Internet Appendix for FinTechs and the Market for Financial Analysis

Jillian Grennan
Duke University Fuqua School of Business
jillian.grennan@duke.edu

Roni Michaely
University of Geneva and SFI Geneva Finance Research Institute
Roni.Michaely@unige.ch

B Additional Figures and Tables

FIGURE B.1

Changes in price informativeness over time

Figure B1 plots the relationship between crowd wisdom from bloggers and price informativeness over time. The dependent variable is price informativeness as proxied by price nonsynchronicity. The dots represent coefficient estimates from instrumental variable (IV) regressions at the equity-quarter level for a given quarter and the bands represent 90% confidence intervals based on robust standard errors clustered at the equity level. The maroon solid line represents the null hypothesis of no change in price informativeness. The instrument for financial blogging is an indicator for whether the equity has below median headline length in the USA Today in a given quarter. Additional control variables include newspaper coverage and lagged values of analyst coverage, firm size, daily return volatility, mean monthly return, log market-to-book ratio, volatility of ROE, profitability, and an indicator for if the stock is a member of the S&P 500.

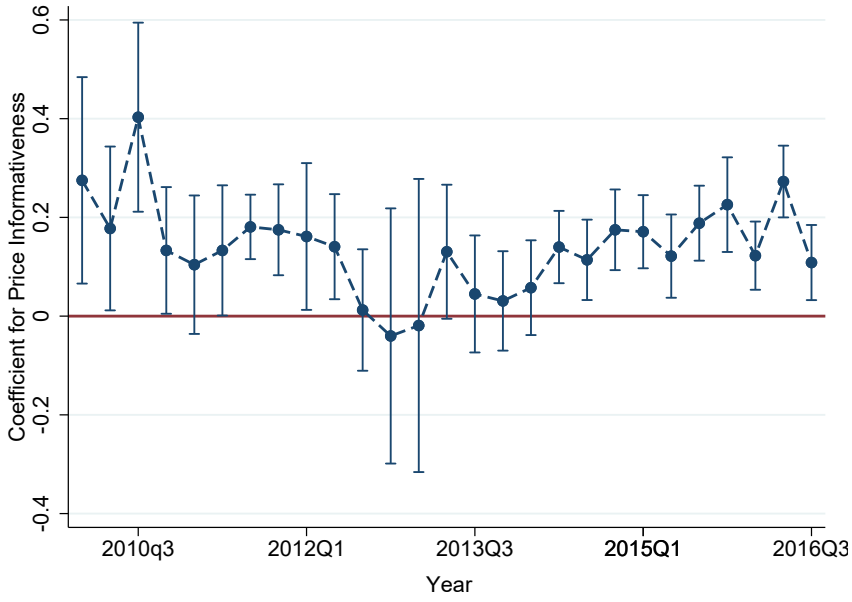

TABLE B.1

Summary Statistics

Table B1 presents summary statistics for the main dependent and independent variables. After combining the data sets, the main sample period is limited to 2010 to 2016. For a detailed description of each variable, see the definitions in Appendix A.

Panel A. Price informativeness Variables

Variable	Freq.	Mean	Median	Std. Dev.	N
	1	2	3	4	5
PRICE_NONSYNCHRONICITY	Q	0.63	0.66	0.25	81,201
ABS_ORDER_IMBALANCE	Q	0.12	0.09	0.10	81,201
PRICE_JUMP_RATIO	Q	0.49	0.47	0.51	20,203
ANALYST_INFO	Q	0.07	0.05	0.08	81,201

Panel B. Explanatory Variables

Variable	Freq.	Mean	Median	Std. Dev.	N
	1	2	3	4	5
CROWD_WISDOM	Q	9.88	4.00	17.04	81,201
TOP_CROWD_WISDOM	Q	2.64	0.00	5.99	81,201
NEWSPAPER	Q	8.31	3.00	15.26	81,201
ANY_KEY_EVENT	Q	13.78	11.00	16.71	81,201
PAYOUT_EVENT	Q	1.17	1.00	1.16	81,201
ACTIVIST_EVENT	Q	0.03	0.00	0.40	81,201
CREDIT_EVENT	Q	0.00	0.00	0.01	81,201
OTHER_EVENT	Q	4.40	3.00	6.55	81,201
ANALYSTS	Q	7.07	5.28	5.81	81,201
FIRM_SIZE	Q	13.90	13.86	1.77	81,201
VOLATILITY	Q	38.0%	32.7%	21.1%	81,201
RETURNS	Q	1.2%	1.2%	6.7%	81,201
ILLIQUIDITYo	Q	14.2	0.2	81.8	81,201
MARKET-TO-BOOKo	Q	0.83	0.75	0.44	81,201
VOLATILITY_OF_ROE	Q	25.9%	0.2%	113.8%	81,201
PROFITABILITY	Q	1.8%	2.2%	4.5%	81,201
MEMBER_OF_S&P_500	Q	14.4%	0.0%	35.2%	81,201
ROE	Q	0.9%	2.1%	13.3%	81,201
MOMENTUM	Q	3.6%	0.6%	15.1%	81,201
INSTITUTIONAL_OWNER	Q	59.1%	66.3%	30.8%	81,201
HEDGE_FUND_OWNER	Q	5.0%	0.5%	8.2%	81,201

TABLE B.2

Summary Statistics for Newspaper Headlines

Table B2 provides summary statistics related to newspaper headline length for the sample period from 2009 through 2016. Headline data comes from Ravenpack. All major newspapers includes: USA Today, the Wall Street Journal, the New York Times, the Los Angeles Times, the Chicago Tribune, the Washington Post, the Financial Times, and the DowJones Newswire. For our analysis, we focus exclusively on the USA Today, because it is a high readership newspaper that is more likely to be read by individual investors including bloggers.

Source	N	Mean	25th Percentile	Median	75th Percentile
NEWSPAPER	7,538,452	56.5	48	57	63
USA_TODAY	431,710	51.3	43	49	57

TABLE B.3

Instrumental Variable (IV) Regression, First Stage Details

Table B3 presents the full details of the first stage of the instrumental variable (IV) regressions for CROWD_WISDOM. Fixed effects (FE) are noted in the bottom rows. Standard errors clustered by equity are reported below the coefficient estimates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For a detailed description of each variable, see the definitions in Appendix A.

Dependent Variable: CROWD_WISDOM	1	2
SHORT_HEADLINE	1.644***	0.504***
	(0.122)	(0.052)
NEWSPAPER	-0.030***	0.013**
	(0.010)	(0.006)
ANALYSTS	0.170***	0.310***
	(0.017)	(0.020)
FIRM_SIZE	0.359***	0.233***
	(0.022)	(0.038)
ILLIQUIDITY	0.051***	0.006**
	(0.005)	(0.003)
VOLATILITY	0.268***	0.206***
	(0.012)	(0.008)
RETURN	-0.035***	-0.020**
	(0.004)	(0.003)
MARKET-TO-BOOK	0.013	0.028*
	(0.011)	(0.017)
VOLATILITY_OF_ROE	0.034***	0.047**
	(0.011)	(0.024)
PROFITABILITY	0.024***	0.006
	(0.008)	(0.008)
MEMBER_OF_S&P_500	0.156***	0.106***
	(0.016)	(0.034)
ANY_KEY_EVENT	0.090***	-0.103***
	(0.018)	(0.027)
OTHER_EVENT	0.145***	0.055***
	(0.022)	(0.016)
PAYOUT_EVENT	0.002	0.013***
	(0.008)	(0.005)
ACTIVIST_EVENT	0.011	0.022***
	(0.008)	(0.004)
CREDIT_EVENT	0.002	0.003
	(0.003)	(0.003)
Other Headline Controls	Yes	Yes
Quarter FE	Yes	Yes
Equity FE	No	Yes
First-stage R^2	53%	79%
First-stage F-statistic	182.7	92.6
Critical Value for Weak Instrument, 10% bias	16.38	16.38
t-statistic on Instrument	12.78	8.66
N	81,201	81,201

TABLE B.4

Price Informativeness, Regression Details

Table B4 presents the full details of the ordinary least squares (OLS) and instrumental variable (IV) regressions for price informativeness. Column 1 shows the OLS estimates and Column 2 shows the IV estimates. Fixed effects (FE) are noted in the bottom rows. Standard errors clustered by equity are reported below the coefficient estimates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For a detailed description of each variable, see the definitions in Appendix A.

