H.O. Stekler Research Program on Forecasting

WORKING PAPER SERIES

Measuring Uncertainty of a Combined Forecast and Some Tests for Forecaster Heterogeneity

Kajal Lahiri
University at Albany, SUNY
klahiri@albany.edu

Huaming Peng
Rensselaer Polytechnic Institute
pengh5@rpi.edu

Xuguang Simon Sheng
American University
sheng@american.edu

Working Paper No. 2021-005

August, 2021

H. O. STEKLER RESEARCH PROGRAM ON FORECASTING
Department of Economics
Columbian College of Arts & Sciences
The George Washington University
Washington, DC 20052
https://www2.gwu.edu/~forcpgm

Measuring Uncertainty of A Combined Forecast and Some Tests for Forecaster Heterogeneity*

Kajal Lahiri[a], Huaming Peng[b] and Xuguang Simon Sheng[c]
[a] *University at Albany, SUNY*
[b] *Rensselaer Polytechnic Institute*
[c] *American University*

Abstract

From the standpoint of a policy maker who has access to a number of expert forecasts, the uncertainty of a combined or ensemble forecast should be interpreted as that of a typical forecaster randomly drawn from the pool. This uncertainty formula should incorporate forecaster discord, as justified by (i) disagreement as a component of combined forecast uncertainty, (ii) the model averaging literature and (iii) central banks' communication of uncertainty via fan charts. Using new statistics to test for the homogeneity of idiosyncratic errors under the joint limits with both T and n approaching infinity simultaneously, we find that some previously used measures can significantly underestimate the conceptually correct benchmark forecast uncertainty.

JEL classification: C12; C33; E37

Keywords: Central Bank Communication, Disagreement, Ensemble, Forecast Combination, Panel Data, Uncertainty

*We are indebted to Hashem Pesaran for giving us many helpful comments at the World Congress of the Econometric Society in Shanghai. Earlier versions of this paper were also presented at the Econometric Society winter meeting, IAAE annual conference, London Conference on "Uncertainty and Economic Forecasting", and CIRET Workshop on "Economic Cycles and Uncertainty". We thank Todd Clark, David Draper, David Hendry, Allan Timmermann, Ken Wallis, Mark Watson and an anonymous referee for many constructive suggestions, and Recai Yucel for sharing his expertise on multiple imputations. Cheng Yang and Herbert Zhao provided able research assistance. For supplemental information, see the online appendix at `https://www.cesifo.org/en/publikationen/2020/working-paper/measuring-uncertainty-combined-forecast-and-some-tests-forecaster`

1 Introduction

Consider the problem of a macro policy maker who often has to aggregate a number of expert forecasts for the purpose of a uniform policy making. A general solution was provided by Bates and Granger (1969) who have inspired extensive research on forecast combination, as evidenced by two comprehensive surveys in Clemen (1989) and Timmermann (2006), and many additional papers since 2006.[1] The solution based on minimizing the mean squared error of the combined forecasts calls for a performance-based weighted average of individual forecasts with precision of the combined forecast that is readily shown to be better than any of the constituent elements under reasonable conditions.[2] Thus, Wei and Yang (2012) characterize this approach as "combination for improvement." However, many studies have found that a simple average is often as good as the Bates-Granger estimator, possibly due to large estimation error of the weights, the variances of individual forecast errors being the same or their pair-wise correlations being the same; see, e.g. Bunn (1985), Clemen and Winkler (1986), Gupta and Wilton (1987), Palm and Zellner (1992) and Smith and Wallis (2009), among many others. Under the standard factor decomposition of a panel of forecasts, where the cross correlations of forecast errors can be attributed to a common aggregate shock, the precision of this equally-weighted average is simply a function of the variance of this common shock that nets out the uncertainty associated with idiosyncratic errors. This precision formula should be enriched with disagreement, as motivated by a variety of theoretical, empirical, and policy factors.

As Timmermann (2006, p.141) has noted, heightened discord among forecasters, *ceteris*

[1] Granger and Jeon (2004) call this approach of making inference based on combined outputs from alternative models as "thick modeling."

[2] The superior performance of the consensus forecast relative to individual forecasts follows from Jensen's inequality, which states that with convex loss functions, the loss associated with the mean forecast is generally less than the mean loss of individual forecasts, cf. Manski (2011). See also Granger (1989), Makridakis (1989), Diebold and Lopez (1996), Newbold and Harvey (2001) and Hendry and Clements (2004) for discussing why combining is beneficial due to unobserved information sets, diversification gains, insurance against structural breaks and misspecifications.

paribus, may be indicative of higher uncertainty in the combined forecast from the standpoint of a policy maker. Thus, the precision formula for the average (or "consensus") forecast should reflect disagreement among experts as part of forecast uncertainty, which is desirable in many situations. On the other hand, the use of disagreement as a sole proxy for forecast uncertainty continues to be debated in other contexts.

Another justification for incorporating disagreement as part of aggregate uncertainty comes from the rich literature on model averaging pioneered by Leamer (1978). Draper (1995) and Buckland, et al. (1997) present cogent explications of the result using Bayesian and Frequentist approaches respectively. See Hansen (2008) and Amisano and Geweke (2017) for more recent advances.

A third consideration for using a theoretically sound uncertainty measure of the consensus forecast comes from the recent advances in the presentation and communication strategies by a number of central banks, pioneered by Bank of England's fan charts to report forecast uncertainty. For the credibility of forecasts in the long run, it is essential that the reported confidence bands for forecasts be properly calibrated. In the U.S., from November 2007, all FOMC members are required to provide their judgments as to whether the uncertainty attached to their projections is greater than, smaller than, or broadly similar to typical levels of forecast uncertainty in the past. In order to aid each FOMC member to report their personal uncertainty estimates, Reifschneider and Tulip (2019) have provided a measure for gauging the average magnitude of historical uncertainty using information on past forecast errors from a number of private and government forecasters. These benchmark estimates for a number of target variables are reported in the minutes of each FOMC meeting and are used by the public to interpret the responses of the FOMC participants. We show how this measure incorporates the disagreement amongst forecasters as a component of forecast uncertainty, but particular formula used may underestimate the true historical uncertainty if the individual forecast errors are heterogeneous. Given that these historical benchmark numbers

are fed into the highest level of national decision making, a careful examination of a number of alternative uncertainty measures relevant for a policy maker cannot be overemphasized.

