VAR-tr models thus reflects a history where the causes of recessions have been dominated by permanent shocks a bit more often than by transitory shocks.
2. The ARMA and VECM outperform the other models because they allow both kinds of shocks. However, the multivariate approach of the VECM is significantly better than the ARMA approach because, unlike the univariate ARMA model, a multivariate approach has the potential to separately identify the two kinds of shocks.
3. The advantage of a multivariate model disappears entirely if it is misspecified as trend stationary. In the tables, the VAR-tr model is generally the worst of all.

Final remarks

I have first provided evidence that, despite a significant literature to the contrary, the null hypothesis of a unit root in U.S. real GDP in postwar real GDP cannot be rejected. The key innovation is to account for the multiple test problem. Nevertheless, failure to reject is not a strong result, and it may well be that lack of power still hobbles the unit root testing approach despite all the data now available. It is also surprising that a large literature testing the trend stationarity null did not arise, although those tests may also lack sufficient power. Regardless, the difficulty of discerning between nearly equivalent null and alternative hypotheses in the unit root context has long been understood (see discussion by Stock 1990).

The second component of my approach shows that, if a unit root in GDP is specified, then estimates of its importance using postwar data are of notable economic magnitude. Although the sizes of permanent shocks estimated by both

the unobserved components model and the VECM average roughly half those of the estimated transitory shocks, the estimated subsequent impact of the permanent shocks is much larger in impulse responses.

Equally if not more convincing, the best forecasting model by far, Cochrane's (1994) VECM, specifies a unit root. But that is not the sole reason for the VECM's success. It allows not only permanent but also transitory shocks, and its multivariate approach allows them to be separately identified. The first-difference ARMA also specifies a unit root and a transitory component, and it also tends to outperform the other models. But its forecasts are, overall, inferior to the VECM's. This reflects its inability to separately identify permanent and transitory shocks. Nevertheless, Campbell and Mankiw's (1987) early application of first-difference ARMAs for real GDP is important because of its recognition of the necessity of specifying a unit root (permanent shocks) *and* transitory shocks. The necessity is echoed in the results the present paper, with the addition of the importance of a multivariate approach (which goes back to Cochrane 1994) to specifically distinguish permanent from transitory shocks.

Appendix 1. Corrected conclusions for two tests in Vougas (2007)

For the PT test, Vougas (2007) reports a value of 8.166 and says it's a rejection, perhaps because it is greater than the 0.05 and 0.10 critical values of 5.66 and 6.86 in Elliott et al. (1996). But the test is *lower tailed*, and so 8.166 does *not* reject. For the LP-CC test, Vougas (2007) reports a test statistic of −6.250. The 0.05 and 0.10 critical values from Table 3 of Lumsdaine and Papell (1997) are −6.82 and −6.49. This test is also lower-tailed, so −6.250 does not reject.

Appendix 2. Methods to calculate joint *p*-values

All the joint *p*-value approaches I use involve comparing the actual bootstrapped *p*-values for the *n* tests in the subset of interest with the bootstrapped joint *p*-value distribution for the *n* tests. Therefore, the first step is to get the bootstrapped joint *p*-value distribution. This is formed from the first 2,500 replications per data-generating process (2,500, not 5,000, because each subset of tests involves at least one test with only the first 2,500 replications available). The 2,500 × *n* matrix of simulated test statistics is then converted into a corresponding matrix of *p*-values. First, I replace each simulated statistic with its rank (*r*) in its own

distribution. Small rank values correspond to small test statistic values (all tests are lower-tailed). Then, the *p*-value of a given simulated test statistic equals $r/2500 - 1/5000$. Because all tests are applied to the same sequence of data sets, the *p*-value matrix reflects the estimated correlations among the test statistics.

