Crowdsourcing Economic Forecasts

Mike Aguilar*, Anessa Custovic, and Amar M. Patel

University of North Carolina at Chapel Hill, Department of Economics

November 26, 2018

Abstract

Economic forecasts are often disseminated via a survey of professionals (i.e. "Consensus"). In this paper we compare and contrast the Consensus with a crowdsourced alternative wherein anyone may submit a forecast. We focus on U.S. Nonfarm Payrolls and find that, on average, Consensus is more accurate, but the best crowdsourced forecasters are superior to the best Consensus forecasters. We also find that information plays a key role. When the Consensus is uncertain and herds together, the crowdsourced forecasts appear to be more accurate. Our findings provide evidence that crowdsourcing might provide a valuable supplement to traditional macroeconomic forecasts.

Keywords: wisdom of crowds; macroeconomic forecasting; expert forecast; macroeconomic news; forecasting

*Correspondence to: University of North Carolina at Chapel Hill, Department of Economics, 141 South Rd, Chapel Hill, NC, United States

E-mail address: maguilar@email.unc.edu

1 Introduction

Economic forecasts are crucial inputs for policy makers, regulators, businesses, and investors. A common method of obtaining such forecasts is via a poll of professionals who are asked to provide their assessments of the future state of the economy. Such polls typically fall under the moniker "Consensus". Within the field of economics and finance, key examples of Consensus polls include the Survey of Professional Forecasters[1] and Consensus Economics[2]. In this paper we explore an alternative to Consensus, wherein the forecasts are crowdsourced.

Defining Crowdsourcing Crowdsourcing can be defined as the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from the on-line community[3].

For our purposes, crowdsourcing differs from Consensus in two key ways: i) openness of the poll, and ii) composition of the participants. Regarding openness, poll administrators may allow anyone to contribute (i.e. an open poll) or restrict contributions to specific individuals (i.e. a closed poll).

Throughout the balance of this paper we consider Consensus polls to be closed to the public, inviting only professional forecasters to participate. Meanwhile, we refer to crowdsourced polls as those that are open to anyone, including professionals and non-professionals alike.

Crowdsourcing is common in applications such as image analysis, biomedical, and urban planning. However, crowdsourcing is rare in economics and finance. A few examples include Chen et al. (2014) that shows that textual analysis of user's posts on seekingalpha.com, a popular opinion forum for stock market investors, has predictive power for future stock returns and earnings surprises. Brown et al. (2017) shows that sentiment analyses via Twitter messages may improve prediction accuracy on sports betting exchanges. Peeters (2018) finds that crowdsourcing in soccer valuations is more accurate

[1]See `https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/`

[2]See `https://www.consensuseconomics.com/`

[3]Adapted from https://www.merriam-webster.com/dictionary/crowdsourcing

than typical models. Adebambo and Bliss (2015) find that crowdsourced earnings forecasts on the Estimize [4] platform are more accurate than Consensus.

Virtues of Crowdsourcing Surowiecki (2005) explores the "wisdom of the crowds" by identifying the circumstance under which crowdsourcing is likely to be successful. Specifically, a wise crowd is necessarily: i) diverse, ii) independent, and iii) decentralized. Diversity of opinion allows for the broadest possible information set to be used. An independent decision making process may prevent individual errors from being correlated with one another. Furthermore, independence is more likely to lead to new information being introduced to the crowd, thereby fostering diversity. Decentralization allows for individuals to specialize and acquire local knowledge.

This characterization of a "wise crowd" is complemented and expanded upon by Simmons et al. (2011) who suggest that a crowd is wise if it is diverse, independent, knowledgeable, and motivated to be accurate. Diversity and independence are criterion shared by both Simmons et al. (2011) and Surowiecki (2005). The decentralized aspect of Surowiecki (2005) rests upon the poll participant's ability to aggregate local information, which is akin to the knowledgeable criterion of Simmons et al. (2011). However, the motivation for accuracy noted by Surowiecki (2005) is unique. In the following we provide a brief survey of the literature, organized by these four criterion (diversity, independence, decentralization, and motivation for accuracy).

First, Batchelor and Dua (1995) attribute the success of their sample of crowdsourced forecasts to the diversity of their participants. Similarly, while studying a crowdsourced platform of company earnings forecasts, Adebambo et al. (2016) finds that crowdsourced forecasts are more accurate than Consensus, and that this relative forecast accuracy increases with forecaster diversity. Specifically, relative to Consensus, the crowdsourced forecasters tend to have more varied professional backgrounds.

Second, regarding independence, Brown (1993) attributes the success of crowdsourcing in the context of earnings forecasts to the averaging across a large pool of forecasts, thereby allowing the idiosyncratic error of individual forecasters to be canceled out. This

[4] See `https://www.estimize.com/`

virtue of crowdsourcing is largely supported by the forecast model averaging work of Clemen (1989), among others, who notes that forecast combination schemes can be more accurate than any single forecasting technique.

Third, Adebambo et al. (2016) suggests that the decentralized nature of the Estimize crowdsourcing platform permits the forecasters to more effectively incorporate local information. Similarly, Lang et al. (2016) find that when forecasting a wide range of market outcomes for an anonymous Fortune 100 company, crowdsourcing is 70% more accurate than standard forecasting techniques. The authors attribute the relative accuracy of crowdsourcing to its ability to exploit specialized information obtained by the forecasters.

Fourth, incentives may influence the behavior of Consensus forecasters in a way that mutes their motivation for accuracy. As shown by Ager et al. (2009) and Gallo et al. (2002), members of the Consensus forecast may be attracted to the mean forecast, inducing a herding behavior. In the context of corporate Earnings per Share (EPS) forecasts, Jame et al. (2016) notes myriad ways in which incentive structures and conflicts of interest would prevent a crowd of traditional Wall Street analysts from being "wise". Specifically, the authors suggest that sell-side analysts are dependent on managers for information and subsidized by investment banking revenues, which causes analysts to have incentives to bias their research to please managers and facilitate investment banking activities. Furthermore, Jame et al. (2016) highlights that approximately half of Estimize EPS forecasts are issued in the two days prior to the earnings announcement date, while less than 2% of Thomson Reuters' Institutional Brokers' Estimate System forecasts, which is used as a Consensus measure, are issued in the same period. The authors suggest that this finding supports the the notion that sell-side analysts are somehow adversely incentivized not to include all relevant information.

Forecast Evaluation There are typically three dimensions along which crowdsourcing and Consensus forecasts are evaluated: accuracy, bias, and efficiency.

Regarding accuracy, Hyndman and Koehler (2006) evaluate the relative merits of myriad measures categorized as follows: i) scale dependent measures, such as Men Squared Error (MSE), ii) measures based on percentage errors, such as Mean Absolute Percentage

Error (MAPE), iii) measures based on relative errors, such as Mean Relative Absolute Error (MRAE), iv) relative measures, such as Relative Mean Absolute Error (RelMAE), and v) scaled errors, such as their preferred Mean Absolute Scaled Error (MASE). Much of the authors' evaluation revolves around the measure's ability to facilitate cross-series comparisons. In our setting, we are forecasting only a single series, and as such, such considerations are not relevant.

Dovern and Weisser (2009) examine forecasts for GDP and inflation for G7 countries provided by Consensus Economics. Their measure of accuracy is RMSE. Loungani (2001) studies Consensus Forecasts for GDP across 69 countries, and uses MAE, RMSE, and Thiel's inequality coefficient. While studying company earnings forecasts, Jame et al. (2016) compares the accuracy of crowdsourced forecasts to a consensus measure using six measures, similar to those explored by Hyndman and Koehler (2006).

As Lamont (2002) suggests, forecasters may produce biased forecasts because they may not have an incentive to report a value that minimizes expected square forecast error, but may optimize profits, wages, credibility, shock value, marketability, or political power. Clements (2018) offers herding as a potential explanation for this behavior. According to Clements (2018) herding occurs when forecasters put undue weight on the views of others, which either moves their forecasts toward or away from the Consensus view in a way that is detrimental to forecast accuracy. Clements et al. (2007) show that a Mincer Zarnowitz type regression can be used to test herding. Holden and Peel (1990) provide an augmented version of this test based off of the forecast error and a test of the associated intercept.

