Can You Improve Upon the GDP Forecasts of Professional Forecasters?

By Dean Croushore*

July 31, 2023

In this paper, I examine the responses by macroeconomic forecasters to various macroeconomic indicators to see whether or not they adjust their forecasts of GDP in an efficient manner. The goal is to investigate, using real-time data, previous research that has found inefficiency in forecasts in several dimensions. The literature suggests that GDP forecasts may not respond appropriately to changes in monetary policy, consumer confidence, retail sales, and claims for unemployment insurance. I use a real-time data set to investigate the response of GDP forecasts to changes in these variables both in-sample and with out-of-sample methods. The results show that for most shocks there is little evidence of forecast inefficiency in-sample. For shocks to monetary policy, in-sample results show evidence of inefficiency in a sample prior to 1984 but the inefficiency disappeared with the Great Moderation. Out-of-sample results show no evidence of inefficiency.

JEL: E37

Keywords: real-time data, forecast efficiency, evaluating forecasts, macroeconomic forecasting, surveys

* Professor of Economics and Rigsby Fellow, University of Richmond, Robins School of Business, 102 UR Drive, University of Richmond, VA 23173, dcrousho@richmond.edu, 804-287-1961. I thank the University of Richmond for summer research support. Thanks also to participants at the Society for Economic Measurement 2023 and the Conference on Computing in Economics and Finance 2023.

I. Introduction

Do forecasters optimally change their forecasts of GDP growth in response to changes in various macroeconomic variables? This question has been answered by a few papers in the literature but mostly using in-sample methods and based on final, revised data. In this paper, we examine the question in a more convincing manner, using real-time data to account more accurately for data revisions, using out-of-sample methods to examine the robustness of in-sample results, and exploring how inefficiency changes over time.

There is a vast literature on the evaluation of forecasts. Point forecasts are evaluated most often by examination of tests of unbiasedness and efficiency. The literature in this area was summed up most clearly by Pesaran and Weale (2006), who suggest that survey forecasts are generally good but show some signs of inefficiency. However, nearly all papers in this literature ignore two facets that I explore in this paper: out-of-sample forecast experiments and the use of real-time data in evaluating forecasts.[1]

The literature suggests that GDP forecasts may not respond appropriately to shocks to a variety of variables. Several papers show that forecasters do not modify their GDP forecasts properly when monetary policy changes. We examine the results of Ball and Croushore (2003) and Rudebusch and Williams (2009) to see how their results hold up when we extend their results to include real-time out-of-sample tests.

In addition to examining how monetary policy may affect forecasts, researchers have also suggested that forecasters may ignore important signals from other variables. Papers including Acemoglu and Scott (1994), Batchelor and Dua (1998), and Souleles (2004) suggest that measures of consumer confidence may improve GDP forecasts. Other papers propose using data on retail sales, such as Koenig, Dolmas and Piger (2003), Zheng and Rossiter (2006), and Diron (2008). Oth-

[1] A recent paper that is complementary to this one using similar methods and evaluates many more variables is Eva and Winkler (2023).

ers suggest that measures related to unemployment insurance claims might also be helpful, including Gavin and Kliesen (2002), Giannone, Reichlin and Small (2008), and Higgins (2014).

Do forecasters already use the information in these variables when they make their output forecasts? To find out, I examine whether or not these variables are related to the forecast errors for GDP, and whether information from these variables could have been used in real time to make better forecasts. So, I use a real-time data set to investigate the response of GDP forecasts to changes in these variables both in-sample and with out-of-sample methods.

II. Data

In this paper, I examine forecasts from the Survey of Professional Forecasters (SPF), which is widely studied.[2] I examine forecasts for real output growth, measured as GNP before 1992 and GDP from 1992 on. The forecasts are made quarterly and the survey asks the respondents to forecast the growth of real output in the current quarter and each of the following four quarters. We examine each of the quarterly forecasts and well as the average output growth forecast over the next four quarters.

Quarterly forecasts for output growth are calculated as in equation (1):

$$y^e_{t,t+h} = (((\frac{Y^e_{t,t+h}}{Y^e_{t,t+h-1}})^4) - 1) \times 100\%, \tag{1}$$

where $h = 0$, 1, 2, 3, and 4, and $Y^e_{t,t+h}$ is the level of the output forecast made at date t for date $t + h$, using data on output through date $t - 1$.

For testing purposes, I compare those forecasts to actual (realized) values, which

[2] The SPF is the only quarterly survey of U.S. macroeconomic forecasters available at no charge, and has been produced on a quarterly basis since 1968. See Croushore and Stark (2019) for a historical discussion of the SPF and the research that uses it.

are calculated as

$$(2) \qquad y^a_{t+h} = ((\frac{Y_{t+h}}{Y_{t+h-1}})^4 - 1) \times 100\%.$$

The definition of "actual" is discussed below. The forecast error is the actual growth rate minus the actual

$$(3) \qquad e_{t,t+h} = y^a_{t+h} - y^e_{t,t+h}.$$

Annual forecasts for output growth are calculated in equation ((4)):

$$(4) \qquad y^{e4}_{t,t+4} = (\frac{Y^e_{t,t+4}}{Y^e_{t,t}} - 1) \times 100\%.$$

