**Technical Paper Series**
**Congressional Budget Office**
**Washington, DC**

# ESTIMATING AND FORECASTING CAPITAL GAINS WITH QUARTERLY MODELS

Jangryoul Kim
Korea Development Institute
Formerly: Federal Reserve Bank of Minneapolis

Preston Miller
Federal Reserve Bank of Minneapolis
preston.miller@mpls.frb.org

Larry Ozanne
Congressional Budget Office
larry.ozanne@cbo.gov

September 2004
2004-14

**Abstract**

At the end of each year, the Congressional Budget Office (CBO) estimates capital gains for the year ending and forecasts them for the next decade. The decade forecast is made using CBO's forecast of GDP and an assumption that gains revert from their current size to their historical size relative to GDP. Our objective in this paper is to describe methods to improve CBO's forecasts, particularly for the first year ahead.

We settled on two procedures. The first is similar to CBO's method for forecasting gains. It uses an equation to forecast gains given forecasts of economic and financial variables. This procedure requires a prior step to forecast the economic and financial variables. The second procedure integrates the forecasting of gains and other variables into a single model. In this model we found it advantageous to work with quarterly data, so we interpolate the reported annual series on capital gains to a quarterly frequency. Forecasting in the prior step of the two-step method and the integrated quarterly method was based on Bayesian-restricted vector autoregressions.

Both of the procedures abstract from the effects of tax changes on forecasts of realizations. CBO's baseline is required to assume that current law continues. We abstract from tax changes by constructing a series of capital gains realizations that assumes taxes remained at their 1998 level throughout the 1948-2000 period used for model development. This tax-adjusted series retains much of the volatility in the growth rate of actual capital gains. Between 1971 and 2000, the period used to test the models, the annual growth rate of tax-adjusted gains ranged from a high of 44 percent to a low of -18 percent. Its mean growth rate was 12 percent with a standard deviation of 16 percent.

We base our model comparisons on their root mean squared errors (RMSE) in 1-year-ahead out-of-sample forecasts of the growth rate of tax-adjusted gains. Our application of CBO's mean reversion method found a RMSE of 18.7 percentage points. The two-step forecasting method reduced the RMSE to 14.8 percentage points, and the integrated quarterly method reduced the error to 11.9 percentage points.

Two additional findings from this investigation suggest improvements to CBO's methods. First, the models we developed may help CBO improve its estimates of gains in the year ending. Second, the models may provide some help in forecasting a second year ahead, but after that, either mean reversion or a simple random walk model with drift appears to be as good or better.

**Table of Contents**

## 1. Introduction[1]

During the last quarter of each year, the Congressional Budget Office (CBO) forecasts revenue and spending for the fiscal year that began at the start of the quarter, and for the next ten fiscal years. Usually the forecasts are published in January or February of the next year in CBO's *Budget and Economic Outlook*. The forecasts are made assuming current law is unchanged to provide a baseline against which the U.S. Congress considers proposals for change.

A large fraction of the revenue collected in the fiscal year beginning at the time of the forecast comes from individual income tax liabilities for the calendar year that is ending. As a result, the forecast of revenue for the current fiscal year builds on an estimate of income tax liability for the calendar year ending. Actual information on that liability is incomplete, however, because the year is not quite complete and people have until the next year to file their tax returns, make final payments, or request refunds. As a result, CBO must estimate income tax liabilities for the year ending as well as forecasting liabilities for the coming decade.

CBO projects income tax liability for the population by projecting the major determinants of tax liability, including aggregate net capital gains realized by the sale of many types of assets.[2] Thus CBO must estimate gains for the calendar year ending (referred to here as the current year) and forecast gains for the coming 10 years. When the gains projections are finalized, usually in early December, CBO has preliminary

[1] The authors thank the Tax Analysis Division of CBO for their comments and encouragement, and Christopher Sims for helpful guidance. John McMurray suggested numerous improvements in the text.

[2] Net losses are also projected, but because these are subject to a limit of $3,000 on most returns, losses are smaller and grow relatively steadily from year to year. As a result, they are forecast separately from gains and are not considered in the current paper.

information on gains for the previous year and most macroeconomic and financial data for the year ending.

We use the term "estimation" to refer to the process of generating a figure for realized gains in the current year because the explanatory variables are largely known at the time. Currently, CBO generates that figure by using macroeconomic, stock market, and tax rate variables for the current year in regression equations estimated through the previous year. We use the term "forecasting" to refer to the process of generating figures for the year about to begin and the succeeding 10 years because all variables must be generated. Currently, the CBO generates realizations for future years by using its official forecast of GDP and an assumption that gains revert to their historical size relative to GDP. In this paper, we seek to improve forecasting of capital gains realizations by the CBO. Our work builds on prior work by Miller and Ozanne on estimating capital gains.[3]

A question naturally arises: why separate estimation and forecasting? Why not use a single model to forecast and get estimates from it by conditioning on the observed values in the current year? The answer is that the variables and relationships that yield the best estimates can be quite different from those that yield the best forecasts. The case of stock prices illustrates this point. There appears to be a contemporaneous relationship over the year between growth in stock prices and realizations.[4] Stock prices for the current year are largely observed by the time that the CBO does its budget forecast, so they can be used to estimate current-year realizations. However, to use stock prices to forecast, they, too, must be forecast. Since the best forecasting model of stock prices is essentially a random walk with drift, they should be of little use in forecasting

[3] Preston Miller and Larry Ozanne, *Forecasting Capital Gains Realizations*, Technical Paper 2000-5, Congressional Budget Office, August 2000.

realizations.[5] The difficulty of forecasting stock prices led CBO to separate estimation from forecasting, a reasonable procedure. It seems that it is simpler to construct separate models for estimation and forecasting rather than trying to integrate separate informational structures into a single model.

In the text that follows, we first describe the current CBO forecast method. We then describe the evolution of our methodology with respect to current year estimation, 1-year-ahead forecasting, and multi-year-ahead forecasting. In addition to discussing some things that seemed to work, we discuss some that did not. Our intent is to provide as much aid as we can to other researchers so that they can improve upon our method.

## 2. Current CBO Forecasting Method

CBO began predicting annual capital gains realizations at the end of 1986 for use in its February 1987 baseline. By the January 1989 baseline, CBO began using a single regression equation to estimate gains for the year ending and forecast gains for the next five years. That equation explained realizations in terms of the value of corporate equity held by households, GNP, interest rates, and a capital gains tax rate. Preliminary values of these explanatory variables were available to estimate gains for the year ending, and forecasts of them by CBO's Macroeconomic Analysis Division were used in the same equation to forecast realizations for each of the next 5 years.

Problems with the predictions became apparent when tax return data revealed that realizations dropped for three successive years between 1989 and 1991. The equations

[4] Miller and Ozanne, *Forecasting Capital Gains Realizations*, pp. 11-12.

[5] We found support for both clauses in this statement. The coefficients in a monthly S&P 500 autoregression approximate the values of a continuous-time random walk process sampled monthly. Although we found that stock prices were not useful in forecasting realizations when examining bivariate relationships, they surprisingly seemed to help marginally in some multivariate quarterly models.

neither forecast the drop nor, once actual data became available, explained it. These errors led CBO to revise its methodology for the January 1992 baseline. New equations were developed to estimate realizations for the year just ending. Forecasting for future years was separated from the estimating equations so that forecasts of stock values would not be necessary and forecasts of realizations would be tied more closely to forecasts of general economic growth.

The forecasting method built on the observation that realizations had grown at about the same average rate as GNP from the mid-1950s to the end of the 1980s, but that their annual growth rate frequently deviated sharply from that of GNP. As a result, the ratio of gains to GNP over the period showed little trend away from its average value of 2.8 percent, but frequently jumped above or fell below that level.

The revised CBO forecasting method incorporated this pattern by assuming that whatever level of gains was estimated for the year ending, gains in future years would trend back to their average size relative to GNP. If the level of realizations estimated for the year ending happened to equal the historical average relative to GNP, then in the next 5 years, gains were forecast to grow at the same rate as GNP. If the estimated value were below that target, then gains would be forecast to grow faster than GNP to restore the average ratio by the last year of the five year forecast period.

Over time, the method became more sophisticated by incorporating an estimated rate of reversion. The rate at which the ratio of gains to GNP reverted to its historical average was estimated by including an error-correction term in the estimating equation. That term's coefficient gave an estimate of how much the gap between last year's ratio and the long-run average was closed in the current year. Depending on the specification and years covered through the mid-1990s, the rate was estimated to be between 20 percent and 30 percent. The reversion rate was used to extend the forecast to 10 years when the entire baseline was extended that far. In addition, the historical average ratio was adjusted to reflect legislated tax rates for the forecast period. When tax rates were below average, the target ratio was above average. Finally, in the mid-1990s when the National Income and Product Accounts gave greater prominence to GDP than GNP, domestic product became the long-run guide for the gains forecast.

The method assumes no change in tax rates. Adjustments were made in three baselines when tax changes seemed imminent or had just happened. The congressional campaign and election in 1994 increased the likelihood that legislation reducing the capital gains tax rate would be passed in 1995. CBO reasoned that taxpayers might defer asset sales in the last two months of 1994 until 1995 in hopes of paying a lower tax rate on gains. Consequently, CBO shifted some of the capital gains it estimated would normally occur in 1994 to 1995. By early 1996, the fate a legislation reducing the capital gains tax rate was still unclear, so in its May 1996 baseline CBO again assumed that taxpayers would defer some asset sales from 1995 to 1996. Another adjustment was made in the January 1998 baseline to reflect the temporary surge in realizations following the capital gains tax reductions of 1997. The January 1997 baseline, however, did not anticipate the tax rate reductions that occurred that year, and therefore its forecast was too low. (The forecast turned out to have been too low for other reasons as well.)

The forecasting method was called into question in the later 1990s by the failure of gains to revert back towards their historical size relative to GDP. Gains surpassed their target size relative to GDP in 1996, and continued growing faster than GDP through

2000. By the end of 2000, the historical pattern showed an upward trend, and suggested that gains had permanently shifted to a new level relative to GDP. The error-correction term in the earlier estimating equations no longer was statistically different from zero, suggesting that no tendency to revert could be identified in the historical data. In the baseline of January 2001, CBO assumed that gains in the year of the baseline would remain at their estimated level for the prior year and then begin reverting in later years.

It now appears that realizations in 2001 and 2002 show substantial reversion, which revives at least temporarily the case for incorporating mean reversion into forecasts of gains. It will be unclear for several years, however, whether realizations will continue to fluctuate around their old size relative to GDP or drift off again.

A related question about the forecasting method has also arisen. Even if mean reversion ultimately reappears, could forecasts for the first year or two be improved by incorporating more information about the economy in the immediately preceding years? The current forecast method starts pulling gains back to their historical size relative to GDP in the first year, but the divergence recently has widened for more than a single year. Gains fell below their target in 1989 and then further below in 1990 and 1991. In addition, as noted above, gains rose further above their target each year from 1997-2000. Perhaps the rest of the economy can provide some guidance about the path of gains in the first year or two ahead.

We address this second question directly. The primary measure of success will be whether new methods could have predicted gains one year ahead more accurately than the methods CBO used. Thus, a starting point is errors in CBO's past forecasts for the year ahead.

The left half of Table 1 shows actual and forecasted growth rates from the year ending to the year ahead for baselines in 1987 to 2000 (Tables appear at the end of the text and before the appendixes). Growth rates are shown instead of dollar levels, because capital gains and their forecast errors tend to grow in dollar terms over time. Recall that the forecasts are for growth from estimated values for the year ending because hard data on gains are available only for the year before that. Over the 14 years, the root mean squared forecast error is 22.3 percentage points. Over the last 9 years, when forecasts were made by mean reversion, the root mean squared error is slightly lower at 22.1 percentage points.

The imprint of mean reversion can be seen in several years. The baselines of January 1992, January 1993, and January 1994 forecast that gains would grow faster than GNP because gains were estimated to be below their historical target in 1991, 1992, and 1993. Baselines in January 1997 through January 2000 forecast that gains would grow less rapidly than GDP because gains were estimated to be above their historical target in each of the prior years. In baselines of 1998-2000, the estimated levels of gains for the year before were so far above their target values that gains were forecast to decline in the year of the baseline even though GDP was growing. As stated earlier, gains actually grew strongly in years 1997-2000, raising doubts about the mean reversion method.

The error in 1997 is exaggerated by the tax reduction enacted and effected during the year. That spurred realizations beyond what could be reflected in the baseline. One estimate is that growth in the absence of a tax change that year would have been 22.6 percent instead of the 39.9 percent that was observed, which would reduce the forecast error to 18 percentage points. Even this lower growth remains at variance with the assumption of mean reversion.

The imprint of mean reversion is obscured in three baseline forecasts. The January 1995 and May 1996 baselines forecast additional gains because taxpayers were assumed to defer gains from the prior year to the forecast year in anticipation of a capital gains tax rate reduction. The decline forecasted in the baseline of January 1998 is larger than mean reversion would predict because taxpayers were assumed to have unlocked a burst of gains immediately after tax rates were reduced in 1997.

The right half of Table 1 shows actual and predicted growth rates from two years before the baseline year to the baseline year. Two years are included because the forecasts on the left half of the table are made from estimated values for the year ending. At the time each forecast is made, a close approximation to actual gains is available only for two years before the year of the baseline. Differences between the actual and predicted growth rates over these two years reflect combined errors in the estimation and forecast methods. The root mean squared errors for these two-year predictions are about double those for the one-year forecasts. The errors in baselines from January 1992 on are slightly lower than in prior years. The error in the January 1997 baseline is overstated by the tax rate reductions in that year.

The inescapable conclusion from Table 1 is that the errors in CBO's forecasts of capital gains have been large. These errors, in turn, have led to large errors in projections of revenue from taxing capital gains. However, to the extent that gains are driven by unforecastable forces, like the stock market, the potential to reduce the forecast errors may be limited. Nevertheless, it seemed to us to be worth trying.

## 3. Initial Steps

Miller and Ozanne found three classes of variables to be useful in estimating current-year capital gains: stock market price and volume variables, macroeconomic variables, and capital gains tax rate variables. Our initial research strategy was to retain the Miller-Ozanne model for estimation and develop a new model for forecasting. Before we began formal modeling, we adjusted the series on capital gains realizations and searched for macroeconomic and financial variables that could help forecast that adjusted series.

### *Tax-Adjusted Capital Gains*

Legislation changing capital gains tax rates can cause large changes in capital gains realizations, as happened, for example, in 1986 and 1997. Yet, legislative changes are difficult to forecast and, more importantly, CBO is legally mandated to assume that current policies remain in place when it does its budget projections. That assumption increases forecast error in years with big tax changes, yet the error isn't the fault of the model. It is due to an error in assumptions about tax policy. Moreover, a model attempting to fit unadjusted capital gains data would mistakenly attribute the effects of tax changes to other variables in the model.