Clements et al. (2007) emphasize that these tests are typically performed for one horizon at a time, suggesting that a more powerful approach would come from pooling across horizons. The situation that Clements et al. (2007) describes is known as Fixed Event Forecasts. In our setting, a poll participant might submit a forecast on December 15th for January's Nonfarm Payrolls number. That same participant might submit another forecast on December 17th for that same January Nonfarm Payrolls number. This is in contrast to Rolling Event Forecasts where, for example, on the 15th of each month the

poll participants are asked to submit their forecast for the following month's Nonfarm Payrolls number. The Estimize database is a type of fixed event forecast. However, the Clements et al. (2007) approach is not applicable since theirs was derived for a single forecaster (i.e. the FED), whereas ours has multiple forecasters. Clements et al. (2007) point to the work of Davies and Lahiri (1995) as a possibility for including multiple poll participants, but their approach is built for rolling event forecasts. Moreover, our measure of Consensus is a rolling event forecast, and so our methodology must be robust to both type of surveys.

Forecast efficiency is typically evaluated via forecast revisions. Ager et al. (2009) build on the pooled approach of Clements et al. (2007) and develop a test of whether or not forecast revisions are predictable. Again, as outlined above, these pooled approaches are not applicable in our setting. Moreover, our database does not include forecast revisions, which limits our ability to explore efficiency.

Forecasting Non Farm Payrolls The particular application of this paper is to explore the accuracy of the Estimize crowdsourcing platform at forecasting U.S. Nonfarm Payrolls (NFP). This is the same platform studied by Jame et al. (2016), but our focus is on the labor markets rather than company earnings. Perhaps closest to our application is Montgomery et al. (1998), which compares the accuracy of a number of forecasting methods for the U.S. unemployment rate. The universe of models considered includes a variety of linear and nonlinear time series models as well as a consensus survey forecast of the U.S. unemployment rate. Their findings show that forecasting performance can be greatly improved by combining subjective methods with traditional time series models.

Our focus on the NFP is motivated by its stated importance as an indicator of the health of the US economy. As indicated by Taylor (2010), the NFP is known to impact Federal Reserve policy. Moreover, Miao et al. (2013) shows that the broad U.S. equity market responds to NFP releases. The Federal Reserve Bank of San Francisco (2004) suggests three main reasons why NFP can be considered the most important indicator for measuring the overall health of the U.S. economy. First, the NFP is highly correlated with the overall economy and is published on a higher frequency (monthly) than GDP

(quarterly). Second, the NFP report can be disaggregated by industry, permitting for a detailed understanding of labor market conditions. Last, monthly national NFP data is complemented by publications at the state and metropolitan level, allowing for regional analysis. Moreover, out of the 80 measures of the economy tracked by Estimize, it is the one with the longest time history and largest number of forecasts.

To the best of our knowledge, we are the first to explore the accuracy of the Estimize platform for economic forecasts. In this paper we uncover two key findings. First, we find that on average, Consensus is slightly more accurate than Estimize for the initial release, and is equally as accurate for the revised. The source of the relative forecasting accuracy does not seem to reside with the differing number of forecasters in each platform, nor in the regularity with which the forecasters participate. Instead, we find that a small group of "all-stars" are important. Specifically, the best Estimize forecasters are significantly more accurate than the best Consensus forecasters. This finding is consistent with the notion of super-forecasters from Tetlock and Gardner (2015) and Clement (1999) who suggests that forecaster ability improved the accuracy of analysts' forecasts. Second, when the pool of Consensus forecasters is uncertain and herds together, the Estimize forecasts appear to be more accurate.

The rest of the paper proceeds as follows. In Section 2 we describe the Estimize platform and the related crowdsourcing forecasts, as well as provide summaries of the Consensus forecasts and our object of interest, the monthly NFP. In Section 3 we evaluate the accuracy of the Consensus and crowdsourced forecasts. In Section 4 we identify the source of the relative forecasting accuracy, and in Section 5 we conclude.

2 Data

In this section we describe the data used in our study. In the first subsection below we detail the Estimize platform and its forecasts of NFP. In the second and third subsections we summarize the Consensus NFP forecasts and the NFP data, respectively. We conclude with exploratory analysis.

2.1 The Estimize Platform

In this subsection we describe the structural details of the Estimize platform. Estimize was launched in 2011 as an "open financial estimates platform designed to collect forward looking financial estimates from independent, buy-side, and sell-side analysts, along with those of private investors and academics" [5]. The platform began sourcing Earnings Per Share (EPS) estimates on equities, which today has more than 50,000 contributors and 650,000 estimates across 2,200 stocks. On its EPS platform, the Estimize user base is split evenly between investment professionals, independent researchers, individual traders, and students (see Drogen and Jha (2013)). This differs from traditional aggregating platforms like the Institutional Brokers Estimate System (I/B/E/S), generally considered to be the Consensus for EPS forecasting, which consists entirely of sell-side equity research analysts (i.e. professionals).

Estimize launched its Economics platform in the first quarter of 2014. The platform provides users with the ability to forecast over 80 economic indicators across developed and major emerging markets. Major US indicators consistently receive close to 50 estimates per release, whereas international indicators typically receive less than 20. Nonfarm Payrolls are the most frequently forecasted indicator on Estimize, with 2,491 forecasts created from 4/2014 through 3/2017, compared with 1,380 for the next-highest indicator.

Estimize provides an intuitive user interface, which should permit parties of varying levels of sophistication to participate easily. The platform offers a visual history of user accuracy, as well as the actual values for the indicators release and subsequent revisions. After at least one forecast has been submitted for a monthly release of an indicator, an average will appear in a "Value" box, allowing users to observe what the community believes. Forecasts that are deemed unrealistic, or are entered in the wrong units, are "flagged" by Estimize as outliers[6].

A useful characterization of the Estimize platform would be that of a real-time Delphi method. The survey literature (for example, Green et al. (2007), Aengenheyster et al.

[5] https://www.estimize.com/about

[6] We note the possibility of an anchoring bias induced by the display of this "Value" box. A detailed examination is beyond the scope of our paper.

(2017) and others) defines a conventional Delphi process as one in which experts within a certain field submit their forecasts on a future topic along with their reasoning behind the forecast. The process is iterative, with multiple rounds where the forecasts and justifications are shared after the completion of each round. The expert participants would then use the information from the previous round to revise their prior forecasts.

The real-time Delphi is similar to the conventional process, where participants are asked to provide a forecast for some future event. However, a real-time Delphi does not have to be iterative, and feedback can be provided within rounds. Since a real-time Delphi does not require rounds, participants can submit as many forecasts as they want at any point in time. The forecasts are accepted up until the actual object of interest is revealed, after which participants are instantly able to determine their forecast accuracy. Aengenheyster et al. (2017) note that a real-time Delphi can only truly be administered on-line.[7]

The Estimize platform conforms to this definition of a real-time Delphi method. It is administered on-line, where participants are free to log in and out as frequently as they choose. They are able to submit as many forecasts as they wish, until the economic data is released. Once participants log onto the platform they are able to see a complete history of previous forecasts and a quantitative assessment of all participant responses.

For the analysis in this paper, Estimize graciously provided to us all participant forecasts for Nonfarm Payrolls (NFP) during the period of May 2014 to February 2017. There are 35 events in our sample, with 3,171 total observations. In those cases where there was more than one forecast by a user for a given month, we followed the procedure used by our Consensus proxy and kept only the most recent forecast.

2.2 Consensus Forecasts

As a proxy for Consensus, we use responses from the survey of professional forecasters conducted by Bloomberg. Note that Bloomberg polls roughly 100 professional forecasters each month. This is similar to services like Reuters and Dow Jones, and greater

[7]See Aengenheyster et al. (2017) for an in depth description of Delphi methods.

than the Survey of Professional Forecasters conducted by the Federal Reserve Bank of Philadelphia. As such, we believe Bloomberg serves as a credible proxy for Consensus.