Actual values over the same period are

$$(5) \qquad y^{a4}_{t+4} = (\frac{Y_{t+4}}{Y_t} - 1) \times 100\%.$$

Thus forecast errors for annual forecasts are equal to

$$(6) \qquad e4_t = y^{a4}_t - y^{e4}_{t,t+4}.$$

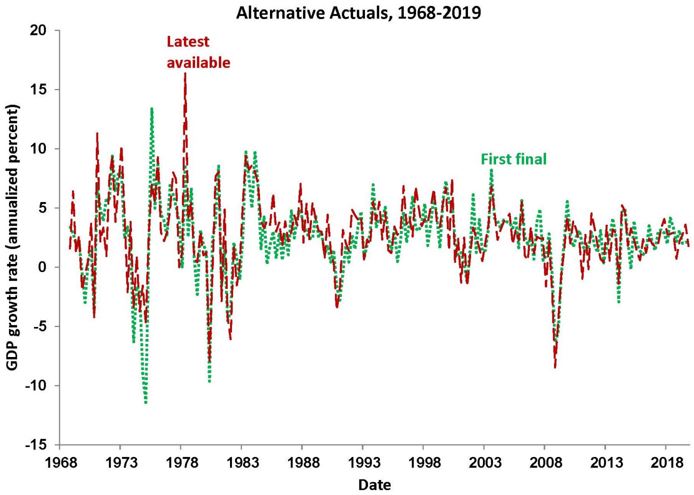
A key question in the forecasting literature is which vintage of the data to use as "actual."[3] There are many alternatives and we explore differences across them, comparing first final (FF) actuals (the release at the end of the third month of the following quarter), to first annual (A) actuals (the release at the end of July of the following year), to pre-benchmark (B) actuals (the last release before a benchmark revision of the National Income and Product Accounts), to latest available (L) actuals (from the latest available vintage of data available when

[3] See Croushore (2011) for a discussion of this issue.

this research started, which was July 2020). I obtain the alternative actuals from the Real-Time Data Set for Macroeconomists (RTDSM), which was created by Croushore and Stark (2001) and made available on the website of the Federal Reserve Bank of Philadelphia. The RTDSM provides information on real output (GNP before 1992, GDP since 1992) and other major macroeconomic variables, as someone standing at the middle of any month from November 1965 to today would have viewed the data. The RTDSM lines up perfectly with the SPF in terms of data availability.

Figure 1 plots GDP growth rates for two of the four alternative actuals, first final and latest available, from 1968Q4 to 2019Q4. You can see that the two series generally move together, but there are quarters when they differ substantially, in one case by over six percentage points. Thus, forecast evaluation conclusions potentially differ significantly depending on the choice of actuals.

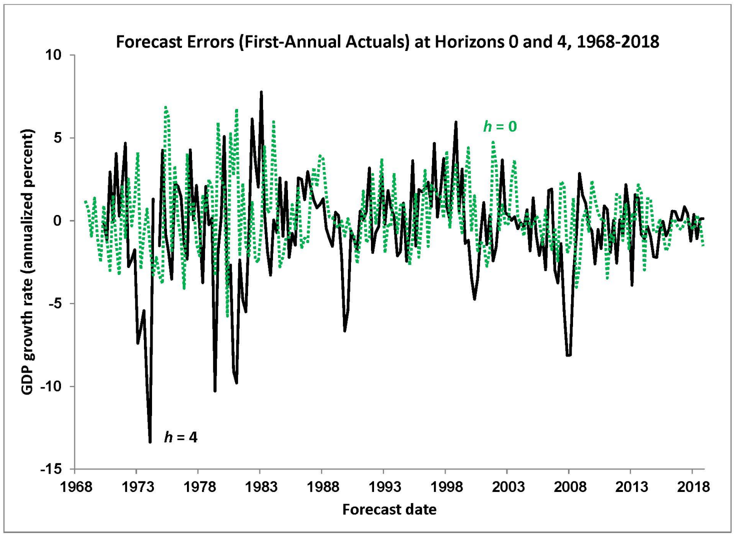
Figure 1. Alternative Actuals



Note: The figure shows the quarterly "actual" GDP growth rates as calculated using Equation (2) based on four alternative concepts: first final, first annual, pre-benchmark, and latest available.

To provide a sense of the size of forecast errors, Figure 2 shows representative forecast errors based on the first-annual concept of actuals at quarterly horizons 0 and 4. The forecast errors are large and volatile, and they change signs frequently, making them difficult to predict. As we might expect, for $h = 4$, the forecast errors are more persistent, as the longer horizon means it takes forecasters longer to respond to shocks.