Dependent Variable: PRICE_NONSYNCHRONICITY	1	2
CROWD_WISDOM	0.062***	0.084***
	(0.010)	(0.029)
NEWSPAPER	-0.032***	-0.032***
	(0.008)	(0.008)
ANALYSTS	0.026**	0.022*
	(0.012)	(0.013)
FIRM_SIZE	-0.665***	-0.673***
	(0.016)	(0.019)
ILLIQUIDITY	0.055***	0.054***
	(0.005)	(0.005)
VOLATILITY	0.009	0.003
	(0.009)	(0.012)
RETURN	0.071***	0.071***
	(0.003)	(0.004)
MARKET-TO-BOOK	0.107***	0.107***
	(0.008)	(0.008)
VOLATILITY_OF_ROE	-0.020*	-0.021*
	(0.011)	(0.011)
PROFITABILITY	-0.009	-0.010
	(0.007)	(0.007)
MEMBER_OF_S&P_500	-0.039***	-0.042***
	(0.011)	(0.012)
ANY_KEY_EVENT	0.014	0.012
	(0.010)	(0.010)
OTHER_EVENT	0.041***	0.038***
	(0.011)	(0.011)
PAYOUT_EVENT	-0.050***	-0.050***
	(0.007)	(0.007)
ACTIVIST_EVENT	0.023***	0.022***
	(0.003)	(0.003)
CREDIT_EVENT	0.001	0.001
	(0.003)	(0.003)
Other headline controls	Yes	Yes
Quarter FE	Yes	Yes
First-stage F-statistic	—	182.7
t-statistic on Instrument	—	13.52
Adj. R^2	49%	—
N	81,201	81,201

TABLE B.5

Price Informativeness by Analyst Coverage Quartiles

Table B5 presents estimates of the coefficient on crowd wisdom from financial bloggers separately for the lowest quartile, the interquartile range, and the highest quartile of analyst coverage. Fixed effects (FE) are noted in the bottom rows. Standard errors clustered by equity are reported below the coefficient estimates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For a detailed description of each variable, see the definitions in Appendix A.

Focal Independent Variable: CROWD_WISDOM	Dependent Variable: PRICE_NONSYNCHRONICITY	
ANALYSTS quartiles	OLS	IV
	1	2
Analyst coverage Q1 (lowest)	0.149***	0.375***
	(0.024)	(0.096)
Analyst coverage Q2-Q3	0.071***	0.182**
	(0.012)	(0.081)
Analyst coverage Q4 (highest)	0.053***	0.070**
	(0.013)	(0.036)
Instrumental Variable Estimate	No	Yes
Controls	Yes	Yes
Quarter FE	Yes	Yes

TABLE B.6

Price Nonsynchronicity (Subsample Test)

Table B6 presents estimates of the change in price informativeness, proxied using PRICE_NONSYNCHRONICITY, as in Panel A of Table 9 for two different subsamples of the data. In Panel A, the subsample comprises observations with an above median price jump ratio. For this subsample of equities, distortions to price informativeness from algorithmic trading are limited. In Panel B, the subsample comprises observations with a below-median price jump ratio. In Columns 1–2, the focal independent variable is CROWD_WISDOM. In Columns 3–4, the focal independent variable is TOP_CROWD_WISDOM. The instrument for the focal independent variables is an indicator for whether the equity has below median headline length in the USA Today newspaper in a given quarter. Control variables include NEWSPAPER, ANALYSTS, FIRM_SIZE, VOLATILITY, RETURNS, ILLIQUIDITY, MARKET-TO-BOOK, VOLATILITY_OF_ROE, PROFITABILITY, MEMBER_OF_S&P_500, ANY_KEY_EVENTS, PAYOUT_EVENT, ACTIVIST_EVENT, CREDIT_EVENT, OTHER_EVENT, and other headline controls include the number of headline appearances for the quarter for the key words determined by the LASSO selection method. The first stage F-statistic is the Kleibergen-Paap Wald F-statistic. Fixed effects (FE) are noted in the bottom rows. Standard errors clustered by equity are reported below the coefficient estimates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For a detailed description of each variable, see the definitions in Appendix A.

Panel A. High PRICE_JUMP_RATIO sample

	Dependent variable: PRICE_NONSYNCHRONICITY			
	1	2	3	4
CROWD_WISDOM	0.006	0.269**		
	(0.049)	(0.137)		
TOP_CROWD_WISDOM			0.005	0.237**
			(0.044)	(0.122)
Controls	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Equity FE	No	Yes	No	Yes
First-stage F-statistic	125.8	21.7	114.9	19.9
t-statistic	11.22	4.66	10.72	4.46
N	10,102	9,566	10,072	9,566

Panel B. Low PRICE_JUMP_RATIO sample

	Dependent variable: PRICE_NONSYNCHRONICITY			
	1	2	3	4
CROWD_WISDOM	0.127***	0.287***		
	(0.038)	(0.131)		
TOP_CROWD_WISDOM			0.117***	0.316**
			(0.028)	(0.144)
Controls	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Equity FE	No	Yes	No	Yes
First-stage F-statistic	88.3	24.2	78.8	13.0
t-statistic on Instrument	9.39	4.92	8.88	3.60
N	10,101	9,577	10,101	9,577

TABLE B.7

Absolute Order Imbalance (Subsample Test)

Table B7 presents estimates of the change in price informativeness, proxied using absolute order imbalance, as in Panel B of Table 9 for two different subsamples of the main data set. In Panel A, the subsample comprises observations with an above median price jump ratio. For this subsample of equities, distortions to price informativeness from algorithmic trading are limited. In Panel B, the subsample comprises observations with a below-median price jump ratio. In Columns 1–2, the focal independent variable is CROWD_WISDOM. In Columns 3–4, the focal independent variable is TOP_CROWD_WISDOM. The instrument for the focal independent variables is an indicator for whether the equity has below median headline length in the USA Today newspaper in a given quarter. Control variables include NEWSPAPER, ANALYSTS, FIRM_SIZE, VOLATILITY, RETURNS, ILLIQUIDITY, MARKET-TO-BOOK, VOLATILITY_OF_ROE, PROFITABILITY, MEMBER_OF_S&P_500, ANY_KEY_EVENTS, PAYOUT_EVENT, ACTIVIST_EVENT, CREDIT_EVENT, OTHER_EVENT, and other headline controls include the number of headline appearances for the quarter for the key words determined by the LASSO selection method. The first stage F-statistic is the Kleibergen-Paap Wald F-statistic. Fixed effects (FE) are noted in the bottom rows. Standard errors clustered by equity are reported below the coefficient estimates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For a detailed description of each variable, see the definitions in Appendix A.

Panel A. High PRICE_JUMP_RATIO sample

	Dependent variable: ABS_ORDER_IMBALANCE			
	1	2	3	4
CROWD_WISDOM	0.080***	0.060**		
	(0.012)	(0.030)		
TOP_CROWD_WISDOM			0.073***	0.053*
			(0.011)	(0.028)
Controls	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Equity FE	No	Yes	Yes	No
First-stage F-statistic	125.8	21.7	114.9	19.9
t-statistic on Instrument	11.22	4.66	10.72	4.46
N	10,102	9,566	10,072	9,566

Panel B. Low PRICE_JUMP_RATIO sample

	Dependent variable: ABS_ORDER_IMBALANCE			
	1	2	3	4
CROWD_WISDOM	0.113***	0.161***		
	(0.019)	(0.049)		
TOP_CROWD_WISDOM			0.104***	0.177***
			(0.018)	(0.064)
Controls	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Equity FE	No	Yes	No	Yes
First-stage F-statistic	88.3	24.2	78.8	13.0
t-statistic on Instrument	9.39	4.92	8.88	3.60
N	10,101	9,577	10,101	9,577

TABLE B.8

Price Informativeness (Weekly Observations)

Table B8 presents estimates of the change in price informativeness using data at a weekly rather than a quarterly frequency. In both panels, the focal independent variable is crowd wisdom from financial bloggers and the dependent variable is absolute order imbalance at a weekly frequency. Panel A shows ordinary least squares (OLS) estimates and Panel B shows instrumental variable (IV) estimates. The instrument for the focal independent variables is an indicator for whether the equity has below median headline length in the USA Today newspaper in a given week. Firm and equity control variables include firm size based on market capitalization at week end, daily equity return volatility for the week, mean weekly equity return, Amihud's illiquidity ratio for the week, an indicator for if the stock is a member of the S&P 500 that week, and quarterly observations for log market-to-book ratio, volatility of ROE, profitability, and analyst coverage. News content controls include an indicator for if an earnings announcement, the count of total key developments reported in Capital IQ for the week, as well as individual counts for payout announcements, activist interventions, credit events, and other key events. Other headline controls include the number of headline appearances for that week for the key words determined by the LASSO selection method. Fixed effects (FE) are noted in the bottom rows. Standard errors clustered by equity are reported below the coefficient estimates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For a detailed description of each variable, see the definitions in Appendix A.