Let us begin with the improved Bonferroni method of Simes (1986). Simes considers the general case where the *p*-values correspond to different null hypotheses. My implementation is a special case where the nulls are all the same. Let P_j (where $j = 1, \ldots, n$) be the ordered *p*-values (bootstrapped in my application) from n tests applied to the actual data. Let a be the significance level. Simes then compares the *p*-values with an adjusted significance level; the jth null is rejected for any $P_j \leq ja/n$. However, I prefer to adjust the *p*-value rather than the significance level. Therefore, I compute adjusted *p*-values: $P_j^A = nP_j$. Since the nulls are all the same in my implementation, the only *p*-value that matters is the smallest one. Therefore my Simes *p*-value is the minimum of the P_j^A values. I refer to this as an "unbootstrapped" joint *p*-value, although it is derived from bootstrapped individual *p*-values, because it is not computed with respect to a bootstrapped distribution of Simes *p*-values. The unbootstrapped Simes *p*-value could be compared with the significance level, a, but I go one step further and get a bootstrapped value. To do so, I compute the Simes *p*-value for every row of the simulated *p*-value matrix, which gives a bootstrapped Simes *p*-value distribution. The bootstrapped Simes *p*-value is the fraction of the simulated Simes *p*-values less than or equal to the actual, unbootstrapped Simes *p*-value.[34] In Table 3, the "Simes" values are bootstrapped, while the "Joint Simes" values are not.

For the two counting methods (Cushman 2008), I first count the number of *p*-values less than 0.05 and less than 0.10 (i.e., I count the number of rejections for the two significance levels) using the actual data. For each significance level in turn, I then find the number of simulated *p*-values less than 0.05 and 0.10 in each row of the simulated *p*-value matrix. The bootstrapped joint *p*-values are then the fraction of the 2,500 rows for which the number of simulated rejections is greater than or equal to the number of actual rejections using 0.05 and 0.10.

The Cushman and Michael (2011) method is most easily explained by example. Suppose we were interested in the subset consisting of just the first four tests in Table 1. The actual *p*-values using the first 2,500 replications and the AR(1) DGP are 0.332, 0.237, 0.051, and 0.234.[35] Therefore, we have one rejection at the 0.051 level, two at the 0.234 level, three at the 0.237 level, and four at the 0.332 level.

34. The classic Bonferroni method adjusts the significance level to a/n, or multiply the *p*-values by n. None of my *p*-values would be remotely significant according to this approach.

35. The reader may notice that these are slightly different from those in Table 1. This is because the *p*-values in Table 1 come from 5000 replications where available.

We then use the bootstrapped joint *p*-value distribution to find that, with k the number of rejections, P($k \geq 1$) at the 0.051 level equals 0.107, P($k \geq 2$) at the 0.234 level equals 0.263, P($k \geq 3$) at the 0.237 level equals 0.172, and P($k \geq 4$) at the 0.332 level equals 0.145. We focus on the smallest probability, 0.107, which is for $k \geq 1$ in this case. The value of 0.107 is (almost) significant at the 0.10 level. But we need to take into account that we have computed four such probabilities. That is, how unlikely is it to find a P($k \geq i$) = 0.107 for one *or more* values of i? To answer this, the next step is to compute the probability that any one or more of the four rejection outcomes ($k \geq i$, i = 1 to 4) could have occurred with the probability of the most significant one, that is, with a probability of 0.107. To do this, we first determine what individual significance level gives a probability of 0.107 for each P($k \geq i$), i = 1 to 4. We already have that P($k \geq 1$) at the 0.051 level equals 0.107. We go on to find that P($k \geq 2$) at the 0.094 level, P($k \geq 3$) at the 0.161 level, and P($k \geq 4$) at the 0.271 level all equal 0.107.[36] The proportion of the 2,500 replications in which one or more of the four outcomes occur turns out to be 0.173, which is the Cushman-Michael joint *p*-value.

Appendix 3. Monte Carlo analysis of the size and power of the joint *p*-value approaches

I analyze joint-test *p*-value performance under a unit root process and several trend stationary processes. The data-generating process (DGP) in each case is based on the actual real gross domestic product (GDP) data. I start with 400 simulated data sets that play the role of "actual" data sets. To each I apply five unit root tests: DF-GLS-MAIC with linear and quadratic trend, the ADF-AIC with linear trend, and the SP-AIC with linear and quadratic trend. I bootstrap the individual and joint *p*-values just as I did with the actual data set. However, the time involved for the bootstrapping, which is particularly long because it must be repeated for 400 simulated actual data sets, requires me to use far less than the 2,500 or 5,000 replications applied to the real data set. I therefore perform 400 replications of the individual tests to get the joint *p*-values for each of the 400 simulated actual data sets.

