Has herding undermined the collective wisdom embodied in the Blue Chip consensus?

J. Peter Ferderer
Edward J. Noble Professor of Economics
and Department Chair
Department of Economics
Macalester College
1600 Grand Avenue
St. Paul, MN 55105, USA
ferderer@macalester.edu

and

Adam Freedman
Macalester College
1600 Grand Avenue
St. Paul, MN 55105, USA
ferderer@macalester.edu

August 23, 2016

The authors are grateful to Stuart Glosser, Marc Tomljanovich and participants at the 2015 Liberal Arts Macro Workshop and 2014 Midwest Economic Association Conference for many useful comments and suggestions. The authors also thank Lee McPheters for sharing information about the Lawrence R. Klein Award for forecasting accuracy. This research was made possible by the Macalester College internal grant program and we enthusiastically acknowledge this support.

J. Peter Ferderer
Edward J. Noble Professor of Economics
and Department Chair
Department of Economics
Macalester College
1600 Grand Avenue
St. Paul, MN 55105, USA
ferderer@macalester.edu

and

Adam Freedman
Macalester College
1600 Grand Avenue
St. Paul, MN 55105, USA
ferderer@macalester.edu

ABSTRACT: The Prediction Diversity Theorem shows that the collective wisdom of a forecasting group deteriorates when its members mimic each other or adopt similar predictive models (i.e., when they herd). We provide two pieces of evidence that herding has undermined the collective wisdom of the Blue Chip forecasting group over the last two decades: (1) a dramatic decline in forecaster disagreement about future GDP growth accompanied by a significant deterioration in the absolute and relative accuracy of the long-term average ("consensus") forecast for GDP growth, (2) identification of exogenous forces that likely decreased the incentive to anti-herd and facilitated model sharing.

JEL Classifications:
E37 Forecasting and Simulation: Models and Applications
D70 Analysis of Collective Decision-Making, General
D83 Search; Learning; Information and Knowledge; Communication; Belief
Keywords: herding, collective wisdom, forecast errors, forecast dispersion

1. Introduction

The disappointing track record of macroeconomic forecasters is well documented (Zarnowitz 1992), Loungani 2001, Juhn and Loungani 2002, Fildes and Stekler 2002, Loungani, et al. 2013). For example, Ahir and Loungani (2014) show that none of the 62 national recessions in 2008 and 2009 were predicted by the *Consensus Forecasts* as the previous year drew to a close. Others have shown that forecasts have little predictive power beyond horizons of 18 months (Isiklar and Lahirir 2007, and Lahiri 2011).

There are two broad explanations for this poor performance. One is that recessions are the result of infrequent and unique shocks (e.g., political crises) which are difficult or impossible to predict. This problem is compounded by the fact that the macroeconomic system is very complex and there are few variables that can serve as reliable indicators for future economic growth (Stock and Watson 2003a).

The second explanation is that herding (i.e., mimicking the behavior of others' or adopting their predictive models) undermines the collective wisdom of forecasting groups. In their seminal work, Bates and Granger (1969) demonstrated that accuracy could be improved by combining forecasts and this finding has led policymakers and others to rely on cross-sectional averages of individual forecasts.[1] However, the superior accuracy of these "consensus" forecasts only results when forecasters remain independent and use diverse predictive models (Batchelor and Dua 1995, Armstrong 2001, Timmermann 2006, Page 2007, Hong and Page 2011, Hong, et al. 2012).[2] When forecasters herd, collective wisdom is undermined.

There are several reasons why people might herd. In information cascade models, they ignore their private information and mimic the publicly revealed judgements of others when the information inferred from these judgments overwhelms their own information (Banerjee 1992, and Bikhchandani, et al. 1992, Easley and Kleinberg 2010, and Eyster and Rabin 2010). This behavior is rational from the individual's perspective, but collectively irrational or inefficient because it makes it more likely the group will zero in on the wrong target. In other models, agents herd because they want to influence the market's perception of their ability (Scharfstein and Stein 1990, Truman 1994, Prendergast and Stole 1996, and Graham 1999). The basic idea in Scharfstein and Stein is that "conformity with other investment professionals preserves the fog—that is, the uncertainty regarding the ability of the manager to manage the portfolio" (Bikhchandani and Sharma 2001). Golub and Jackson (2010) study social networks and show that the

[1] Surowiecki (2005) provides an insightful discussion of these issues in his popular book, *The Wisdom of Crowds*.

[2] Bates and Granger (1969) also emphasized the critical importance of model diversity and state that it existed when "(i) One forecast is based on variables or information that the other forecast has not considered. (ii) The forecast makes a different assumption about the form of the relationship between the variables."

wisdom of the crowd declines when "opinion leaders" emerge because their information is over-weighted by others and the idiosyncratic errors of the opinion leaders lead everyone astray.[3]

While some incentive structures cause people to herd, others induce anti-herding where they place more weight on their own information and deliberately deviate from the consensus. For example, participation in winner-take-all prediction contests or strong preferences for publicity can lead to anti-herding (Laster, et al. 1999, and Ottaviani and Sørensen 2006).[4] When agents over-weight private information, the public knowledge bias contained in forecast averages decrease and collective wisdom increases (Crowe 2010, and Lichtendahl, et al. 2013).[5]

In all of these models, information could be aggregated efficiently if an outside entity was able to observe the private information which is dispersed across forecasters. This is not possible in the market for professional forecasts where predictions, rather than the information they were based on, are reported publicly.

A number of researchers have speculated that herding undermines the collective wisdom of forecasting groups. For instance, Zarnowitz (1992) suggests that large forecast errors around business cycle turning points "cannot be explained away by a general reference to random disturbances" and occur, in part, because "few forecasters take the risk of signaling a recession prematurely ahead of others." Gallo, Granger, and Jeon (2002) show that members of the Consensus Economics survey herd and conclude that "the forecasting performance of these groups may be severely affected by the detected imitation behavior and lead to convergence to a value that is not the 'right' target." There is also evidence that model diversity affects collective wisdom. In particular, Batchelor and Dua (1995) show that the accuracy of forecast averages constructed with individual forecasts from the Blue Chip survey increases

[3] Experimental work by Lorenz, et al. (2011) shows that subtle changes in the structure of the social environment undermine collective wisdom. Compared to a control condition where subjects provide judgments about factual questions in social isolation, treatments where they are provided with information about the responses of others display much less diversity of opinion and larger collective errors.

[4] Because payoffs (prize money or publicity) decline sharply when forecasters share the spot light with others in winner-take-all contests, they have an incentive to decrease the weight they place on public information used by others. That is, they deliberately move away from the mean forecast when "the first-order reduction in the expected number of winners with who the prize must be shared more than compensates for the second-order reduction in the probability of winning" (Marinovic, et al. 2013).

[5] Kim, et al. (2001) first discussed the concept of the public knowledge bias. To see how it works, consider a case in which a group of agents generate independent (non-strategic) forecasts using Bayes' rule by taking a weighted average of their private signal and a common prior (public knowledge). Efficient aggregation involves pooling information in the disparate private signals with the common prior, but forecast averages over-weight the latter because it is an input in each of the individual forecasts. Crowe (2010) suggests the public knowledge bias could be reduced if we were to "provide incentives for strategic forecasting such that forecasters attempt to differentiate their forecasts from those of others" (p. 33).

when the component forecasts are based on a more diverse set of modeling assumptions (Keynesian, monetarist, supply-side, etc.) and forecasting techniques (judgment, structural econometric models, etc.).

This paper examines whether increased herding has caused the collective wisdom of the Blue Chip group, measured by the accuracy of the Blue Chip consensus, to decline over time. Professional forecasters compete in markets for their services and this motivates them to exploit new information and differentiate their forecasts (Batchelor and Dua 1990, Batchelor 2007, and Page 2007). However, they also operate in social and economic networks which make herding possible and changes in ideology, technology and institutions can alter the incentive to herd over time. We argue that three particular changes may have been important over the last few decades.

First, advances in computer technology have elevated the role of formal models in forecasting and made it easier for models to be shared. As Zarnowitz (1992) pointed out long ago:

> A free market exists in economic data and ideas, and advances in forecasting technology are soon open to all practitioners. This openness tends to reduce the diversity of individual predictions created by the undoubted fact that forecasters differ greatly in theoretical orientation and training and talent and experience. (p. 133)

Human judgment, which is partly innate but also improved through education and experience, is an important input into the forecasting process. Another is the computer which allows experts to exploit statistical knowledge to uncover complex relationships in data and encode them in predictive algorithms. While it is difficult to transfer human judgment between people, this is not true for computer models. As the reliance on computers in the forecasting industry has increased, it is plausible that the diversity of predictive models has decreased.

Second, it is generally accepted that ideological agreement within macroeconomics increased prior to the Great Recession. For example, Michael Woodford (2009, p. 267) concludes that:

> While macroeconomics is often thought of as a deeply divided field, with less of a shared core and correspondingly less cumulative progress than other areas of economics, in fact, there are fewer fundamental disagreements among macroeconomists now than in past decades. This is due to important progress in resolving seemingly intractable debates.

If professional forecasters' views were also converging, this could have reduced their collective wisdom.

Finally, there is evidence that the return to anti-herding has decreased. Beginning in 1981, an award – now called the Lawrence R. Klein Award – was given annually to a member of the Blue Chip group who made the most accurate forecasts over the previous four years, with winners receiving a $5,000 cash prize and significant media attention. However, the monetary component was never increased (it would have been $13,000 in 2015 had it kept up with inflation), was declined by winners starting in 2009, and has recently been discontinued. Moreover, media attention given to award winners decreased starting in the 1990s. Ceteris paribus, decreases in the returns to this winner-take-all competition should have reduced the incentive to anti-herd and the collective wisdom of the Blue Chip group.

To explore these issues, we examine the relationship between (i) the disagreement about future GDP growth among members of the Blue Chip group and (ii) the accuracy of the Blue Chip consensus for GDP growth. According to the Prediction Diversity Theorem, which we discuss in the next section, herding cause the predictive diversity and collective wisdom of a group to decrease. This implies that the size of consensus errors and the dispersion of forecasts across the group's members will be negatively correlated if changes in dispersion largely reflect variation in herding propensities.

In contrast, models that ignore social interactions between forecaster and focus on the impact of uncertainty predict a positive (or non-negative) relationship between forecast dispersion and the size of consensus errors. For instance, the sticky information model of Mankiw and Reis (2002) and Mankiw, et al. (2003) assumes that forecasters face a fixed cost to acquire and process new information, making it optimal to update infrequently. Following a shock, some fail to update and this increases forecast dispersion and the persistence and size of forecast errors.[6] Lahiri and Sheng (2010) show that forecast dispersion and volatility of the economic environment are positively related when private and public signal variances are positively correlated.[7] These models predict that the Great Moderation, which began in the mid-1980s, should have decreased both forecast errors and disagreement about future GDP growth.[8] Other work in this area shows that forecast dispersion and consensus errors both decline as forecast horizons shrink and information accumulates (Lahiri and Sheng 2008 and 2010, and Patton and Timmermann 2010).[9]

We use forecasts for real GDP growth published in the Blue Chip Economic Indicators newsletter to examine the covariation between consensus errors and forecast dispersion across 24 forecast horizons between 1976 and 2011 and provide three important findings. First, there was a large secular decline in the dispersion of GDP growth forecasts across all 24 forecast horizons between 1977 and 2011. For

---

[6] Similarly, Coibion and Gorodnichenko (2012) consider a model where forecasters continuously update, but observe noisy signals about the true state of the economy. This model also predicts a positive relationship between forecast dispersion and error size when agents have heterogeneous noise-to-signal ratios.

[7] In their seminal paper, Zarnowitz and Lambros (1987) use different terminology but discuss the same competing sources of forecast dispersion considered here: conformity and uncertainty (see their Figure 1) and provide evidence that dispersion and uncertainty (and thus macroeconomic volatility) are positively correlated. This has spawned a large literature (Bomberger 1996, Rich, et al. 1992, Giordani and Söderlind 2003, Lahiri and Liu 2006, Rich and Tracy 2010, Dovern et al. 2012, Boero, et al. 2014).