2.3 NFP Data

The U.S. Nonfarm Payrolls indicator is part of the monthly Employment Situation Report produced by the Bureau of Labor Statistics (BLS).[8] The NFP indicator measures the change in the number of workers in the non-farm sector of the U.S. economy. The NFP data is compiled on a monthly basis by the Bureau of Labor Statistics (BLS). Each report contains information on the coverage month's payrolls data (i.e. initial release) as well as revisions to the payrolls data for the two months immediately prior (i.e. revised) that take into account new government and business reports as well as a recalculation of seasonality factors. There are also annual benchmark revisions based on unemployment insurance tax records. In this paper, we examine both the initial release as well as the revised numbers.

Although U.S. data on NFP is available as early as 1948, our sample begins in 2014 due to the short tenure of the Estimize sample.

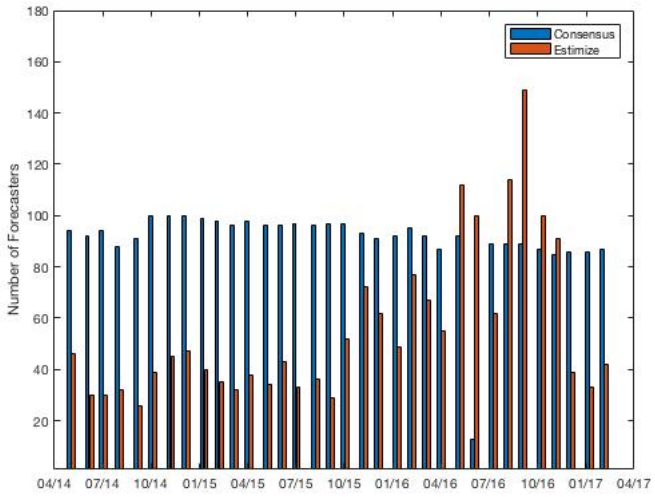
2.4 Exploratory Analysis

Figure 1 depicts the evolution of the number of forecasters over time. We find that the number of forecasters participating in the Consensus poll remains quite constant over our sample, with about 90 participants each month. Meanwhile, the number of Estimize participants is generally growing over the sample, but highly variable.

Table 1 characterizes the sample moments of all four data sets (As Reported, Revised, Estimize, Consensus). We run two-tailed t-tests of equal means of Estimize and Consensus for each of the monthly sample statistics listed in Table 1. Though we see that on average Estimize tends to forecast a higher number of payrolls each month, with a mean forecast (averaged over forecasts and then over time) of 215,000 compared to that of 208,000

[8]Further details on the BLS revisions can be found on the BLS website at: https://www.bls.gov/web/empsit/cesbmart.htm

Figure 1: Monthly Participants for Estimize and Consensus
This figure depicts the numbers of participants making NFP forecasts in each month from April 2014 - February 2017. The number of Estimize forecasters is indicated in red, and the number of Consensus forecasters, as measured by Bloomberg, is indicated in blue.



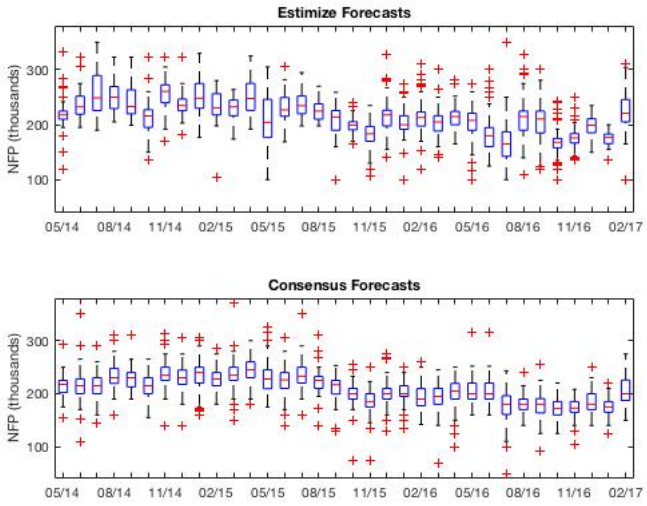
for Consensus, these differences are not statistically significant. We find that the only statistically significant differences are with the standard deviations and IQR.

The sample statistics suggest that Estimize might be more variable with a standard deviation of 28 compared to Consensus' value of 23. Estimize's IQR and skew is also higher than that of Consensus, suggesting it has a wider spread in forecasts than Consensus does. We these claims via test the hypothesis that Estimize appears to have a wider spread than Consensus does in the forecasts. We run one-tailed t-tests of equal means for the monthly standard deviations and the monthly IQR. The results indicate that Estimize does in fact have a wider spread in forecasts than Consensus. The p-value from testing that Estimize has a greater standard deviation than Consensus is $< 0.01\%$ and the p-value that Estimize has a greater IQR than Consensus is 1%. This finding is depicted in Figure 2, which shows that the variability of forecasts for Estimize is much larger than the spread for Consensus. This could potentially suggest less herding amongst the participants, which we discuss in Section 4.

Table 1: NFP (As Reported, Revised), Estimize, and Consensus
This table provides sample statistics for the As Reported and Revised NFP data, as well as for the forecasts from Estimize and Consensus, as proxied by the survey of professional forecasts conducted by Bloomberg.
The sample is from May 2014 - February 2017. Data is in thousands ('000)

	As Reported	Revised	Estimize	Consensus	P-value
Mean	221	221	215	208	0.23
Median	220	225	214	207	0.21
Max	321	423	287	278	0.33
Min	126	24	145	135	0.24
Std. Dev.	55	71	28	23	<0.01***
IQR	110	91	32	27	0.03**
Skew	-0.05	0.06	0.05	-0.03	0.64
N	35	35	35	35	

Figure 2: Estimize and Consensus Monthly Forecasts
This figure provides box plots of individual NFP forecasts for each month of the sample (May 2014 - February 2017). The top panel captures forecasts from Estimize, while the bottom panel captures forecasts from Consensus, as measured by Bloomberg.



3 Measuring Forecast Accuracy

Are crowdsourced forecasts of NFP more accurate than Consensus forecasts? To address this question we follow Jame et al. (2016) by evaluating forecasts with several measures of accuracy. In Table 2 we detail the accuracy measures, wherein we denote F as the forecast, which could come either from Estimize or Consensus, and we denote A as the actual NFP value, which could either be "As Reported" or "Revised". The Absolute Forecast Error (AFE), Mean Squared Forecast Error (MSFE), and Proportional Mean

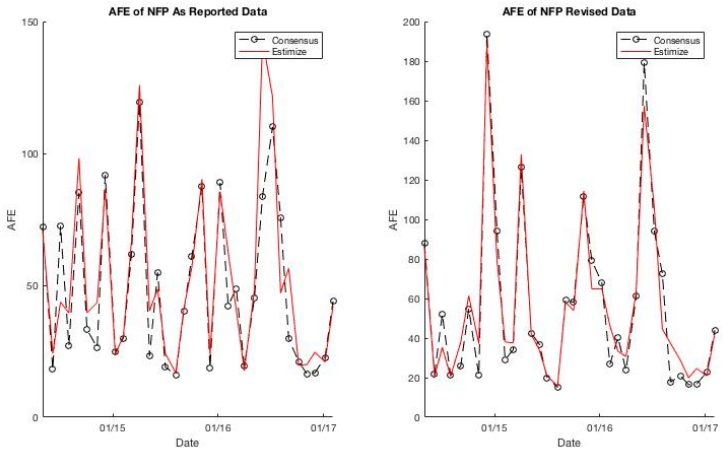
Absolute Forecast Error (PMAFE) all provide a sense of accuracy that is agnostic to the sign of the forecast error. Meanwhile, Forecast Error (FE), Percent Forecast Error (PFE), and Standard Forecast Error (SFE) all provide a measure of the bias of the forecasts, which accommodates for the sign of the error.

Table 2: Definitions of forecasting accuracy and bias where F denotes the forecast of the NFP value, and A denotes the actual NFP value. Subscripts i capture individual forecasts, and t indicates the calendar month of the NFP event.