In order to address the problem of forecasting gains in years affected by tax changes, we removed an estimate of the tax effects from the reported series on capital gains. The estimate comes from the Miller-Ozanne estimating equation that contains real GDP, S&P 500, and tax rates. With that equation, we calculated what capital gains would have been in each year had tax rates been at their level between 1998 and 2000. We call this series tax-adjusted capital gains.

For our initial investigation conducted during the end of 2000 and the beginning of 2001, we had data on capital gains through 1999, although the 1999 figure was preliminary. Thus our initial series on tax-adjusted gains extended through 1999 and was estimated with data on stock prices and GDP available around the end of 2000 and the beginning of 2001. By the time we completed our formal forecasting models a year later, we had a final figure for capital gains in 1999 and a preliminary figure for 2000, and other data as reported early in 2002. The equation used to construct tax-adjusted gains through 2000 appears in Table 2.

Tax-adjusted capital gains are compared to reported capital gains for the years 1948-2000 in Figure 1. Adjusted gains are above reported gains in most years because tax rates in 1998-2000 were lower than in most years. The main exception is between 1982 and 1986 when the top tax rate was slightly lower than in 1998. The largest single adjustment is in 1986. The spike in reported gains that year was caused by taxpayers realizing many additional gains to beat the tax rate increase legislated to begin in 1987. Adjusted gains iron out that spike.

The formal models developed below will attempt to forecast the annual growth rates of tax-adjusted gains between 1971 and 2000. In judging their success, it will be helpful to keep in mind that over that period tax-adjusted gains grew at an average annual rate of 12.2 percent with a standard deviation of 16.4 percent. The largest decline was 18.5 percent and the largest increase was 44 percent.

### *Potential Explanatory Variables*

Our next step in late 2000 was to assemble a set of variables that could help to forecast gains. We would then incorporate these variables into a forecasting model. We reasoned that the greatest chance for success was at the one-year horizon. If the dependent variable were annual gains in year t+1 and the explanatory variables were quarterly values in the 4th quarter of year t and earlier, we could check out the usefulness of various series in forecasting gains without actually constructing a model. We could just deal with single equations having (growth of) annual gains on the left-hand side and a constant and predetermined variables on the right-hand side.[6] We initially examined whether adding lag distributions of alternative variables would improve the fit of an equation with only a constant on the right-hand-side (a random-walk specification). We experimented with adding more than one variable at a time and with restricting the lag distributions. The results from all of these specifications were discouraging: none did appreciably better than the random walk specification.

We reasoned that if gains were related contemporaneously to other variables, then forecasts of gains would be related to forecasts of other variables. It could be that the failure of our first exercise reflects an inability of the univariate lag distributions we specified to imply adequate forecasts of the variable being examined. We sought to separate the ability of a variable to explain gains from the ability of that variable itself to be forecast.

We attempted to gather information on the usefulness to gains forecasts of macroeconomic and financial variables by determining their contribution assuming that they, themselves, could be forecast without error. If a variable were useful in forecasting under the perfect foresight assumption, we could later direct our efforts to forecasting that variable as accurately as possible. However, if it were not useful, there would be no reason to proceed further with it. We realized that the first part of our inquiry can be considered estimation of tax-adjusted capital gains based on observed values of macroeconomic and (other than stock market) financial variables. As a result, knowledge gained from this first part had the potential to lead to improvements in estimation as carried out in Miller-Ozanne.

The results from this exercise were surprising. Although realizations is a nominal, or current-dollar, series, it is best estimated with real, or constant-dollar, macroeconomic series. Current financial variables or prices were not useful. Whenever a nominal series helped to estimate gains, it was dominated by its real counterpart.[7] The three series among those we examined that contributed most to fit were real GDP, private nonfarm output-per-hour, and the ratio of real consumer durable expenditures to real personal disposable income. Moreover, the variables contributed the most when they entered jointly rather than individually. Although it is purely conjecture, we believe the first two variables have explanatory power for gains because they indicate something about the state of the economy and the profitability of firms. We believe the third variable has power because there is a connection, possibly going in both directions, between sales of financial assets and purchases of big, discretionary items. For instance, when people decide to buy a luxury car, they may pay for it in part by selling stock. Or, when people realize a big profit in the stock market, they may use some of it to purchase

---

[6] As described in Miller-Ozanne, the data seem to prefer a growth-rate specification for gains.

[7] An exception is that the growth in nominal stock prices explains the growth of gains better than the growth of real stock prices. We ignored stock prices at this stage of our analysis, however, because they are so difficult to forecast.

a luxury car. In either case, the third variable would change as the numerator is affected but not the denominator.

## 4. Two-Step Forecasting Approach

Since our point of departure was the Miller-Ozanne estimation model, a two-step approach to forecasting gains seemed logical. In estimating gains, current-year gains are regressed on actual values of explanatory variables. In a two-step approach to forecasting, next-year's gains would be regressed on forecast values of explanatory variables. The explanatory variables for forecasting would be from one of all possible subsets of the variables found useful in our preliminary exercise: real GDP, labor productivity, and the ratio of consumer durables to disposable income.

Our next task was to forecast the three explanatory variables to determine the best combination to forecast adjusted capital gains. Since all three potential explanatory variables are available quarterly, we judged that they could be best forecast using a quarterly model. We still had two major decisions to make:

- what form the model should take, and
- which variables in addition to the explanatory variables should be included to help forecast.

Since our approach is not explicitly based on economic theory, the nonstructural forecasting methodology of vector autoregressions seemed like a natural choice for modeling strategy. And because we contemplated including as many as six variables (three explanatory variables and three auxiliary variables useful in forecasting the explanatory variables), improved forecasting accuracy could be expected by using

Bayesian-restricted vector autoregressions (BVARs). See the procedures in Sims and Zha and in Robertson and Tallman.[8]

The tightness of prior restrictions in a BVAR can be varied. While the data and the restrictions determine the estimated coefficients in a BVAR, the tightness of the prior determines the relative weights given to each determinant. A loose prior gives more weight to the data; a tight prior gives more weight to the restrictions. One set of restrictions essentially maintains that each variable in the BVAR can be modeled as a univariate random walk with drift. Another restriction takes the form of a dummy observation that implies the variables in the model are cointegrated. As a result, the BVAR retains a form of error correction, like the CBO forecasting model, but it is implemented in a more flexible way.

The set of auxiliary variables we considered to aid in forecasting the three explanatory variables share two common properties. First, they are available on a quarterly basis back to 1948, which allows a significant period to estimate a model for annual capital gains and evaluate its out-of-sample forecasts. Second, they are related to variables that other researchers have found useful in forecasting business cycle data. Our method to select auxiliary variables was to specify four classes of variables and choose at most one variable from each class. The variables within each class are shown in Table 3 along with the explanatory variables.[9]

---

[8] Christopher A. Sims and Tao Zha, "Bayesian Methods for Dynamic Multivariate Models" *International Economic Review*, (1998) Vol. 39 Issue 4, pp. 949-968. John C. Robertson and Ellis W. Tallman, "Vector Autoregressions: Forecasting and Reality" *Federal Reserve Bank of Atlanta Economic Review* (First Quarter, 1999): Vol. 4, Issue 18.

[9] In all the models we describe in the following text, level variables enter as logarithms, while rate and ratio variables enter without transformation.

Only the interest-rate spread is included in the first auxiliary class. It is included because researchers have found it useful in forecasting business cycle turning points.[10] The main reason for specifying the Moody's seasoned long-term rate in the spread is that it goes back to 1948.

The second auxiliary class contains either a real wage or a labor share variable. Our thinking was that such a variable would close a supply-side subsector of the model composed of output, productivity, and wages.

The third auxiliary class contains a mix of variables, all thought to provide leading indicators of demand. It is common to find either investment variables or crude price variables in macro BVARs.[11]

We originally planned to:

a. use the Miller-Ozanne equations to estimate gains,
b. estimate a BVAR incorporating explanatory and auxiliary variables, and
c. use the forecasts of explanatory variables from that BVAR in a regression with gains as the dependent variable.

As work proceeded, we made two changes to this plan. The first was to estimate gains with our own equations rather than the Miller-Ozanne equations. The reason was that the dependent variable in our equations is tax-adjusted gains, while that in the latter is unadjusted gains. That change caused us to add the stock market as a fourth auxiliary

---

[10] See, for example, Dan M. Chin, John Geweke, and Preston J. Miller, "Predicting turning points" Federal Reserve Bank of Minneapolis Staff Report Number 267, (June 2000). Arthuro Estrella and Frederic S. Mishkin, 1998. Predicting U.S. recessions: Financial variables as leading indicators. *Review of Economics and Statistics* (1998) 80, no. 1, pp. 45–61. James H. Stock, and Mark W. Watson, "New Indexes of Coincident and Leading Indicators" In *National Bureau of Economic Research Macroeconomic Annual 4*, edited by Olivier Blanchard and Stanley Fisher. Cambridge, Mass.: National Bureau of Economic Research. (1989).

[11] See, for example, Robert B. Litterman, "A Bayesian Procedure for Forecasting with Vector Autoregressions" Massachusetts Institute of Technology Department of Economics Working Paper (1980);

class of variables. Since stock market volume is not available on a quarterly basis back to 1948, only the S&P 500 was included in the seventh class. The second change was to remove any distinction among our classes of variables. We made this change because we found estimates and forecasts of capital gains sometimes could be improved by adding variables from classes other than the explanatory variables. Thus, one variable from any class could be included in the estimation and forecasting equations for gains.

Next we describe our revised methodology in greater detail and then present evidence on the approach's success. We distinguish in our description between estimation and forecasting. Experimentation confirmed our prior that it is better to use separate models to estimate and to forecast than to use a single model to do both. The models we chose are identified by the variables they contain and by the hyperparameter values used in the BVAR restrictions. The identifying features associated with the best-performing models are reported in Appendix A, along with background information on the hyperparameters.

***Methodology***

Presentation of the methodology and discussion of the empirical results is facilitated by introducing some mathematical notation and our error measures. The notation is first used to develop our estimation methodology and then our forecasting methodology.

***Notation and Measures of Success.*** For any variable Z, we use the subscript t to denote year, so that $Z_t$ is the annual value of Z in year t. We use the subscript t:i to denote the ith

---

Robertson and Tallman, and Sims and Zha.

quarter of year t, i= 1,2,3,4, so that $Z_{t:i}$ is the value of Z in the ith quarter of year t. Throughout the text we denote the value of tax-adjusted capital gains in a year by CG and the log of CG simply by Y. We often refer to CG as "gains." While capital gains are reported only annually, all the other variables used in our models are reported at least quarterly. We refer to the other variables as the vector X.

Our data include annual values of CG and quarterly values of X from 1948-2000. These data are as reported in early 2002. At that time, we had preliminary information on gains in 2000. Tax-adjusted gains equaled actual gains in 2000 because tax rates on capital gains between 1998 and 2000 were the base rates for the series on tax-adjusted gains.

Much of what we do is guided by the CBO's process for annual revenue projections. The CBO finalizes its estimates and forecasts of capital gains in the late fall of the year. So, if s is the current year, we suppose the information at hand consists of the history of CG through year s-1, denoted by $[CG]_{s-1}$ and the history of X through the 4th quarter of s, denoted by $[X]_{s:4}$ .[12] Consequently, for any of the approaches we develop, we estimate $(CG_s)^e$ based on the information $[CG]_{s-1}$ and $[X]_{s:4}$. We forecast $(CG_{s+j})^f$ where j =1….10, based on the same information and conditional on $(CG_s)^e$.

For the most part, the focus is on 1-year-out forecasts, and all of our estimation and forecasting exercises are done out-of-sample. That is, at each date t, the models' coefficients are estimated on only $[CG]_{t-1}$ and $[X]_{t-1:4}$. The objectives and performance measures are stated in terms of 3 error measures:

[12] Our information assumption deviates from the situation faced by the CBO in 2 ways. First, CBO had only partial data for the 4th quarter of each year, while we assume full information for the quarter. Second, our data set is as it existed early in 2002. When the CBO made its forecasts, it had to use data that existed at the time and which subsequently have been revised, perhaps, substantially.

1. Estimation error, $E1_t = CG_t^e/CG_{t-1} - CG_t/CG_{t-1}$ ,

2. (Forecast | Estimation) error, $E2_t = CG_{t+1}^f/CG_t^e - CG_{t+1}/CG_t$ , and

3. (Forecast | Actual) error, $E3_t = CG_{t+1}^f/CG_{t-1} - CG_{t+1}/CG_{t-1}$ .

For any method, we first estimate capital gains to minimize the out-of-sample root mean squared error (RMSE) in terms of E1. Then, given that estimate of capital gains, we forecast 1-year-ahead capital gains to minimize the out-of-sample RMSE in terms of E2. We also report overall forecasting performance in terms of the RMSE of the 2-year error measure E3.

***Methodology for Estimation***. In this part of the methodology, we examine the performance of estimation equations incorporating different sets of variables. The objective is to find the set of variables that minimizes the RMSE in terms of the estimation error E1. We begin by selecting a set of variables, with at most 1 from each of the 7 classes, to form the vector X. We next estimate by OLS over the period 1949-1969 the relationship:

$$(1) \qquad \Delta Y_t = b_0 + b_1(X_t - X_{t-1}) + \varepsilon_t ,$$

where the variables in X are annualized and where level variables are logged and rate or ratio variables are untransformed. Then using the estimated values of $b_0$ and $b_1$ and the actual $X_{1970} – X_{1969}$, the estimate of $\Delta Y_{1970}$ is calculated as $(\Delta Y_{1970})^e = b_0 + b_1(X_{1970} – X_{1969})$. This in turn yields the estimate of log(capital gains) in 1970 of $(Y_{1970})^e = Y_{1969} + (\Delta Y_{1970})^e$, and finally $(CG_{1970})^e = \exp[(Y_{1970})^e]$. We add a year to the estimation period and proceed this way year-by-year until $(Y_{2000})^e$ is obtained. We then compute the E1 errors for each year and compute the RMSE in terms of these errors. We choose a different set of variables to be included in X and repeat this process until all possible subsets have been examined. We choose as best the one that has the minimum RMSE.

Unfortunately, there is a complication due to an apparent break in the statistical process for gains in the 1990s. Specifications that perform best, in terms of either estimation or forecasting, over the whole period 1971-2000 are found to be different from those that perform best in the 1990s. As time passes, the question will have to be confronted of whether the process for gains has permanently changed to that of the 1990s or whether it will return to the way it was earlier. One way of dealing with this problem is to carry along estimates/forecasts of models best over each time period and put weights on them according to their relative errors, i.e., a likelihood calculation. Because we do not attempt to answer the question, we carry along 2 models—one best for the whole period 1971-2000 and one best for 1991-2000.