Panel A. Ordinary Least Squares

	Dependent Variable: ABS_ORDER_IMBALANCE			
	1	2	3	4
CROWD_WISDOM	0.040***	0.010***		
	(0.002)	(0.001)		
TOP_CROWD_WISDOM			0.034***	0.006***
			(0.002)	(0.000)
Controls	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Equity FE	No	Yes	Yes	No
Adj. R^2	56%	73%	56%	73%
N	1,289,420	1,289,419	1,289,420	1,289,419

Panel B. Instrumental Variable

	Dependent Variable: ABS_ORDER_IMBALANCE			
	1	2	3	4
CROWD_WISDOM	0.103***	0.056***		
	(0.008)	(0.010)		
TOP_CROWD_WISDOM			0.096***	0.046***
			(0.007)	(0.008)
Controls	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Equity FE	No	Yes	No	Yes
First-stage F-statistic	163.3	74.9	150.2	102.2
t-statistic on Instrument	12.78	8.66	12.26	10.11
N	1,289,420	1,289,419	1,289,420	1,289,419

TABLE B.9

Price Informativeness (Annual Observations)

Table B9 presents estimates of the change in price informativeness but uses data at an annual frequency rather than a quarterly frequency. In both panels, the focal independent variable is crowd wisdom from financial bloggers. In Panel A, the dependent variable is price nonsynchronicity and in Panel B, the dependent variable is absolute order imbalance. Columns 1 and 3 show the ordinary least squares (OLS) estimates and Columns 2 and 4 show the instrumental variable (IV) estimates. The instrument for the focal independent variables is an indicator for whether the equity has below median headline length in the USA Today newspaper in a given quarter. Fixed effects are noted in the bottom rows. Standard errors clustered by equity are reported below the coefficient estimates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For a detailed description of each variable, see the definitions in Appendix A.

Panel A. Dependent Variable: PRICE_NONSYNCHRONICITY

	1	2	3	4
CROWD_WISDOM	0.072***	0.107***	0.007	0.150**
	(0.012)	(0.036)	(0.010)	(0.073)
Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Equity FE	No	No	Yes	Yes
First-stage F-statistic	—	156.3	—	57.5
t-statistic on Instrument	—	12.50	—	7.58
Adj. R^2	53%	—	78%	—
N	19,033	19,033	19,032	19,032

Panel A. ABS_ORDER_IMBALANCE

	1	2	3	4
CROWD_WISDOM	0.087***	0.135***	0.024***	0.112***
	(0.006)	(0.014)	(0.004)	(0.021)
Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Equity FE	No	No	Yes	Yes
First-stage F-statistic	—	156.3	—	57.5
t-statistic on Instrument	—	12.50	—	7.58
Adj. R^2	70%	—	91%	—
N	19,033	19,033	19,032	19,032

C Additional blogger data details

In this section, we provide additional details about the blogger tools available from TipRanks. Then, we explore if the bloggers who historically have issued more accurate buy and sell ratings are subsequently rewarded by the market when they make new ratings. We find evidence consistent with this fact, suggesting that some bloggers do have superior forecasting abilities and that they are rewarded for their costly activity of generating such a rating.

One of the unique data sets analyzed by TipRanks is its financial blogger data. Financial bloggers conduct research on stocks. Financial bloggers are independent writers and their research methods are more flexible than financial analysts. While financial bloggers do not carry the same prestige as financial analysts, they can potentially provide useful information with their provision of independent stock research and diverse priors. That being said, there are many bloggers and not all bloggers have the same abilities and incentives; and both might affect the accuracy of their stock performance prediction (Kogan, Moskowitz and Niessner, 2020). For users to be able to quickly assess the merits of the financial blog posts, TipRanks uses machine learning technology to read, measure and rank the stock recommendations of over 7,000 bloggers from the leading financial blogs.

To extract the wisdom of bloggers, TipRanks provides tools that investors can use to quickly evaluate bloggers' opinion on any stock. Figure C1 shows what a user would see when he or she clicks on the financial blogger tool for Facebook. The blogger tool includes blogger sentiment relative to the sector average, the distribution of financial blogs reporting on Facebook, a summary of bullish and bearish headlines with date stamps from the best performing financial bloggers, as well as table with sortable data on all financial bloggers posts. The table shows the blogger's name

and their star rating – with five stars being the bloggers with the best success rates and average returns. The table also shows which blog the blogger is affiliated with, the blogger's sentiment for the stock, when they last stated an opinion online about the stock, and a link to the blogger's article about this stock.

One can also click on the blogger's name to see their full profile, including their full stock coverage. Figure C2 shows individual blogger's profiles and the additional information beyond the star rating. In particular, the blogger's rank out of all bloggers and out of all experts based on a user-specified benchmark is provided. For example, the blogger on the left ranks #4,783 among bloggers and the blogger on the right ranks #6 among bloggers based on the selected time period of one month and compared to the selected benchmark of the industry sector. The profile also shows the average return associated with blogger's recommendation and the average success rate for the blogger. Listed next to the blogger profile is the raw performance data of the blogger for each rating. Figure C3 shows an example from a blogger's ratings of Alteryx. It shows that the blogger has had two successful ratings and the average profit associated with each rating. Finally, if a user is not interested in searching for a specific stock and drilling down into individual blogger ratings and instead would rather know what stocks bloggers are most bullish on, they can examine the top blogger stock predictions that reflect the stocks with the highest blogger rating.

Next, we present and discuss results for whether blog posts are associated with rating returns, especially for bloggers who make more accurate ratings. On one hand, research suggests investors may react positively to blogger buy rating, because investors are rationally responding to the revelation of private information conveyed by the recommendation (Chen, De, Hu and Hwang, 2014, Farrell, Green, Jame and Markov, 2019). On the other hand, other research suggests that bloggers'

ratings are based on ad hoc heuristics, rather than on sound economic analysis, and so reveal no useful information (Antweiler and Frank, 2004, Das and Chen, 2007).

To operationalize the bloggers' historical performance, we calculate the cumulative abnormal returns (CARs) for the bloggers' previous ratings in various event windows surrounding their latest rating. This two-step approach of first calculating the unexpected portion of historical performance, and then using that result to evaluate the newest rating returns is similar to methods developed by other researchers to examine skill in analysts' recommendations (Loh and Mian, 2006).

In the first step, we estimate the CARs from the posting of the blogger's ratings. We use daily data to estimate the parameters of a Carhart four-factor model in which the four factors are 1 the market return, which is the CRSP value-weighted index, 2 SMB (small minus big), which is a mimicking portfolio to capture risk related to size, 3 HML (high minus low), which is a mimicking portfolio to capture risk associated with book-to-market characteristics, and 4 UMD (up minus down), which is a mimicking portfolio designed to address risk associated with prior returns by subtracting a portfolio of low prior return firms from a portfolio of high prior return firms. The event periods that we estimate are the event window [0,+2], one week [0,+5], one month [0,+22], six months [0,+126] and one year [0,+252]. These are all measured relative to the posting day 0.

In the second step, we use the ordinary least squares (OLS) estimates of the average historical CARs that would be known at the time of the latest blogger's rating as well as indicators for historical blogger success (defined as a CAR greater than 0). For example, consider a blogger with a new rating, a rating that is eight months old, and a rating that is one year old. The one week, one month, and six month CARs would be averages of the previous two ratings, but the one year rating would only include the oldest rating. We use this information to investigate the degree of

performance persistence in the blogger's ratings. The regression specification is:

$$CAR_{ibjt} = \alpha_{ibjt} + \sum_{k=1}^{K} \beta_{C\bar{A}R_{ik}} C\bar{A}R_{ik} + \beta_{A\bar{c}c_{ik}} A\bar{c}c_{ik} + e_{ibjt}, \tag{C.1}$$

where CAR_{ibjt} is the [0,+2] event-window CAR for blog post i on blog b about stock j at date t. The CAR is in excess of benchmark portfolios matched on size, book-to-market, and momentum and blogger ratings are adjusted for buy versus sell ratings. The coefficients on $C\bar{A}R_{ik}$ represent the investors' unexpected return from following blogger i's rating at date t based on that blogger's average historical abnormal returns over the periods $k = 1, .., 4$ where k indexes our four event windows of one week [0,+5], one month [0,+22], six months [0,+126] and one year [0,+252]. Similarly, the coefficients on $A\bar{c}c_{ik}$ represent the investors' unexpected return from following blogger i's rating at date t based on that blogger's average historical accuracy over the periods $i = 1, .., 4$.

Table C1 shows the average excess returns for the 0 to +2 event window for posting a new blog post can be predicted using historical performance data. Blogger rating announcement returns are higher when the blogger has a history of ratings that performance successfully over a one-month horizon. Other factors are also predictive but the largest economic magnitude is associated with the historical one month excess returns from the blogger's prior ratings. This examination proves useful for multiple reasons. First, it provides evidence that some bloggers' investment advice is profitable and is associated with some type of superior forecasting skill. Second, it suggests that the resources that some bloggers devote to forecasting future stock performance is a useful tool to gauge the investment potential of a company's stock and moreover, that these resources provide the blogger with a competitive advantage. Finally, these results suggest that tools like that provided by TipRanks help to identify the crowd wisdom in financial blogging. To the extent that

such tools exist, it then seems plausible that bloggers post could contribute to changes in price informativeness.