The unit root DGP is an AR(1) first-difference model, and the 400 simulated actual data sets are the first 400 for this DGP used in the bootstrapping of the tests in the main text. Next, there are four trend stationary DGPs. They are based on estimated OLS trends from the actual real GDP data, first a linear trend and then

36. For each value of i, this requires applying a trial-and-error, iterative search to the joint *p*-value matrix.

a quadratic trend. An AR(2) model is then estimated for the two sets of residuals, and 400 new sets of residuals are then simulated. These are added to the previously estimated trends to get 400 sets of simulated actual data. I thus end up with (1) a linear trend model using the linear trend and linear trend residuals, (2) a linear trend model using the linear trend and quadratic trend residuals, (3) a quadratic trend model using the quadratic trend and linear trend residuals, and (4) a quadratic trend model using the quadratic trend and quadratic trend residuals. The reason for the various mixtures of trends and residuals is that the quadratic trend residuals are much smaller than the linear trend residuals. Thus, I am able to separate the effects on power of trends and residuals.

Table A3 shows the proportion of joint *p*-values from the five DGPs that exceed 0.05 and 0.10 among the 400 'actual' data sets for each DGP. For the unit root DGP the results estimate the sizes of the joint tests, and for other DGPs the results estimate the powers. The results give no firm evidence that the joint tests suffer from significant size distortion, as none of the proportions are very far from the nominal values of 0.05 or 0.10. I also judge that power is reasonable in most cases. The exception is that all the joint approaches have trouble with a nonlinear trend with the large residuals from the linear trend model. The differences among the joint approaches do not seem sufficiently large to draw any strong conclusion that one is generally superior to any other. A possible exception is the power disadvantage for the 'count 0.10 rejections' method under quadratic trends.

TABLE A3. Size and power of joint *p*-value procedures

		Joint procedure				
Data generating procedure		Simes NB	Simes	Count 0.05	Count 0.10	CM
Unit root	0.05 size	0.050	0.058	0.038	0.038	0.050
Linear trend with linear trend residuals	0.05 power	0.218	0.240	0.205	0.225	0.263
Quadratic trend with linear trend residuals	0.05 power	0.105	0.120	0.015	0.013	0.090
Linear trend with quadratic trend residuals	0.05 power	0.730	0.755	0.708	0.735	0.770
Quadratic trend with quadratic trend residuals	0.05 power	0.478	0.510	0.115	0.088	0.458
Unit root	0.10 size	0.090	0.108	0.065	0.063	0.100
Linear trend with linear trend residuals	0.10 power	0.333	0.415	0.295	0.340	0.433
Quadratic trend with linear trend residuals	0.10 power	0.173	0.205	0.123	0.040	0.165
Linear trend with quadratic trend residuals	0.10 power	0.853	0.883	0.828	0.845	0.880
Quadratic trend with quadratic trend residuals	0.10 power	0.618	0.673	0.538	0.208	0.620

Notes: "Simes NB" is the Simes approach, not bootstrapped. "CM" is the Cushman-Michael approach.

Appendix 4. Unobserved components estimates adjusted for mean bias

The variance and lag parameter estimates from STAMP are: $s_\eta^2 = 3.62 \times 10^{-5}$; $s_\varepsilon^2 = 2.66 \times 10^{-5}$; $r_1 = 0.65$; $r_2 = 0.03$; $r_3 = -0.17$ (where r_j are the estimated lag j serial correlation parameters). Suppose we repeatedly simulate the model of equations (1a) and (1b) including the substitution for serial correlation, and use a DGP with random, independent, heteroskedastic errors (heteroskedasticity estimated from the data as with the unit root tests), and the above variance and lag parameter estimates. With 50 simulations, the means of the distributions of estimated parameters in the replications are mostly quite different from the above results from the actual data: $s_\eta^2 = 6.33 \times 10^{-5}$; $s_\varepsilon^2 = 1.01 \times 10^{-5}$; $r_1 = 0.45$; $r_2 = 0.07$; $r_3 = -0.15$. This strongly hints that the initial estimates are biased. The process generating the data's results apparently has a much smaller permanent component, a much larger transitory component, and a larger autoregressive first order lag coefficient.