Figure 2. Alternative Actuals

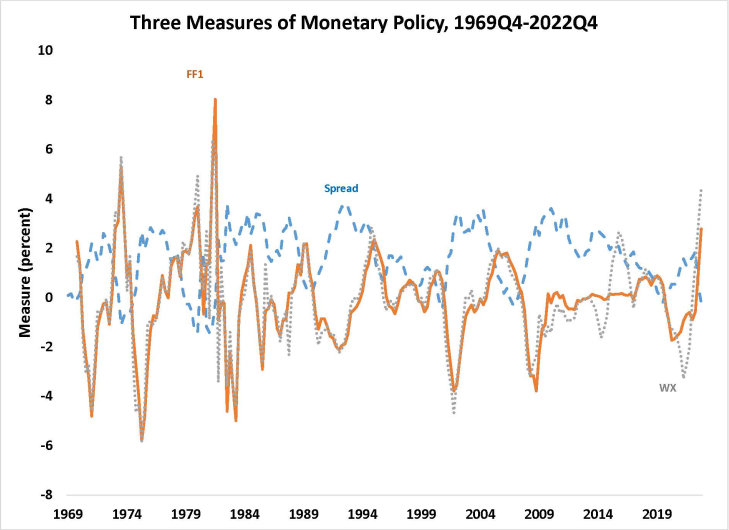


Note: The figure shows the quarterly forecast errors for GDP growth rates as calculated using (3) for two horizons: current quarter ($h = 0$) and four quarters ahead ($h = 4$).

To examine whether measures of monetary policy might be used to improve GDP forecasts, I consider three alternative measures of monetary policy: the yield spread, changes in the real federal funds rate, and changes in the shadow real federal funds rate. For the yield spread, I use the measure of Rudebusch and Williams (2009), which is the interest rate on 10-year Treasury notes minus the interest rate on 3-month Treasury bills, using the constant-maturity series for each security. For the change in the real federal funds rate, I use the Ball and Croushore (2003) measure, which is the average federal funds rate in the previous quarter, minus the expected inflation rate over the coming year in the SPF.[4] However, in the late 2000s, the federal funds rate became constrained by the effective lower bound on interest rates, so changes in the real federal funds rate may not be as useful as a measure of monetary policy as they were before. To remedy that, we use the shadow real fed funds rate of Wu and Xia (2016), which accounts for nontraditional monetary policy tools and creates an effective federal funds rate based on the impact of those tools.[5]

The three measures of monetary policy differ somewhat over time but their major movements are correlated, as you can see in Figure 3.

Figure 3. Three Measures of Monetary Policy



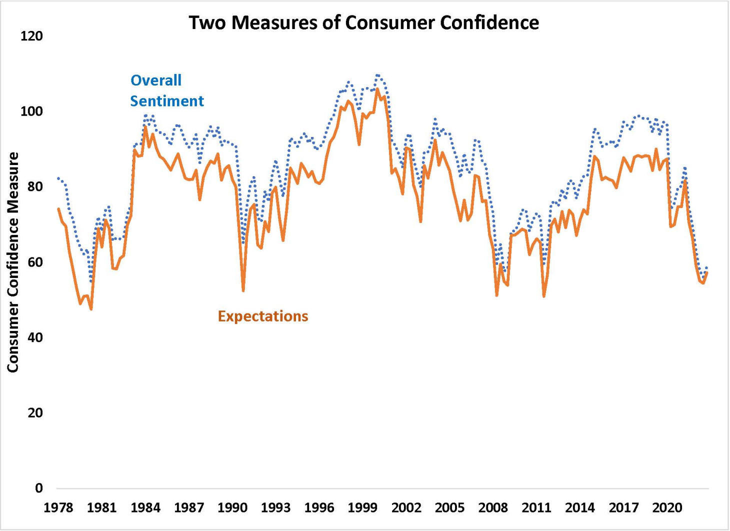
Note: The figure shows the three alternative measures of monetary policy that we use: the term spread between 10-year T-notes and 3-month T-bills (Spread), the change in the real fed funds rate over the previous year ($FF1$), and the change in the Wu-Xia measure of the real fed funds rate over the previous year (WX).

[4] Ball and Croushore (2003) examined alternatives to this measure and found that the results were not sensitive to the proxy used.

[5] Updated data on the Wu-Xia shadow rate are available online at www.atlantafed.org/cqer/research/wu-xia-shadow-federal-funds-rate.

Data on consumer sentiment is measured, as in Batchelor and Dua (1998), by the University of Michigan Indexes of Consumer Expectations. We use both the overall sentiment index and the index of consumer expectations. The former is based on all five survey questions about sentiment, while the latter is based on questions about expectations in the future.[6] The two measures are highly correlated, as Figure 4 shows.

Figure 4. Two Measures of Consumer Confidence



Note: The figure shows the two alternative measures of consumer confidence that we use: the overall Michigan index (Overall Sentiment), and the expectations measure of the Michigan index (Expectations).