In this paper we establish the asymptotic limits for these alternative measures of uncertainty under the joint limits with both the time series (T) and cross section (n) dimensions approaching infinity simultaneously, and develop tests to check if the uncertainty measures are statistically different and the forecasters are exchangeable. We build on Issler and Lima (2009), who have shown the optimality of the (bias-corrected) simple average forecast using panel data sequential asymptotics. Our tests identify the differences in the idiosyncratic error variances, in addition to the differences in the means, and thus shed new light on the heterogeneity of expectation formation processes.[3] A Monte Carlo study confirms that the test performs well in our context. We use individual forecasts from the Survey of Professional Forecasters (SPF) and the Michigan Survey of Consumers (MSC) to show that the uncertainty measure conventionally attached to a consensus forecast using the Bates-Granger approach and the Reifschneider and Tulip (2019) [hereafter RT] benchmark measure can underestimate the true uncertainty under certain circumstances. Similar to Rossi and Sekhposyan (2015) and Jo and Sekkel (2019), our measure is based on subjective forecasts of market participants and reflects their perceived uncertainty. In contrast to these two papers, but like RT, we include both common and idiosyncratic uncertainty in the measurement and provide the typical levels of uncertainty seen on average over history. Our test also confirms these results at the 1% level for multiple forecast horizons.

The plan of the paper is as follows. Section 2 derives the relationship between disagreement and overall forecast uncertainty. Section 3 compares different measures of historical uncertainty and develops a new test for forecaster homogeneity. In Section 4 we use SPF data on real GDP and inflation forecasts by experts and the MSC data on price expectations

[3]See, e.g. Lahiri and Sheng (2008), Patton and Timmermann (2010), and Andrade, et al. (2016). Pesaran and Weale (2006) contains an early elaboration of many of these issues.

made by households to highlight the differences in the alternative uncertainty measures, and implement our test for forecaster homogeneity. Pesaran (1987) established the value of using survey data in measuring uncertainty and testing for rationality. Finally, Section 5 summarizes the results and presents some concluding remarks. Proofs of theorems and corollaries in Section 3 are relegated to the unpublished mathematical appendix in Lahiri, Peng and Sheng (2020).

2 Uncertainty and Disagreement

Let Y_t be the random variable of interest, F_{ith} be the forecast of Y_t made by individual i at time $t-h$. Then individual i's forecast error, e_{ith}, can be defined as

$$e_{ith} = A_t - F_{ith}, \tag{1}$$

where A_t is the actual realization of Y_t. Following a long tradition, e.g., Davies and Lahiri (1995) and Gaglianone and Lima (2012), we write e_{ith} as the sum of an individual bias, μ_{it}, a common component, λ_{th} and idiosyncratic errors, ε_{ith}:

$$e_{ith} = \mu_{ith} + \lambda_{th} + \varepsilon_{ith}, \tag{2}$$

where μ_{ith} is nonrandom and time-varying, λ_{th} represents the cumulative weighted effect of all independent shocks that occurred from h-period ahead to the end of target year t. Thus, even if forecasters make "perfect" forecasts, the forecast error may still be nonzero due to shocks (λ_{th}), which are, by nature, unpredictable. Forecasters, however, do not make "perfect" forecasts even in the absence of unanticipated shocks. This "lack of perfection" is due to other factors (e.g., differences in information processing, loss functions, interpretation, judgment, and forecasting models) specific to a given individual at a given point in time and

is represented by the idiosyncratic error, ε_{ith}.

In order to establish the relationship between different measures of uncertainty and derive their asymptotic limits, we make the following simplifying assumptions:

Assumption 1 (Bias)

μ_{it} is nonstochastic for all i and all t with $\sup_i \frac{1}{T}\sum_{t=1}^{T}\mu_{it}^4 = O\Big((Tn)^{-\alpha}\Big)$ for some $\alpha \geq 2$.

Assumption 2 (Common Shocks)

$\lambda_{th} = \sum_{k=0}^{h-1}\theta_k\zeta_{thk}$ with $\theta_0 = 1$, $|\theta_k| < \infty$ for $k = 1, \ldots, h-1$, where ζ_{thk}, occurred from k-period ahead to the end of target year t, are economic shocks that are uncorrelated across k, and stationary ergodic over t such that $E(\zeta_{thk}) = 0$, $E(\zeta_{thk}^2) = \sigma_{\zeta hk}^2$, $E|\zeta_{thk}|^{4+\delta} < \infty$ and $var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\sum_{k=0}^{h-1}\theta_k\zeta_{thk}\right)^2\right) \to \varphi_{\lambda h}$ with $0 < \varphi_{\lambda h} < \infty$ as $T \to \infty$.

Assumption 3 (Idiosyncratic Shocks)

ε_{ith} is independent identically distributed over t, and independent potentially non-identically distributed across i with $E\varepsilon_{ith} = 0$, $E\varepsilon_{ith}^2 = \sigma_{\varepsilon ih}^2$, $\sigma_{\varepsilon h}^2 = \lim \frac{1}{n}\sum_{i=1}^{n}\sigma_{\varepsilon ih}^2$, $E\varepsilon_{ith}^3 = 0$, $E(\varepsilon_{ith}^4) = \omega_{\varepsilon ih}$ such that $\inf_i \sigma_{\varepsilon ih}^2 > 0$, $\sup_i E\varepsilon_{ith}^8 < \infty$. In addition, $\omega_{\varepsilon ih} = \omega_{\varepsilon jh}$ whenever $\sigma_{\varepsilon i}^2 = \sigma_{\varepsilon j}^2$.

Assumption 4 (Relations)

λ_{th} is independent of ε_{ish} for all i, t and s.

Remark 1. *Assumption 1 allows for time-varying nonrandom bias, which is more general than the time-invariant assumption made in the literature (for example, Issler and Lima (2009)) and hence potentially has a wider range of applications. The condition* $\frac{1}{T}\sum_{t=1}^{T}\mu_{ith}^4 = O_p((Tn)^{-\alpha})$ *for some* $\alpha \geq 2$ *helps to ensure that individual bias is negligible in the asymptotic limits involving various ex post measures of forecast uncertainty. The eventually vanishing bias condition is in line with the spillover effect that the bias gets smaller as more forecasters learn from each other, and consistent with the empirical evidence that forecasters' biases diminish over time as they gain experience, cf. Pesaran (1987) and Lahiri and Sheng (2008).*[4]

[4]Reifschneider and Tulip (2019) report the biases to be transitory. See also Clark, et al. (2020) who make a similar assumption. Note that our bias condition allows for heterogenous rates of individual biases approaching zero.

