Testing for and Estimating Arbitrarily Time-varying Forecast Bias*

Neil R. Ericsson†

Abstract
Impulse indicator saturation (IIS) generalizes the standard Mincer–Zarnowitz test of time-invariant forecast bias by allowing for arbitrarily time-varying forecast bias. Using both approaches, the current paper analyzes potential biases in different U.S. government agencies' one-year-ahead forecasts of U.S. gross federal debt over 1984–2012. Standard tests typically fail to detect biases, whereas IIS detects economically large and highly significant time-varying biases, particularly at turning points in the business cycle. IIS defines a generic procedure for examining forecast properties; it explains why standard tests fail to detect bias; and it provides a mechanism for potentially improving forecasts.

Key Words: Autometrics, bias, debt, federal government, forecasts, impulse indicator saturation, projections, United States

1. Introduction

Government debt has attracted considerable attention during the recent financial crisis and Great Recession. In the United States, federal debt limits, sequestration, and the federal government shut-down have posed substantial economic, political, and policy challenges; see *The Economist* (November 20, 2010), Podkul (2011), Bernanke (2011, 2013), and Chokshi (2013) *inter alia*. In Europe, government debt and fiscal policy are central to current discussions about the euro-area crisis. Because future outcomes of government debt are unknown, forecasts of that debt may matter in government policy, so it is of interest to ascertain how good those forecasts are, and how they might be improved. A central focus in forecast evaluation is forecast bias, especially because forecast bias is systematic, and because ignored forecast biases may have substantive adverse consequences for policy.

Building on Martinez (2011, 2015), the current paper analyzes potential biases in different U.S. government agencies' one-year-ahead forecasts of the U.S. gross federal debt over 1984–2012. Standard tests typically do not detect biases in these forecasts. However, a recently developed technique—impulse indicator saturation—detects economically large and highly statistically significant time-varying biases in the forecasts, particularly for 1990, 1991, 2001–2003, and 2008–2011. Biases differ according to the agency making the forecasts as well as over time. Biases are

*The views in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System. The author is grateful to Danny Bachman, Russell Davidson, Ed Gamber, David Hendry, Stedman Hood, Søren Johansen, Fred Joutz, Kajal Lahiri, Jeffrey Liebner, Prakash Loungani, Jaime Marquez, Andrew Martinez, Toshihiko Mukoyama, Bent Nielsen, John Rogers, Tara Sinclair, Herman Stekler, Ben Taylor, and Christopher Williams for helpful discussions and comments; and in addition to Stedman Hood for invaluable research assistance, and to Andrew Martinez for providing the data and forecasts analyzed and for stimulating my interest in this topic. All numerical results were obtained using PcGive Version 14.0B3, Autometrics Version 1.5e, and Ox Professional Version 7.00 in OxMetrics Version 7.00: see Doornik and Hendry (2013) and Doornik(2009).

†Division of International Finance, Board of Governors of the Federal Reserve System, Washington, DC 20551 USA (ericsson@frb.gov), and Research Program on Forecasting, Department of Economics, The George Washington University, Washington, DC 20052 USA (ericsson@gwu.edu)

typically associated with turning points in the business cycle and (to a lesser degree) economic expansions. Impulse indicator saturation defines a generic procedure for examining forecast properties; it explains why standard tests fail to detect forecast bias; and it provides a mechanism for potentially improving the forecasts.

This paper is organized as follows. Section 2 describes the data and the forecasts being analyzed. Section 3 discusses different approaches to testing for potential forecast bias and proposes impulse indicator saturation as a generic test of forecast bias. Section 4 describes indicator saturation techniques, including impulse indicator saturation and several of its extensions. Section 5 presents evidence on forecast bias, using the methods detailed in Sections 3 and 4. Section 6 concludes. More extensive results appear in Ericsson (2014, 2015).

2. The Data and the Forecasts

This section describes the data on the United States gross federal debt and the three different one-year-ahead forecasts of that debt that are analyzed herein. The forecasts are denoted by their sources:

- CBO (Congressional Budget Office) in its *Budget and Economic Outlook*,
- OMB (Office of Management and Budget) in its *Budget of the U.S. Government*, and
- APB (*Analysis of the President's Budget*).

The Congressional Budget Office and the Office of Management and Budget are different agencies within the U.S. federal government. The *Analysis of the President's Budget* is produced by the Congressional Budget Office, but the forecast in the *Analysis of the President's Budget* is referred to as the "APB forecast" in order to distinguish it from the "CBO forecast", which appears in the CBO's *Budget and Economic Outlook*. The agencies' publications detail how debt is forecast and the assumptions made in generating those forecasts. Significantly, the CBO forecast assumes that current law remains unchanged, whereas the OMB and APB forecasts assume that the president's proposed budget is implemented. The assumptions underlying the forecasts, the complex process involved in generating the forecasts, and the goals and objectives of that process are of considerable interest in their own right and merit detailed examination. However, in the spirit of Stekler (1972), Chong and Hendry (1986), and Fildes and Stekler (2002) *inter alia*, the current paper focuses on the properties of the forecasts themselves. The data on the debt are published by the Financial Management Service at the U.S. Department of the Treasury in the *Treasury Bulletin*.

The data on debt are annual (end of fiscal year) over 1984–2012 (29 observations) and are for total gross federal debt outstanding held by the public and the government. The CBO, OMB, and APB forecasts typically are published in late January, early February, and early March respectively, where those months directly precede the end of the fiscal year (September 30); see Martinez (2011, Table 2; 2015) for details. For convenience, these forecasts are called "one-year-ahead", even though the actual horizon is somewhat less than one year, differs for the three forecasts, and varies somewhat from one year to the next. Debt and its forecasts are in billions of U.S. dollars (nominal), and the analysis below is of the logs of debt and of its forecasts.