Absolute Forecast Error	AFE	$\lvert F_{i,t} - A_t \rvert$
Mean Squared Forecast Error	MSFE	$(F_{i,t} - A_{i,t})^2$
Proportional Mean Absolute Forecast Error	PMAFE	$\frac{(AFE_{i,t} - \frac{1}{N}\sum_{i=1}^{N} AFE_{i,t})}{\frac{1}{N}\sum_{i=1}^{N} AFE_{i,t}}$
Forecast Error	FE	$F_{i,t} - A_{i,t}$
Percent Forecast Error	PFE	$\frac{F_{i,t} - A_{i,t}}{A_{i,t}}$
Standardized Forecast Error	SFE	$\frac{FE_{i,t}}{\sigma_{FE_t}}$

In Figure 3's plots, we overlay the As Reported (Revised) release of the NFP data with the Estimize and Consensus averages for each month. A quick visual inspection suggests little difference between the forecasting platforms. In fact, the correlation between Estimize and Consensus is 0.86. However, closer inspection reveals important differences.

Figure 3: Monthly average AFE of Estimize and Consensus for NFP Data
Figure 3 plots the monthly average AFE (as defined in Table 2) for Consensus forecasts (in black) and Estimize forecasts (in red) during the period May 2014- February 2017. Panel a) computes the AFE relative to the As Reported NFP data, while panel b) computes the AFE relative to the Revised Release.



In the top panel of Table 3 we present each of the six measures outlined in Table 2 for both the Estimize and Consensus forecasts as they pertain to the As Reported NFP data. In the last row of the table we list the p-value of a two tailed t-test examining the equality of these monthly sample measures. The bottom panel of Table 3 repeats this structure for the Revised NFP data. We find that there are indeed statistically significant differences between Estimize and Consensus for all of our sample measures for both As Reported and Revised data.

With regard to the top panel of As Reported data, Estimize appears approximately 8.5% less accurate than Consensus, with Estimize AFE of 53 versus the Consensus AFE of 48. However, Consensus appears to have much larger bias, with a FE of -12.26, the negative sign implying that Consensus tends to underestimate the As Reported value. Meanwhile, Estimize has a slight positive bias, with FE of 0.72. We confirm that Consensus has a much larger bias than Estimize with one-tailed t-tests.

The bottom panel of Table 3 tells a slightly different story. Consensus and Estimize are about equally as accurate in forecasting the Revised NFP data. Consensus remains highly negatively biased, but Estimize is now negatively biased as well. Note that Consensus is approximately twice as biased as Estimize, with a FE of -12.94 versus -6.28 for Estimize. Though Estimize has a larger with Revised NFP data than with As Reported, one-tailed t-tests again confirm that Consensus remains more biased than Estimize.

Also note that the Consensus appears more accurate at forecasting the As Reported value than at forecasting the Revised number. This finding is consistent with Loungani (2001) who notes that previous studies have suggested it is more likely that forecasters are trying to forecast the initial release of the data rather than revisions.

4 The Source of Relative Forecasting Accuracy

In this section we examine several potential sources of the relative forecasting accuracy between Consensus and Estimize. In subsection 4.1 we examine whether the ability and activity of the forecaster impacts the relative forecasting accuracy of our samples. In

Table 3: Sample Statistics for As Reported and Revised NFP Data

This table provides sample statistics for each of the accuracy and bias measures outlined in Table 2 for both the Estimize and Consensus forecasts. The top panel bases the computations off of As Reported NFP data and the bottom panel bases the computations off of Revised NFP data. The statistics reported are the averages of monthly values. The column n reports the average number of participants throughout our sample and p reports the p-value for a two-tailed t-test of the equality of means. The *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

As Reported	n	AFE	$MSFE$	$PMAFE$	FE	PFE	SFE
Estimize	56	52.50	4366.83	0.35	0.72	0.15	0.02
Consensus	91	48.43	3622.11	0.23	-12.26	0	-0.43
$p-value$		$< .001^{***}$	$< .001^{***}$	$< .001^{***}$	$< .001^{***}$	$< .001^{***}$	$< .001^{***}$
Revised							
Estimize	56	55.84	5300.34	0.43	-6.28	0.21	-0.20
Consensus	91	55.60	5427.60	0.44	-12.94	0.20	-0.42
$p-value$		0.07^{*}	$< .001^{***}$	$< .001^{***}$	$< .001^{***}$	$< .001^{***}$	$< .001^{***}$

subsections 4.2, 4.3, 4.4 we explore whether forecasting horizon, number of forecasters, and boldness of forecasts impact forecast accuracy. In subsection 4.5 we examine the role of public and private information. Finally, in subsection 4.6 we run a series of multivariate regressions to explore the marginal explanatory power of these various potential drivers of relative forecast accuracy.

4.1 Forecaster Ability and Activity

In this subsection we examine whether the relative forecasting power can be attributed to a specific group of forecasters. This type of analysis is similar in spirit to Clement (1999) who found that forecaster ability and resources impact the accuracy of analysts' forecasts. First, we ask if the average forecasting power comes from only the best forecasters i.e. the "All Stars"? We begin by identifying individual forecasters within the Consensus sample. We restrict our analysis to all those who had at least two forecasts over the entire sample. We compute the mean AFE for each forecaster over the sample. We then rank the analysts from lowest AFE ("best") to highest AFE ("worst"). We repeat for the Estimize analysts.

Table 4 provides the AFE for each of the "All Star" forecasters. The first column provides the forecaster's rank, the second column represents the AFE for Consensus, the third column is the AFE for Estimize, the fourth column gives the difference between those AFE's, and the fifth column is the p-value for a two-tailed t-test of equal means.

The results of the t-tests in the top panel of Table 4 suggest that there are differences among the All Stars from each forecasting pool. The positive signs of the difference suggest that the AFE for Estimize tends to be smaller than AFE for Consensus, implying greater accuracy for Estimize. These findings are echoed in the bottom panel of Table 4 for Revised data.

Of particular note is the dramatically lower AFE for the top 3 Estimize All Stars with AFE values equal to 2, 5, 8, for As Reported data compared to 23, 24, 24, for Consensus, as well as AFE values equal to 8, 8, 9 for Estimize Revised data compared to 17, 23, 26 for Consensus. This may not be a fair comparison. The average number of forecasts submitted by the top 10 Consensus All Stars is 19 while the top 10 Estimize All Stars submitted only 3 forecasts. Hence, it may not be surprising that the top Consensus forecasters have higher AFE's.

Due to this participation disparity we conduct a similar analysis with the most "regular" forecasters from Consensus and from Estimize. By looking at the most "regular" forecasters, we hope to glean if the apparent out-performance of the Estimize All-Stars is only driven by their low participation rates within our sample size. Clement (1999) constructed a similar forecast regularity variable as a proxy for forecaster ability. He reasons that a forecaster with more experience would have likely acquired more knowledge which would in turn improve their forecasting ability.

Within the entire pool of Consensus and Estimize forecasters we count the number of times each individual forecaster submits forecasts over our sample of 34 months. We create the following four bins of forecasting "regularity": those who participate less than 25%[9] of the time, those who participate between $25-49\%$ of the time, those who participate between $50-74\%$ of the time, and those who participate at least 75% of the time. Table 5 computes the AFE for each bin for the Consensus and Estimize samples.

Table 5 reveals several important aspects of our data. First, Consensus forecasters participate much more regularly than Estimize forecasters. Second, among the most regular forecasters, the t-test results suggest that there are differences, and the positive

[9]As mentioned earlier, we drop submissions from any participants that only contribute a single forecast during our sample.