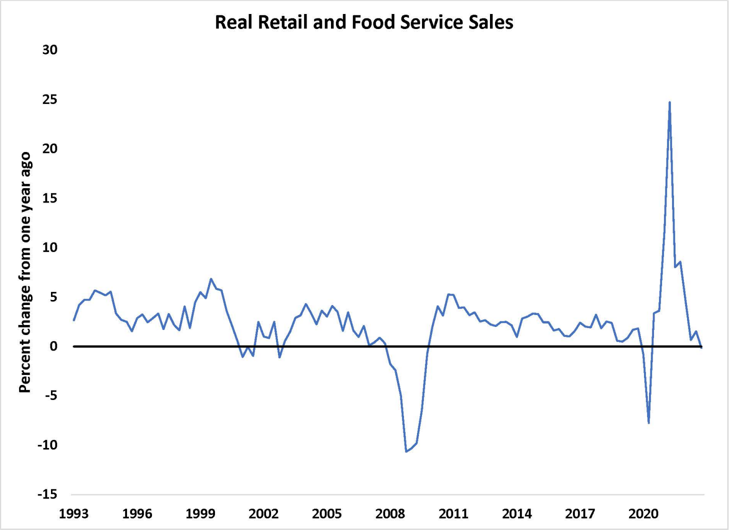
Data on retail sales is measured using real retail and food service sales.[7] I use the quarterly annualized growth rate of the series in our tests. Figure 5 plots the series.

Data on unemployment insurance claims come from two different series, initial

[6] The overall index is FRED variable UMCSENT, while the expectations index comes from the University of Michign web site at data.sca.isr.umich.edu.

[7] The data come from FRED, series RRSFS.

Figure 5. Real Retail and Food Service Sales



Note: The figure shows the level of real retail sales.

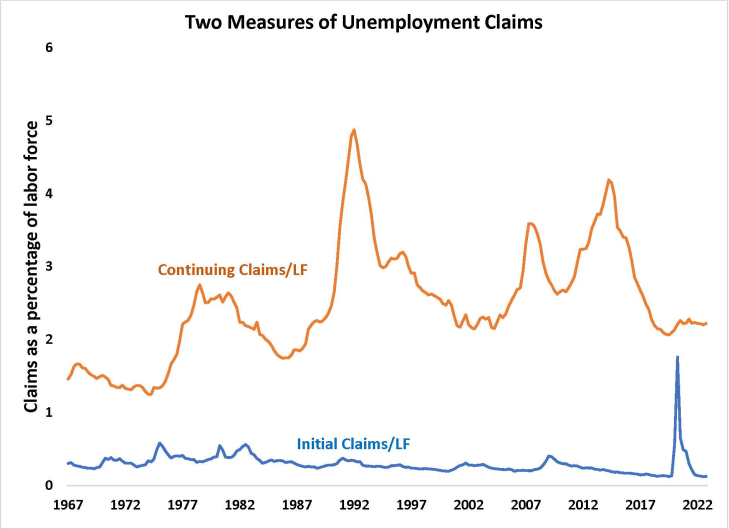
claims and continuing claims. In both cases, I divide the number of claims by the civilian labor force, then take the quarterly annualized growth rate.[8] Figure 6 shows the two series.

III. Forecast Errors and Monetary Policy

In this section, I investigate whether our three measures of monetary policy could be used to improve real output forecasts in real time. We begin with in-sample results to see if the variables are related to GDP forecast errors in the data set, then we move to quasi out-of-sample forecasting to see if the in-sample relationship can be used to improve GDP forecasts. The sample uses all SPF forecasts made from 1968Q4 to 2018Q4, so that four-quarter-ahead forecasts end

[8] Initial claims is FRED series ICSA, continuing claims is FRED series CCSA, and the civilian labor force is FRED series CLF16OV.

Figure 6. Two Measures of Unemployment Insurance Claims



Note: The figure shows the two measures of unemployment insurance claims, both as a percentage of the labor force.

before COVID begins in 2020.[9]

First, we run a regression of each of the forecast errors for the six horizons and four different measures of realizations for each of the three different measures of monetary policy. The regression is simply:

$$e_{t,t+h} = \alpha + \beta MP_t + \epsilon_t, \tag{7}$$

where MP_t is one of the measures of monetary policy available at date t and $e_{t,t+h}$ is a forecast error from equation (3) or (4).

The results are summarized in Table 1.

[9] Including the COVID period in the sample affects the numerical values of the results that follow but does not change the overall conclusions.

Table 1—In-Sample Results for Monetary Policy

Horizon	0	1	2	3	4	1-4
first final	x x x	x M x	M S M	M S S	S S S	x M S
first annual	x x x	M S S	S S S	S S S	S S S	S S S
pre-benchmark	x x x	M S S	S S S	S S M	S S S	S S S
latest-available	x x x	x x x	x x x	S S x	S S S	S S S

Note:

x = measure of monetary policy not statistically significant in regression
S = measure of monetary policy statistically significant in regression ($p < 0.05$)
M = measure of monetary policy marginally statistically significant in regression ($0.05 < p < 0.10$)
First term: yield spread; second term: lagged change in real fed funds rate; third term: lagged change in effective (Wu-Xia) real fed funds rate
The sample uses SPF forecasts from 1968Q4 to 2018Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

In Table 1, we see that about half of all the cases show a statistically significant coefficient in regression equation (7) on the monetary-policy measure, which suggests that forecasters are not using information about monetary policy efficiently in forming their forecasts. In addition, about 15 percent of the slope coefficients also show marginal significance. The coefficients on monetary policy are most often significant at longer horizons, which is consistent with the literature allowing for a lag in the effect of monetary policy on output. In terms of the alternative measures of actuals, the coefficients on monetary policy are more often significant for using first annual or pre-benchmark actuals. Coefficients using the Wu-Xia measure of monetary policy are somewhat less likely to be significant than the other two measures of monetary policy.