Figure 1 plots actual U.S. gross federal debt and its forecasts by the CBO, OMB, and APB (in logs). Actual and forecast values appear close, reflecting in part the

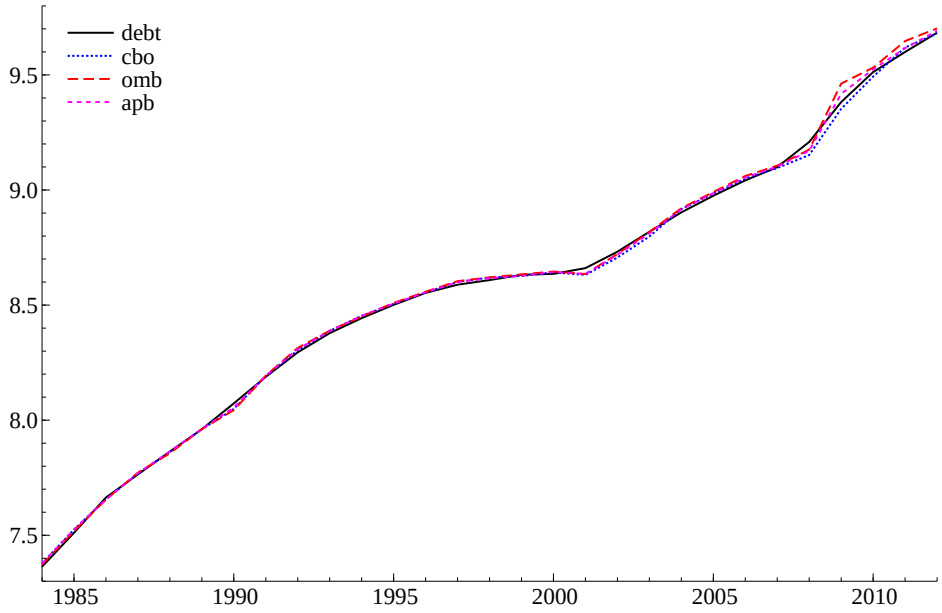

Figure 1: Actual U.S. gross federal debt and its forecasts by the CBO, OMB, and APB (in logs).

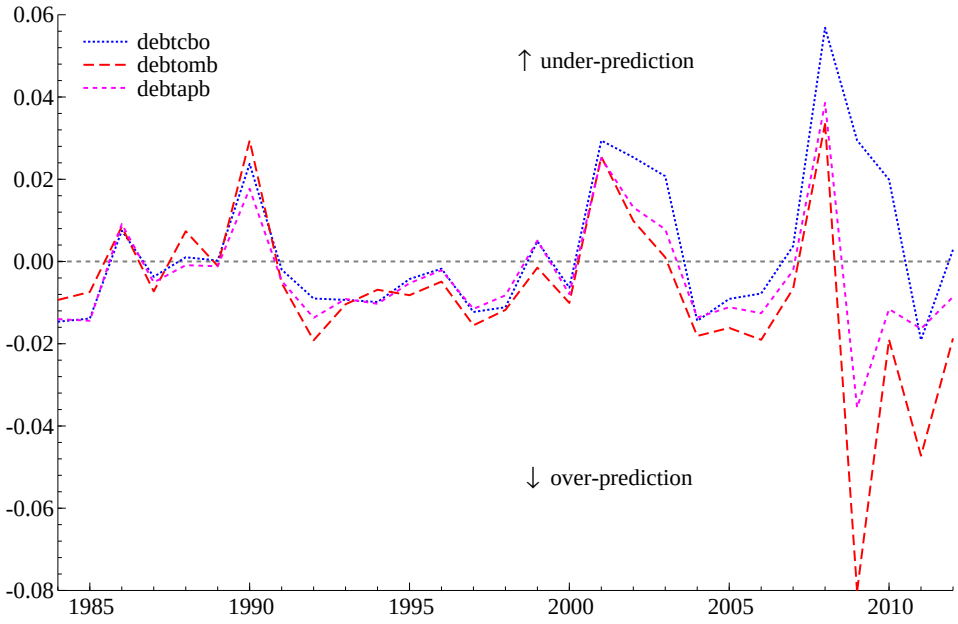

Figure 2: Forecast errors for the log of U.S. gross federal debt.

scale of the graph: debt increases by approximately an order of magnitude over the sample. Figure 2 plots the forecast errors for the log of U.S. gross federal debt. The forecast errors for all three forecasts are often small—under 2% in absolute value—but sometimes they are much larger, and with the magnitude and even the sign differing across agency as well as by forecast date. Forecast errors are often persistent, suggestive of systematic biases in the forecasts. For comparison, the average growth rate of debt and that growth rate's standard deviation are 8.5% and 4.2% respectively.

The presence of forecast bias has both economic significance and statistical significance. That said, the particular sense in which forecast bias is significant depends in part on whether an agency's forecasts are interpreted as "forecasts" or as "projections", where "projections" are in the sense of being policy simulations conditional upon a certain set of assumptions. If the agency's forecasts are interpreted *qua* forecasts, then forecast bias implies potential room for improvement in terms of standard performance measures such as the root mean squared error. If the forecasts are interpreted *qua* projections, then forecast bias implies a limited usefulness of the forecasts as representing interesting hypothetical paths for economic policy. With that in mind, the agencies' forecasts are always referred to as "forecasts" below, while recognizing that some of these forecasts may be more usefully viewed as projections. This broader usage of the term "forecast" is also in line with Clements and Hendry (2002b, p. 2): "A forecast is any statement about the future". For some previous analyses of these and other governmental and institutional forecasts, see Corder (2005), Engstrom and Kernell (1999), Frankel (2011), Joutz and Stekler (2000), Nunes (2013), Sinclair, Joutz, and Stekler (2010), Romer and Romer (2008), and Tsuchiya (2013). Finally, many prior studies compare forecasts whose assumptions differ from each other. Hence, the differing assumptions of the CBO, OMB, and APB forecasts are not grounds *per se* for not comparing the forecasts.

3. Approaches to Detecting Forecast Bias

This section considers different approaches for assessing potential forecast bias, starting with the standard test of forecast bias by Mincer and Zarnowitz (1969). This section then discusses how Chong and Hendry's (1986) forecast-encompassing test is interpretable as a test of time-varying forecast bias. Finally, this section proposes using impulse indicator saturation as a generic test of arbitrarily time-varying forecast bias. This generic test generalizes the Mincer–Zarnowitz test, which is a test of a constant (i.e., time-invariant) forecast bias.

Mincer and Zarnowitz (1969, pp. 8–11) suggest testing for forecast bias by regressing the forecast error on an intercept and testing whether the intercept is statistically significant. That is, for a variable y_t at time t and its forecast $\hat{y}_t$, estimate the equation:

$$(y_t - \hat{y}_t) \;=\; a \;+\; e_t \qquad t = 1, \ldots, T, \tag{1}$$

where a is the intercept, e_t is the error term at time t, and T is number of observations. A test of $a = 0$ is interpretable as a test that the forecast $\hat{y}_t$ is unbiased for the variable y_t. For one-step ahead forecasts, the error e_t may be serially uncorrelated, in which case a t- or F-statistic may be appropriate. For multi-step ahead forecasts, e_t generally will be serially correlated; hence inference about the intercept a may require some accounting for that autocorrelation.

