Online Appendix For

"The Voice of Monetary Policy"

By Yuriy Gorodnichenko, Tho Pham, Oleksandr Talavera

Appendix A

Appendix Table A1. Voice tone for responses during Q&A sessions

Press conference date	Speaker	Positive responses	Neutral responses	Negative responses	Tone
April 27, 2011	Bernanke	17	0	1	0.89
June 22, 2011	Bernanke	19	0	0	1.00
November 2, 2011	Bernanke	19	0	0	1.00
January 25, 2012	Bernanke	18	0	0	1.00
April 25, 2012	Bernanke	19	0	0	1.00
June 20, 2012	Bernanke	22	0	1	0.91
September 13, 2012	Bernanke	23	0	0	1.00
December 12, 2012	Bernanke	20	0	3	0.74
March 20, 2013	Bernanke	14	0	7	0.33
June 19, 2013	Bernanke	10	0	11	-0.05
September 18, 2013	Bernanke	1	0	17	-0.89
December 18, 2013	Bernanke	18	0	3	0.71
March 19, 2014	Yellen	7	5	4	0.27
June 18, 2014	Yellen	2	0	14	-0.75
September 17, 2014	Yellen	2	1	9	-0.64
December 17, 2014	Yellen	1	4	10	-0.82
March 18, 2015	Yellen	15	0	5	0.50
June 17, 2015	Yellen	1	3	13	-0.86
September 17, 2015	Yellen	16	1	1	0.88
December 16, 2015	Yellen	4	1	13	-0.53
March 16, 2016	Yellen	12	1	3	0.60
June 15, 2016	Yellen	11	0	4	0.47
September 21, 2016	Yellen	5	0	14	-0.47
December 14, 2016	Yellen	12	5	3	0.60
March 15, 2017	Yellen	9	1	8	0.06
June 14, 2017	Yellen	7	0	9	-0.12
September 20, 2017	Yellen	4	1	9	-0.38
December 13, 2017	Yellen	1	5	12	-0.85
March 21, 2018	Powell	0	0	20	-1.00
June 13, 2018	Powell	0	0	22	-1.00
September 26, 2018	Powell	9	0	15	-0.25
December 19, 2018	Powell	0	0	21	-1.00
January 30, 2019	Powell	16	2	7	0.39
March 20, 2019	Powell	25	0	1	0.92
May 1, 2019	Powell	18	0	5	0.57
June 19, 2019	Powell	0	0	20	-1.00

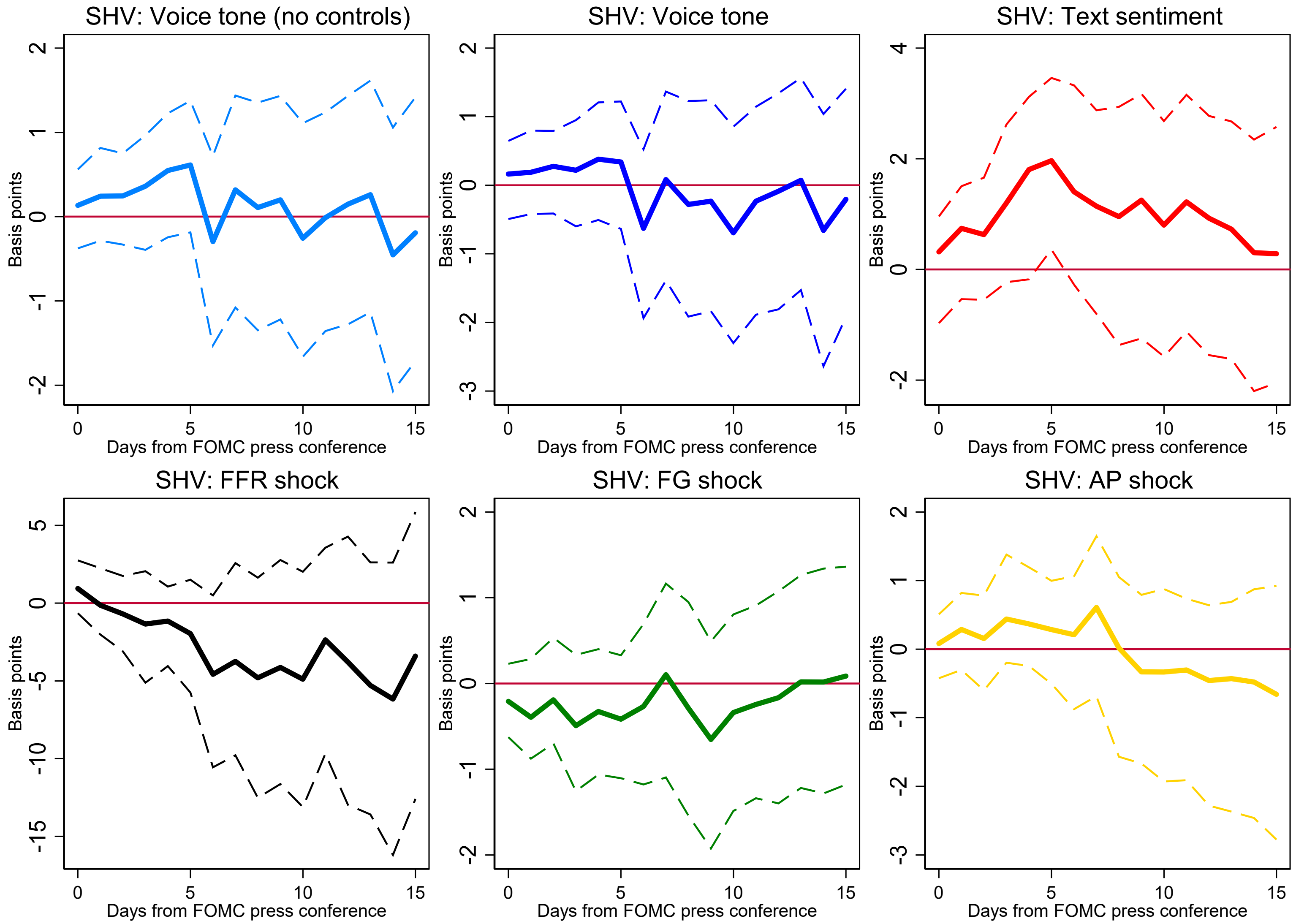
Notes: This table shows the number of positive, negative, and neutral responses as well as the aggregate voice tone for each press conference in the sample.

Appendix Table A2. Text sentiment for statement, remarks and Q&A

Press conference date	Speaker	Text Sentiment	Press conference date	Speaker	Text Sentiment
26/01/2011	Bernanke	1.00	29/04/2015	Yellen	0.60
15/03/2011	Bernanke	1.00	17/06/2015	Yellen	0.50
27/04/2011	Bernanke	0.24	29/07/2015	Yellen	1.00
22/06/2011	Bernanke	0.41	17/09/2015	Yellen	0.31
09/08/2011	Bernanke	1.00	28/10/2015	Yellen	0.60
21/09/2011	Bernanke	1.00	16/12/2015	Yellen	0.06
02/11/2011	Bernanke	0.61	27/01/2016	Yellen	1.00
13/12/2011	Bernanke	1.00	16/03/2016	Yellen	0.25
25/01/2012	Bernanke	0.55	27/04/2016	Yellen	0.33
13/03/2012	Bernanke	1.00	15/06/2016	Yellen	0.54
25/04/2012	Bernanke	0.71	27/07/2016	Yellen	1.00
20/06/2012	Bernanke	0.80	21/09/2016	Yellen	0.20
01/08/2012	Bernanke	1.00	02/11/2016	Yellen	1.00
13/09/2012	Bernanke	0.30	14/12/2016	Yellen	0.45
24/10/2012	Bernanke	1.00	01/02/2017	Yellen	1.00
12/12/2012	Bernanke	0.19	15/03/2017	Yellen	-0.04
30/01/2013	Bernanke	0.60	03/05/2017	Yellen	1.00
20/03/2013	Bernanke	0.47	14/06/2017	Yellen	0.26
01/05/2013	Bernanke	1.00	26/07/2017	Yellen	1.00
19/06/2013	Bernanke	0.58	20/09/2017	Yellen	0.48
31/07/2013	Bernanke	1.00	01/11/2017	Yellen	1.00
18/09/2013	Bernanke	0.67	13/12/2017	Yellen	0.00
30/10/2013	Bernanke	1.00	31/01/2018	Yellen	0.00
18/12/2013	Bernanke	0.53	21/03/2018	Powell	0.10
29/01/2014	Bernanke	0.20	02/05/2018	Powell	1.00
19/03/2014	Yellen	0.33	13/06/2018	Powell	-0.03
30/04/2014	Yellen	0.67	01/08/2018	Powell	-1.00
18/06/2014	Yellen	0.46	26/09/2018	Powell	-0.07
30/07/2014	Yellen	0.00	08/11/2018	Powell	0.00
17/09/2014	Yellen	0.50	19/12/2018	Powell	0.17
29/10/2014	Yellen	0.50	30/01/2019	Powell	0.52
17/12/2014	Yellen	0.28	20/03/2019	Powell	0.60
28/01/2015	Yellen	0.60	01/05/2019	Powell	0.50
18/03/2015	Yellen	0.26	19/06/2019	Powell	-0.15