Assumption 2 implies that λ_{th} *is a stationary ergodic moving average process of order at most* $h-1$ *with* $E\lambda_{th}=0$, $E\lambda_{th}^2=\sigma_{\lambda h}^2=\sum_{k=0}^{h-1}\theta_k\sigma_{\zeta hk}^2$, $E\left|\lambda_{th}\right|^{4+\delta}<\infty$ *for some* $\delta>0$, *and* $var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\lambda_{th}^2\right)\rightarrow\varphi_\lambda$ *as* $T\rightarrow\infty$. *Thus, this condition is almost identical to the assumption in Issler and Lima (2009) except for the higher moment condition, which, together with the higher moment assumption of* ε_{ith}, *is required to establish the asymptotic limits in Theorem 1. Assumption 3 is standard in errors component or factor analysis. It can be readily extended, at the expense of some technical complication, to allow for both some weak time dependence and cross-sectional dependence of groupwise block form brought by some residual group-wide influences.*[5] *The requirement,* $\omega_{\varepsilon ih}=\omega_{\varepsilon jh}$ *whenever* $\sigma_{\varepsilon ih}^2=\sigma_{\varepsilon jh}^2$, *though slightly restrictive, still allows for a wide range of probability distributions such as normal,* t, *and uniform distributions with zero mean. The independence of* λ_{th} *and* ε_{ish} *in Assumption 4 is common in errors component or factor models.*

Taken together, the assumptions 1-4 imply that the individual forecast error is not only an asymptotic stationary and ergodic process for any given horizon h, but also has a factor structure interpretation. Given a panel of forecasts, Lahiri and Sheng (2010) decompose the average squared individual forecast errors as

$$\frac{1}{n}\sum_{i=1}^{n}e_{ith}^2=(A_t-F_{\cdot th})^2+\frac{1}{n}\sum_{i=1}^{n}(F_{ith}-F_{\cdot th})^2, \tag{3}$$

where $F_{\cdot th}=\frac{1}{n}\sum_{j=1}^{n}F_{jth}$. The simple average $\frac{1}{n}\sum_{i=1}^{n}e_{ith}^2$ can be viewed as the volatility associated with a representative forecaster, selected randomly from among all forecasters, e.g. Giordani and Söderlind (2003), Lahiri and Sheng (2010), and Ozturk and Sheng (2018). This decomposition of the uncertainty of a typical forecaster is consistent with the vast literature on the capital asset pricing model that decomposes the return volatility of a typical

[5]For example, some residual group-wide influences, resulting from the facts that groups of forecasters may adopt similar models, loss functions, judgements of interpretations under certain circumstances, may not be strong enough or explicit enough to be embodied in specific common factors.

stock into market volatility and firm-specific volatility; see, e.g. Campbell, et al. (2001).

By taking time average on both sides of equation (3), we get an empirical measure of historical forecast uncertainty based on past errors such that

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} e_{ith}^2 = \frac{1}{T}\sum_{t=1}^{T}(A_t - F_{\cdot th})^2 + \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(F_{ith} - F_{\cdot th})^2. \tag{4}$$

Equation (4) states that the squared measure can be decomposed into two components: uncertainty that is common to all forecasters and uncertainty that arises from heterogeneity of individual forecasters. The first component is the empirical variance of the average that is conventionally taken as the uncertainty of the consensus forecast; see, e.g. Patton and Timmermann (2011) and Clements (2014). The second component is the disagreement among forecasters. Similar decomposition of uncertainty is also obtained by Draper (1995) in assessing model uncertainty via Bayesian approach. Geweke and Amisano (2014) presented a parallel decomposition of predictive variance from Bayesian model averaging in terms of intrinsic and extrinsic variances.

By virtue of the assumptions 1-4, the population analog of equation (4) is given by

$$\frac{1}{n}\sum_{i=1}^{n} E(e_{ith}^2) = \sigma_{\lambda h}^2 + \frac{1}{n}\sum_{i=1}^{n}\sigma_{\varepsilon ih}^2 + \frac{1}{n}\sum_{i=1}^{n}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mu_{ith}^2. \tag{5}$$

It is now obvious from equation (5) that the squared uncertainty of a typical forecaster arises from the variance of the aggregate shock common to all forecasters and from the heterogeneity of individual forecasters that contains both the average idiosyncratic variance and the average of the variance of individual biases. What is not readily recognized in the literature is that apart from the disagreement coming from time-varying systematic biases (i.e., $\frac{1}{n}\sum_{i=1}^{n}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mu_{ith}^2$), the average of individual variances also contains a disagreement component coming from $\frac{1}{n}\sum_{i=1}^{n}\sigma_{\varepsilon ih}^2$. In the context of the empirical examples on real GDP and inflation forecasts that we report in Section 4, a *model uncertainty audit* reveals that

the variance explained by the systematic bias component is tiny compared to the other two components in equation (5). A similar result on the transitory nature of the individual bias terms is also reported by Reifschneider and Tulip (2019).

3 Measures of Historical Uncertainty and Tests for Forecaster Homogeneity

3.1 Measures of Forecast Uncertainty and their Asymptotic Properties

A common practice in the uncertainty literature is to quantify uncertainty in terms of standard deviation. In line with this tradition, taking the square root of the average of the individual variances observed over the sample period in equation (4) gives the historical uncertainty faced by a policy maker while using a typical forecaster.

Definition (Forecast combination uncertainty)

The historical uncertainty of a combined forecast from n experts is given by

$$RMSE_{LPS} = \sqrt{\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} e_{ith}^2}. \tag{6}$$

The historical uncertainty measure, $RMSE_{LPS}$ in equation (6), in which the uncertainties add in quadrature is consistent with the standard error propagation formula used for calculating uncertainties among experimental scientists in engineering, physics, chemistry and biology, cf. Draper (1995).

On the other hand, the conventional choice as suggested by Bates and Granger (1969) is the

root mean squared error (RMSE) of the average forecast

$$RMSE_{AF} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{n}\sum_{i=1}^{n} e_{ith}\right)^2}. \tag{7}$$

With the stated objective of using forecast errors made by a panel of forecasters to generate a benchmark estimate of historical forecast uncertainty, Reifschneider and Tulip (2019) propose the following measure

$$RMSE_{RT} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{T}\sum_{t=1}^{T} e_{ith}^2}. \tag{8}$$

They explicitly recognized that the empirical uncertainty faced by a typical forecaster is the average of the estimated individual uncertainty. Along this line, Jurado, et al. (2015) proposed an ex post analog of aggregate uncertainty measure. However, as Boero, et al. (2008) pointed out, aggregating individual standard deviations as a measure of collective uncertainty would violate the identify in equation (4). Obviously, $RMSE_{RT}$ is distinct from $RMSE_{LPS}$, and by construction, incorporates partially the disagreement as a component of uncertainty as shown in the following theorem and corollary.