Table C2 provides some descriptive statistics about the full sample of blog posts. The sample includes 1,315,898 blog posts between 2010 and 2017. About 35% of blog posts provide a buy or sell recommendation on a stock. One-fifth of those blog posts have sell recommendations while four-fifths are buy recommendations. Among all blog posts, there are 14,754 unique bloggers that cover 6,722 stocks. Among those that make buy or sell recommendations, there are 10,488 unique bloggers covering 6,385 stocks. Finally, among those that make at least 25 recommendations, there are 1,585 unique bloggers covering 6,210 unique stocks. We consider these bloggers that are making multiple buy and sell recommendations across a variety of different stocks to be the most similar to financial analysts.

In regard to the stocks covered in blogs posts, we observe 196 posts per stock and 12 posts per stock per quarter, on average. We observe 73 recommendations per stock and five buy or sell recommendations per stock per quarter, on average. Among those bloggers that make at least 25 recommendations, we see that they post to 1.4 different websites, on average, and make a total of 268 posts over our seven-year sample period. This translates into a new blog post approximately every 16 days. Similar to the performance at the blog-level, the performance of the bloggers with at least 25 recommendations (i.e., those we consider to be the most similar to equity analysts) demonstrate significant noise. Table C3 provides a list of the top equities discussed by financial bloggers. In comparison to recent research from the online platform StockTwits (Cookson and Niessner, 2020), our sample of bloggers are not as concentrated on big name stocks and appear to more evenly allocate their time across equities.

FIGURE C.1

TipRanks financial blogger opinion and predictions tool

Figure C1 shows the financial blogger page for Facebook from TipRanks stock analysis tool. The blogger page includes blogger sentiment relative to the sector average, the distribution of financial blogs reporting on Facebook, a summary of bullish and bearish headlines with date stamps from the best performing financial bloggers, as well as table with sortable data on all financial bloggers posts. The table shows you the blogger's name and their star rating – with five stars being the bloggers with the best success rates and average returns. One can then click on the blogger's name to see their full profile, including their full stock coverage. The table also shows which blog the blogger is affiliated with, the blogger's sentiment for the stock, when they last stated an opinion online about the stock, and a link to the blogger's article about this stock.

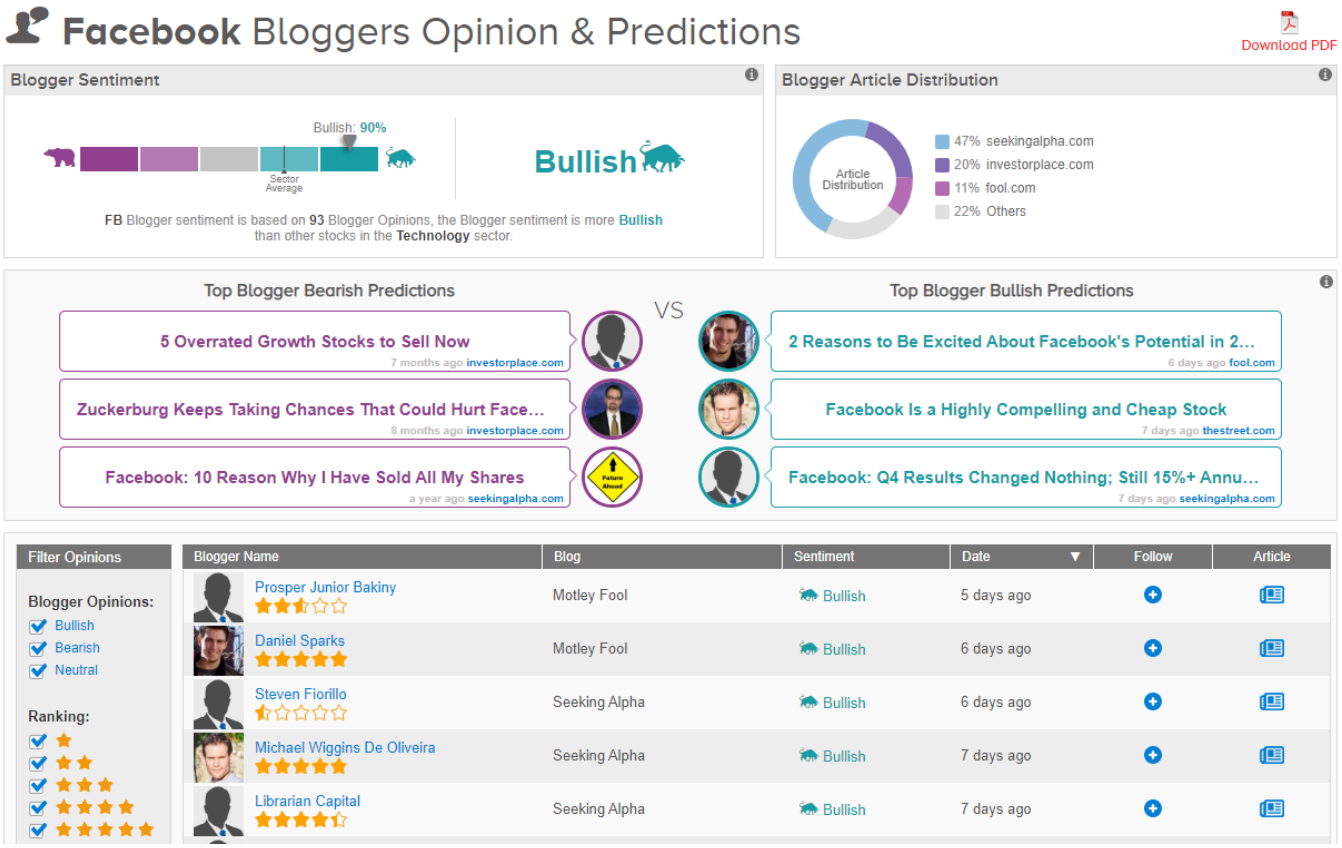

FIGURE C.2

TipRanks financial blogger profiles

Figure C2 shows example financial blogger profiles from TipRanks stock analysis tool. Each blogger is given a number ranking among bloggers and among all experts. In addition, the blogger profiles reveal the performance in terms of average returns and success relative to the selected benchmark.

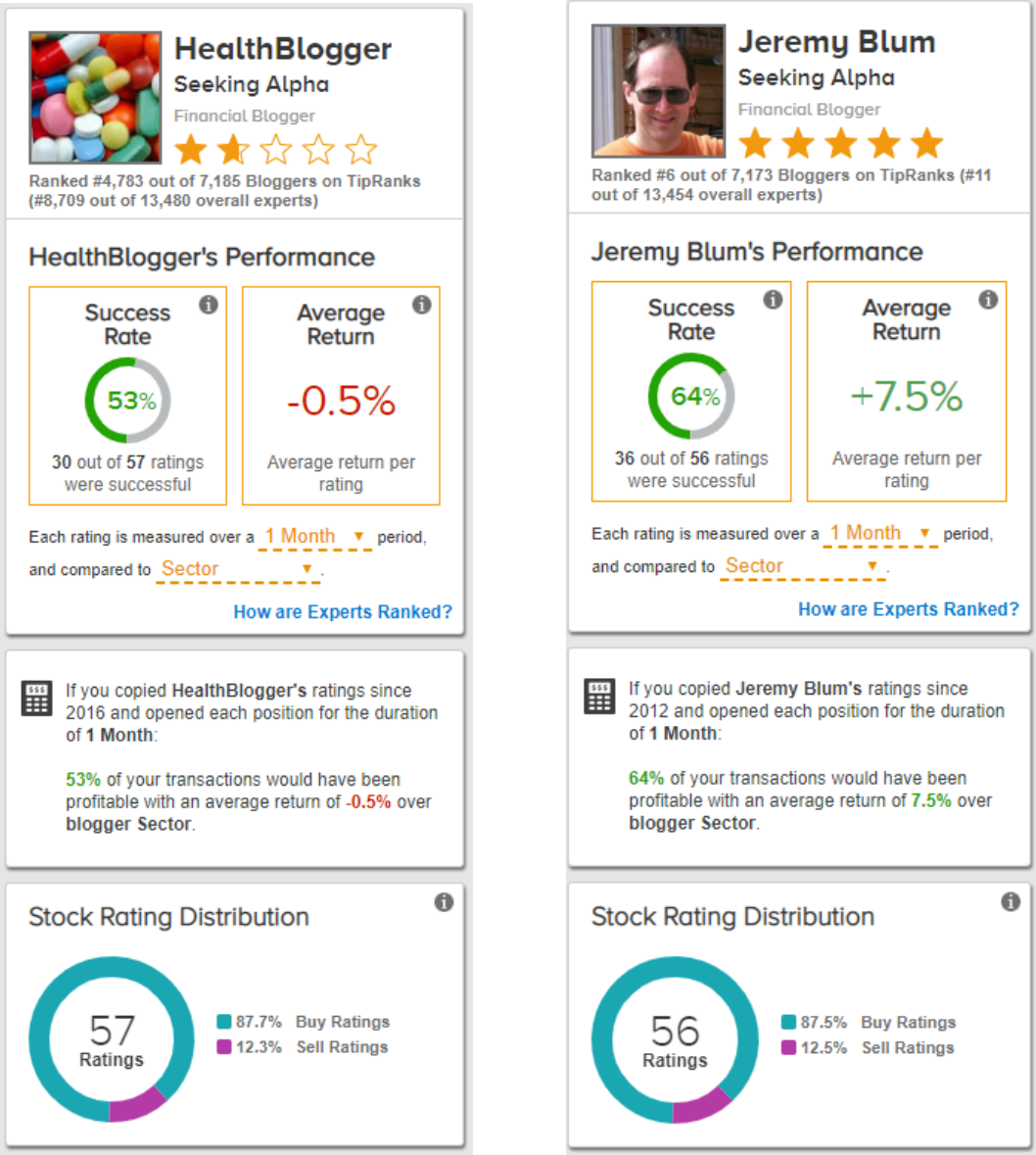

FIGURE C.3

TipRanks financial blogger performance history

Figure C3 shows an example of the raw performance data associated with an individual blogger's profile. It is presented as a table which includes the name of each individual company, the type of rating (buy versus sell), and the date of rating. One can then click on an individual stocks to see the raw performance, success rate, and average profit of the recommendations.