Table 4: Accuracy of the All Star Forecasters for As Reported and Revised NFP Data
The top 10 rows are results using reported NFP releases and the bottom 10 rows are results using revised NFP releases. Rank is determined by calculating individual forecaster's AFE over time and then rank ordering from lowest AFE to highest. Only those who had at least two forecasts over the entire sample were considered. AFE^C is the AFE as defined in Table 2 for Consensus and AFE^E is defined analogously for Estimize. "Diff." is the difference $AFE^C - APE^E$, and *pval* reports the p-value for the test of the equality of means. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

As Reported	Rank	AFE^C	AFE^E	$Diff.$	*pval*
	1	23.00	2.00	21.00	0.30
	2	23.71	5.25	18.46	0.02**
	3	24.20	8.33	15.87	0.01***
	4	25.22	10.30	14.92	0.01***
	5	27.61	12.24	15.37	0.00***
	6	29.26	13.87	15.39	0.00***
	7	30.55	15.03	15.52	0.00***
	8	31.53	16.03	15.50	0.00***
	9	32.39	16.91	15.48	0.00***
	10	33.08	17.64	15.44	0.00***
Revised					
	1	17.29	7.50	9.79	0.20
	2	23.48	8.00	15.48	0.01***
	3	25.56	8.72	16.84	0.01***
	4	27.42	10.42	17.00	0.01***
	5	28.58	11.88	16.70	0.01***
	6	29.73	13.32	16.41	0.00***
	7	31.05	14.42	16.63	0.00***
	8	32.10	15.49	16.61	0.00***
	9	32.92	16.36	16.56	0.00***
	10	33.63	17.13	16.50	0.00***

sign of the difference (6.73) suggests greater accuracy for Estimize. This finding holds for the Revised numbers as well. By contrast, the least regular forecasters have a negative sign of the difference in AFE (-6.73). This finding holds for the Revised data as well.[10]

4.2 Forecasting Horizon

Adebambo et al. (2016) notes that the forecast horizon might be a useful determinant of the relative forecasting power because forecasts made closer to the announcement date should contain more information, and thus be more accurate. We investigate this claim by counting the number of days between a particular month's NFP forecast and the associated NFP release date. We group the counts into discrete buckets of horizon: 1

[10]We note that due to the small size of the Estimize sample, any inferences should be made cautiously.

Table 5: "Regular" Forecasters for As Reported and Revised NFP Releases

This table presents forecasting accuracy within four bins: those who participate less than 25% of the time, those who participate between $25-49\%$ of the time, those who participate between $50-74\%$ of the time, and those who participate at least 75% of the time. AFE^C is the AFE as defined in Table 2 for Consensus and AFE^E is defined analogously for Estimize. N^C and N^E are the number of forecasters in each bin. "Diff." is the difference $AFE^C - AFE^E$, and *pval* reports the p-value for the test of the equality of means. The top panel presents results for the As Reported data while the bottom panel presents results for the Revised data. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

As Reported	AFE^C	N^C	AFE^E	N^E	$Diff.$	*pval*
$\geq 75\%$	54.62	83	47.89	4	6.73	0.09*
$50-74\%$	49.43	12	52.01	16	-2.63	0.38
$25-49\%$	55.66	10	50.57	24	5.09	0.24
$< 25\%$	45.24	17	54.94	122	-9.69	0.03**
Revised						
$\geq 75\%$	60.21	83	52.06	4	8.15	0.10*
$50-74\%$	55.49	12	55.16	16	0.33	0.93
$25-49\%$	62.72	10	56.40	24	6.32	0.25
$< 25\%$	51.1	17	60.43	122	-9.36	0.10*

week prior to the NFP release, 2 weeks prior to the NFP release, 3 weeks prior to the NFP release, 4 weeks prior to the NFP release and over 4 weeks prior to the NFP release. Within each bucket we compute the absolute forecast error (AFE) for the associated consensus forecasts, as well as for the associated Estimize forecasts. The column labeled "Diff." in Table 6 computes the mean Consensus AFE minus the mean Estimize AFE. The column labeled "p" captures the values from the two tailed t-test with the null hypothesis of that difference being equal to zero.

In general, we find that for both As Reported and Revised NFP data Consensus forecasters are more accurate than Estimize forecasters. This advantage tends to grow as the horizon lengthens. Importantly, however, there is a reversal of relative accuracy at very long horizon forecasts. When forecasts are made more than 4 weeks prior to the event, Estimize tends to be more accurate, with this being especially true for Revised data.

It is possible that the advantage of Consensus at shorter horizons reflects their superior ability to access and/or incorporate high frequency public information. Analogously, the advantage of Estimize forecasters with Revised NFP data at very long horizons, might reflect their ability to incorporate private information. We explore these ideas further in subsection 4.5.

Table 6: Forecasting Horizon for As Reported and Revised NFP Data

In this table, horizon is defined as the difference between a particular month's NFP forecast and the associated NFP release. The horizon is grouped into buckets: 1 week prior to release, 2 weeks prior to release, 3 weeks prior to release, 4 weeks prior to release and over 4 weeks prior to the NFP release. AFE^C is the AFE as defined in Table 2 for Consensus and AFE^E is defined analogously for Estimize. N^C and N^E are the number of forecasters in each horizon. "Diff." is the difference $AFE^C - AFE^E$, and *pval* reports the p-value for the test of the equality of means. The top panel presents results for As Reported NFP data while the bottom panel presents results for Revised data. As usual *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

As Reported	AFE^C	N^C	AFE^E	N^E	$Diff.$	*pval*
week 1	46.48	2374	53.23	1748	-6.75	0.00***
week 2	49.01	681	56.51	90	-7.50	0.11
week 3	53.63	83	60.47	37	-6.84	0.42
week 4	46.10	20	69.67	47	-23.57	0.04**
>week 4	82.48	92	77.96	63	4.52	0.54
Revised						
week 1	50.41	2374	57.08	1748	-6.68	0.00***
week 2	56.21	681	61.64	90	-5.44	0.29
week 3	62.61	83	60.73	37	1.88	0.85
week 4	48.75	20	69.93	47	-21.18	0.05**
>week 4	181.13	92	70.07	63	111.06	0.00***

4.3 Number of Forecasters

Does the size of a group of forecasters (i.e. the crowd) impact it's accuracy? The forecast model averaging work of Timmermann (2006) suggests that this might be the case. Moreover, Adebambo et al. (2016) also notes that the number of forecasters might influence the relative accuracy of differing forecasting groups. A group with more forecasters could result in less herding and thus possibly influence forecasting accuracy.

We examine the impact of the size of the crowd upon relative forecast accuracy. For each month in our sample we compute the number of forecasters in the Estimize sample (N^E). We also compute the number of forecasters in the Consensus sample (N^C). We then compute the difference in that month: $N^{Diff} = N^C - N^E$. We repeat this process for each month in our sample, and create three buckets of roughly equal size: the top tercile of N^{Diff}, the middle tercile, and the bottom tercile. Within each bucket we compute the mean AFE, compute the difference, and test if that difference is equal to zero.

We find that the difference in the number of analysts is not significant. We run the same analysis for Revised NFP data, and find that the number of analysts again remains insignificant.

This finding that the number of analysts does not impact forecast accuracy is in

line with that of Ashton and Ashton (1985) and Batchelor and Dua (1995). Ashton and Ashton (1985) found that when aggregating subjective forecasts one only needs to combine a small number of forecasts in order to realize most of the benefits. Batchelor and Dua (1995) demonstrate how combining and aggregating individual forecasts can increase the accuracy and the utility of the forecast. They also found diminishing gains with the inclusion of an additional forecast after a certain N.

4.4 Boldness of Forecasts

As per Jame et al. (2016), the boldness of a forecast might be relevant to explaining relative accuracy since it is a key attribute in the theories of herding and reputation. The boldness statistic measures the extent to which the individual forecast deviates from the Consensus. The larger the boldness statistic is, the less herding there is between analysts. Boldness is computed as follows,

$$\frac{|F_{i,t}^{pop} - \overline{F}_t^{full}|}{\overline{F}_t^{pop}}$$

where $F_{i,t}^{pop}$ is the forecast i of the population (either Estimize or Consensus) made for month (t), $\overline{F}_t^{full}$ is the average forecasts made for month (t) among all populations (i.e. Estimize and Consensus combined), and $\overline{F}_t^{pop}$ is the average forecast for month (t) within that population (Estimize or Consensus).

We compute boldness for every forecast from the Estimize and the Consensus samples. The average monthly boldness for the Estimize sample is 0.13 and the average monthly boldness for the Consensus sample is 0.12.

To examine if boldness is associated with relative forecasting accuracy, we group the boldness for Consensus (Estimize) into tercile bins. For each bin we compute the AFE, take the difference between the Consensus and Estimize AFE and test the equality of means via a two tailed t-test.