[8] The Great Moderation has been documented by a number of scholars including Kim and Nelson (1999), McConnell, et al. (2000), Stock and Watson (2003b) and others. More recently, Gamber, et al. (2010) show that the standard deviation of annualized real GDP growth rates, measured at quarterly frequencies, fell from nearly 5% between 1947 and 1983 to 2.2% from 1984 to 2008.

[9] The decline in disagreement observed in all of these cases is not the result of herding as described earlier where forecasters deliberately mimic one another or adopt similar models. Rather, it would be the consequence of "spurious" or unintentional herding "where groups facing similar decision problems and information sets take similar decisions" (Bikhchandani and Sharma 2001).

example, the cross-sectional variance of forecasts fell by more than 70 percent over the sample. The decline in dispersion is consistent with the hypothesis that uncertainty fell over the sample, herding increased, or some combination of these two possibilities.

Second, we show that the absolute and relative accuracy of long-term (horizons of 18 to 24-months) Blue Chip consensus forecasts decreased over the sample. This finding, when combined with the evidence of decreased forecast dispersion, supports the hypothesis that herding diminished the collective wisdom of the Blue Chip group. This conclusion is supported by regression analysis which shows that there is a negative and statistically significant relationship between absolute consensus errors and forecast dispersion at horizons of 18 through 24 months.

Finally, there is evidence that forecast dispersion declined in responses to exogenous changes in the economic environment. These include ideological convergence in the field of macroeconomics, a reduction in the cost of using and sharing computer models, and a significant decline in the value of the Lawrence Klein forecasting award. The experience of one forecasting firm, Lawrence H. Meyers & Associates, illustrates how independence may have been compromised by technological changes and the Klein Award. By the early 1990s when forecast dispersion began its sharp secular decline, this firm was openly sharing the PC version of its model and encouraging others to use the Blue Chip consensus as a starting point for developing their own forecasts. A few years later, Lawrence H. Meyers & Associates won the Klein award twice in a three-year period, an event which would have increased the incentive for others to mimic this firm or use their model. While this evidence is largely anecdotal, it points to exogenous forces that promoted herding.

Our results relate to previous findings using Blue Chip forecasts. Batchelor and Dua (1990) show that individual members of the Blue Chip group persistently issued forecasts above or below the consensus and attribute this to product differentiation motives. Batchelor and Dua (1992) find that forecasters place relatively little weight on the lagged consensus and large weights on their own past forecasts when updating their forecasts. They conclude that members of the Blue Chip group are variety-seeking rather than consensus-seeking. Laster, et al., (1999) find that independent forecasters (those not employed by banks, manufacturing firms, etc.) issue forecasts further from the consensus and argue that they are trading off accuracy for publicity. All of these findings are consistent with the hypothesis that members of the Blue Chip group were anti-herding. However, it is important to keep in mind that they are based on forecasts with relatively short-term (12-month) horizons and on data from the 1970s through early 1990s, and it is possible that incentives to herd changed over time for the reasons we discussed above. In fact, Lahiri and Sheng (2008) are unable to replicate the findings of Laster, et al. (1999) using more recent data.

The paper is organized as follows. The next section presents the Prediction Diversity Theorem which is used to illustrate the disparate impact that intentional and unintentional herding have on collective wisdom. Section 3 discusses the data employed in the study and the next three sections present the findings. Section 7 discusses developments in the economy and forecasting industry which could explain changes in herding and Section 8 concludes the paper.

2. A Model of Collective Wisdom

This section presents a simple model to illustrate how changes in herding and uncertainty have very different effects on the relationship between the collective wisdom of a forecasting group and the level of disagreement its members exhibit.[10] We show that, all else held equal, changes in herding propensities produce a negative relationship between the magnitude of consensus errors and dispersion of forecasts. In contrast, changes in uncertainty produce a positive relationship between consensus error size and forecast dispersion. This is the basic theoretical insight we exploit throughout the paper.

Assume that a group of n agents is attempting to predict the future value of a variable such as GDP growth. Each agent *i* receives a noisy signal s_i which is distributed conditional on the "truth" – the actual value of the variable given by *A*. That is, the signal distribution is given by $f_i(s_i|A)$. Agent i's squared prediction error is,

$$(1) \qquad SqE(s_i) = (s_i - A)^2$$

Two factors influence the variance of $f_i(s_i|A)$ and the size of prediction errors. The first is the agent's ability, which is determined by innate talent, formal training, experience, and the forecasting technology (e.g., computers and software). The second is the volatility of the forecasting environment. Even high-ability agents will make large errors when the shocks hitting the system are more variable.

Next, we define four variables that measure different characteristics of the forecasting group. The average squared individual prediction error is,

$$(2) \qquad SqE(\vec{s}) = \frac{1}{n}\sum_{i=1}^{n}(s_i - A)^2$$

where $\vec{s} = (s_1, s_2, \ldots s_n)$ is the vector of signals received by the n agents. The consensus is the equal-weighted mean of individual predictions,

[10] The following is largely based on the model presented in Hong and Page (2012).

(3) $$c = \frac{1}{n}\sum_{i=1}^{n} s_i$$

and the squared consensus error is,

(4) $$SqE(c) = (c - A)^2$$

Finally, predictive diversity is the variance of the agents' predictions around the consensus,

(5) $$PDiv(\vec{s}) = \frac{1}{n}\sum_{i=1}^{n} (s_i - c)^2$$

Predictive diversity is synonymous with forecaster disagreement and forecast dispersion.

The *Prediction Diversity Theorem* (PDT), a mathematical identity, states that the squared consensus error is equal to the average individual squared error minus predictive diversity:[11]

(6) $$SqE(c) = SqE(\vec{s}) - PDiv(\vec{s})$$

The PDT illustrates three important results. First, it shows why a crowd is wiser than many or most of its members. Note that the squared consensus error, a measure of the crowd's wisdom, is smaller than the average individual squared error when predictive diversity is positive. The basic idea is that signal noise cancels out when forecasts are averaged.[12] Second, predictive diversity and ability are *equally* important for collective wisdom. This is seen in (6) where the coefficients in front of both terms on the right-side are equal to one. Adding an expert to the group who increases predictive diversity without lowering its average ability, increases the wisdom of the crowd. Finally, the PDT is a general result which allows for a wide range of possibilities. The majority of theoretical models in the literature which analyze forecaster

[11] For a proof, see Hong and Page (2012). Engle (1983) and Bomberger (1996) provide earlier derivations in the context of macroeconomic forecasting.

[12] Alternatively, we can think of individuals using models that are misspecified in different ways. In this case, averaging forecasts lowers squared consensus errors because each model captures an important element of the underlying system (Hendry and Clements 2004 and Hong and Page 2012). Others emphasize that forecasters have different loss functions or use models that differ in their degree of adaptability to structural breaks and these sources of heterogeneity increase the potential for accuracy to be improved by combining forecasts (Timmerman 2006). Page (2014) argues that this interpretive framework, where agents are viewed as using different models rather than observing noisy signals of the "truth," is more realistic and intuitive because it relates predictive diversity to the different models employed by people rather than noise. We agree but present the signal-based model because it produces the same results and is familiar to more readers.

disagreement or forecast combinations assume signal independence, which implies that $PDiv(\vec{s}) > 0$ and $SqE(c) < SqE(\vec{s})$. Compared to this benchmark case, herding induces positive signal correlation which decreases predictive diversity and increases consensus errors.

To gain further insight, we make two changes to the model. First, we allow for bias where agent *i's* bias is $b_i = \mu_i - A$ and μ_i is the mean of *i's* signal distribution. Second, we focus on expected squared errors. In this case, the variance of agent *i's* signal is $v_i = E[(s_i - \mu_i)^2]$ where is $E[\cdot]$ is the expectation over all possible signal realizations given the outcome.

We define three new group-level variables. The first is the average individual bias,

$$(7) \qquad \bar{b} = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - A)$$

The second is the average individual variance,

$$(8) \qquad \bar{V} = \frac{1}{n}\sum_{i=1}^{n}E[s_i - \mu_i]^2$$

The third is the average covariance of individual signals,

$$(9) \qquad \overline{COV} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}E[s_i - \mu_i][s_j - \mu_j]$$

Using these terms, the expected squared consensus error is[13]

$$(10) \qquad E[SqE(c)] = \bar{b} + \frac{1}{n}\bar{V} + \frac{n-1}{n}\overline{COV}$$

This identity, referred to as the bias-variance-covariance decomposition (BVCD), illustrates three important results.

First, the expected squared consensus error is increasing in average bias $\bar{b}$. The logic is similar to that for noisy signals: individual biases cancel out when predictions are averaged as long as biases are both positive and negative. In contrast, collective wisdom declines when individuals have similar biases. This could occur if an "opinion leader" emerged and his or her idiosyncratic bias spread to others.

Second, the expected squared consensus error is increasing in the average signal variance $\bar{V}$. As forecaster ability increases or the economic environment becomes less uncertain, consensus accuracy increases ceteris paribus. However, note that the coefficient on $\bar{V}$ is small and converges to zero as *n*

[13] See Hong and Page (2012) for a proof.

becomes large. The minor impact of $\bar{V}$ reflects the fact that uncertainty (i.e., noise variance) has two competing effects on consensus accuracy. On the one hand, less uncertainty means that individual predictions are more accurate and this increases collective wisdom. On the other, less uncertainty causes predictive diversity to decrease as individual predictions cluster closer to the truth and with less noise the benefits from averaging predictions – canceling out noise – are reduced. The key point is that the impact of $\bar{V}$ is positive and we should observe a positive relationship between forecast dispersion and the size of consensus errors if exogenous changes in uncertainty are the primary factor driving both series. That is, consensus errors and forecaster disagreement should both decrease when forecast horizons shrink and information accumulates and during periods like the Great Moderation where the macroeconomic environment is less volatile.

Finally, the expected squared consensus error is increasing in the average signal covariance $\overline{COV}$. Note that the coefficient on this variable, (*n-1)/n*, is much larger than the coefficient in front of $\bar{V}$. This shows the powerful impact that the signal covariance has on collective wisdom. If the covariance is zero, the expected squared consensus error is simply equal to the average bias plus a fraction *1/n* of the average signal variance. In contrast, herding causes the signal covariance to be positive and increases consensus errors ceteris paribus. The key prediction is that the relationship between the size of consensus errors and forecast dispersion will be negative if there are large changes in herding propensities in the sample.

3. The Data Set

To examine the relationship between forecast dispersion and consensus accuracy, we use real GDP growth forecasts from the monthly *Blue Chip Economic Indicators* (BCEI) newsletter. Since 1976, the newsletter has surveyed a panel of approximately 50 professional forecasters and published their forecasts for different macroeconomic variables, as well as the cross-sectional average of individual forecasts—the consensus. The newsletter's founder, Robert Eggert, was a pioneer in the use of combination methods to improve forecast accuracy and the panelists are drawn from different sectors of the economy to promote predictive diversity.[14]

We focus on fixed-target forecasts for year-over-year real GDP growth made between August 1976 and December 2011. The panelists make their first forecast of GDP growth for year t (the target year) in January of year t-1 and revise it each month until a final forecast is made in December of year t. Thus there are 24 forecasts for each target year with the forecast horizon (h) shrinking from 24 months to one

[14] See Clemen (1989) for a discussion of Eggert's contributions.

month over a two-year period. In other words, panelists report two forecasts each month: a *year-ahead* forecast with $h = 24, 23, \ldots 13$ and a *current-year* forecast with $h = 12, 11, \ldots 1$.[15]

There are several features of the Blue Chip survey that make it a fertile testing ground for examining the impact of herding on collective wisdom. First, the newsletter publishes forecasts one month after they are made and identifies the forecaster (e.g., Ford Motor Company, Lawrence H. Meyer & Associates, etc.). This makes it possible for panelists to mimic the consensus or, due to the lack of anonymity, particular opinion leaders. Second, the forecasts are for fixed targets so the information they contain has a long half-life and remains valuable for many months. This contrasts with fixed-horizon forecasts where the target continually changes (e.g., inflation the next month) and the value of information contained in them decays rapidly. Third, there is turnover in the Blue Chip group. This is important because it has been shown that less experienced forecasters are more likely to cluster (see Lamont 2002).