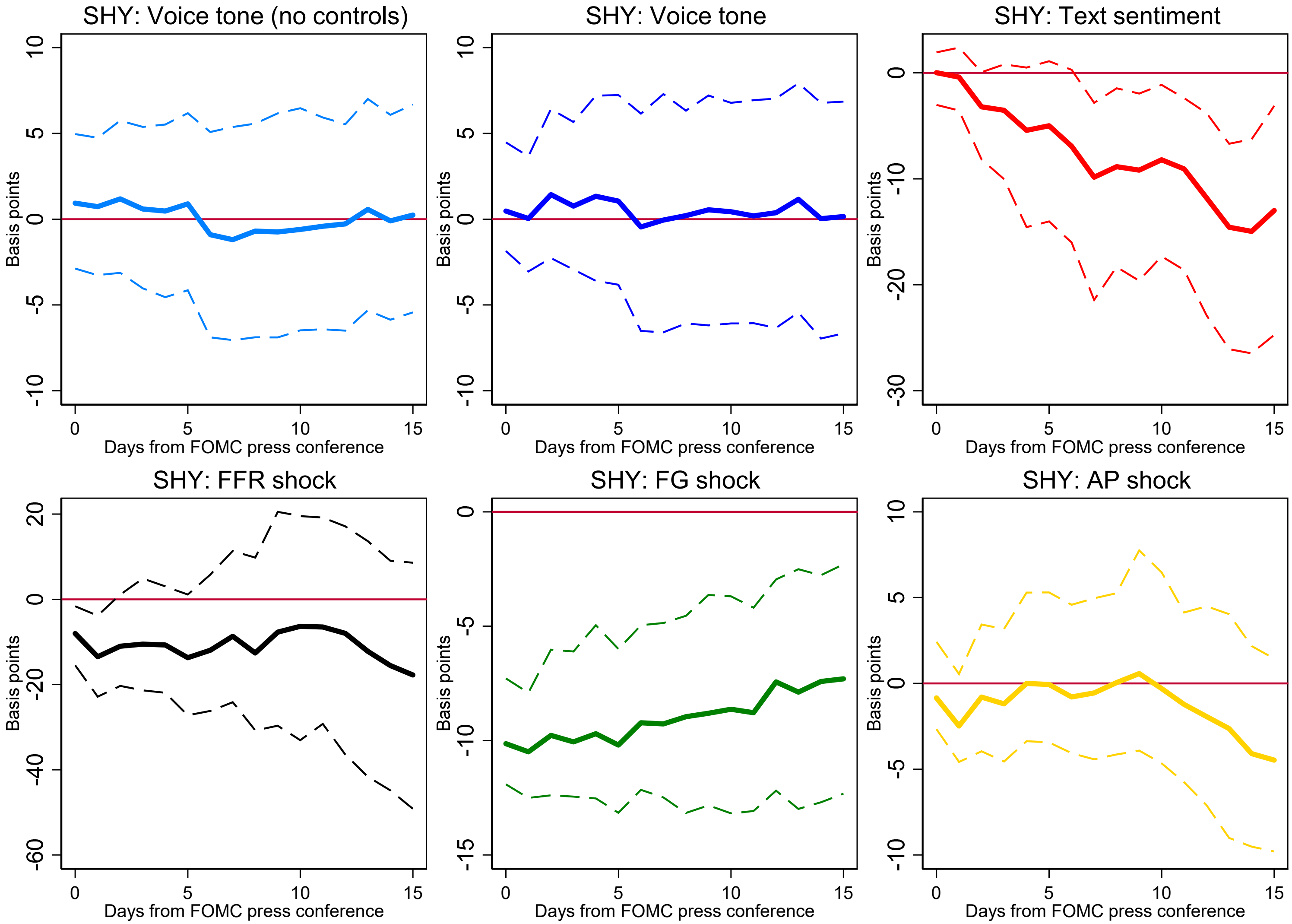
Notes: This table shows the aggregate text sentiment for each FOMC meeting in the sample.

Appendix Figure A1. Response of SHV ETF (Short Treasury Bond ETF; maturities one year or less) to policy actions and messages



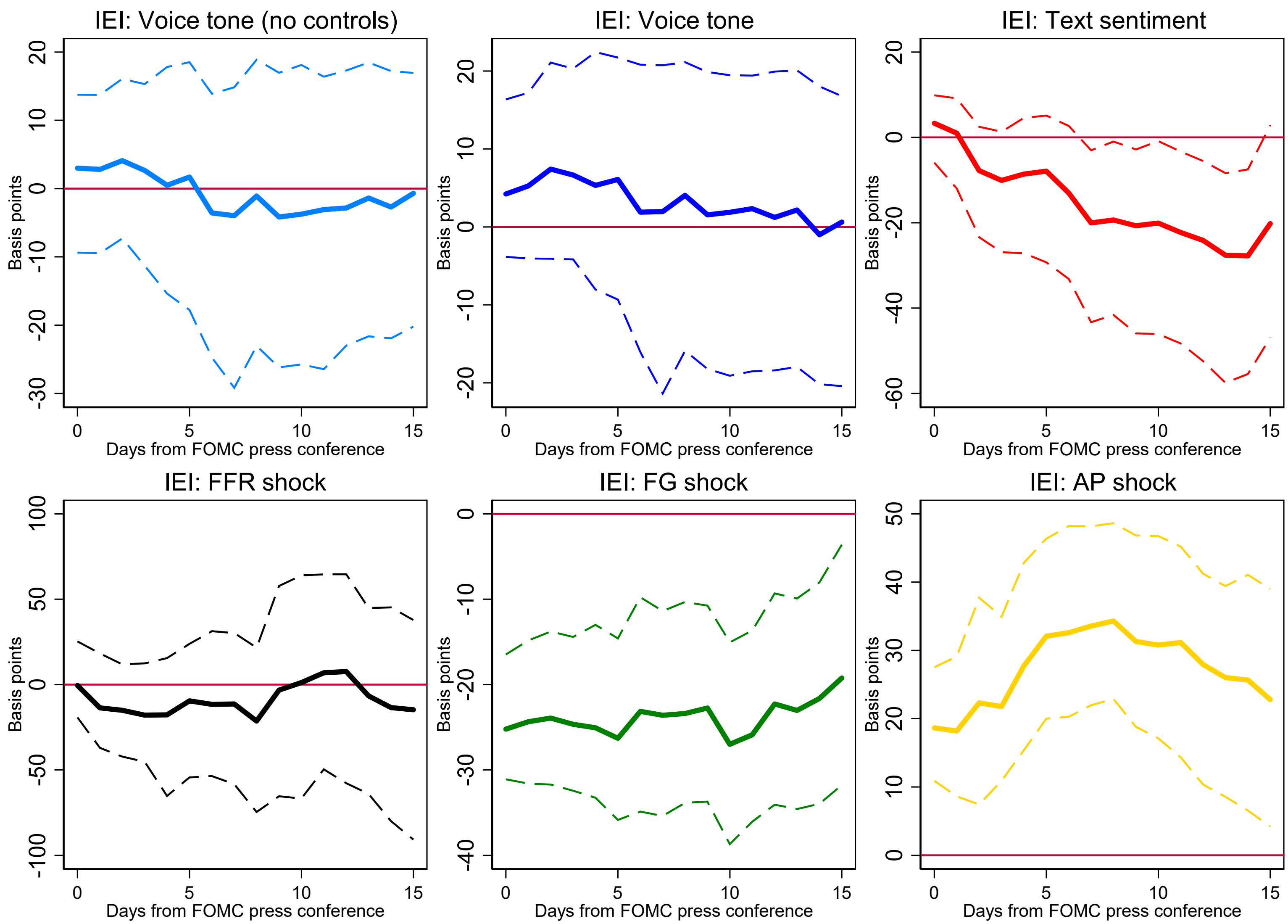
Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for policy communication/actions. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure A2. Response of SHY ETF (1-3 Year Treasury Bond ETF) to policy actions and messages



