Identifying Monetary Policy Shocks: A Natural Language Approach*

S. Borağan Aruoba — University of Maryland

Thomas Drechsel — University of Maryland, CEPR

May 10, 2024

Abstract

We develop a novel method for the identification of monetary policy shocks. By applying natural language processing techniques to documents that Federal Reserve staff prepare in advance of policy decisions, we capture the Fed's information set. Using machine learning techniques, we then predict changes in the target interest rate conditional on this information set and obtain a measure of monetary policy shocks as the residual. We show that the documents' text contains essential information about the economy which is not captured by numerical forecasts that the staff include in the same documents. The dynamic responses of macro variables to our monetary policy shocks are consistent with the theoretical consensus. Shocks constructed by only controlling for the staff forecasts imply responses of macro variables at odds with theory. We directly link these differences to the information that our procedure extracts from the text over and above information captured by the forecasts.

Keywords: Monetary policy; Federal Reserve; Greenbook; Natural Language Processing; Machine learning.
JEL Classification: C10; E31; E32; E52; E58.

---

*We would like to thank Juan Antolin-Diaz, Michael Bauer, Gabriel Chodorow-Reich, Pierre De Leo, Burcu Duygan-Bump, Marty Eichenbaum, Simon Freyaldenhoven, Friedrich Geiecke, Yuriy Gorodnichenko, Tarek Hassan, Jonathon Hazell, Kilian Huber, Matteo Iacoviello, Diego Kaenzig, Anil Kashyap, Guido Kuersteiner, Michele Lenza, Kevin Lee, Xiang Li, Michael McMahon, Vitaly Meursault, Emi Nakamura, Ivan Petrella, Chris Redl, Frank Schorfheide, Eric Swanson, Minchul Shin, Jon Steinsson, Lumi Stevens, Silvana Tenreyro, Fabian Winkler, Harald Uhlig, Chris Wolf, Johannes Wieland and Jonathan Wright, as well as seminar and conference participants at the NBER Summer Institute, University of Chicago, University of Copenhagen, University of Hamburg, Indiana University, the Federal Reserve Board, the Federal Reserve Banks of Cleveland, Dallas, and Philadelphia, the Bank of England, the BIS, the ECB, the Reserve Bank of Australia, the Reserve Bank of New Zealand, the Central Bank of Chile, Sveriges Riksbank, Norges Bank, Danmarks Nationalbank, the IMF, the IWH Halle, FGV EESP Sao Paulo, the Barcelona Summer Forum, the Glasgow Workshop on Recent Advances in Econometrics, the LSE Conference in Honor of Wouter Den Haan, and the DC-MD-VA Econometrics Workshop for helpful comments. Eugene Oue, Danny Roth, Mathias Vesperoni, Eric Youmans and Jialing Zhang provided excellent research assistance. Contact: Department of Economics, University of Maryland, Tydings Hall, College Park, MD 20742, USA; aruoba@umd.edu and drechsel@umd.edu. *We make our estimated monetary policy shocks, the Fed sentiment time series, the FOMC composition data set, as well as the Greenbook forecast errors available to other researchers at http://econweb.umd.edu/-drechsel/research.html.*

1 Introduction

To study how monetary policy affects the economy, macroeconomists isolate changes in interest rates that are not systematic responses to economic conditions, but occur in a nonsystematic way. This paper proposes a novel method for the identification of such monetary policy shocks. Our method is based on the idea that exogenous movements in interest rates are the difference between actual and systematic changes in the central bank's target interest rate, and that systematic changes can be estimated using measures of the central bank's information set. We propose an identification approach that captures the large amount of high-quality numerical and textual information in documents that economists at the Federal Reserve prepare for Federal Open Market Committee (FOMC) meetings.

Our idea is inspired by Romer and Romer (2004) who propose running a regression of the change in the target Federal Funds Rate (FFR) on forecasts of inflation, output and unemployment, and retrieving a measure of monetary policy shocks as the residual. The forecasts are contained in the *Greenbook* created by Fed staff economists for the FOMC and are intended to serve as a proxy for the FOMC's information about the economy. Instead of using only a handful of forecasts, our methodology converts the *natural language* in the documents prepared by staff for the FOMC into many economically meaningful time series, which capture the much larger information that the FOMC has available ahead of policy decisions. We orthogonalize changes in the FFR with respect to all available forecasts and these text-based time series in order to extract a measure of monetary policy shocks. We implement our approach with natural language processing and machine learning techniques.

We show econometrically that the language in the documents produced by Fed staff contains information that the staff's numerical forecasts do not incorporate. Including *at least some* text-based information is required for identification, as this correctly captures the Fed's mean expectation of the variables of interest, a condition for the original Romer and Romer (2004) approach to work correctly (Cochrane, 2004). Beyond this requirement, we argue that to estimate monetary policy shocks that are exogenous to all available information and can be used to study many macro variables, it is beneficial to control for *a large set* of information.

Our identification procedure estimates monetary policy shocks as the residuals from a prediction of changes in the FFR target using *(i)* all numerical forecasts in the documents that Fed economists prepare for the FOMC; *(ii)* textual information in the documents converted into time series, including lagged information from documents prepared for previous meetings; and *(iii)* nonlinearities in *(i)* and *(ii)*. *(i)* includes the original forecasts used by Romer and Romer (2004) but we expand the set to include additional variables that Fed economists provide forecasts for, such as industrial production, housing and government spending. To obtain *(ii)*, we first identify the most commonly mentioned economic terms in the documents. This results in a set of 296 single or multi-word expressions, such as "inflation," "economic activity" or "labor force participation." We then construct sentiment indicators that capture the degree to which these concepts are associated with positive or negative language using a dictionary. The documents we use are carefully crafted by Fed economists, with precise wording and consistency in language over time, so this type of natural language processing is particularly applicable. Our collection of sentiment time series paints a rich picture of the historical assessment of economic conditions by Fed staff.

A regression of FFR target changes on *(i)*, *(ii)* and *(iii)* is infeasible given that there are more regressors than observations. To overcome this issue, we resort to machine learning techniques. Specifically, we employ a ridge regression to predict changes in the FFR target using our large set of forecast and sentiment regressors. This choice is guided by recent insights about alternative types of machine learning methods for economic data (Giannone, Lenza, and Primiceri, 2022). A ridge regression minimizes the residual sum of squares plus an additional term that penalizes squared deviations of each coefficient from zero to achieve shrinkage. We select the ridge penalty parameter using k-fold cross-validation, a standard way in the machine learning literature to validate a model's predictive ability in alternating subsets of the data.

We assess the informational value added of our sentiment indicators. In a discussion of Romer and Romer (2004), Cochrane (2004) points out that for the purpose of studying the effects of monetary policy on a given variable, it would be enough to orthogonalize FFR changes with respect to the Fed's forecast of that variable alone. The logic is that the forecast should incorporate all other relevant information efficiently. The argument is correct when Greenbook forecasts exactly correspond to the Fed's *conditional mean* expectation. We present anecdotal and econometric evidence that the Fed staff produce *modal* forecasts, in combination with a description of expected changes around the mode in words, such as discussions of asymmetric tail risks. Furthermore, Greenbook forecasts assume a specific future path of the policy rate, which alters the conditioning set (Faust and Wright, 2008). To show econometrically that the language content of the documents reflects valuable information beyond what is incorporated in the numerical staff forecasts, we demonstrate that our sentiment indicators predict the errors of these forecasts. The forecast error for the unemployment rate is predictable with one of the 296 sentiment indicators alone, across horizons and to an economically significant degree. In consequence, it is essential that FFR target changes are orthogonalized with respect to at least some of the textual information, even for the original Romer and Romer (2004) method to correctly fulfill the purpose of recovering the response of unemployment to a monetary policy shock. Using natural language and machine learning, we include a much wider set of information to create a plausibly exogenous "all purpose" monetary policy shock time series that can be used to study the effects any economic variable, including economic variables for which forecasts are not produced by the Fed.

The application of our method yields three sets of findings. First, we examine the relative contribution of systematic and exogenous variation in the FFR target since 1982. A linear regression that contains only numerical forecasts for output, inflation and the unemployment rate yields an R^2 of around 0.5, suggesting that half of the variation in the FFR target is attributed to systematic policy, while the other half is included in the monetary policy shock. The R^2 of our ridge regression is 0.94, implying that the exogenous component of FFR changes is reduced almost ten fold from 50% to 6%, when a larger set of forecasts, text-based sentiments, as well as nonlinearities are included. While our analysis of forecast error predictability already supports the view that more information about the systematic component of monetary policy should be included, a high R^2 is also economically appealing. Macroeconomists typically think of monetary policy decisions to be largely taken systematically, with a small role for exogenous shocks, as discussed for example by Leeper, Sims, and Zha (1996).

Second, we provide an interpretation of what our estimated monetary policy shocks capture. We do so by closely analyzing the discussion that took place among the FOMC participants in meetings where the estimated shock is large in magnitude. It turns out that in these episodes the FOMC made decisions based on considerations not directly related to the staff's analysis, in an unsystematic manner. For example, in the November 1994 meeting, the material prepared by the staff economists is supportive of a 50 basis point rate hike. However, in the FOMC meeting Chairman Alan Greenspan advocates a 75 basis point hike, arguing that "a mild surprise would be of significant value", in order to emphasize long-run credibility. Our procedure estimates almost the entire 25 basis point difference to be a nonsystematic contractionary shift in policy.

Alongside our interpretation of monetary policy shocks, we provide a comparison with an alternative measure of nonsystematic changes in monetary policy extracted from high-frequency (HF) surprises in interest rate futures around FOMC announcements as computed by Swanson (2021). There is a positive correlation between our shock measure and his HF identified surprises to the FFR, and our method increases this correlation relative to the original Romer-Romer approach. One practical advantage of our procedure compared to using high-frequency surprises is that we obtain shocks over a longer time period, while the availability of futures data restricts HF measures to start around the early 1990's.

Third, we study impulse response functions (IRFs) of macro variables to our monetary policy shocks. We include our monetary policy shock series in a state-of-the-art Bayesian vector autoregression (BVAR) as an external instrument.[1] We find that a monetary policy tightening leads to a reduction in economic activity, a fall in the price level, an increase in bond premia and a decline in stock prices. These findings are in line with what economic theory predicts. Notably, following a tightening there is a relatively swift decline in real output, while the reduction in the price level builds up sluggishly over time. We also show that IRFs resulting from shocks computed using the original Romer-Romer methodology lead to responses not in line with the theoretical consensus. We discuss potential interpretations, in particular by drawing a direct connection to our findings on the insufficient information content of the Greenbook forecasts. This allows us conclude that natural language processing and machine learning deliver a cleanly identified estimate of monetary policy shocks.

Finally, in an extension we demonstrate how our method can be applied to extract monetary policy shocks from the most recent FOMC meetings. The Tealbooks and associated forecasts are available to the public only with a 5-year lag, so our preferred ridge regression cannot include the latest FOMC decisions.[2]

[1] We also use local projections (Jordà, 2005) as an alternative methodology and find similar results.

[2] "Tealbook" is a recent label for the documents that the staff prepares for FOMC meetings. We sometimes use the terms Tealbook and Greenbook interchangeably, but in description of our

However, the Beigebooks are publicly available prior to contemporaneous FOMC meetings. These summarize regional economic conditions in each Federal Reserve district, and are already part of the set of documents that we process over our main estimation sample. For the most recent FOMC meetings, constructing sentiment indicators based on the Beigebooks alone provides at least a limited proxy for the FOMC's information set. Leveraging the Beigebooks is not possible in the original Romer and Romer (2004) approach, as they do not contain any numerical forecasts and are composed of only textual information.

Literature. We contribute to three branches of research. The first is the literature that seeks to identify monetary policy shocks, most notably the seminal work of Romer and Romer (2004). Their method is still widely used, see e.g. Tenreyro and Thwaites (2016) and Wieland and Yang (2020).[3] There is a wide array of other approaches to identifying monetary policy shocks, as surveyed by Ramey (2016). One approach uses structural vector autoregressions (SVARs) identified in different ways.[4] Another approach is based on HF surprises in market interest rates, e.g. Gürkaynak, Sack, and Swanson (2005), Gertler and Karadi (2015), Swanson (2021) and Bauer and Swanson (2023). We compare our shock measures with those extracted from HF interest rate surprises. We contribute to the literature on identifying monetary policy shocks by applying natural language processing and machine learning to achieve identification through a large set of information in economic data and text. We show that including the additional text-based information is critical for identification.[5]

The second branch of research we contribute to is a fast-growing literature that applies textual analysis to documents produced by the Fed. Hansen, McMahon, and Prat (2018) show that communication in the FOMC changes after public transparency increased in the early 1990's. Similar to us, Sharpe, Sinha, and Hollrah (2020) carry out sentiment analysis using documents produced by

method we precisely define which types of documents we process over which sample periods.

[3] Bachmann, Gödl-Hanisch, and Sims (2022) suggest summarizing the Fed's information set using forecast errors. The Romer-Romer methodology has also been applied to other countries, e.g. Cloyne and Hürtgen (2016) use it for the UK and Holm, Paul, and Tischbirek (2021) for Norway.

[4] Identification in SVARs is obtained e.g. through zero restrictions (Christiano, Eichenbaum, and Evans, 1999), sign restrictions (Uhlig, 2005), or narrative sign restrictions (Antolin-Diaz and Rubio-Ramirez, 2018). Coibion (2012) compares SVAR approaches to that of Romer and Romer (2004).

[5] Our emphasis on a large information set has parallels to Bernanke, Boivin, and Eliasz (2005) who incorporate many time series in a factor-augmented VAR (FAVAR), but do not consider text.

Fed economists and a pre-defined dictionary. Different from us, these authors construct a single sentiment index rather than sentiments for individual economic concepts (or 'aspect-based' sentiments). Shapiro and Wilson (2021) analyze FOMC transcripts, minutes, and speeches in order to draw inference about central bank objectives.[6] Cieslak and Vissing-Jorgensen (2020) employ textual analysis on FOMC documents to understand if monetary policy reacts to stock prices.[7] Cieslak et al. (2021) construct text-based measures of policy makers' uncertainty. None of the aforementioned studies identify monetary policy shocks, which is the goal of our methodology. Two complementary papers use textual analysis on Fed documents for purposes similar to ours. Handlan (2020) estimates a "text shock" that separates the difference between forward guidance and current assessment of the FOMC in driving FFR futures since 2005. We instead estimate a more conventional series of monetary policy shocks over several decades. Ochs (2021) uses publicly available FOMC documents to extract surprise changes in monetary policy from the point of view of private agents. We orthogonalize interest rates changes with respect to the central bank's information set as captured by the documents prepared internally for the FOMC. In that sense, our procedure is closer to the original Romer and Romer (2004) approach to estimating monetary policy shocks. Natural language processing and machine learning enable us to capture the central bank's information set in a more comprehensive way and to a degree that we show is required for identification.

The third branch of research we contribute to studies the Fed's Greenbook forecasts, including Romer and Romer (2000), Faust and Wright (2008, 2009), and Nakamura and Steinsson (2018). This literature points to the high quality of the Greenbook forecasts and the Fed's informational advantage over the private sector. We emphasize that Greenbook forecasts are best interpreted as modal, and text-based explanations by staff economists incorporate information about asymmetric risks. We show that as a result there is useful information, expressed in words, that can explain Greenbook forecast errors on average.

[6] Acosta (2022) studies how the FOMC responded to calls for transparency. A further paper using Fed is Doh, Song, and Yang (2022).

[7] Several others study the reverse, whether financial markets react to Fed language. Gardner, Scotti, and Vega (2021) study the response of equity prices to FOMC statements using sentiment analysis. Gorodnichenko, Pham, and Talavera (2023) use deep learning techniques to capture emotions in FOMC press conference.

Structure. Section 2 lays out our method. Section 3 shows why the sentiment indicators contain essential information. Section 4 discusses the results of our identification procedure, including the contribution of systematic policy and our estimated shocks. Section 5 presents our results on the responses of macro variables to monetary policy shocks. In Section 6 we apply a our method to the Fed's most recent interest rate hikes using the Beigebook only. Section 7 concludes.

2 A new method to identify monetary policy shocks

This section first provides the motivation for our approach, explains the relevant institutional setting, and lays out the main idea of our methodology. It then gives an in-depth description of the full shock identification procedure.