Mincer and Zarnowitz (1969, p. 11) also propose a variant of equation (1) in

which the coefficient on $\hat{y}_t$ is estimated rather than imposed. That variant is:

$$y_t \;=\; a_0 \;+ a_1 \hat{y}_t + \;e_t \qquad t = 1, \ldots, T, \tag{2}$$

where a_0 is the intercept, and a_1 is the coefficient on $\hat{y}_t$. Mincer and Zarnowitz (1969) interpret a test that $a_1 = 1$ as a test of the efficiency of the forecast $\hat{y}_t$ for the outcome y_t. The joint hypothesis $\{a_0 = 0, a_1 = 1\}$ is of interest to test as well. Subtracting $\hat{y}_t$ from both sides, equation (2) may be conveniently rewritten as:

$$(y_t - \hat{y}_t) \;=\; a_0 \;+ a_1^* \hat{y}_t + \;e_t \qquad t = 1, \ldots, T, \tag{3}$$

where $a_1^* = a_1 - 1$. Hence, the hypothesis $\{a_0 = 0, a_1^* = 0\}$ in equation (3) is equivalent to $\{a_0 = 0, a_1 = 1\}$ in equation (2).

Below, "Mincer–Zarnowitz A" denotes the regression-based test of $a = 0$ in equation (1), whereas "Mincer–Zarnowitz B" denotes the regression-based test of $\{a_0 = 0, a_1^* = 0\}$ in equation (3). While equations (2) and (3) are equivalent, equation (3) is reported below because it parallels the structure of equation (1), with $y_t - \hat{y}_t$ as the dependent variable. Mincer–Zarnowitz A (i.e., testing that $a = 0$ in equation (1)) is itself equivalent to testing $a_0 = 0$ in equation (3), subject to the restriction that $a_1^* = 0$. See Holden and Peel (1990) and Stekler (2002) for expositions on these tests as tests of unbiasedness and efficiency, and Sinclair, Stekler, and Carnow (2012) for a recent discussion.

Chong and Hendry (1986) propose another test about forecast errors, namely, a test of whether one model's forecasts provide information about another model's forecast errors. If one model's forecasts do provide information about another model's forecast errors, then those forecast errors are in part predictable. If not, then the latter model "forecast-encompasses" the first model. As Ericsson (1992) discusses, a necessary condition for forecast encompassing is having the smallest mean squared forecast error (MSFE); Granger (1989) and Diebold and Mariano (1995) propose tests of whether one model's MSFE is less than another model's MSFE.

Chong and Hendry (1986) and subsequent authors implement many versions of the forecast-encompassing test. One appealing version is based on the regression:

$$\begin{aligned} (y_t - \hat{y}_t) \;&=\; b_0 \;+\; b_1 \cdot (\tilde{y}_t - \hat{y}_t) \;+\; e_t \\ &=\; a_t \;+\; e_t \qquad\qquad\qquad\qquad t = 1, \ldots, T, \end{aligned} \tag{4}$$

where $\hat{y}_t$ is the forecast of y_t by model 1 (say), $\tilde{y}_t$ is the forecast of y_t by model 2, and b_0 and b_1 are regression coefficients. A test of $b_1 = 0$ is interpretable as a test of whether discrepancies between the two models' forecasts are helpful in explaining model 1's forecast errors. The joint hypothesis $\{b_0 = 0, b_1 = 1\}$ is also of interest to test. Equation (4) can be extended to compare several forecasts at once, in which case the right-hand side of (4) includes the differential of each alternative model's forecast relative to model 1's forecast; see Ericsson and Marquez (1993).

Tests of forecast encompassing are interpretable as tests of time-varying forecast bias, as the second line in equation (4) indicates. The subscript t on the intercept a_t emphasizes the time dependence of the potential bias, which here is parameterized as $b_0 + b_1 \cdot (\tilde{y}_t - \hat{y}_t)$. The forecast-encompassing test thus focuses on a specific time-varying form of potential forecast bias.

The time dependence of the forecast bias could be completely general, as follows:

$$\begin{aligned} (y_t - \hat{y}_t) \;&=\; \textstyle\sum_{i=1}^{T} c_i I_{it} \;+\; e_t \\ &=\; a_t \;+\; e_t \qquad\qquad\qquad\qquad t = 1, \ldots, T, \end{aligned} \tag{5}$$

where the impulse indicator I_{it} is a dummy variable that is unity for $t = i$ and zero otherwise, and c_i is the corresponding coefficient for I_{it}. Because the $\{c_i\}$ may have any values whatsoever, the intercept a_t in (5) may vary arbitrarily over time. In this context, a test that all coefficients c_i are equal to zero is a generic test of forecast unbiasedness. Because equation (5) includes T coefficients, equation (5) cannot be estimated unrestrictedly. However, the question being asked can be answered using impulse indicator saturation, as discussed in the following section.

4. Indicator Saturation Techniques

Impulse indicator saturation (IIS) is a general procedure for model evaluation, and in particular for testing parameter constancy. As this section shows, IIS also can be used to test for time-varying forecast bias. Doing so provides a new application of impulse indicator saturation—as a generic test of forecast bias—noting that IIS has previously been employed for model evaluation, model design, and robust estimation. The current section first discusses IIS and its extensions as a procedure for testing parameter constancy and then, drawing on the framework in Section 3, re-interprets existing tests of forecast bias as special cases of IIS and shows how IIS can be used to detect possibly time-varying forecast bias. Section 5 then applies IIS to analyze time-varying bias in government forecasts of the U.S. gross federal debt.

Impulse indicator saturation and extensions. Impulse indicator saturation provides a general procedure for analyzing a model's constancy. Specifically, IIS is a generic test for an unknown number of breaks, occurring at unknown times, with unknown duration and magnitude, anywhere in the sample. IIS is a powerful empirical tool for both evaluating and improving existing empirical models, and use of IIS in forecast development is consistent with a progressive modeling approach; see White (1990). Hendry (1999) proposes IIS as a procedure for testing parameter constancy. See Hendry, Johansen, and Santos (2008), Doornik (2009a), Johansen and Nielsen (2009, 2011), Hendry and Santos (2010), Ericsson (2011a, 2011b, 2012), Ericsson and Reisman (2012a, 2012b), Bergamelli and Urga (2013), Hendry and Pretis (2013), Castle, Doornik, Hendry, and Pretis (2013), and Hendry and Doornik (2014) for further discussion and recent developments.