***Methodology for Forecasting.*** In this part of the methodology, we search anew for a set of variables that leads to the most accurate 1-year-out forecasts of gains given the current years' estimates obtained above. The set of variables found to be best can differ from that selected in the estimation part. In the first step of the forecasting methodology, each set of variables is incorporated into a quarterly BVAR to generate forecasts over the next year. In the second step, these forecasts are then included in a forecasting equation for gains. The best set of variables for forecasting is the one that leads to the minimum RMSE in terms of out-of-sample E2 errors.

As before, we begin this part by selecting a set of variables, with at most 1 variable from each of the 7 classes, to form the vector X, and a vector of hyperparameter values. Given X and the hyperparameter values, we estimate a quarterly BVAR with 5 lags from 1949:2 through 1959:4 and use that estimated model to dynamically forecast $(X_{1960:1})^f$, $(X_{1960:2})^f$, $(X_{1960:3})^f$, and $(X_{1960:4})^f$. After appropriately converting the 4 quarterly forecasts to a forecast of $(X_{1960})^f$, we estimate the model through 1960:4 and follow the same procedure to generate a forecast of $(X_{1961})^f$. We continue this process iteratively until we have forecasts of $(X_t)^f$ conditional on $[X]_{t-1:4}$ for t = 1960,….,2000.

The forecasts of the X variables are used to forecast gains by estimating an equation that relates the growth of capital gains to the forecast changes of X variables:

$$(2) \qquad Y_t - Y_{t-1} = c_0 + c_1[(X_t)^f - X_{t-1}] + \mu_t.$$

The regression (2) is estimated by OLS initially over the eleven years 1960-1970, based on information available at the end of 1970. Since $Y_{1970}$ is not available based on the information assumption, it is replaced by its estimate determined in the previous step. Then, based on the estimated coefficients, $c_0$ and $c_1$, next year's forecasts of X, $(X_{1971})^f$, and the current year's estimate of capital gains, $(Y_{1970})^e$, the relationship is used to forecast gains in 1971. We then add a year to the estimation period for (2) and proceed as before to get a forecast of gains for the 1972 (again substituting the estimated gain for the unobserved actual in 1971). The iteration continues until equation (2) is estimated through 1999 and the last forecast is for year 2000. With these thirty forecasts of gains, we calculate year-by-year E2 errors and the associated RMSE. Now, with the variables in X still fixed, the hyperparameter values are varied and the process repeated to find the hyperparameters that minimize the RMSE in terms of E2 errors. In practice, for each specification of X, we do a coarse grid search around standard hyperparameter values to improve forecasting results. Finally, the whole process is repeated for each set of variables in X to get the best model: the best set of X variables, and given those, the best set of hyperparameter values.

***Results for Two-Step Approach***

In Table 4, we compare the out-of-sample, one-year-ahead estimation and forecast RMSEs for adjusted capital gains of five alternative models. All of the models are conditional on $[CG]_{t-1}$ and $[X]_{t:4}$ . The last two models listed are the winners in the above process of selecting two-step models. TS-all is the combination of the model that estimated best and the model that forecast best for the entire 1971-2000 period. The two components of TS-all differ in variables included and hyperparameters, as noted above, but otherwise employ the same methodology. TS-90 is the corresponding combination of models that estimated and forecast best for the 1991-2000 period.

The three models listed on the top of Table 4 provide standards against which the errors of the two-step method can be judged. The first of these, CBO, replicates the CBO method. Gains are estimated for year t using a simplified version of a CBO equation, and gains for t+1 are forecast using the mean reversion methodology.[13] The second of these, RW (random walk with drift) assumes that gains grow at a constant rate and that rate is the average of the historical rates known at the time of the estimation and forecast. The third and final comparison model, AR (autoregression), regresses the log of gains on a constant and an unconstrained 1-year lag of gains. It uses the estimated coefficients to estimate and forecast gains.

The RMSEs are shown for the E1 estimation errors, the E2 current-year estimate-to-forecast, errors, and the E3 previous-year actual-to-forecast, errors. The average RMSEs are shown for each decade running say, from 1971-1980 and for the entire period, 1971-2000.

Each two-step model estimates gains more accurately than its standards of comparison. That is, given actual values for exogenous variables in a year and gains in the prior year, the two-step models estimate gains with lower RMSEs. The simplified CBO equation comes in second place in the full period and in the 1990s, with RMSEs that are 30 percent to 35 percent higher. The better estimates of the two-step models may come from their having been developed to explain tax-adjusted gains while the CBO models were developed to explain actual capital gains. Changes in tax rates account for substantial changes in actual capital gains. The narrower focus of the two-step equations may have identified additional explanatory variables that will be helpful in explaining

[13] The simplified CBO equation is the equation used to construct tax-adjusted capital gains, see Table 2, with the simplification that the two variables measuring tax rates are eliminated. The mean reversion forecast follows the current CBO procedure described earlier, but is applied retroactively to years from 1971 forward. For full details, see memo from Larry Ozanne, revised December 12, 2002.

actual capital gains. The leading candidates are productivity, the interest rate spread, real compensation per hour, and real fixed private domestic investment.

The two-step approach also generally forecasts better than the standards of comparison. Over the full period and in terms of E2 errors, TS-all forecasts have the lowest RMSE at 14.80 percentage points; the random-walk forecasts have the second lowest RMSE at 16.78 percentage points. The AR(1) forecasts, closely followed by the CBO mean reversion forecasts, have larger RMSEs of 18.31 percentage points and 18.57 percentage points respectively. The TS-all forecast errors are 20 percent lower than the mean reversion errors. The advantage of the TS-all model occurs primarily in the 1971-1980 and 1991-2000 decades. In the 1981-1990 decade, however, TS-all has the highest RMSE, although it is not much worse than the errors of the random-walk and autoregressive models.

When the focus is narrowed to forecasting the decade 1991-2000, the TS-90 has the lowest RMSE for E2 errors at 12.55 percentage points. TS-all is second, and the three standards of comparison follow in the same relative ranking as for the full period. The poor performance of mean reversion forecasts in the 1990s is not surprising given the surge in gains in the later half of the decade. In fact, it was the large forecast errors CBO made in the 1990s, documented in Table 1, that led us to search for better forecasts.[14]

The models with the lowest E2 error also have the lowest E3 error. That means that models that best forecast the growth rate of gains conditional on their estimate of

[14] The RMSE from applying the CBO method to tax adjusted gains between 1992 and 2000 is only slightly lower than the RMSE CBO had in forecasting actual gains over the same period with its variants of mean reversion. Thus the primary source of error during those years was from economic and behavioral factors, not from unanticipated tax changes.

gains in the prior year also get the closest to the actual value of gains in the year ahead. That outcome is not necessary, as we will see below.

Neither two-step model forecasts better than the other over all decades. TS-90 does considerably better than TS-all in the 90s, but it does much worse in the 70s and 80s.

The lower forecasting error between 1971 and 2000 of the random walk than the mean reversion model suggests that forecasting gains to grow at their average rate as known up to that year would have been superior to using the mean reversion model.

## 5. Integrated Quarterly Approach

Although the two-step approach met with some success, we reasoned that it had some shortcomings. One of them relates to how it deals with mixed-frequency data. With realizations available annually and all other macroeconomic and financial variables available at least quarterly, any forecasting model has to deal with a mixed-frequency data problem. The two most common procedures to address this problem are either to aggregate the high-frequency data to the lowest frequency of any variable or to interpolate the low-frequency data to the highest frequency of any variable. Our two-step approach followed the former procedure by aggregating up all quarterly data to produce annual growth rates or changes. A shortcoming with this procedure is that it can cause a loss in information. For instance, had real GDP shown no growth over the year, it could matter for gains whether GDP was flat over the whole year, had fallen in the first half and risen in the second, or risen in the first half and fallen in the second. The two-step approach makes no distinction among these alternatives.

Another shortcoming with our two-step approach is the way it chews up degrees of freedom. Many observations are required in the first step to estimate the BVAR before it can be used to reliably generate forecasts. Then further observations are required in the second step to reliably estimate the regression of gains on forecast values of the explanatory variables.

The second approach with interpolation of gains circumvents the shortcomings of our two-step approach. However, it introduces a new one. The interpolated values of gains must differ from the true, unreported quarterly gains. Consequently, interpolation introduces some measurement error. Since both the two-step approach and second approach using interpolation face conceptual problems, the determination of which works better in practice is an empirical issue.

Our research turned to developing the second approach to estimating and forecasting gains. The findings are clear: interpolating gains and incorporating them into a unified model outperforms the two-step approach. We now describe that research.

In order to estimate a single model that contains macroeconomic and financial data as well as gains, we need to construct a quarterly series of gains. We call the constructed series interpolated gains and require that the average of the four quarterly values (at annual rates) equals the reported annual value. Since there obviously are many reasonable ways to interpolate gains, we require a way to choose among the alternatives. We consider three alternatives and choose among them by which leads to the most accurate estimates and forecasts of capital gains in the current and next year, respectively. As in the previous section, the form of model used for the unified approach is taken to be a BVAR, and once again, we allow for using different models to estimate and forecast gains.

With two of the alternatives for interpolating gains, we translate annual gains into quarterly gains once-and-for-all before we do any modeling for estimation or forecasting. As a result, the quarterly gains constructed from these alternatives are external to our models and are incorporated into them in the same way as any other quarterly series. With the third alternative, quarterly gains are constructed "as you go" and are internal to our model. Nevertheless, for all three alternative interpolation schemes, the objectives are: first to minimize RMSEs in terms of E1 errors conditional on the information sets $[CG]_{t-1:4}$ and $[X]_{t:4}$, and second, to minimize the RMSEs in terms of E2 errors conditional on the same information plus $\{(CG_{t:1})^e, (CG_{t:2})^e, (CG_{t:3})^e, (CG_{t:4})^e\}$.

***Three Interpolation Methods for Integrated Approach***

The first interpolation alternative is linear. It is constructed by assuming that gains grow linearly from the fourth quarter of one year through the fourth quarter of the next year. The slope is set so that the average of the four quarters of the next year equals the actual annual amount of gains for that year.[15] For example, if gains in the 4th quarter of one year are \$100 and they total \$162.50 for the next year, the path of quarterly gains in the next year is \$125, \$150, \$175, and \$200. This method requires an alternative way of interpolating gains in the first year, 1948, because no value is available for the fourth quarter of the prior year. The alternative we use is to assume that realizations within 1948 grew at the average annual rate of gains over the 1948-2000 period. It is possible

[15] Quarterly gains are at annual rates so that the average of the 4 quarters of a year equals the annual total.

then to solve for the quarterly levels of gains consistent with both that growth rate and the 1948 total.[16]

Since linear interpolation forces the changes in gains to be the same in each quarter of a year, these changes cannot reflect changing economic conditions within the year. We reasoned that there could be an advantage to letting quarterly changes in conditions affect quarterly patterns of gains. The next two alternatives were designed with that in mind.

The second interpolation alternative that we consider is dubbed "economic interpolation." Like the first alternative, annual gains are allocated at the outset to individual quarters in each year, and a complete quarterly series is constructed that can be included with all other macroeconomic and financial variables in our data set.

To construct economically interpolated gains, we use the same Miller-Ozanne equation as was used to remove tax effects from the unadjusted capital gains series (see Table 2). The equation is used to calculate a growth rate of gains from one quarter to the corresponding quarter in the next calendar year. That growth depends on changes in the business cycle and stock prices over eight quarters. Repeated application of the equation provides an estimate of gains growth from each quarter to the same quarter in the preceding and the following calendar year. Then, by somewhat arbitrarily allocating actual gains for one year among four quarters, the calculated growth rates are used to

[16] We investigated, in the context of univariate models, the difference it made to the forecasting of gains of using linearly interpolated quarterly gains instead of annual gains. This investigation allowed us to determine how much of the difference in forecasting success of the two-step and integrated approaches stems solely from the difference in using the annual and interpolated gains series.

Our findings, which may seem counterintuitive, show that the interpolated series does better. The intuition is that since the interpolated series is just an algebraic manipulation of the original series and contains no additional information, the two series should do equally well. We give an explanation and geometrical interpretation of our results in Appendix B.

estimate gains in all quarters from the first quarter of 1948 to the last quarter of 2000. In a final step, the amount of gains in the four quarters of each year is adjusted proportionately to hit the known total of gains for each year. This alternative has the advantage of letting the quarterly pattern of gains respond to the pattern of macroeconomic and financial variables. (Appendix C describes the procedure in greater detail and identifies key assumptions and limitations.)

The third, and final, interpolation alternative that we consider is dubbed "in-model" interpolation. Unlike the other two alternatives, for which the interpolated series is created once-and-for-all before the BVAR construction begins, this one creates the series on the go as the BVAR is constructed. The idea behind this alternative is to make the interpolation endogenous by attempting to determine the quarterly gains series, consistent with the model's structure, that yields the best estimate of annual gains. The interpolation series using this alternative is constructed in three stages:

a. Initial model estimation with an interpolated "seed",
b. Out-of-sample estimation of annual gains, and
c. Reconciliation of estimated gains with actual annual gains.

Since in-model interpolation is done within an estimated model, the latter must exist before the series construction can begin. We choose 1948-1960 as the initial estimation period. We refer to the initial gains series as the seed, because it is used to get the process started but then is extended with a new series in the period after 1960. The seeds are either the linearly interpolated gains series or the economic interpolated series. The estimation and reconciliation stages to construct the in-model interpolated series are described in the next section as part of the general discussion of estimating gains with all three interpolation methods.

Appendix D compares gains interpolated by each of the three methods to limited data on the timing of actual transactions that generated gains. The comparison finds limited evidence that linear and in-model interpolation can pick up some trend in gains within a year, but no method can match quarter-to-quarter movement of the transactions data. Differences in the types of gains included in of our interpolated variable and the transactions data cloud our findings

***Estimation of Annual Tax-Adjusted Gains***

In this section we seek to minimize RMSEs in terms of E1 errors, when we estimate $CG_t$ conditional on $[CG]_{t-1:4}$ and $[X]_{t:4}$. For each interpolation alternative, we search for the best model, where a model is identified by a choice of variables and a set of hyperparameter values. Our method is to select a set of variables, and then given that set, to search for the hyperparameters that lead to the smallest out-of-sample estimation errors. Finally, we choose among the different models based on which estimates the most accurately.

In order to estimate annual gains, we have to account for the partial information structure in the current year. We do this for each interpolation method by straightforward application of Kalman filtering. However, our method can be understood as a series of three steps, as described below.

The first step is to estimate the model based on complete quarterly data through the historical period that ends in the fourth quarter of the previous year, i.e.,

$[CG]_{t-1:4}$ and $[X]_{t-1:4}$.