Given the in-sample results, we proceed to investigate the possibility of using the regression results from equation (7) to improve upon the SPF forecasts in a simulated real-time out-of-sample exercise; we call this a forecast improvement exercise. Taking the estimated $\hat{\alpha}$ and $\hat{\beta}$, and recalling from equation (3) that $e_{t,t+h} = y^a_{t+h} - y^e_{t,t+h}$, we create, at each date t, an improved forecast $y^f_{t,t+h}$, where

$$y^f_{t,t+h} = y^e_{t,t+h} + \hat{\alpha} + \hat{\beta} MP_t. \tag{8}$$

An important consideration here is to think about the timing of when this type of forecast improvement exercise can be undertaken. It is easiest to use the first-final version of actuals, for which new data become available every three months. If we use first annual data for this exercise, then we get new data just one each year when the annual revision occurs, so there is a longer lag and thus it may be less likely for us to find any ability to improve on the SPF forecasts. So, we try the forecast improvement exercise based on first-final actuals. Using regression equations (7) and (8), we simulate the activity of a real-time forecaster beginning in 1980Q1 and proceeding to 2018Q4, forming improved forecasts at each date based only on the real-time data and past forecast errors available at each date. We collect all the improved forecasts over that period and calculate RMSEs for each different horizon and each different measure of monetary policy. We compare those RMSEs to those of the SPF forecast. Table 2 reports the results, showing the RMSE of the SPF survey and the relative RMSEs (RRMSE) from trying to improve on the forecast using our method. An RRMSE greater than one means the attempt to improve on the SPF forecasts actually made them worse, while an RRMSE less than one means the attempt to improve on the SPF succeeded.[10]

Despite the strong evidence of significant coefficients on monetary policy measures in the in-sample regression, our attempt to improve on the SPF forecasts fails dramatically. The attempt makes the forecasts statistically significantly worse, mostly for current-quarter forecasts; and all the attempts to improve the forecast lead to higher RMSEs. One possibility for this is that the explanatory power of monetary policy may have changed over time, as Rudebusch and

[10] Statistical significance of differences between the surveys is tested using the Harvey, Leybourne and Newbold (1997) modified Diebold and Mariano (1995) test statistic of the corresponding null hypothesis.

Table 2—Real-time Forecast Improvement Exercise with First Final Actuals

Horizon	Survey RMSE	RRMSE Spread	RRMSE FF1	RRMSE WX
0	1.964	1.242	1.242	1.248
		(0.006)	(0.005)	(0.004)
1	2.334	1.137	1.135	1.162
		(0.059)	(0.126)	(0.091)
2	2.407	1.113	1.140	1.220
		(0.078)	(0.188)	(0.136)
3	2.520	1.042	1.071	1.113
		(0.137)	(0.084)	(0.002)
4	2.533	1.067	1.042	1.083
		(0.297)	(0.314)	(0.023)
1-4	2.016	1.269	1.007	1.031
		(0.061)	(0.879)	(0.670)

Note:

The sample period is based on simulations of the forecast improvement exercise from 1980Q1 to 2018Q4. The *p*-values reported in parentheses come from the Harvey, Leybourne and Newbold (1997) modified Diebold and Mariano (1995) test statistic.

Williams (2009) suggest. Because of the Great Moderation, which began around 1984, RMSEs of forecasts generally declined because of the decline in macroeconomic volatility.[11] To see if the responsiveness of forecasts to monetary policy might have changed, we split the sample in 1983Q4 and re-run the in-sample tests. The first subsample runs from 1968Q4 to 1983Q4 and the second subsample runs from 1984Q1 to 2020Q2.

Table 3 shows that the evidence of a relationship between forecast errors and monetary policy is almost nonexistent in the sample that begins in 1984Q1, whereas it is common in the sample that ends in 1983Q4. This result suggests that the nature of the SPF changed in the Great Moderation, at least with respect to monetary policy. It may be that the forecasters did not understand the impact of the expansionary monetary policy pursued by the Fed in the 1970s. But the forecasters seem to have adjusted after that period, or perhaps monetary policy became more sensible.

[11] See Stock and Watson (2002) for details.

Table 3—In-sample results, with split in 1984Q1

Horizon	0	1	2	3	4	1-4
Sample ending 1983Q4						
first final	x x x	S S S	S S S	M S S	S S S	S x x
first annual	x x x	S S S	S S S	S S S	S S S	S S S
pre-benchmark	x x x	S S S	S S S	S S S	S S S	S S S
latest-available	x x x	x S S	x M x	S S M	S S S	S S S
Sample Beginning 1984Q1						
first final	x x x	x x x	x x x	x x x	x x x	x S S
first annual	x M M	x x x	x x x	x x x	x x x	S M x
pre-benchmark	x x x	x x x	x x x	x x x	x x x	M x x
latest-available	x x x	x x x	x x x	x x x	x x x	S x x

Note:

x = measure of monetary policy not statistically significant in regression
S = measure of monetary policy statistically significant in regression ($p < 0.05$)
M = measure of monetary policy marginally statistically significant in regression ($0.05 < p < 0.10$)
First term: yield spread; second term: lagged change in real fed funds rate; third term: lagged change in effective (Wu-Xia) real fed funds rate
The sample periods are 1968Q4 to 1983Q4 in the top table and 1984Q1 to 2018Q4 in the bottom table. Standard errors are adjusted following the Newey and West (1987) procedure.