Impulse indicator saturation uses the zero-one impulse indicator dummies $\{I_{it}\}$ to analyze properties of a model. For a sample of T observations, there are T such dummies, so unrestricted inclusion of all T dummies in a model (thereby "saturating" the sample) is infeasible. However, blocks of dummies *can* be included, and that insight provides the basis for IIS. A simple "bare-bones" example with two equal-sized blocks motivates the generic approach in IIS.

Imagine estimating a specification such as the Mincer–Zarnowitz regression (1) in three steps. First, estimate that model, including impulse indicator dummies for the first half of the sample. That estimation is equivalent to estimating the model over the second half of the sample, ignoring the first half. Drop all statistically insignificant impulse indicator dummies and retain the statistically significant dummies. Second, repeat this process, but start by including impulse indicator dummies for the second half of the sample; and retain the statistically significant ones. Third, re-estimate the original model, including all dummies retained in the two block searches; and select the statistically significant dummies from that combined set. Hendry, Johansen, and Santos (2008) and Johansen and Nielsen (2009) have shown that, under the null hypothesis of correct specification, the fraction of impulse indicator dummies retained is roughly αT, where α is the target size. For

instance, if $T = 100$ and the target size is 1%, then (on average) only one impulse indicator dummy is retained when the model is correctly specified.

If the model is mis-specified such that its implied coefficients are nonconstant over time, IIS has power to detect that nonconstancy. See Hendry and Santos (2010, Section 4) for an example. Interestingly, the residuals of the estimated model *without* any impulse indicator dummies need not lie outside their estimated 95% confidence region, even with a statistically and economically large break in the underlying parameters of the data generation process. Also, the IIS procedure can have high power to detect the break, even though the nature of the break is not utilized in the procedure itself.

In practice, IIS in the Autometrics routine of Doornik and Hendry's (2013) OxMetrics utilizes many blocks, and the partitioning of the sample into blocks may vary over iterations of searches; see also Hendry and Krolzig (1999, 2001, 2005), Hoover and Perez (1999, 2004), and Krolzig and Hendry (2001). IIS is a statistically valid procedure for integrated, cointegrated data; see Johansen and Nielsen (2009). IIS can also serve as a diagnostic statistic for many forms of mis-specification.

Many existing procedures can be interpreted as "special cases" of IIS in that they represent particular algorithmic implementations of IIS. Such special cases include recursive estimation, rolling regression, the Chow (1960) predictive failure statistic (including the 1-step, breakpoint, and forecast versions implemented in OxMetrics), the Andrews (1993) unknown breakpoint test, the Bai and Perron (1998) multiple breakpoint test, tests of extended constancy in Ericsson, Hendry, and Prestwich (1998, pp. 305ff), tests of nonlinearity, intercept correction (in forecasting), and robust estimation. IIS thus provides a general and generic procedure for analyzing a model's constancy, allowing for an unknown number of structural breaks occurring at unknown times with unknown duration and magnitude anywhere in the sample. Algorithmically, IIS also solves the problem of having more regressors than observations by testing and selecting over blocks of variables.

IIS has numerous extensions; see Ericsson (2011b, 2012) and Ericsson and Reisman (2012a) for expositions and developments. Three immediate extensions include step indicator saturation (SIS), which searches across all possible one-off step functions; super saturation, which searches across all impulses and one-off step functions; and ultra saturation, which searches across all impulses, one-off step functions, and one-off broken trends. Higher-order broken trends are also feasible. Other extensions include sequential and non-sequential pairwise impulse indicator saturation; zero-sum pairwise IIS; many many variables for a set of K potential regressors ($K > T$), as in the aggregation test of Ericsson (2011a); factors; principal components; and multiplicative indicator saturation. A saturation procedure may itself be a combination of extensions; and that choice may affect the power of the procedure to detect specific alternatives. See Castle, Clements, and Hendry (2013a), Castle, Doornik, Hendry, and Pretis (2013), Doornik, Hendry, and Pretis (2013), and Ericsson (2011b, 2012) for details, discussion, and examples in the literature.

Re-interpretation and generalization. IIS and its extensions provide a conceptual framework for re-interpreting existing tests of forecast bias. Equally, IIS-type procedures generalize those existing tests to allow for time-varying forecast bias. Ericsson (2014, 2015) discusses how the interpretation of an IIS-based test as a test of forecast *bias* faces certain challenges, and how outside information and the dating of the retained impulses may address those challenges.

The Mincer–Zarnowitz A test (based on equation (1)) is an explicit special case of super saturation in which only the first step dummy (equivalent to the intercept)

is included. The Mincer–Zarnowitz A test is also interpretable as the IIS test based on equation (5), but where $c_1 = c_2 = \ldots = c_T$ is *imposed*, and the hypothesis $c_1 = 0$ is tested.

The Mincer–Zarnowitz B test (based on equation (3)) is a special case of multiplicative indicator saturation in which the dependent variable is the forecast error, the regressors are the intercept and the forecast, and the only multiplicative indicators considered are those multiplied by the first step indicator. Multiplicative indicator saturation also includes the forecast encompassing test and standard tests of strong efficiency as special cases; cf. Holden and Peel (1990) and Stekler (2002).

As equation (5) emphasizes, IIS generalizes the Mincer–Zarnowitz tests to allow for time-varying forecast bias. This observation and the observations above highlight the strength of the Mincer–Zarnowitz tests (that they focus on detecting a constant nonzero forecast bias) and also their weakness (that they assume that the forecast bias *is* constant over time). These characteristics of the Mincer–Zarnowitz tests bear directly on the empirical results in the next section.

5. Evidence on Biases in the Forecasts of Debt

This section examines the CBO, OMB, and APB forecasts of U.S. gross federal debt for potential bias over 1984–2012. Standard (Mincer–Zarnowitz) tests of forecast bias typically fail to detect economically and statistically important biases. By contrast, IIS tests detect economically large and statistically highly significant time-varying biases in the CBO, OMB, and APB forecasts, particularly for 1990, 1991, 2001–2003, and 2008–2011. Forecast biases for a given year differ numerically across the CBO, OMB, and APB, albeit with some similarities. Tests of and estimates of forecast bias using extensions to IIS appear in Ericsson (2014, 2015).