The second step is to generate quarter-by-quarter estimates of gains in the current year based on the information set

$[CG]_{t-1:4}$ and $[X]_{t:q}$ for $q = 1, 2, 3, 4$,

available at the start of each quarter. We do this using a sequence of forecast and revision operations. Based on the data and the estimated coefficients as of the end of the historical period, forecasts are generated for all variables in the first quarter of the current year. Then, those forecasts are revised based on differences between actual and observed values of macroeconomic and financial variables in the quarter. Essentially, this is done by pre-multiplying the one-step-forecast errors for the macroeconomic and financial variables by their covariances with gains, where the covariance matrix is constructed from the estimated residuals (up to the prior on the covariance matrix of the BVAR). Then, the resulting revision term is added to the forecast of gains to get an initial estimate of gains in the first quarter. Next, we forecast all variables in the second quarter based on information through the first quarter, of which the data in the last quarter consist of actual values of macroeconomic and financial variables and the initial estimate of gains. We continue this process of forecasting and revising to get estimates of quarterly gains based on information through the fourth quarter.

The third step involves year-end revisions to generate final quarterly estimates of gains. Conceptually, the initial estimates in the first three quarters of the year do not utilize information available later in the year when the CBO would be estimating current-year gains. After observing their fourth quarter values, therefore, it is possible to improve further on the estimates from the second round of operations. We do this by Kalman smoothing, and take the average of the resulting final quarterly estimates to get the current-year estimate of annual gains. When we move to the next year, the process is repeated with estimated quarterly gains being replaced by interpolated gains, where the latter is based on observed annual gains for the year.

With linear or economic interpolation, there is no ambiguity about what is meant by interpolated gains, but there is with in-model interpolation. With the two former interpolation alternatives, estimated gains for the current year are replaced by the linearly or economically interpolated gains contained in the data file, the model is reestimated, and the estimation process for the next year begins. For these two alternatives, the interpolated gains, which by construction add up to the observed annual gains, are in the data set and are treated as actual data. In contrast, with the in-model alternative, there are no quarterly gains in the data set for the current year. There are only estimates. By the fourth quarter of the next year, however, annual gains for the current year will be known, and the sum of the quarterly estimates will differ from it. Thus, in order to make use of all observations, we need to reconcile the best estimates of quarterly gains with the total for the year. After the estimation of gains in the current year is completed and before the estimation of gains in the next year begins, a reconciliation stage is required with in-model interpolation.

After some experimentation, we found the best way of reconciling ("best" in terms of lowest estimation errors) is to use the structure of the model. In concept, we shock the gains equation by a constant amount each quarter and let the model's dynamics (its impulse response) determine the revised quarterly gains. We require that the constant shock be such that the sum of the revised gains equals the observed annual figure. Since the values of macroeconomic and financial variables are already observed, only gains are affected by this exercise. Hence, we use the lag coefficients in the gains equation to determine the response. Because a prior is that the log of gains follows a random walk, its estimated own lag coefficients will be approximately 1 on the first lag and zeros elsewhere. For this reason, a shock of $\varepsilon$ applied each quarter will approximately produce changes in the log of gains, revised minus estimated, of $\varepsilon$ in the first quarter, $2\varepsilon$ in the second quarter, $3\varepsilon$ in the third, and $4\varepsilon$ in the fourth. Now, based on actual values of macroeconomic and financial variables through the current year and reconciled gains figures over that period, we reestimate the model and proceed to estimate gains in the next year.

In Table 5, we compare the estimation errors from each of our best models using the different methods. The models are the two-step annual model and the 4 quarterly interpolated models: linear, economic, in-model with linear seed, and in-model with economic seed. The three comparison models included in Table 4 are omitted because they generally did less well than the included two-step models. The RMSEs of the three comparison models shown in Table 4 can be compared to the RMSEs in Table 5. In panel A, we include the models that were best over the entire 1971-2000 period. In panel B, we include those that were best over the 1990s.

All five models have similar estimation errors when fit over the 1971-2000 period. The linear interpolation model is best with a RMSE of 0.1025, and the model with in-model interpolation from a linear seed is worst with a RMSE of 0.1190. The two-step method is in the middle of the pack. All five models also have similar RMSEs when fit to the 1991-2000 period, but the model with in-model interpolation and a linear seed has moved from worst to best. Again, the two-step method is comparable to the interpolation models.

***Forecasting of Tax-Adjusted Gains One Year Ahead***

In this section, the objective is to minimize RMSEs in terms of E2 errors conditional on the information sets $[CG]_{t-1:4}$, $[X]_{t:4}$ and $\{(CG_{t:1})^e,\ldots,(CG_{t:4})^e\}$. Each forecasting model we examine uses the same gains series as the associated estimation model; i.e. linearly interpolated gains or in-model with economic seed gains, for example. However, the variables and hyperparameter values can vary in the associated models for estimation and forecasting.

The models' coefficients then are estimated using data through t-1:4. The forecasts are conditioned on actual values of macroeconomic and financial variables through t:4 and on estimated values of gains for the quarters of t. The forecasts are generated dynamically, quarter by quarter, for year t+1.

All models have more difficulty forecasting a year ahead than estimating the current year, but some have more difficulty than others. When the models are judged by how well they forecast 1971-2000, the integrated model with simple linear interpolation does best with a RMSE of 11.92 percentage points (see Table 6). The three other models with interpolated gains follow with gradually higher RMSEs. The two-step model has a larger RMSE than all of the interpolation models. The mean reversion model that mimics CBO's method has a RMSE of 18.57 percentage points. Thus, the linear interpolation model reduces the RMSE from that of mean reversion by one-third. Recall from Table 4 that mean reversion forecasts had a lower RMSE than the two-step model for the 1981-1990 decade. All four models with quarterly interpolation beat the mean reversion model in that decade as well as the other decades.

Among models that fit the 1991-2000 years best, the superiority of integrated models over the two-step models is reduced (see Table 7). The model with simple linear interpolation again has the lowest RMSE, but the two-step model beats one of the interpolation models and all models are closer in their RMSEs. One interpretation of these results is that the advantage of the quarterly methodology over the two-step approach is that it conserves on degrees of freedom. That advantage would be expected to dissipate over time as more observations become available.

Although the model with linear interpolation has the lowest RMSE for the E2 error over the 1991-2000 period, it does not have the lowest RMSE for the E3 error. Both models with in-model interpolation have lower E3 RMSEs. The lower E3 error means that those models get closer to the level of gains being forecast than does the model with simple linear interpolation. The model with simple linear interpolation gets closer to the growth rate from the year of estimation to the year of forecast, but either has larger errors in the year of estimation or its errors from estimation and forecasting are more additive while the other models' errors are more offsetting.

## 6. Applications of the Model

After the models were developed, we applied them to forecast tax-adjusted gains in 2001. We also used the integrated models to forecast more than one year ahead. In each case, we compared the model errors to errors of CBO's mean reversion method.

### *Forecasting 2001*

After the above models were developed, information on gains in 2001 became available showing that gains declined by almost 46 percent that year. The small reduction in tax rates on capital gains occurring in that year would have minimal effect, so a similar drop in tax-adjusted gains should be expected. The forecasts of the above models for 2001 provide a true out-of-sample test, since the data for 2001 were not available when the models were constructed.

No model came close to forecasting the apparent decline, and the differences among the forecasts is much smaller than any of their errors. The integrated models forecasting best from 1971 through 2000 predict 2001 better than the ones forecasting best from 1991 through 2000. Three of the four models that were best over the full period correctly predicted the turning point in gains, although the largest predicted decline was only 6.9 percent (see Table 8). The four integrated models that forecast best over the 1991-2000 period all predicted that gains would continue to grow in 2001. In each group of integrated models, linear interpolation performed best. Both two-step models predicted gains would grow, with the one forecasting best for the full period predicting gains would grow a strong 13.0 percent in 2001. That was the worst forecast.

CBO's mean reversion method, as applied in this paper, predicted that gains would rise a slight 0.7 percent. It predicted growth because the estimated reversion rate available at the end of 2000 would have been just 13 percent and CBO was predicting strong enough growth in GDP at the time to more than offsets that slight degree of reversion. In its January 2001 baseline, CBO actually predicted that gains in 2001 would remain at the level estimated for 2000. In the three previous years CBO had continued to assume a 20 percent reversion rate in spite of the recent experience, and at a 20 percent reversion rate, gains had been predicted to decline in each of the previous 3 baselines. CBO dropped that methodology in the January 2001 baseline because gains had failed to revert in the previous years.

While 2001 provided a true out-of-sample test, it happened to be a most difficult test. The 46 percent decline in 2001 was matched previously only in 1987, when gains subsided after the tax-induced rush to realize gains in 1986. Abstracting from tax-induced changes, the largest previous decline since 1949 was 24 percent in 1970. Not only is the decline in 2001 extreme, it is unexplained by standard determinants. When actual values of economic and financial variables for 2001 are used in equations from Miller and Ozanne, those equations predict declines averaging just 20 percent. That means some omitted variables powerfully affected realizations in 2001. The year was certainly unusual with the attacks on the World Trade Towers and the Pentagon and with the subsequent closing of stock markets. It was also unusual coming at the end of a dramatic bull market. While it is too early to tell what factors accounted for the sharp decline in realizations, it is clear that the year provided a difficult test for historically based models. In addition, a single year cannot provide a definitive test of statistical models.

***Forecasting Multiple Years Ahead***

The integrated quarterly models can generate dynamic forecasts over an arbitrarily long horizon. Thus, under the same assumptions and methodology used in the previous section to generate 1-year-ahead forecasts, the models can forecast gains in quarter t + j using previous forecasts of all variables in quarters t + j – 1, t + j – 2, …, in place of actuals. With a single model, it is possible to generate forecasts of gains over the 10-year horizon that the CBO currently is required to do.

Although it is possible to forecast indefinitely with these models, there is good reason to believe that their forecast accuracy will deteriorate fairly rapidly. That is because the search for variables and hyperparameter values was targeted at getting the best 1-year forecasts of gains. If accuracy over a longer horizon were desired from the outset, different variables and hyperparameter values would have been found to be best. For instance, the weight on the cointegration dummy is almost certain to increase as the targeted forecast horizon is extended.

These considerations lead to the question, for how long could these models be used to forecast more accurately than viable alternatives? We explore the answer to this question with respect to 2 alternatives. The first is an annual random walk with drift model. The second is the CBO mean reversion method as applied earlier in this paper, except that reversion is continued out additional years.

We measure errors in terms of a multiyear version of E2. We define the error in forecast at horizon j as the difference in growth rates of forecast gains at t + j relative to the estimate at t:

$$E2_t(j) = (CG_{t+j})^f/(CG_t)^e - CG_{t+j}/CG_t, \qquad j = 1,2,\ldots,10.$$

The capital gains estimation and forecasts are all based on the standard information sets $[CG]_{t-1}$ and $[X]_{t:4}$. It is readily seen that $E2(1) \equiv E2$. All error calculations are done from out-of-sample forecasts over the years 1971-2000.

Note that the further ahead the forecast, the fewer are the number of forecasts for which errors can be calculated, and more importantly, the fewer are the number of independent errors. The thirty-year period provides 30 tests of the one-year ahead forecasts, and all of them are independent. The same period provides 29 tests of the two-year ahead forecasts, but only 15 independent ones. That is because the two-year ahead forecasts done for adjacent years, say 1981 and 1982, have a common year between them, 1982. At the extreme of 10-year ahead forecasts, only 21 forecasts can be made and only three of them are completely independent. Clearly, the fewer the number of independent forecasts, the less confidence that can be placed in the robustness of the RMSEs.

The quarterly model selected for the comparison is the one with the lowest RMSEs over the 1971-2000 period from Table 6. It uses linear interpolation of gains. We did not select any of the models that were best in the 1990s because they predicted poorly in the earlier decades. And we did not select any two-step models because the two-step approach does not easily extend to multiyear horizons.

In Table 9, we compare the RMSEs in terms of E2(j) errors for CBO's mean reversion method, the random walk with drift model, and the linearly interpolated model. The CBO forecasts go six years ahead, the other two go ten.[17] The CBO forecasts are conditioned on estimates of gains in the base year made with a simplified CBO equation, as described in the above section on results from applying the two-step method. The random walk and BVAR models are conditioned on the same estimates from the best linearly interpolated estimation model.

In terms of these RMSEs, the interpolated model outperforms the other two methods for only 3 years, and then falls farther behind with each successive year forecast. The comparison also shows that the random walk and CBO methods have similar errors. The random walk model's errors are slightly lower than the mean reversion's errors in the first two years forecast and become progressively higher in the next four. Although the difference becomes larger for five and six-year forecasts, the decreasing number of independent observations at these horizons makes it impossible to conclude that one method is systematically better than the other.[18]

---

[17] Six years ahead are forecast because that is the number of years that CBO forecasts in its annual reports from 1976 through the mid-1990s. The CBO methodology described previously forecasts gains using the assumption that gains revert to their historically expected size relative to GDP. Thus the forecasts of gains rely on a forecast of GDP, or, before the mid-1990s, of GNP. The forecast errors reported here are based on actual CBO forecasts when they exist. In earlier years, imitations have been substituted by growing GNP for six years at its average rate during the prior 5 years. CBO has made 10-year forecasts in recent years, but too little time has lapsed to evaluate the accuracy of gains forecasts seven or more years ahead.

[18] The limited number of independent forecasts must account for the anomaly that the random walk method has lower errors forecasting 10 years ahead than 9.

The comparisons support our reasoning above that models selected on the basis of forecasting one-year ahead lose their superiority after a few years. Why they become worse than the alternatives probably reflects the tendency for gains to fluctuate widely from year to year but to, at least in the period of our data, return to its average size relative to the size of the economy. The BVAR models seem to capture more of the factors causing the annual fluctuations but at the expense of slighting the longer-range trends. The random walk model forecasts only the longer-run trend because its forecast is for gains to grow at the historical average rate known at the time the forecast is made. The mean reversion model captures the longer run trend through the forecast of GDP and some movement around trend to the extent reversion occurs in the forecast interval.

## 7. Conclusion

The primary objective of this paper has been to improve on the method of forecasting gains employed by the CBO. We find that both the two-step model and the integrated quarterly model forecast tax-adjusted gains better than a mean reversion method similar to that used by CBO. Both models also forecast better than two other basic forecasting methods: a random walk method with drift, or a slightly more general autoregressive model.

The integrated quarterly model achieves the greatest improvement in forecasting. Over the 1971-2000 period, the RMSE from forecasting with the best integrated quarterly model is 36 percent below that from forecasting with the mean reversion method. The RMSE from forecasting with the best two-step method is 20 percent below the error from forecasting with the mean reversion method.  Similar improvements occurred for models best at forecasting over the 1991-2000 decade.