To get unbiased estimates, I employ an iterative process as follows (based on Rudebusch 1992). After a set of simulations, re-simulate the model after adjusting each DGP parameter by a fraction of the distance between the desired distributional mean (the parameter estimate from the actual data) and the mean of the simulated distribution using the current DGP's parameter values. Repeat until the means of the latest DGP parameter estimates are all very close to the estimates from the actual data set ("convergence"). The DGP parameters of the final set of simulations are the unbiased estimates.

For example, the permanent variance component for the first iteration's DGP is $[3.62 + 0.5 \times (3.62 - 6.33)] \times 10^{-5} = 2.26 \times 10^{-5}$. The mean of the permanent component variance from this DGP is 4.98×10^{-5}, which is closer than before to the desired value of 3.62×10^{-5}. For the second iteration's DGP, the permanent component variance is set to $[2.26 + 0.5 \times (3.62 - 4.98)] \times 10^{-5} = 1.58 \times 10^{-5}$. Convergence for all five parameters occurs with the seventh iteration, where convergence is defined as the mean of the iteration's simulated variance values being within 5 percent of the estimates from the actual data, and the mean of the iteration's simulated lag parameter values being within 0.02 of the estimates from the actual data. There are 1,200 replications in the final, convergence-achieving set of simulations. In the earlier iterations, the fractional adjustment is 0.5 and in the later ones 0.25. The final DGP variances and lag parameters are: $s_\eta^2 = 8.36 \times 10^{-6}$; $s_\varepsilon^2 = 4.71 \times 10^{-5}$; $r_1 = 0.94$; $r_2 = -0.05$; $r_3 = -0.18$.

The final iteration's distributions for the 1,200 replications show very high variability for the variances. Thus, the ratios of permanent to transitory standard deviations also have very high variability. The 5th to the 95th percentile range for the permanent to transitory standard deviation ratio is 0.16 to an 'indefinitely large' number (i.e., the denominator of the ratio is zero).

Appendix 5. Simulation evidence for the out-of-sample forecast contest of Figure 2

Let us postulate a world where real GDP evolves over its 1947:1–2007:3 period according to the VECM of section 6b, with two breaks. The first break occurs in 1970:2, which is the upper 67 percent confidence limit for the break in services consumption in Stock and Watson (2002), and which could be considered a compromise between their point estimate of the break (1969:4) and Perron and Wada's (2009) real GDP trend break date of 1973:1. The second break occurs in 1991:4, which is Stock and Watson's (2002) break date for nondurable goods consumption. Now let us simulate the model of this world using random residuals drawn from normal distributions with moving three-period covariance matrices (a wild bootstrap to account for contemporaneous correlation and heteroskedasticity including the Great Moderation). Replicate this world and time period 1,000 times, each time performing the forecasting contest with 181 forecasts for each of the six models. Then compute the proportion of times each model that allows for a unit root (four models) has a lower forecast standard deviation than a trend stationary model (two models). The results for horizons 4 and 8:

- VECM < AR-tr = 0.91, 0.80
- VECM < VAR-tr = 0.88, 0.74
- ARMA < AR-tr = 0.82, 0.83
- ARMA < VAR-tr = 0.74, 0.69
- AR-dif < AR-tr = 0.93, 0.87
- AR-dif < VAR-tr = 0.86, 0.74
- VAR-dif < AR-tr = 0.96, 0.93
- VAR-dif < VAR-tr = 0.93, 0.82

The models specifying unit roots usually beat the models that don't.

Appendix 6. Data and code

The data and programs used in this research are contained in two files available for download. **The first file (1.3 MB, .zip)** includes the documentation and contains everything except for five large TSP databanks with bootstrapped data sets, which are in **the second file (44 MB, .zip)**.