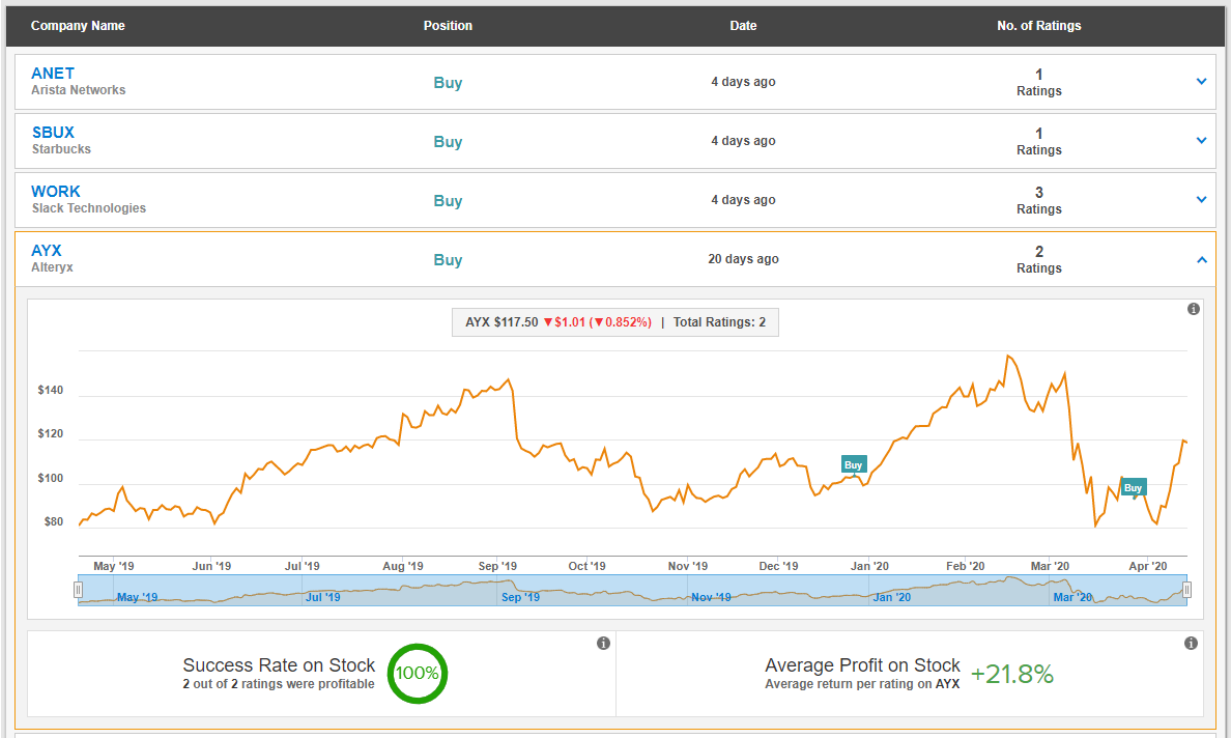

TABLE C.1

Blogger Performance

Table C1 presents estimates of blogger rating-induced abnormal returns as a function of past blogger performance. Abnormal returns are in excess of benchmark portfolios matched on size, book-to-market, and momentum. Abnormal returns from blogger ratings are adjusted for buy versus sell ratings. The estimates in the table present within-equity estimates where the cumulative abnormal returns (CARs) from the blog post are a function of historical performance representing CARs and accuracy of the blogger that would have been available on TipRanks website at the time of the post. The dependent variable is CARs from a 0 to +2 event window surrounding the blog post. Below the coefficient estimates are test statistics from robust standard errors clustered by blogger. Fixed effects (FE) are noted in the bottom rows. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	Estimated average CAR for [0,+2] event window					
	1	2	3	4	5	6
Blogger's historical average market-adjusted one-week return	0.065**					-0.059
	(0.031)					(0.066)
Blogger's historical average accuracy at one week	0.002***					0.002***
	(0.001)					(0.001)
Blogger's historical average market-adjusted one-month return		0.093**				0.195**
		(0.046)				(0.078)
Blogger's historical average accuracy at one month		0.001**				0.000
		(0.001)				(0.000)
Blogger's historical average market-adjusted three-month return			0.005			-0.038**
			(0.005)			(0.018)
Blogger's historical average accuracy at three months			0.001			0.000
			(0.001)			(0.001)
Blogger's historical average market-adjusted six-month return				0.004		0.018***
				(0.003)		(0.007)
Blogger's historical average accuracy at six months				0.001		0.000
				(0.001)		(0.001)
Blogger's historical average market-adjusted one-year return					0.000	-0.001
					(0.002)	(0.002)
Blogger's historical average accuracy at one year					0.003***	0.002**
					(0.001)	(0.001)
Blog FE	Yes	Yes	Yes	Yes	Yes	Yes
Equity FE	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	7%	7%	8%	9%	11%	11%
N	375,201	362,607	337,382	300,507	239,340	239,340

TABLE C.2

Characterizing Financial Bloggers' Posts

Table C2 presents summary statistics for our sample of financial blogs that make stock recommendations over the sample period from 2010 to 2017. This table provides descriptive statistics about bloggers' posts, their recommendations, the stocks they cover, the number of sites the bloggers post to, the days between posts, and the market-adjusted returns associated with their recommendations. Market-adjusted returns are in excess of benchmark portfolios matched on size, book-to-market, and momentum. For a detailed description of each variable, see the definitions in Appendix A.

Panel A. Blog Posts by Year

Year	Frequency
2010	46,360
2011	110,606
2012	144,868
2013	180,293
2014	257,444
2015	291,201
2016	196,637
2017	88,489

Panel B. Recommendation Sentiment

Sentiment	Frequency	Percent
Bearish	81,603	6%
Neutral	851,708	65%
Bullish	383,127	29%

Panel C. Stocks Covered in Blog Posts

Per Stock	Mean
Blog posts per stock	196
Blog posts per stock per quarter	12
Recommendations per stock	73
Recommendations per stock per quarter	5

Panel D. Blog Posts

	Among All Bloggers		Bloggers with 25 Recs.	
Blog Posts	Mean	Median	Mean	Median
Number of sites bloggers post to	1.1	1.0	1.4	1.0
Number of posts per blogger	89.2	4.0	267.8	70.0
Days between blog posts	65.8	23.2	16.3	10.2
Number of stocks covered	24.8	3.0	94.6	43.0
No. of bloggers	14,754	14,754	1,585	1,585
No. of stocks	6,722	6,722	6,210	6,210

Panel E. Performance of Bloggers with 25 Recs.

Performance Horizon	Mean	Median
Market-adjusted 1-month Return	0.2%	0.0%
Market-adjusted 3-month Return	-1.5%	-0.2%
Market-adjusted 6-month Return	-3.1%	-0.8%
Market-adjusted 12-month Return	-6.1%	-2.0%

TABLE C.3

Top Stocks Financial Bloggers Post About

Table C3 presents a list of the top stocks blogged about by financial bloggers.