2.1 Motivation, institutional setting, and main idea

Definition of monetary policy shocks. The challenge of studying how monetary policy affects the economy is that policy is set endogenously, by taking current economic conditions and the outlook for the economy into account. An influential literature has addressed this challenge by isolating changes in monetary policy that are orthogonal to the information that policy makers react to. In this line of work, the central bank is assumed to set its policy instrument s_t, according to a rule

$$s_t = f(\Omega_t) + \varepsilon_t, \tag{1}$$

where Ω_t is the information set of the central bank, $f(\cdot)$ is the systematic component of monetary policy, and ε_t is the monetary policy shock, or the nonsystematic component. The systematic component of policy is endogenous, so the only way to understand the causal effect of monetary policy on the economy is to consider changes in ε_t. The formalization of the endogeneity challenge in equation (1) is the explicit or implicit starting point of most studies in the literature.

The Romer-Romer approach. One approach to estimating monetary policy shocks, following Romer and Romer (2004), is to run the linear regression

$$\Delta i_t = \alpha + \beta i_{t-1} + \boldsymbol{\gamma} \boldsymbol{X}_t + \varepsilon_t^{RR}, \tag{2}$$

where i_t is the FOMC's FFR target, and $\boldsymbol{X}_t$ contains economic forecasts that the FOMC has at its disposal at time t, where time evolves at meeting frequency. In their original work, these include forecasts of output growth, inflation, and the unemployment rate of various horizons, and enter in both levels and changes. Running regression (2) results in the residuals $\hat{\varepsilon}_t^{RR}$, which provide an empirical measure for ε_t in (1).

Two key assumptions underlie this approach. First, the forecasts included in $\boldsymbol{X}_t$ need to be a good proxy for the information set Ω_t. The FOMC reviews a large amount of information on the economic and financial conditions of the US economy, prepared by staff economists as part of different documents. These documents contain numerical forecasts but also many pages of text. The numerical forecasts can by themselves provide a suitable proxy (or "sufficient statistic") for the information set. For this to be the case they need to correspond to the FOMC's mean expectation conditional on incorporating all the other information efficiently (Cochrane, 2004). The second assumption is that the mapping $f(\cdot)$ from the information to decisions is well captured by a linear relationship.

We revisit the first assumption by the enhancing the proxy for the information set Ω_t. The documents produced around FOMC meetings contain a vast amount of high-quality information, both in textual form and in the form of numerical forecasts. They are crafted by the Fed staff in a careful and analytical manner, with consistency in language over time, so natural language processing (NLP) techniques are well suited for extracting valuable information from them. Importantly, we will show that the language with which Fed economists describe the subtleties around the economic outlook provides valuable information *beyond* what is contained in purely numerical predictions.

We revisit the second assumption by examining the presence of nonlinearities in $f(.)$. We do so by including higher order terms in our econometric counterpart of (1). Since considering numerical forecasts, text-based information, as well as nonlinearities requires us to include a large number of variables on the right hand side of a regression model, we apply machine learning (ML) techniques to cope with this dimensionality problem. We then estimate monetary policy shocks as the residuals from a prediction of changes in the FFR using a large amount of numerical and textual information.

2.2 Step-by-step description of our method

Our procedure to estimate monetary policy shocks consists of four steps. First, we process the text of relevant FOMC meeting documents. Second, we identify frequently discussed economic concepts in these documents. Third, we construct sentiment indicators for each economic concept. Fourth, we run a regression that includes sentiment indicators and numerical forecasts, linearly and nonlinearly.

Step 1: Process FOMC documents

In FOMC meetings, scheduled 8 times per year, the committee discusses monetary policy decisions.[8] We first retrieve historical documents associated with FOMC meetings from the website of the Federal Reserve Board of Governors. We start with the meeting on October 5, 1982, to capture the period over which the Fed targeted the FFR as their main policy instrument, according to Thornton (2006). Coibion (2012) points out that including the earlier period in which the FOMC targeted nonborrowed reserves is problematic, as the FFR displays extremely large swings. Most of FOMC meeting documents are available with a 5-year lag, the latest document used in our analysis is for the last FOMC meeting of 2016.

For each FOMC meeting, several documents are available. We include the following: *Greenbook 1 and 2* (until June 2010), *Tealbook A* (after June 2010), *Redbook* (until 1983), *Beigebook* (after 1983).[9] We focus on these documents to capture the Fed's information set just prior to the meeting. We do not include minutes, transcripts or announcements because these might capture the decision process rather than the information set of policy makers going into the meeting. Our choice results in 772 PDF files for 276 meetings (630 files for 210 meetings before the zero lower bound), containing tens of thousands of pages of text and numbers.

We read each document into a computer and process it as follows. We remove stop words ("the", "is", "on", etc.); remove numbers (that are not forecasts, e.g.

[8]There are also unscheduled meetings or conference calls during which the FOMC makes policy decisions. Since no new documents are prepared for these meetings, they do not contribute to the monetary policy shock time series that we estimate.

[9]The *Greenbooks*, later replaced by *Tealbook A*, contain staff analysis for the US economy. We exclude the *Bluebook*, later replaced by *Tealbook B*, as these contain different hypothetical scenario analyses, where outcomes conditional on alternative policy actions are described, and which we judged might obfuscate our sentiment extraction. The *Redbooks* (until 1983) / *Beigebooks* (from 1983) discuss economic conditions for each Federal Reserve district. We use the Beigebooks by themselves in our analysis of recent Fed meetings in Section 6.

Figure 1: ECONOMIC CONCEPTS MENTIONED FREQUENTLY IN FOMC DOCUMENTS

Notes. Word cloud of the 75 most frequently mentioned economic concepts in documents prepared by Federal Reserve Board economists for FOMC meetings between 1982 and 2016. The size of concept reflects the frequency with which it occurs across the documents.

dates, page numbers); remove "erroneous" words. We then retrieve *singles*, *doubles* and *triples*. Singles are individual words. Doubles and triples are joint expressions not interrupted by stop words or sentence breaks. For example, "... consumer price inflation ..." is a triple, and also gives us two doubles ("consumer price" and "price inflation") and three singles ("consumer", "price" and "inflation"). "... inflation and economic activity ..." gives us three singles and one double. "... for inflation. Activity on the other hand..." only gives us three singles ("inflation", "activity" and "hand"). For the 276 meetings there are roughly 18,000 singles, 450,000 doubles, and 600,000 triples. For comparison, the Oxford English dictionary has roughly 170,000 single words. We then calculate the frequency at which each single, double and triple occurs for each meeting date and each document.

Step 2: Identify frequently used economic concepts

We rank all singles, doubles and triples from Step 1 by their total frequency of occurrence over the whole time period. We then start from the most frequent ones, move downwards and select those singles, doubles and triples that are economic concepts, such as "credit", "output gap", or "unit labor cost".[10] Sometimes there are economic concepts that overlap across singles, doubles and triples. For

[10]Both authors went through this selection independently and discussed disagreements. When moving down the frequency ranking, we stop at a lower bound, e.g. one mention per meeting on average. We discuss advantages of imposing judgmental restrictions at the end of Section 2.

Table 1: EXAMPLES OF WORDS ASSOCIATED WITH POSITIVE AND NEGATIVE SENTIMENT

Positive sentiment	**Negative sentiment**
adequate	adversely
advantage	aggravate
benefit	bad
boost	burdensome
confident	collapse
conducive	concerning
desirable	decline
diligent	deficient
encouraging	eroded
excellent	exacerbate
...	...

Notes. Selected examples of words that are classified as expressing positive or negative sentiments in our augmented version of the dictionary of Loughran and McDonald (2011). The total number of classified words is 2,882.

example, should "commercial real estate" be an economic concept or just "real estate" or both separately? To address this, we follow a precise selection algorithm that we describe in Appendix A. Our selection procedure results in 296 economic concepts. Figure 1 shows a word cloud for the 75 most frequent economic concepts, where the size of the concepts reflects its frequency across the documents.

Step 3: Construct sentiment indicators for each economic concept

For each of the 296 individual economic concepts, we apply a method to capture the *sentiment* surrounding them, inspired by Hassan, Hollander, van Lent, and Tahoun (2022). For each occurrence of a concept in a document, we check whether any of the 10 words mentioned before and after the concept's occurrence are associated with positive or negative sentiment.[11] This classification builds on the dictionary of positive and negative terms in Loughran and McDonald (2011). This is a widely used dictionary in the literature, which is especially constructed for financial text, so it should already be reasonably suitable for the economic content discussed in the Fed documents. For our application we make several modifications to this dictionary.[12] Based on our augmented dictionary,

[11] Further below, we explore robustness with sentiment indicators constructed using an alternative distance of 5 words. We also show that constructing sentiments based on positive and negative words within the same sentence, rather than inside a 10-word window, yields time series that are highly correlated with the ones we use. See Appendix C for two examples.

[12] We modify the Loughran and McDonald (2011) dictionary along two dimensions. First, we enhance the list of words by adding terms typical for Fed language, such as *tightening*. We also add some more variations of existing terms, for example the original dictionary contains *boom* and *booming*, and we add *booms* and *boomed*. Second, we remove some terms, either because they

each positive word then adds a score of +1 and each negative word a score of -1 towards the sentiment of the concept. Table 1 provides a few examples of positive and negative words. For each of our concepts, we then sum up the sentiment scores within the documents associated with an FOMC meeting, and scale by the total number of words in the documents to obtain a sentiment indicator. The final product of this procedure is a sentiment indicator time series for each economic concept, where the time variation is across FOMC meetings. For the purpose of entering these indicators in a regression, we also standardize all indicators.

Figure 2 presents the sentiment indicators for selected economic concepts. These indicators are standardized, but not otherwise smoothed or filtered. They clearly display meaningful variation. For example, Panel (a) shows that the sentiment surrounding "economic activity" falls sharply in recessions. Furthermore, comparisons across concepts reveal meaningful information about the Fed economists' view on the nature of different recessions. For example, the sentiment around credit appears to fall both in the 1991 recession and the Great Recession of 2007-09, while negative sentiment surrounding mortgages plays a role primarily in the Great Recession and its aftermath (see Panels (e) and (f)). Another insight coming from the figure is that some concepts gain importance over time. For example, the sentiment around inflation expectations in Panel (b) moves relatively little for most of the sample, but displays larger volatility since the 2000's. While we use the full set of 296 sentiment indicators in a multivariate econometric analysis, a by-product of our analysis is a rich descriptive picture of the Fed's assessment of various aspects of the US economy over the last few decades. Appendix B contains sentiment plots for additional economic concepts.

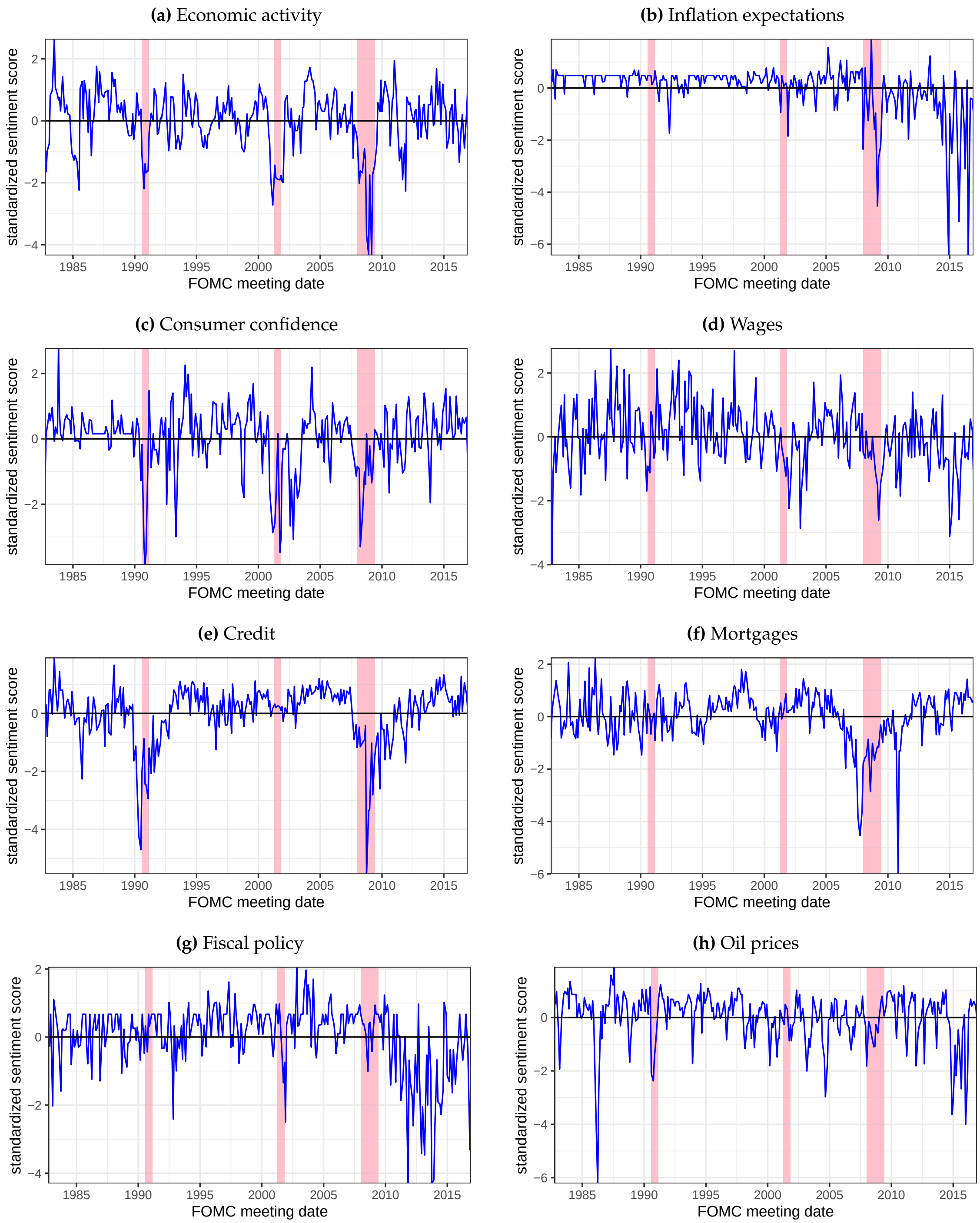
Step 4: Specify and estimate the empirical model

Nonlinear specification using forecasts and sentiments. Our empirical counterpart of equation (1) includes the Fed's policy instrument on the left hand side, and both numerical forecasts and sentiment indicators from FOMC documents on the right hand side. Both sets of variables can enter nonlinearly. Formally, we define

$$\Delta i_t = \alpha + \Gamma(i_{t-1}, \widetilde{\boldsymbol{X}}_t, \boldsymbol{Z}_t) + \varepsilon_t^*. \tag{3}$$

are among our selected economic concepts, such *unemployment* and *unemployed*, or because we think they should not necessarily be interpreted as positive or negative in the context of the Fed's analysis, such as the term *unforeseen*.

Figure 2: SELECTED SENTIMENT INDICATORS



Notes. Sentiment indicators for a selection of economic concepts discussed in FOMC meeting documents, out of our full list of 296. The sentiments are constructed using the dictionary of positive and negative words in financial text of Loughran and McDonald (2011). Each indicator is standardized across the sample. Shaded areas represent NBER recessions.

Δi_t are changes in the FOMC's FFR target, which for simplicity we mostly refer to as just the FFR.[13] $\widetilde{\boldsymbol{X}}_t$ contains augmented set of Fed forecasts, which includes additional production, investment, housing and government spending variables relative to $\boldsymbol{X}_t$ in (2).[14] Following Romer and Romer (2004)'s specification, we enter forecasts in levels and first differences, across several forecast horizons, which amounts to 132 forecast time series. $\boldsymbol{Z}_t$ contains our 296 sentiment indicators. We also allow 4 lags of the sentiment indicators to enter, as the path of the economy, which includes recent historical performance, may have an influence on how the current state of the economy translates into policy changes.[15] $\Gamma(\cdot)$ is a nonlinear mapping. In our main analysis, we specify this as a linear-quadratic function. Together with the level of the FFR, i_{t-1}, which we also allow to enter quadratically, (3) includes 3,226 variables on the right hand side. We analyze different lag structures and alternative nonlinear specifications of $\Gamma(\cdot)$ for robustness.