Another important feature of the data is the presence of the Lawrence R. Klein Award given each year since 1981 (except 2001) to the most accurate member of the Blue Chip group.[16] Until recently, the winner received a $5,000 cash prize and considerable publicity. In theory, this created a winner-take-all competition which promoted anti-herding and collective wisdom. However, as we discuss in Section 7, the monetary prize did not keep up with inflation and was eventually jettisoned, while the publicity generated by the award has diminished. We posit that these changes led to an exogenous decline in the return to anti-herding, which should have reduced forecast dispersion and consensus accuracy.

This paper explores the possibility that the incentive to herd changed over time, but it is likely that it also varied over the forecast horizons. Two influences are at work here. First, forecasters have less information at longer horizons and this uncertainty increases the incentive to exploit the information contained in the consensus or the forecasts of opinion leaders. In addition, incorrect information cascades, where agents collectively zero in on the wrong target, are more likely when information is imprecise. Second, it is likely that the incentive to anti-herd varied across the horizons. As postulated

---

[15] The newsletter began reporting a full set of 12 current-year and 12 year-ahead forecasts each year in 1984. Prior to 1984, the panelists made current-year forecasts between January and June and year-ahead forecasts between July and December.

[16] The name of the award has changed over the years due to changes in sponsorship. Originally, it was called the Silbert Economics Forecasting Award, named after Theodore E. Silbert who was the chairman of the Sterling National Bank and Trust Company. At some point this sponsorship ended and it was renamed the Blue Chip Economic Forecasting Award. In 2003 the WP Carey Foundation agreed to sponsor the award and asked it be named for Lawrence R. Klein who was a board member. For sake of clarity, we refer to it as the Lawrence Klein Award throughout the paper. Despite the name changes, the method for measuring forecast accuracy and determining the winner has remained the same. Accuracy is measured by taking the absolute difference between the forecasts for four different variables (real GDP growth, CPI inflation rate, T-bill rate and unemployment rate) issued in January of the target year (h = 12) and the actual values of the variables being forecasted. An equally-weighted average of the absolute errors is taken for the four variables over the previous four years.

earlier, the Klein Award creates a winner-take-all competition, which motivates forecasters to deviate from the consensus. However, the award is based solely on 12-month forecasts made in January and performance at longer horizons provides forecasters with little opportunity to stand out from the pack. Thus there should have been less incentive to anti-herd at longer horizons.

To evaluate the accuracy of the Blue Chip consensus, we need to measure actual real GDP growth (GNP before 1992) and select between available vintages. One option is to use the latest available data, but as Croushore (2006) argues it is not reasonable to expect that forecasters anticipated all the revisions and benchmark changes made to this series in the years since they made their forecasts. Also, Croushore (2011) shows that evaluations of survey forecasts are sensitive to choices researchers make about data vintages. Given this, we use two different real-time vintages to measure forecast errors. The first, which we refer to as *vintage 1*, is the year-over-year real GDP growth rate reported six months after the end of the target year in the June issue of the *Survey of Current Business*. The second, *vintage 2*, is the growth rate reported 18 months after the end of the target year in the *Survey of Current Business*.

4. Temporal Variation in Forecast Dispersion and Consensus Accuracy

We begin by examining whether there have been secular changes in the accuracy of the consensus and level of disagreement among members of the Blue Chip group over the 1977 to 2011 sample period. As we discussed above, there are good reasons to believe that herding propensities and uncertainty vary over the forecast horizons. Thus we examine each horizon separately and do not pool our data. Given this, and the fact that our sample period is relatively short (35 years), it is difficult to identify the dates of structural break with much precision. As a consequence, we simply divide the sample in half and examine changes between the 1977-1993 and 1994-2011 subsamples.

4.A. Forecast Dispersion

Figures 1 and 2 illustrate the Blue Chip consensus for real GDP growth, the range of forecasts across the Blue Chip group and actual real GDP growth (vintage 1) between 1977 and 2011. Figure 1 shows the 18-month forecasts made in July of the year preceding the target year. This is the longest horizon for which forecast data is available over the entire sample. Figure 2 shows 9-month forecasts made in April of the target year.

The figures show a large decline in forecast dispersion measured by the difference between the most optimistic and most pessimistic GDP growth forecast. Prior to 1994, the range of 18-month forecasts exceeded five percent in seven of 17 years. In contrast, a range this large is observed only once in the later subsample (in 2000) and the smallest 18-month range (1.3 percent) is observed during the Great Recession in 2008. A similar pattern is observed for the 9-month forecasts. Ranges over 4 percent were

common prior to the mid-1990s, but rare in the second half of the sample. The 9-month range hit its lowest value (around 1 percent) in the three years preceding the Great Recession (2005, 2006 and 2007).

To test for structural changes in forecast dispersion over the sample, we regress it on a constant and dummy variable, 94BREAK, which takes on values of zero from 1977 to 1993 and one from 1994 to 2011. The constant term provides an estimate of the mean level of dispersion for the 1977-93 subsample and the coefficient on 94BREAK shows how dispersion changed between the subsamples. We only consider the 11 horizons where forecasts are available for the entire 1977-2011 period: h = 18, 17, … 8. To measure forecast dispersion, we use both the range of forecasts discussed above and cross-sectional variance of forecasts.

The results in Table 1 show that both measures of forecast dispersion decline as the horizon shortens. For example, the average range for the 18-month forecasts over the 1977-93 subsample was 4.44 percent compared to 2.8 percent for the 8-month forecasts. Even larger decreases are observed for the cross-sectional forecast variances: the average variance over the 1977-93 subsample was 0.83 at the 18-month horizon and 0.27 at the 8-month horizon. These horizon effects are consistent with models of forecaster disagreement that focus on the role of uncertainty (Lahiri and Sheng 2008 and 2010, and Patton and Timmermann 2010). That is, forecasts converge as the target date approaches because the panelists all use the same public information to revise their forecasts. This is unintentional or spurious herding.

More importantly, Table 1 provides strong evidence that forecast dispersion declined over time. The coefficient on 94BREAK indicates that post-1994 dispersion was significantly smaller than its pre-1994 counterpart and this is true for both measures of dispersion and all horizons considered. Moreover, the decreases are economically meaningful. For example, the variance of the 18-month forecasts fell from an average of 0.83 between 1977 and 1993 to an average of 0.25 between 1994 and 2011. This is a 70 percent decrease and, as we show below, the percentage decline in disagreement is more than 2.5 times greater than the percentage decline in GDP growth volatility across these two subsamples.

Figure 3 shows the cross-sectional variance of the 18- and 9-month forecasts over the full sample. Three observations warrant discussion. First, while disagreement appears to move countercyclically as many researchers have noted, the ebbs and flows are overshadowed by the secular decline. Second, the sharp decrease begins around 1990 and this timing suggests that our estimate of a 70 percent decline in forecast dispersion understates the clustering which has taken place. Third, the dispersion of 18-month forecasts declines much more than that for the 9-month forecasts.

Why did forecast dispersion decline over the sample? As we discussed earlier, there are two basic possibilities. The first is that uncertainty declined as the Great Moderation reduced the volatility of GDP growth. The second is that herding increased. While both explanations are plausible, they have very different implications for consensus accuracy: decreases in volatility and uncertainty should be associated

with smaller consensus errors, while increased herding with larger ones. The next section examines which of the two possibilities is consistent with the data.

4.B. Consensus Accuracy

Did the accuracy of the Blue Chip consensus decline over the sample? Examination of Figure 1 suggests that, at least for the longer-term forecasts, it may have. With the exception of 1982, the 18-month-ahead consensus is close to realized GDP growth throughout the late-1970s and 1980s. Starting in the mid-1990s, consensus errors appear to increase in size and become more persistent. For example, the consensus under-forecasted GDP growth by around 2 percentage points four years in a row from 1997 to 2000. In contrast, the group over-forecasted GDP growth by large magnitudes in 2001 and again during the Great Recession in 2008 and 2009.

One way to examine whether consensus accuracy changed over time is to simply compare forecast errors in different subsamples. To do this, we regress absolute consensus errors on a constant and 94BREAK, the structural break dummy discussed above. The constant term provides an estimate of the mean absolute error (MAE) for the consensus over the 1977-93 subsample, while the coefficient on 94BREAK shows how the MAE changed from the first subsample to the second.

Table 2 presents results for the structural break regressions. Two important findings are shown. First, the MAE for the Blue Chip consensus falls as the horizon shrinks. This is not surprising and indicates that the uncertainty decreased as the target date approached. Second, there is evidence that the absolute consensus errors increased over time. The MAE is larger in the second subsample for 18 of the 22 horizons considered and the changes are large for several of the long-term forecasts. For instance, the MAE increased by 20 percent or more at horizons of 12, 13, 15 16 and 17 months when we use vintage 1 data. When vintage 2 is employed, the MAE increases by over 40 percent for horizons between 15 and 18 months. While, these changes are not statistically significant as shown by the relatively low t-statistics for 94BREAK, it should be recognized that the temporal variation in forecast errors is large relative to their mean error and our sample size is small. In addition, the 1994 break date, chosen to split the sample in half, is somewhat arbitrary. If we had more data and were able to estimate precise break dates using Chow-tests, statistical significance would likely rise.

Another problem with this approach is that it does not control for changes in the volatility of GDP growth which influence the size of the consensus errors independent of changes in herding propensities. To assess the *relative* accuracy of the consensus we compare its performance to a naïve forecast based on GDP growth in year t-1. While the naïve forecast provides a low bar for the consensus to jump over, it allows us to control for changes in the volatility of the economic system.

The top row of 2 shows that the MAE for the naïve forecast falls by 27 percent during the second subsample for both data vintages. Although we cannot, once again, reject the null that the coefficient on

94BREAK is significantly different from zero at conventional levels, the decrease in the size of the naïve forecast errors are consistent with the hypothesis that the Great Moderation decreased GDP volatility over the sample. In addition, they put the 20-40 percent *increase* in long-term Consensus errors in a new light given that GDP growth volatility *fell* by 27 over the same period.

To explore this issue further, we use the approach developed by Diebold and Mariano (1995) and test the null hypothesis that there is no significant difference in the accuracy of the consensus relative to the naïve forecast.[17] Tables 3A and 3B report the MAE for the naïve forecasts and Blue Chip consensus, the Diebold-Mariano statistic S(1) used to test the null hypothesis of equal accuracy, and the probability level at which the null can be rejected. Results are provided for our two subsamples (1977-93 and 1994-2011) and vintage 1 (Table 3A) and vintage 2 (Table 3B) data.

Both Tables show that the Blue Chip consensus outperforms the naïve forecast by wide margins during the first subsample. This is true for all horizons and the differences in the MAE are large and highly significant. For example, using vintage 1 data (Table 3A) the MAE for the naïve model is 2.08 percent while that for the 18-month consensus is 1.13 percent. This is a 46 percent improvement in performance that is significant at the 0.013 level. In contrast, during the 1994-2011 subsample the MAE for the naïve forecast falls to 1.51 percent, while the MAE for the 18-month consensus rises to 1.27 percent. As a result, the consensus errors are only 16 percent lower than the naïve forecast errors in the later subsample and the difference is not significant. Similar differences across the two subsamples are observed when we use vintage 2 data (Table 3B). Overall, these findings provide evidence that the relative accuracy of the Blue Chip consensus declined during the second half of the sample.

Competing theories about the forces driving the magnitude of consensus errors and forecast dispersion imply different relationships between these two variables. If fluctuations in uncertainty are the dominant factor driving both, forecast dispersion and absolute consensus errors should move in the same direction. In contrast, we should observe a negative relationship between absolute consensus errors and forecast dispersion if variations in herding propensities are more powerful. For the long-term forecasts, the evidence presented in this section is consistent with this second explanation.

5. Bracketing and Relative Consensus Accuracy

5.A. Bracketing

[17] The null is tested under the assumption that the ratio of the sample mean loss differential (i.e., the differences in mean squared errors generated by the two forecasts) to the consistent estimate of the standard deviation of the sample mean loss differential has a limiting N(0,1) distribution. To calculate the long-run variance, we use the uniform kernel with the Schwert criterion used to section the maximum lag length order.