Ticker	Company Name	Percent of Total Blog Posts
1	2	3
AAPL	Apple Inc.	2.78%
AMZN	Amazon.com, Inc.	0.86%
FB	Facebook Inc.	0.70%
MSFT	Microsoft Corporation	0.70%
TSLA	Tesla Motors Inc.	0.67%
BAC	Bank of America	0.67%
INTC	Intel	0.66%
NFLX	Netflix Inc.	0.56%
F	Ford Motor Company	0.51%
GOOG	Alphabet Inc.	0.42%
BBRY	BlackBerry Limited	0.41%
BA	Bank of America Corp.	0.39%
DIS	The Walt Disney Company	0.37%
IBM	International Business Machines Corp.	0.36%
CSCO	Cisco Systems, Inc.	0.36%
WMT	Wal-mart	0.36%
MCD	McDonald's	0.35%
T	AT&T Inc.	0.35%
TWTR	Twitter Inc.	0.34%
GE	General Electric	0.34%
GILD	Gilead Sciences, Inc.	0.33%
JNJ	Johnson & Johnson	0.32%
JPM	J.P. Morgan	0.31%
XOM	Exxon Mobil	0.30%
GM	General Motors	0.30%
CVX	Chevron Corporation	0.29%
YHOO	Yahoo! Inc.	0.29%

D Instrumental variable identification checks

In this Appendix, we report the results of several tests that we conducted to support the exclusion restriction assumption for our IV identification strategy. To illustrate our IV, an example of a short Apple headline in the USA Today is "Apple unveils IPad Mini, new Macs," and a long headline is "Apple CEO Cook mum on new products, says 'we have more game changers in us."' The USA Today often also publishes general interest articles such as "What exactly is Apple Music anyway?" While it is possible that headline length might be correlated with stuff about stock or news event, we test these ideas and find that they are either not systematically correlated, or the magnitude is so small that it does not affect the estimated relationship between blogging and informativeness.

For example, consistent with the identifying assumption, Table D1 shows that the IV is uncorrelated with variation in stock characteristics such as momentum, market-to-book ratio, profitability, ROE, and firm size. We evaluate over 7 million headlines over all newspapers in our sample and find no variable associated with firm characteristics is statistically significant at the 95th percentile. Moreover, the R-squared is only 0.10%. We do find evidence that newspaper fixed effects are significant, suggesting that the editor plays an important role.

A potential limitation of our IV assumption is if the USA Today editor exhibits selection bias in that she picks a shorter headline in order shift readers' attention to more important firm news. To explore this possibility, we examine news-level rather than firm-level characteristics. First, we examine example headlines for Wells Fargo that have similar content, viz., information about Wells Fargo's scandal and settlement with the Consumer Financial Protection Bureau announced September 9, 2016. As Table D2 shows the headline length varies from 27 to 111 characters. That is,

holding the content of the news constant, we see a wide variation in headline lengths, suggesting factors other than the importance of the news dictate their length.

Second, we do a case study of Apple by manually classifying the news content of all USA Today headlines in our sample from 2009 to 2016 with the company name "Apple" in the headline. The five categories of news content conveyed by the headlines include (i) financial, (ii) indirect financial, (iii) product, (iv) people, and (v) legal. We also manually code the headlines to identify ones that had a negative implication for the stock price. Table D3 reveals that in no case, is the type of news content statistically significantly associated with headline length. Moreover, the point estimates are relatively small and go in both directions. While we interpret these case studies with caution, they are consistent with editors headline choices being driven by space limitations.

The logic behind the IV identification is that a shorter headline in the USA Today generates more posts by bloggers from which more crowd wisdom can be extracted. To determine if this temporal pattern is consistent with the data, we evaluate daily data and see if there are more blog posts immediately following news with a shorter headline. Table D4 reveals a significant increase in blog posts in the day, three days, and week after a short headline about Apple is posted. This holds when we control for day of week and week fixed effects as well as the type of news content. Further, we see this relationship is evident for blog posts by top bloggers.

To explore if headline length is independent of news content outside of these case studies and instead is a more generalizable phenomenon as practitioners would argue, we use Capital IQ's Key Developments database, which provides summaries of material news and events that may affect the market value of securities. It monitors over 100 key development types including executive changes, M&A rumors, SEC inquiries, etc. Each key development item includes announcement date and type. We focus on value-relevant events and match the dates of the value relevant events

to USA Today headlines about that firm on that day. We then examine these 431,000 headlines to determine whether value-relevant news content is an important determinant of headline length.

Table D5 shows that there is no association between headline length and value-relevant events either in the cross-section or within-firm over time. Further, when we focus on earnings events as these likely represent the most important news for firms, we find that headline length is associated with 0.7 to 0.8 fewer characters – while statistically significant, this magnitude is too small to alter our results. When we add controls for positive and negative earnings surprises[1] we see no difference in headline length either in magnitude or statistical significance. Finally, when we examine non-earnings key events such as payout announcements, targeting by activist investors, and other key non-earnings events (e.g., announcements of M&A deals), we again see little difference in headline length for these value-relevant events. In each case the point estimate is one character or less, which likely would not noticeable to a reader. Again, including controls for these event types does not affect our estimated relationship between blogging and price informativeness.

[1] We follow the methodology outlined in Livnat and Mendenhall (2006) to define earnings surprises. $\text{SUE1}_{jt} = \frac{X_{jt} - X_{jt-4}}{P_{jt}}$ where X_{jt} is the primary Earnings Per Share (EPS) before extraordinary items for firm j in quarter t , and P_{jt} is the price per share for firm j at the end of quarter t from Compustat. We adjust for stock splits using Compustat's adjustment factor (AJEXQ). We use Compustat's primary (EPSPXQ) or diluted (EPSFXQ) EPS for X_{jt} depending on if the majority of analyst EPS forecasts are based on primary or diluted basis. Similarly, we also then adjust to divide by the number of shares used to calculate primary EPS (CSHPRQ) or diluted EPS (CSHFDQ). To link IBES and CRSP, we use the IBES-CRSP link table available on WRDS. SUE2_{jt} excludes special items from EPS which is equivalent to SPIQ $\times$ 0.65. SUE3_{jt} is defined similarly to SUE1, except X_{jt-4} and X_{jt} are replaced with a measure of analyst's expectations and actual earnings as reported by IBES. The measure for analysts' expectations is the median of latest individual analysts forecasts issued within the 90 days prior to the earnings announcement date. A positive earnings surprise is defined as the intersection of the set of positive observations for SUE1, SUE2, and SUE3. Similarly, a negative earnings surprise is defined as the intersection of the set of negative observations for SUE1, SUE2, and SUE3.

Finally, we employ the model selection technique of LASSO (Efron, Hastie, Johnstone and Tibshirani, 2004) to determine whether particular headline words that predict headline length systematically convey something meaningful about the firm. Table D6 shows the words selected by the variable selection model along with how much variation they explain. Inspecting the words reveals that they are not associated with content but with their own length. For example, the word "available" or "financial" are associated with longer headlines while the words "talk" and "mgmt" are associated with shorter headlines. In contrast, if we thought that having a short headline was conveying important news, we might expect to see words like "beat" as in "beat earnings estimates" or "fear" as in "fear trade conflicts," but we do not see any of these words.

TABLE D.1

Headline Length and Firm Characteristics

Table D1 presents OLS estimates in which the dependent variable is HEADLINE_LENGTH and the explanatory variables are firm characteristics. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	Dependent Variable: HEADLINE_LENGTH
	1
MARKET_TO_BOOK	0.00
	(0.00)
PROFITABILITY	-0.53
	(0.77)
ROE	0.01
	(0.00)
MOMENTUM	1.32*
	(0.79)
FIRM_SIZE	-0.02
	(0.05)
Adj. R^2	0.1%
N	7,538,452

TABLE D.2

Example Newspaper Headlines for Wells Fargo

Table D2 presents example headlines for Wells Fargo from September 2016 from different major newspapers. The articles have similar content, viz., information about Wells Fargo's scandal and settlement with the Consumer Financial Protection Bureau. The table also includes headline length and orders headlines from shortest to longest.

Headline and headline length	
Wells Fargo Is Getting Heat	27
Trust Was Broken at Wells Fargo	31
Wells Fargo Fined for Sales Scam	32
Wild West at Wells Fargo: Our View	34
Wells Fargo Fined $185m for Fake Accounts	41
Wells Fargo Fined $185m; 5,300 Were Fired	41
Wells Fargo to Pay $185 Million Settlement	42
5,300 Wells Fargo Staff Fired Over Bogus Accounts	49
Wells Fargo Fined $185m for Unauthorized Accounts	49
Wells Fargo to Pay $185 Million Fine Over Sales Tactics	55
Wells Fargo Fined $185m for Opening Unauthorized Accounts	57
Wells Fargo to Pay $185 Million Fine Over Account Openings	58
Wells Fargo Fined $185m for Fake Accounts; 5,300 Were Fired	59
What Wells Fargo's $185 Million Settlement May Mean for You	59
Wells Fargo Fined $185 Million for Improper Account Openings	60
Wells Fargo Fined $185m for Unauthorized Accounts; Fires 5,300	62
Wells Fargo Fined $185 Million Over Unwanted Customer Accounts	62
Wells Fargo Fined $185m for Unauthorized Accts That Hurt Customers	66
Wells Fargo Cuts Bank Sales Goals After $185m Fine for Fake Accounts	68
Wells Fargo CEO Defends Bank Culture, Lays Blame With Bad Employees	68
How Wells Fargo's High-Pressure Sales Culture Spiraled out of Control	69
Wells Fargo Fined $185m Over Unauthorized Accounts That Harmed Customers	72
Wells Fargo to Pay $185 Million Settlement for 'Outrageous' Sales Culture	73
Wells Fargo Fined $185m for Unauthorized Accounts; Says It Has Fired 5,300	74
Wells Fargo Fires 5,300 People Over Improper Account Openings; Company Fined $185m	82
Wells Fargo to Pay $185 Million to Settle Allegations Its Workers Opened Fake Accounts	86
Wells Fargo CEO John Stumpf Puts on a Clinic: How to Weasel out of Real Accountability	86
Wells Fargo Fires 5,300 People for Opening Millions of Phony Accounts; Company Fined $185m	90
Wells Fargo Fined $185 Million for Improper Account Openings; 5,300 People Fired in Connection	94
Wells Fargo Settled Over Its Bogus Accounts, but It Still Faces a Fight From Customers and Ex-Employees	103
Wells Fargo Fired 5,300 Workers for Improper Sales Push, the Executive in Charge Is Retiring With $125 Million.	111