Ridge regression. While we construct sentiments until 2016, we focus on the period before the zero lower bound (ending with the meeting on October 29, 2008) to estimate (3). This avoids running a regression with many zeros for the dependent variable. Our sample from October 1982 to October 2008 captures 210 FOMC meetings (observations). Thus an ordinary least squares (OLS) regression with several thousands of regressors is infeasible. To overcome this issue, we resort to ML techniques. Specifically, we employ a ridge regression. The idea of a ridge regression is to minimize the residual sum of squares and an additional term that penalizes the squared deviations of each regression coefficient from zero. Formally, in the model $y_i = \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik} + \varepsilon_i$, the ridge minimizes $\sum_i \varepsilon_i^2 + \lambda \sum_j^k \gamma_j^2$. The Bayesian interpretation of a ridge regression is Bayesian OLS with a normal prior on each coefficient, centered around 0, with scale of the prior variance equal to λ. Unlike its close sibling, the LASSO regression, a ridge regression results in estimated coefficients for all regressors. An optimal λ (in a predictive sense) can be found using k-fold cross-validation. This is done as follows: randomly divide the sample into k subsamples, so-called folds, of

[13] In the part of our sample that overlaps with Romer and Romer (2004), our left hand side is identical to theirs. Afterwards, we use the series built by Thornton (2005) (updated by FRED).

[14] This forecast data is conveniently made available by the Philadelphia Fed here.

[15] The text often assumes knowledge of the previous meetings' analysis, which calls for including lagged sentiment indicators in (3). For example, the language in the second Greenbook following the 9/11 terrorist attacks appears to take knowledge about the attacks and their impact on the economy as given, with reference to the previous meeting's documents.

equal size; use each subsample to evaluate the model when it is fit on the $k-1$ other subsamples; in each case, compute a mean-squared error (MSE); compute an average MSE across these k MSEs; find the smallest average MSE by changing λ. We follow this procedure using $k = 10$. Note that all variables that enter the ridge regression are standardized.

Discussion of NLP and ML choices. We conclude the step-by-step description of our method with two remarks. First, relative to the rich variety of modern NLP and ML methods, we opt for an approach with restrictions to reduce the complexity of the information. We carry out sentiment analysis for hand-selected economic concepts, sometimes referred to as Aspect-Based Sentiment Analysis. One alternative to our Steps 2 and 3 would be to capture the entirety of the FOMC documents in (3), for example through term-document matrices, in which rows correspond to documents, columns correspond to any English-language term, and entries in the matrix contain the frequency of each term.[16] This alternative would involve hundreds of thousands of regressors, and might be more suitable for less structured text. Instead, we build on the fact that the Greenbook documents we use contain very structured and carefully worded text with consistency through time. An advantage of our procedure is that the model retains interpretability and echoes the spirit of the original idea of Romer and Romer (2004).

Second, the ridge regression in Step 4 is one of several related ML techniques that could be applied here. One natural alternative would be the LASSO regression, which instead minimizes $\sum_i \varepsilon_i^2 + \lambda \sum_j^k |\gamma_j|$, or the elastic net, which is a mixture between ridge regression and LASSO. A key difference is that LASSO results in a *sparse* model that contains only a subset of the right-hand-side variables, while ridge results in a *dense* model, containing all regressors and associated coefficients. In this sense, ridge regressions are more related to dynamic factor models, which are often employed for macro data. We prefer ridge regression on the grounds that dense rather than sparse prediction techniques tend to be preferable for economic data, which typically consists of many correlated regressors with relatively small number of time series observations. This is confirmed by the in-depth analysis of Giannone, Lenza, and Primiceri (2022). These authors develop a Bayesian prior that allows for both shrinkage and variable

[16] Kalamara et al. (2020) discuss and compare different prediction models based on high-dimensional text analysis methods in an application to newspaper text.

selection, and find that including many predictors, rather than reducing the set of possible predictors, improves accuracy in several different economic applications. Although their study does not consider text data specifically, it shows that sparse method can become unstable in the presence of high collinearity between the predictors. This is clearly the case across the numerical forecasts and our sentiment measures based on text, and within both groups of variables.[17]

3 Examining the information content of the sentiment indicators

Before we apply our method, this section presents a discussion and econometric validation exercise to assess the informational value added of our sentiment indicators. We examine the "sufficient statistic" argument laid out by Cochrane (2004) in a discussion of Romer and Romer (2004): suppose the forecasts in the Greenbook efficiently incorporate all information that the FOMC has available about a variable of interest. In that case, it would not be necessary to include additional information in (2) to retrieve a shock measure that can be used to recover the true response of that variable to changes in monetary policy. In fact, keeping the additional information in the residual would be desirable for statistical power. The analysis that follows shows where this arguments fails and why it is thus essential to include relevant additional information coming from the text. We also argue that it is advantageous to include a large set of information, beyond just the essential information about one specific variable.

3.1 Mean vs. mode forecasts

Cochrane (2004)'s argument starts with the assumption that the Greenbook forecasts correspond to a conditional mean expectation. This assumption does not hold if the Greenbook forecasts are instead interpreted as *modal* predictions. For example, if there is asymmetric tail risk, a conditional mean and conditional mode predictions are not the same. We systematically examine the transcripts of FOMC meetings and find ample evidence that support the view of the Greenbook forecasts representing modal predictions. For example, in the FOMC meeting on July 2-3, 1996, Michael Prell, the director of Research and Statistics at the time

[17] In a macroeconomic forecasting context Bianchi, Ludvigson, and Ma (2022) find that elastic net methods, which weight ridge and LASSO penalties, perform best among various ML techniques including random forest techniques. They also emphasize the collinearity of macroeconomic data.

clarifies: "*I would characterize our forecasts over the years as an effort to present a meaningful, modal forecast of the most likely outcome. When we felt that there was some skewness to the probability distribution, we tried to identify it. In this instance, as we looked at the recent data, we felt that there was a greater thickness in the area of our probability distribution a little above our modal forecast.*" Appendix D provides numerous additional quotes from the FOMC transcripts across our sample period.

In other words, the staff's forecasts are not designed to be correct on average, but rather they provide the most likely outcome. They are designed to predict the realization of macroeconomic variables in a modal scenario, which the staff provides in combination with a description about expected changes around the mode in words, such as the emergence of asymmetric tail risks. Indeed, the Tealbook A nowadays contains a "Risks and Uncertainties" section, where the asymmetric balance of risks around the numerical forecasts is described explicitly by the staff. This general insight is in line with some complementary research that alludes to the modal nature of Greenbook forecasts, for example by Reifschneider and Tulip (2019), and recently by Cieslak et al. (2021).

Another important aspect of the Greenbook forecasts is that the staff produce them based on a specific future path for the policy rate, as explained by Faust and Wright (2008). This property is another reason why they are a different object from the mean expectation conditional on the FOMC's full information set, which would integrate over all possible future policy paths. This feature of the forecasts provides an additional argument for including text-based information.

3.2 Forecast error predictability

Even in the presence of modal forecasts, Cochrane (2004)'s reasoning might be valid in practice if conditional mean and mode mostly coincide. We therefore verify econometrically whether our sentiment indicators predict errors in the staff's numerical forecasts on average. If that is the case, then there is valuable information about the conditional mean available to the FOMC that is not captured by the forecasts, and thus should be removed from FFR variation to obtain valid monetary policy shocks. The tests presented here focus on the Greenbook unemployment rate forecast, which is particularly suitable because there is little definitional change over time and it is subject to only small data revisions. We provide analogous results for output and inflation forecasts in Appendix E.

Table 2: GREENBOOK FORECAST ERROR PREDICTABILITY TESTS

	Dependent variable: Greenbook unemployment rate forecast errors							
	(1) current quarter	(2) 1-quarter ahead	(3) 1-year ahead	(4) 2-years ahead	(5) current quarter	(6) 1-quarter ahead	(7) 1-year ahead	(8) 2-years ahead
First PC of all sentiments	-0.029*	-0.114**	-0.445**	-0.622**				
	[0.016]	[0.049]	[0.190]	[0.238]				
Economic activity sentiment					-0.026	-0.098**	-0.285*	-0.363**
					[0.016]	[0.048]	[0.165]	[0.171]
Constant	-0.019	-0.070**	-0.082	0.059	-0.019	-0.069**	-0.077	0.160
	[0.014]	[0.033]	[0.121]	[0.201]	[0.014]	[0.035]	[0.145]	[0.258]
R^2	0.045	0.149	0.248	0.208	0.033	0.097	0.090	0.055
Number of observations	210	210	210	62	210	210	210	62

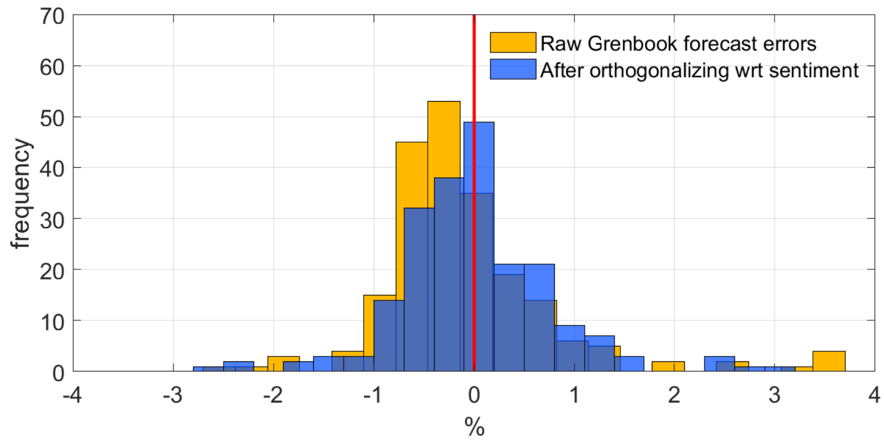
Notes. The forecast errors on the left hand side are constructed by subtracting the Greenbook forecast of the quarterly unemployment rate from the actual unemployment rate (final vintage). Regressions are run at FOMC meeting frequency 1982-2008. Newey-West standard errors with optimal bandwidth are provided in brackets. *, **, *** indicate significance at the 10%, 5%, 1% level.

Table 2 presents estimates from different forecast error regressions, over the same sample period we use to estimate equation (3). The left hand side is the unemployment rate forecast error in percent (defined as final vintage minus forecast). Columns (1)-(4) include the first principal component (PC) of all 296 sentiment indicators on the right hand side. Columns (5)-(8) include one single sentiment indicator on the right hand side, the one for "economic activity" shown in Panel (a) of Figure 2. In both sets of regressions, we focus on the current quarter, 1-quarter, 1-year and 2-year ahead forecast errors. Note that for the 2-year horizon, the number of observations is lower because two-year ahead forecasts are not produced for all FOMC meetings.

The table reveals that Greenbook forecast errors are predictable with our text-based sentiment measures. The error in the Greenbook unemployment rate forecast is predictable with the first PC of our sentiments, and even with one of the 296 sentiment indicators alone, at various forecast horizons and to an economically significant degree. The economic significance increases with the forecast horizon, and the R^2 of the regression can be as high as 0.25. To give an example for how the magnitude of the coefficients should be interpreted, column (3) indicates that a one standard deviation increase in the sentiment PC is associated with an almost 0.5 percentage point negative forecast error in the unemployment rate.[18]

[18] Appendix E provides results for real output growth and inflation. For these variables, forecast errors are predictable as well, though at fewer horizons. The same appendix also shows results based on using the first release instead of the final vintage in the construction of the forecast errors.

Figure 3: UNEMPLOYMENT FORECAST ERRORS BEFORE & AFTER ADJUSTING FOR SENTIMENT



Notes. The orange bars represent a histogram of the Greenbook forecast errors for the unemployment rate at the 1-year horizon. The blue bars show the residuals of regression the forecast errors on the first principal component of our text-based sentiments (column (3) of Table 2). Both histograms are constructed based on 1982-2008 sample, using 20 bins.

These estimates are in line with our argument that the staff construct modal forecasts. Assume for illustration that there is a well-calibrated distribution of unemployment rate forecasts with a lower bound of 4%, a mode of 6%, an upper bound of 8%, and a mean greater than 6%. That is, there is more mass to the right of the mode. If one computes a forecast error using the modal forecast, there would likely be a positive average forecast error, because on average the outcome should be greater than the mode forecast. This positive average error would, according to our regressions, be significantly negatively correlated with the economic activity sentiment. This is consistent with negative economic activity sentiment in the staff documents capturing the thicker upper tail of the unemployment rate distribution, and therefore predicting the positive forecast error on average. In other words, while the Fed staff provides a numerical modal forecast, through their narrative that accompany this forecast, they relay what is in this case an upside risk in unemployment, which, in turn, is captured by our sentiments.

In this illustrative example, the text-based sentiments capture risk of higher than predicted unemployment. Of course the example applies equally in the opposite direction, where positive activity sentiment captures the left tail of the unemployment rate distribution and potentially negative average forecast errors. This begs the question whether over the full sample period the Greenbook

The results are similar. Finally, note that we also tried including lags in the regressions, and found that the predictive power is mostly concentrated in the contemporaneous sentiment measures.

forecasts on average over- or underpredict the unemployment rate. Figure 3 focuses on the 1-year ahead prediction and shows that the forecast errors are negative on average, shown by the orange bars. The blue bars represent the residuals from regressing the forecasts on our text-based sentiment. After this orthogonalization, the distribution becomes more symmetric and more centered around zero, highlighting also graphically the relevance of the information content we extract from the text.

Our findings about Greenbook forecast errors are in line with complementary work by Sharpe, Sinha, and Hollrah (2020), who find that language "tonality" surrounding forecasts predicts errors of both Fed and private sector forecasts. We make clear that this means text-based information is crucial to inform the systematic component of policy when estimating monetary policy shocks.

3.3 Essential vs. comprehensive measurement of information

Given the modal nature of Greenbook forecasts and our analysis of forecast errors, it is essential that FFR target changes are orthogonalized with respect to *at least some information*, namely the subset of text-based information that is relevant to "correct" the forecast and be able to control FFR changes for the FOMC's conditional mean expectation of a variable. This is true even for the original Romer and Romer (2004) approach to work correctly, applying the Cochrane (2004) logic.

Beyond this requirement, the philosophy of our approach is that a more comprehensive estimate of the FOMC's information set with a *large amount of information* has additional advantages. We want to create a monetary policy shock measure that is exogenous to all available information and can be used to study the effect of monetary policy on many macroeconomic variables, including those for which forecasts are not produced by the Fed. For example, based on the argument of Cochrane (2004), studying the effects of monetary policy on credit spreads would not strictly be feasible using Romer-Romer residuals, as the staff does not produce credit spread forecasts. However, credit spreads are discussed in detail in the text and their fluctuations are captured by our sentiment indicators. In other words, our procedure allows us to construct an "all purpose" monetary policy shock time series that is portable to any other econometric setting.

The downside of an approach based on a large information set might be that the resulting monetary policy shock has lower statistical power. We examine whether this is an issue in our practical application. Our findings below indicate that this downside does not appear to outweigh the benefits of a cleaner shock estimate.

4 Results of the identification procedure

This section discusses the estimation results for the empirical model represented by equation (3). The findings we present include measures of fit, properties of the estimated shock time series, an interpretation of monetary policy shocks as well as a comparison to surprises in market interest rates.

4.1 Systematic vs. nonsystematic changes in the target rate

Table 3, column (1) presents the R^2 of alternative empirical specifications. First, as the simplest benchmark it includes equation (2), the restricted version of (3) where only the staff forecasts used in the original Romer and Romer (2004) specification enter in a linear OLS estimation. Second, a model that includes the expanded set of 132 forecast variables, and is estimated as a ridge rather than an OLS regression. Third, a ridge model where the augmented set of forecasts and our sentiment indicators are included, but function $\Gamma(\cdot)$ is still linear. Fourth, a ridge model with the same forecasts and sentiments variables entering linearly and quadratically. Fifth, the linear ridge model which also contains 4 lags of the sentiment indicators. Sixth, our main specification in which forecasts and sentiments enter linear, nonlinearly and with 4 lags.