In Table 7 we find that the average forecast error is in fact different for Consensus than it is for Estimize, across all boldness bins. However, the differences do not appear to vary systematically across bins for either As Reported nor Revised, suggesting little

impact of boldness upon forecast accuracy.

Table 7: Boldness for As Reported and Revised NFP Data

The top 3 rows are results using As Reported NFP releases and the bottom 3 rows are results using Revised NFP releases. Boldness is computed as $\frac{|F_{i,t}^{pop} - \overline{F}_t^{full}|}{\overline{F}_t^{pop}}$ for every individual forecast, i, from the Consensus and Estimize sample and grouped into terciles. AFE^C is the AFE as defined in Table 2 for Consensus and AFE^E is defined analogously for Estimize. N^C and N^E are the number of forecasters in each tercile. "Diff." is the difference $AFE^C - AFE^E$, and *pval* reports the p-value for the test of the equality of means. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

As Reported	AFE^C	N^C	AFE^E	N^E	$Diff.$	*pval*
1st	46.38	988	50.55	623	-4.17	0.00***
2nd	46.34	1182	54.31	637	-7.97	0.00***
3rd	51.93	1080	59.91	658	-7.97	0.00***
Revised						
1st	61.26	988	53.17	623	8.09	0.00***
2nd	49.90	1182	58.71	637	-8.82	0.00***
3rd	56.69	1080	63.56	658	-6.87	0.01***

4.5 The Role of Information

Following Barron et al. (1998) we investigate whether the information sets of Estimize and Consensus can explain their relative forecasting accuracy. We define information diversity, ρ, as follows:

$$\rho = \frac{C}{V},$$

where $V = \frac{1}{N}\sum_{i=1}^{N} V_i$ and $C = \frac{1}{N}\sum_{i=1}^{N} C_i$. Barron et al. (1998) defines V_i as the level of uncertainty for analyst i and C_i as the mean covariance between analyst i's beliefs and the rest of the analysts' beliefs. V can be thought of as the overall uncertainty for all analysts, and C can be interpreted as the common uncertainty. This common uncertainty, C, is different from the overall uncertainty, V, because C is the common uncertainty based off of the analysts' reliance on incorrect common information. All analysts are relying on some degree of public information, which is potentially inaccurate, to form their individual forecasts.

Within this framework, ρ is then the ratio of common uncertainty to overall uncertainty. As $\rho \to 1$, all analysts' beliefs are converging to some common belief based off of the publicly available information used. Herding is generally defined as a high degree of agreement among predictions by analysts. Thus, a high ρ potentially indicates a high degree of herding among analysts. As $\rho \to 0$, all analysts' beliefs are diverging from the imprecise common information available to all.

Unfortunately, the above defined ρ requires data that is unobservable, and so cannot be directly calculated. To circumvent this, Barron et al. (1998) shows that the measures of V and ρ can be recovered with observable data as such:

$$V = (1 - \frac{1}{N})D + SE$$

$$\rho = \frac{SE - \frac{D}{N}}{(1 - \frac{1}{N})D + SE}.$$

In the above definition, D is the unconditional sample variance of the forecasts and SE is the expectation of mean squared forecast. The interpretations of V and ρ remain

unchanged despite using the above construction.

Barron et al. (1998) suggests that it is not meaningful to look at individual values of ρ, nor should one really draw much inference from that measure. Instead, he suggests using a large number of calculated ρ statistics when conducting any sort of analysis. As such, we do not partition the sample, as we did in previous sections. Instead, we calculate the statistics at the aggregate level.

In Table 8 we report the aggregate ρ and V for both Consensus and Estimize samples. We compute differences and test the equality of those differences. The top panel reports results for As Reported NFP data, while the bottom panel reports results for Revised data. We find no statistically significant difference at the monthly level between the calculated measures of ρ and V, for neither As Reported nor Revised, indicating that information sets are not a driver of relative forecast accuracy.

Table 8: Uncertainty & Diversity for As Reported and Revised NFP Data
The top 2 rows are results using As Reported NFP releases and the bottom 2 rows are results using Revised NFP releases. V and ρ are calculated as follows: $V = (1 - \frac{1}{N})D + SE$ and $\rho = \frac{SE - \frac{D}{N}}{(1-\frac{1}{N})D+SE}$. The column called Consensus contains the calculated statistics for the Consensus group while the other column is for Estimize. "Diff." is the difference $AFE^C - AFE^E$, and *pval* reports the p-value for the test of the equality of means. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

As Reported	Consensus	Estimize	$Diff.$	$pval$
ρ	0.98	0.98	0.00	0.72
V	3644.50	4394.10	-749.60	0.49
Revised				
ρ	0.98	0.98	0.00	0.80
V	5450.00	5327.60	122.40	0.95

4.6 Multivariate Regressions

Following the approach of Clement (1999) we implement a series of multivariate regressions to explore further the relative forecasting accuracy of Estimize and Consensus.

Similar to Adebambo et al. (2016) we construct our dependent variable as the monthly mean AFE for Consensus minus the mean AFE for Estimize. If this dependent variable is positive (negative), this implies that Estimize is more accurate than (less) Consensus. We examine several independent variables as highlighted in the previous subsections. The horizon is calculated as the average number of days prior to release that forecasts are

made for Estimize ($\overline{Hor}^E$) and Consensus ($\overline{Hor}^C$). We compute the number of Estimize forecasters (N^E) and Consensus forecasters (N^C) that submitted a forecast in any given month t. Measures of dispersion, ρ^C (ρ^E), and uncertainty, V^C (V^E), are calculated as defined above at each time t.

We also calculate the All Stars and the Regulars as described in the subsections above and include them in the multivariate regression. The regressor $Best^C\%_t - Best^E\%_t$ is defined as the percentage of All-Stars at month t for Consensus minus the percentage of All Stars at month t for Estimize.[11] The difference in Regulars, where Regulars are characterized as those who participated in at least half of our sample period, is defined analogously.

$$\begin{aligned}\overline{AFE}_t^C - \overline{AFE}_t^E &= \alpha_0 + \alpha_1(\overline{Hor}_t^C - \overline{Hor}_t^E) + \alpha_2(N_t^C - N_t^E) + \alpha_3(\rho_t^C - \rho_t^E) + \alpha_4(V_t^C - V_t^E) \\ &\quad + \alpha_5(Best^C\%_t - Best^E\%_t) + \alpha_6(Reg^C\%_t - Reg^E\%_t) + e_t \end{aligned} \tag{1}$$

Table 9: Multivariate Regression for As Reported NFP Data

Results for Equation 1 estimated via OLS on As Reported data. The column "Estimate" reports the estimated coefficient for each variable. $Hor^C - Hor^E$ is the difference in horizon, where horizon is defined as in the previous subsection, for Consensus and Estimize. $N^C - N^E$ is the difference in individual participants for Consensus and Estimize. $\rho^C - \rho^E$ and $V^C - V^E$ are defined as in the previous subsection. $Best^C\% - Best^E\%$ is the difference in the percentage of top forecasters in each month for Consensus and Estimize. $Reg^C\% - Reg^E\%$ is the difference in the percentage of regular forecasters in each month for Consensus and Estimize. SE reports the standard error of each estimated coefficient. T-stat reports the calculated t-statistic for each estimate and *pval* reports the p-value for the corresponding t-statistic. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

	Estimate	SE	t-stat	*pval*
(Intercept)	-1.50	6.68	-0.22	0.82
$Hor^C - Hor^E$	1.19	0.34	3.52	0.00***
$N^C - N^E$	-0.0039	0.027	-0.15	0.89
$\rho^C - \rho^E$	590.5	103.33	5.71	0.00***
$V^C - V^E$	0.0061	0.0006	10.97	0.00***
$Best^C\% - Best^E\%$	-3.43	10.52	-0.33	0.75
$Reg^C\% - Reg^E\%$	-6.38	8.51	-0.75	0.46

Table 9 presents results for the As Reported NFP data. The horizon is significant

[11] Due to the nature of this analysis, and the irregularity with which the best forecasters participate over the sample, we augment the cutoff for All Star classification that was defined in Section 4.1 to be the top 30 analysts. Supplemental exercises suggest that our general conclusions are robust to this choice of cutoff.