Theorem 1. *Suppose Assumptions 1-4 hold. Then as* $(n, T \to \infty)$,

(i) $\sqrt{T}\left(RMSE_{AF}^2 - \sigma_{\lambda h}^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sigma_{\varepsilon ih}^2\right) \to_d N(0, \varphi_{\lambda h})$.

(ii) $\sqrt{T}\left(RMSE_{RT}^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\sigma_{\lambda h}^2 + \sigma_{\varepsilon ih}^2}\right)^2\right) \to_d N(0, \phi\varphi_{\lambda h})$,

where $\phi = \left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_{\lambda h}^2 + \sigma_{\varepsilon ih}^2\right)^{1/2}\right)^2 \left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_{\lambda h}^2 + \sigma_{\varepsilon ih}^2\right)^{-1/2}\right)^2$.

(iii) $\sqrt{T}\left(RMSE_{LPS}^2 - \left(\sigma_{\lambda h}^2 + \frac{1}{n}\sum_{i=1}^{n}\sigma_{\varepsilon ih}^2\right)\right) \to_d N(0, \varphi_{\lambda h})$.

An immediate consequence of Theorem 1 is Corollary 1.

Corollary 1. *Suppose Assumptions 1-4 hold. Then as* $(n, T \to \infty)$,

(i) $RMSE_{AF} \to_p \sigma_{\lambda h}$.

(ii) $RMSE_{RT} \to_p \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sqrt{\sigma_{\lambda h}^2 + \sigma_{\varepsilon ih}^2}$.

(iii) $RMSE_{LPS} \to_p \sqrt{(\sigma_{\lambda h}^2 + \sigma_{\varepsilon h}^2)}$.

Remark 2. *$RMSE_{AF}$ tends to ignore the uncertainty associated with the idiosyncratic shocks, especially when n is large since $RMSE_{AF}^2 = \sigma_{\lambda h}^2 + \frac{1}{n^2} \sum_{i=1}^{n} \sigma_{\varepsilon ih}^2 + O_p(T^{-1/2})$ with $\frac{1}{n^2} \sum_{i=1}^{n} \sigma_{\varepsilon ih}^2 = O(\frac{1}{n})$. By contrast, for $RMSE_{LPS}$, we have $RMSE_{LPS}^2 \to_p \sigma_{\lambda h}^2 + \sigma_{\varepsilon h}^2$ as $(n, T \to \infty)$. Finally, it is trivial to see that $RMSE_{LPS}$ and $RMSE_{AF}$ yield identical asymptotic limit if and only if $\sigma_{\varepsilon h}^2 = 0$.*

Remark 3. *Corollary 1 implies, in light of Jensen's inequality, $RMSE_{RT} \leq RMSE_{LPS}$ in the limit. $RMSE_{RT}$, though allows for some disagreement among forecasters, underestimates the historical uncertainty especially in the presence of unequal idiosyncratic error variances. The amount of underestimation in the limit, obtained via applying second-order Taylor's expansion to $\sqrt{1 + (\sigma_{\lambda h}^2 + \sigma_{\varepsilon h}^2)^{-1} (\sigma_{\varepsilon ih}^2 - \sigma_{\varepsilon h}^2)}$, is given by*

$$RMSE_{LPS} - RMSE_{RT} \to_p \frac{1}{8} \left(\sigma_{\lambda h}^2 + \sigma_{\varepsilon h}^2\right)^{-3/2} var(\sigma_{\varepsilon ih}^2), \tag{9}$$

where $var(\sigma_{\varepsilon ih}^2) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\sigma_{\varepsilon ih}^2 - \sigma_{\varepsilon h}^2)^2$. Thus, $RMSE_{LPS}$ and $RMSE_{RT}$ are equal in the limit if and only if $var(\sigma_{\varepsilon ih}^2) = 0$ since $\frac{1}{8} (\sigma_{\lambda h}^2 + \sigma_{\varepsilon h}^2)^{-3/2}$ is a positive constant.

Remark 4. *Interestingly, $RMSE_{RT}$, as a measure of the typical level of historical uncertainty, is potentially more volatile than $RMSE_{LPS}$ because of $\phi \geq 1$ by virtue of Cauchy-Schwarz inequality.*