We compare the fit of these alternative models to understand what they imply about the contribution of the systematic component of monetary policy. The first line in Table 3, column (1) shows that over the sample period October 1982 to October 2008 that we consider, the Romer-Romer OLS model implies an R^2 of 0.5. In other words, this empirical model attributes 50% of the variation in the FFR target to systematic policy, while 50% is attributed to monetary policy shocks. This seems undesirable – as Leeper, Sims, and Zha (1996) put it: "*Even the harshest critics of monetary authorities would not maintain policy decisions are unrelated to the economy.*"

The remaining lines in column (1) of the table reveal that expanding the information set in the empirical model increases the implied fit. Bear in mind that the ridge regression does not maximize fit, but instead optimizes predictive performance based on the 10-fold cross-validation so increase in R^2 is not purely

Table 3: R^2 ACROSS DIFFERENT SPECIFICATIONS

	Number of regressors	(1) R^2 **with 10-word sentiment (main specification)**	(2) R^2 **with 5-word sentiment (robustness)**
Romer-Romer original OLS with subset of forecasts	19	0.50	
Ridge with extended set of forecasts	133	0.55	
Ridge with all forecasts & sentiments (linear)	429	0.65	0.66
Ridge with all forecasts & sentiments (nonlinear)	858	0.75	0.77
Ridge with all forecasts & sentiments (linear with lags)	1,613	0.87	0.88
Ridge with all forecasts & sentiments (nonlinear with lags)	3,226	0.94	0.95

Notes. Implied goodness-of-fit, measured by R^2, from estimating different empirical specifications of equation (3). For the first two specifications, sentiments are not included so the 10-word/5-word distinction does not apply. Our preferred specification is the last one presented in the table, with forecasts and sentiments entering nonlinearly and with 4 lags.

mechanical. Nevertheless, each step of enriching the empirical model – going from OLS to ridge regression, including more numerical forecasts and sentiment indicators, and allowing for nonlinearities and lagged sentiments – delivers some additional improvement in the fit of the model. Our preferred specification, the bottom line in Table 3 implies an R^2 of 0.94, suggesting that 94% of FFR variation is systematic, and 6% are explained by shocks. Relative to the Romer-Romer OLS model, this reduces the contribution of exogenous shocks almost ten fold.

Besides the economic appeal of these findings, our analysis of forecast error predictability in Section 3 already supported the view that more information about the systematic component of monetary policy should be included, i.e. that the R^2 should be higher than in a specification with forecasts only. One downside of a higher R^2 and therefore less variation in the shocks could be low statistical power when studying the responses to the shocks in a finite sample of macroeconomic data. Leeper et al. (1996) describe this challenge quite pointedly, by saying "*This is what one would expect of good monetary policy, but it is also the reason why it is difficult to use the historical behavior of aggregate time series to uncover the effects of monetary policy.*" Our remaining results, in particular the IRFs we estimate in Section 5, show that lack of statistical power in the shock measure is not an issue in our application.

Further variations in the specification. We check robustness of the results above along several dimensions. Column (2) of Table 3 focuses on empirical models in which our sentiment indicators are constructed using a 5-word instead of a 10-word window around economic concepts. By construction, the first two rows in each column remain unchanged, as these specifications do not incorporate sentiment indicators. The meaningful increase in fit from expanding the information set remains present when we vary our way to construct sentiment indicators. We verified that constructing sentiments based on positive and negative words within the same sentence, rather than a fixed word window, yields time series that are highly correlated with the ones we use, see examples in Appendix C . We also experimented with the lag structure of those specifications that include lags. We found that increasing the number of lags, starting at 0 lags, increased the R^2 for a given specification, but the increases becomes fairly small around 4 lags.Furthermore, we constructed an auxiliary data set about the FOMC's composition, in order to verify whether personal dynamics between FOMC members drive FFR changes. We found that this was not the case: the R^2 from including this information in the ridge regression increased by less than 0.1%.[19] Finally, we tried alternative nonlinear specifications of $\Gamma(\cdot)$. We found that the model fit and time series of the residuals we obtained were similar to the quadratic version. For example, the residuals obtained with a cubic version were 99% correlated with the corresponding quadratic specification. A specification in which we added all possible linear interaction terms between all sentiment indicators and all forecasts, as well as squared terms – amounting to almost 40,000 variables on the right hand side of (3) – gave residuals that were 96% correlated with the corresponding quadratic version.

Predictive power vs. interpreting coefficients. The methodology we employ is designed to capture as much of the systematic policy changes as possible. In a setting with many more variables than observations as ours, the ridge regression has a lot of flexibility in matching the policy changes with discipline coming from cross-validation. An output of this process is a list of parameter estimates for each of the 3,226 regressors, which one may think is useful to analyze. However, a variable-by-variable analysis of the estimated coefficients our ridge regression is difficult given the high correlation between many of the regressors. Moreover the presence of multiple lags and the quadratic terms also complicate pinpointing the contribution of individual variables to the fit. To use an analogy with a popular application of ML techniques, a goal of a self-driving car is to recognize obstacles on the road and avoid hitting them. In order to do so it uses a large number of measurements from its sensors. It would be hard for an engineer to answer the

[19]More details on the construction of this data set are provided in Appendix F.

question of why the car did stop when it stopped as a complicated combination of such measurements are at play. Mullainathan and Spiess (2017) in their review of ML techniques, conclude that ML belongs in the part of the toolbox marked $\hat{y}$ rather than in the more familiar $\hat{\beta}$ compartment.

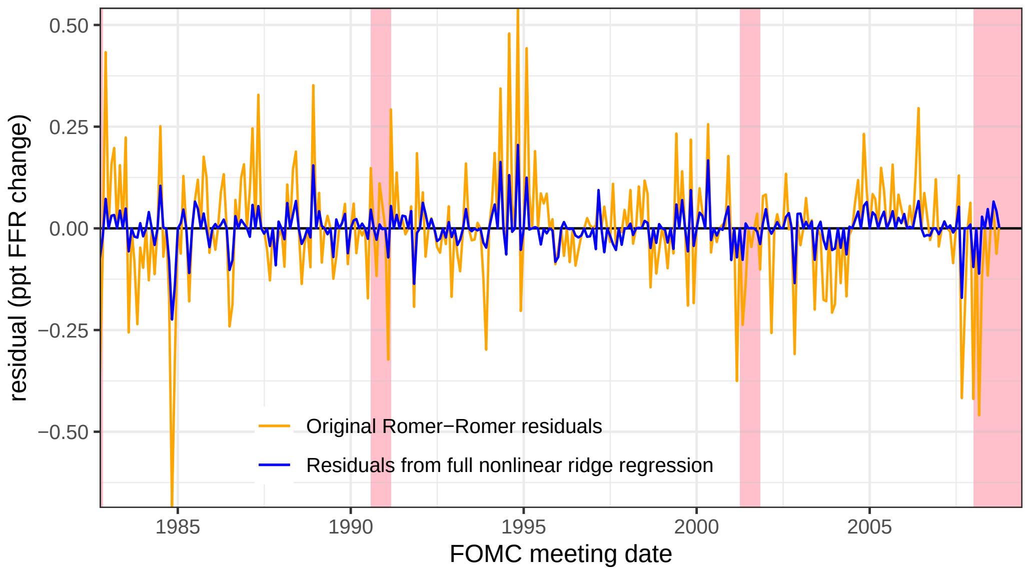
4.2 What are monetary policy shocks?

The dark blue line in Figure 4 plots the estimated time series of monetary policy shocks, that is, the residuals $\hat{\varepsilon}_t^*$ from our preferred empirical specification which includes forecasts, sentiments and nonlinearities in a ridge model. The figure compares this with the estimated residuals from the Romer-Romer OLS model as the lighter orange line. The residuals have the same unit as that of the left hand side of the regression, so can be interpreted in percentage point changes in the FFR. Recall that the shocks represented by the blue line explain 6% of FFR variation while those represented by the orange line explain 50%. Related to the lower contribution to FFR variation, the figure shows that our measure of monetary policy shock displays lower volatility. We also find it to display a lower degree of autocorrelation, with a correlation with its first lag of 0.066 as opposed to 0.204 for the Romer-Romer residuals. It is also visible in the figure that our estimate of shocks is not simply a scaled-down version of the shocks implied by the Romer-Romer OLS model. In many instances, the orange line implies a larger shock in the same direction, while in others the shock measures go in different directions.

Case studies of largest FFR changes. For those episodes in which the estimated shocks are particularly large in magnitude, we closely inspect the discussion that took place in the FOMC. Here we provide two examples, which shed light on what estimated monetary policy shocks capture. Further below, we show that the effects of monetary policy on the economy that we estimate hold when restricting our shock time series to only its largest realizations. This underlines the relevance of our interpretation of monetary policy shocks for estimated IRFs.

The largest shock in absolute value is estimated for the November 7, 1984 FOMC meeting. The policy change is a decline in the FFR of 75 basis points (bp) and our shock measure is minus 22 bp, indicating that based on staff forecasts and sentiments, we predict a decline of 53 bp. This is a period that has a mixed economic outlook: employment shows the smallest rise since the expansion began

Figure 4: ESTIMATED MONETARY POLICY SHOCKS



Notes. Time series of estimated monetary policy shocks. Dark blue: our preferred version, the residuals from predicting changes in the FFR based on numerical forecasts, sentiment indicators and nonlinearities in FOMC documents. Orange: benchmark version based on a specification that follows Romer and Romer (2004). Shaded areas represent NBER recessions.

at the end of 1982, yet investment and consumption show robust increases. The staff conclude that the "slowdown may only be a pause in a recovery that has not run its full course" in the Beigebook. They forecast an increase of 3.5% in GDP for the current quarter, compared to 2.75% in the previous quarter. Their inflation forecast is also flat relative to recent quarters and it is expected to pick up in 1985. When we read the transcript of the FOMC meeting, it becomes clear that several participants find the staff forecast too optimistic. As a result, at the end of the meeting, FOMC's policy actions are consistent with a sizable easing of policy, which is contrary to what one may have decided by simply reading the staff documents. In fact, one of the two policy options put forward by the staff involved no changes in policy. This episode is a good example of a situation where the FOMC participants' views about the economy are different from the staffs', and the policy action is far from what would be implied by the latter. It is important to emphasize that this is an unusual situation. If the disagreement happened more often, then our procedure would have picked it up as a systematic part of policy, and it would not show up in our shocks.

Our second example is the November 15, 1994 meeting, where a 75 basis point FFR increase was decided, and our analysis shows 21 bp of this was a monetary policy shock. The staff analysis paints the picture of robust growth: they forecast an acceleration in output, final demand is high and banks are lending. They conclude that the economy is above its full capacity with the inflationary consequences not yet realized. The staff proposes two policy options: a no change option and one where the FFR increases by 50 bp. In their forecasts, the staff assume "appreciable further tightening" with a cumulative increase of 150 bp in the following 6 months. During the meeting, Chairman Greenspan suggests that "they are behind the curve" and since the market already built in a significant rate hike "a mild surprise would be of significant value." He proposes a rate increase of 75 bp to get "ahead of general expectations." Most participants agree with this proposal, with several participants emphasizing the credibility of keeping inflation under control. Once again this is a situation where the FOMC decided on an action not simply based on the current economic outlook but also other considerations, and our procedure implies that this reflects a monetary policy shock. The difference between the 75 bp decision and the staff's suggested 50 bp option almost exactly matches the 21 bp contractionary shock we estimate.

Consistent with the low variation in our shock measure, our interpretation of monetary policy shock is also narrower than the one of Romer and Romer (2004). In their original study, they describe changes in the Fed's target (monetary aggregates vs. the FFR), as well as political interactions between the Fed and Presidents as potential sources of shocks. Both are unlikely to cause meaningful shocks in the post-1982 sample. The change in the target occurred before 1982 and political pressure on the Fed was most salient in the 1970s (Drechsel, 2023).

One might argue that credibility concerns such as the ones motivating the strong hike in November 1994, are in some way a feature of the Fed's policy rule and should therefore not constitute a shock. However, the types of decisions that imply monetary policy shocks in our procedure must occur in a completely nonsystematic way, not in response to changes in the Fed's information set. Systematic credibility concerns, that arise based on available information, will be picked up by our ridge regression as part of systematic policy.

Our case studies imply that one can give a "surprise" interpretation to our shock measure, that is, FFR target decisions by the FOMC that constitute *surprises to the Fed staff*. In instances of monetary policy shocks where the FOMC makes a decision that is orthogonal to its information set – as summarized by the staff's forecasts and language – this should be unpredictable by the staff.

4.3 Shocks vs. market surprises

An alternative branch of research identifies monetary policy shocks from surprise movements in *market* interest rates in tight windows around FOMC announcements. Early contributions include Gürkaynak et al. (2005) and Gertler and Karadi (2015). Our approach is different from high-frequency approaches, as our left hand side variable is the *target* FFR that the FOMC sets directly, rather than a market price that reacts to FOMC decisions and announcements. Of course surprise movements in market interest rates themselves may be of interest for researchers. When it comes to identifying monetary policy shocks using HF approaches, one challenge is that other effects might cause market interest rate surprises, for example the "Fed information effect", see e.g. Romer and Romer (2000), Campbell et al. (2012) and Nakamura and Steinsson (2018).[20]

To examine how our shocks compare to this alternative methodology, we retrieved the FFR surprises constructed by Swanson (2021) and provide a comparison in Table 4.[21] The table provides the correlation between our shocks and the surprises for all scheduled FOMC meetings between 1991 and 2008. As a benchmark, we also compute the same correlations with the original Romer-Romer shock measures. Across the entire sample, the correlation is 0.49, compared to 0.36 for the original Romer-Romer measure. We also focus on the largest observations, in order to cut out the potential noise coming from smaller shocks. When we focus on the 10 largest shocks from our procedure, the correlation with the corresponding surprise-based measure of shocks is 0.77. When we focus on the largest surprises, the correlation is 0.51. In both cases this significantly exceeds the corresponding correlation for the original Romer-Romer shocks. This makes clear that by better controlling for the Fed's information set, our methodology reduces the difference between alternative approaches to identify monetary policy shocks.

To put the size of the 0.49 correlation coefficient in Table 4 into context, we emphasize that we are comparing the output from two completely different

[20] Jarocinski and Karadi (2020) and Miranda-Agrippino and Ricco (2021) separate HF surprises in market interest rates between pure monetary policy shocks and informational shocks. Bauer and Swanson (2023) highlight a "Fed response to news" mechanism.

[21] We thank the author for making the data publicly available. While there are alternative surprise series, we focused on this one for two reasons. First, Swanson (2021) captures surprises to the FFR separately from surprises about unconventional monetary policy. This makes it conceptually similar to our methodology. Second, it is available at the meeting frequency. Other surprise series are only available monthly, aggregating scheduled and unscheduled meetings.

Table 4: MONETARY POLICY SHOCKS FROM OUR METHODOLOGY VS. FROM MARKET SURPRISES

	(1)	**(2)**
	Our measure	**Original Romer-Romer**
Correlation shocks with surprises	0.49	0.36
Correlation top 10 shocks with surprises	0.77	0.61
Correlation top 10 surprises with shocks	0.51	0.18

Notes. Comparison with the FFR surprises constructed by Swanson (2021) (1991-2008). These can directly be matched to our shocks and Romer-Romer shocks for scheduled FOMC meetings.

methodological approaches. Both seek to identify exogenous shifts in monetary policy, so it is reassuring that they deliver correlated time series. However, at a deeper level, it is not clear that they necessarily need to get at the same underlying concept of "true" shocks. Instead, both approaches isolate variation in rates that is plausibly exogenous, and allows to study the effect of monetary policy on the economy. In other words, it is possible that researchers find two valid and relevant instruments for the same variable without the instruments being highly correlated.

Practical considerations relative to HF surprise measures. One advantage of our procedure is that we obtain a shock series that spans a long time period, while the availability of FFR futures data restricts HF measures to start in the 1990's. Prior to 1994 the FOMC did not announce interest rate changes publicly, which further complicates pinpointing monetary policy surprises. A key advantage of surprise-based measures, on the other hand, is that they can be constructed for unscheduled meetings, while the Greenbooks are only produced for schedule meeting. Furthermore, surprises in market rates can be observed around other events such as speeches by FOMC members (Jayawickrema and Swanson, 2023). It could be a useful practical consideration to combine both approaches econometrically, for example as multiple external instruments.