Another way to measure the relative accuracy of the consensus is to examine the extent to which forecasts "bracket the truth."[18] When individual forecasts are scattered on opposite sides of the actual value of the variable being forecasted, they bracket the truth and the consensus will always outperform the average member of the group. In contrast, when forecasts cluster on either side of the actual value they do not bracket the truth and the consensus cannot be more accurate than the average individual.[19] Forecasts are less likely to bracket the truth when herding increases and informational cascades form.

Figure 1 shows that the frequency with which the 18-month Blue Chip forecasts bracketed the truth diminished considerably over the sample. Prior to 1994, actual GDP growth fell outside the range between the most optimistic and pessimistic forecast on only two occasions (1982 and 1991). In contrast, the forecasts failed to bracket actual GDP growth 9 times in the second half of the sample (1994, 1997, 1998, 1999, 2000, 2001, 2008, 2009 and 2011). For example, the most optimistic forecaster underestimated GDP growth four years in a row between 1997 and 2000. Then the entire group overestimated growth in 2001 and during the Great Recession in 2008 and 2009.

Figure 2 shows that the 9-month forecasts also fail to bracket actual GDP growth at an increasing rate over the sample. Prior to 1994, actual GDP growth fell between the highest and lowest forecasts in every year except 1985. Starting in 1994, they failed to bracket GDP growth on five occasions (1995, 1997, 1998, 2003 and 2011). Although 9-month forecasts bracket the truth more frequently than their 18-month counterparts, both bracket the truth less over the sample. Overall, the decline in bracketing provides evidence that the ability of the consensus to improve on the average or typical forecaster has declined.

5.A. Relative Consensus Accuracy

[18] For a discussion of bracketing and its implications for the wisdom of the crowd, see Larrick and Soll (2006), Larrick, et al. (2011) and Mannes, et al. (2014).

[19] To illustrate, consider a simple example. Suppose Alice predicts GDP growth will be 3 percent and Bill predicts 8 percent. If actual GDP growth is 6 percent, the two forecasts bracket the truth and Alice's and Bill's absolute errors are 3 and 2 percent, respectively, and their average error is 2.5 percent. The consensus prediction in this case is 5.5 percent (0.5x3 + 0.5x8) and the absolute consensus error is 0.5 percent. Thus the consensus is more accurate than the average individual accuracy. Put differently, if we selected a forecaster based on a coin flip because we did not know which would be more accurate ex ante, the expected error (2.5 percent) would exceed the consensus error (0.5). Now consider what happens when Alice's prediction rises to 7 percent and Bill's remains at 8 percent. This might occur if Bill became an opinion leader. If actual growth continues to be 6 percent, the forecasts no longer bracket the truth. Alice's and Bill's absolute errors are 1 and 2 percent, respectively, and their average absolute error falls to 1.5 percent. In contrast, the absolute consensus error rises to 1.5 percent. In general, the consensus cannot outperform the average member of the group when predictions do not bracket the truth.

To further explore the relative accuracy of the consensus, we use the Prediction Diversity Theorem to compare the size of consensus and average individual errors. That is, relative accuracy is measured by dividing both sides of (6) by the average individual squared forecast error,

$$(11) \qquad \theta = \frac{SqE(c)}{SqE(\vec{s})} = 1 - \frac{PDiv(\vec{s})}{SqE(\vec{s})}$$

Low values of θ indicate high relative consensus accuracy and the ratio on the right-side shows the two factors that influence it. Decreases in the average individual squared error – which we expect to have occurred during the Great Moderation – should lead to an increase in relative consensus accuracy (a decline in θ) if not matched by a proportional decline in predictive diversity (forecast dispersion). On the other hand, decreases in predictive diversity relative to the average individual squared error, which would occur if herding intensified, increases θ and causes relative consensus accuracy to decline. When predictive diversity falls to zero, θ = 1 and the consensus is unable to outperform the average forecaster.

Figures 4 and 5 illustrate the three components used to construct our measure of relative consensus accuracy. The height of the blue bar illustrates the squared consensus error, the height of the red bar reflects predictive diversity, and the heights of the two bars combined is the average individual squared error. Relative consensus accuracy, θ, is represented by the height of the blue bar relative to the height of the blue and red bars combined. To test whether θ changed over the sample, we follow the approach employed above and regress it on a constant and the structural break dummy, 94BREAK.

The results are presented in Table 4 and show there is strong evidence that relative consensus accuracy fell over time. The coefficient on 94BREAK is positive in each of the 22 regressions and significantly different from zero at the 10 percent level or better in 17 of them. In many cases, particularly for the regressions that use vintage 2 GDP data, the coefficients on 94BREAK are large and highly significant. The size of the coefficients is economically meaningful as well. For example, estimates using the 18-month forecasts and vintage 1 data indicate that the squared consensus error was 46 percent of the average individual squared error between 1977 and 1993 and 71 percent from 1994 to 2011.

Overall, the findings show that large declines in forecast dispersion over time were not matched by proportionate decreases in average individual squared errors. As a consequence, the relative accuracy of the consensus fell. These findings are consistent with the hypothesis that increases in herding reduced the collective wisdom embodied in the Blue Chip consensus.

6. Contemporaneous Regression Analysis

This paper exploits the idea that uncertainty and herding have very different effects on the relationship between forecast dispersion and the size of consensus errors. Changes in the level of

uncertainty – due, among other factors, to variations in the volatility of the economic environment – produce a positive relationship between forecast dispersion and the size of consensus errors. In contrast, changes in herding propensities produce a negative relationship between these two variables. The previous sections of the paper examined these relationships by testing for changes in consensus errors and forecast dispersion over time. This section examines the contemporaneous relationship between them.

To do this, we regress absolute consensus errors on a constant and the cross-sectional variance of GDP forecasts. The coefficient on the cross-sectional variances in this simple regression reflects the competing effects of changes in uncertainty and herding: if variations in uncertainty are the dominate force driving forecast dispersion and absolute consensus errors the coefficient should be positive, and if variations in herding are more quantitatively important it should negative. Because there is good reason to believe that herding behavior and uncertainty depend on the length of time to the target date, we estimate a separate regression for each horizon. We use data from 1986 to 2011 because this is the longest sample that allows us to measure consensus errors and compare results across all 24 horizons. Thus each regression is estimated with 26 annual observations and the two data vintages.

The results are presented in Table 5 and show two important findings. First, as we saw earlier, the consensus error decreases as the target date approaches. For instance, the second column shows the constant term for regressions using vintage 1 data is 1.45 percent at the 24-month horizon, 0.72 percent at the 12-month horizon, and 0.15 percent at the 1-month horizon. Clearly, uncertainty diminishes as the target date approaches. Thus this result provides further evidence that changes in uncertainty are the key factor driving the relationship between the size of the consensus error and forecast dispersion over the forecast horizons.

The second important finding is that the sign of the relationship between forecast dispersion and absolute consensus errors is conditional on the length of the forecast horizon. For horizons of 14-months or shorter, the coefficients are positive in all but one case. However, in only five cases are they significantly different from zero (9 months for vintage 1 data and 11, 9, 7 and 6 months for vintage 2 data). Once again, the positive relationship between dispersion and absolute consensus errors is consistent with the hypothesis that changes in uncertainty are the primary force driving both variables for the short-run horizons.

In contrast, the relationship between forecast dispersion and absolute consensus errors is negative at longer horizons. For each horizon of 15 months or longer, the coefficient on the cross-sectional forecast variance is negative and in 9 of these 20 regressions the coefficient is significantly different from zero (horizons of 23 and 24 months for vintage 1 data and 18-24 months for vintage 2). This is an important finding and consistent with the hypothesis that variations in herding propensities were large and affected both the dispersion of forecasts and collective wisdom of the Blue Chip group. At the long-term

horizons, these results support the hypothesis that increases in herding caused forecast dispersion to fall and consensus errors to increase in the second half of the sample.

The estimates for the 23-month horizon are particularly noteworthy and provide additional evidence that herding plays a key role. For both data vintages, the coefficient on the cross-sectional forecast variance takes on its largest negative value at this horizon and we can reject the null that it is equal to zero at high levels of significance. In addition, the estimated intercept rises as we move from the 24- to 23-month horizon, suggesting that consensus errors increased as the group moved one month closer to the target date. These findings are consistent with the hypothesis that members of the Blue Chip group were able to observe the forecasts of others for the first time at the 23-month horizon and responded by adjusting their forecasts toward that of the consensus or forecasts or mimicking opinion leaders. This behavior would cause forecast dispersion to decline and consensus errors to increase at the 23-month horizon, thus producing the strong effect seen in Table 5.

Our findings are consistent with two important results in the literature. First, Lahiri and Sheng (2008) and Patton and Timmermann (2010) find that forecaster disagreement at longer horizons (e.g., 24-months) are largely driven by differences in the models used by forecasters (or their prior beliefs). Then, as the target date approaches and forecasters update using the same public information, disagreement declines. If model heterogeneity underlies differences in long-term forecasts, then the dispersion of long-term forecasts should be more susceptible to changes in model diversity than is the dispersion of short-term forecasts. Thus we should observe a negative relationship forecast dispersion and the size of consensus errors at long-term horizons, but not the short-term ones, if variations in model diversity are quantitatively important over the sample. This is what Table 5 shows.

Finally, a group of decision makers is more likely to form an incorrect information cascade – that is, to collectively zero in on the wrong target – when information is less precise.[20] Given that information is less precise at longer forecast horizons we expect incorrect cascades to occur more frequently here. Our finding of a negative relationship between forecast dispersion and absolute consensus errors is consistent with the hypothesis that incorrect information cascades have developed at the longer horizons.

7. Why did forecast dispersion decline?

We have shown that forecast dispersion decreased dramatically starting in the 1990s and was accompanied by a reduction in the absolute and relative accuracy of the Blue Chip consensus at long-term horizons. The Great Moderation could account for some of the decline in forecast dispersion, but a structural change like this should have also been associated with smaller forecast errors, not the larger

[20] See Figure 1 in Bikhchandani, et al. (1992).

ones we observed in the data. Moreover, we showed that GDP growth volatility fell by 27 percent over the second part of our sample, while forecast dispersion declined by over 70 percent. This implies that reduced uncertainty cannot fully explain the decline in forecaster disagreement and suggests that increased herding played a role. This section presents evidence that exogenous forces may have caused herding to increase over the sample.

To begin, we consider the history of the BCEI newsletter and the way it was marketed to the public. The first advertisement it placed in *Business Economics* in April 1987 quotes Victor Zarnowitz of the University of Chicago and NBER on the benefits of averaging forecasts and includes testimonials by prominent policymakers and thought leaders.[21] Another ad in March 1988 quoted economist Stephen McNees who had published a study of the Blue Chip consensus three months earlier:

> [consensus forecasting] is particularly appealing to those who reject the idea of one 'true' model and sympathize with the view that all models are flawed yet contain a grain of partial truth… 'consensus' forecasts are more accurate than most, sometimes virtually all, of the individual forecasters that constitute the consensus.[22]

While earlier ads touted the accuracy of the consensus from October in the year preceding the target year, by 1994 focus shifted to the 12-month forecasts made in January of the target year. The last ad in January 2001 did not emphasize the absolute accuracy of the consensus, but claimed it had "become the benchmark against which all other forecasts are measured." This shift to shorter horizons and decline in the emphasis on accuracy coincides with the diminished absolute and relative performance we showed earlier.

More importantly, this history suggests a possible explanation for why long-term consensus accuracy fell over time. If the newsletter was successful convincing the Blue Chip panelists that the consensus was more accurate than they were, each had an incentive to adjust his or her forecast toward the consensus during the revision process—behavior that would have reduced independence and collective wisdom.[23] Put differently, a "tragedy of the commons" could have emerged where the freedom of group members to ignore negative externalities created by their overuse of the common resource allowed the early predictive success of the consensus to undermine its future accuracy.

---

[21] These include Beryl Sprinkel, Chairman of the Council of Economic Advisors, Henry Wallich, former member of the Federal Reserve Board, and Louis Rukeyser of "Wall Street Week." Rudolf Penner, former Director of the Congressional Budget Office, stated that "the CBO begins its work by reading the Blue Chip Consensus Forecasts."