Table 1 reports various statistics for testing for bias in the CBO, OMB, and APB forecasts. The first four rows report statistics for the Mincer–Zarnowitz A regression, the Mincer–Zarnowitz B regression, forecast encompassing (generalized for three forecasts), and IIS: equations (1), (3), (4), and (5) respectively. Entries within a given block of numbers are the F-statistic for testing the null hypothesis against the designated maintained hypothesis, the tail probability associated with that value of the F-statistic (in square brackets), the degrees of freedom for the F-statistic (in parentheses), and (for IIS statistics) the retained dummy variables. Superscript asterisks * and ** denote rejections of the null hypothesis at the 5% and 1% levels respectively, and the null hypothesis includes setting the coefficient on the intercept to zero. K denotes the number of *potential* regressors for selection; and the target size for IIS is chosen much smaller than $1/K$ in order to help ensure that few if any indicators are retained fortuitously.

The Mincer–Zarnowitz results in the first two rows of Table 1 provide little evidence of forecast bias for any of the forecasts. The estimated forecast bias for the CBO is statistically insignificantly different from zero, with F-statistics of 0.67 and 1.30 respectively for the two types of Mincer–Zarnowitz statistic. The Mincer–Zarnowitz statistics for OMB and APB are likewise insignificant, except that the Mincer–Zarnowitz B statistic for OMB is significant at around the 1% level. Thus, the Mincer–Zarnowitz A test fails to detect bias in all three forecasts, and the Mincer–Zarnowitz B test fails to detect bias in two of three forecasts. Standard tests thus provide little evidence of forecast bias.

Row three and four report the forecast-encompassing and IIS statistics, which *always* detect bias for all three forecasts; cf. Martinez (2011, 2015). Notably, IIS

Table 1: Statistics for testing for bias in the CBO, OMB, and APB forecasts.

Statistic (target size)	K	CBO	OMB	APB
Mincer–Zarnowitz A	1	0.67 [0.421] $F(1, 28)$	3.96 [0.056] $F(1, 28)$	1.80 [0.191] $F(1, 28)$
Mincer–Zarnowitz B	2	1.30 [0.290] $F(2, 27)$	5.21* [0.012] $F(2, 27)$	1.05 [0.363] $F(2, 27)$
Forecast-encompassing	3	8.38** [0.000] $F(3, 26)$	19.44** [0.000] $F(3, 26)$	3.12* [0.043] $F(3, 26)$
Impulse indicator saturation (1%)	29	18.50** [0.000] $F(8, 21)$ I_{1990}, $I_{2001}, I_{2002}, I_{2003}$, $I_{2008}, I_{2009}, I_{2010}$	28.04** [0.000] $F(6, 23)$ I_{1990}, I_{2001}, $I_{2008}, I_{2009}, I_{2011}$	14.40** [0.000] $F(5, 24)$ I_{1990}, I_{2001}, I_{2008}, I_{2009}
Intercept from OLS	1	0.27 (0.33) {0.82}	−0.79 (0.40) {−1.99}	−0.36 (0.27) {−1.34}
Intercept from IIS (1%)	29	−0.58 (0.15) {−3.78}	−0.79 (0.18) {−4.43}	−0.60 (0.16) {−3.74}

detects bias for historically and economically consequential years. The dates of several retained impulse dummies are indicative of important events that potentially affected the actual federal debt after its forecasts were made.

1990: Iraq invasion of Kuwait on August 2, 1990; July 1990–March 1991 recession.

2001: March–November 2001 recession; September 11, 2001.

2008, 2009: December 2007–June 2009 recession.

Recessions are dated per the National Bureau of Economic Research (2012); and business-cycle turning points are prominent among the events listed. The four years listed also highlight the difficulties in forecasting the debt, especially in light of unanticipated events that affect both government expenditures and revenues; cf. Alexander and Stekler (1959) and Stekler (1967).

Ericsson (2014, 2015) re-interprets and re-analyzes the estimated biases in light of the dates for the peaks and troughs of the business cycle, as determined by the National Bureau of Economic Research (NBER). This re-interpretation leads to a standardized reformulation of the estimated forecast biases in terms of business-cycle turning points, augmented by a few additional adjustments. This approach draws on Sinclair, Joutz, and Stekler (2010), who analyze the Fed's Greenbook

forecasts similarly; and on Hendry (1999), who re-interprets IIS-detected outliers in an economic and institutional framework; see also Ericsson, Hood, Joutz, Sinclair, and Stekler (2013). Thus, an initially "theory-neutral" analysis using IIS leads to a "theory-guided" interpretation based on historical knowledge of the economy, highlighting the informative date-specific information that IIS provides; see also Hendry and Johansen (2011, 2012). Ericsson (2014, 2015) finds that the biases differ across the agencies making the forecasts and across the peaks (and the troughs) of different business cycles, but appear little affected by other factors and, in particular, do not appear to be politically related. Ericsson (2014, 2015) also discusses some potential implications of forecast bias, noting that a forecast bias is a conditional expectation and examining the importance of the information set on which that expectation is taken.

Table 1 focuses on the statistical significance of the biases for each set of forecasts. The corresponding regressions also permit assessing the extent and economic and numerical importance of the bias for the forecasts. Figure 2 plots the CBO, OMB, and APB forecast errors; Figure 3 plots the IIS estimates of forecast bias.

The forecast biases vary markedly over time, and they exhibit some similarities across agencies. For the CBO forecasts, the bias is approximately 2.5% for 1990 and 2001–2003, 5.6% for 2008, declining thereafter, and –0.6% (and statistically detectably so) for all other years. For the OMB forecasts, the bias is approximately 3% for 1990, 2001, and 2008, –8% for 2009, –4.7% for 2011, and –0.8% for all other years. For the APB forecasts, the bias is approximately 1.8% for 1990, 2.5% for 2001, 4% for 2008, –3.6% for 2009, and –0.6% for all other years. As a reference, the residual standard errors for the IIS regressions are 0.72%, 0.87%, and 0.80% respectively. In several instances, forecast biases exceed 2% in absolute value. These biases are economically large, especially considering that debt is a stock (not a flow), and that the forecasts are made less than nine months prior to the end of the fiscal year.