5 The effects of monetary policy shocks on the economy

This section uses our new shock measure to study the effects of changes in monetary policy on the US economy in a state-of-the-art BVAR model, following Jarocinski and Karadi (2020). The BVAR is estimated at monthly frequency, and includes the 1-year Treasury yield, the log of the S&P500, the log of real GDP, the log of the GDP deflator, the unemployment rate, and the excess bond premium

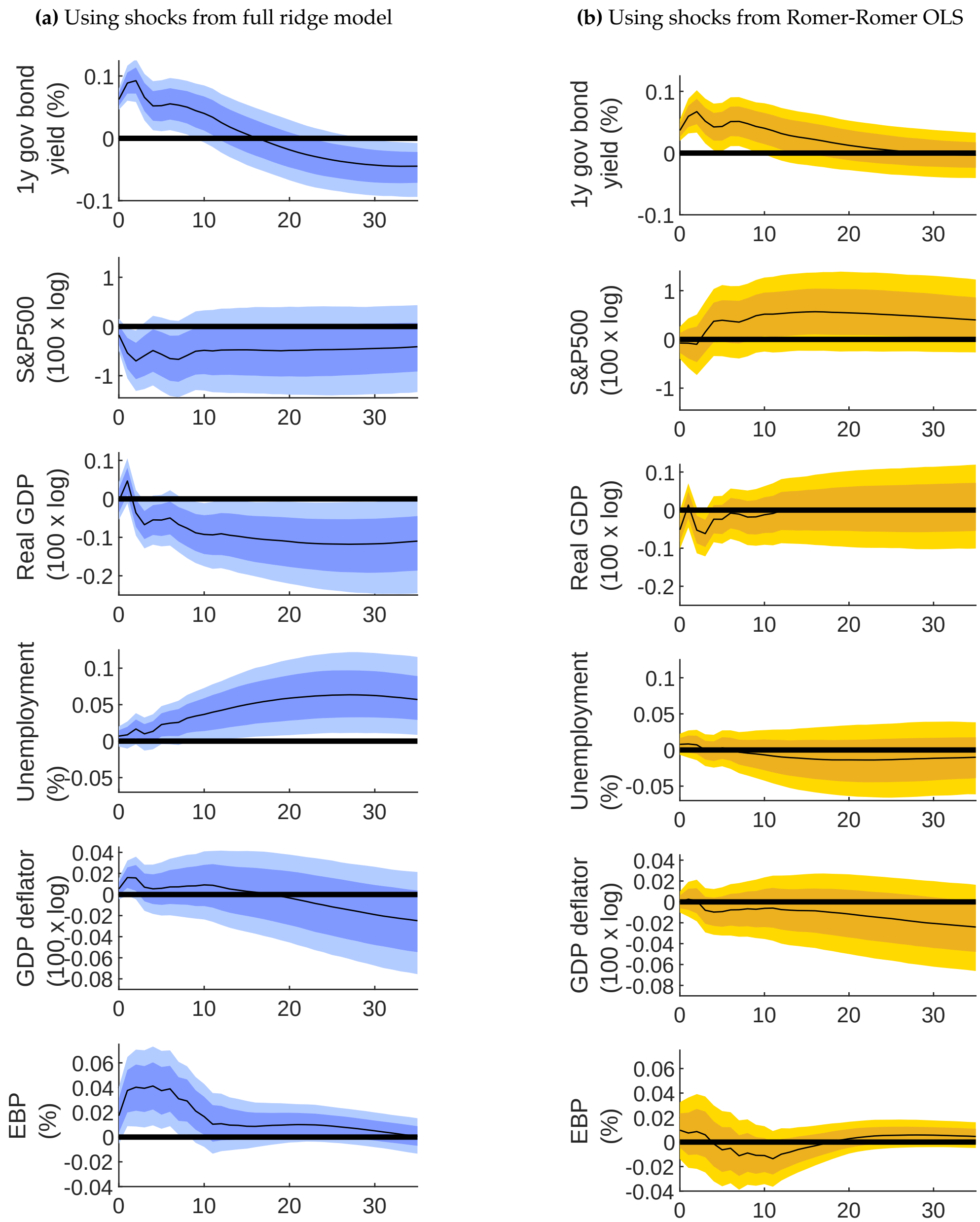
(EBP). Our time series of shocks enters as an exogenous variable, ordered first in a Cholesky ordering, which yields asymptotically identical results to using the shock series as an external instrument (Plagborg-Moller and Wolf, 2021).While our shocks span the period 1982:10-2008:10, applying it as an external instrument allows us to estimate the system over a longer sample. The 1-year yield is included as it is mostly free to move while the target FFR is at the zero lower bound for part of the sample. GDP and its deflator are included to capture the effect of monetary policy on activity and prices. We use their monthly versions, interpolated using the Kalman filter. We include the unemployment rate, given that we found the Greenbook forecast error predictability to be particularly strong for this variable. The S&P500 and EBP are included as forward-looking financial variables. For comparability, we use the same sample period (1984:02-2016:12), settings and priors as in Jarocinski and Karadi (2020).[22]

Figure 5, Panel (a) presents IRFs of macro variables to our preferred measure of monetary policy shocks. We find that a monetary policy tightening is characterized by a relatively persistent increase in yields, lasting for about 20 months. The rise in interest rates leads to a reduction in real economic activity and a fall in the price level, directly in line with what standard economic theory predicts. The reduction in real output and the increase in unemployment take about a year to materialize and are very persistent. The price level response displays a mild version of a "price puzzle" in the first months, but is persistently negative thereafter. It takes about 18 months for the point estimate to be visibly negative, and 30 months for the response to be significantly negative. Bond premia increase sharply and significantly after a monetary policy tightening, a finding in line with models of monetary policy and external finance premia. Furthermore, our identified monetary policy shocks imply a fall in stock prices following a tightening in monetary policy, consistent with theory (Jarocinski and Karadi, 2020).

These results contrast with Panel (b), which presents IRFs to residuals constructed using the original Romer-Romer OLS specification, in which only a handful of numerical forecasts are used to predict the systematic component of monetary policy. There is a similar path for market interest rates, as well as a comparable reduction in the price level, but the effect on real GDP and the unemployment rate are completely flat. This is different from the IRFs in

[22] We thank these authors for making their Gibbs sampler codes available. Their sample starts in February 1984, the end of the Volcker disinflation period.

Figure 5: IRFS TO DIFFERENT MONETARY POLICY SHOCK MEASURES



Notes. IRFs to different estimated monetary policy shocks in BVAR model (without additional sign restrictions imposed). Panel (a) uses our proposed measure of monetary policy shocks, estimated using the full nonlinear ridge model on the extended set of numerical forecasts and our sentiment indicators from FOMC documents. Panel (b) shows the analogous IRFs when a simpler empirical specification is used to estimate the shocks, which includes only the original set of numerical forecasts in a Romer-Romer OLS regression. The solid line represents the median, the 16th and 84th percentiles are represented by the darker bands, and the 5th and 95th percentiles by the lighter bands. The sample period to estimate the shocks is 1982:10-2008:10. The sample used to estimate the IRFs is 1984:02-2016:12.

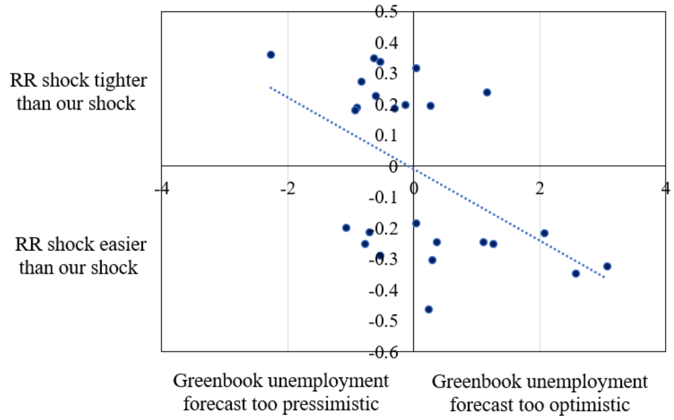
the original Romer and Romer (2004) paper, using the 1969-1996 sample, where economic activity is significantly reduced after a tightening. This contrast connects to earlier findings that the IRFs to their original shocks give results at odds with theory in more recent samples. Ramey (2016), Barakchian and Crowe (2013) and Caldara and Herbst (2019) provide discussions about possible reasons behind these results. Moreover, the shocks computed using the original Romer-Romer methodology imply an insignificant response of the EBP to a policy tightening, and positive comovement between the S&P500 and interest rates conditional on a monetary policy shocks, both of which are inconsistent with standard theory.

The differences between Panels (a) and (b) suggest that some systematic policy variation is still present in the shock measure only based on controlling for numerical forecasts, whereas our measure based on a larger information set is more plausibly exogenous. We provide three complementary interpretations of this finding, focusing on different variables in the BVAR. Since the unemployment response is particularly different, the first interpretation draws a direct connection to the evidence in Section 3 that text-based sentiment is predictive of Greenbook unemployment rate forecast errors. Given the fact that Greenbook unemployment forecasts are modal in nature and therefore is not the conditional mean expectation, the standard Romer-Romer OLS regression cannot fully incorporate the effects of asymmetric changes in the balance of risks around unemployment forecasts on the systematic conduct of policy. This results in an "incorrect" IRF of unemployment.

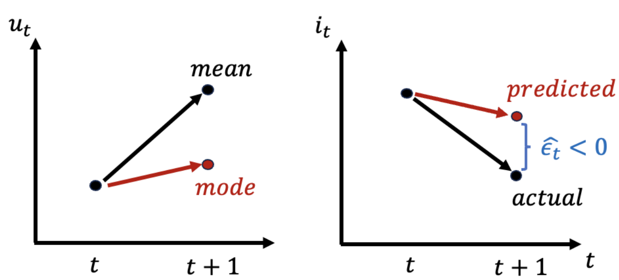
To support the first interpretation, we compare the direction of the unemployment forecast errors (Table 2) with the unemployment IRF differences. Figure 6, Panel (a) shows that when the Greenbook unemployment forecast is too optimistic, the Romer-Romer residual implies more easing (less tightening) relative to our shock. Those are instances in which the unemployment rate turns out higher than expected by the modal forecast and the conditional mean forecasts based on all information, including that in the text, would predict higher unemployment. Panel (b) shows how this pattern is directly consistent with an unemployment rate IRF that is *lower* than the true response, as in the BVAR. Suppose for simplicity that the interest rate i_t is set based only on the unemployment forecast $\mathbb{E}(u_{t+1})$ and we are in a situation where $\mathbb{E}(u_{t+1}) > mode(u_{t+1})$. Predicting i_t only with the modal forecast of u_t implies an easing shock, as made clear by Panel (a). But this means easing shocks are estimated when unemployment goes up. If this happens frequently in the sample, the resulting

Figure 6: HOW GREENBOOK FORECAST ERRORS AFFECT SHOCK ESTIMATES

(a) Relation between forecast errors and shock estimates



(b) Effect of too optimistic mode forecast on shock estimate



Notes. Panel (a) is a scatter plot of Greenbook forecast errors against the differences between the Romer-Romer residual and our preferred shock measure, for the 25 meetings with the largest differences. Panel (b) illustrates that when the Greenbook (i.e. mode) forecast is too optimistic, an easing shock is estimated (u_t denotes unemployment, i_t the target interest rate).

IRF will be incorrect because monetary policy easing and high unemployment are spuriously correlated. Using the conditional mean expectation, informed by the text-based sentiment, eliminates this spurious relation, giving an accurate IRF.

The second interpretation, in light of the response of stock prices, is that the Fed systematically reacts to equity markets, e.g. lowers the FFR after contractions in stock prices (Cieslak and Vissing-Jorgensen, 2020). If orthogonalizing the FFR only with respect to a small set of numerical forecasts does not control for this systematic feature of monetary policy, then the implied residuals might spuriously pick up a positive correlation between stock prices and the FFR, as observed in Panel (b). Instead, our sentiment indicators might reflect the relevant information about financial market developments that the FOMC considers.

To support the second interpretation, we implement the sign identification suggested by Jarocinski and Karadi (2020) where a monetary policy shock is identified as one which creates a negative comovement between interest rates and stock prices, while an informational shock creates a positive comovement. This is accomplished by using a second instrument, HF changes in the S&P500 Index around FOMC meetings in addition to the monetary policy instrument. Identification is achieved by ordering these two instruments first in a recursive scheme and imposing the sign restrictions. Figure G.1 in Appendix G shows the responses to a monetary policy shock obtained using this methodology. The IRFs using our shock measure look similar to their counterpart in Figure 5, making clear that our preferred shock measure already satisfies the additional sign restrictions.On the contrary, the sign restrictions alter the IRFs based on the Romer-Romer shock drastically. Imposing a negative comovement between interest rates and stock prices also "corrects" the activity, price and bond premia responses, which are now similar to our preferred measure and to what theory would predict.

The third interpretation of the difference between Panels (a) and (b) in Figure 5 is that the Romer-Romer residual based on forecasts only still contains endogenous variation with regards to credit spreads. Our sentiment indicators, on the other hand, account for the fact that the FOMC closely monitors developments in credit spreads. Indeed our set of sentiment indicators contains the sentiments around "spreads", "credit standards" as well as "credit quality". This third interpretation is supported by findings of a separate study by Caldara and Herbst (2019). These authors show that not accounting for the Fed's reaction to credit spreads attenuates the responses of several macroeconomic variables to monetary policy shocks measures, including the original Romer and Romer (2004) approach.

Additional results. Appendix G presents additional IRFs. First, it shows IRFs constructed with shocks from intermediate specifications of (3) (see Table 3). One noteworthy observation is that monetary policy shocks retrieved using the extended set of numerical forecasts, but without including sentiments (Romer-Romer ridge), already render the IRFs to be in line with theory. We emphasize that IRFs in line with consensus of the economic literature should be a necessary, but not a sufficient criterion for a good measure of monetary policy shocks.[23] Second,

[23] As shown in Section 3, errors from numerical forecasts are predictable using our sentiment indicators, a strong argument for including them when retrieving a monetary policy shock measure. Furthermore, Section 4 shows that the Romer-Romer ridge specification has an R^2 of only 0.55, as opposed to 0.94 in our preferred specification, implying an unappealingly strong

we construct IRFs based only on the 10 largest shocks in absolute value, setting all other elements of the shock time series to zero. These IRFs are of course more noisy, but we find that they display a very similar pattern to our main results in Figure 5. This finding underlines the relevance of our interpretation of large monetary shock episodes in Section 4.2. Third, the same appendix presents results analogous to Figure 5, but instead constructed using local projections (Jordà, 2005). As one would expect without the shrinkage imposed by the BVAR, the IRFs are generally noisier, but we confirm the general results using this alternative approach. Most notably, the shocks from the original Romer-Romer specification again result in responses of real activity and stock prices that are not in line with theory.

6 Extracting monetary policy shocks from recent FOMC meetings

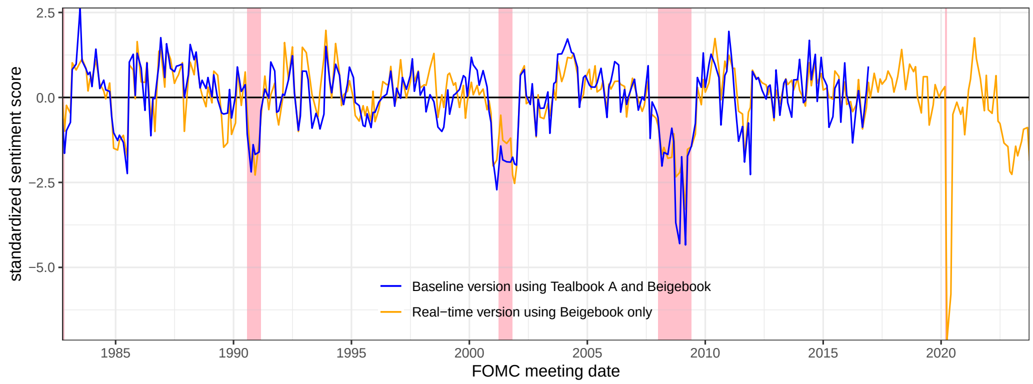
As an extension, we demonstrate how our method can be used to extract monetary policy shocks from the FOMC's more recent decisions. While the Tealbooks are available only with a five-year delay, the Beigebooks are available prior to every FOMC meeting. These summarize regional economic conditions for each individual Federal Reserve district. We already use the Beigebooks alongside the Tealbooks over our main sample period 1982-2008. The idea behind this section is to show that constructing our sentiment indicators only from the Beigebook text provides at least a limited proxy for the FOMC's information set.

We verify how well this proxy works: while in our main analysis we use both the Tealbook A and the Beigebook, we find that using only the Beigebook over our main 1982 to 2008 sample gives us strongly correlated sentiment indicators, as illustrated for "economic activity" in Figure 7. Running our main ridge regression with these sentiments, we find that the R^2 from using only Beigebook sentiments to estimate (3) is 0.68, compared to 0.94 with information from Tealbooks and Beigebooks combined. The resulting shocks have a correlation of 0.92 with each other. We further confirm that the BVAR IRFs we study in the previous section look qualitatively similar for the shocks constructed using only the Beigebook. It is important to emphasize that leveraging the Beigebooks is not possible in the original Romer and Romer (2004) approach, as the Beigebooks do not contain any numerical forecasts. This is a further advantage of our NLP approach.