The same span of 30 years provides less opportunity to test nonoverlapping forecasts of two or more years ahead. Within those limits, the models developed here appear to forecast better than the simple models for two or three years ahead, but not for longer periods. An exercise similar to the one conducted here would be necessary to determine whether models could be found that would forecast three or more years ahead better than the simpler methods.

It can be argued that our models forecast 1971-2000 better than the simple methods because our models employ more variables and hyperparameters whose influences are determined through extensive searches for what works. We agree, but hope that the models succeed because they reflect actual economic behavior that will continue to influence capital gains in the future. In that case, our models should continue to forecast better than the simpler alternatives.

So far, only one true year of out-of-sample forecasts is available, and that year proved to be an historical outlier. One year is too few to draw conclusions about the relative success of our models and the simpler methods. The main lesson from 2001 is that none of the models picked up the extent of the decline in capital gains that occurred.

Finally, the two-step and the integrated quarterly models were able to estimate current year tax-adjusted gains more accurately than a simple Miller-Ozanne equation. This result suggests that the additional variables uncovered in our search could be used to improve the full Miller-Ozanne equations as used on unadjusted capital gains. The primary candidates discovered are productivity, the interest-rate spread, real compensation per hour, and real fixed private domestic investment.

**Table 1: Errors in Forecasts of Growth Rates for Baseline Year (in Percentage Points)**

| Date of Baseline | Growth from Year Before: Actual | Forecast | Error | Growth from Two Years Before: Actual | Est-Frcst | Error |
|---|---|---|---|---|---|---|
| Feb-87 | -54.7 | -47.1 | -7.6 | -13.7 | -31.4 | 17.7 |
| Feb-88 | 9.5 | -8.3 | 17.9 | -50.4 | -59.7 | 9.3 |
| Jan-89 | -5.3 | 21.2 | -26.5 | 3.8 | 33.5 | -29.7 |
| Jan-90 | -19.6 | 13.4 | -33.0 | -23.9 | 57.1 | -81.0 |
| Jan-91 | -9.8 | 11.2 | -21.0 | -27.6 | 24.6 | -52.2 |
| Jan-92 | 13.5 | 9.1 | 4.4 | 2.4 | 20.0 | -17.6 |
| Jan-93 | 20.2 | 10.6 | 9.6 | 36.4 | 35.2 | 1.3 |
| Jan-94 | 0.3 | 10.7 | -10.4 | 20.5 | 17.4 | 3.1 |
| Jan-95 | 17.9 | 18.7 | -0.7 | 18.3 | 22.8 | -4.5 |
| May-96 | 44.7 | 17.3 | 27.4 | 70.7 | 33.4 | 37.3 |
| Jan-97 | 39.9 | 4.6 | 35.3 | 102.5 | 13.9 | 88.6 |
| Jan-98 | 24.8 | -7.1 | 31.9 | 74.6 | 36.0 | 38.6 |
| Jan-99 | 21.4 | -5.5 | 26.9 | 51.5 | 6.8 | 44.7 |
| Jan-00 | 16.6 | -3.9 | 20.5 | 41.5 | 9.2 | 32.4 |
| RMSE | | | | | | |
| 1987-2000 | | | 22.3 | | | 42.0 |
| 1992-2000 | | | 22.1 | | | 39.6 |

NOTES:

Forecasts are typically completed the month before the baseline is released.

Forecast from year before is from estimated value in year before.

Est-Frcst is estimate of growth from preliminary data two years before baseline to year before, and forecast from year before to year of baseline.

Starting in baseline for 1992, forecasts are made using a form of mean reversion.

Forecasts for baselines of 1995 and 1996 contain adjustments for anticipated tax changes.

Forecast for baseline of 1997 takes no account of tax reduction enacted in 1997.

Forecast for baseline of 1998 contains adjustment for responses to 1997 tax reduction.

**Table 2: Equation Used to Construct Tax-Adjusted Capital Gains**

Linear Regression - Estimation by Least Squares
Dependent Variable DLRATIOFE
Annual Data From 1949 To 2000
Usable Observations 52 Degrees of Freedom 47
Centered R**2 0.747748 R Bar **2 0.726280
Uncentered R**2 0.750266 T x R**2 39.014
Mean of Dependent Variable 0.0244196989
Std Error of Dependent Variable 0.2455856527
Standard Error of Estimate 0.1284861497
Sum of Squared Residuals 0.7759084613
Regression F(4,47) 34.8304
Significance Level of F 0.00000000
Durbin-Watson Statistic 2.024232

| Variable | Coefficient | Std Error | T-Stat | Significance |
|---|---|---|---|---|
| 1. Constant | -0.0490061 | 0.0208595 | -2.34934 | 0.02306045 |
| 2. DMTRNEXT | -2.5742500 | 0.7479459 | -3.44176 | 0.00122346 |
| 3. DMTRTRANS | -117.8239911 | 15.1468711 | -7.77877 | 0.00000000 |
| 4. DLGAP | 3.6162592 | 0.7936502 | 4.55649 | 0.00003702 |
| 5. DLSP500Q4 | 0.8308007 | 0.1262210 | 6.58211 | 0.00000004 |

NOTES:

DLRATIOFE is the change in the logarithm of the ratio of capital gains to potential Gross Domestic Product (GDP)

DMTRNEXT is the change in our measure of the permanent tax rate on capital gains.

DMTRTRANS is the change in our measure of the transitory tax rate on capital gains.

DLGAP is the change in the logarithm of the ratio of actual to potential GDP.

DLSP500Q4 is the change in the logarithm of the average S&P 500 closing price during the fourth quarter of each year.

Equation is estimated with data available as of July 2002.

Development of equation is described in Preston Miller and Larry Ozanne, *Forecasting Capital Gains Realizations*, Congressional Budget Office, Technical Paper 2000-5, August 2000.

**Table 3: Explanatory and Auxiliary Variables**

| Code | Class | Description |
|---|---|---|
| X1 | Explanatory | Output Per Hour in Non-farm Business Sector |
| X2 | Explanatory | Real GDP = Nominal GDP/ GDP deflator |
| X3 | Explanatory | Real PCE on Durable Goods / Real DPI |
| X4 | Auxiliary 1 | Spread, Moody's Seasoned Corp. Bonds (AAA) minus TB Rate (3M) |
| X5-1 | Auxiliary 2 | Real Compensation per Hour in Non-farm Business Sector |
| X5-2 | Auxiliary2 | Wage and Salary Disbursement (All Industries) / GDP |
| X6-1 | Auxiliary 3 | Producer Price index (Crude materials) / GDP Deflator |
| X6-2 | Auxiliary 3 | Producer Price index (Crude materials) / PCE Deflator |
| X6-3 | Auxiliary 3 | Nominal Effective Exchange rate/PCE Deflator |
| X6-4 | Auxiliary 3 | Real Fixed Private Domestic Investment |
| X6-5 | Auxiliary 3 | Real Gross Private Domestic Investment |
| X6-6 | Auxiliary 3 | Nominal Effective Exchange Rates |
| X6-7 | Auxiliary 3 | Real Private Nonresidential Fixed investment |
| X7 | Auxiliary 4 | S&P 500 Index |

**Table 4: Root Mean Squared Errors in Estimating and Forecasting with Two-Step and Comparison Models (in percentage points)**

| | | RMSE by Decade and Full Period | | | |
|---|---|---|---|---|---|
| Model | Type of Error | 1971-1980 | 1981-1990 | 1991-2000 | 1971-2000 |
| **CBO** | E1 | 17.46 | 12.30 | 11.94 | 14.13 |
| | E2 | 21.13 | 13.28 | 20.29 | 18.57 |
| | E3 | 26.86 | 22.47 | 27.34 | 25.65 |
| **RW** | | | | | |
| | E1 | 19.24 | 15.15 | 15.49 | 16.73 |
| | E2 | 19.13 | 15.26 | 15.69 | 16.78 |
| | E3 | 33.27 | 24.03 | 32.54 | 30.24 |
| **AR(1)** | | | | | |
| | E1 | 21.51 | 15.20 | 17.06 | 18.12 |
| | E2 | 21.29 | 15.55 | 17.64 | 18.31 |
| | E3 | 38.28 | 25.11 | 36.17 | 33.68 |
| **TS-all** | | | | | |
| | E1 | 9.34 | 11.61 | 11.7 | 10.94 |
| | E2 | 15.47 | 15.76 | 13.03 | 14.80 |
| | E3 | 26.92 | 15.05 | 25.14 | 22.98 |
| **TS-90** | | | | | |
| | E1 | 11.17 | 22.18 | 8.79 | 15.21 |
| | E2 | 28.41 | 21.51 | 12.55 | 21.81 |
| | E3 | 36.91 | 43.48 | 21.29 | 35.15 |

**Table 5: Root Mean Squared Errors in Estimating with Two-Step and Integrated Models (E1 Errors in Percentage Points)**

| | RMSE by Decade and Full Period | | | |
|---|---|---|---|---|
| Model | 1971-1980 | 1981-1990 | 1991-2000 | 1971-2000 |
| | Best for 1971-2000 | | | |
| **Annual** | | | | |
| Two Step | 9.34 | 11.61 | 11.70 | 10.94 |
| **Quarterly** | | | | |
| Linear | 9.36 | 11.66 | 9.56 | 10.25 |
| Economic | 9.13 | 9.13 | 10.98 | 10.63 |
| In-Model Lin. | 10.73 | 12.22 | 12.66 | 11.90 |
| In-Model Econ. | 11.08 | 10.36 | 12.09 | 11.12 |
| | Best for 1991-2000 | | | |
| **Annual** | | | | |
| Two Step | 11.17 | 22.18 | 8.79 | 15.21 |
| **Quarterly** | | | | |
| Linear | 17.07 | 15.86 | 8.90 | 14.40 |
| Economic | 19.24 | 15.23 | 9.78 | 15.25 |
| In-Model Lin. | 19.96 | 19.96 | 8.06 | 16.81 |
| In-Model Econ. | 11.90 | 16.27 | 10.17 | 13.04 |

**Table 6: Root Mean Squared Errors in Years 1971-2000 from Forecasting One Year Ahead (in percentage points)**

| Model | Type of Error | RMSE by Decade and Full Period | | | |
|---|---|---|---|---|---|
| | | 1971-1980 | 1981-1990 | 1991-2000 | 1971-2000 |
| **Annual** | | | | | |
| CBO | E2 | 21.13 | 13.28 | 20.29 | 18.57 |
| Two Step | E2 | 15.47 | 15.76 | 13.03 | 14.80 |
| | E3 | 26.92 | 15.05 | 25.14 | 22.98 |
| **Quarterly** | | | | | |
| Linear | E2 | 10.57 | 12.47 | 12.62 | 11.92 |
| | E3 | 18.04 | 19.92 | 21.01 | 19.69 |
| Economic | E2 | 11.39 | 11.73 | 15.91 | 13.17 |
| | E3 | 20.14 | 14.92 | 28.36 | 21.85 |
| In-Model Lin. | E2 | 10.71 | 11.80 | 15.33 | 12.77 |
| | E3 | 19.45 | 20.31 | 27.79 | 22.82 |
| In-Model Econ. | E2 | 10.25 | 13.00 | 15.32 | 13.03 |
| | E3 | 19.50 | 18.60 | 27.86 | 22.38 |

**Table 7: Root Mean Squared Errors in Years 1991-2000 from Forecasting One Year Ahead (in percentage points)**

| Model | Type of Error | RMSEs by Decade and Full Period | | | |
|---|---|---|---|---|---|
| | | 1971-1980 | 1981-1990 | 1991-2000 | 1971-2000 |
| **Annual** | | | | | |
| CBO | E2 | 21.13 | 13.28 | 20.29 | 18.57 |
| Two Step | E2 | 28.41 | 21.51 | 12.55 | 21.81 |
| | E3 | 36.91 | 43.48 | 21.29 | 35.15 |
| **Quarterly** | | | | | |
| Linear | E2 | 22.11 | 18.50 | 11.46 | 17.91 |
| | E3 | 38.34 | 32.13 | 20.26 | 31.16 |
| Economic | E2 | 18.85 | 16.27 | 13.41 | 16.33 |
| | E3 | 36.35 | 24.14 | 24.16 | 28.79 |
| In-Model Lin. | E2 | 23.85 | 17.26 | 11.72 | 18.30 |
| | E3 | 36.81 | 34.25 | 18.62 | 30.96 |
| In-Model Econ. | E2 | 16.17 | 14.97 | 11.55 | 15.54 |
| | E3 | 24.46 | 29.84 | 19.50 | 24.96 |

**Table 8: Gains Estimates for 2000 and Forecasts for 2001**
**(Billions of dollars)**

| | Estimate of 2001 | Forecast of 2001 | Growth Rate |
|---|---|---|---|
| CBO Baseline | 652 | 652 | 0.0% |
| CBO Reversion Methodology | 563 | 567 | 0.7% |
| | | | |
| Models best for 1971-2000 | | | |
| Two Step | 620 | 701 | 13.0% |
| Linear | 615 | 573 | -6.9% |
| Economic | 635 | 643 | 1.2% |
| In-Model Lin. | 651 | 639 | -1.9% |
| In Model Econ. | 642 | 640 | -0.2% |
| | | | |
| Models best for 1991-2000 | | | |
| Two Step | 657 | 702 | 6.9% |
| Linear | 646 | 653 | 1.0% |
| Economic | 647 | 707 | 9.2% |
| In-Model Lin. | 663 | 699 | 5.4% |
| In-Model Econ. | 598 | 611 | 2.2% |
| | | | |
| Addendum: Actual Values | 644 | 349 | -45.8% |