As Figure 3 shows, the estimated forecast biases are sometimes positive and are othertimes negative. The Mincer–Zarnowitz tests have particular difficulty in detecting such biases because the Mincer–Zarnowitz tests average all biases (both negative and positive) over time, and because the Mincer–Zarnowitz tests assign any time variation in bias to the residual rather than to the bias itself. As an extreme example, the Mincer–Zarnowitz A test has no power to detect a forecast bias that is $+\$10^{100}$ for the first half of the sample and $-\$10^{100}$ for the second half of the sample, even though this bias would be obvious from (e.g.) graphing the data.

Mincer–Zarnowitz tests also can lack power to detect forecast bias if forecast errors have thick tails or are heteroscedastic. Ericsson (2014, 2015) shows that residual diagnostic statistics reject either normality or homoscedasticity for every Mincer–Zarnowitz regression of the CBO, OMB, and APB forecasts. As follows from Johansen and Nielsen (2009), IIS provides robust inference about the intercept in such a situation. While heteroscedasticity-consistent standard errors may provide consistent inference, they fail to improve efficiency of coefficient estimates, whereas robust estimation techniques such as IIS can. Those differences are highlighted in the bottom two rows of Table 1, which compare the estimated intercepts in the (OLS) Mincer–Zarnowitz A regressions with their IIS equivalents. The intercepts in the standard Mincer–Zarnowitz A regressions are statistically insignificant, whereas the intercepts estimated using IIS are highly significant. Estimated standard errors appear in parentheses $(\cdot)$ under regression coefficients, and t-ratios appear in curly brackets $\{\cdot\}$. Even when IIS is viewed purely as a robust esti-

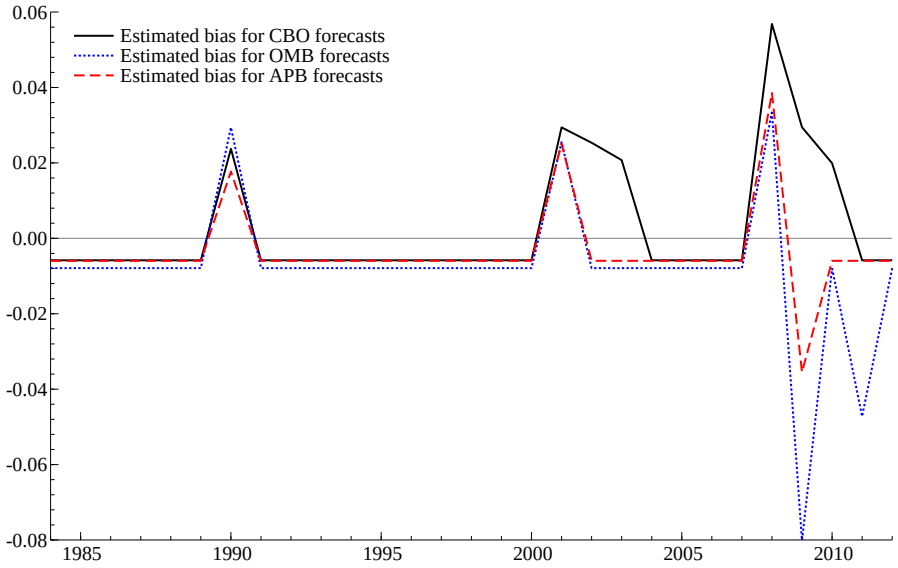

Figure 3: IIS estimates of forecast bias for the log of U.S. gross federal debt.

mation procedure, empirical inferences about bias alter dramatically for the CBO, OMB, and APB forecasts. Bias is present in all three forecasts, and the standard Mincer–Zarnowitz A test fails to detect that bias.

6. Conclusions

Government debt and its forecasts feature prominently in current economic and political discussions. The properties of these forecasts are thus of interest, and it matters how these properties are assessed. Mincer–Zarnowitz tests typically fail to detect biases in the CBO, OMB, and APB one-year-ahead forecasts of U.S. gross federal debt over 1984–2012. By contrast, more general tests based on impulse indicator saturation detect economically large, systematic, and statistically highly significant time-varying biases in the CBO, OMB, and APB forecasts, particularly for 1990, 1991, 2001–2003, and 2008–2011. These biases differ according to the agency making the forecasts, and these biases are closely linked to turning points in the business cycle and (to a lesser degree) economic expansions. The IIS approach also explains *why* Mincer–Zarnowitz tests may fail to detect bias. The Mincer–Zarnowitz tests average over the biases for *all* observations, but those biases may be positive for some observations and negative for others, thereby reducing the tests' power.

Impulse indicator saturation defines a generic procedure for examining forecast properties and, in particular, for detecting and quantifying forecast bias. Forecast bias can be systematic yet time-varying; forecast bias can be difficult to detect in a timely fashion; and forecast bias may have substantive implications for policy analysis. IIS aims to address these issues, with IIS characterizing systematic properties in forecast bias. The IIS approach also links directly to techniques for robustifying forecasts, noting that intercept correction is a variant of super saturation; see Clements and Hendry (1996, 1999, 2002a), Hendry (2006), Castle, Fawcett, and

Hendry (2010), and Castle, Clements, and Hendry (2013b).

The IIS approach has many potential applications, beyond its initial roles in model evaluation and robust estimation. Ericsson (2012) considers its uses for detecting crises, jumps, and changes in regime. IIS also provides a framework for creating near real-time early-warning and rapid-detection devices, such as of financial market anomalies; cf. Vere-Jones (1995) on forecasting earthquakes and earthquake risk, and Goldstein, Kaminsky, and Reinhart (2000) on early warning systems for emerging market economies. Relatedly, the model selection approach in IIS is applicable to nowcasting with a large set of potential explanatory variables, such as those generated from Google Trends; and it generalizes to systems; see Doornik (2009b), Choi and Varian (2012), and Hendry and Doornik (2014).

REFERENCES

Alexander, S. S., and H. O. Stekler (1959) "Forecasting Industrial Production—Leading Series versus Autoregression", *Journal of Political Economy*, 67, 4, 402–409.

Andrews, D. W. K. (1993) "Tests for Parameter Instability and Structural Change with Unknown Change Point", *Econometrica*, 61, 4, 821–856.

Bai, J., and P. Perron (1998) "Estimating and Testing Linear Models with Multiple Structural Changes", *Econometrica*, 66, 1, 47–78.