Are the forecast errors in the pre-Great Moderation period large enough that we can successfully improve on the SPF forecasts in that period? To investigate, we re-run the forecast-improvement exercise for that period. We begin with the forecasts made in 1975Q1 and carry out our exercise through 1983Q4.[12] The results are shown in Table 4.

The results show little evidence of ability to improve on the forecasts in the pre-Great Moderation period. In one of 18 cases, the forecasts are statistically significantly worse at the five-percent level, and in three additional cases they are statistically significantly worse at the ten-percent level. In all 18 cases, the "improved" forecasts have a root-mean-squared error that is more than 10 percent higher. In addition, for the full sample, rolling 10-year windows for the forecast

[12] Note that we are only evaluating over nine years, which is a fairly short sample, and for evaluating forecasts from 1975Q1 we only have six years of forecast data and outcomes on which to estimate equation (7), which may be insufficient to get precise parameter estimates. But this is the best we can do, given the constraints on forecast availability before the Great Moderation.

Table 4—Real-time Forecast Improvement Exercise with First Final Actuals, Pre-Great Moderation

Horizon	Survey RMSE	RRMSE Spread	RRMSE FF1	RRMSE WX
0	3.355	1.437	1.439	1.458
		(0.058)	(0.069)	(0.078)
1	4.159	1.265	1.179	1.163
		(0.143)	(0.166)	(0.151)
2	4.103	1.332	1.286	1.315
		(0.223)	(0.213)	(0.138)
3	4.119	1.115	1.123	1.130
		(0.463)	(0.446)	(0.331)
4	3.995	1.201	1.157	1.159
		(0.287)	(0.557)	(0.572)
1-4	2.356	1.733	1.311	1.294
		(0.001)	(0.262)	(0.175)

(p-values in parentheses)

Note:

The sample period is based on simulations of the forecast improvement exercise from 1975Q1 to 1983Q4. The p-values reported in parentheses come from the Harvey, Leybourne and Newbold (1997) modified Diebold and Mariano (1995) test statistic.

improvement exercise (details not reported here but available on request) made the root-mean-squared errors even worse than reported in Table 2. Thus, even in a period in which in-sample results show a significant relationship between monetary policy and forecast errors, that relationship cannot be exploited to improve upon the SPF forecasts.

IV. Forecast Errors and Consumer Confidence Indexes

In this section, I investigate whether our two measures of consumer confidence could be used to improve real output forecasts in real time. We begin with in-sample results to see if the variables are related to GDP forecast errors in the data set, then we move to quasi out-of-sample forecasting to see if the in-sample relationship can be used to improve GDP forecasts. I follow the same procedure as in the previous section but use the consumer confidence indexes instead of measures of monetary policy. The results are summarized in Table 5.

Table 5—In-Sample Results for Consumer Confidence Indexes

Horizon	0	1	2	3	4	1-4
first final	x x	x x	x x	x x	x x	M x
first annual	x x	x x	x x	x x	x x	x x
pre-benchmark	x x	x x	x x	x x	x x	x x
latest-available	M M	x x	x x	S S	x x	x x

Note:

x = measure of consumer confidence not statistically significant in regression
S = measure of consumer confidence statistically significant in regression ($p < 0.05$)
M = measure of consumer confidence marginally statistically significant in regression ($0.05 < p < 0.10$)
First term: overall sentiment; second term: expectations
The sample uses SPF forecasts from 1968Q4 to 2018Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

In Table 5, we see that about one-sixth of all the cases show a statistically significant coefficient in the regression on the consumer-confidence measure. About half of those are for the one-quarter-ahead forecasts. Splitting the results up into sub-samples before and during the Great Moderation shows similar results to those found for monetary policy, with most of the significant outcomes in the period before 1984, as shown in Tables 6 and 7.

Table 6—In-Sample Results for Consumer Confidence Indexes, Pre-Great Moderation

Horizon	0	1	2	3	4	1-4
first final	x x	x x	x x	x x	x x	S S
first annual	x x	x x	x x	x x	x x	x x
pre-benchmark	x x	x x	x x	x x	x x	x x
latest-available	S S	x x	x x	S M	M x	x x

Note:

x = measure of consumer confidence not statistically significant in regression
S = measure of consumer confidence statistically significant in regression ($p < 0.05$)
M = measure of consumer confidence marginally statistically significant in regression ($0.05 < p < 0.10$)
First term: overall sentiment; second term: expectations
The sample uses SPF forecasts from 1968Q4 to 1983Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

Table 7—In-Sample Results for Consumer Confidence Indexes, During Great Moderation

Horizon	0	1	2	3	4	1-4
first final	x x	x x	x x	x x	x x	x x
first annual	x x	x x	x x	x x	x x	x x
pre-benchmark	x x	x x	x x	x x	x x	x x
latest-available	x x	x x	x x	M M	x x	x x

Note:

x = measure of consumer confidence not statistically significant in regression
S = measure of consumer confidence statistically significant in regression ($p < 0.05$)
M = measure of consumer confidence marginally statistically significant in regression ($0.05 < p < 0.10$)
First term: overall sentiment; second term: expectations
The sample uses SPF forecasts from 1984Q1 to 2018Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

With very marginal results for the period after the Great Moderation begins, it seems unlikely that the forecast improvement exercise will find any ability for the out-of-sample forecasts to show any improvement, but I try, with the results shown in Table 8.