As a "proof of concept", we run the Beigebook-only ridge, with 4 lags and

contribution of shocks to variation in the FFR target when only forecasts are included.

Figure 7: SENTIMENT SURROUNDING ECONOMIC ACTIVITY: BASELINE VS. BEIGEBOOK-ONLY



Notes. Sentiment around economic activity over time. Dark blue: indicator used for our main analysis based on Tealbook A and Beigebook. Orange: alternative version based on Beigebook only. The 5-year period after the blue line stops corresponds to the publication lag of the Tealbook and associated forecasts. Shaded areas represent NBER recessions.

squared terms, over the period December 2015 to October 2023, after the 2008 to 2015 zero lower bound period.[24] This is not feasible in our baseline because Tealbooks are not yet available over this sample. Towards the end of this period, our procedure measures sharp changes in the sentiment indicators around various economic concepts in the Beigebooks. For example, the sentiment around "inflation" drops massively in late 2021, with a reduction of more than 6 standard deviations (in terms of its 1982-2023 variability). A main contributors to this pattern is a sharp increase in the use of the negatively connotated word "concern" from the Loughran and McDonald (2011) in proximity to inflation. Other concepts around which the sentiment deteriorates strongly into negative territory in the runup to the first tightening decisions are "recession", "fuel", and "China".

We find that the R^2 from estimating the Beigebook-only version of (3) over the 2015 to 2023 sample is 0.98, suggesting only a small role for monetary policy shocks. Recall that this is the case despite the fact that we can only include the Beigebook sentiments, without using Tealbook sentiments and numerical forecasts which add significant predictive power in the 1982 to 2008 period. While the total increase in the FFR target between March 2022 and October 2023 amounted to 525 bp, the estimated shock component cumulates to around 21 bp over this period. In other words, our method implies that the tightening starting in 2022 entailed only mild contractionary monetary policy shocks.

[24]When estimating equation (3) in that sample, we exclude observations corresponding to the second zero lower bound period between March 2020 and December 2021.

To conclude, we think researchers should use our baseline measure whenever they can, even if it means dropping a number of observations at the end of their sample due to the availability of the Tealbooks. In situations where this will be very costly, the Beigebook-only version provides a viable alternative.

7 Conclusion

This paper develops a method for the identification of monetary policy shocks using natural language processing and machine learning. We show that including text-based information from the Greenbooks is crucial to summarize the Fed's information set. In response to our estimated shocks, economic activity and prices decline, bond premia rise, and stock prices fall after a tightening, in line with theory. Our analysis as a whole shows that the novel procedure proposed in this paper delivers a cleanly estimated series of monetary policy shocks.

References

ACOSTA, M. (2022): "A New Measure of Central Bank Transparency and Implications for the Effectiveness of Monetary Policy," *Working Paper*.

ANTOLIN-DIAZ, J. AND J. F. RUBIO-RAMIREZ (2018): "Narrative Sign Restrictions for SVARs," *American Economic Review*, 108, 2802–29.

BACHMANN, R., I. GÖDL-HANISCH, AND E. R. SIMS (2022): "Identifying monetary policy shocks using the central bank's information set," *Journal of Economic Dynamics and Control*, 145, 104555.

BARAKCHIAN, S. M. AND C. CROWE (2013): "Monetary policy matters: Evidence from new shocks data," *Journal of Monetary Economics*, 60, 950–966.

BAUER, M. D. AND E. T. SWANSON (2023): "An Alternative Explanation for the "Fed Information Effect"," *American Economic Review*, 113, 664–700.

BERNANKE, B. S., J. BOIVIN, AND P. ELIASZ (2005): "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach*," *The Quarterly Journal of Economics*, 120, 387–422.

BIANCHI, F., S. C. LUDVIGSON, AND S. MA (2022): "Belief distortions and macroeconomic fluctuations," *American Economic Review*, 112, 2269–2315.

CALDARA, D. AND E. HERBST (2019): "Monetary Policy, Real Activity, and Credit Spreads: Evidence from Bayesian Proxy SVARs," *American Economic Journal: Macroeconomics*, 11, 157–92.

Campbell, J. R., C. L. Evans, J. D. Fisher, and A. Justiniano (2012): "Macroeconomic effects of federal reserve forward guidance [with comments and discussion]," *Brookings papers on economic activity*, 1–80.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1999): "Chapter 2 Monetary policy shocks: What have we learned and to what end?" Elsevier, vol. 1 of *Handbook of Macroeconomics*, 65–148.

Cieslak, A., S. Hansen, M. McMahon, and S. Xiao (2021): "Policymakers' Uncertainty," *Available at SSRN 3936999*.

Cieslak, A. and A. Vissing-Jorgensen (2020): "The Economics of the Fed Put," *The Review of Financial Studies*, 34, 4045–4089.

Cloyne, J. and P. Hürtgen (2016): "The Macroeconomic Effects of Monetary Policy: A New Measure for the United Kingdom," *American Economic Journal: Macroeconomics*, 8, 75–102.

Cochrane, J. (2004): "Comments on 'A new measure of monetary shocks: Derivation and implications' by Christina Romer and David Romer," *July 17, 2004, presented at NBER EFG Meeting*.

Coibion, O. (2012): "Are the Effects of Monetary Policy Shocks Big or Small?" *American Economic Journal: Macroeconomics*, 4, 1–32.

Doh, T., D. Song, and S.-K. Yang (2022): "Deciphering federal reserve communication via text analysis of alternative fomc statements," *Working Paper*.

Drechsel, T. (2023): "Estimating the Effects of Political Pressure on the Fed: A Narrative Approach with New Data," *Working paper*.

Faust, J. and J. H. Wright (2008): "Efficient forecast tests for conditional policy forecasts," *Journal of Econometrics*, 146, 293–303, honoring the research contributions of Charles R. Nelson.

——— (2009): "Comparing Greenbook and Reduced Form Forecasts Using a Large Realtime Dataset," *Journal of Business & Economic Statistics*, 27, 468–479.

Gardner, B., C. Scotti, and C. Vega (2021): "Words speak as loudly as actions: Central bank communication and the response of equity prices to macroeconomic announcements," *Journal of Econometrics*.

Gertler, M. and P. Karadi (2015): "Monetary Policy Surprises, Credit Costs, and Economic Activity," *American Economic Journal: Macroeconomics*, 7, 44–76.

Giannone, D., M. Lenza, and G. E. Primiceri (2022): "Economic predictions with big data: The illusion of sparsity," *Econometrica*.

GORODNICHENKO, Y., T. PHAM, AND O. TALAVERA (2023): "The voice of monetary policy," *American Economic Review*, 113, 548–584.

GÜRKAYNAK, R. S., B. SACK, AND E. SWANSON (2005): "The Sensitivity of Long-Term Interest Rates to Economic News: Evidence and Implications for Macroeconomic Models," *American Economic Review*, 95, 425–436.

HANDLAN, A. (2020): "Text Shocks and Monetary Surprises: Text Analysis of FOMC Statements with Machine Learning," *Working Paper*.

HANSEN, S., M. MCMAHON, AND A. PRAT (2018): "Transparency and deliberation within the FOMC: a computational linguistics approach," *The Quarterly Journal of Economics*, 133, 801–870.

HASSAN, T. A., S. HOLLANDER, L. VAN LENT, AND A. TAHOUN (2022): "The Global Impact of Brexit Uncertainty," *The Journal of Finance (Forthcoming)*.

HOLM, M., P. PAUL, AND A. TISCHBIREK (2021): "The Transmission of Monetary Policy under the Microscope," *Journal of Political Economy*, 129, 2861–2904.

JAROCINSKI, M. AND P. KARADI (2020): "Deconstructing Monetary Policy Surprises—The Role of Information Shocks," *American Economic Journal: Macroeconomics*, 12, 1–43.

JAYAWICKREMA, J. AND E. T. SWANSON (2023): "Speeches by the Fed Chair Are More Important than FOMC Announcements: An Improved High-Frequency Measure of US Monetary Policy Shocks," *Working paper*.

JORDÀ, O. (2005): "Estimation and Inference of Impulse Responses by Local Projections," *American Economic Review*, 95, 161–182.

KALAMARA, E., A. TURRELL, C. REDL, G. KAPETANIOS, AND S. KAPADIA (2020): "Making text count: economic forecasting using newspaper text," .

LEEPER, E. M., C. A. SIMS, AND T. ZHA (1996): "What does monetary policy do?" *Brookings papers on economic activity*, 1996, 1–78.

LOUGHRAN, T. AND B. MCDONALD (2011): "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, 66, 35–65.

MIRANDA-AGRIPPINO, S. AND G. RICCO (2021): "The Transmission of Monetary Policy Shocks," *American Economic Journal: Macroeconomics*, 13, 74–107.

MULLAINATHAN, S. AND J. SPIESS (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

NAKAMURA, E. AND J. STEINSSON (2018): "High-frequency identification of monetary non-neutrality: the information effect," *The Quarterly Journal of Economics*, 133, 1283–1330.

OCHS, A. C. (2021): "A New Monetary Policy Shock with Text Analysis," *Faculty of Economics, University of Cambridge*.

PLAGBORG-MOLLER, M. AND C. K. WOLF (2021): "Local Projections and VARs Estimate the Same Impulse Responses," *Econometrica*, 89, 955–980.

RAMEY, V. A. (2016): "Macroeconomic shocks and their propagation," *Handbook of macroeconomics*, 2, 71–162.

REIFSCHNEIDER, D. AND P. TULIP (2019): "Gauging the uncertainty of the economic outlook using historical forecasting errors: The Federal Reserve's approach," *International Journal of Forecasting*, 35, 1564–1582.

ROMER, C. D. AND D. H. ROMER (2000): "Federal Reserve Information and the Behavior of Interest Rates," *American Economic Review*, 90, 429–457.

——— (2004): "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 94, 1055–1084.

SHAPIRO, A. H. AND D. J. WILSON (2021): "Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis," *The Review of Economic Studies*, 89, 2768–2805.

SHARPE, S. A., N. R. SINHA, AND C. A. HOLLRAH (2020): "The Power of Narratives in Economic Forecasts," *Working Paper*.

SWANSON, E. T. (2021): "Measuring the effects of federal reserve forward guidance and asset purchases on financial markets," *Journal of Monetary Economics*, 118, 32–53.

TENREYRO, S. AND G. THWAITES (2016): "Pushing on a String: US Monetary Policy Is Less Powerful in Recessions," *American Economic Journal: Macroeconomics*, 8, 43–74.

THORNTON, D. L. (2005): "A new federal funds rate target series: September 27, 1982-december 31, 1993," *FRB of St. Louis Working Paper No*.

——— (2006): "When Did the FOMC Begin Targeting the Federal Funds Rate? What the Verbatim Transcripts Tell Us," *Journal of Money, Credit and Banking*, 38, 2039–2071.

UHLIG, H. (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure," *Journal of Monetary Economics*, 52, 381–419.

WIELAND, J. F. AND M.-J. YANG (2020): "Financial dampening," *Journal of Money, Credit and Banking*, 52, 79–113.

ONLINE APPENDIX TO

Identifying Monetary Policy Shocks: A Natural Language Approach

by S. Borağan Aruoba and Thomas Drechsel

A Algorithm to combine and exclude concepts

The below algorithm describes how we deal with overlapping economic concepts in Step 2 of our procedure, which is described in Section 2 of the main text.

1. Start with triples. Go through the list of triples that have at least 250 mentions (around one per meeting on average). Select triples that are economic concepts (based on judgment).

2.a) Go through the list of doubles that have at least 500 mentions. Select doubles that are economic concepts (based on judgment).

2.b) **IF** a selected double is a subset of one or several triples:

- Unselect the double and keep the triple(s) **IF**
  [***Criterion 1***] the triples close to add up to the double **AND**
  [***Criterion 2***] the triples are sufficiently different concepts
  OR
  [***Criterion 3***] the double by itself is too ambiguous
- **ELSE**: keep the double and unselect the triple(s)

3.a) Go through the list of singles that have at least 2000 mentions. Select singles that are economic concepts (based on judgment).

3.b) **IF** a selected single is a subset of one or several doubles:

- Unselect the single and keep the double(s) **IF**
  [***Criterion 1***] the doubles close to add up to the single **AND**
  [***Criterion 2***] the doubles are sufficiently different concepts
  OR
  [***Criterion 3***] the single by itself is too ambiguous
- **ELSE** Keep the single and unselect the double(s)

END

An example of ***Criterion 1*** and ***Criterion 2*** being satisfied is for: "commercial real estate" and "residential real estate". The occurrences of these two triples almost exactly add up to the occurrences of the double "real estate". Since they are also sufficiently different concepts (e.g. capture meaningfully different markets and thus span richer information), we kept the two triples.

An example ***Criterion 1*** not being satisfied and ***Criterion 3*** not being satisfied is for the single "credit". While there are doubles such as "consumer credit" and "bank credit", the overall occurrence of credit is much larger than the associated doubles. So we decided to keep credit.

An example ***Criterion 1*** not being satisfied and ***Criterion 3*** satisfied is for the single "expenditures". Unlike credit, this single by itself is too vague based on our judgment (as "capital expenditures" and "government expenditures" are quite different). We therefore selected the doubles, even though their added-up occurrence is well below the one of "expenditures" by itself.

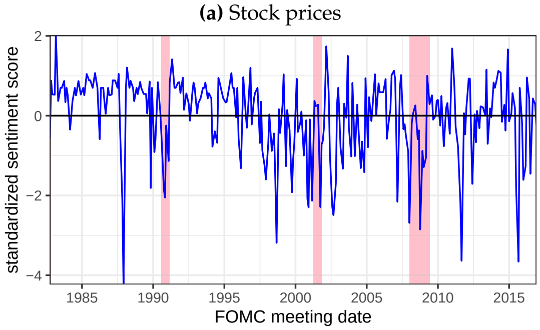
After going through algorithm, we also applied to following additional steps to clean up the list:

- Sometimes a concept occurred as a singular and a plural, for example "oil price" and "oil prices". In this case, we add them up.
- Sometimes the algorithm produced different concepts that are quite similar, which we unified. For example "stock prices" and "equity prices". We add them up.
- In a few instances we selected singles and doubles separately for the same single. For example "employment" and "employment cost".
- We also added one quadruple: "money market mutual funds."