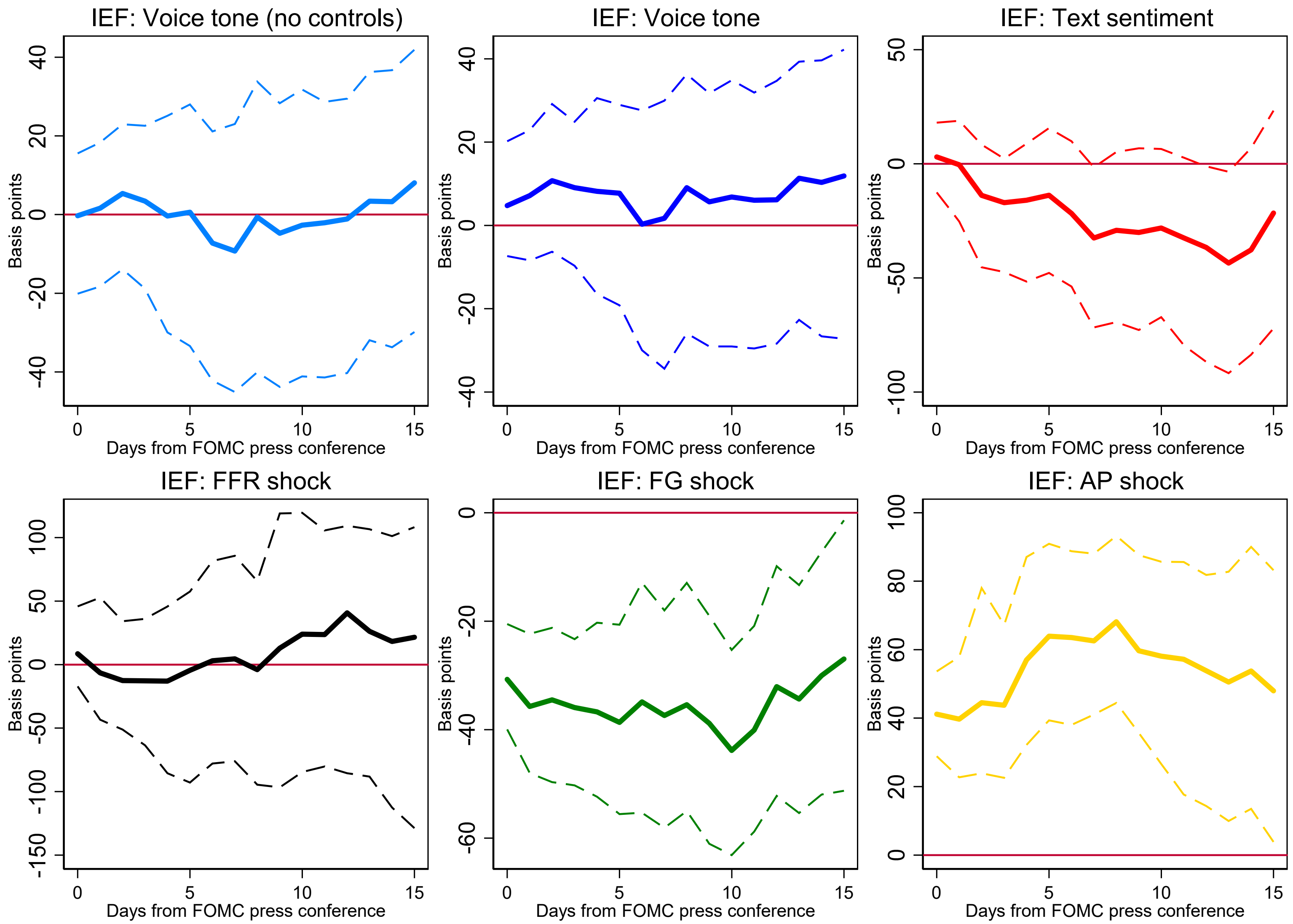
Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for policy communication/actions. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure A3. Response of IEI ETF (3-7 Year Treasury Bond ETF) to policy actions and messages



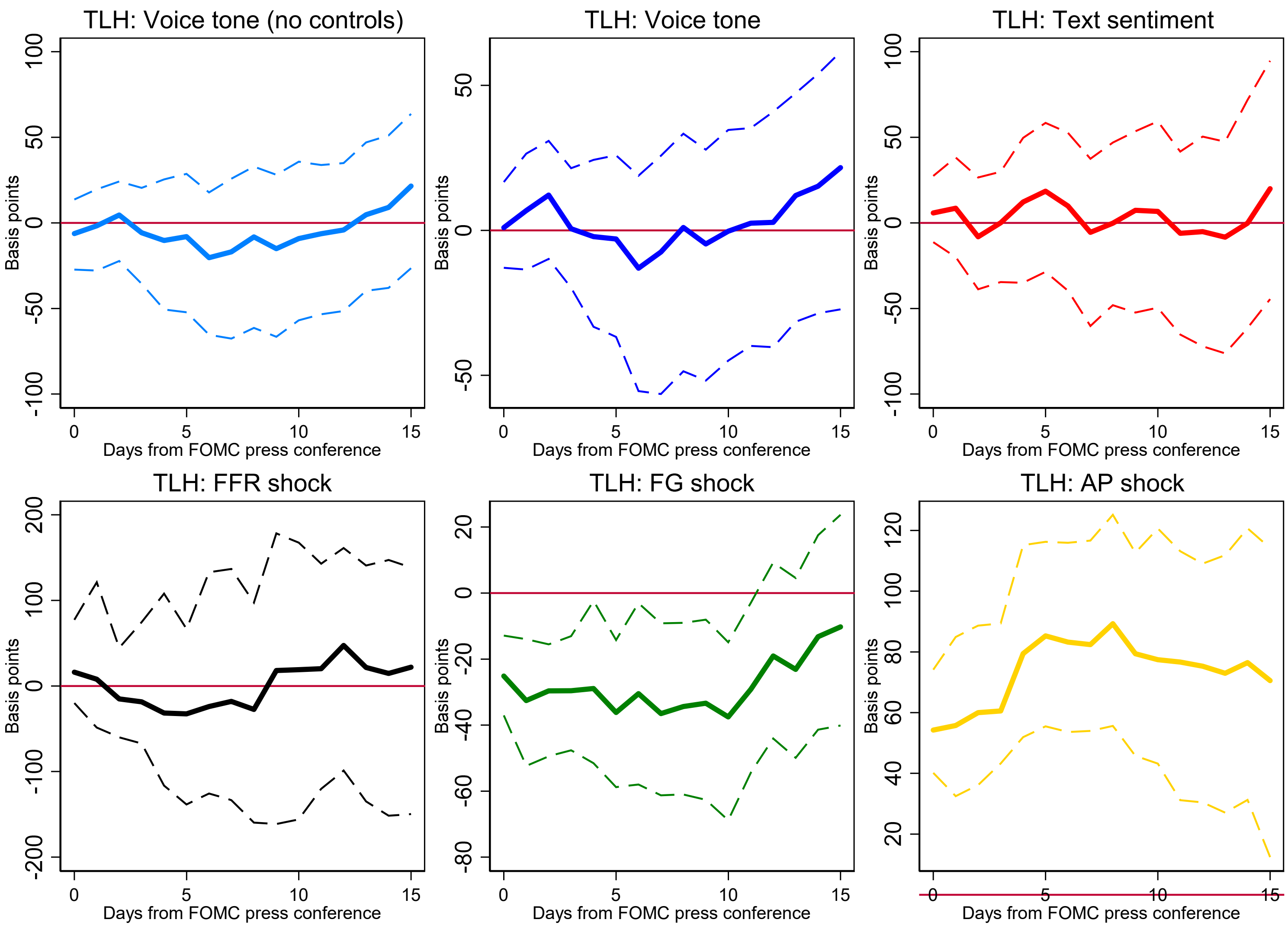
Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for policy communication/actions. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure A4. Response of IEF ETF (7-10 Year Treasury Bond ETF) to policy actions and messages