[22] McNees (1987). Subsequent empirical work also showed that the Blue Chip consensus consistently outperformed the individual members of the group (see Bauer, et al. 2003).

[23] Hirshleifer and Teoh (2008) discuss the possibility of multiple equilibria created by the interaction between the incentives to herd, independence and collective wisdom.

It is also possible that exogenous advances in computer technology reduced model diversity. Initially, the Blue Chip newsletter made an important distinction between "econometric modelers" (e.g., DRI, Wharton Econometrics, the Michigan Quarterly U.S. Model, etc.) and other forecasters who presumably relied on computers less. The emergence of low-cost computing starting in the 1980s blurred this distinction and made it possible for all forecasters to rely more on computer models and less on human judgment. To the extent that computer models can be shared with others more easily than human judgment, the emergence of low-cost computing could have reduced predictive diversity.

In fact, there is evidence that model sharing occurred in the Blue Chip group. Laurence H. Meyer & Associates, which won the Klein Award for forecasting accuracy in 1993 and 1995, provided clients with a PC version of their model starting in the mid-1980s. When developing their own forecasts, clients were encouraged to "begin from the LHM&A base forecast or from the Blue Chip consensus forecast on the PC" (*Business Economics*, July 1985). To the extent that members of the Blue Chip group engaged in this practice, predictive diversity would have decreased.

After LHM&A won the Klein Award in 1993 and 1995, it is plausible that they became an opinion leader, which could have further decreased the accuracy of the Blue Chip consensus. As Golub and Jackson (2010) show, the presence of "opinion leaders" in social and economic networks reduces collective wisdom because others place too much weight on the opinion leader's idiosyncratic information when judgments are revised. LHM&A used a Keynesian, demand-side model which was not well suited for producing accurate forecasts in an environment with large and unanticipated increases in productivity. There is strong agreement that supply-side shocks of this nature played an important role in the second half of the 1990s (see Schuh 2001 and Krane 2003) and overreliance on Keynesian models could, therefore, have produced the large under-predictions of real GDP growth seen in Figures 1 and 2.

Another possible explanation for the decline in predictive diversity is that the incentive to anti-herd fell over time. As we discussed in the Introduction, the real value of the monetary component of the Klein Award declined because it did not keep up with inflation and was eventually discontinued. In addition, there is evidence that the positive publicity associated with the award diminished over time. For example, articles about the award winner appear in the *New York Times* every year between 1981 (the first year it was given) and 1986 and again in 1988. The *Wall Street Journal* published its first article about the winner in 1987 (Saul Hymans, the first two-time winner), another in 1994 (Donald Ratajczak), and two articles in 1996 mention that Laurence H. Meyer, who had been appointed to the Federal Reserve board that year, won the Klein Award in 1993 and 1995. The final article in the *Wall Street Journal* to discussing a Klein Award winner (Martin Zimmerman of Ford) was published in 1997. If the monetary and publicity benefits associated with the award declined, the forecasters had less incentive to place a

large weight on their own information, thereby increasing the public information bias and lowering the collective wisdom embodied in the Blue Chip consensus.

Finally, a number of economists have argued that ideological conformity in macroeconomics has increased over the past few decades and reduced the collective ability of economists to anticipate the Great Recession. For example, Jean-Claude Trichet, former ECB President, concluded that

> The key lesson I would draw from our experience is the danger of relying on a single tool, methodology or paradigm. Policymakers need to have input from various theoretical perspectives and from a range of empirical approaches. Open debate and a diversity of views must be cultivated.[24]

Similarly, Wieland and Wolters (2012) and Wieland et al. (2012) discuss how reliance on a particular class of models (dynamic stochastic general equilibrium models) may have contributed to failures to predict the Great Recession and call for (i) greater openness to a variety of modeling paradigms and (ii) forecasting competitions.

8. Conclusion

Professional forecasters compete in a market for their services and face incentives to be accurate as well as different. However, they also operate in social and economic networks and, as is well known from both social psychology and economics, humans often mimic the actions and thoughts of others when they are uncertain. This herding can undermine collective wisdom and lead to large and persistent forecast errors.

This paper provides evidence that there has been a dramatic decline in the level of disagreement about future GDP growth among members of the Blue Chip forecasting group over the past few decades, which has been accompanied by a deterioration in the absolute and relative accuracy of the long-term (18-24 month horizons) Blue Chip consensus forecasts of GDP growth. We have also identified three exogenous changes in the economic environment that may have increased the propensity to herd: (i) greater ideological conformity within the field of macroeconomics, (ii) greater reliance on computer models which could be shared more easily than human judgment, and (iii) decreases in the incentive to anti-herd produced by a reduction in the value of the Lawrence Klein Forecasting Award. Finally, we argue that the decline in the accuracy of the Blue Chip consensus could have been an endogenous response to its earlier success which increased the incentive to herd. While the evidence is based on a relatively small sample and should thus be viewed with some caution, it suggests that increases in herding propensities have caused the collective wisdom of the Blue Chip group to decline.

[24] Quoted in Wieland and Wolter (2012).

Our findings help explain why policymakers and others did not anticipate the Great Recession. In addition, they raise the specter that economic agents mistakenly interpreted the high levels of agreement between macroeconomic forecasters as evidence that uncertainty had diminished. If true, this could help explain the origins of the excessive risk-taking and financial fragility which made such a large economic contraction possible. An important implication of our work is that stronger incentives to promote diverse perspectives among forecasters could improve social welfare.

References

Ahir, H., Loungani, P., 2014. There will be growth in the spring: How well do economists predict turning points? Voxeu.org (April 14).

Armstrong, J., 2001. Combining Forecasts. In: Armstrong, J. (Eds.). Principles of Forecasting: A Handbook for Researchers and Practitioners, Norwell, MA: Kluwer Academic Publishers, 1-1.

Banerjee, A., 1992. A Simple Model of Herding Behavior. Quarterly Journal of Economics 107, 797-817.

Batchelor, R., 2007. Bias in Macroeconomic Forecasts. International Journal of Forecasting 23, 189-203.

Batchelor, R., Dua, P., 1990a. Product Differentiation in the Economic Forecasting Industry. International Journal of Forecasting, 311-16.

Batchelor, R., Dua, P., 1990. Forecaster Ideology, Forecasting Technique, and the Accuracy of Economic Forecasts. International Journal of Forecasting 6, 3-10.

Batchelor, R., Dua, P., 1992. Conservatism and Consensus-Seeking Among Economic Forecasters. Journal of Forecasting 11, 169-81.

Batchelor, R., Dua, P., 1995. Forecaster Diversity and the Benefits of Combining Forecasts. Management Science 41, 68-75.

Bates, J., Granger, C., 1969. The Combination of Forecasts. Operational Research Quarterly 20, 451-468.

Bauer, A., Eisenbeis, R., Waggoner, D., Zha, T., 2003. Forecast Evaluation with Cross-Sectional Data: The Blue Chip Surveys. Federal Reserve Bank of Atlanta Economic Review, 17-31.

Bikhchandani, S., Hirshleifer, D., Welch, I., 1992. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. Journal of Political Economy 100, 992-1026.

Bikhchandani, S., Sharma, S., 2001. Herd Behavior in Financial Markets. IMF Staff Papers 47, 279-310.

Boero, G., Smith, J., Wallis, K., 2014. The Measurement and Characteristics of Professional Forecasters' Uncertainty. Journal of Applied Econometrics.

Bomberger, W., 1996. Disagreement as a Measure of Uncertainty. Journal of Money, Credit and Banking 38, 381-92.

Business Economics, various issues.

Clemen, R., 1989. Combining Forecasts: A Review and Annotated Bibliography. International Journal of Forecasting 5, 559-583.

Clement, M., Tse, S., 2005. Financial Analyst Characteristics and Herding Behavior in Forecasting. The Journal of Finance 60, 307-341.

Coibion, O., Gorodnichenko, Y., 2008. What can survey forecasts tell us about informational rigidities? Journal of Political Economy 120(1), 116-59.

Coibion, O., Gorodnichenko, Y., 2015. Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. American Economic Review 105(8), 2644-2678.

Croushore, D., 2006. Forecasting with real-time macroeconomic data. Handbook of Economic Forecasting 1, 961-982.

Croushore, D., 2010. An evaluation of inflation forecasts from surveys using real-time data. *The BE Journal of Macroeconomics 10*(1).

Crowe, C., 2010. Consensus Forecasts and Inefficient Information Aggregation. IMF Working Paper 178.

Diebold, F., 2015. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. Journal of Business and Economic Statistics, *33*(1), 1-1.

Diebold, F., Mariano, R., 1995. Comparing Predictive Accuracy. Journal of Business and Economic Statistics 13, 253-263.

Dovern, J., Fritsche, U., Slacalek, J., 2012. Disagreement among Forecasters in G7 Countries. The Review of Economics and Statistics 94(4), 1081-1096.

Easley, D., Kleinberg, J., 2010. Information cascades. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press.

Engle, R., 1983. Estimates of the Variance of U.S. Inflation Based Upon the ARCH Model. Journal of Money, Credit and Banking 15, 286-301.

Eyster, E., Rabin, M., 2010. Naïve Herding in Rich-Information Settings. American Economic Journal: Microeconomics 2, 221-243.

Fildes, R., Stekler, H., 2002. The state of macroeconomic forecasting. Journal of Macroeconomics *24*(4), 435-468.

Gallo, M., Granger, C., Jeon, Y., 2002. Copycats and Common Swings: The Impact of the Use of Forecasts in Information Sets. IMF Staff Papers 49, 4-21.

Gamber, E., Smith, J.K., Weiss, M.A., 2010. Forecast Errors Before and During the Great Moderation. Journal of Economics and Business 63, 278–289.

Giordani, P., Söderlind, P., 2003. Inflation Forecast Uncertainty. European Economic Review 47, 1037-1059.

Golub, B., Jackson, M., 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics 2*(1), 112-149.

Graham, J.R., 1999. Herding Among Investment Newsletters: Theory and Evidence. The Journal of Finance 54, 237–268.

Hendry, D., Clements, M., 2004. Pooling of Forecasts. Econometrics Journal 7, 1-31.

Hirshleifer, D.A., Teoh, S.H., 2009. Thought and Behavior Contagion in Capital Markets. In: Hens, T., Schenk-Hoppe, K.R. (Eds.). Handbook of Financial Markets: Dynamics and Evolution, Amsterdam: North Holland, 1-46.

Hong, L., Page, S., 2012. Some Microfoundations of Collective Wisdom. In: Landemore, H., Elster, J. (Eds.). Collective Wisdom: Principles and Mechanisms, Cambridge: Cambridge University Press, 56-71.

Hong, L., Page, S., Riolo, M., 2012. Incentives, Information, and Emergent Collective Accuracy. Managerial and Decision Economics 33, 323-334.

Isiklar, G., Larhiri, K., 2007. How far ahead can we forecast? Evidence from cross-country surveys. International Journal of Forecasting 23, 167-187.

Juhn, G., Loungani, P., 2002. Further cross-country evidence on the accuracy of the private sector's output forecasts. *IMF staff papers 49*(1), 49-64.

Kim, C.J., Nelson, C.R., 1999. Has the U.S. Economy Become More Stable? A Bayesian Approach Base on a Markov-Switching Model of the Business Cycle. Review of Economics and Statistics 81, 1-10.

Kim, O., Lim, S.C., Shaw, K.W., 2001. The Inefficiency of Mean Analyst Forecast as a Summary of Forecast of Earnings. Journal of Accounting Research 39, 329-335.

Krane, S., 2003. An Evaluation of real GDP Forecasts: 1996-2001. Federal Reserve Bank of Chicago Economic Perspectives, 2-19.

Krane, S., 2011. Professional Forecasters' Views of Permanent and Transitory Shocks to GDP. American Economic Journal: Macroeconomics *3*(1), 184-211.

Krishnan, M., Lim, S.C., Zhou, P., 2006. Analysts' Herding Propensity: Theory and Evidence from Earnings Forecasts. Working Paper.

Lahiri, K., 2011. Limits to Economic Forecasting. In: Higgins, M.L. (Eds.). Advances in Economic Forecasting, Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 25-50.