TABLE D.3

Headline Length and News Content for Apple

Table D3 presents OLS estimates in which the dependent variable is headline length and the explanatory variables are indicator variables for different types of news content for Apple. This regressions use headlines from the USA Today between 2009 and 2016 that have the company name "Apple" included in the headline. Robust standard errors are reported below the coefficient estimates. Fixed effects (FE) are noted in the bottom rows. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	Dependent Variable: HEADLINE_LENGTH		
Type of headline content	1	2	3
Financial	0.81	0.31	-0.58
	(1.06)	(1.39)	(2.90)
Negative		2.02	1.20
		(1.94)	(1.39)
Financial × negative		-0.06	
		(2.78)	
People			0.60
			(2.11)
Product			-3.43
			(3.04)
Legal			2.00
			(3.17)
Miscellaneous			2.85
			(2.94)
Quarter FE	Yes	Yes	Yes
Adj. R^2	17%	17%	19%
N	524	524	524

TABLE D.4

Headline Length, Subsequent Blogging, and Price Informativeness for Apple

Table D4 presents OLS estimates showing more posts from bloggers in the near future after Apple is featured in news with shorter headline in Panel A. Then, in Panel B, these additional posts from bloggers are linked to price informativeness. In Panel A, the main explanatory variable is an indicator variable for having a short headline. The dependent variables in Columns 1 and 2 are total blog posts on the first day and first three days, respectively. In Column 3 the dependent variable is total blog posts in the first three days by top bloggers. In Columns 4 and 5, the dependent variables are the total blog posts in the first week and the total blog posts with buy or sell recommendations made in the first week after the headline. In Panel B, the dependent variables from Panel A become the focal independent variables. The dependent variable in Panel B is ABS_ORDER_IMBALANCE and it is scaled by a constant (100,000) to make the coefficients easier to read. These regressions use observations at the headline-level from the USA Today between 2009 and 2016 that have the company name "Apple" included in the headline. Robust standard errors are reported below the coefficient estimates. Fixed effects (FE) are noted in the bottom rows. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A. Blog Posts After Short Headline

	POSTS_D_1	POSTS_D_1-3	TOP_POSTS_D_1-3	POSTS_WK_1	REC_POSTS_WK_1
	1	2	3	4	5
SHORT_HEADLINE	2.43**	5.35**	0.77*	10.93**	2.74**
	(1.22)	(2.22)	(0.43)	(4.39)	(1.22)
Controls	Yes	Yes	Yes	Yes	Yes
Day of Week FE	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	No	No
Adj. R^2	41%	50%	64%	6%	3%
N	524	524	524	524	524

Panel B. Change in Price Informativeness

	Dependent Variable: ABS_ORDER_IMBALANCE				
	1	2	3	4	5
POSTS_D_1	2.82*				
	(1.64)				
POSTS_D_1-3		2.00**			
		(0.92)			
TOP_POSTS_D_1-3			10.40***		
			(3.48)		
POSTS_WK_1				2.61***	
				(0.60)	
REC_POSTS_WK_1					9.58***
					(1.92)
Day of Week FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Adj. R^2	48%	1%	2%	5%	5%
N	524	524	524	524	524

TABLE D.5

Headline Length for Value Relevant Events

Table D5 presents OLS estimates in which the dependent variable is HEADLINE_LENGTH and the explanatory variables are value relevant events from Capital IQ's key development data set. This regressions use headlines from the USA Today between 2009 and 2016. In Columns 1 and 2, the focal explanatory variable is an indicator for if a value relevant event occurred that day. In Columns 3 and 4, the focus is on earnings events. We define an earnings announcement as the day of and day after to allow time for a headline to publish. To define POSITIVE_SURPRISE and SURPRISE we follow the methodology outlined in Livnat and Mendenhall (2006). In Columns 5–8, we examine non-earnings key events such as payout announcements, targeting by activist investors, and other non-earnings events. In each regression, controls for MARKET_TO_BOOK, PROFITABILITY, ROE, MOMENTUM, and FIRM_SIZE are included. For a detailed description of each variable, see the definitions in Appendix A. Standard errors clustered by equity are reported below the coefficient estimates. Fixed effects (FE) are noted in the bottom rows. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	Dependent Variable: HEADLINE_LENGTH							
Value-relevant Events	1	2	3	4	5	6	7	8
ANY_KEY_EVENT	0.217	-0.081						
	(0.290)	(0.097)						
EARNINGS_EVENT			-0.679***	-0.756***				
			(0.233)	(0.206)				
POSITIVE_SUE			-0.198	-0.389				
			(0.337)	(0.270)				
NEGATIVE_SUE			-0.827	-0.336				
			(0.764)	(0.543)				
NON-EARNINGS_EVENT					0.424	0.133		
					(0.308)	(0.098)		
PAYOUT_EVENT							1.158***	0.727**
							(0.370)	(0.337)
ACTIVIST_EVENT							0.257	1.050
							(1.141)	(0.936)
OTHER_EVENT							0.256	0.032
							(0.332)	(0.093)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Equity FE	No	Yes	No	Yes	No	Yes	No	Yes
Adj. R^2	9%	23%	9%	23%	9%	23%	9%	23%
N	431,710	431,240	431,710	431,240	431,710	431,240	431,710	431,240

TABLE D.6

LASSO Selection of Words Associated with Headline Length

Table D6 presents estimates connecting common words to headline length. The estimates are based on a LASSO regression. This technique helps with the problem of picking out the relevant words from a larger set (i.e., variable selection) by pushing estimates of some coefficients to be exactly zero. The words are listed in the order in which they are selected to be included in the model. Column 1 shows the LASSO adjusted coefficient estimate for the word, and Column 2 displays the cumulative variance explained when that word is included. Given that the variance explained plateaus toward the end, only the first twenty words selected into the model are listed.

	Dependent variable: HEADLINE_LENGTH	R^2 when variable is included
Key headline words	1	2
quarterly	24.89	2.25%
available	10.69	7.42%
annual	5.76	7.94%
stories	-6.21	8.11%
market	4.18	8.34%
talk	-14.78	8.41%
events	-8.25	8.71%
financial	10.47	10.78%
agreement	11.25	10.87%
million	8.84	10.96%
morning	5.07	11.08%
mgmt	-4.32	11.26%
billion	6.79	11.31%
investors	6.53	11.59%
capital	6.95	11.62%
sells	-0.28	11.76%
china	5.32	11.89%
week	7.13	12.22%
fund	5.79	12.37%
bank	4.40	12.37%
Additional Word Controls	Yes	
Firm Characteristic Controls	Yes	
N	7,538,452	

E Additional Robustness Tests

In this Appendix, we use two recent advances in the literature on IV estimation to test and relax the exclusion restriction (Angrist, Lavy and Schlosser, 2010, Conley, Hansen and Rossi, 2012, Kippersluis and Rietveld, 2018). First, we follow the approach emphasizing the identification of subgroups for which the IV is irrelevant as a test of the exclusion restriction. Second, we explore the consequences of relaxing the exclusion restriction assumption using a partial identification approach. If the IV correlates with some unobserved variable affecting price informativeness, then the point estimate for crowd wisdom will be biased. Thus, what would be useful is to investigate the robustness of the IV estimator in relation to the extent of potential bias from violating the exclusion restriction in our estimates. To achieve this goal, we estimate a bound for this potential bias by using an approach that produces a plausible estimate (a term proposed by (Conley, Hansen and Rossi, 2012)).

The first approach we use to detect and investigate sensitivity to violations of the exclusion restriction is to estimate the direct effect of the IV on the outcome in a subsample for which the IV likely does not affect the treatment variable, which following the prior literature we refer to as the "zero-first-stage test" (Angrist et al., 2010). This test provides potential evidence for or against the exclusion restriction. This informal test can never verify the exclusion restriction, but it can build confidence that the exclusion restriction is satisfied. The intuition for the test follows the logic of a placebo test. If the first stage is zero, then the reduced-form effect of the IV on the outcome variable should be zero too if the exclusion restriction is satisfied. In our setting, a zero first stage implies that short headlines have no effect. We would expect short headlines to have no effect if there is no blogging associated with an equity. Given that firm size is the most important

determinant of financial blogging, we use equities in the lowest decile of market capitalization as our zero-first-stage test group.