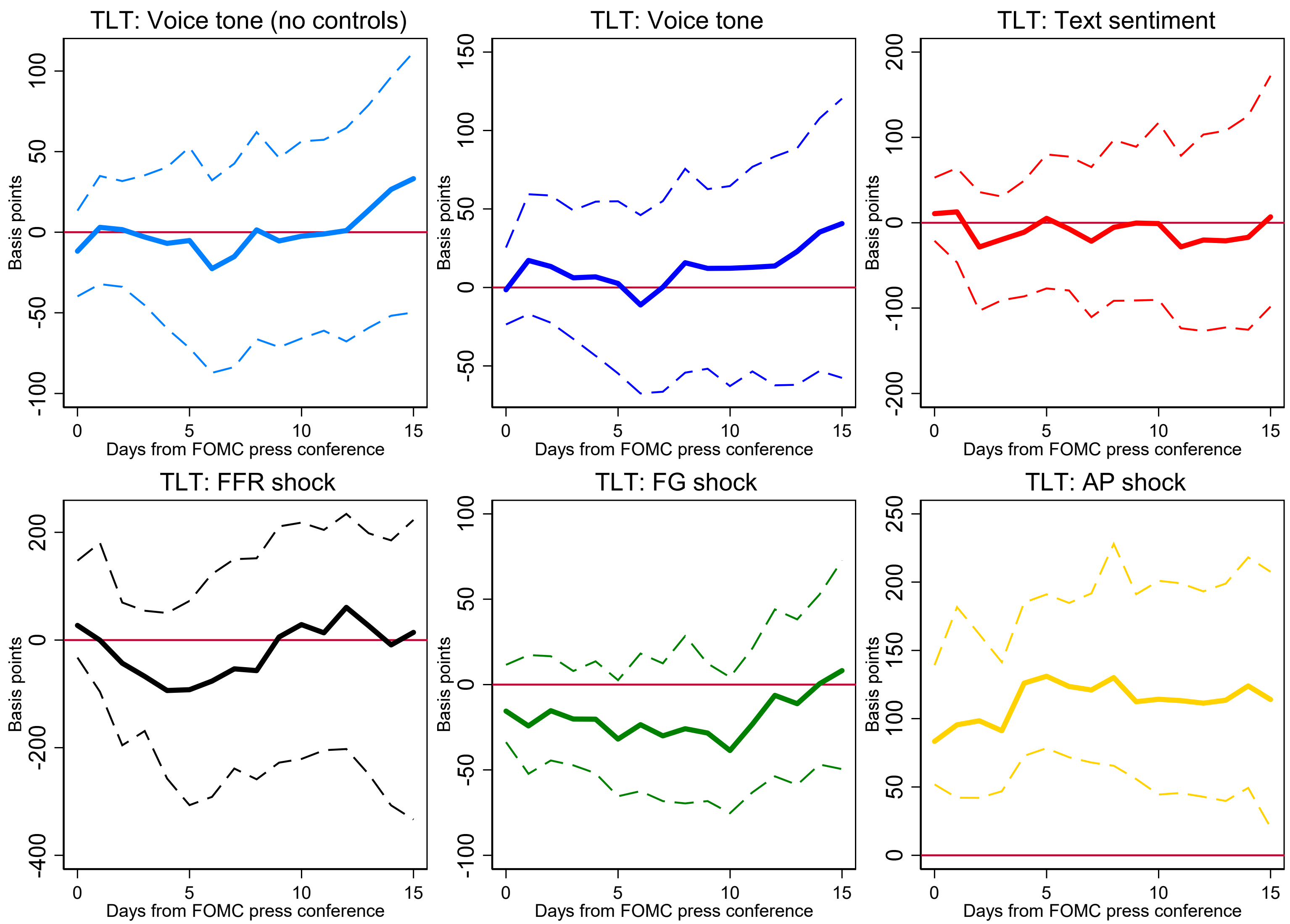
Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for policy communication/actions. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure A5. Response of TLH ETF (10-20 Year Treasury Bond) to policy actions and messages



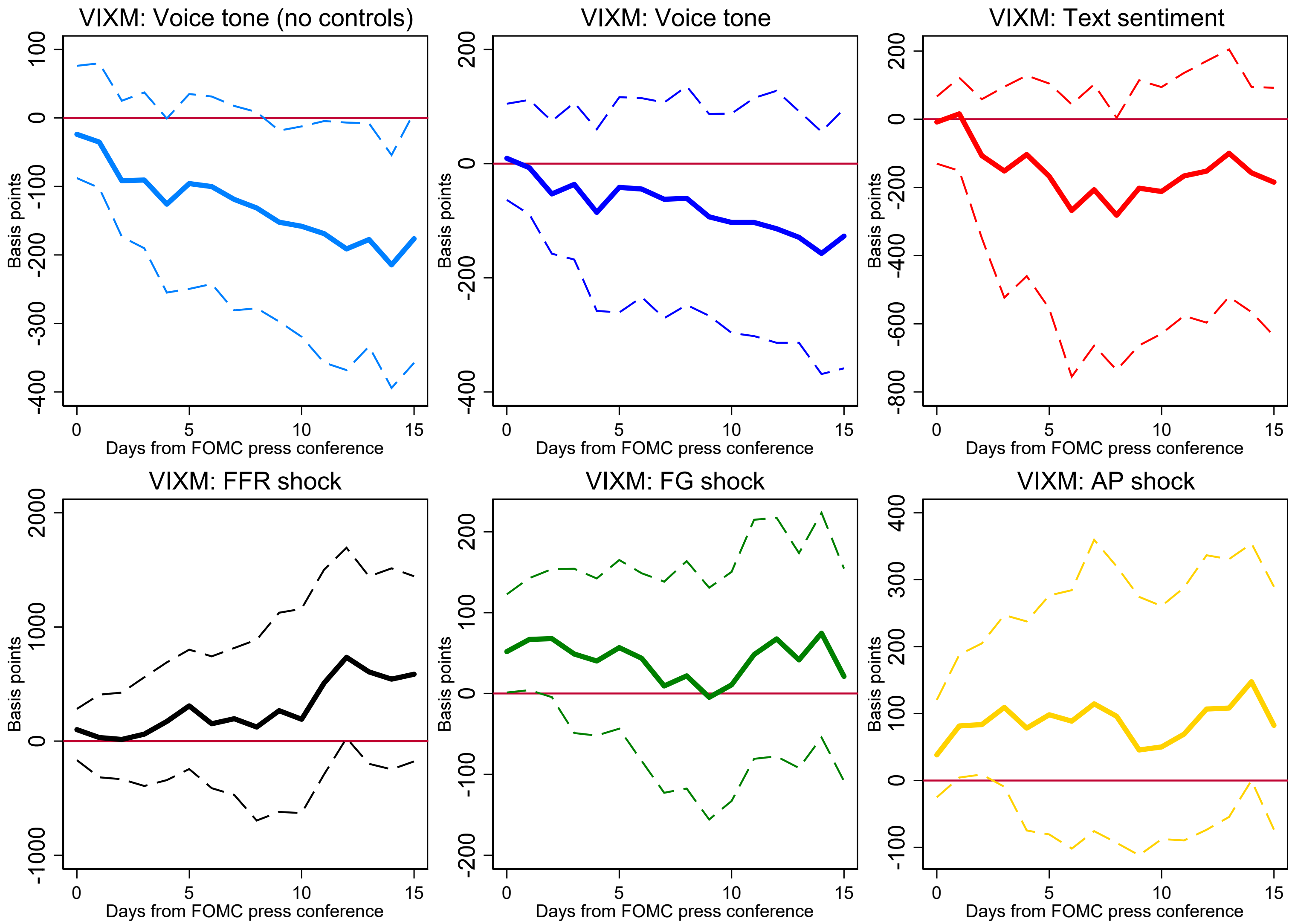
Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for policy communication/actions. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure A6. Response of TLT ETF (20+ Year Treasury Bond ETF) to policy actions and messages