Lahiri, K., Liu, F., 2006. Modeling Multi-Period Inflation Uncertainty Using a Panel of Density Forecasts. Journal of Applied Econometrics 21, 1199-1219.

Lahiri, K., Sheng, X., 2008. Evolution of Forecast Disagreement in a Bayesian Learning Model. Journal of Econometrics 144, 325-340.

Lahiri, K., Sheng, X., 2010. Learning and Heterogeneity in GDP and Inflation Forecasts. International Journal of Forecasting 26, 265-292.

Lamont, O., 2002. Macroeconomic Forecasts and Microeconomic Forecasters. Journal of Economic Behavior & Organization 48, 265-280.

Larrick, R., Soll, J., 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. Management science *52*(1), 111-127.

Larrick, R., Mannes, A., Soll, J. (2011). The social psychology of the wisdom of crowds. In Krueger, J. I. (Ed.), Frontiers in social psychology: Social judgment and decision making. New York: Psychology Press.

Laster, D., Bennett, P., Geoum, I.S., 1997. Rational Bias in Macroeconomic Forecasts. The Federal Reserve Bank of New York Staff Reports 21.

Laster, D., Bennett, P., Geoum, I.S., 1999. Rational Bias in Macroeconomic Forecasts. The Quarterly Journal of Economics 114, 293-318.

Lichtendahl, K.C., Grushka-Cockayne, Y., Pfeifer, P.E., 2013. The Wisdom of Competitive Crowds. Operations Research 61, 1383-1398.

Lichtendahl, K.C., Winkler, R.L., 2007. Probability Elicitation, Scoring Rules, and Competition Among Forecasters. Management Science 53, 1745-1755.

Lorenz, J., Rauhut, K., Schweitzer, F., Helbing, D., 2011. How Social Influence Can Undermine the Wisdom of Crowd Effect. Proceedings of the National Academy of Sciences 108, 9020-25.

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.

Mankiw, N., Reis, R., 2002. Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve. Quarterly Journal of Economics 117, 1295-1328.

Mankiw, N., Reis, R., Wolfers, J., 2004. Disagreement about inflation expectations. In NBER Macroeconomics Annual 2003, Volume 18 (pp. 209-270). The MIT Press.

Mannes, A., Soll, J., Larrick, R., 2014. The wisdom of select crowds. Journal of personality and social psychology *107*(2), 276.

Marinovic, I., Ottaviani, M., Sørensen, P.N., 2013. Forecasters' Objectives and Strategies. In: Elliot, G., Timmermann, A. (Eds.). Handbook of Economic Forecasting, Vol. 2, Part B, Elsevier, 690-720.

McConnell, M.M., Quiros, G.P., 2000. Output Fluctuations in the United States: What Has Changed Since the Early 1980's? The American Economic Review 90, 1464-1476.

McNees, S., 1987. Consensus forecasts: Tyranny or Majority? *New England Economic Review*.

McNees, S., Fine, L. K., 1994. Diversity, uncertainty, and accuracy of inflation forecasts. New England Economic Review, (July), 33-44.

Newbold, P., Harvey, D.I., 2002. Forecast Combination and Encompassing. In: Clements, M., Hendry, D.F. (Eds.). A Comparison of Economic Forecasting, Oxford: Blackwell Press, 268-283.

Ottaviani, M., Sørensen, P.N., 2006. The Strategy of Professional Forecasting. Journal of Financial Economics 81, 441–466.

Page, S., 2007. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton, NJ: Princeton University Press.

Page, S., 2014. Where Diversity Comes From and Why it Matters? European Journal of Social Psychology 44, 267-279.

Patton, A., Timmermann, A., 2010. Why Do Forecasters Disagree? Lessons from the Terms Structure of Cross-Sectional Dispersion. Journal of Monetary Economics 57, 803-820.

Potter, S. 2011. The failure to forecast the Great Recession. Federal Reserve Bank of New York, Liberty Street Economics Blog (November 25).

Prendergast, C., Stole, L., 1996. Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning. Journal of political Economy, 1105-1134.

Rich, R., Raymond, J., Butler, J., 1992. The Relationship between Forecast Dispersion and Forecast Uncertainty: Evidence from a Survey Data-ARCH Model. Journal of Applied Econometrics 7. 131-148.

Rich, R., Tracy, J., 2010. The Relationships Among Expected Inflation, Disagreement, and Uncertainty: Evidence from Matched Point and Density Forecasts. The Review of Economics and Statistics 92, 200-207.

Scharfstein, D., Stein, J., 1990. Herd Behavior and Investment. American Economic Review 80, 465-79.

Schuh, S., 2001. An evaluation of recent macroeconomic forecast errors. New England Economic Review, 35-56.

Soll, J., Larrick, R., 2009. Strategies for Revising Judgment: How (and How Well) People Use Others' Opinions. Journal of Experimental Psychology: Learning, Memory and Cognition 35, 780-805.

Stock, J., Watson, M., 2003a. Forecasting output and inflation: The role of asset prices. Journal of Economic Literature *41*(3), 788-829.

Stock, J., Watson, M., 2003b. Has the Business Cycle Changed? Evidence and Explanations. In: Monetary Policy and the Economy: Adapting to a Changing Economy. Federal Reserve Bank of Kansas City, 9-56.

Surowiecki, J., 2005. The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economics, Society, and Nations. London: Little, Brown.

Timmermann, A., 2006. Forecast Combinations. In: Elliot, G., Granger, C., Timmermann, A. (Eds.). Handbook of Economic Forecasting, Vol. 1, Elsevier, 135-154.

Trichet, J., 2010. Reflections on the nature of monetary policy non-standard measures and finance theory, Opening address at the ECB Central Banking Conference. Frankfurt, Germany, 18.

Truman, B., 1994. Analyst forecasts and herding behavior. Review of financial studies 7(1), 97-124.

Welch, I., 2000. Herding Among Security Analysts. Journal of Financial Economics 58, 369-396.

Wieland, V., Wolters, M., 2012. Macroeconomic model comparisons and forecast competition. Voxeu.org (February 13)

Wieland, V., Cwik, T., Müller, G., Schmidt, S., Wolters, M., 2012. A new comparative approach to macroeconomic modeling and policy analysis. Journal of Economic Behavior & Organization *83*(3), 523-541.

Wieland, V., Wolters, M., 2011. The diversity of forecasts from macroeconomic models of the US economy. Economic Theory *47*, 247-292.

Woodford, M., 2009. Convergence in Macroeconomics: Elements of the New Synthesis. American Economic Journal: Macroeconomics 1, 267-279.

Zarnowitz, V., 1992. Has Macro-Forecasting Failed? Cato Journal 12, 129-160.

Zarnowitz, V., Lambros, L., 1987. Consensus and Uncertainty in Economic Prediction. Journal of Political Economy 95, 591-621.

Table 1

Structural Break Tests for Forecast Dispersion

Horizon (months)	Range of Forecasts			Variance of Forecasts		
	1977-93 Average	94BREAK	R-Squared	1977-93 Average	94BREAK	R-Squared
18	4.44***	-1.70***	0.34	0.83***	-0.58***	0.52
	(13.23)	(4.02)		(8.56)	(5.65)	
17	4.62***	-1.81***	0.34	0.84***	-0.58***	0.50
	(12.91)	(4.10)		(8.63)	(5.66)	
16	4.64***	-1.96***	0.38	0.83***	-0.58***	0.50
	(12.82)	(4.46)		(8.37)	(5.61)	
15	4.39***	-1.66***	0.30	0.78***	-0.50***	0.41
	(12.11)	(3.73)		(7.93)	(4.73)	
14	4.18***	-1.52***	0.22	0.75***	-0.48***	0.30
	(11.47)	(3.07)		(6.59)	(3.73)	
13	4.16***	-1.63***	0.31	0.70***	-0.48***	0.39
	(11.53)	(3.78)		(7.01)	(4.46)	
12	3.85***	-1.75***	0.40	0.64***	-0.46***	0.41
	(11.39)	(4.63)		(6.65)	(4.65)	
11	3.52***	-1.48***	0.35	0.52***	-0.35***	0.37
	(11.64)	(4.30)		(6.57)	(4.26)	
10	3.38***	-1.42***	0.36	0.46***	-0.31***	0.39
	(12.22)	(4.30)		(7.13)	(4.50)	
9	3.28***	-1.52***	0.36	0.39***	-0.27***	0.40
	(10.83)	(4.26)		(7.02)	(4.55)	
8	2.80***	-1.35***	0.40	0.27***	-0.18***	0.40
	(11.53)	(4.62)		(7.15)	(4.57)	

Notes: The table shows regressions of two measures of forecast dispersion (the cross-sectional range and cross-sectional variance of forecasts) on a constant (shown as the 1977-93 average) and dummy variable (94BREAK) which takes on values of zero from 1977 to 1993 and one from 1994 to 2011. All regressions have 35 annual observations with the exception of the one estimated with 18-month horizons which has 34. T-statistics are shown in parentheses and are constructed with robust Huber-White standard errors that correct for possible heteroscedasticity.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 2
Structural Break Tests for Absolute Forecast Errors

	Vintage 1				Vintage 2			
Forecast	1977-93 MAE	94BREAK	R-Squared	Percent Change	1977-93 MAE	94BREAK	R-Squared	Percent Change
$Naïve_{h=12}$	2.08***	-0.57	0.04	-27.4%	2.11***	-0.58	0.04	-27.5%
	(5.97)	(1.17)			(5.67)	(1.10)		
$Consensus_{h=18}$	1.13***	0.13	0.00	11.5%	1.00***	0.44	0.03	44.0%
	(3.65)	(0.34)			(2.80)	(0.97)		
$Consensus_{h=17}$	0.99***	0.20	0.01	20.0%	0.92***	0.44	0.03	48.0%
	(3.60)	(0.55)			(3.05)	(1.05)		
$Consensus_{h=16}$	0.91***	0.21	0.01	23.0%	0.87***	0.42	0.03	48.0%
	(3.66)	(0.60)			(3.19)	(1.04)		
$Consensus_{h=15}$	0.86***	0.22	0.01	26.0%	0.84***	0.38	0.03	45.0%
	(3.77)	(0.70)			(3.34)	(1.04)		
$Consensus_{h=14}$	0.83***	0.18	0.01	18.0%	0.87***	0.27	0.02	31.0%
	(4.65)	(0.70)			(4.24)	(0.86)		
$Consensus_{h=13}$	0.76***	0.17	0.02	22.0%	0.86***	0.23	0.02	27.0%
	(5.07)	(0.78)			(4.97)	(0.85)		
$Consensus_{h=12}$	0.78***	0.10	0.01	22.0%	0.88***	0.17	0.02	19.0%
	(5.62)	(0.52)			(5.51)	(0.71)		
$Consensus_{h=11}$	0.74***	-0.01	0.00	-1.4%	0.81***	0.08	0.00	10.0%
	(5.51)	(0.08)			(5.22)	(0.38)		
$Consensus_{h=10}$	0.71***	-0.10	0.01	-14.0%	0.77***	0.00	0.00	0.0%
	(5.83)	(0.65)			(5.65)	(0.01)		
$Consensus_{h=9}$	0.58***	-0.06	0.01	-10.0%	0.65***	0.11	0.01	17.0%
	(5.78)	(0.45)			(5.42)	(0.69)		
$Consensus_{h=8}$	0.47***	-0.02	0.00	-4.0%	0.61***	0.09	0.01	15.0%
	(5.10)	(1.15)			(6.02)	(0.58)		

Notes: The table shows results from regressions of absolute forecast errors for year-over-year real GDP growth on a constant (1977-93 Average) and dummy variable (94BREAK) that takes on values of zero from 1977 to 1993 and one from 1994 to 2011. All regressions have 35 annual observations with the exception of the one estimated with 18-month horizons which has 34. T-statistics are shown in parentheses and constructed with robust Huber-White standard errors that correct for possible heteroscedasticity. Percent Change shows the percentage change in average absolute errors between 1977-93 and 1994-2011. Forecast errors are measured in the target year using actual real GDP growth estimates reported in the June issue of the *Survey of Current Business* in the year following the target year (Vintage 1) or second year following the target year (Vintage 2).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 3A