Table 10: Multivariate Regression for Revised NFP Data

Results for Equation 1 estimated via OLS on Revised data. The column "Estimate" reports the estimated coefficient for each variable. $Hor^C - Hor^E$ is the difference in horizon, where horizon is defined as in the previous subsection, for Consensus and Estimize. $N^C - N^E$ is the difference in individual participants for Consensus and Estimize. $\rho^C - \rho^E$ and $V^C - V^E$ are defined as in the previous subsection. $Best^C\% - Best^E\%$ is the difference in the percentage of top forecasters in each month for Consensus and Estimize. $Reg^C\% - Reg^E\%$ is the difference in the percentage of regular forecasters in each month for Consensus and Estimize. SE reports the standard error of each estimated coefficient. T-stat reports the calculated t-statistic for each estimate and *pval* reports the p-value for the corresponding t-statistic. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

	Estimate	SE	t-stat	p
(Intercept)	7.66	8.38	0.914	0.37
$Hor^C - Hor^E$	-0.83	0.26	-3.16	0.00***
$N^C - N^E$	-0.036	0.032	-1.12	0.27
$\rho^C - \rho^E$	579.90	96.00	6.04	0.00***
$V^C - V^E$	0.0052	0.00058	8.91	0.00***
$Best^C\% - Best^E\%$	-13.22	14.18	-0.93	0.36
$Reg^C\% - Reg^E\%$	0.47	8.69	0.054	0.96

and positive, implying that as Estimize forecasters tend to make forecasts closer to the event, they tend to be more accurate than Consensus. We also find that ρ differences are significant and positive, implying that more herding in Consensus relative to Estimize implies less accuracy of Consensus relative to Estimize. Similarly, the significant positive coefficient on V implies that the higher uncertainty of Consensus relative to Estimize is associated with a loss of relative forecasting accuracy for Consensus. The results in Table 10 for Revised NFP data are largely consistent with those in Table 9.

In both Tables 9 and 10 we see that neither the difference in the share of Regulars nor the difference in the share of All Stars are statistically significant determinants of relative forecasting accuracy. This finding stands in contrast to Timmermann (2006) who found that the source of relative accuracy can be traced in part to the abilities of the forecasters. It is possible that the explanatory power of All Stars and Regulars is subsumed by other regressors, or that our inference is contaminated by a small sample size in either length of calender time (T) and/or the number of participants (N).

To explore this second possibility, we construct a pseudo matched pairs sample around the Regulars and All Stars. We use the seven most active forecasters from Consensus and the seven most active from Estimize, thereby ensuring that we have multiple observations for each month of the data sample[12]. By using the top seven most active we are also en-

[12]Our general findings are robust to this choice.

suring that each month in our data sample has roughly the same number of observations. Since the seven most active forecasters from Consensus all participated in each month of our data sample, we were able to match them with the seven most active forecasters from Estimize. Note, only one of the most active forecasters from Estimize had participated in each month of our sample. The least active Estimize participant (out of the most active seven) participated a total of 23 out of 34 months in our sample period. This technique allowed us to expand our sample from 34 observations to a total of 189 observations.

With this expanded sample we specify our model similar to that above with the exception that, we do not include the number of analysts, but instead include a measure of boldness for each forecast. The measures of dispersion and uncertainty remain fixed over each i and only vary with time, t.

$$\overline{AFE}^{C}_{i,t} - \overline{AFE}^{E}_{i,t} = \alpha_0 + \alpha_1(\overline{Hor}^{C}_{i,t} - \overline{Hor}^{E}_{i,t}) + \alpha_2(\rho^{C}_{t} - \rho^{E}_{t}) + \alpha_3(V^{C}_{t} + V^{E}_{t}) + \alpha_3(bold^{C}_{i,t} + bold^{E}_{i,t}) + e_{i,t} \tag{2}$$

The results in Table 11 show some similarities with the multivariate regressions reported in Table 9. In particular, we see that the difference in uncertainty remains highly significant and are associated with higher relative forecasting accuracy for Estimize.

Table 11: "Regulars" Multivariate Regression for As Reported NFP Data
Results for Equation 2 estimated via Panel OLS for As Reported data. All regressors are constructed at the individual forecaster level. $Hor^C - Hor^E$ is the difference in forecasting horizon, $bold^C - bold^E$ is the difference in the measure of boldness, while $\rho^C - \rho^E$ and $V^C - V^E$ are as defined previously. The column "Estimate" reports the estimated coefficient for each variable. SE reports the standard error of each estimated coefficient. T-stat reports the calculated t-statistic for each estimate, and *pval* reports the p-value for the corresponding t-statistic. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

	Estimate	SE	t-stat	*pval*
(Intercept)	2.11	2.46	0.86	0.39
$Hor^C - Hor^E$	0.72	0.39	1.84	0.07*
$\rho^C - \rho^E$	101.02	168.23	0.60	0.55
$V^C - V^E$	0.0023	0.0007	3.55	0.00***
$bold^C - bold^E$	20.89	16.33	1.28	0.20

We repeat this exercise with the Revised data. The results depicted in 12 echo those of Table 11, with the exception that boldness matters. Specifically, a bolder Consensus is related to less relative forecasting accuracy for Consensus, or said differently, herding amongst the regular Consensus forecasters would improve their performance compared

to that of Estimize.

Table 12: Regulars Multivariate Regression for Revised NFP Data
Results for Equation 2 estimated via Panel OLS for Revised data. All regressors are constructed at the individual forecaster level. $Hor^C - Hor^E$ is the difference in forecasting horizon, $bold^C - bold^E$ is the difference in the measure of boldness, while $\rho^C - \rho^E$ and $V^C - V^E$ are as defined previously. The column "Estimate" reports the estimated coefficient for each variable. SE reports the standard error of each estimated coefficient. T-stat reports the calculated t-statistic for each estimate, and *pval* reports the p-value for the corresponding t-statistic. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

	Estimate	SE	t-stat	p
(Intercept)	-1.28	2.42	-0.53	0.60
$Hor^C - Hor^E$	-0.33	0.34	-0.98	0.33
$\rho^C - \rho^E$	-96.01	148.18	-0.65	0.52
$V^C - V^E$	0.0049	0.00086	5.76	0.00***
$bold^C - bold^E$	36.11	15.70	2.30	0.02**

Based upon our findings above with the Regulars, we also explore the "All-Stars". We create a panel of the $AFE^C_{\lfloor i \rfloor} - AFE^E_{\lfloor i \rfloor}$ for $i = \{1, ..., 30\}$.[13] Using this approach, we were able to expand our panel to 79 observations. Within this panel dataset we no longer have all 34 months as before, but rather 17 months since there were a number of months that only had one All Star from Estimize and one from Consensus. Since many of the regressors we include require calculations based off of multiple forecasters, we removed these months as they could potentially provide highly unreliable regressors. Further, calculating monthly averages and standard deviations would be meaningless with so few observations.

We then project these differences in the AFE values upon the differences in horizons, differences in boldness, and differences in ρ and V, noting that ρ and V are computed with the universe of the All Stars at each point in time.

Our results in Table 13 support the previously reported multivariate regression results within this section. For the As Reported NFP data, uncertainty (V) is highly statistically significant, positively related to relative forecasting accuracy, and similar in magnitude to the results in Table 11. Similarly, the results of Table 14 echo the results of Table 12 for Revised NFP data. Differences in uncertainty (V) and boldness are important. In contrast to earlier findings, ρ is important for the Revised data, implying that for All

[13]Our choice of 30 analysts is a compromise between focusing on the few elite forecasters from each group, while ensuring that we have sufficient regularity of forecasts. Supplemental exercises suggest that our general findings are not sensitive to this choice.

Stars there is some differences in information that impacts the accuracy of the forecasts.