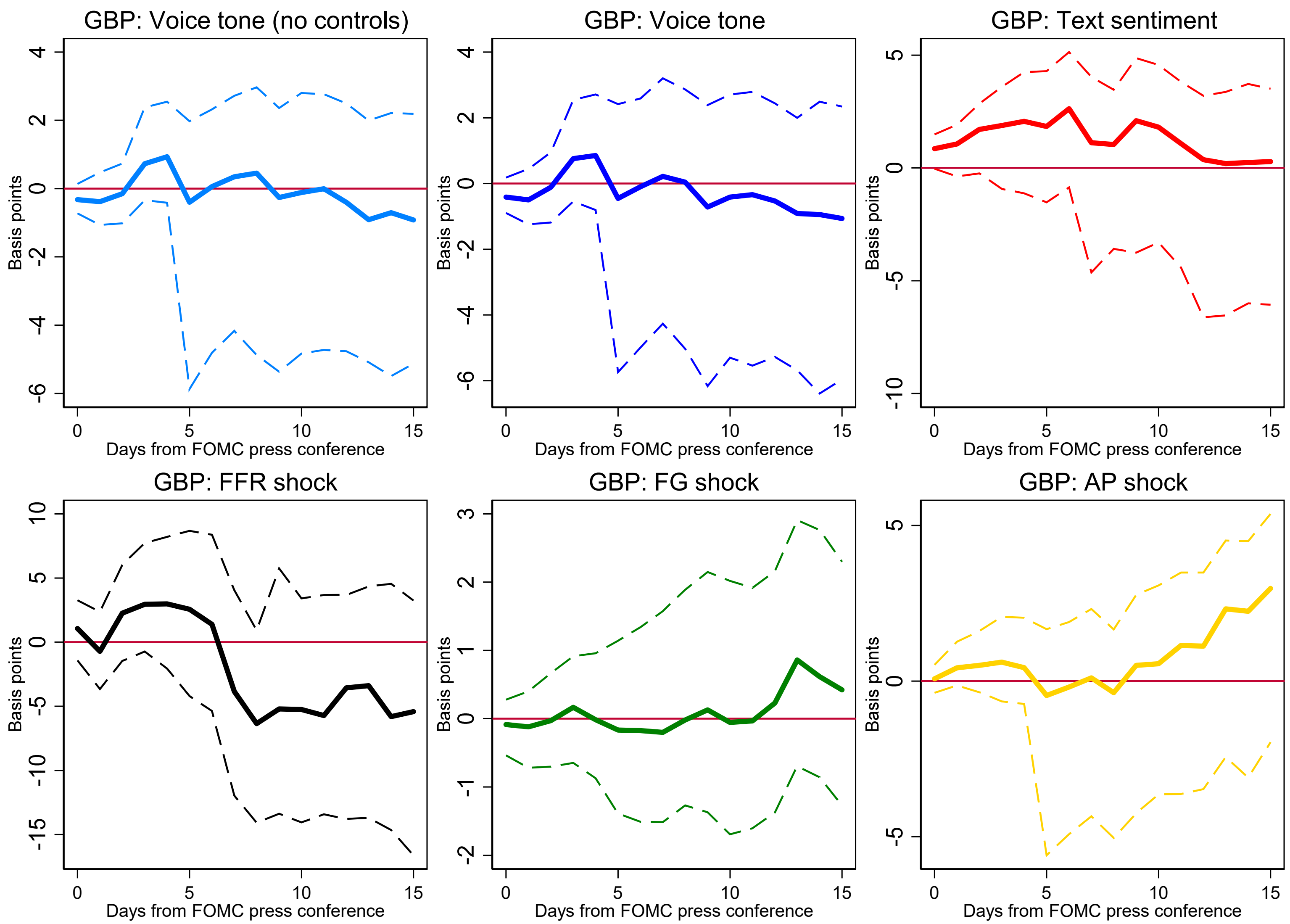
Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for policy communication/actions. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure A7. Response of VIXM ETF (VIX Mid-Term Futures ETF) to policy actions and messages



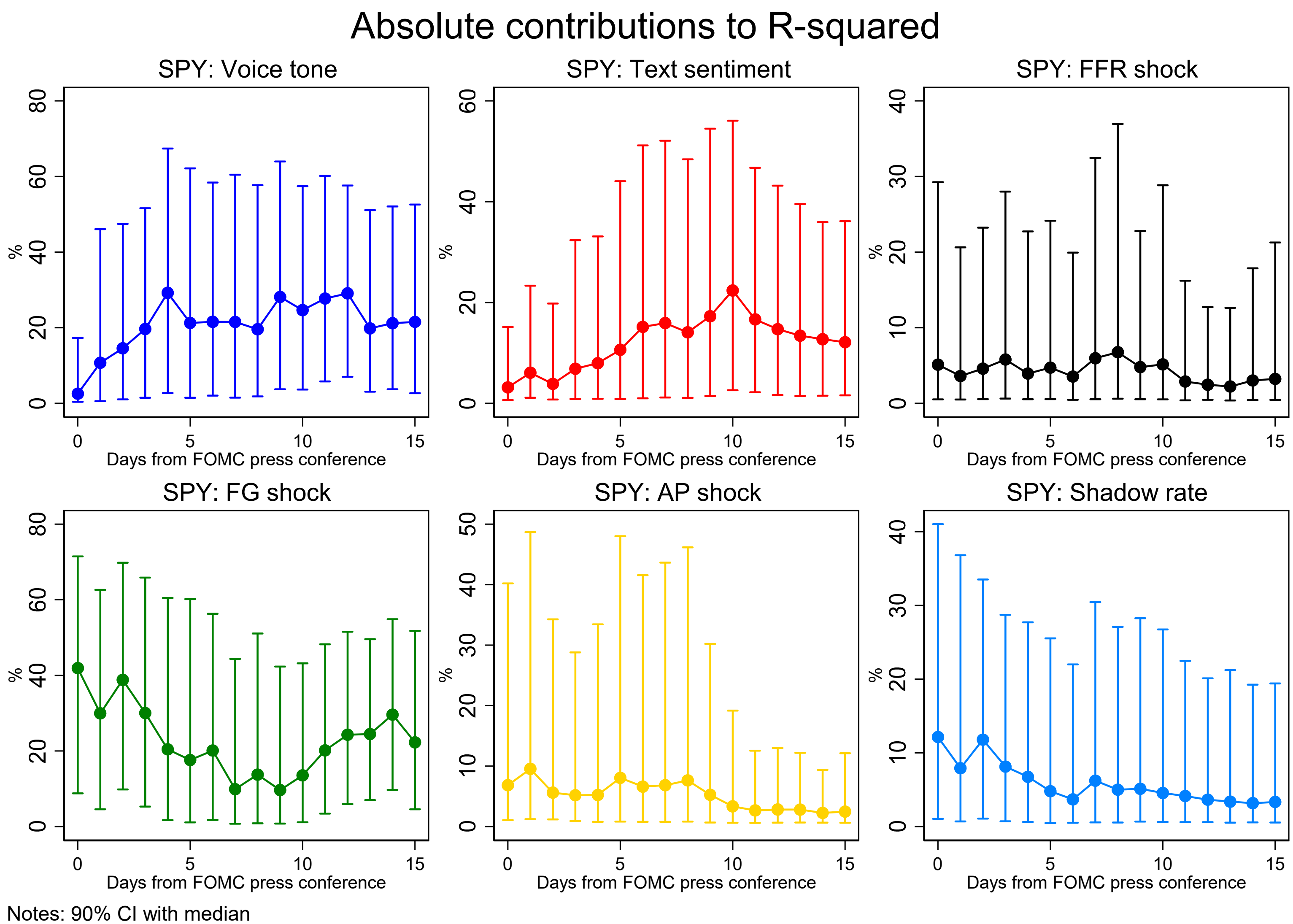
Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for policy communication/actions. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure A8. Response of the British Pound to one U.S. Dollar (pound/dollar) exchange rate to policy actions and messages



Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for policy communication/actions. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure A9. Variance decomposition (SPY ETF)



Notes: This figure reports the Shapley decomposition for the estimations with SPY ETF.

Appendix B. Neural network to classify audio tracks into emotions

Audio data

All audio files are converted to 16,000 Hz sample rate and mono channel. When passing the audios into the Librosa package for feature extraction, we use the default frame length (the number of samples in a frame) and the hop length (the number of samples between successive frames) of 2,048 and 512, respectively. Thus, for each audio, the number of frames (or "slices") used for feature extraction is calculated as:

$$frames_s = \frac{duration_s(in\ seconds) \times 16{,}000}{512} \qquad \text{(B1)}$$

Feature extraction

The inputs of our neural network algorithm are essentially the representations of two important vocal aspects, namely frequency (or pitch/highness) and amplitude (or volume/loudness). For an audio signal we can extract the following characteristics:

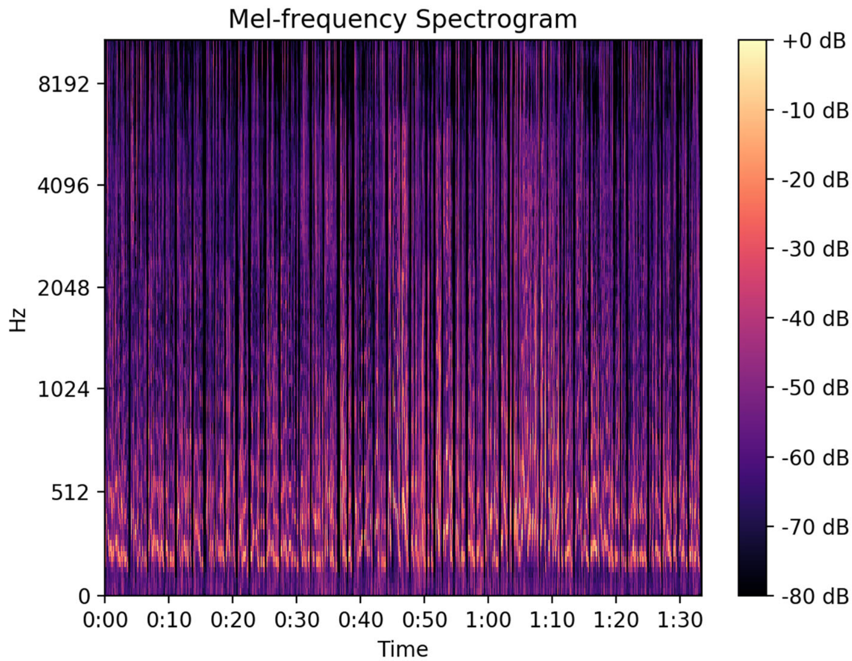
- A Mel frequency spectrogram is the spectrum of frequencies of an audio signal mapped onto a Mel scale (instead of the frequency scale) time. It allows us to determine the level of loudness of a particular frequency at a particular time.[1]
- A Chromagram is a representation of an audio signal in which the spectrum of frequencies is projected onto 12 equal-tempered pitch classes or 12 chroma bands (i.e., C, C#, D, D#, E, F, F#, G, G#, A, A#, and B). The Chromagram reflects the distribution of energy along 12 chroma bands over time and, hence, it can capture the melodic and harmonic characteristics of an audio signal.
- A Mel-frequency cepstral coefficients (MFCC) is a discrete cosine transformation of the Mel frequency spectrogram.