3.2 Tests for Forecaster Homogeneity and their Asymptotic Distribution

It is evident from Remark 3 that testing for the equality of $RMSE_{LPS}$ and $RMSE_{RT}$ in the limit is equivalent to testing for $var(\sigma^2_{\varepsilon ih}) = 0$, that is, $\sigma^2_{\varepsilon ih} = \sigma^2_{\varepsilon h}$ for almost all i with n approaching infinity. But to examine whether $RMSE_{RT}$ and $RMSE_{LPS}$ give statistically different measures of uncertainty in the context of a particular data set, it is necessary to restrict the null hypothesis to $\sigma^2_{\varepsilon ih} = \sigma^2_{\varepsilon h}$ for all i.[6] To derive the test, we first obtain various bias-corrected estimators by letting $e_{\cdot th} = \frac{1}{n}\sum_{i=1}^{n} e_{ith}$, $\hat{\varepsilon}_{ith} = e_{ith} - e_{\cdot th}$, and defining $\widehat{\sigma}^2_{\varepsilon ih} = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}^2_{ith}$, $\widehat{\sigma}^2_{\varepsilon h} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\sigma}^2_{\varepsilon ih}$, $\widehat{\omega}_{\varepsilon ih} = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}^4_{ith}$, $\widehat{\omega}_{\varepsilon h} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\omega}_{\varepsilon ih}$, $\widehat{\psi}_{\varepsilon ih} = \widehat{\omega}_{\varepsilon ih} - \widehat{\sigma}^4_{\varepsilon ih}$, $\widehat{\psi}_{\varepsilon h} = \widehat{\omega}_{\varepsilon h} - \widehat{\sigma}^4_{\varepsilon h}$, $\widetilde{\sigma}^2_{\varepsilon ih} = \left(1-\frac{1}{n}\right)^{-2}\widehat{\sigma}^2_{\varepsilon ih} - \left(1-\frac{1}{n}\right)^{-2}\frac{1}{n^2}\sum_{j\neq i}^{n}\widehat{\sigma}^2_{\varepsilon jh}$, $\widetilde{\sigma}^2_{\varepsilon h} = \left(1-\frac{1}{n}\right)^{-2}\widehat{\sigma}^2_{\varepsilon h} - \frac{1}{n}\left(1-\frac{1}{n}\right)^{-1}\widehat{\sigma}^2_{\varepsilon h}$, $\widehat{\phi}_{1ih} = 6(1-\frac{1}{n})^2\left[\widetilde{\sigma}^2_{\varepsilon ih}\left(\frac{1}{n}\sum_{j\neq i}^{n}\widetilde{\sigma}^2_{\varepsilon jh}\right)\right]$, $\widehat{\phi}_{2ih} = \frac{1}{n^2}\sum_{j\neq i}^{n}\widehat{\omega}_{\varepsilon ih} + \frac{6}{n^2}\sum_{j\neq i}^{n}\sum_{k\neq i,j}^{n}\widetilde{\sigma}^2_{\varepsilon jh}\widetilde{\sigma}^2_{\varepsilon kh}$, $\widehat{\phi}_{1h} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\phi}_{1ih}$, $\widehat{\phi}_{2h} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\phi}_{2ih}$, $\widetilde{\psi}_{\varepsilon h} = \left(1-\frac{1}{n}\right)^{-4}\widehat{\psi}_{\varepsilon h} - \widehat{\gamma}_{hn}$, and $\widehat{\gamma}_{hn} = \frac{1}{n}\left[\widehat{\phi}_{1h} - 2\left(1-\frac{1}{n}\right)^{3}\widetilde{\sigma}^4_{\varepsilon h}\right] + \frac{1}{n^2}\left[\widehat{\phi}_{2h} + \left(1-\frac{1}{n}\right)^{2}\widetilde{\sigma}^4_{\varepsilon h}\right]$. The resulting test is presented in the following theorem.

Theorem 2. *Suppose Assumptions 1-4 hold. Then under the null hypothesis that $\sigma^2_{\varepsilon i} = \sigma^2_{\varepsilon}$ for all i,*

$$Z^o_{nT} = \frac{1}{s_{nT}}\sum_{i=1}^{n}\left\{\left[T\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}^2_{ith} - \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\varepsilon}^2_{ith}\right)^2\right] - (1-\frac{1}{n})^4\widetilde{\psi}_{\varepsilon h}\right\} \to_d N(0,1)$$

as $(n, T \to \infty)$ and $\frac{T}{n} \to 0$, where $s^2_{nT} = 2n\widetilde{\psi}^2_{\varepsilon h}$.

Theorem 2 has established the joint limit distribution of the statistic for testing the null hypothesis that $\sigma^2_{\varepsilon ih} = \sigma^2_{\varepsilon h}$ for all i. By construction, λ_{th} plays no role in the test since it is completely removed in the process of computing consistent estimates for ε_{ith}. Along with

[6] Indeed, it would be enough to test the null hypothesis that $\frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2_{\varepsilon ih} - \frac{1}{n}\sum_{i=1}^{n}\sigma^2_{\varepsilon ih}\right)^2 = o(T^{-1}n^{-1})$ for our purpose. But $\sigma^2_{\varepsilon ih}$ does not depend on T by Assumption 3. So a natural choice for the null would then be $\sigma^2_{\varepsilon ih} = \sigma^2_{\varepsilon h}$ for all i.

λ_{th}, all between- and within-forecaster correlations are also removed. Moreover, the impact of the nonstochastic bias μ_{it} is asymptotically negligible as implied by assumption. Thus, our test is essentially an unconditional test for homogeneity of idiosyncratic variances in large panel data framework. A rejection of the null hypothesis based on Theorem 2 can be interpreted as a signal of the need for using unequal weights in computing the measures of uncertainty. Future research is warranted in exploring an optimally weighted (over i) version of $RMSE_{LPS}$, which might be lower than the $RMSE_{LPS}$ based on a simple average.

Remark 5. *The statistical literature for testing equality of variances is huge. The most widely used procedure among these is an F test proposed by Levene (1960) in the form of the classic ANOVA method applied to the absolute differences between each observation and the mean of its group. Brown and Forsythe (1974) suggested using median instead of the mean, and this version of the Levene test has been found to have excellent power properties even under asymmetric distributions, see Gastwirth, et al. (2009). However, as Iachine, et al. (2010) have pointed out this family of tests assume independence of observations, and hence are not suitable in our context where forecast errors are sticky and correlated across forecasters due to common shocks.*

An alternative approach assumes that the individual variance ($\sigma^2_{\varepsilon ih}$) can be approximated by a function of covariates. Testing for homoscedasticity then reduces to a joint testing for zero coefficients using Lagrange Multiplier tests, see Baltagi, et al. (2006) and Baltagi, et al. (2010). These tests require a prior knowledge of what might be causing the heteroskedasticity, and have statistical power provided $\sigma^2_{\varepsilon ih}$ can be well explained by a few proxies. Since we have very little information on the characteristics of professional forecasters and how they make forecasts, this approach is not feasible in our case.

Remark 6. *It is clear from expression (9) that we are in essence testing the following null*

hypothesis

$$8\left(\sigma_{\lambda h}^2+\sigma_{\varepsilon h}^2\right)^{3/2}[\operatorname*{plim}_{(n,T\to\infty)} RMSE_{LPS}-\operatorname*{plim}_{(n,T\to\infty)} RMSE_{RT}]=0.$$

That is, we are testing the significance of the scaled difference between the asymptotic limits of $RMSE_{LPS}$ and $RMSE_{RT}$.