In all cases, the attempt to improve the forecasts leads to a higher RMSFE.

Table 8—Real-time Forecast Improvement Exercise with First Final Actuals, Consumer Confidence

Horizon	Survey RMSE	RRMSE CS	RRMSE ICE
0	1.964	1.257 (0.023)	1.218 (0.026)
1	2.334	1.645 (0.270)	1.206 (0.079)
2	2.407	1.388 (0.268)	1.145 (0.129)
3	2.520	1.394 (0.221)	1.422 (0.247)
4	2.533	1.206 (0.242)	1.365 (0.296)
1-4	2.016	1.490 (0.326)	1.877 (0.331)

(p-values in parentheses)

Note:

The sample period is based on simulations of the forecast improvement exercise from 1980Q1 to 2018Q4. CS = overall consumer sentiment; ICE = index of consumer expectations. The p-values reported in parentheses come from the Harvey, Leybourne and Newbold (1997) modified Diebold and Mariano (1995) test statistic.

For the current-quarter forecasts, the higher RMSFE is statistically significant. Many of the other horizons show even larger RMSFEs but are not statistically significantly larger. Thus, it does not appear possible to improve on the SPF forecasts using data on consumer confidence.

V. Forecast Errors and Retail Sales

In this section, I investigate whether data on retail sales could be used to improve real output forecasts. We begin with in-sample results to see if the variables are related to GDP forecast errors in the data set. I follow the same procedure as in the previous section but use retail sales data instead of the consumer confidence indexes. The results are summarized in Table 9.

Table 9—In-Sample Results for Growth Rate of Real Retail Sales

Horizon	0	1	2	3	4	1-4
first final	x	x	x	x	x	x
first annual	x	x	x	x	x	x
pre-benchmark	x	x	M	x	x	x
latest-available	x	x	x	x	x	x

Note:

x = measure of retail sales not statistically significant in regression
S = measure of retail sales statistically significant in regression ($p < 0.05$)
M = measure of retail sales marginally statistically significant in regression ($0.05 < p < 0.10$)
The sample is based on SPF forecasts from 1968Q4 to 2018Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

In Table 9, we see that almost none of the cases show a statistically significant coefficient in the regression on the growth rate of retail sales. Testing alternative functional forms, such as the growth rate over the previous year (which is a less noisy measure of retail sales growth) show even fewer cases (none) with even marginally significant coefficients. The results may be explained by the fact that the retail sales data begin in 1992, which is after the Great Moderation began. Because the in-sample results show little relationship, there is no point running out-of-sample tests.

VI. Forecast Errors and Claims for Unemployment Insurance

In this section, I investigate whether our two measures of unemployment insurance could be used to improve real output forecasts. I begin with in-sample results to see if the variables are related to GDP forecast errors in the data set. I follow the same procedure as in the previous sections but use the measures of unemployment claims instead of the other measures of monetary policy. The results are summarized in Table 10.

Table 10—In-Sample Results for Unemployment Insurance Claims

Horizon	0	1	2	3	4	1-4
first final	x x	x x	x x	x x	x x	S M
first annual	x x	x x	x x	x x	x x	x x
pre-benchmark	x x	x x	x x	x x	x x	x x
latest-available	x x	x x	x x	x x	x x	x x

Note:

x = measure of unemployment insurance claims not statistically significant in regression
S = measure of unemployment insurance claims statistically significant in regression ($p < 0.05$)
M = measure of unemployment insurance claims marginally statistically significant in regression ($0.05 < p < 0.10$)
First term: initial claims; second term: continuing claims
The sample is based on SPF forecasts from 1968Q4 to 2018Q4. Standard errors are adjusted following the Newey and West (1987) procedure.

In Table 10, we see that only for the four-quarter average forecasts using first-final data is there even a marginally statistically significant coefficient in the regression on either unemployment insurance claims measure.

VII. Summary and Conclusions

We have tested the ability of four different variables (and variations on them) to forecast GDP forecast errors: measures of monetary policy, consumer sentiment, retail sales, and unemployment insurance claims. We found some cases where there appears to be a relationship in-sample between one of the variables and GDP forecast errors. But we find no evidence that such a relationship could be exploited in real time, and most attempts to do so make forecasts worse, not better. We conclude that it is hard to improve on SPF forecasts of GDP growth.