Bergamelli, M., and G. Urga (2013) "Detecting Multiple Structural Breaks: A Monte Carlo Study and an Application to the Fisher Equation for the US", draft, Cass Business School, London, March.

Bernanke, B. S. (2011) "Fiscal Sustainability", speech, Annual Conference, Committee for a Responsible Federal Budget, Washington, D.C., June 14.

Bernanke, B. S. (2013) "Chairman Bernanke's Press Conference", transcript, Board of Governors of the Federal Reserve System, Washington, D.C., September 18.

Castle, J. L., M. P. Clements, and D. F. Hendry (2013a) "Forecasting by Factors, by Variables, by Both or Neither?", *Journal of Econometrics*, 177, 2, 305–319.

Castle, J. L., M. P. Clements, and D. F. Hendry (2013b) "Robust Approaches to Forecasting Inflation", draft, Economics Department and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford, April.

Castle, J. L., J. A. Doornik, D. F. Hendry, and F. Pretis (2013) "Detecting Location Shifts by Step-indicator Saturation", draft, revised, Economics Department and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford, October.

Castle, J. L., N. W. P. Fawcett, and D. F. Hendry (2010) "Forecasting with Equilibrium-correction Models During Structural Breaks", *Journal of Econometrics*, 158, 1, 25–36.

Choi, H., and H. Varian (2012) "Predicting the Present with Google Trends", *Economic Record*, 88, Special Issue, 2–9.

Chokshi, N. (2013) "Beware Obama's Budget Predictions: Many Forecasts Are Wrong", *National Journal*, April 10 (www.nationaljournal.com).

Chong, Y. Y., and D. F. Hendry (1986) "Econometric Evaluation of Linear Macro-economic Models", *Review of Economic Studies*, 53, 4, 671–690.

Chow, G. C. (1960) "Tests of Equality Between Sets of Coefficients in Two Linear Regressions", *Econometrica*, 28, 3, 591–605.

Clements, M. P., and D. F. Hendry (1996) "Intercept Corrections and Structural Change", *Journal of Applied Econometrics*, 11, 5, 475–494.

Clements, M. P., and D. F. Hendry (1999) *Forecasting Non-stationary Economic Time Series*, MIT Press, Cambridge.

Clements, M. P., and D. F. Hendry (2002a) "Explaining Forecast Failure in Macroeconomics", Chapter 23 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 539–571.

Clements, M. P., and D. F. Hendry (2002b) "An Overview of Economic Forecasting", Chapter 1 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 1–18.

Corder, J. K. (2005) "Managing Uncertainty: The Bias and Efficiency of Federal Macroeconomic Forecasts", *Journal of Public Administration Research and Theory*, 15, 1, 55–70.

Diebold, F. X., and R. S. Mariano (1995) "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics*, 13, 3, 253–263.

Doornik, J. A. (2009a) "Autometrics", Chapter 4 in J. L. Castle and N. Shephard (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, Oxford University Press, Oxford, 88–121.

Doornik, J. A. (2009b) "Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data", draft, Economics Department, University of Oxford, Oxford, September 8 (www.doornik.com/flu/Doornik%282009%29_Flu.pdf).

Doornik, J. A., and D. F. Hendry (2013) *PcGive 14*, Timberlake Consultants Press, London (3 volumes).

Doornik, J. A., D. F. Hendry, and F. Pretis (2013) "Step-indicator Saturation", Discussion Paper No. 658, Department of Economics, University of Oxford, Oxford, June.

Engstrom, E. J., and S. Kernell (1999) "Serving Competing Principals: The Budget Estimates of OMB and CBO in an Era of Divided Government", *Presidential Studies Quarterly*, 29, 4, 820–829.

Ericsson, N. R. (1992) "Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration", *Journal of Policy Modeling*, 14, 4, 465–495.

Ericsson, N. R. (2011a) "Improving Global Vector Autoregressions", draft, Board of Governors of the Federal Reserve System, Washington, D.C., June.

Ericsson, N. R. (2011b) "Justifying Empirical Macro-econometric Evidence in Practice", invited presentation, online conference *Communications with Economists: Current and Future Trends* commemorating the 25th anniversary of the *Journal of Economic Surveys*, November.

Ericsson, N. R. (2012) "Detecting Crises, Jumps, and Changes in Regime", draft, Board of Governors of the Federal Reserve System, Washington, D.C., November.

Ericsson, N. R. (2014) "How Biased Are U.S. Government Forecasts of the Federal Debt?", International Finance Discussion Paper, Board of Governors of the Federal Reserve System, Washington, D.C., forthcoming.

Ericsson, N. R. (2015) "How Biased Are U.S. Government Forecasts of the Federal Debt?", *International Journal of Forecasting*, forthcoming.

Ericsson, N. R., D. F. Hendry, and K. M. Prestwich (1998) "The Demand for Broad Money in the United Kingdom, 1878–1993", *Scandinavian Journal of Economics*, 100, 1, 289–324 (with discussion).

Ericsson, N. R., S. B. Hood, F. Joutz, T. M. Sinclair, and H. O. Stekler (2013) "Greenbook Forecasts and the Business Cycle", draft, Board of Governors of the Federal Reserve System, Washington, D.C., December.

Ericsson, N. R., and J. Marquez (1993) "Encompassing the Forecasts of U.S. Trade Balance Models", *Review of Economics and Statistics*, 75, 1, 19–31.

Ericsson, N. R., and E. L. Reisman (2012a) "Evaluating a Global Vector Autoregression for Forecasting", *International Advances in Economic Research*, 18, 3, 247–258.

Ericsson, N. R., and E. L. Reisman (2012b) "Evaluating a Global Vector Autoregression for Forecasting", International Finance Discussion Paper No. 1056, Board of Governors of the Federal Reserve System, Washington, D.C., November.

Fildes, R., and H. O. Stekler (2002) "The State of Macroeconomic Forecasting", *Journal of Macroeconomics*, 24, 4, 435–468.

Frankel, J. (2011) "Over-optimism in Forecasts by Official Budget Agencies and Its Implications", *Oxford Review of Economic Policy*, 27, 4, 536–562.

Goldstein, M., G. L. Kaminsky, and C. M. Reinhart (2000) *Assessing Financial Vulnerability: An Early Warning System for Emerging Markets*, Institute for International Economics, Washington, D.C.

Granger, C. W. J. (1989) *Forecasting in Business and Economics*, Academic Press, Boston, Massachusetts, Second Edition.