Remark 7. *The restriction that $\frac{T}{n}\to 0$ as $(n,T\to\infty)$ controls for the approximation errors in panel estimation and prevents them to have a non-trivial effect on the limit distribution. Moreover, from the proof of Theorem 2 in the online appendix, we see the presence of two bias terms of magnitude order $O_p(n^{-1/2})$ - one positive and one negative, and two positive bias terms of order $O_p(n^{-3/2})$.*

Remark 8. *Under the null hypothesis, the term $\left[T\left(\frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{ith}^2-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\varepsilon}_{ith}^2\right)^2\right]$ roughly follows a χ_1^2 distribution for large T and large n, leading to potential size distortions and slow convergence to standard normality due to its large positive skewness (close to $\sqrt{8}$).*

To address the bias and skewness issues pointed out in Remarks 7 and 8, we define $\widetilde{B}_{RT}=-(1-\frac{1}{n})^4\widetilde{B}_1+4(1-\frac{1}{n})^2\widetilde{B}_2+\widetilde{B}_3$, where $\widetilde{B}_1=n^{-1/2}\widetilde{\psi}_{\varepsilon h}$, $\widetilde{B}_2=n^{-1/2}(1-\frac{1}{n})^2\widetilde{\sigma}_{\varepsilon h}^4$, and $\widetilde{B}_3=3n^{-3/2}(1-\frac{1}{n})^2(1-\frac{2}{n})\widetilde{\sigma}_{\varepsilon h}^4+n^{-5/2}(1-\frac{1}{n})(\widetilde{\omega}_{\varepsilon h}-5\widetilde{\sigma}_{\varepsilon h}^4)$, and modify the test statistic proposed in Theorem 2. The result is then summarized in the following theorem.

Theorem 3. *Suppose Assumptions 1-4 hold. Then under the null hypothesis that $\sigma_{\varepsilon ih}^2=\sigma_{\varepsilon h}^2$ for all i,*

$$\begin{aligned} Z_{nT}^{bsc} &= \left\{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(1-\frac{1}{n})^4\widetilde{\psi}_{\varepsilon h}}\left[T(\frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_{ith}^2-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{\varepsilon}_{ith}^2)^2-\frac{1}{n^{1/2}}\widetilde{B}_{RT}\right]\right)^{1/3}-1+\frac{2}{9n}\right\}\left(\frac{2}{9n}\right)^{-1/2} \\ &\to_d \quad N(0,1) \end{aligned}$$

as $(n,T\to\infty)$ and $\frac{T}{n}\to 0$.

Remark 9. *The statistic in Theorem 3 reduces the asymptotic bias by subtracting the estimated means for the four bias terms discussed in Remark 7 and by scaling with the factor*

$(1-\frac{1}{n})^{-4}$. *A similar approach was adopted by Pesaran and Yamagata (2008) in their test of slope homogeneity in large random coefficient panel data models, see also Hsiao and Pesaran (2008). In addition, it addresses the issues of positive skewness and slow convergence by adopting the popular Wilson-Hilferty cube root transformation; see Chen and Deo (2004) for a general discussion on power transformation to tackling skewness and slow convergence problems.*

3.3 Monte Carlo Simulation

To assess the performance of our tests, we conduct Monte Carlo simulations. We consider all combinations of $T = 20, 60, 120$ and $n = 20, 60, 120$. Data are generated according to $e_{ith} = \lambda_{th} + \varepsilon_{ith}$.[7] Since λ_{th} plays no role in the test, for simplicity, it is generated as a moving average process of order one such that $\lambda_{th} = \xi_{th} - 0.5\xi_{(t-1)h}$, with $\xi_{th} \overset{iid}{\sim} U(-1, 1)$. ε_{ith} are randomly generated from either a normal distribution $N(0, \sigma^2_{\varepsilon ih})$ or a uniform distribution $U(-\sqrt{3}\sigma_{\varepsilon ih}, \sqrt{3}\sigma_{\varepsilon ih})$. To assess the size of our tests, we let $\sigma^2_{\varepsilon ih} = \sigma^2_{\varepsilon h}$ for all i and set $\sigma^2_{\varepsilon h} = 0.05, 0.25$, and 1.25, respectively. To evaluate the power, we first set the value for the average of idiosyncratic variances ($\sigma^2_{\varepsilon h}$), and then let $100r$ percentage of idiosyncratic variances differ from $\sigma^2_{\varepsilon h}$, with half of them greater and the other half smaller than $\sigma^2_{\varepsilon h}$. The magnitude of the difference is measured by $100p$ percentage. Our simulation design allows us to explore the effect of changes in r and p on the performance of the test statistics, as large values of r and/or p introduce increasing heterogeneity of idiosyncratic variances. In our simulation study, we consider all combinations of $r = 0.3, 0.5, 0.7$ and $p = 0.3, 0.5, 0.7$. For brevity, we report the results for $\sigma^2_{\varepsilon h} = 0.05$ only, since other values of $\sigma^2_{\varepsilon h}$ (namely, 0.25 and 1.25) yield very similar power. All results are obtained from $5,000$ replications.

Since the results for the original test in Theorem 2 are slightly inferior to those for

[7] We also consider the following data generating process $e_{ith} = \mu_{ith} + \lambda_{th} + \varepsilon_{ith}$ with $\mu_{ith} = O((nT)^{-1/2})$. We find that the size and power of our tests are almost identical to those of the process $e_{ith} = \lambda_{th} + \varepsilon_{ith}$, implying that the impact of μ_{ith} on our tests are insignificant and thus can be safely ignored.

the bias and skewness corrected test (Z_{nT}^{bsc}) in Theorem 3, and for the sake of brevity, we report only the simulation results for the latter. Table 1 summarizes the size of the test. When the idiosyncratic errors are assumed to be normally distributed, the Z_{nT}^{bsc} test yields good empirical size, though slightly oversized when $n = 20$. With a uniform distribution for the error terms, the test exhibits size distortions, especially for $n = 20$, but the size distortion becomes less as n increases to 60. Turning to the power, Table 2 shows that the Z_{nT}^{bsc} test becomes more powerful when either r or p increases. Recall that r indicates the fraction of heterogeneous idiosyncratic variances in the panel and p captures the deviation of the individual variances from the average of idiosyncratic variances on the whole. Taken together, r measures the relative amount of evidence against the null (or "patterns"), and p measures the overall amount of evidence against the null (or "strength"). Moreover, the power tends to increase when T and/or n increase for given values of r and p, which justifies our proposed test for the use in large panels. Finally, the Z_{nT}^{bsc} test performs better under a uniform distribution than a normal distribution for the idiosyncratic error terms.

4 Empirical Illustrations: Underestimation of Uncertainty in US GDP and Inflation Forecasts

In this section, we present estimates of historical uncertainty in inflation and output growth forecasts using $RMSE_{LPS}$, and compare it to $RMSE_{AF}$ and $RMSE_{RT}$. The use of various types of survey data in measuring forecast uncertainty is well elaborated in Pesaran and Weale (2006).