**Table 9: Root Mean Squared Errors in Multiple Year Forecasts (in percentage points)**

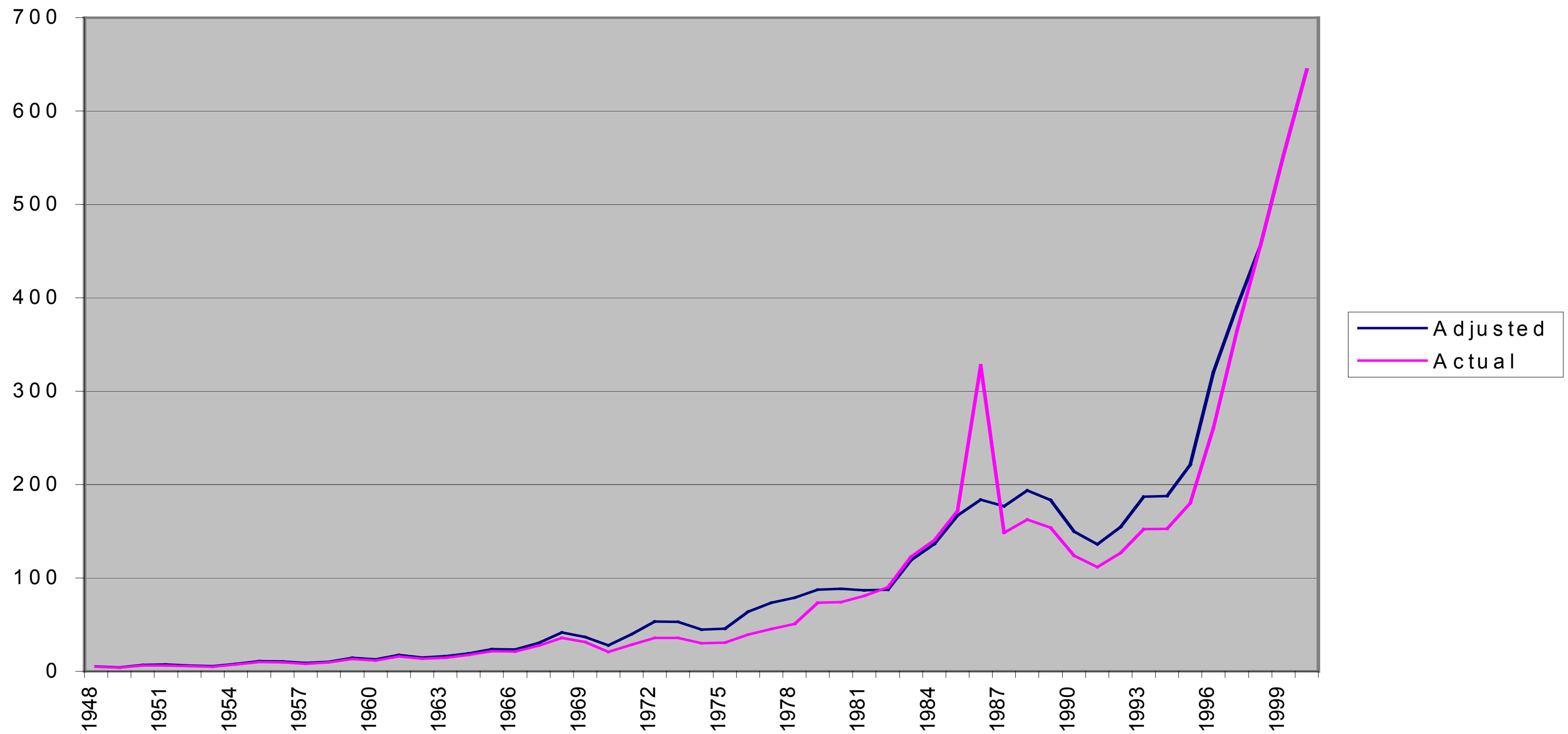
| | Years Ahead Forecast | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| CBO | 18.53 | 31.83 | 39.54 | 48.41 | 60.79 | 69.55 | | | | |
| RW | 16.76 | 30.69 | 40.09 | 51.11 | 64.55 | 74.42 | 80.17 | 88.94 | 92.06 | 84.75 |
| BVAR | 11.92 | 24.90 | 37.60 | 52.42 | 69.44 | 82.68 | 92.36 | 106.06 | 120.36 | 126.17 |

NOTES:

The row titled CBO forecasts with the mean reversion method described earlier. Here, however, forecasts of GDP are taken from CBO's winter baselines. These baselines are usually published in January. The first of these baselines is for January 1976, and the last is January 1999. We substituted baseline forecasts for 1971-1975 in which GDP was assumed to grow at its average rate for the previous five years. The forecasts in the table are limited to horizons of six years because CBO forecasts covered only 6 years until the mid-1990s. In addition, the forecasts made before the mid-1990s were of GNP rather than GDP.

The RMSE on one-year ahead forecasts reported for CBO is shown in the table as 18.53 while the same conceptual RMSE reported as E2 on Table 4 and Table 7 is slightly higher at 18.57. The difference arises because the mean reversion forecasts in this table are based on historical CBO forecasts of aggregate output whereas those in Table 4 and Table 7 are based on forecasts of aggregate output taken from a BVAR. The similarity of the RMSEs indicates the similarity of the independent forecasts of aggregate output.

**Figure 1: Annual Gains and Tax-Adjusted Gains (Billions of dollars)**



## Appendix A: Specifications of Forecasting Models

In this appendix, we first summarize the specifications of the models compared in the text and then describe the role of the hyperparameters.

Each model is identified by i) the set of variables used for current year estimation, ii) the set of variables used for one-year-ahead forecasting, and iii) the set of hyperparameters used for estimation and/or forecasting. See Table A1. As shown in the first column of Table 3 of the main text, variables are divided into seven categories. Therefore, we use the following convention for the *variable IDs* to describe whether a variable is used in the estimation or forecasting step:

1. If the $i^{th}$ element in a *variable ID* is 0 , no variable in the $i^{th}$ category is used.
2. If the $i^{th}$ element in a *variable ID* is $j$, the $j^{th}$ variable in the $i^{th}$ category is used.

For example, a *variable ID* of [1 1 0 0 2 7 0] corresponds to a model that employs [X1 X2 X52 X67] in Table 3.

The BVAR method used in the present work features a set of inexact prior restrictions on the coefficients and the covariance matrix, treating them as random quantities with given mean values. The tightness of their distributions around prior means is determined by a set of hyperparameters.[19] Roughly, the following three prior restrictions are imposed:

[19] See Robertson and Tallman for formal definitions of the hyperparameters. The six hyperparameters ($\mu_1$, $\mu_2$, $\mu_3$, $\mu_4$, $\mu_5$, $\mu_6$) described in this appendix correspond to ($\lambda_0$, $\lambda_1$, $\lambda_4$, $\lambda_3$, $\lambda_5$, $\lambda_6$) in that paper.

i) The variables in a VAR follow a multivariate random walk with drift. In other words, the means of the coefficients on the first and higher-order lags are an identity and zero matrices, respectively.

ii) The VAR in question is expressed entirely in terms of differenced series with an order lower by one. In other words, variables are difference stationary without cointegration among them.

iii) The VAR system being considered has a single unit root. In other words, there is possible cointegration among variables.

The three priors can coexist because none needs to be imposed completely.

The first four hyperparameters control the tightness of the prior of random walk with drift. Among them, the first hyperparameter, $\mu_1$, controls the overall tightness of the random walk prior, or equivalently that of the prior on the innovation covariance matrix. Increasing $\mu_1$ renders less tight the random walk prior as a whole. The second hyperparameter, $\mu_2$, is the prior standard deviations of the diagonal elements of the coefficients on the first lag, reflecting how closely the random walk approximation is imposed. As $\mu_2$ decreases toward zero, the diagonal elements of the first lag coefficients get closer to one and other off-diagonal coefficients to zero. The third hyperparameter, $\mu_3$, determines the degree by which the intercept term shrinks toward mean of zero, controlling its standard deviation. The fourth hyperparameter, $\mu_4$, determines the extent to which coefficients on the lags beyond the first one are likely to be different from zero. As $\mu_4$ increases, the coefficients on higher order lags shrinks toward zero more rapidly.

The fifth hyperparameter, $\mu_5$, is the weight on a VAR in terms of differenced series. As $\mu_5$ increases to infinity, the model tends to have as many unit roots as variables and no cointegration among them.

The last hyperparameter, $\mu_6$, is the weight on the single unit root prior. As $\mu_6$ increases, more and more weight is put on a VAR in which all variables share a single stochastic trend and the intercept is close to zero.

**Table A1: Model Specifications**

| | 1971-2000 | | 1991-2000 | |
|---|---|---|---|---|
| Model | Variable ID | Hyperparameters | Variable ID | Hyperparameters |
| Two Step | | | | |
| Estimation | 1 1 0 1 1 4 0 | NA | 1 0 1 0 3 1 1 | NA |
| Forecasting | 1 1 1 0 0 4 0 | 0.7 0.15 0.05 0.5 6 4 | 0 0 0 1 0 3 1 | 0.8 0.125 0.125 1.5 3 6 |
| Lin. Interpol. | | | | |
| Estimation | 0 0 0 1 0 4 1 | 0.8 0.15 0.05 0.5 7 7 | 0 0 1 0 0 1 1 | 0.6 0.05 0.075 0.5 4 7 |
| Forecasting | 0 1 1 1 2 6 0 | 0.8 0.15 0.05 1.25 7 3 | 0 0 1 0 0 1 1 | 0.7 0.125 0.05 0.75 3 7 |
| Econ. Interpol. | | | | |
| Estimation | 1 0 0 1 0 4 0 | 0.8 0.15 0.125 0.5 7 7 | 0 0 1 0 0 4 0 | 0.8 0.15 0.05 1.5 7 7 |
| Forecasting | 0 0 1 1 2 7 1 | 0.8 0.15 0.05 0.75 7 3 | 0 0 1 0 0 1 1 | 0.8 0.05 0.1 0.5 3 7 |
| In-Model Lin. | | | | |
| Estimation | 0 0 0 1 0 4 0 | 0.5 0.075 0.1 1.5 3 7 | 0 0 1 1 1 1 1 | 0.5 0.075 0.1 1.5 3 7 |
| Forecasting | 0 0 1 0 0 1 1 | 0.7 0.15 0.075 1.25 3 7 | 0 0 1 0 0 1 1 | 0.7 0.15 0.075 1.25 3 7 |
| In-Model Econ. | | | | |
| Estimation | 0 0 0 1 0 4 0 | 0.8 0.125 0.05 0.75 7 4 | 0 0 0 1 1 4 1 | 0.4 0.075 0.1 1.5 7 7 |
| Forecasting | 0 0 1 1 2 6 1 | 0.7 0.125 0.1 0.75 7 3 | 0 0 1 0 0 0 0 | 0.7 0.15 0.05 0.75 6 3 |

## Appendix B: Annual vs. Quarterly Univariate Models of Gains

In this appendix, we offer an explanation for the forecasting superiority of a random walk model with drift when it is estimated with linearly interpolated quarterly gains instead of with actual annual gains. Our explanation is not general: it is thought to hold for particular series over particular times. For the gains series it holds over the 1971-2000 period. However, it holds over neither the 1960s decade nor the period 1961-2000. Our explanation requires that the forecast errors from the model for annual gains be positively serially correlated. We first show with a diagram and simple algebra why this condition leads to forecasting superiority of the interpolated gains model and then verify that this condition holds for the period 1971-2000.

In the diagram, we illustrate the difference in the two univariate models by contrasting their forecasts in the gains-time space. In this space, a random-walk-with-drift model is represented by a line with slope equal to the drift and intercept determined by the initial observation.

$$X(t) = \text{Drift} + X(t-1)$$

To simplify the presentation, we refer to X(t) as gains in period t, however, in our actual estimates of the models referred to footnote 16 and in the tables, X(t) is measured by the logarithm of tax-adjusted gains. To further simplify, we suppose that the drift term for the annual model is equal to 1, and that two models have made no errors over the years prior to t.

The annual model is represented in the upper graph by the two lines passing through midpoints (i.e., the end of the second quarters) of each year. That is, annual gains are plotted in the middle of each year, while the values in other quarters should be ignored. By assumption, in year t-2 the forecast from the model is accurate; gains are forecast to rise to 9 in year t-1 and they do. In year t-1, gains are forecast to rise to 10 in year t, and if no shock occurs in year t, the annual model forecasts gains will rise to 11 in year t+1, shown on the dashed line labeled "No shock".

The quarterly model is represented in the lower chart. The requirement that the average of the quarterly gains in a year equal the annual gains implies that the midpoint of a line in a year equals the annual value of gains. Since, by assumption, annual gains prior to t are linearly related, the quarterly model's forecasts and actuals also coincide through t-1. Also, the midpoint of the "No shock" line in year t is the model's forecast of gains for the year, and the midpoint in year t+1 is the model's forecast for the following year assuming no forecasting error in year t.

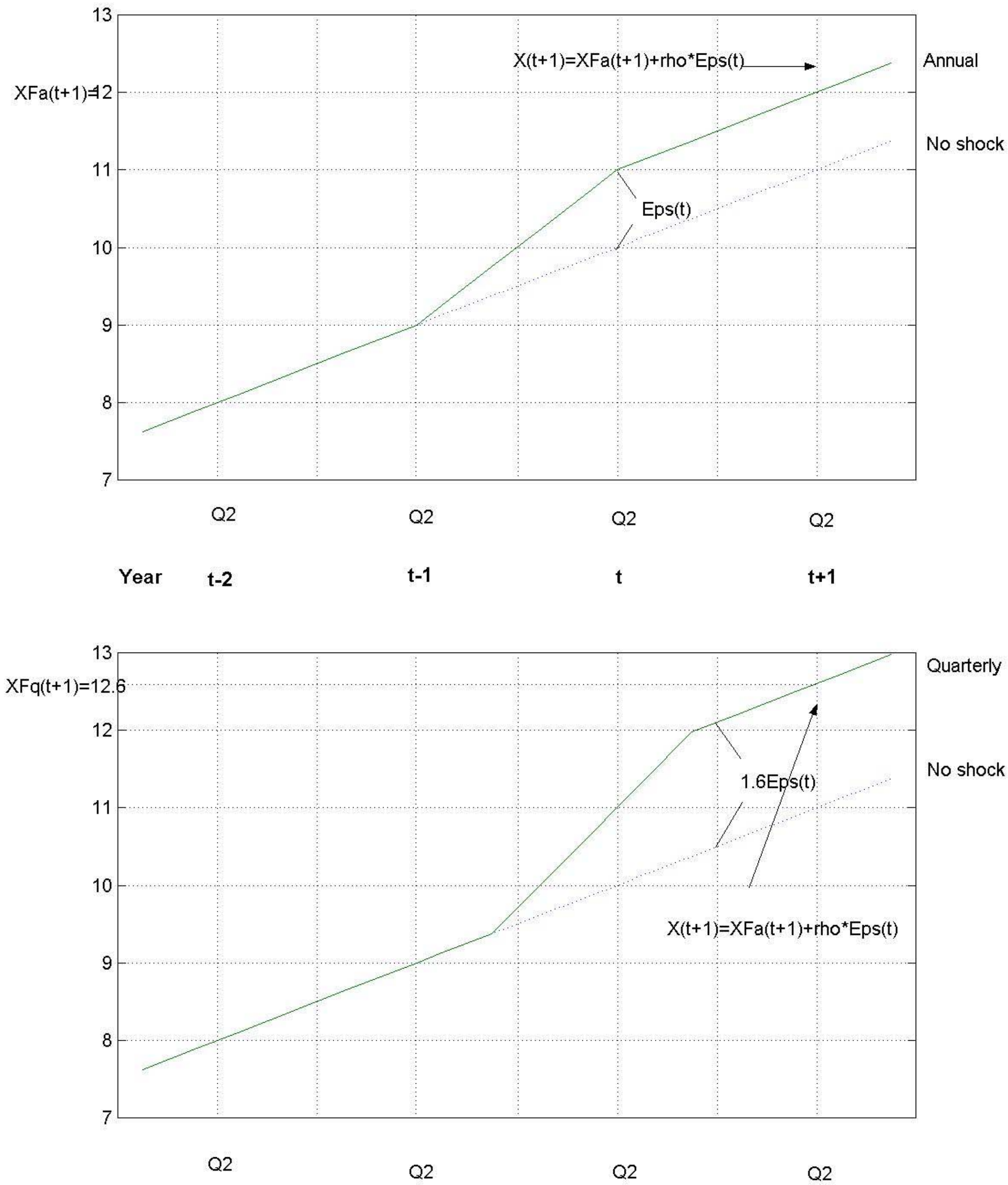
The difference in the two models is seen when there occurs a shock Esp(t)=1 in the annual value of gains for year t. Assuming the drift term is unchanged, the annual model now starts from the actual value of X(t)=11 at the midpoint of the year t and adds drift to generate a revised forecast of 12 in t+1. (The forecast is shown as Xfa(t+1) in the diagram.) Thus, for the annual model, the revision in its forecast for X(t+1) equals Eps(t), reflected by the upward shift of the line "Annual" from the "no shock" line. For the quarterly model, however, the error of Eps(t) must be reflected at the midpoint by the average of revisions for the 4 quarters of the year. Since the quarterly values of actual gains must lie along a line segment in year t, the revisions in forecasts from quarter 1 to quarter 4 must be .4Eps(t), 0.8Eps(t), 1.2Eps(t), and 1.6Eps(t), respectively. Since the forecast from this model starts from the 4$^{th}$ quarter of the year t, the line "Quarterly" shows an upward shift by 1.6Eps(t) from the "No shock" line. Its forecast for year t+1 is 12.6, labled XFq(t+1) in the diagram.