References

Ayat, Leila, and Peter Burridge. 2000. Unit Root Tests in the Presence of Uncertainty about the Non-Stochastic Trend. *Journal of Econometrics* 95(1): 71–96.

Barro, Robert J., and Xavier Sala-i-Martin. 1992. Convergence. *Journal of Political Economy* 100(2): 223–251.

Beechey, Meredith, and Pär Österholm. 2008. Revisiting the Uncertain Unit Root in GDP and CPI: Testing for Non-Linear Trend Reversion. *Economics Letters* 100(2): 221–223.

Ben-David, Dan, Robin L. Lumsdaine, and David H. Papell. 2003. Unit Roots, Postwar Slowdowns and Long-Run Growth: Evidence from Two Structural Breaks. *Empirical Economics* 28(2): 303–319.

Bierens, Herman J. 2011. EasyReg International. Department of Economics, Pennsylvania State University (State College, Pa.). **Link**

Blanchard, Olivier, Eugenio Cerutti, and Lawrence Summers. 2015. Inflation and Activity—Two Explorations and Their Monetary Policy Implications. In *Inflation and Unemployment in Europe: Conference Proceedings*, 25–46. Frankfurt am Main: European Central Bank. **Link**

Bordo, Michael G., and Joseph G. Haubrich. 2012. Deep Recessions, Fast Recoveries, and Financial Crises: Evidence from the American Record. *NBER Working Paper* 18194, National Bureau of Economic Research (Cambridge, Mass.). **Link**

Brüggemann, Ralf, and Helmut Lütkepohl. 2001. Lag Selection in Subset VAR Models with an Application to a U.S. Monetary System. In *Econometric Studies: A Festschrift in Honour of Joachim Frohn*, eds. Ralph Friedmann, Lothar Knüppel, and Helmut Lütkepohl, 107–128. Münster: LIT Verlag.

Campbell, John Y., and N. Gregory Mankiw. 1987. Are Output Fluctuations Transitory? *Quarterly Journal of Economics* 102(4): 857–880.

Cheung, Yin-Wong, and Menzie D. Chinn. 1997. Further Investigation of the Uncertain Unit Root in GNP. *Journal of Business and Economic Statistics* 15(1): 68–73.

Chinn, Menzie D. 2009. Interesting Econometric Result of the Day: And the Prospects for a Growth Bounceback. *Econbrowser* (econbrowser.com), March 4. **Link**

Christiano, Lawrence J., and Martin Eichenbaum. 1990. Unit Roots in Real GNP: Do We Know, and Do We Care? *Carnegie-Rochester Conference Series on Public Policy* 32(1): 7–62.

Cochrane, John H. 1991a. Comment on "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots," by John Y. Campbell and Pierre Perron. *NBER Macroeconomics Annual* 6(1): 201–210. **Link**

Cochrane, John H. 1991b. A Critique of the Application of Unit Root Tests. *Journal of Economic Dynamics and Control* 15(2): 275–284.

Cochrane, John H. 1994. Permanent and Transitory Components of GNP and Stock Prices. *Quarterly Journal of Economics* 109(1): 241–265.

Cochrane, John H. 2012. Just How Bad Is the Economy? *The Grumpy Economist* (johnhcochrane.blogspot.com), July 31. **Link**

Cochrane, John H. 2015a. Unit Roots, Redux. *The Grumpy Economist* (johnhcochrane.blogspot.com), April 24. **Link**

Cochrane, John H. 2015b. Unit Roots in English and Pictures. *The Grumpy Economist* (johnhcochrane.blogspot.com), April 27. **Link**

Cochrane, John H. 2015c. Email correspondence with David Cushman, November 2.

Cook, Steven. 2008. More Uncertainty: On the Trending Nature of Real GDP in the US and UK. *Applied Economics Letters* 15(9): 667–670.

Council of Economic Advisers. 2009. Economic Projections and the Budget Outlook. *Council of Economic Advisers Fact Sheets and Reports* (White House, Washington, D.C.), February 28. **Link**

Cushman, David O. 2002. Nonlinear Trends and Co-trending in Canadian Money Demand. *Studies in Nonlinear Dynamics and Econometrics* 6(1): 1–27.