4.1 Survey of Professional Forecasters

The first data set used in this study to examine the alternative uncertainty estimates comes from the US Survey of Professional Forecasters (SPF) over 1991Q1 to 2017Q4. We focus on forecasts of GDP price deflator and real GDP growth, with horizon varying from one to five quarters. In order to calculate the forecast errors, we used the first-announced actual values in real time from the Real Time Data Set for Macroeconomists (RTDSM) provided by the Federal Reserve Bank of Philadelphia. The forecast data set fits our need well because it covers 90-100 forecasters over 108 quarters. The SPF is a quality-assured and widely used quarterly survey on macroeconomic forecasts in the United States. The American Statistical Association (ASA) and the National Bureau of Economic Research (NBER) initiated the survey in 1968Q4. Due to a rapidly declining participation rate in the late 1980s, the Federal Reserve Bank of Philadelphia took over the survey in 1990 with a new infusion of forecasters. Thus, in order to minimize the missing data problem, our sample starts from 1991Q1; even then nearly 70% of the potentially observable forecasts are unavailable, cf. Engelberg, et al. (2011).

The missing values pose a potential challenge for empirically implementing our test statistics since many of the asymptotic inequalities that we established are not necessarily valid in the context of incomplete panels. Following a lead from Genre, et al. (2013), we impute the missing values, but allow for uncertainty in inference due to missing data by multiple imputations (MIs). Using Markov chain Monte Carlo (MCMC) techniques, a predictive distribution of missing data conditional on observed forecasts is simulated leading to the creation of MIs, see Little and Rubin (2002). Our model of imputation for each variable and for each horizon is specified as a linear mixed-effects model

$$e_{ith} = \alpha + \beta e_{.th} + \gamma_i (e_{ih.} - e_{h..}) + \epsilon_{ith}, \tag{10}$$

where $e_{.th}$ is the average forecast error for period t made by the participating forecasters, $e_{ih.}$ is the average forecast error by forecaster i during the periods for which he/she forecasted, and the overall mean of forecast errors is $e_{h..}$. ϵ_{ith} is the error in the imputation equation. It is presumed that parts of e_{ith} are missing that we need to impute. Whereas β is specified as a fixed effect with expected value 1, $\gamma_i(e_{ih.} - e_{h..})$ is treated as random effects allowing for time invariant individual biases. Note that in Genre, et al. (2013), the second term in (10) is a function of recent average deviation of forecasts made by a forecaster from the mean forecasts.[8] However, as discussed in Lahiri, et al. (2017), due to excessive missing observations in SPF data, we took individual means instead. Since our aim is to fill in the missing values retrospectively for calculating the ex post RMSEs, we did not have to impute recursively in real time, even though our scheme in principle can allow for this. After each imputation, we replaced the right-hand-side variables based on the imputed data set, and the missing observations were imputed again. In this way, the three variables in equation (10) will be pairwise consistent. This is a sensible imputation scheme in our context since the original time series of mean forecasts will be preserved, and the structure of correlations in the forecast errors between and within individuals will be largely maintained. What is most noteworthy is that the mean squared forecast errors based on the original incomplete panels and the imputed data sets were very close.[9]

[8]Following Davies and Lahiri (1999), we also experimented with a number of alternative imputation schemes including using known lagged actuals and aggregate forecast revisions from last forecasts. But these variables were found to be redundant in specification (10).

[9]We did 100 imputations for each data set using packages *pan* and *mitml* in R (version 3.6.0). Specifically, the calculated 100 test statistics Z^o_{nT} and Z^{bsc}_{nT} from the imputed data sets are combined in such a way that they reflect the variabilities due to both within and between imputations, see Little and Rubin (2002, pp. 86-87). For asymptotically valid inference in this context, one needs the assumption that the missingness mechanism is ignorable or missing at random (MAR). The MAR assumption merely means that the mechanism generating missing values can be ignored while preforming statistical inference. Identifying the mechanism generating attrition is difficult in our case because we have very little information on the forecasters except for their past forecast performance and the number of quarters they have been responding. Capistrán and Timmermann (2009), Genre, et al. (2013) and Lahiri, et al. (2017) found little association between participation and performance. While the assumption of MAR is almost impossible to test, it does not seem to be unreasonable in this example, see Yucel (2011).

Tables 3 and 4 report various statistics for inflation and output growth forecasts respectively using multiple imputed data. Two points are worth noting. First, the RMSEs associated with output are uniformly higher compared to inflation due to a differential incidence of common shocks. Both the idiosyncratic and common shocks are more variable for GDP growth forecasts than those for inflation, and the latter for GDP is comparatively very high. This phenomenon, which makes real GDP growth a difficult variable to predict, has been documented by Lahiri and Sheng (2008) using a heterogeneous learning model. More importantly, as expected, for all five horizons and for both GDP growth and inflation, $RMSE_{RT}$ is less than $RMSE_{LPS}$, but the differences between these two measures are very small. Yet, these differences are statistically significant for almost all cases. To understand the latter finding, note that we are testing the significance of the scaled difference between the asymptotic limits of $RMSE_{LPS}$ and $RMSE_{RT}$, as noted in Remark 6. Indeed, the scaled differences range from 0.09 to 0.13 for inflation and from 0.12 to 0.36 for output growth forecasts, resulting in a rejection of the null hypothesis that the scaled difference (i.e., the variance of idiosyncratic variances) is zero by both the original and bias-corrected test statistics (Z^{o}_{nT} and Z^{bsc}_{nT}). The power of the tests comes from the fact that they hone into the individual forecast variances after netting out the more formidable variability of the common shocks in constructing the statistics.

Note that the RMSE figures that are reported by Reifschneider and Tulip (2019) and those in this paper are not directly comparable. RT used the simple averages of the individual projections in SPF, Blue Chip and FOMC panels, together with Greenbook, Congressional Budget Office (CBO) and the Administration forecasts giving $n = 6$ in their calculation. Specifically, their measure is expressed as $RMSE^{group} = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{T}\sum_{t=1}^{T}(A_t - F^{m}_{.t})^2}$, where $F^{m}_{.t}$ is the mean forecast for the group m, for the target year t and h-period ahead to the end of the target year. By averaging across individual projections, most of idiosyncratic differences and disagreement in FOMC, SPF and Blue Chip forecasts have inadvertently

been washed away. They found very little heterogeneity in these six forecasts. On the other hand, their simultaneous use of Greenbook, CBO, Adminstration, consensus FOMC, SPF, and Blue Chip forecasts meant that RT had to meticulously sort out important differences in the comparability of these six forecasts due to data coverage, timing of forecasts, reporting basis for projections, and forecast conditionality. Despite all these differences, these two sets of uncertainty estimates are very close in the context of SPF dataset. At least a part of the explanation for this similarity is due to the use of dataset from professional forecasters. For non-professionals, such as surveys of households, where the idiosyncratic errors are expected to be more heteroskedastic, we may see a substantial difference between RT and LPS uncertainty measures. Indeed, if the cross sectional variance of idiosyncratic error variances, defined as $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}e_{ith}^2-\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}e_{ith}^2\right)^2$, were to increase from 0.0004 to 0.004 at 1-quarter ahead inflation forecast, $RMSE_{RT}$ would decrease from 0.209 in Table 5 to 0.163, resulting in an underestimation of the correct benchmark uncertainty by 23%. Clements and Galvao (2017) compare RT measure against two *ex ante* uncertainty measures from survey forecasts and find that for both inflation and output growth at within-year horizon, RT uncertainty underestimates *ex ante* uncertainty measures.[10]