For both models, it is not quite right that the estimated drift terms will be unaffected by the error in year t. Nevertheless, we would expect the two estimated drift terms on an annual basis to move similarly. Moreover, the change in drift caused by one error should diminish as the number of observations increases, and the difference in forecasts caused by the difference between Eps(t) and 1.6Eps(t) should dominate.

If the annual model is correctly specified, its errors will be uncorrelated and its forecasts should be more accurate than those from the quarterly model: without a further shock in year t+1, the quarterly model overpredicts the gain in t+1, while the annual forecasts accurately. However, if the annual model's errors are sufficiently positively correlated, the quarterly model can be more accurate: if Eps(t) follows an AR(1) process with coefficient of rho, for example, the actual gain in t+1 will be higher than XFa(t+1) by rho times Eps(t), and the degree of underprediction by the annual model can outweigh that of overprediction by the quarterly model. That can occur in our example when rho is closer to 0.6 than to zero.

We computed the out-of-sample forecast errors from an annual random walk-with-drift model over the years 1971-2000. The variable modeled is the logarithm of tax-adjusted gains. We judge the degree of first-order serial correlation by the Durbin-Watson statistic. Over this 30-year period, the statistic is 0.7822, which suggests a significant degree of positive serial correlation.

Diagram: Forecasting with an Annual Model (Upper Graph) and an Interpolated Quarterly Model (lower Graph)



## Appendix C:  Economic Interpolation of Quarterly Gains

Our initial thought on how to interpolate gains with economic data was to apply the Miller-Ozanne equation of Table 2 to predict the growth rate of gains from one quarter to the next. Applying the equation quarterly would require the assumption that the growth rate of stock prices and the business cycle over a quarter have the same affects on the growth of gains in that quarter as their growth over a year has on the growth of gains over the year.  This assumption seemed unlikely to us.  We thought that the growth rate of gains in one quarter was likely to depend on changes in stock prices and the business cycle over more than one quarter.

Consequently, we chose to interpolate gains to quarters by maintaining the four-quarter span of time per observation, and then applying the growth rate for that span to the last quarter in the span.  This keeps the data used in the equation covering the same span of time as that used to estimate the equation, but requires us to arbitrarily assign the predicted growth to specific quarters.

More specifically, we compute both explanatory variables for each span of 4 quarters ending in the fourth quarter of 1948 through the fourth quarter of 2000.  Then we apply the equation to predict the growth rate of gains between adjacent periods of four quarters.  We assign the growth rate between the first and the second four-quarter span to the growth between the last quarter of the first span and the last quarter of the second span.

For example, the growth rate of gains from the third quarter of 1962 to the third quarter of 1963 is computed by first computing the levels of the business cycle and stock price variables over the four quarters ending in the third quarter of 1962 and again for the four quarters ending in the third quarter of 1963. Next, the change in the log of each explanatory variable is computed and used in the equation to compute a growth rate for gains. (Tax rates are assumed to be constant so as to obtain a growth rate for tax-adjusted gains.) The growth rate computed in this manner technically applies to the growth of gains from the first to the second span of four quarters. However, we arbitrarily say it is the growth from the third quarter of 1962 to the third quarter of 1963.

The level of the business cycle variable for the span from the fourth quarter of 1962 through the third quarter of 1963 is computed by averaging the quarterly values of GDP and also of potential GDP and then taking the ratio of average GDP to average potential GDP. A similar computation is done for the four quarters ending in the third quarter of 1962. The change in the logarithm of the ratios so computed matches the time span of the variables in the Miller-Ozanne equation.

The level of the stock price variable for the four quarters ending in the third quarter of 1963 is simply the level for the third quarter of 1963, and that for the four quarters ending in the third quarter of 1962 is the level for that last quarter. The reason is that the stock price variable in the Miller-Ozanne equation is specified as the change in the logarithm of the stock prices from the fourth quarter of one year to the fourth quarter of the next.

The dependent variable of the Miller-Ozanne equation is the change in the logarithm of the ratio of gains to potential GDP. Consequently, once the equation has been applied to adjacent spans of four quarters, the predicted change in the logarithm of the ratio must be converted to the change in the logarithm gains. The conversion is done by adding to the prediction from the equation the change in the logarithm of potential GDP computed over the same interval. The change in the logarithm of a variable is what we refer to in this appendix as the growth rate of the variable.

The process for computing the growth between the third quarters of 1962 and 1963is repeated for all four-quarter intervals between the fourth quarter of 1948 and the fourth quarter of 2000. Every fourth interval corresponds to a calendar year, and so has the growth rate computed from the annual data.

The next step in the interpolation converts the estimated growth rates to quarterly values of gains. Conversion is done by distributing the annual tax-adjusted gains for 1960 evenly among its quarters, and then using the growth rates computed above to get the gains in all other quarters. We want quarterly gains to be measured at annualized rates, because that is the way several of the macroeconomic variables are measured that we will combine with quarterly gains in our integrated models. Consequently, we set the gain for each quarter of 1960 to the year's total tax-adjusted gain. Applying the growth rates to those levels makes the predicted quarterly gains for other years measured at anual rates as well.

The year 1960 is chosen because gains were plausibly realized at a constant rate throughout that year. Gains in 1960 were at a trough between 1959 and 1961, and growth rates of both the business cycle and the S&P were relatively flat during the year's four quarters. Based on this evidence, we distribute annual tax adjusted gains evenly across the four quarters.

Given gains in, say, the first quarter of 1960, gains in the first quarter of 1961 are determined by applying the computed growth rate of gains for the first quarter of 1961.

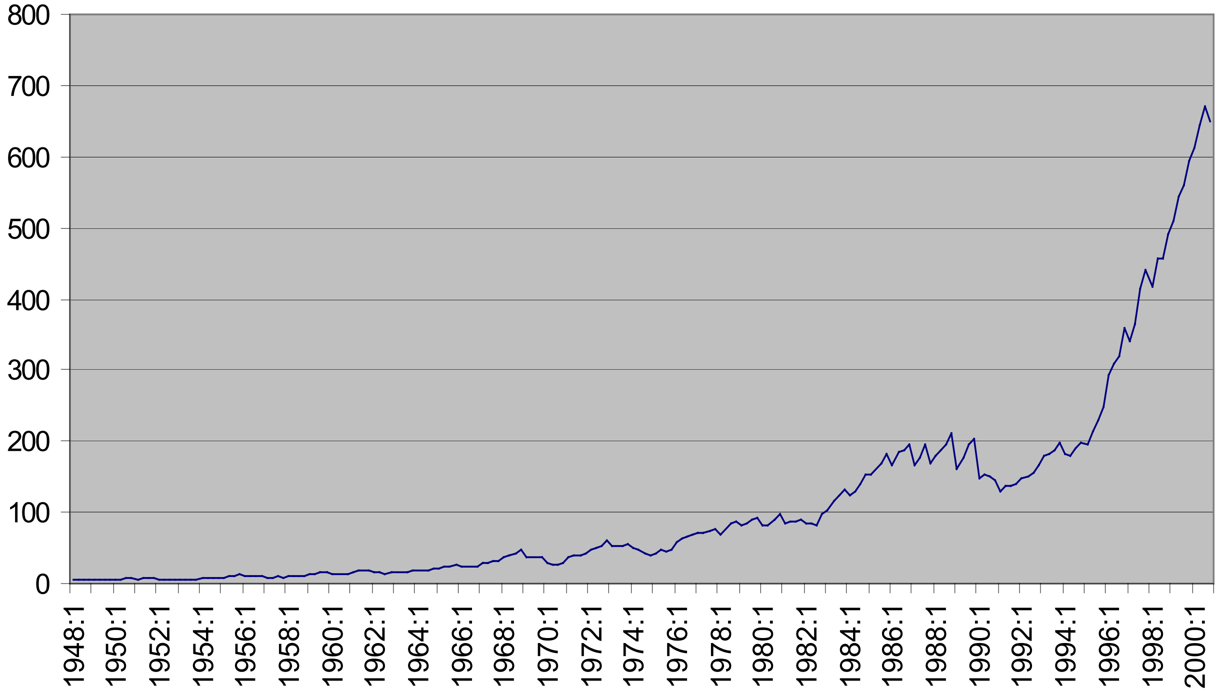
And gains in the first quarter of 1959 are determined by subtracting the computed growth rate of gains for the first quarter of 1960 from the level of gains in that quarter.   Gains in each other quarter of 1961 and 1959 are determined by applying the appropriate growth rate to gains in the corresponding quarter of 1960.  Once gains are determined for each quarter of 1961 and 1959, a similar process determines gains for each quarter of 1962, and 1958.  Repeating the process for each adjacent year determines gains forward through the fourth quarter of 2000 and backward to the first quarter of 1948.

Finally, the four quarters of gains in each calendar year are raised or lowered proportionately so that their average equals the value of tax-adjusted gains for that year. This adjustment causes the growth rates between the fourth quarters of one year and the first quarter of the next to differ from the rates indicated by the equation.  As a result, the plot of quarterly gains shows smooth changes from one quarter to the next within a year but frequently shows abrupt shifts between the fourth quarter of one year and the first quarter of the next year.  This pattern creates a saw-tooth pattern in the graph of these interpolated quarterly gains (see Figure C1).

While the above process arrives at a value of tax-adjusted gains for each quarter, the process involves strong assumptions.  One assumption is that the annual growth rate estimated for eight quarters applies to growth to a single quarter from the same quarter in the preceding year.  No constraint is imposed to make growth for three preceding quarters consistent with the annual growth rates estimated for that quarter.  Some consistency should arise because growth rates for four adjacent quarters use overlapping quarters in their measures of growth in business cycles and stock prices.  A second assumption is that unmeasured factors that cause actual tax-adjusted gains in one calendar year to differ from the average of the four quarterly gains estimated through the above process affect all quarters proportionately.

In principle, economic interpolation has an unfair advantage over the two other interpolation methods used in this paper when it comes to testing models over the 1971-2000 period. The Miller-Ozanne equation was estimated over the years 1948 to 2000 and then used to interpolate quarterly gains in all years from 1948 to 2000. Thus the interpolated gains in all years before 2000 reflect future information that would not have been available to forecasters at the time. That future information could help economically interpolated gains to better reflect the influence of economic activity on quarterly gains than could the other methods of interpolation. The better reflection could reduce errors in estimating and forecasting gains during the testing years of 1971-2000. In practice, however, models using economic interpolation never had the lowest errors for estimating or forecasting. Either the future information was not particularly useful or other limitations of economic interpolation offset this advantage.

**Figure C1: Economic Interpolaton of Quarterly Tax-Adjusted Gains**
**(Billions of Dollars)**



## Appendix D: Actual and Interpolated Gains

No information on the timing of gains within a year is available for most years, hence the need to interpolate. In a few years, related data are available, and they provide limited insight to the accuracy of interpolation. A comparison of quarterly patterns in 5 years shows a weak tendency for interpolated gains to move in the same direction as actual gross gains from start to finish, but no tendency to move with quarter-by-quarter changes in actual gross gains. These findings are limited both because of the limited number of years with data on timing of gains and because of conceptual difficulties in converting the actual annual gains variable into a quarterly version. As a result of these limitations, the possibility that interpolated gains provide useful quarterly variation cannot be rejected.

The accuracy of interpolation is important primarily because of its contribution to identifying model parameters. If interpolated gains consistently reflect quarterly movements in actual gains, then interpolation will improve the estimation of parameters. If not, interpolation will add noise that estimation must overcome to arrive useful parameters. Even if interpolation adds only noise about movements within a year, quarterly models may still forecast better than annual ones if the benefits from using other variables on a quarterly basis is large enough.

### *Net Positive Gains and Data on Individual Transactions*

Taxpayers enter information on Schedule D and related schedules about each sale of capital assets they have conducted during the year. The information includes the data of sale as well as the amount of gains. In addition to reporting their own sales, taxpayers report net gains or losses they receive from "pass-through" entities such as mutual funds, partnerships, and trusts. They also include unused losses carried forward from prior years. The form directs taxpayers to combine information from all of these sources into a net gain or loss for the year. If a taxpayer's net is a gain, it is entered into adjusted gross income (AGI). Actually, prior to 1987, taxpayers were allowed to exclude a portion of their long-term gains in excess of their losses before entering the remainder into AGI.

If the net value from the above computations is a loss, only the amount under the loss limit can be included in the current year's AGI, and the remainder can be carried forward to future years. The loss limit was $1,000 between 1955 and 1976, $2,000 in 1977, and has been $3,000 since 1978.

The annual capital gains amount that we are attempting to forecast is the sum of net gains across all taxpayers. The long-term gains that taxpayers excluded in years prior to 1987 are added back to keep the annual totals consistent over time. Net gains summed across all taxpayers is also the figure that we are attempting to interpolate to quarters within a year. (Net losses are forecast separately because the limit makes them steadier from year to year.)