Panel A of Table E1 summarizes these regression results. Column 1 displays the estimates for price nonsynchronicity and Column 2 for absolute order imbalance. We see that the equities in this zero-first-stage group have no relationship with the IV. In contrast, for the remaining sample, which we expect to be related to the IV, we see positive, statistically significant relationship. Thus, these subsample tests provide support in favor of the exclusion restriction being satisfied. Further, we find the results in support of this identifying assumption being satisfied are evident for both proxies of price informativeness.

Next, we explore the importance of the exclusion restriction holding precisely for the IV coefficients on crowd wisdom that we estimate. Specifically, we follow the local-to-zero method outlined in Conley et al. (2012) that incorporates different degrees of exclusion restriction violation to generate a plausibly exogenous IV estimate. Their method involves estimating the following equation:

$$\text{(E.1)}\quad \text{PRICE\_NONSYNCHRONICITY}_{it} = \alpha + \beta \text{CROWD\_WISDOM}_{it} + \theta X_{it} + \zeta Z_{it} + \delta_t + \epsilon_{it}$$

where ζ is a parameter measuring the plausibility of the exclusion restriction and Z_{it} is the IV. The difference between the primary specification and this specification is the presence of the term ζZ_{it}. Typically, ζZ_{it} is removed from the specification, because the exclusion restriction holds and $\zeta = 0$. If the exclusion restriction is not violated, then this method produces the same confidence interval as a traditional instrumental variable regression.

If the exclusion restriction is violated, then different degrees of violation can be incorporated into the original estimate. Mathematically, the estimates that allow for exclusion restriction viola-

tion are Bayesian. The estimates come from updating a prior without exclusion restriction violation to a posterior with violations. To see this more clearly, let the distribution for β be approximated as follows:

$$\hat{\beta} \sim N\left(\beta, Var_{2SLS}\right) + A\zeta$$

$$Prior = \zeta \sim N\left(0, \Omega_{\zeta}\right) \tag{E.2}$$

$$\hat{\beta} \sim N\left(\beta, Var_{2SLS} + A\Omega_{\zeta}A'\right)$$

where Var_{2SLS} is the variance-covariance matrix and A is the projection matrix from estimating the two-stage least-squares estimator from Equation E.1. Exclusion restriction violations are represented by ζ. In contrast to the traditional IV approach, where ζ is assumed to be zero, we can replace that assumption with the assumption that ζ is close to, but not necessarily equal to, zero. We do this by specifying a distribution for ζ. This can either be symmetrical around zero, for example, by specifying a normal distribution centered at zero as shown in the equation above. If the underlying economic arguments suggest that the direction of potential exclusion restriction violations are ambiguous, an uninformative prior of 0 is reasonable.

Alternatively, Kippersluis and Rietveld (2018) develop a method for estimating a sensible prior distribution to use as an input when generating plausibly exogenous IV estimates Conley et al. (2012). The intuition for their approach to generating an informative prior is to use observable changes in coefficient estimates and standard errors associated with the IV stemming from different subsamples of the data. In particular, Kippersluis and Rietveld (2018) suggest using the zero-first-stage test discussed above in conjunction with a formula based on Imbens and Rubin (2015) for

estimating the variance. Specifically, $\Omega_\zeta = \left(0.125\sqrt{S_Z^2 + S_R^2}\right)$ where S_Z is the standard error for the zero-first-stage test and S_R is the standard error from the remaining sample.

Panel B of Table E1 summarizes a variety of plausibly exogenous IV estimates for price nonsynchronicity in Column 1 and for absolute order imbalance in Column 2. For comparison, the OLS and IV results are also included in the top two rows. The results from the plausibly exogenous IV estimates reveal that crowd wisdom is positively and significantly associated with price informativeness even if the exclusion restriction does not hold precisely. To put the mean used in the prior distribution into context, we can compare it with known determinants of price informativeness. For example, the estimated effect of profitability on price nonsynchronicity is 0.024 with a standard error of 0.008. The sensitivity tests suggest that excluding an unobservable factor as important as profitability would result in a point estimate for crowd wisdom that is about 15% smaller but still significant at the 95th percentile. While certainly different from the estimate under the assumption that the exclusion restriction holds precisely, the plausibly exogenous IV estimate remains economically meaningful. Thus, these analyses lead us to conclude that our IV is useful, even if there are potentially minor exclusion restriction violations.

TABLE E.1

Assessing the Exclusion Restriction Assumption

Table E1 summarizes tests for evaluating the impact of potential violations of the exclusion restriction assumption. Panel A examines the direct effect of the IV on price informativeness for the zero-first-stage and remaining group. The zero-first-stage groups equities in the lowest decile of market capitalization. Panel B reports the OLS, IV, and a variety of plausibly exogenous IV estimates. The plausibly exogenous IV is estimated using Conley et al. (2012) and the prior distribution with Imbens and Rubin uncertainty follows the procedure in Kippersluis and Rietveld (2018). The dependent variable is PRICE_NONSYNCHRONICITY and ABS_ORDER_IMBALANCE in Columns 1 and 2, respectively. The IV is an indicator for whether the equity has below median headline length in the USA Today in a given quarter. Controls are the same equity, news content, and headline controls as in previous regressions. Fixed effects (FE) are noted in the bottom rows. Standard errors clustered by equity are reported below the coefficient estimates. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. For a detailed description of each variable, see the definitions in Appendix A.

Panel A. Short Headline and Price Informativeness

	Dependent Variable:	
	PRICE_NONSYNCHRONICITY	ABS_ORDER_IMBALANCE
	1	2
Zero-first-stage Group (i.e., microcap stocks)	-0.235	0.088
	(0.209)	(0.398)
N	8,120	8,120
Remaining Group	0.123***	0.161***
	(0.048)	(0.018)
N	73,081	73,081

Panel B. Crowd Wisdom and Price Informativeness

	1	2
OLS	0.062***	0.073***
	(0.010)	(0.004)
IV	0.084***	0.137***
	(0.029)	(0.012)
Plausibly Exogenous IV with $\zeta \sim N(0.00, 0.001)$	0.084***	0.137***
	(0.035)	(0.022)
Plausibly Exogenous IV with $\zeta \sim N(0.01, 0.001)$	0.078**	0.131***
	(0.035)	(0.022)
Plausibly Exogenous IV with $\zeta \sim N(0.03, 0.001)$	0.066*	0.118***
	(0.035)	(0.022)
Plausibly Exogenous IV with$\zeta \sim N(0.05, 0.001)$	0.054	0.106***
	(0.035)	(0.022)
Plausibly Exogenous IV with $\zeta \sim N(0.00, 0.003)$	0.084*	0.137***
	(0.044)	(0.035)
Plausibly Exogenous IV with $\zeta \sim N(0.00, 0.005)$	0.084	0.137***
	(0.052)	(0.045)
Plausibly Exogenous IV with Imbens and Rubin uncertainty	0.084**	0.137***
	(0.033)	(0.033)
Controls	Yes	Yes
Quarter FE	Yes	Yes
N	81,201	81,201

References

Angrist, J.; V. Lavy; and A. Schlosser. "Multiple Experiments for the Causal Link between the Quantity and Quality of Children." *Journal of Labor Economics*, 28 (2010), 773–824.

Antweiler, W., and M. Z. Frank. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *Journal of Finance*, 59 (2004), 1259–1294.

Chen, H.; P. De; Y. Hu; and B.-H. Hwang. "Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media." *Review of Financial Studies*, 27 (2014), 1367–1403.

Conley, T. G.; C. B. Hansen; and P. E. Rossi. "Plausibly Exogenous." *Review of Economics and Statistics*, 94 (2012), 260–272.

Cookson, A. J., and M. Niessner. "Why don't we agree? Evidence from a social network of investors." *Journal of Finance*, 75 (2020), 173–228.

Das, S. R., and M. Y. Chen. "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web." *Management Science*, 53 (2007), 1375–1388.

Efron, B.; T. Hastie; I. Johnstone; and R. Tibshirani. "Least angle regression." *The Annals of Statistics*, 32 (2004), 407–499.

Farrell, M.; T. C. Green; R. Jame; and S. Markov, "The Democratization of Investment Research and the Informativeness of Retail Investor Trading." (2019), unpublished working paper.

Imbens, G. W., and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences.*, New York, NY: Cambridge University Press (2015).

Kippersluis, H. V., and C. A. Rietveld. "Beyond plausibly exogenous." *Journal of Econometrics*, 21 (2018), 316–331.

Kogan, S.; T. J. Moskowitz; and M. Niessner, "Fake News: Evidence from Financial Markets." (2020), unpublished working paper.

Livnat, J., and R. R. Mendenhall. "Comparing the Post-Earnings Announcement Drift for Surprises Calculated from Analyst and Time Series Forecasts." *Journal of Accounting Research*, 44 (2006), 177–205.

Loh, R. K., and G. M. Mian. "Do accurate earnings forecasts facilitate superior investment recommendations?" *Journal of Financial Economics*, 80 (2006), 455–483.