Diebold-Mariano Forecasting Accuracy Tests: Vintage 1 GDP Data

	Sample: 1977-1993			Sample: 1994-2011		
Forecast	**MAE**	**S(1)**	**p-value**	**MAE**	**S(1)**	**p-value**
$Naïve_{h=12}$	2.08			1.51		
$Consensus_{h=18}$	1.13	-2.49	**0.013**	1.27	-0.72	0.471
$Consensus_{h=17}$	0.99	-3.11	**0.002**	1.19	-0.94	0.347
$Consensus_{h=16}$	0.91	-3.50	**0.001**	1.13	-1.11	0.266
$Consensus_{h=15}$	0.86	-3.75	**0.002**	1.08	-1.26	0.206
$Consensus_{h=14}$	0.83	-3.81	**0.000**	1.01	-1.40	0.160
$Consensus_{h=13}$	0.76	-4.11	**0.000**	0.94	-1.60	0.111
$Consensus_{h=12}$	0.82	-4.20	**0.000**	0.88	-1.71	0.087
$Consensus_{h=11}$	0.78	-4.35	**0.000**	0.73	-2.11	**0.035**
$Consensus_{h=10}$	0.75	-4.37	**0.000**	0.60	-2.43	**0.015**
$Consensus_{h=9}$	0.61	-4.61	**0.000**	0.52	-2.72	**0.007**
$Consensus_{h=8}$	0.50	-5.05	**0.000**	0.45	-3.10	**0.002**

Notes: The table shows mean absolute errors (MAE) for forecasts of year-over-year real GDP growth based on a Naive model (previous year's GDP growth), and Blue Chip Consensus at the 18- through 8-month horizons. S(1) is the Diebold-Mariano statistic to test the null hypothesis of no difference in accuracy between the Blue Chip Consensus and the naive forecast. The p-values indicate the level of significance at which we can regject the null of equal accuracy. The uniform kernel was used to calculate the long-run variance, with the maximum lag order calculated from the Schwert criterion as a function of the sample size. All regressions are estimated over the 1977-2011 sample period and have 35 annual observations with the exception of the one estimated with 18-month horizons which has 34. Forecast errors are measured in the target year using actual real GDP growth estimates reported in the June issue of the *Survey of Current Business* in the year following the target year (Vintage 1) or second year following the target year (Vintage 2).

Table 3B

Diebold-Mariano Forecasting Accuracy Tests: Vintage 2 GDP Data

	Sample: 1977-1993			Sample: 1994-2011		
Forecast	**MAE**	**S(1)**	**p-value**	**MAE**	**S(1)**	**p-value**
$Naïve_{h=12}$	2.11			1.53		
$Consensus_{h=18}$	1.00	-2.77	0.007	1.44	-0.23	0.819
$Consensus_{h=17}$	0.92	-3.00	**0.003**	1.36	-0.42	0.672
$Consensus_{h=16}$	0.87	-3.18	**0.002**	1.29	-0.60	0.545
$Consensus_{h=15}$	0.84	-3.33	**0.001**	1.21	-0.82	0.413
$Consensus_{h=14}$	0.87	-3.30	**0.001**	1.14	-1.01	0.310
$Consensus_{h=13}$	0.86	-3.47	**0.001**	1.09	-1.17	0.243
$Consensus_{h=12}$	0.93	-3.52	**0.000**	1.05	-1.25	0.211
$Consensus_{h=11}$	0.85	-3.73	**0.000**	0.89	-1.65	0.098
$Consensus_{h=10}$	0.82	-3.77	**0.000**	0.78	-1.93	**0.053**
$Consensus_{h=9}$	0.68	-3.98	**0.000**	0.76	-2.00	**0.047**
$Consensus_{h=8}$	0.65	-4.09	**0.000**	0.70	-2.25	**0.024**

Notes: The table shows mean absolute errors (MAE) for forecasts of year-over-year real GDP growth based on a Naive model (previous year's GDP growth), and Blue Chip Consensus at the 18- through 8-month horizons. S(1) is the Diebold-Mariano statistic to test the null hypothesis of no difference in accuracy between the Blue Chip Consensus and the naive forecast. The p-values indicate the level of signifiance at which we can regject the null of equal accuracy. The uniform kernel was used to calculate the long-run variance, with the maximum lag order calculated from the Schwert criterion as a function of the sample size. All regressions are estimated over the 1977-2011 sample period and have 35 annual observations with the exception of the one estimated with 18-month horizons which has 34. Forecast errors are measured in the target year using actual real GDP growth estimates reported in the June issue of the *Survey of Current Business* in the year following the target year (Vintage 1) or second year following the target year (Vintage 2).

Table 4

Test for Structural Break in the Ratio of the Squared Consensus Error to the Average Individual Square Error

Horizon (months)	Vintage 1			Vintage 2		
	Average θ 1977-93	94BREAK	R-Squared	Average θ 1977-93	94BREAK	R-Squared
18	0.46***	0.25**	0.14	0.33***	0.44***	0.38
	(5.89)	(2.30)		(3.54)	(4.26)	
17	0.42***	0.23*	0.11	0.35***	0.38***	0.29
	(5.44)	(1.98)		(4.08)	(3.68)	
16	0.41***	0.21*	0.09	0.35***	0.33***	0.24
	(5.44)	(1.81)		(4.28)	(3.22)	
15	0.42***	0.20	0.08	0.38***	0.28**	0.16
	(5.26)	(1.64)		(4.36)	(2.51)	
14	0.43***	0.19*	0.08	0.42***	0.23**	0.12
	(6.01)	(1.71)		(5.70)	(2.16)	
13	0.41***	0.22*	0.10	0.45***	0.24**	0.15
	(5.66)	(1.96)		(6.30)	(2.45)	
12	0.43***	0.20*	0.10	0.48***	0.26***	0.20
	(6.88)	(1.92)		(7.12)	(2.84)	
11	0.45***	0.17	0.08	0.46***	0.22**	0.12
	(6.55)	(1.66)		(5.78)	(2.14)	
10	0.45***	0.12	0.04	0.47***	0.19	0.07
	(6.77)	(1.24)		(6.09)	(1.64)	
9	0.41***	0.14	0.05	0.45***	0.28***	0.19
	(6.03)	(1.35)		(5.99)	(2.80)	
8	0.38***	0.19*	0.09	0.51***	0.26***	0.21
	(4.71)	(1.79)		(7.01)	(2.91)	

Notes: The table shows regressions of the ratio of the squared consensus error to the average individual squared error (θ) on a constant and dummy variable that takes on values of zero from 1977 to 1993 and one from 1994 to 2011. T-statistics are shown in parentheses and constructed with robust Huber-White standard errors to correct for possible heteroscedasticity. All regressions are estimated over the 1977-2011 sample period and have 35 annual observations with the exception of the one estimated with 18-month horizons which has 34. Forecast errors are measured in the target year using actual real GDP growth estimates reported in the June issue of the *Survey of Current Business* in the year following the target year (Vintage 1) or second year following the target year (Vintage 2).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 5
Regressions of Blue Chip Absolute Consensus Errors on Forecast Dispersion

Horizon (month)	Vintage 1					Vintage 2				
	Constant	p-val.	Dispersion	p-val.	R^2	Constant	p-val.	Dispersion	p-val.	R^2
24	1.45***	0.000	-0.62*	0.062	0.07	1.65***	0.000	-0.72*	0.054	0.07
23	1.52***	0.000	-0.80***	0.008	0.07	1.76***	0.000	-1.00***	0.001	0.08
22	1.37***	0.000	-0.42	0.292	0.02	1.69***	0.000	-0.83**	0.021	0.06
21	1.35***	0.000	-0.39	0.327	0.02	1.69***	0.000	-0.93**	0.016	0.07
20	1.35***	0.000	-0.37	0.314	0.02	1.66***	0.000	-0.88**	0.041	0.06
19	1.36***	0.000	-0.42	0.233	0.02	1.68***	0.000	-1.00**	0.043	0.07
18	1.32***	0.000	-0.46	0.218	0.02	1.65***	0.000	-1.05**	0.038	0.08
17	1.17***	0.000	-0.29	0.408	0.01	1.43***	0.000	-0.61	0.160	0.03
16	1.07***	0.001	-0.18	0.646	0.00	1.27***	0.002	-0.36	0.444	0.01
15	0.96***	0.001	-0.01	0.996	0.00	1.14***	0.002	-0.18	0.691	0.00
14	0.73***	0.001	0.49	0.101	0.08	0.91***	0.004	0.32	0.478	0.02
13	0.70***	0.001	0.40	0.283	0.04	0.88***	0.002	0.32	0.533	0.02
12	0.72***	0.000	0.26	0.525	0.02	0.84***	0.001	0.43	0.378	0.03
11	0.54***	0.004	0.62	0.210	0.08	0.56***	0.004	1.14**	0.037	0.15
10	0.46***	0.004	0.66	0.165	0.10	0.58***	0.002	0.88	0.144	0.10
9	0.40***	0.003	0.70*	0.089	0.11	0.54***	0.000	1.10**	0.027	0.16
8	0.40***	0.001	0.35	0.467	0.01	0.56***	0.000	1.02	0.101	0.06
7	0.35***	0.000	0.38	0.301	0.02	0.55***	0.000	0.94**	0.021	0.07
6	0.33***	0.000	0.47	0.270	0.02	0.55***	0.000	1.10**	0.019	0.07
5	0.28***	0.000	0.33	0.561	0.00	0.57***	0.000	1.03	0.409	0.02
4	0.25***	0.000	0.78	0.307	0.01	0.50***	0.000	1.50	0.371	0.01
3	0.22***	0.002	0.34	0.788	0.00	0.48***	0.000	1.43	0.620	0.01
2	0.19***	0.000	1.03	0.632	0.00	0.45***	0.001	3.22	0.532	0.01
1	0.15***	0.005	0.92	0.696	0.00	0.50***	0.000	-0.78	0.904	0.01

Notes: The table contains results for regressions of absolute consensus errors for year-over-year real GDP growth on a constant and forecast dispersion. p-values show the probability at which we can reject the null that a coefficient estimate is zero and are based on robust Huber-White standard error estimates that correct for possible heteroscedasticity. All regressions are estimated over the 1986-2011 sample period and have 26 annual observations. Forecast errors are measured in the target year using actual real GDP growth estimates reported in the June issue of the Survey of Current Businessin the year following the target year (Vintage 1) or second year following the target year (Vintage 2).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