Hendry, D. F. (1999) "An Econometric Analysis of US Food Expenditure, 1931–1989", Chapter 17 in J. R. Magnus and M. S. Morgan (eds.) *Methodology and Tacit Knowledge: Two Experiments in Econometrics*, John Wiley and Sons, Chichester, 341–361.

Hendry, D. F. (2006) "Robustifying Forecasts from Equilibrium-correction Systems", *Journal of Econometrics*, 135, 1–2, 399–426.

Hendry, D. F., and J. A. Doornik (2014) *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*, MIT Press, Cambridge, Massachusetts.

Hendry, D. F., and S. Johansen (2011) "The Properties of Model Selection When Retaining Theory Variables", Discussion Paper No. 11–25, Department of Economics, University of Copenhagen, Copenhagen, September.

Hendry, D. F., and S. Johansen (2012) "Model Discovery and Trygve Haavelmo's Legacy", Discussion Paper No. 598, Department of Economics, University of Oxford, Oxford, March.

Hendry, D. F., S. Johansen, and C. Santos (2008) "Automatic Selection of Indicators in a Fully Saturated Regression", *Computational Statistics*, 23, 2, 317–335, 337–339.

Hendry, D. F., and H.-M. Krolzig (1999) "Improving on 'Data Mining Reconsidered' by K. D. Hoover and S. J. Perez", *Econometrics Journal*, 2, 2, 202–219.

Hendry, D. F., and H.-M. Krolzig (2001) *Automatic Econometric Model Selection Using PcGets 1.0*, Timberlake Consultants Press, London.

Hendry, D. F., and H.-M. Krolzig (2005) "The Properties of Automatic *Gets* Modelling", *Economic Journal*, 115, 502, C32–C61.

Hendry, D. F., and F. Pretis (2013) "Anthropogenic Influences on Atmospheric CO_2", Chapter 12 in R. Fouquet (ed.) *Handbook on Energy and Climate Change*, Edward Elgar, Cheltenham, 287–326.

Hendry, D. F., and C. Santos (2010) "An Automatic Test of Super Exogeneity", Chapter 12 in T. Bollerslev, J. R. Russell, and M. W. Watson (eds.) *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, Oxford University Press, Oxford, 164–193.

Holden, K., and D. A. Peel (1990) "On Testing for Unbiasedness and Efficiency of Forecasts", *The Manchester School*, 58, 2, 120–127.

Hoover, K. D., and S. J. Perez (1999) "Data Mining Reconsidered: Encompassing and the General-to-specific Approach to Specification Search", *Econometrics Journal*, 2, 2, 167–191 (with discussion).

Hoover, K. D., and S. J. Perez (2004) "Truth and Robustness in Cross-country Growth Regressions", *Oxford Bulletin of Economics and Statistics*, 66, 5, 765–798.

Johansen, S., and B. Nielsen (2009) "An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator", Chapter 1 in J. L. Castle and N. Shephard (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, Oxford University Press, Oxford, 1–36.

Johansen, S., and B. Nielsen (2011) "Asymptotic Theory for Iterated One-step Huber-skip Estimators", Discussion Paper No. 11–29, Department of Economics, University of Copenhagen, Copenhagen, November.

Joutz, F., and H. O. Stekler (2000) "An Evaluation of the Predictions of the Federal Reserve", *International Journal of Forecasting*, 16, 1, 17–38.

Krolzig, H.-M., and D. F. Hendry (2001) "Computer Automation of General-to-specific Model Selection Procedures", *Journal of Economic Dynamics and Control*, 25, 6–7, 831–866.

Martinez, A. B. (2011) "Comparing Government Forecasts of the United States' Gross Federal Debt", RPF Working Paper No. 2011–002, Research Program on Forecasting, Center of Economic Research, Department of Economics, The George Washington University, Washington, D.C., February.

Martinez, A. B. (2015) "How Good Are U.S. Government Forecasts of the Federal Debt?", *International Journal of Forecasting*, forthcoming.

Mincer, J., and V. Zarnowitz (1969) "The Evaluation of Economic Forecasts", Chapter 1 in J. Mincer (ed.) *Economic Forecasts and Expectations: Analyes of Forecasting Behavior and Performance*, National Bureau of Economic Research, New York, 3–46.

National Bureau of Economic Research (2012) "US Business Cycle Expansions and Contractions", webpage, National Bureau of Economic Research, Cambridge, MA, April (www.nber.org/cycles.html).

Nunes, R. (2013) "Do Central Banks' Forecasts Take Into Account Public Opinion and Views?", International Finance Discussion Paper No. 1080, Board of Governors of the Federal Reserve System, Washington, D.C., May.

Podkul, C. (2011) "Bernanke Rejects Alternatives to Raising the U.S. Debt Ceiling", *Washington Post*, July 15, p. A.13.

Romer, C. D., and D. H. Romer (2008) "The FOMC versus the Staff: Where Can Monetary Policymakers Add Value?", *American Economic Review*, 98, 2, 230–235.

Sinclair, T. M., F. Joutz, and H. O. Stekler (2010) "Can the Fed Predict the State of the Economy?", *Economics Letters*, 108, 1, 28–32.

Sinclair, T. M., H. O. Stekler, and W. Carnow (2012) "A New Approach for Evaluating Economic Forecasts", *Economics Bulletin*, 32, 3, 2332–2342.

Stekler, H. O. (1967) "The Federal Budget as a Short-Term Forecasting Tool", *Journal of Business*, 40, 3, 280–285.

Stekler, H. O. (1972) "An Analysis of Turning Point Forecasts", *American Economic Review*, 62, 4, 724–729.

Stekler, H. O. (2002) "The Rationality and Efficiency of Individuals' Forecasts", Chapter 10 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 222–240.

The Economist (2010) "America's Budget Deficit: Speak Softly and Carry a Big Chainsaw", *The Economist*, November 20, leader article.

Tsuchiya, Y. (2013) "Are Government and IMF Forecasts Useful? An Application of a New Market-timing Test", *Economics Letters*, 118, 1, 118–120.

Vere-Jones, D. (1995) "Forecasting Earthquakes and Earthquake Risk", *International Journal of Forecasting*, 11, 4, 503–538.

White, H. (1990) "A Consistent Model Selection Procedure Based on m-testing", Chapter 16 in C. W. J. Granger (ed.) *Modelling Economic Series: Readings in Econometric Methodology*, Oxford University Press, Oxford, 369–383.