In some years, the Internal Revenue Service publishes greater detail on the transactions taxpayers report, including the date of sale. In those years, it is possible to group many individual transactions by quarter of the year. The IRS or Treasury Department typically prepare tables of transactions data by month of sale, and that data has been used here for five years: 1962, 1973, 1985, 1997, and 1998.

The difficulties of comparing interpolated net gains to distributions of actual transactions are both conceptual and practical. A conceptual difficulty is that net gains include losses that are predominantly realized at the end of the year. In the five years analyzed here, losses in December ranged from 22 percent to 35 percent of all losses, and losses for the last quarter averaged around 40 percent of all losses. The concentration in the end of the year does not result from big drops in asset prices at the end of each year. Instead, it reflects a timing decision by taxpayers to realize losses so they can be counted for tax purposes in the current year. Since this is an annual decision, it would be misleading to count all losses in the quarter they occur. For this reason, we compare our interpolated net gains only to transactions that resulted in a gain. Fortunately these gross gains are much larger than gross losses. In 1998, for example, $584 billion of gross gains were reported compared to $152 billion of gross losses.

The quarterly pattern of gross gains does not show a marked pattern like that of gross losses. With only 5 years of data, however, we cannot be sure that a weaker but still regular pattern is not present, such as taxpayers selling more gains early in the year. If a regular pattern were present, it would tend to show up as an inconsistency between gross gains and our interpolated net gains. Such an inconsistency would not reflect a shortcoming of our interpolated gains, however, because they are intended to represent seasonally adjusted gains. Interpolated gains should be seasonally adjusted because they are used in models with other seasonally adjusted quarterly variables. In other words, if we find a consistent divergence between gross gains and our interpolated gains, it might be due to seasonal patterns rather than an inadequacy in our interpolation methods.

Another conceptual difficulty is that net gains include losses carried over from prior years. Since these losses occurred in another year, they cannot be assigned to any particular time within the year they enter net gains. Of course transactions data for the year to which losses are carried do not reflect those losses either. For this reason alone, interpolated gains cannot be expected to match transactions data on gross gains. Fortunately, losses carried over are usually a small component of net gains and hence the comparison of interpolated gains and gross gains should not be badly distorted by this incompatibility.

A practical problem in comparing interpolated net gains and actual gains is that taxpayers do not report dates for many transactions. In 1998, the IRS could not identify a date for 52 percent of gross gains. In 1985, the IRS could not date 23 percent of gross gains. The problem is not entirely lax reporting by taxpayers. Gains taxpayers receive from the pass-through entities noted above do not identify the date for taxpayers to report.

Reporting of dates is more common for stock transactions than for other assets. In 1998, when 52 percent of all gains lacked a date, 26 percent of stock transactions lacked a date. Stocks also account for a large portion of all gains; in 1998 they accounted for 43 percent of all gains including those without dates. As a result, stocks account for the preponderant share of gains with dates. For that reason, the distributions by quarter of stocks and of all assets look very similar for the three years in which both are available: 1985, 1997, and 1998. In 1962 and 1973, only gains on stocks are reported. For simplicity, then, only gains on stocks with valid dates are distributed by quarter and compared to our interpolated net gains.

***Comparisons***

Three types of interpolations are described in the body of the paper. Briefly, one is linear interpolation, where net gains are grown linearly from the last quarter of one year to the last quarter of the next year at a slope sufficient to make net gains for the four quarters of the next year average to the known annual net gain for that year. A second interpolation, which we call economic interpolation, is constructed with the equation in Table 2 of the paper (see Appendix C for a more complete explanation). The third interpolation is done with BVAR models as they are updated from year to year. This model interpolation yields many different interpolated net gains variables, depending on the other variables and hyperparameters used in the model. For the comparison here, the interpolated gains are from the model with a linear seed that had the lowest E2 RMSE over the period 1971-2000, as reported in Table 7.

The comparisons are summarized in graphs for each year. In 1962, 33 percent of gross stock gains with dates were realized in the first quarter. The fraction fell slowly in the second quarter, sharply in the third, and then barely at all in the fourth quarter, when 18 percent of such gains were realized. Two of the interpolation schemes show similar total declines, but in nearly equal sized steps. One of these is linear interpolation, which is constrained to constant steps, and the other is model interpolation. Economic interpolation starts with substantially fewer gains in the first quarter (26 percent), has small declines in the next two quarters and then an increase in the fourth quarter to about the share of the first quarter.

In 1973, gross gains start at 25 percent, jump above in the second quarter, fall below in the third, and then return in the fourth. The linear and model interpolations again get the general direction (they are flat) but they miss the jump and dip. Economic interpolation also is essentially flat.

In 1985, gross gains decline from 28 percent of gains in the first quarter to 22 percent in the third before rebounding to 25 percent in the fourth. All three interpolations show steady growth from around 22 percent of net gains in the first quarter to about 28 percent in the fourth.

In 1997, gross gains decline slightly in the second quarter, jump to a peak in the third quarter, and recede slightly in the fourth quarter, still up from the first quarter. The linear and model interpolations again match the annual change in shares, but do so with constant step increases. The economic interpolation also rises steadily, but from a lower starting point to a higher finish.

The pattern for 1998 is similar to that for 1985. Gross gains decline for three quarters and rebound slightly in the fourth. All three interpolated gains rise linearly throughout the year.

Overall, linear and model interpolation match the change in shares between the first and fourth quarters 3 out of 5 times. In none of these years does model interpolation pick up on the quarter-to-quarter fluctuations of gross gains, and of course linear interpolation cannot do so. Economic interpolation is less linear than the other two, but its fluctuations are not synchronized with the fluctuations in gross gains. Thus, the comparison suggests that linear and model interpolation can catch the trend of gains within some years, but that no interpolation catches the quarterly fluctuations of gross gains.

Model interpolation appears to be very similar to linear interpolation, for all of the machinery used to generate it. Perhaps it is simply reflecting the finding of serial correlation of gains among years that by coincidence makes linear interpolation successful. In addition, its error reconciliation process should impart a liner pattern within a year.

The limited number of years for which we have data and the conceptual and practical limitations of the comparisons within a year limit the degree of confidence we can place in our findings.

**Figure D1: Distributions of Gross Gains and 3 Interpolated Net Gains by Quarters in 1962**

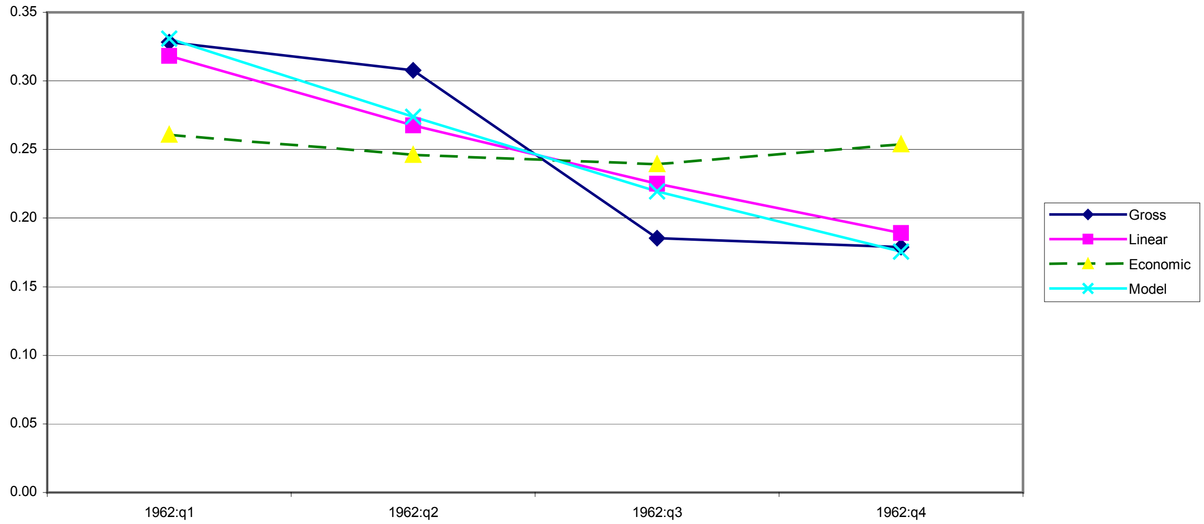



**Figure D2: Distributions of Gross Gains and 3 Interpolated Net Gains by Quarters in 1973**

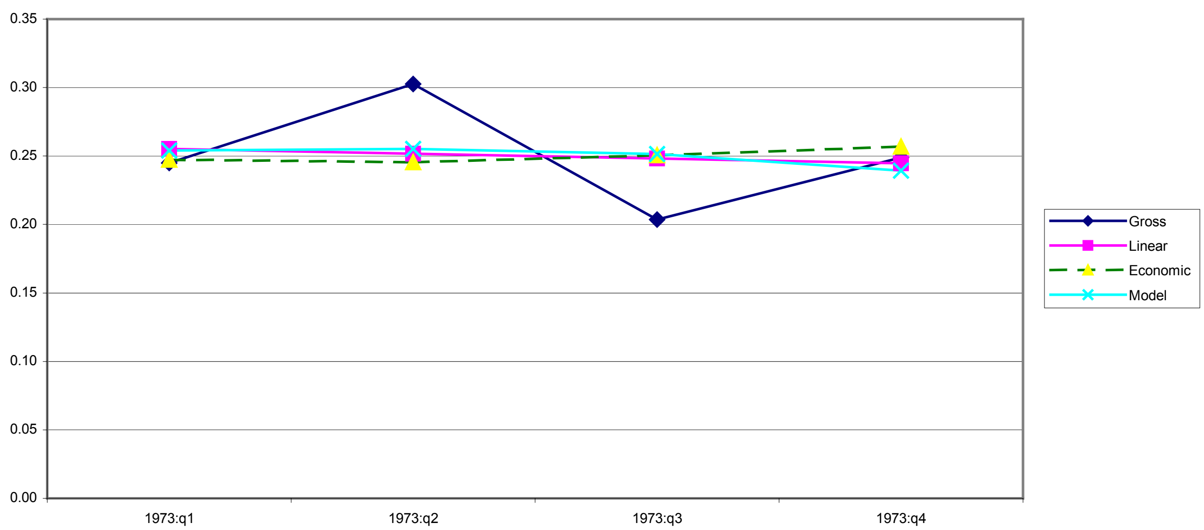


**Figure D3: Distributions of Gross Gains and 3 Interpolated Net Gains by Quarters in 1985**

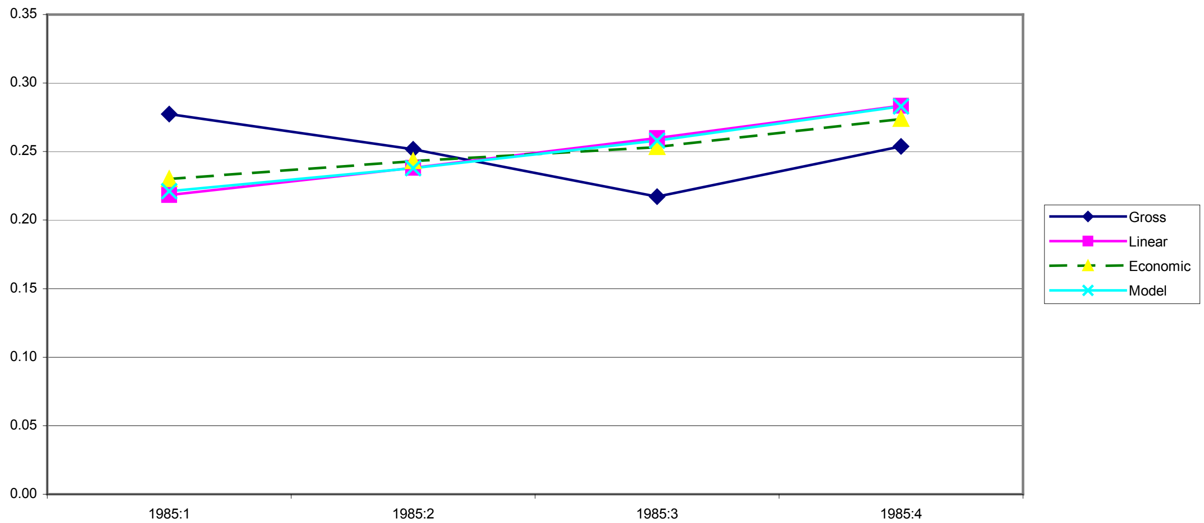



**Figure D4: Distributions of Gross Gains and 3 Interpolated Net Gains by Quarters in 1997**

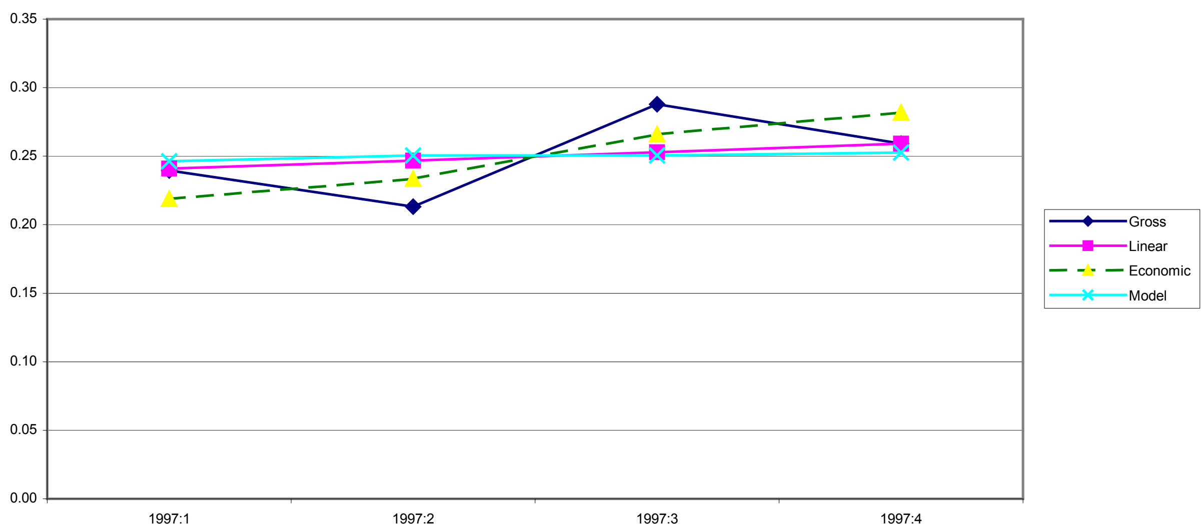


**Figure D5: Distributions of Gross Gains and 3 Interpolated Net Gains by Quarters in 1998**