Why might in-sample results show a relationship between macroeconomic variables and forecast errors, but out-of-sample results do not? It may be that forecasters do not recognize the importance of a variable for forecasting until some time passes, so there is an in-sample relationship that is not useful for forecasting for very long. Or, as Cukierman, Lustenberger and Blinder (2018) suggest, a permanent-transitory confusion may lead to in-sample correlations, even if forecasters have rational expectations.

The structure of the forecast-improvement exercises in this paper is based on the in-sample results reported by others in the literature, cited in the Introduction. Some possible future extensions of this work include: (1) Considering instabilities and modifying forecasts to account for changing relationships, as suggest by Rossi and Sekhposyan (2010); (2) Looking at forecasts errors and their relationship to forecast revisions, as in Coibion and Gorodnichenko (2015); and (3) Modifying the forecast-improvement exercises to allow for time-varying coefficients in the equations that test for efficiency.

REFERENCES

Acemoglu, Daron, and Andrew Scott. 1994. “Consumer Confidence and Rational Expectations: Are Agents’ Beliefs Consistent with the Theory?” *The Economic Journal*, 104: 1–19.

Ball, Laurence, and Dean Croushore. 2003. “Expectations and the Effects of Monetary Policy.” *Journal of Money, Credit and Banking*, 35: 473–484.

Batchelor, Roy, and Pami Dua. 1998. “Improving Macro-economic Forecasts: The Role of Consumer Confidence.” *International Journal of Forecasting*, 14(1): 71 – 81.

Coibion, Olivier, and Yuriy Gorodnichenko. 2015. “Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts.” *American Economic Review*, 105(8): 2644–2678.

Croushore, Dean. 2011. “Frontiers of Real-Time Data Analysis.” *Journal of Economic Literature*, 49: 72–100.

Croushore, Dean, and Tom Stark. 2001. “A Real-Time Data Set for Macroeconomists.” *Journal of Econometrics*, 105: 111–130.

Croushore, Dean, and Tom Stark. 2019. “Fifty Years of the Survey of Professional Forecasters.” *Federal Reserve Bank of Philadelphia Economic Insights*, 1–11.

Cukierman, Alex, Thomas Lustenberger, and Allan Blinder. 2018. “The Permanent-Transitory Confusion: Implications for Tests of Market Efficiency and For Expected Inflation During Turbulent and Tranquil Times.” Working Paper, Tel Aviv University.

Diebold, Francis X., and Roberto S. Mariano. 1995. “Comparing Predictive Accuracy.” *Journal of Business and Economic Statistics*, 13: 253–263.

Diron, Marie. 2008. “Short-Term Forecasts of Euro Area Real GDP Growth: An Assessment of Real-Time Performance Based on Vintage Data.” *Journal of Forecasting*, 27: 371–390.

Eva, Kenneth, and Fabian Winkler. 2023. “A Comprehensive Empirical Evaluation of Biases in Expectation Formation.” Working Paper, Federal Reserve Board.

Gavin, William T., and Kevin L. Kliesen. 2002. “Unemployment Insurance Claims and Economic Activity.” *Federal Reserve Bank of St. Louis Review*, 15–27.

Giannone, Domenico, Lucrezia Reichlin, and David Small. 2008. “Nowcasting: The Real-Time Informational Content of Macroeconomic Data.” *Journal of Monetary Economics*, 55: 665–676.

Harvey, David S., Stephen J. Leybourne, and Paul Newbold. 1997. “Testing the Equality of Prediction Mean Squared Errors.” *International Journal of Forecasting*, 13: 281–291.

Higgins, Patrick. 2014. “GDPNow: A Model for GDP ‘Nowcasting’.” *Federal Reserve Bank of Atlanta Working Paper 2014-7.*

Koenig, Evan, Sheila Dolmas, and Jeremy Piger. 2003. “The Use and Abuse of ‘Real-Time’ Data in Economic Forecasting.” *Review of Economics and Statistics*, 85: 618–628.

Newey, W.K., and K.D. West. 1987. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, 55: 703–708.

Pesaran, M. Hashem, and Martin Weale. 2006. “Survey Expectations.” In *Handbook of Economic Forecasting.* , ed. G. Elliott, C. W. J. Granger and A. Timmermann, 715–776. Elsevier.

Rossi, Barbara, and Tatevik Sekhposyan. 2010. "Have Economic Models' Forecasting Performance for US Output Growth and Inflation Changed Over Time, and When?" *International Journal of Forecasting*, 26(4): 808–835.

Rudebusch, Glenn D., and John C. Williams. 2009. "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve." *Journal of Business and Economic Statistics*, 27(4): 492–503.

Souleles, Nicholas S. 2004. "Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence from the Michigan Consumer Sentiment Surveys." *Journal of Money, Credit and Banking*, 36(1): 39–72.

Stock, James H., and Mark W. Watson. 2002. "Has the Business Cycle Changed and Why?" *NBER Macroeconomics Annual*, 17: 159–218.

Wu, Jing Cynthia, and Fan Dora Xia. 2016. "Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound." *Journal of Money, Credit and Banking*, 48: 253–291.

Zheng, Isabel Yi, and James Rossiter. 2006. "Using Monthly Indicators to Predict Quarterly GDP." Working Papers 06-26, Bank of Canada.