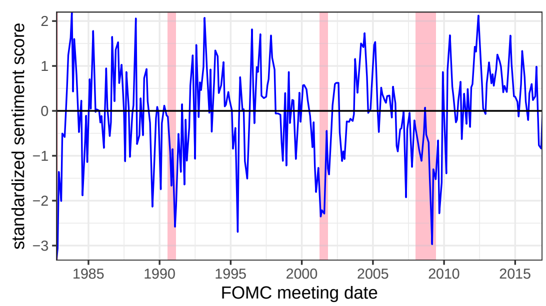
B Additional sentiment indicators

Figure B.1: SELECTED SENTIMENT INDICATORS

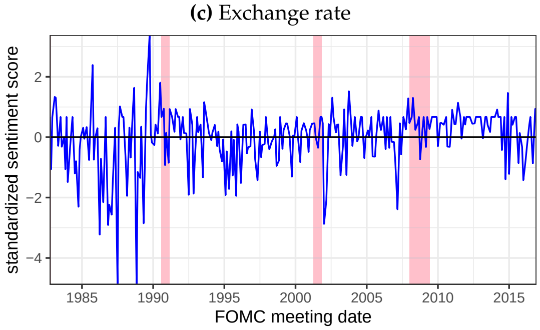
(a) Stock prices



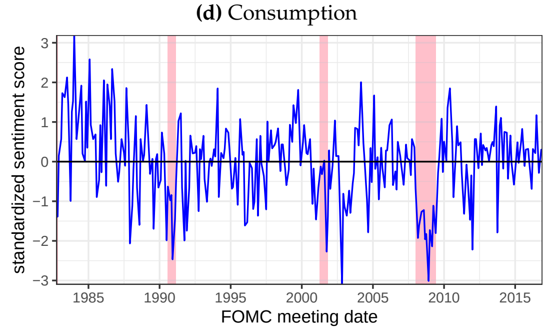
(b) Inventories



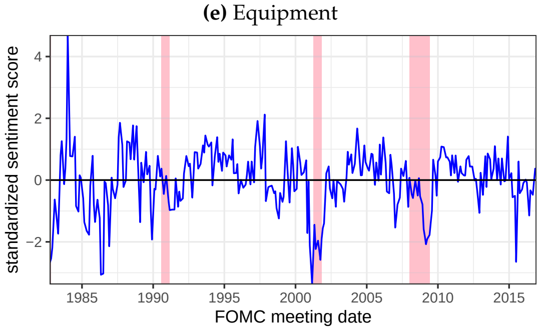
(c) Exchange rate



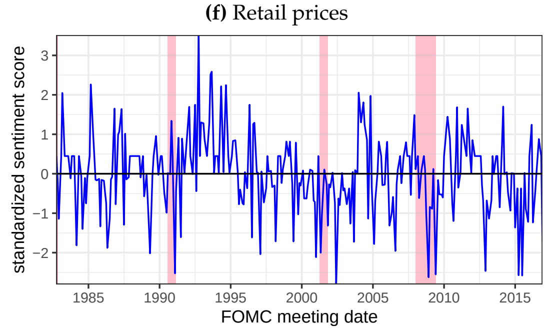
(d) Consumption



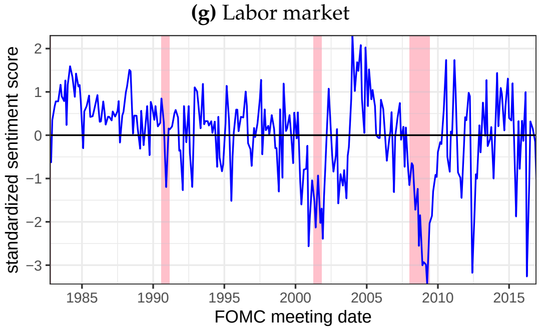
(e) Equipment



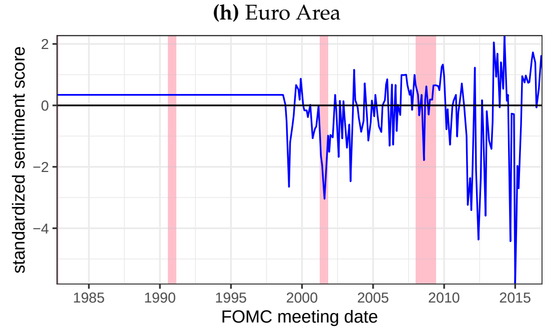
(f) Retail prices



(g) Labor market



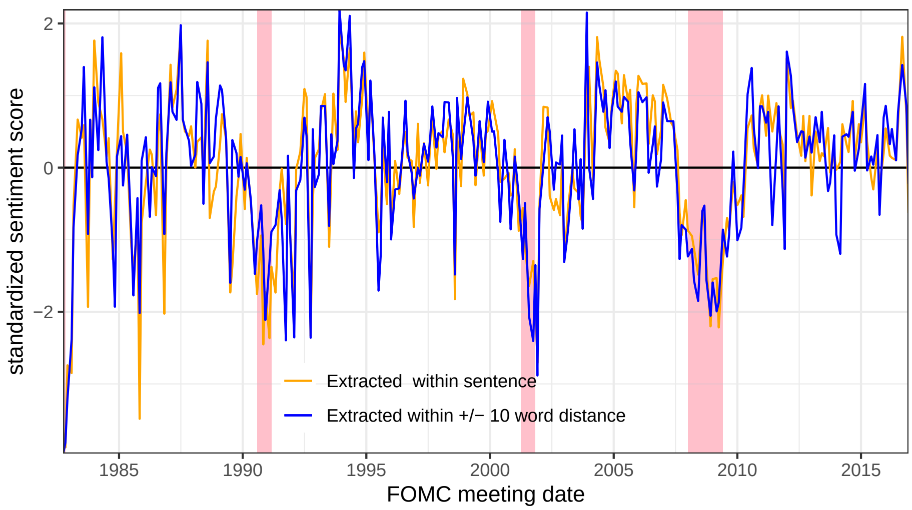
(h) Euro Area



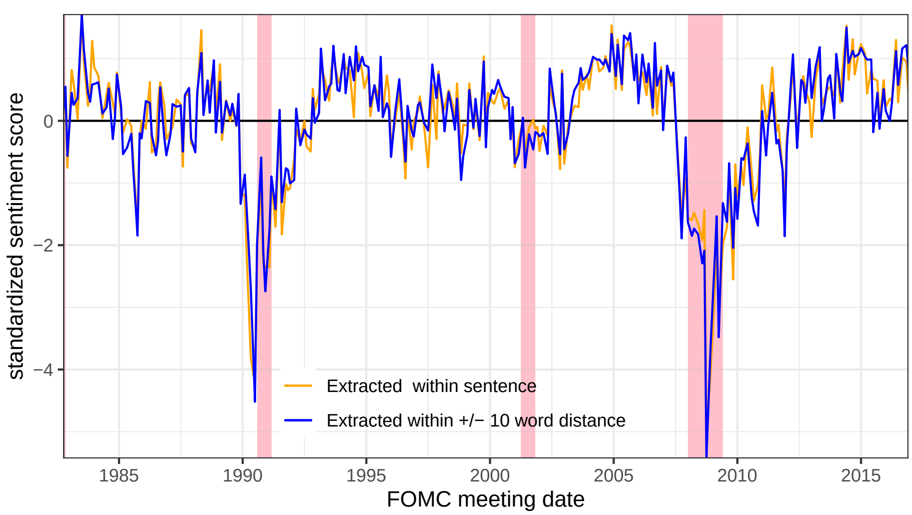
C Sentiments in +/- 10 word distance vs. in sentences

Figure C.1: SENTIMENT INDICATORS CONSTRUCTED IN ALTERNATIVE WAYS

(a) Employment



(b) Credit



Notes. Two examples of sentiment indicators constructed based on positive and negative words within +/- 10 word window vs. based on positive and negative words within the same sentence. See discussion in Section 2.2. For the sentiment surrounding employment the correlation across the two alternative indicators is 0.875. For the case of credit sentiment, the correlation is 0.959. Shaded areas represent NBER recessions.

D Evidence for the modal nature of Greenbook forecasts

We systematically check the transcripts of the FOMC meetings in our sample period 1982 to 2016 for mentions of the terms "modal" and "modal forecasts" and then read the discussions around those instances. Below we provide several examples, spanning all decades over our sample period, that indicate that the staff and members of the FOMC interpret the Greenbook forecasts as modal in nature.

- In the **February 1985** meeting, Governor Wallich asks the staff "Could I ask a question on that? The greater probability is the number on a skewed distribution. Presumably, the probability distribution of inflation is that it can't go much below zero but it can go up quite far; it has a long right hand tail. Are you thinking in terms of the mode–the most likely single value–or the mean, including the tail?"

  The director of Research and Statistics James L. Kichline responds "We have alleged for years that we have a modal forecast. I would say that it's very difficult, but basically, if we use the model and try to come out with confidence intervals, the model comes out with substantially lower rates of inflation. In fact, if you put a 70 percent confidence interval around our deflator estimate, a couple of times we drift out of that range on the high side. So with the same policy assumptions for 1985, the model forecast, for whatever it's worth, is a rate of increase in the deflator one percentage point less than in the staff forecast. I view that information as saying that the risks tend to be skewed on the down side. We think 3-1/2 percent is the most likely outcome; but if we're wrong, I'd say we're probably too high rather than too low."

- In the **July 1996** meeting Michael Prell, the director of Research and Statistics clarifies: "I think there have been some occasions when we have indicated that the risks in our outlook were asymmetric. I would characterize our forecasts over the years as an effort to present a meaningful, modal forecast of the most likely outcome. When we felt that there was some skewness to the probability distribution, we tried to identify it. In this instance, as we looked at the recent data, we felt that there was a greater thickness in the area of our probability distribution a little above our modal forecast."

  [This is the example we provide in the main text.]

- In the **November 2001** meeting, Governor Meyer states, in reference to the 9/11 terrorist attacks that "The Greenbook, like most forecasts, seems to assume a one-time terrorist attack with a near-term effect on confidence that dissipates over time.

That might be appropriate for a modal forecast. But relative to this assumption, there seems to be significant asymmetric downside risks, specifically of further terrorist attacks that affect confidence in the economy or perhaps for other reasons as well. The forecast for the first state of the world is therefore likely to be biased in an optimistic direction though, as David Stockton noted, we would be hard pressed to parameterize the downside risks associated with the second state of the world. Still this analysis suggests that the mean of the forecast might be interpreted as being below the mode in this case. So the question is how policy should respond to this type of uncertainty and whether policy should be set to err on the side of ease relative to the modal forecast."

- In the **March 2005** meeting, President of the Federal Reserve Bank of San Francisco Janet Yellen states that "While the Greenbook expectation of a relatively flat path for bond rates through the end of next year may be a reasonable modal forecast, I don't think the risks here are balanced."

- In the **June 2009** meeting, FOMC secretary Brian Madigan lays out different policy options, with reference to the forecasts: "With both a modal outlook for weak growth and low inflation, and downside risks around the outlook for activity, macroeconomic considerations would seem to argue for providing additional monetary policy stimulus at this juncture. However, with the federal funds rate at the zero bound, the Committee has limited policy options at its disposal."

- In the **June 2011** meeting, President of the Federal Reserve Bank of San Francisco John Williams explains "Furthermore, despite the deep cuts to the output projection, the Tealbook has also shifted to a downside skew to the risks of the growth outlook. This combination of a downward modal revision to the growth forecast and downside risk assessment is a truly sobering development, but it's consistent with what we see in financial markets."

- In the **December 2016** meeting, Vice Chairman Dudley says "I guess my view of the risks to the forecast is that you have a modal forecast and then you ask, where is the skew of the distribution? It's not about where the lower bound lies relative to the funds rate. So I guess I interpret the balance of the risks differently (...)."

E More results on forecast error predictability

E.1 Additional results for output and inflation forecasts

Table E.1: ADDITIONAL GREENBOOK FORECAST ERROR PREDICTABILITY TESTS

	Panel (a): unemployment rate forecast errors on LHS							
	current quarter	1 quarter ahead	1 year ahead	2 years ahead	current quarter	1 quarter ahead	1 year ahead	2 years ahead
First PC of all sentiments	-0.029*	-0.114**	-0.445**	-0.622**				
	[0.016]	[0.049]	[0.190]	[0.238]				
Economic activity sentiment					-0.026	-0.098**	-0.285*	-0.363**
					[0.016]	[0.048]	[0.165]	[0.171]
Constant	-0.019	-0.070**	-0.082	0.059	-0.019	-0.069**	-0.077	0.160
	[0.014]	[0.033]	[0.121]	[0.201]	[0.014]	[0.035]	[0.145]	[0.258]
R^2	0.045	0.149	0.248	0.208	0.033	0.097	0.090	0.055
Number of observations	210	210	210	62	210	210	210	62
	Panel (b): output forecast errors on LHS							
	current quarter	1 quarter ahead	1 year ahead	2 years ahead	current quarter	1 quarter ahead	1 year ahead	2 years ahead
First PC of all sentiments	0.121	0.411	0.540*	-0.171				
	[0.220]	[0.325]	[0.310]	[0.402]				
Economic activity sentiment					0.036	0.146	0.079	-0.485
					[0.228]	[0.272]	[0.251]	[0.403]
Constant	0.300*	0.139	-0.252	-0.380	0.298*	0.131	-0.268	0.442
	[0.167]	[0.276]	[0.340]	[0.750]	[0.163]	[0.299]	[0.374]	[0.717]
R^2	0.005	0.030	0.049	0.003	0.000	0.003	0.001	0.021
Number of observations	206	204	198	54	206	204	198	54
	Panel (c): inflation forecast errors on LHS							
	current quarter	1 quarter ahead	1 year ahead	2 years ahead	current quarter	1 quarter ahead	1 year ahead	2 years ahead
First PC of all sentiments	0.148	0.170	0.142	-0.011				
	[0.101]	[0.133]	[0.173]	[0.164]				
Economic activity sentiment					0.263***	0.222*	0.236*	0.013
					[0.092]	[0.126]	[0.141]	[0.214]
Constant	-0.163	-0.136	-0.267	0.056	-0.167	-0.140	-0.271	-0.019
	[0.109]	[0.167]	[0.208]	[0.216]	[0.103]	[0.160]	[0.201]	[0.207]
R^2	0.029	0.032	0.017	0.013	0.081	0.049	0.041	0.000
Number of observations	210	210	210	62	210	210	210	62

Notes. Panel (a) repeats Table 2 from the main text. Panels (b) and (c) show analogous results for real output growth and inflation forecasts.

E.2 Results for first release instead of final vintage

Table E.2: GREENBOOK FORECAST ERROR PREDICTABILITY TESTS FOR FIRST RELEASE

	Panel (a): unemployment rate forecast errors on LHS							
	current quarter	1 quarter ahead	1 year ahead	2 years ahead	current quarter	1 quarter ahead	1 year ahead	2 years ahead
First PC of all sentiments	-0.025*	-0.104**	-0.433**	-0.637**				
	[0.013]	[0.045]	[0.189]	[0.242]				
Economic activity sentiment					-0.020	-0.089*	-0.272	-0.376**
					[0.014]	[0.044]	[0.166]	[0.173]
Constant	-0.032***	-0.084***	-0.097	0.048	-0.032***	-0.083*	-0.093	0.150
	[0.011]	[0.031]	[0.119]	[0.240]	[0.011]	[0.032]	[0.142]	[0.260]
R^2	0.038	0.129	0.240	0.214	0.020	0.084	0.084	0.058
Number of observations	210	210	210	62	210	210	210	62
	Panel (b): output forecast errors on LHS							
	current quarter	1 quarter ahead	1 year ahead	2 years ahead	current quarter	1 quarter ahead	1 year ahead	2 years ahead
First PC of all sentiments	-0.093	0.172	0.327	-0.291				
	[0.125]	[0.256]	[0.282]	[0.245]				
Economic activity sentiment					-0.144	0.052	-0.069	-0.551*
					[0.131]	[0.235]	[0.228]	[0.318]
Constant	0.214**	0.070	-0.236	-0.348	0.218**	0.067	-0.241	-0.374
	[0.103]	[0.192]	[0.256]	[0.568]	[0.106]	[0.200]	[0.283]	[0.535]
R^2	0.006	0.009	0.024	0.015	0.012	0.001	0.001	0.045
Number of observations	206	204	198	54	206	204	198	54
	Panel (c): inflation forecast errors on LHS							
	current quarter	1 quarter ahead	1 year ahead	2 years ahead	current quarter	1 quarter ahead	1 year ahead	2 years ahead
First PC of all sentiments	0.104	0.049	0.116	0.062				
	[0.091]	[0.093]	[0.126]	[0.155]				
Economic activity sentiment					0.201**	0.098	0.232**	0.130
					[0.087]	[0.093]	[0.115]	[0.196]
Constant	-0.167**	-0.133	-0.281*	-0.483**	-0.170**	-0.135	-0.285**	-0.470**
	[0.079]	[0.123]	[0.155]	[0.214]	[0.073]	[0.120]	[0.143]	[0.212]
R^2	0.018	0.003	0.013	0.004	0.059	0.010	0.046	0.012
Number of observations	210	210	210	62	210	210	210	62

Notes. This table repeats Table E.1, based on the outcome being the first release (constructed from ALFRED) rather than the final vintage of each variable.

F Construction of committee composition variables

The additional data set that captures information on the composition of the FOMC in each meeting, which we use for robustness, is constructed as follows. For each FOMC meeting, we record the list of participants. This list consists of the governors at the board as well as the representatives from each regional bank. Typically, regional bank representatives are their respective presidents, except in cases where there is an interim president. We classify the participants by their voting status: they are either voting members, alternate members, or non-voting members. The governors always vote and the regional bank presidents alternate between the three roles. For each governor, we create a dummy variable that equals 1 if he/she attended a given meeting and 0 otherwise. We record the attendance of each regional bank representative in a similar way. Here we create three sets of dummy variables. The first set of variables are constructed at the participant-position-voting status level, meaning for example that we distinguish between Mr. Boehne (president of the FRB of Philadelphia) when he is attending as a voting member and when he is attending as a non-voting member. The second set of variables are constructed only at the participant-position level, without regard to their voting statuses. The last set of variables recorded whether a regional bank's representative voted during the meeting for each of the 12 banks. For governors, we also record information on who appointed them. We tally the total number of governors in attendance by the US president who made the appointment, as well as the number of governors appointed by a Republican and Democratic administration respectively.[1] In addition to attendance, for each meeting we record the number of motions voted upon and the results of each vote. Indicator variables are constructed for whether there is only one vote during the meeting, whether there is not a vote at all, and in the case that there is one vote, whether the voting result was unanimous. Lastly, we tally the total number of female participants in attendance at each meeting. Over the sample period 1982:10 to 2008:10, this results in 298 variables.