Cushman, David O. 2008. Real Exchange Rates May Have Nonlinear Trends. *International Journal of Finance and Economics* 13(2): 158–173.

Cushman, David O. 2012. Mankiw vs. DeLong and Krugman on the CEA's Real GDP Forecasts in Early 2009: What Might a Time Series Econometrician Have Said? *Econ Journal Watch* 9(3): 309–349. **Link**

Cushman, David O. 2013. Paul Krugman Denies Having Concurred with an Administration Forecast: A Note. *Econ Journal Watch* 10(1): 108–115. **Link**

Cushman, David O., and Nils Michael. 2011. Nonlinear Trends in Real Exchange Rates: A Panel Unit Root Test Approach. *Journal of International Money and Finance* 30(8): 1619–1637.

DeLong, J. Bradford. 2009. Permanent and Transitory Components of Real GDP. *Brad DeLong's Semi-Daily Journal* (delong.typepad.com), March 3. **Link**

Duggal, Vijaya G., Cynthia Saltzman, and Lawrence R. Klein. 1999. Infrastructure and Productivity: A Nonlinear Approach. *Journal of Econometrics* 92(1): 47–74.

Easterly, William. 2012. US Election Depends on Whether Voters Believe Output Has a Unit Root. August 8. Development Research Institute at NYU (New York). **Link**

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock. 1996. Efficient Tests for an Autoregressive Unit Root. *Econometrica* 64(4): 813–836.

Evans, George W. 1989. Output and Unemployment Dynamics in the United States: 1950–1985. *Journal of Applied Econometrics* 4(3): 213–237.

Farmer, Roger E. A. 2015. There is No Evidence that the Economy is Self-Correcting (Very Wonkish). *Roger Farmer's Economic Window* (rogerfarmer-blog.blogspot.com), April 16. **Link**

Gonçalves, Sílvia, and Lutz Kilian. 2004. Bootstrapping Autoregressions with Conditional Heteroskedasticity of Unknown Form. *Journal of Econometrics* 123(1): 89–120.

Gordon, Stephen. 2009. The Unit-Root Zombie: Dead, but Still Wreaking Havoc. *Worthwhile Canadian Initiative* (worthwhile.typepad.com), March 14. **Link**

Hanck, Christoph. 2012. Multiple Unit Root Tests Under Uncertainty Over the Initial Condition: Some Powerful Modifications. *Statistical Papers* 53(3): 767–774.

Hansen, Bruce E. 2007. Least Squares Model Averaging. *Econometrica* 75(4): 1175–1189.

Harvey, David I., Stephen J. Leybourne, and A. M. Robert Taylor. 2009. Unit Root Testing in Practice: Dealing with Uncertainty Over the Trend and Initial Condition. *Econometric Theory* 25(3): 587–636.

Harvey, David I., Stephen J. Leybourne, and A. M. Robert Taylor. 2011. Testing for Unit Roots and the Impact of Quadratic Trends, with an Application to Relative Primary Commodity Prices. *Econometric Reviews* 30(5): 514–547.

Hendry, David F., and Hans-Martin Krolzig. 2001. *Automatic Econometric Model Selection Using PcGets 1.0*. London: Timberlake Consultants.

Kapetanios, George, Yongcheol Shin, and Andy Snell. 2003. Testing for a Unit Root in the Nonlinear STAR Framework. *Journal of Econometrics* 112(2): 359–379.

Kim, Chang-Jin, and Charles R. Nelson. 1999. Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle. *Review of Economics and Statistics* 81(4): 608–616.

Kling, Arnold. 2015. Do Unit Roots Ruin the Concept of Potential GDP? *Askblog* (arnoldkling.com), April 21. **Link**

Koopman, Siem Jan, Andrew C. Harvey, Jurgen A. Doornik, and Neil Shephard. 2000. *Stamp: Structural Time Series Analyser, Modeller and Predictor*. London: Timberlake Consultants Press.