As mentioned in Section 2.1.1, we first extracted a vector of 128 Mel spectrogram coefficients. Appendix Figure B1 presents an example of a Mel spectrogram: the brighter colors around the frequency range of 256 – 512 Hz suggest the stronger (or "louder") amplitude of such a range. Second, the extracted features also include a vector of 40 MFCCs, which are considered to be the decorrelated versions of the Mel spectrogram. The negative MFCCs indicate that the spectral energy is concentrated at the high frequencies, while the positive MFCCs represent the concentration around the low frequencies. This is illustrated in Appendix Figure B2: the majority of cepstral coefficients are positive, corresponding to the stronger amplitude of the 256-512 Hz range suggested in Appendix Figure B1. Finally, the Chromagrams with 12 chroma coefficients are extracted from the audio signals. In the example shown in Appendix Figure B3, the pitches are scattered and distributed over all pitch classes, which reflects the fact that the examined audio is a sample audio book.[2] All obtained features are then averaged over all frames, meaning that we obtain a set of 180 features, or inputs, for each audio file.

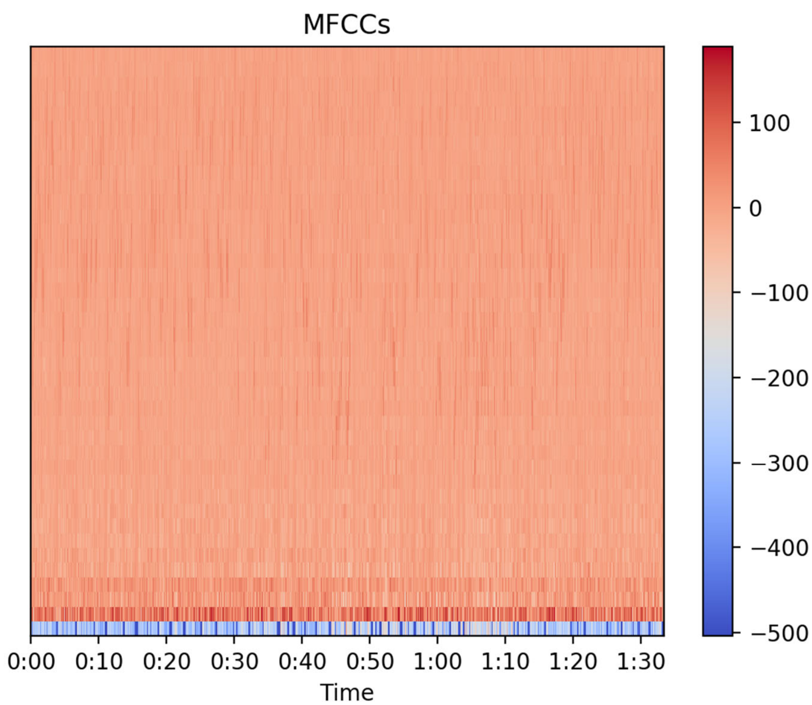
[1] Mel scale is a log transformation of frequencies which "mimic" the human perception of sound. That is, pitches of equal distance on the Mel scale are of equal distance when judged by humans.

[2] This audio can be found at https://secure.toolkitfiles.co.uk/clients/24554/sitedata/files/AudioBook-Tanya-S-Bartlett.mp3.

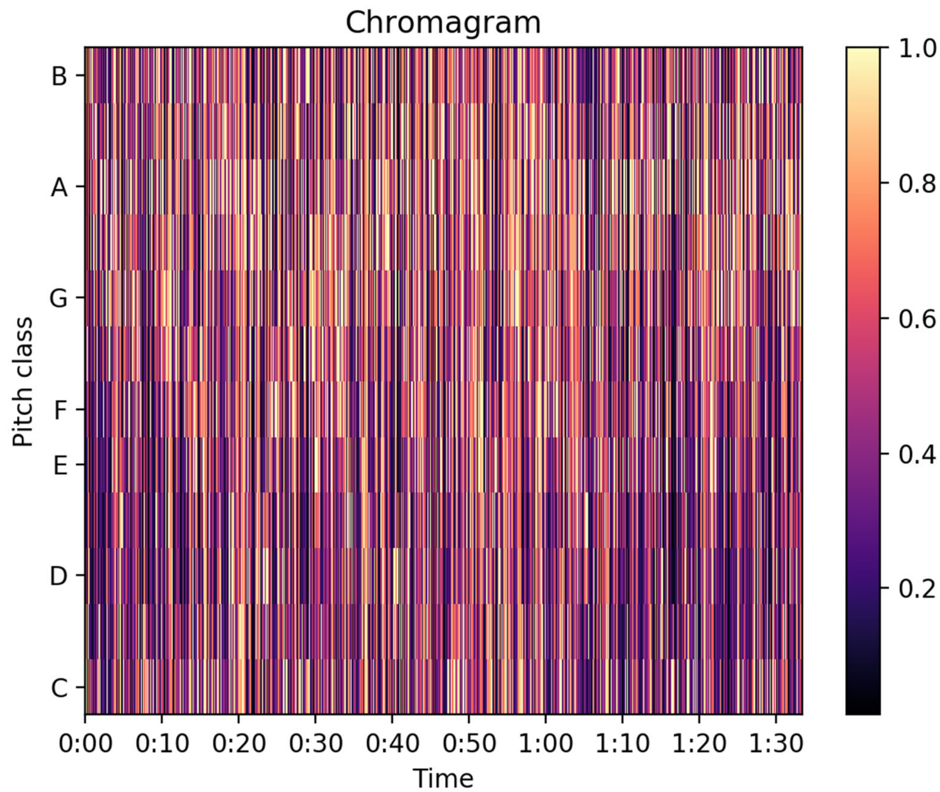
Appendix Figure B1. Example of Mel-frequency Spectrogram



Appendix Figure B2. Example of MFCCs



Appendix Figure B3. Example of Chromagram



It should be noted that the number of Mel spectrogram coefficients, the number of MFCCs, and the number of chroma coefficients are hyperparameters that can be adjusted to achieve a more effective algorithm. Similarly, one might ask whether it is necessary to use both Mel Spectrogram Frequencies and MFCCs as the inputs since the latter is essentially derived from the former. Within the scope of our fine-tuning exercise, we find that using both types of features helps to improve the accuracy of the model.

In addition, there are other spectral features that could be extracted and used in the neural network. For example, a spectral contrast (Contrast) is the level of difference between the mean energy in the top and bottom quantiles of the spectrum. One could also compute the tonal centroid of a chroma vector (Tonnetz), in which the chroma features are projected onto a 6-dimensional basis representing the perfect fifth, minor third, and major third.[3] As part of the fine-tuning exercise, we also experimented with using all five spectral features (Mel spectrogram coefficients, MFCCs, Chromagram, Tonnetz, and Contrast) as the inputs. However, this combination did not improve the accuracy rate.

The neural network

We use Keras, a deep learning API run on top of Google's machine learning platform TensorFlow, to build our neural network. In what follows, we will describe the specific model and training parameters of our network. This network is trained on 80% of TESS and RAVDESS data and tested on the remaining 20%.

Network structure

[3] See https://librosa.org/doc/main/feature.html for more information on various spectral features.

Our neural network is a fully connected network with four layers. This means that every node in one layer is connected to every node in the next layer through an activation function. Particularly, a node in the next layer is connected with all inputs *I* in the previous layer through weight ($w_{k,i}$) and bias (b_k): $\sum_{i=1}^{j} I_i \times w_{k,i} + b_k$.