Table 13: All Stars Multivariate Regression for As Reported NFP Data
OLS regression output for As Reported NFP data using only the best forecasters. All regressors are also constructed at the individual level. $Hor^C - Hor^E$ is the difference in forecasting horizon, $bold^C - bold^E$ is the difference in the measure of boldness, while $\rho^C - \rho^E$ and $V^C - V^E$ are defined as previously. The column "Estimate" reports the estimated coefficient for each variable. SE reports the standard error of each estimated coefficient. T-stat reports the calculated t-statistic for each estimate and *pval* reports the p-value for the corresponding t-statistic. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

	Estimate	SE	t-stat	*pval*
(Intercept)	-4.26	2.90	-1.47	0.15
$Hor^C - Hor^E$	1.0086	0.43	2.35	0.02**
$\rho_C - \rho^E$	92.96	96.44	0.96	0.34
$V^C - V^E$	0.0090	0.0017	5.14	0.00***
$bold^C - bold^E$	2.13	21.27	0.10	0.92

Table 14: All Stars Multivariate Regression for Revised NFP Data
OLS regression output for As Reported NFP data using only the best forecasters. All regressors are also constructed at the individual level. $Hor^C - Hor^E$ is the difference in forecasting horizon, $bold^C - bold^E$ is the difference in the measure of boldness, while $\rho^C - \rho^E$ and $V^C - V^E$ are defined as previously. The column "Estimate" reports the estimated coefficient for each variable. SE reports the standard error of each estimated coefficient. T-stat reports the calculated t-statistic for each estimate and *pval* reports the p-value for the corresponding t-statistic. As usual, *, **, *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The sample used is from May 2014 - February 2017.

	Estimate	SE	t-stat	p
(Intercept)	-1.85	2.89	-0.64	0.52
$Hor^C - Hor^E$	0.65	0.412	1.56	0.12
$\rho^C - \rho^E$	231.20	86.74	2.67	0.01***
$V^C - V^E$	0.0050	0.00087	5.74	0.00***
$bold^C - bold^E$	46.80	20.61	2.27	0.03**

5 Conclusion

To the best of our knowledge, we are the first to examine the accuracy of Estimize's crowdsourced forecasts of U.S. Nonfarm payrolls. Estimize appears to offer a forecast aggregating platform that is comparable to the traditional "Consensus".

Similar to Batchelor and Dua (1995) and Adebambo et al. (2016), on average we find that Estimize crowdsourced forecasts are similar to those of Consensus. However, closer inspection reveals important differences. When considering the As Reported NFP number, Consensus appears more accurate in terms of mean squared error, but tends to have a large negative bias. Meanwhile, when considering the Revised NFP number, Estimize and Consensus are equally accurate, with the Consenus maintaining a negative

bias. Consensus tends to be more accurate at shorter horizons, while Estimize tends to be more accurate with very long horizons.

Consistent with Timmermann (2006) the source of these differences can be traced in part to the relative abilities of participants in each platform. For instance, the most accurate Estimize forecasters seem to be superior to the most accurate Consensus forecasters. Moreover, consistent with Ager et al. (2009), we find that during episodes of Consensus herding, the crowdsourced platform appears relatively more accurate. Ager et al. (2009) link this herding to the lack of anonymity among professional forecasters and the associated reputational concerns associated with poor, out-of-consensus forecasts.

Our study is limited by the short tenure of the Estimize crowdsourcing platform. In time, a natural extension of our work would follow the approach of Adebambo et al. (2016) who judges the ability for forecasting platforms to reflect market expectations of the underlying event. In addition, since Estimize maintains crowdsourced forecasts for over 80 U.S. economic indicators, the approach of our paper could easily be repeated for other aspects of the economy. Lastly, again following Adebambo and Bliss (2015), associated trading strategies could be developed for economically sensitive asset prices.

References

Adebambo, B. and Bliss, B. (2015). The value of crowdsourcing: Evidence from earnings forecasts.

Adebambo, B., Bliss, B., and Kumar, A. (2016). Geography, diversity, and accuracy of crowdsourced earnings forecasts.

Aengenheyster, S., Cuhls, K., Gerhold, L., Heiskanen-Schttler, M., Huck, J., and Muszynska, M. (2017). Real-time delphi in practice a comparative analysis of existing software-based tools. *Technological Forecasting and Social Change*, 118:15–27.

Ager, P., Kappler, M., and Osterloh, S. (2009). The accuracy and efficiency of the consensus forecasts: A further application and extension of the pooled approach. *International Journal of Forecasting*, 25(1):167 – 181.

Ashton, A. H. and Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, 31(12):1499–1508.

Barron, O. E., Kim, O., Lim, S. C., and Stevens, D. E. (1998). Using analysts' forecasts to measure properties of analysts' information environment. *The Accounting Review*, 73(4):421–433.

Batchelor, R. and Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, 41(1):68–75.

Brown, A., Rambaccussing, D., Reade, J. J., and Rossi, G. (2017). Forecasting with social media: Evidence from tweets on soccer matches. *Economic Inquiry*, 56(3):1748–1763.

Brown, L. D. (1993). Earnings forecasting research: its implications for capital markets research. *International Journal of Forecasting*, 9(3):295 – 320.

Chen, H., De, P., Hu, Y., and Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559 – 583.

Clement, M. B. (1999). Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter? *Journal of Accounting and Economics*, 27(3):285–303.

Clements, M. P. (2018). Do macroforecasters herd? *Journal of Money, Credit and Banking*, 50(2-3):265–292.

Clements, M. P., Joutz, F., and Stekler, H. O. (2007). An evaluation of the forecasts of the federal reserve: a pooled approach. *Journal of Applied Econometrics*, 22(1):121–136.

Davies, A. and Lahiri, K. (1995). A new framework for analyzing survey forecasts using three-dimensional panel data. *Journal of Econometrics*, 68(1):205–227.

Dovern, J. and Weisser, J. (2009). Accuracy , unbiasedness and efficiency of professional macroeconomic forecasts : An empirical comparison for the g 7 by jonas dovern.

Drogen, L. and Jha, V. (2013). Generating abnormal returns using crowdsourced earnings forecasts from estimize.

Federal Reserve Bank of San Francisco (2004). Why does the federal reserve consider nonfarm payroll employment to be an important economic indicator? `https://www.frbsf.org/education/publications/doctor-econ/2004/june/nonfarm-jobs-payroll-employment/`.

Gallo, G., Granger, C., and Jeon, Y. (2002). Copycats and common swings: The impact of the use of forecasts in information sets. *IMF Staff Papers*, 49(1):167 – 181.

Green, K., Armstrong, J., and Graefe, A. (2007). Methods to elicit forecasts from groups: Delphi and prediction markets compared. *Foresight: The International Journal of Applied Forecasting*, (8):17–20.

Holden, K. and Peel, D. A. (1990). On testing for unbiasedness and efficiency of forecasts*. *The Manchester School*, 58(2):120–127.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688.

Jame, R., Johnston, R., Markov, S., and Wolfe, M. C. (2016). The value of crowdsourced earnings forecasts. *Journal of Accounting Research*, 54(4):1077–1110.

Lamont, O. A. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior & Organization*, 48(3):265–280.

Lang, M., Bharadwaj, N., and DiBenedetto, C. A. (2016). How crowdsourcing improves prediction of market-oriented outcomes. *Journal of Business Research*, 69(10):4168 – 4176.

Loungani, P. (2001). How Accurate Are Private Sector Forecasts: Cross-Country Evidence From Consensus Forecasts of Output Growth. *International Journal of Forecasting*, 17(3):419–432.

Miao, H., Ramchander, S., and Zumwalt, J. K. (2013). S&p 500 index-futures price jumps and macroeconomic news. *Journal of Futures Markets*, 34(10):980–1001.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1998). Forecasting the u.s. unemployment rate. *Journal of the American Statistical Association*, 93(442):478–493.

Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, 34(1):17 – 29.

Simmons, J. P., Nelson, L. D., Galak, J., and Frederick, S. (2011). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, 38(1):1–15.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Taylor, N. (2010). The determinants of future u.s. monetary policy: High-frequency evidence. *Journal of Money, Credit and Banking*, 42(2/3):399–420.

Tetlock, P. E. and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown Publishing Group, New York, NY, USA.

Timmermann, A. (2006). Forecast combinations. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 4, pages 135–196. Elsevier.