Krane, Spencer D. 2011. Professional Forecasters' Views of Permanent and Transitory Shocks to GDP. *American Economic Journal: Macroeconomics* 3(1): 184–211.

Krugman, Paul. 2009. Roots of Evil (Wonkish). *The Conscience of a Liberal, New York Times*, March 3. **Link**

Leybourne, Stephen, Paul Newbold, and Dimitrios Vougas. 1998. Unit Roots and Smooth Transitions. *Journal of Time Series Analysis* 19(1): 83–98.

Lumsdaine, Robin L., and David H. Papell. 1997. Multiple Trend Breaks and the Unit Root Hypothesis. *Review of Economics and Statistics* 79(2): 212–218.

MacKinnon, James G. 1994. Approximate Asymptotic Distribution Functions for Unit Root and Cointegration Tests. *Journal of Business and Economic Statistics* 12(2): 167–176.

Mankiw, N. Gregory. 2009a. Team Obama on the Unit Root Hypothesis. *Greg Mankiw's Blog* (gregmankiw.blogspot.com), March 3. **Link**

Mankiw, N. Gregory. 2009b. Wanna Bet Some of That Nobel Money? *Greg Mankiw's Blog* (gregmankiw.blogspot.com), March 4. **Link**

McCallum, Bennett. 1991. Discussion of "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots," by John Y. Campbell and Pierre Perron. *NBER Macroeconomics Annual* 6: 218–219. **Link**

McCloskey, Deirdre N., and Stephen T. Ziliak. 2012. Statistical Significance in the New Tom and the Old Tom: A Reply to Thomas Mayer. *Econ Journal Watch* 9(3): 298–308. **Link**

Mitra, Sinchan, and Tara N. Sinclair. 2012. Output Fluctuations in the G-7: An Unobserved Components Approach. *Macroeconomic Dynamics* 16(3): 396–422.

Morley, James C., Charles R. Nelson, and Eric Zivot. 2003. Why Are the Beveridge-Nelson and Unobserved-Components Decompositions of GDP So Different? *Review of Economics and Statistics* 85(2): 235–243.

Murray, Christian J., and Charles R. Nelson. 2000. The Uncertain Trend in U.S. GDP. *Journal of Monetary Economics* 46(1): 79–95.

Murray, Christian J., and Charles R. Nelson. 2004. The Great Depression and Output Persistence: A Reply to Papell and Prodan. *Journal of Money, Credit and Banking* 36(3): 429–432.

Nelson, Charles R., and Charles I. Plosser. 1982. Trends and Random Walks in Macroeconomic Time Series. *Journal of Monetary Economics* 10(2): 139–162.

Ng, Serena, and Pierre Perron. 1995. Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag. *Journal of the American Statistical Association* 90(429): 268–281.

Ng, Serena, and Pierre Perron. 2001. Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power. *Econometrica* 69(6): 1519–1554.

Nunes, Joao Marcus Marinho. 2013. Never Reason from the Previous Peak. *Historinhas* (thefaintofheart.wordpress.com), December 29. **Link**

Ouliaris, Sam, Yoon Y. Parl, and Peter C. B. Phillips. 1989. Testing for a Unit Root in the Presence of a Maintained Trend. In *Advances in Econometrics and Modelling*, ed. Raj Baldev, 7–28. Amsterdam: Kluwer.

Papell, David H. 1997. Searching for Stationarity: Purchasing Power Parity under the Current Float. *Journal of International Economics* 43(3–4): 313–332.

Papell, David H., and Ruxandra Prodan. 2004. The Uncertain Unit Root in U.S. Real GDP: Evidence with Restricted and Unrestricted Structural Change. *Journal of Money, Credit and Banking* 36(3): 423–427.

Perron, Pierre. 1989. The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis. *Econometrica* 57(6): 1361–1402.

Perron, Pierre, and Tatsuma Wada. 2009. Let's Take a Break: Trends and Cycles in US Real GDP. *Journal of Monetary Economics* 56(6): 749–765.

Phillips, Peter C. B. 1987. Time Series Regression with a Unit Root. *Econometrica* 55(2): 277–301.