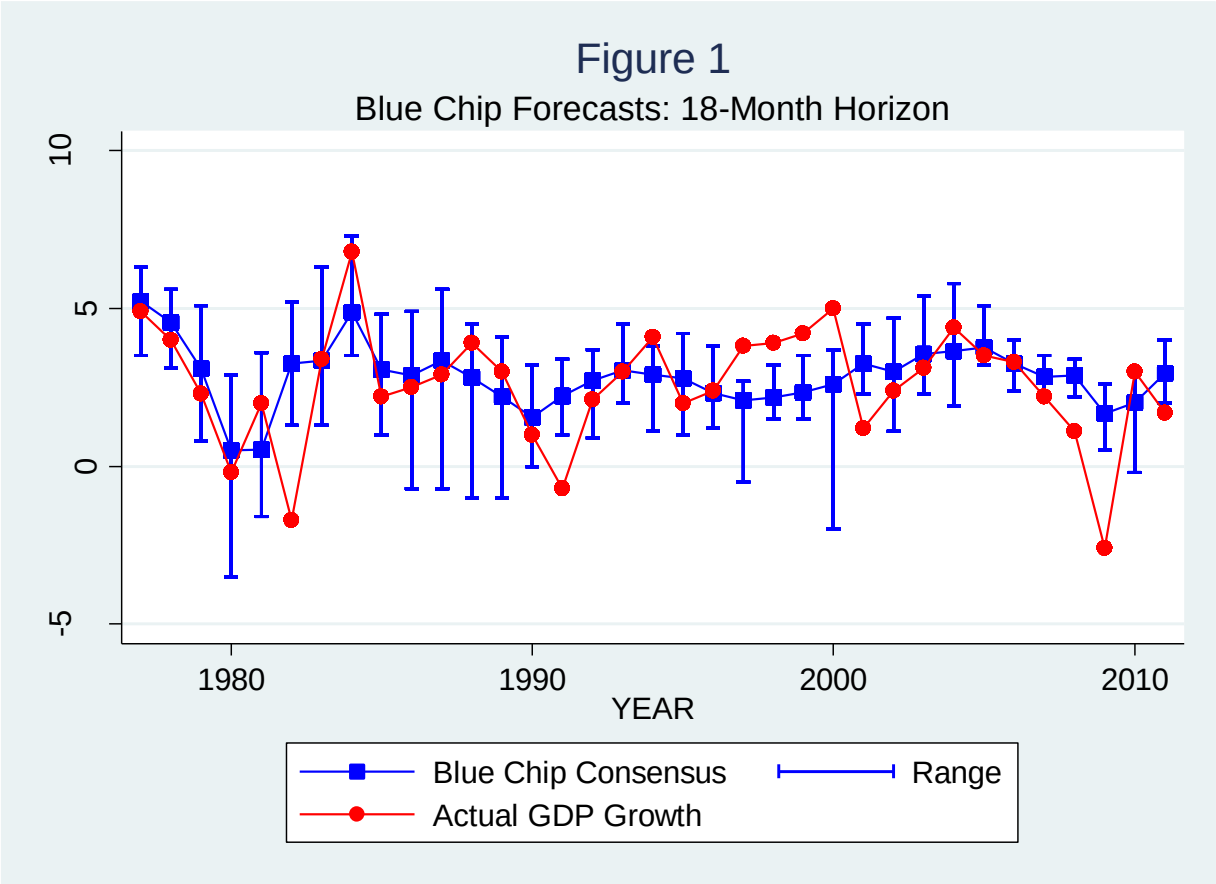
Figure 1
Blue Chip Forecasts: 18-Month Horizon
10
5
0
-5
1980
1990
2000
2010
YEAR
Blue Chip Consensus
Range
Actual GDP Growth

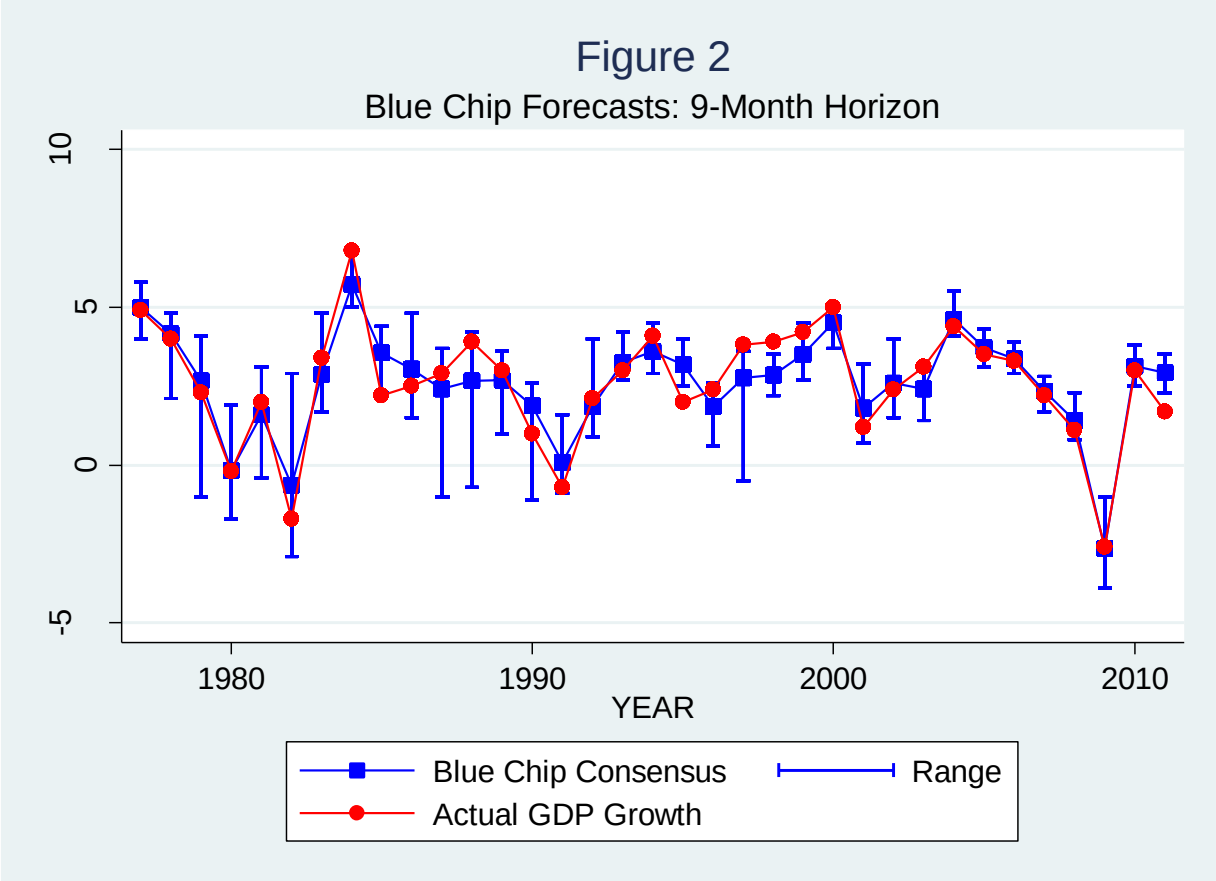
Figure 2
Blue Chip Forecasts: 9-Month Horizon
10
5
0
-5
1980
1990
2000
2010
YEAR
Blue Chip Consensus
Range
Actual GDP Growth

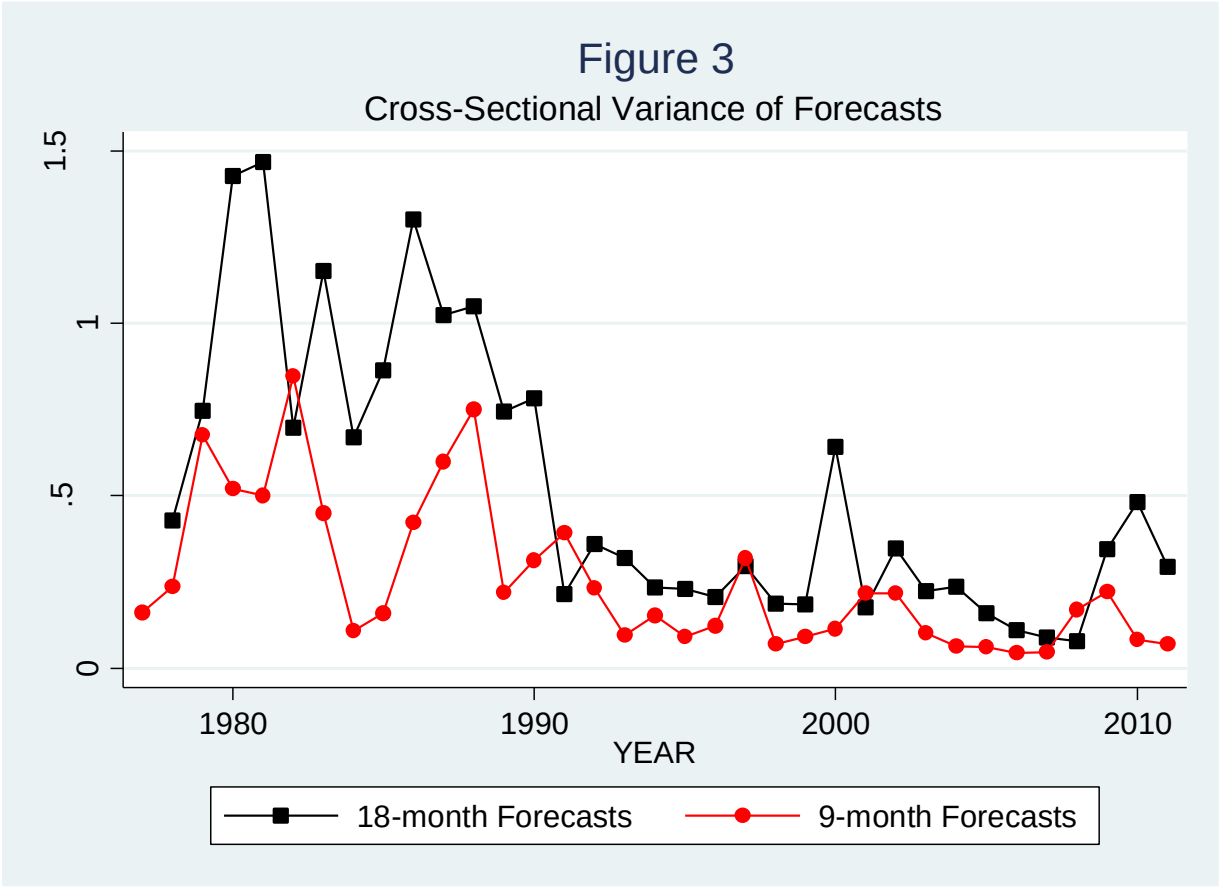
Figure 3
Cross-Sectional Variance of Forecasts
1.5
1
.5
0
1980
1990
2000
2010
YEAR
18-month Forecasts
9-month Forecasts

Figure 4

Average Squared Individual Errors: 18-Month Horizon

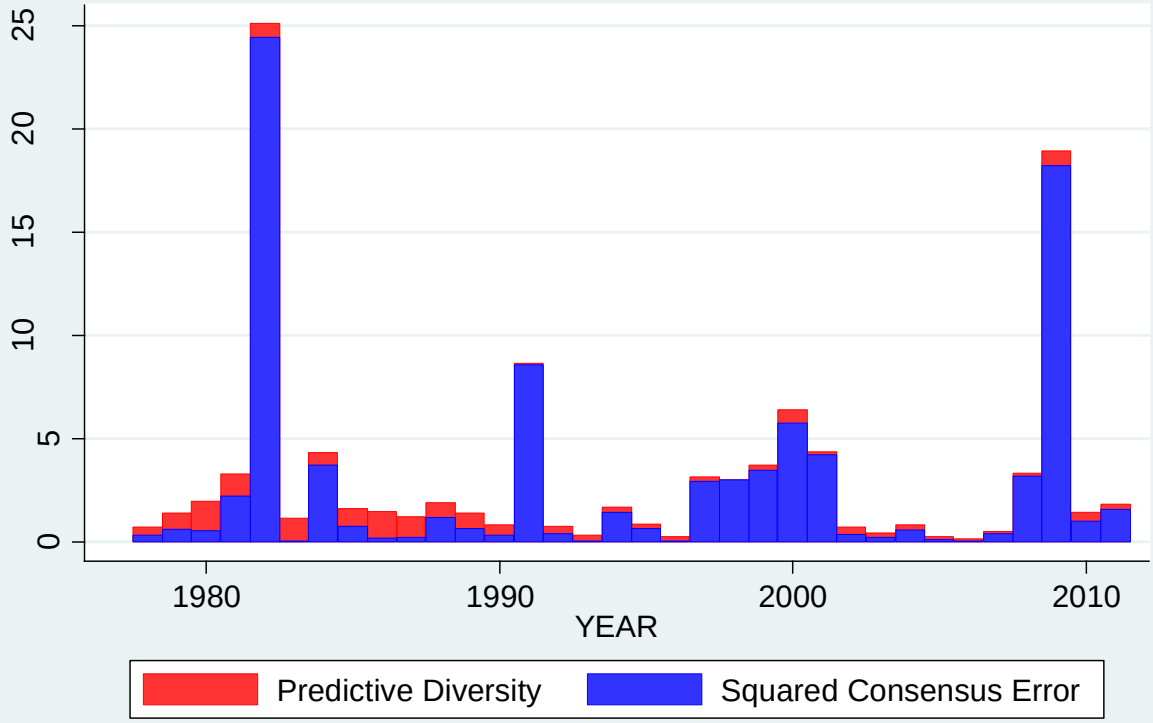

Figure 5

Average squared Individual Errors: 9-Month Horizon