- The first layer is a dense layer that takes 180 features (128 Mel coefficients, 40 MFCCs, and 12 chroma coefficients) as inputs and passes them through the linear activation function to produce 200 nodes as outputs.
- The second layer has 200 nodes that are connected with 200 nodes in the first layer through the linear activation function.
- The third layer has 200 nodes that are connected with 200 nodes in the second layer through the linear activation function.
- The output layer has five nodes representing five emotions (happy, pleasantly surprised, neutral, sad, and angry). Given that our task is a multi-class classification task, we use the softmax activation function (normalized exponential function), a logistic function, to connect the nodes in this layer with 200 nodes in the previous layer.
- To prevent overfitting, three Dropout layers with a dropout rate of 0.3 are added after each layer before the output layer. This means that 30% of inputs are randomly set to 0 at each step during the training time (hence, only 70% of inputs are retained for training).

Training parameters

- The number of training epochs is 2,000. This means that the entire training dataset is passed forward and backward through the network 2,000 times.
- The batch size is 64. This means that 64 training audio files are propagated through the network (i.e., processed) before the model's weights are updated.
- How the weights are updated is determined by an optimization algorithm. In this study, we use the adaptive moment estimation (Adam) with the default learning rate of 0.001 as the optimizer.
- The loss function, or the error function, is used to optimize the parameter values. Given the multi-class classification task, we use the categorical cross-entropy function which minimizes the distance between the distribution over pre-defined emotions and the "model" distribution over predicted emotions.

To evaluate the model, we use the following formula to calculate the accuracy rate:

$$Accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} 1(\hat{y}_i = y_i) \qquad \text{(B2)}$$

where y and $\hat{y}$ are the true emotion and the predicted emotion, respectively. *n* is the number of audio files in the testing dataset.

The accuracy rate of the model used for analysis is 84%. When applying this formula for each of the emotion classes, we obtain accuracy scores of 87%, 84%, 74%, 87%, and 80% for angry, sad, neutral, pleasantly surprised, and happy, respectively.

Appendix C. Neural network to classify (central bank) text sentiment

Text embeddings

We used two different BERT models to extract the word embeddings from texts. The first model is the base uncased model, which has 12 layers, 768 hidden states, 12 heads, 110M parameters, and was trained on lower-case English text. The second model is the RoBERTa model, which has 12 layers, 768 hidden states, 12 heads, and 125M parameters.

The neural network

The sequence of the hidden states at the output of the last layer of the BERT model is used as inputs for the text classification model, which is specified below. This network is trained on 80% of our unique (balanced) labelled FOMC statements data, validated on 10% of the sample, and tested on the remaining 10%.

Network structure

Our neural network's structure is as follows.

- Input layer is the sequence of the hidden states at the output of the last layer of the BERT model.
- The first hidden layer is a bidirectional long short-term memory (LSTM) layer created by wrapping a LSTM layer with a Bidirectional layer. The LSTM has 512 units, a dropout rate of 0.1, and a recurrent dropout rate of 0.1. We use the default activation function (hyperbolic tangent). Following the bidirectional LSTM is a dropout layer with a dropout rate of 0.1
- The second hidden layer is a global average pooling 1D layer which is added to flatten the 2-dimensional data into 1-dimensional data, followed by a dropout layer (dropout rate is 0.1).
- The third hidden layer is a dense layer, which has 512 nodes (we use the rectified linear unit activation function). This hidden layer is followed by a dropout layer with a dropout rate of 0.1.
- The fourth hidden layer is a dense layer, which has 128 nodes. The rectified linear unit activation function is used.
- The output layer has three nodes representing three sentiment classes (hawkish, neutral, dovish). We use the softmax activation function for this multi-class classification task.

Training parameters

- The number of training epochs is 200.
- The batch size is 10.
- The optimization algorithm is Adam with the default learning rate of 0.001.
- The loss function (categorical cross-entropy function) is used to optimize the parameter values.

Evaluation

We use formula (B2) to calculate the accuracy score when applying the trained text sentiment model on the testing data. The performance of the model is as follows:

Appendix Table C1

Embeddings	Accuracy score			
	Average	Hawkish	Neutral	Dovish
BERT	81%	85%	77%	79%
RoBERTa	78%	88%	68%	78%

Appendix D. Textual analysis

Appendix D1. Search and count approach

We build four lists of nouns, adjectives, and verbs (Appendix Table D1), combinations of which will indicate either tight monetary policy/strong economic outlook (i.e., hawkish) or expansionary monetary policy/weak economic outlook (i.e., dovish). A phrase combined of (1) A1 and A2 or (2) B1 and B2 is classified as "dovish" while a phrase combined of (1) A1 and B2 or (2) B1 and A2 is classified as "hawkish". To increase the accuracy of our classification, the search and count approach is performed on each part of a sentence and then aggregated over the whole document. For example, the sentence "With inflation running persistently below this longer-run goal, the Committee will aim to achieve inflation moderately above two percent for some time so that inflation averages two percent over time and longer-term inflation expectations remain well anchored at two percent" contains two parts: "With inflation running persistently below this longer-run goal" and "the Committee will aim to…two percent". The search and count approach is performed on each part separately, then aggregated over the whole sentence, and then aggregated over the whole document. Since negations such as "won't" or "aren't" can alter the meaning of the text, for each part of a text, a hawkish (dovish) phrase is only counted as hawkish (dovish) if the text does not contain a negation word/phrase. In contrast, if a hawkish phrase is accompanied by a negation word/phrase, then it is counted as dovish and vice versa. A similar approach was applied in Cieslak and Vissing-Jorgensen (2021), where a negative word accompanied by "not" is considered positive. The aggregate sentiment of the text of an FOMC statement/remarks/Q&A is measured as:

$$TextSentiment = \frac{Dovish\ phrases - Hawkish\ phrases}{Dovish\ phrases + Hawkish\ phrases}$$

where $Dovish\ phrases$ and $Hawkish\ phrases$ are the counts of respective phrases in the FOMC statements as well as transcripts when a press conference is held.

Appendix Table D1. Dictionary for hawkish and dovish words

Panel A1	Panel A2
inflation expectation, interest rate, bank rate, fund rate, price, economic activity, inflation, employment	anchor, cut, subdue, declin, decrease, reduc, low, drop, fall, fell, decelarat, slow, pause, pausing, stable, non-accelerating, downward, tighten
Panel B1	**Panel B2**
unemployment, growth, exchange rate, productivity, deficit, demand, job market, monetary policy	ease, easing, rise, rising, increase, expand, improv, strong, upward, raise, high, rapid
Panel C	
weren't, were not, wasn't, was not, did not, didn't, do not, don't, will not, won't	

Notes: This table shows the words/phrases used to classify text into dovish/hawkish.

Appendix D2. Measuring the intensity of text sentiment

As an additional robustness check, we adopt the approach used in Kozlowski et al. (2019) and Jha et al. (2021) to measure the text sentiment's intensity. The steps of this approach can be summarized as follows. First, based on the dictionary in Neuhierl and Weber (2019), we build a dovishness-hawkishness dataset of sentence/phrase pairs with opposite monetary policy stances (see Appendix Table D2 below). Second, we use the BERT algorithm to extract embeddings for the text in this dataset and the policy texts. Third, for each pair of embedding vectors in the dovishness-hawkishness dataset, we calculate the embedding difference between the dovish sentence/phrase and the hawkish counterpart. The average of these dovish-minus-hawkish vectors is considered a dovishness dimension. Finally, the degree of dovishness (or hawkishness) of a given policy text is the cosine similarity score between the policy text's embedding vector and the vector of the dovishness dimension. By construction, this continuous score ranges from -1 to 1 where a positive score indicates a dovish connotation and a negative score represents a hawkish connotation. A higher absolute value of a positive (negative) score means a higher degree of dovishness (hawkishness).