Phillips, Peter C. B., and Pierre Perron. 1988. Testing for a Unit Root in Time Series Regression. *Biometrika* 75(2): 335–346.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2010. Multiple Testing. In *The New Palgrave Dictionary of Economics*, online ed., eds. Steven N. Durlauf and Lawrence E. Blume. Basingstoke, UK: Palgrave Macmillan. **Link**

Rudebusch, Glenn D. 1992. Trends and Random Walks in Macroeconomic Time Series: A Re-Examination. *International Economic Review* 33(3): 661–680.

Rudebusch, Glenn D. 1993. The Uncertain Unit Root in Real GNP. *American Economic Review* 83(1): 264–272.

Schmidt, Peter, and Peter C. B. Phillips. 1992. LM Tests for a Unit Root in the Presence of Deterministic Trends. *Oxford Bulletin of Economics and Statistics* 54(3): 257–287.

Schwert, G. William. 1989. Tests for Unit Roots: A Monte Carlo Investigation. *Journal of Business and Economic Statistics* 7(2): 147–159.

Shelley, Gary L., and Frederick H. Wallace. 2011. Further Evidence Regarding Nonlinear Trend Reversion of Real GDP and the CPI. *Economics Letters* 112(1): 56–59.

Simes, R. J. 1986. An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* 73(3): 751–754.

Stock, James H. 1990. Unit Roots in Real GNP: Do We Know and Do We Care? A Comment. *Carnegie-Rochester Conference Series on Public Policy* 32(1): 63–82.

Stock, James H., and Mark W. Watson. 1986. Does GNP Have a Unit Root? *Economics Letters* 22(2–3): 147–151.

Stock, James H., and Mark W. Watson. 1999. Business Cycle Fluctuations in U.S. Macroeconomic Time Series. In *Handbook of Macroeconomics*, vol. 1A, eds. John B. Taylor and Michael Woodford, 3–64. Amsterdam: Elsevier.

Stock, James H., and Mark W. Watson. 2002. Has the Business Cycle Changed and Why? *NBER Macroeconomics Annual* 17: 159–230. **Link**

Stulz, René M., and Walter Wasserfallen. 1985. Macroeconomic Time Series, Business Cycles and Macroeconomic Policies. *Carnegie-Rochester Conference Series on Public Policy* 22: 9–54.

Summers, Lawrence. 2015. Advanced Economies Are So Sick We Need a New Way to Think About Them. *Larry Summers' Blog* (larrysummers.com), November 3. **Link**

Sumner, Scott. 2014. The Forces of Evil Easily Triumph over Krugman and DeLong. *TheMoneyIllusion* (themoneyillusion.com), January 31. **Link**

Taylor, John B. 2012. Debate and Evidence on the Weak Recovery. *Economics One* (economicsone.com), May 2. **Link**

Taylor, John B. 2014. Rapid Growth or Stagnation: An Economic Policy Choice. *Journal of Policy Modeling* 36(4): 641–648.

Taylor, John B. 2015. Can We Restart This Recovery All Over Again? *Economics One* (economicsone.com), September 12. **Link**

Tsay, Ruey S. 2005. *Analysis of Financial Time Series*, 2nd ed. Hoboken, N.J.: John Wiley & Sons.

Vougas, Dimitrios V. 2007. Is the Trend in Post-WW II US Real GDP Uncertain or Non-Linear? *Economics Letters* 94(3): 348–355.

Ziliak, Stephen T., and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, Mich.: University of Michigan Press.

Zivot, Eric, and Donald W. K. Andrews. 1992. Further Evidence On the Great Crash, the Oil-Price Shock, and the Unit Root Hypothesis. *Journal of Business and Economic Statistics* 10(3): 251–270.

About the Author

David O. Cushman is Professor Emeritus of Economics at Westminster College (Pennsylvania) and the University of Saskatchewan. He received a Ph.D. from Vanderbilt University. A list of his papers is found at his Google Scholar page (link). His email address is cushmado@gmail.com.

Go to archive of Economics in Practice section
Go to January 2016 issue