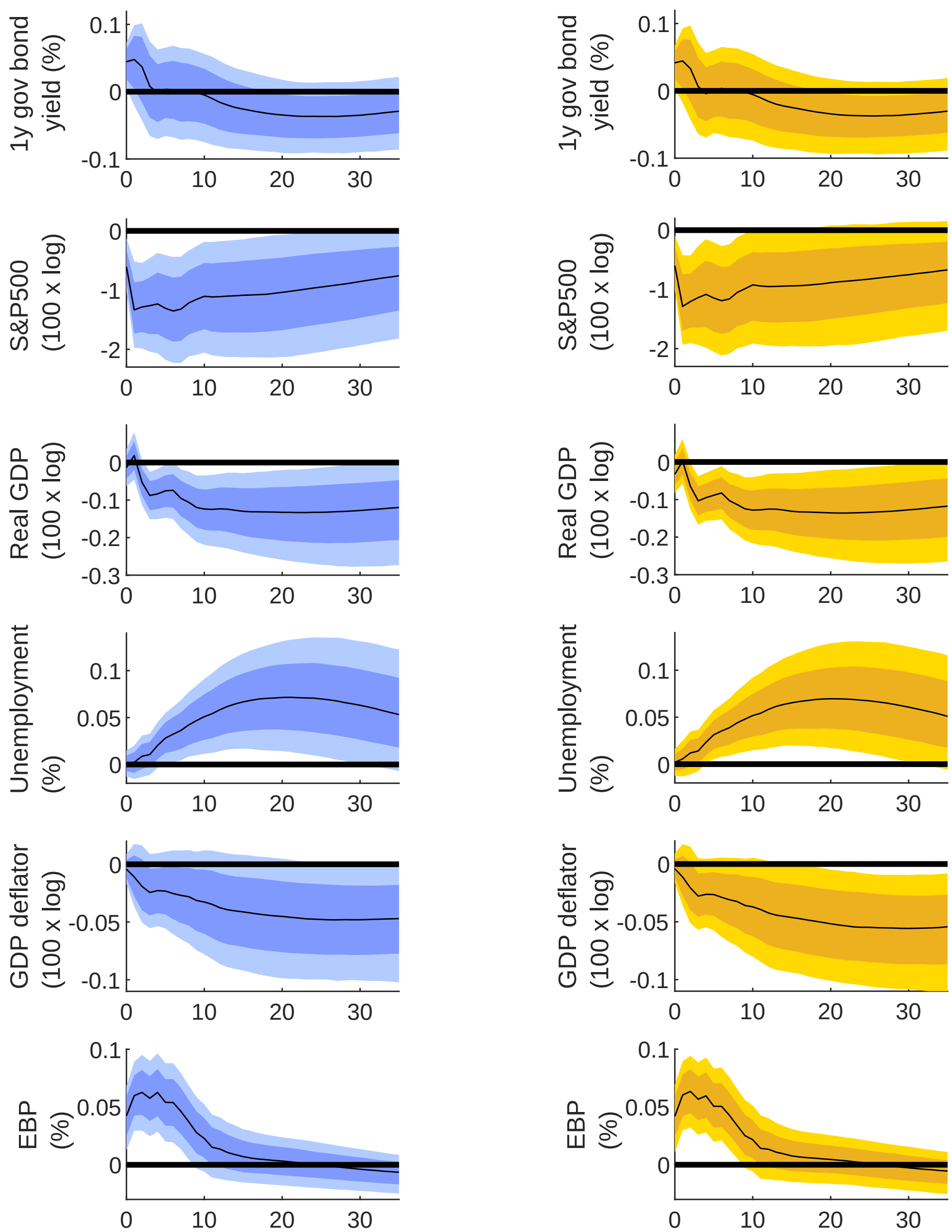
[1] In the case that a governor served multiple tenures appointed by different US presidents, we make that distinction. For example, Janet Yellen was appointed by Bill Clinton to serve as a governor in 1994 and then by Barack Obama in 2010 – and these are recorded separately.

G Additional IRFs

Figure G.1: IRFS CONSTRUCTED WITH ADDITIONAL SIGN RESTRICTIONS

(a) Using shocks from full ridge model

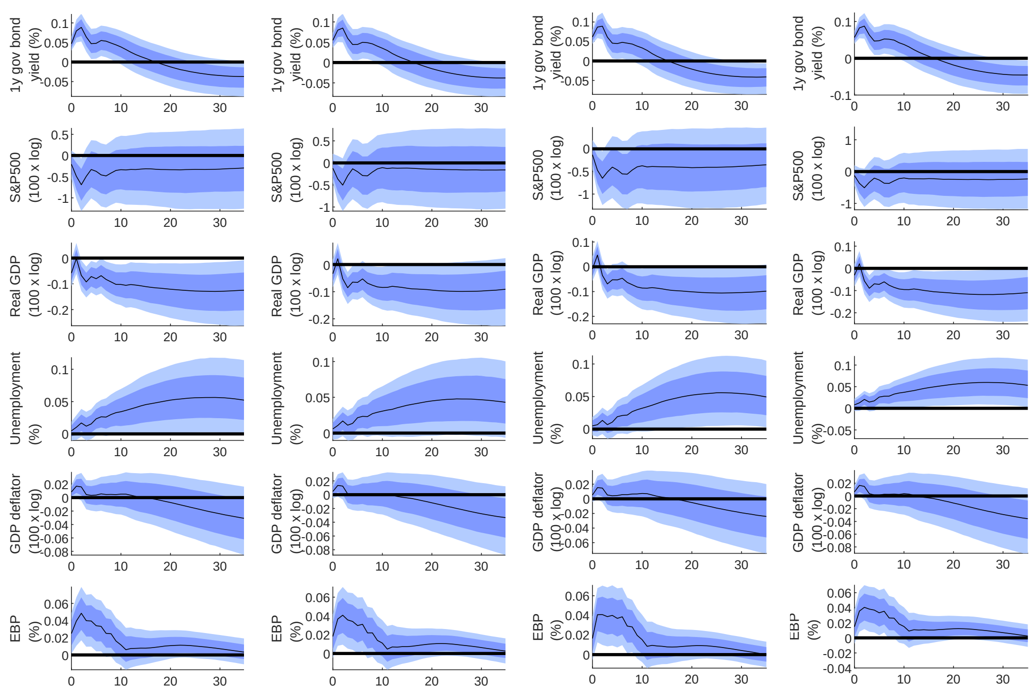
(b) Using shocks from Romer-Romer OLS



Notes. The two panels correspond to those in Figure 5, but impose the additional sign restrictions suggested by Jarocinski and Karadi (2020) to separate monetary policy shocks from central bank information shocks. Specifically, the IRFs shown here are for monetary policy shocks which are assumed to create a negative covariance between interest rates and stock prices.

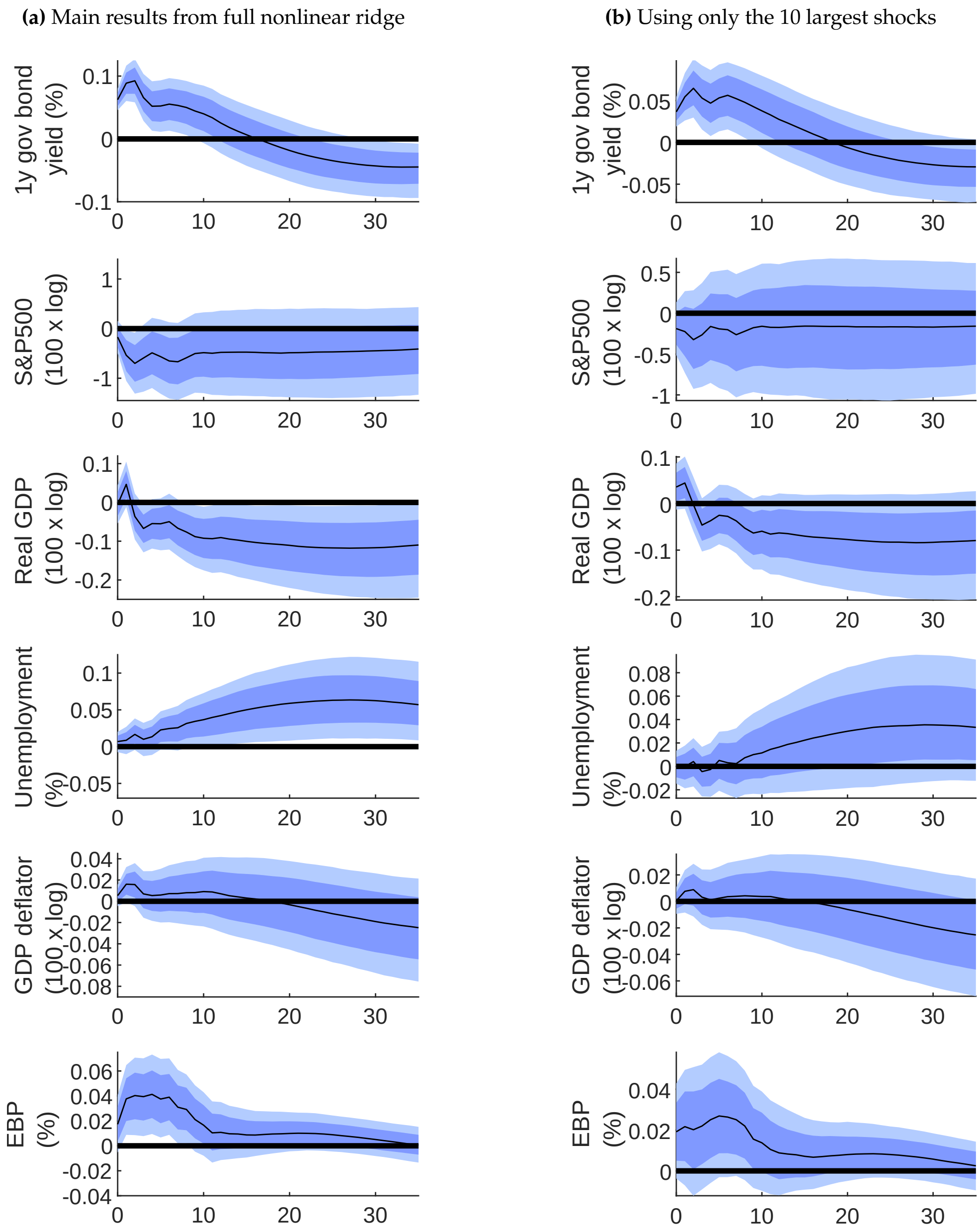
Figure G.2: IRFS ESTIMATED FROM INTERMEDIATE SHOCK VERSIONS

(a) Romer-Romer ridge **(b)** Linear only **(c)** Linear and lags **(d)** Nonlinear but no lags



Notes. IRFs to different intermediate versions of the estimated monetary policy shocks, computed from the BVAR model. Panel (a) shows the IRFs to the shocks from an empirical specification where only the extended set of forecasts are used in a ridge regression. Panel (b) uses the measure of monetary policy shocks retrieved from a linear instead of nonlinear ridge model using the extended set of numerical forecasts and sentiment indicators, but where no lags or squared sentiment indicators are included. Panel (c) is similar to Panel (b) but the specification to estimate the shocks also adds lagged sentiments. Panel (d) is similar to Panel (b) but the specification to estimate the shocks also adds squared terms. The sample period to estimate the shocks is 1982:10-2008:10. The solid line represents the median, the 16th and 84th percentiles are represented by the darker bands, and the 5th and 95th percentiles by the lighter bands. The sample used to estimate the IRFs is 1984:02-2016:12.

Figure G.3: IRFS TO ALL SHOCKS AND THE 10 LARGEST SHOCKS IN COMPARISON

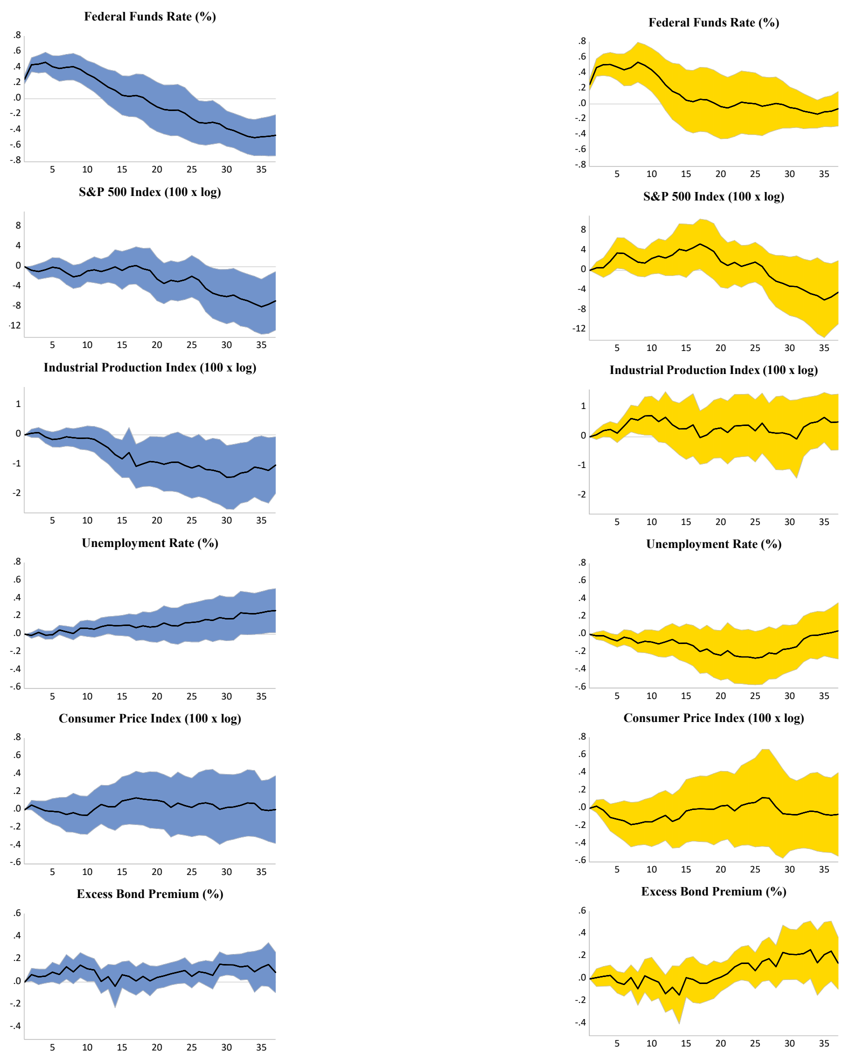


Notes. Panel (a) repeats our main IRFs (Figure 5, Panel (a)). Panel (b) applies the same BVAR specification but only using the 10 largest observations in absolute value for the time series of the monetary policy shocks, setting the shock for all other meetings to zero. The sample period to estimate the shocks is 1982:10-2008:10. The solid line represents the median, the 16th and 84th percentiles are represented by the darker bands, and the 5th and 95th percentiles by the lighter bands. The sample used to estimate the IRFs is 1984:02-2016:12.

Figure G.4: IRFS TO DIFFERENT MONETARY POLICY SHOCKS USING LOCAL PROJECTIONS

(a) Using shocks from full ridge model

(b) Using shocks from Romer-Romer OLS



Notes. IRFs analogous to Figure 5 in the main text, but based on a frequentist local projections approach (Jordà, 2005) rather than a BVAR. Panel (a) uses our proposed measure of monetary policy shocks, estimated using the full nonlinear ridge model on the extended set of numerical forecasts and our sentiment indicators from FOMC documents. Panel (b) shows the analogous IRFs when a simpler empirical specification is used to estimate the shocks, which includes only the original set of numerical forecasts in a Romer-Romer OLS regression. The solid line represents the median, and the 5th and 95th percentiles are captures by the bands. The sample used to estimate the IRFs is 1984:02-2008:10.