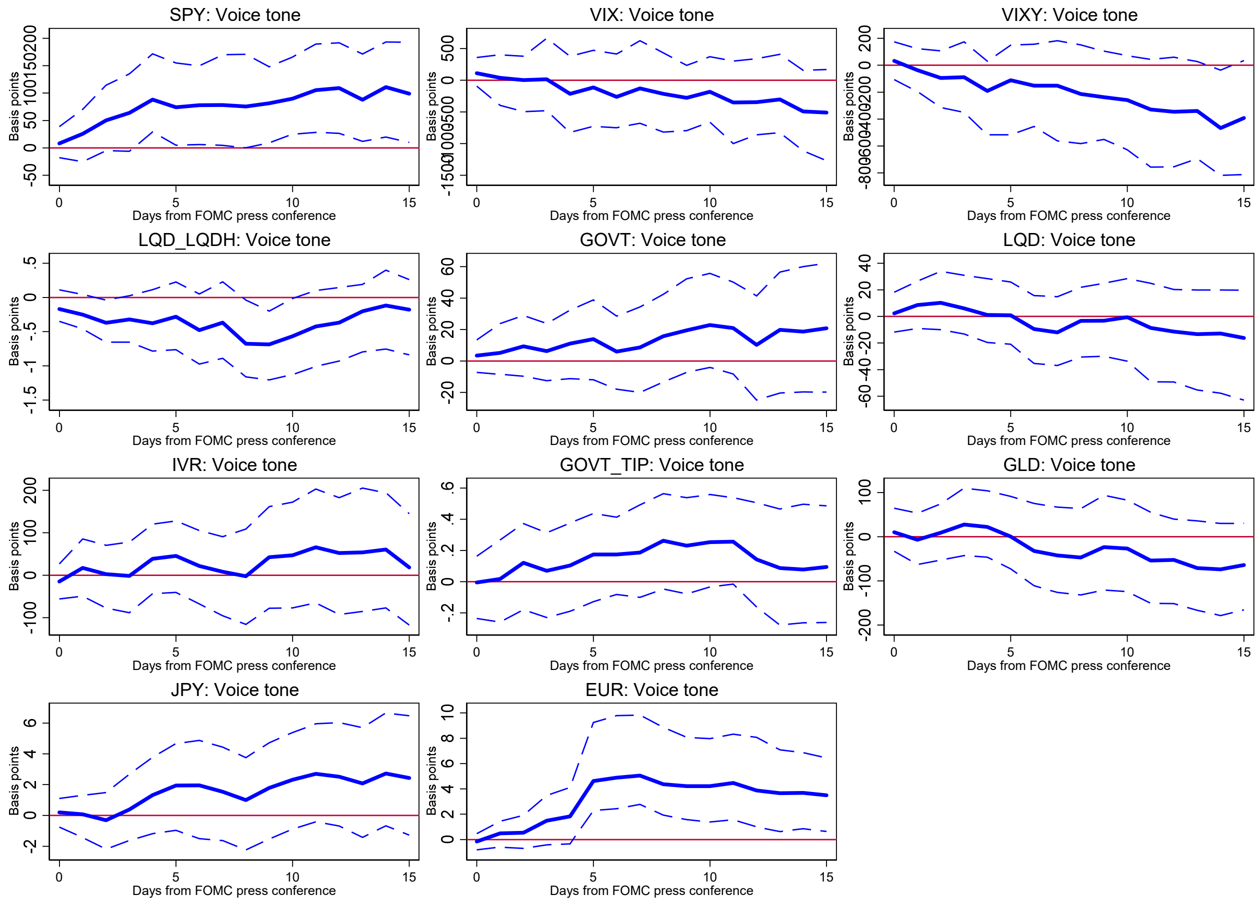
As shown in Appendix Figure D1, the results for the tone of voice measure are consistent when controlling for the degree of dovishness/hawkishness of the policy texts. The consistent findings are also observed when we allow for the non-linear terms in text sentiment (Appendix Figure D2).

Appendix Table D2. Policy stance pairs to measure dovish dimension

Dovish phrases	Hawkish phrases
inflation expectations anchor	inflation expectations increase
anchor inflation expectations	rise inflation expectations
inflation expectations decline	inflation expectations increase
inflation expectations remain stable	inflation expectations higher
inflation expectations stable	inflation expectations higher
stable inflation expectations	rise inflation expectations
lower inflation expectations	higher inflation expectations
reduction inflation expectations	increase inflation expectations
cut federal funds rate	raise federal funds rate
lower federal funds rate	higher federal funds rate
reduce federal funds rate	raise federal funds rate
decrease federal funds rate	increase federal funds rate
reduction federal funds rate	rise federal funds rate
cut interest rate	raise interest rate
lower interest rate	higher interest rate
reduce interest rate	raise interest rate
decrease interest rate	increase interest rate
reduction interest rate	rise interest rate
decline economic activity	increase economic activity
stable inflation	rise inflation
downward pressure inflation	upward pressure inflation
decrease inflation	increase inflation
declined employment	higher employment
employment fallen	employment increased
employment fell	employment increased
unemployment rate rising	unemployment rate lower
increases unemployment rate	declines unemployment rate
rise unemployment rate	drop unemployment rate
higher unemployment rate	lower unemployment rate
dovish monetary policy	hawkish monetary policy
easing monetary policy	tightening monetary policy

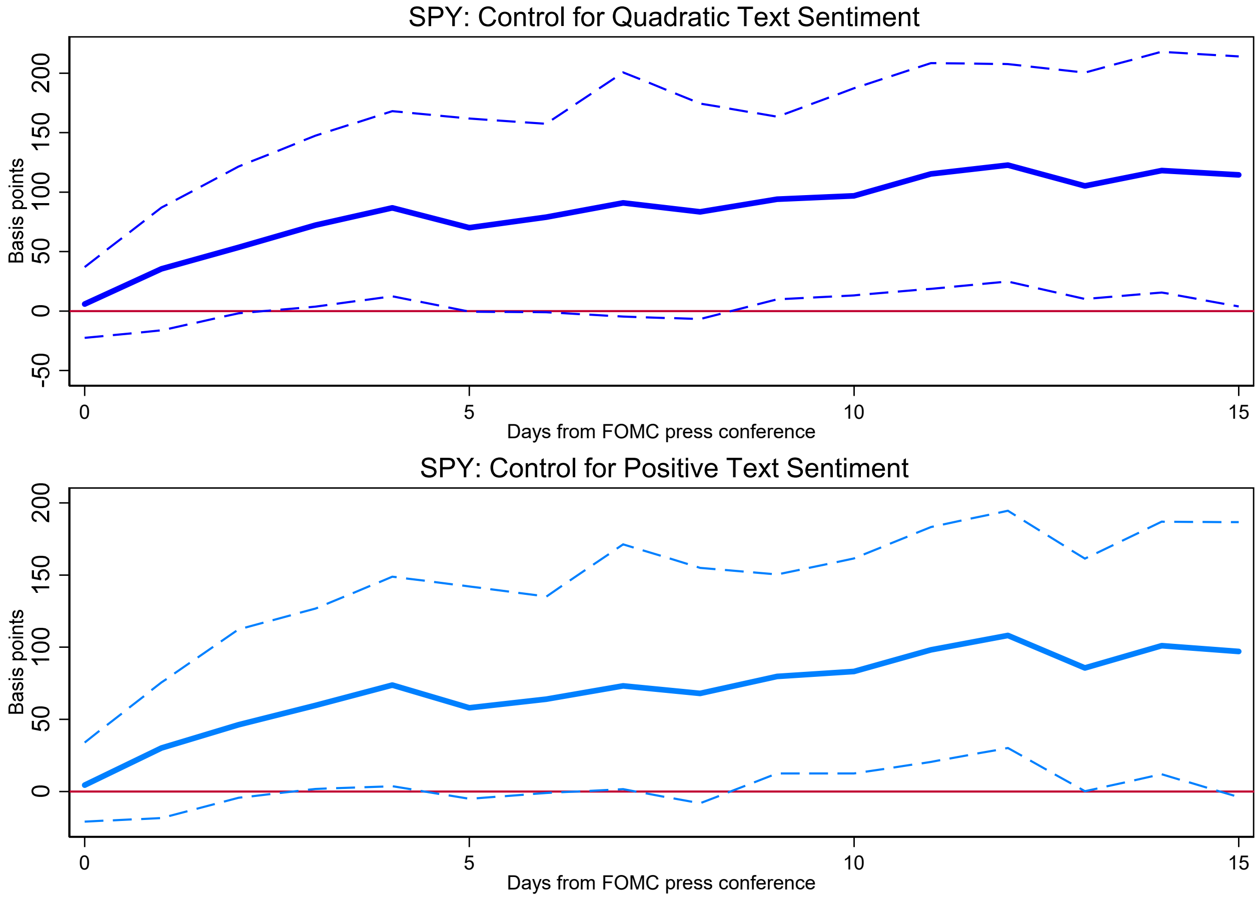
Notes: This table shows the words/phrases used to classify text into dovish/hawkish.

Appendix Figure D1. Control for the intensity of text sentiment



Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for the tone of voice while controlling for the intensity of text sentiment. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.

Appendix Figure D2. Control for the non-linear terms of text sentiment



Notes: This figure reports the estimated slope coefficients *b* (Specification (4)) for the tone of voice while adding the non-linear terms of text sentiment. Dashed lines show 90% bias-corrected and accelerated bootstrap confidence intervals.