4.2 University of Michigan Survey of Consumers

To gain further insight into heterogeneous idiosyncratic errors, we conduct a separate experiment using data from the University of Michigan Survey of Consumers (MSC). We choose the Michigan survey since household expectations in this survey are often used in the macroeconomics literature; see, e.g. Carroll (2003), Ang, et al. (2007) and Coibion and Gorodnichenko

[10]Clark, et al. (2020) have compared the RT approach based on past errors with a stochastic multi-horizon volatility model of nowcasts and successive forecast updates, and found that the former yields incorrect coverage rates. However, with smaller rolling window sizes around 40 quarters, the two approaches gave comparable results. That said, RT measure was proposed not as a stand-alone measure of uncertainty, but rather as a historical benchmark against which the FOMC participants would form their own forward-looking evaluations and downside risks.

(2015). Each month households give their forecasts of price changes over the next 12 months.[11] In order to build a balanced panel, we followed Deaton (1985) to convert the repeated cross-sections MSC data to a pseudo panel. Thus, we classify each household into different cohorts according to their age (at five-year intervals), gender (male vs. female) and income (quartiles). Souleles (2004), Bruine de Bruin, et al. (2011) and Lahiri and Zhao (2016) provide mounting evidence on the heterogeneity in the household price expectations along these dimensions. Then we construct a pseudo balanced panel of 104 forecasters, with each of them calculated as the average inflation forecast in the corresponding age/gender/income cohort. To increase the number of observations for each cohort, we pool monthly observations for each quarter. The sample in this study comprises 153 quarterly surveys from the fourth quarter of 1979 through the fourth quarter of 2017. There are about 1,400 participants in each year/quarter and 13 participants in each cohort, on average.[12] For our purpose, the structure of heterogeneity should be maintained in the pseudo panel. Indeed, the correlation between disagreement from the pseudo panel and from the entire sample is about 0.79.

To further explore the heterogeneity across cohorts, in Table 5 we report the RMSE and test statistics. For both the whole sample period and various subsamples, we see substantial differences between $RMSE_{RT}$ and $RMSE_{LPS}$, and these differences are statistically significant at the 1% level. Depending on the sample period, $RMSE_{RT}$ underestimates the correct benchmark uncertainty by 2% to 14%. One potential concern with the above analysis is that there are not enough participants for each cohort for valid asymptotic inference. To address this issue, we drop the gender category and form cohorts by only age and income categories. Also, we now construct 6 age cohorts by ages 18-30, 31-40, 41-50, 51-60, 61-70,

[11] Specifically, households are first asked, "During the next 12 months, do you think that prices in general will go up, or go down, or stay where they are now?" If the respondent answers "go up" or "go down", point forecasts are requested: "By about what percent do you expect prices to go up/down on the average, during the next 12 months?"

[12] For 204 cohorts (accounting for about 1% of all cohorts) where there are no participants, we replace the missing value by the corresponding mean forecast across all participants in that year/quarter.

71 and above. Thus, we now have 24 cohorts (= 6 age cohorts × 4 income cohorts) in this alternative dataset, and there are about 58 participants in each cohort on average. Table 6 reports the RMSE and test statistics, and the results based on 24 cohorts are qualitatively the same as those based on 104 cohorts. This simple experiment confirms our conjecture that there exists substantial differences in the idiosyncratic forecast variances among households, and suggests the need to construct the correct benchmark uncertainty by incorporating heterogeneous individual error variances.

5 Concluding Remarks

A number of surveys of professional forecasters and households are regularly conducted in many countries around the world, and a widespread interest in these surveys suggests that the aggregate macroeconomic forecasts reported by these organizations are considered useful by policy makers, investors and other stakeholders. Even though it is now recognized in the forecasting profession that a point forecast by itself is of limited use and should be reported with an indication of the associated uncertainty, currently the consensus forecasts from these surveys are not reported with uncertainty bands.

The dominant methodology of forecast combination in econometrics is due to Bates and Granger (1969) whose basic criterion for optimal combination is based on minimizing the mean squared error of combined forecasts that rule out any consideration of the cross sectional distribution of forecasts. From the standpoint of a policy maker who has access to a number of expert forecasts, the uncertainty of a combined or ensemble forecast should be interpreted as that of a typical forecaster randomly drawn from the pool. This uncertainty formula should incorporate forecaster discord, as justified by (i) disagreement as a component of combined forecast uncertainty, (ii) the model averaging literature and (iii) central banks' communication of uncertainty via fan charts. This is not entirely a new idea, but

the asymptotic results that we have provided in this paper will help crystallize the role of forecaster disagreement in measuring uncertainty of combined forecast from the standpoint of a policy maker.

We have identified two layers of heterogeneity in individual forecast errors, arising from i) systematic individual biases, and ii) random individual errors with heteroskedasticity. We develop two new statistics to test the heterogeneity of idiosyncratic errors under the joint limits with both n and T approaching infinity simultaneously. We find significant heterogeneity in professional forecasters, which is due primarily to the heterogeneity in individual error variances. However, for this set of professional forecasters, the observed heterogeneity does not translate into a significant underestimation of true uncertainty if one uses the benchmark uncertainty formula suggested by Reifschneider and Tulip (2019). However, when we implement our test on the household inflation expectations, the cross sectional heterogeneity is found to be considerable, and perhaps not surprisingly, the RT formula significantly underestimates the theoretical value by as much as 10% for one-year ahead forecasts.

One potential concern in incorporating disagreement as part of aggregate uncertainty is that the prediction intervals will get wider, making inter-temporal movements in consensus forecasts less meaningful. Why would practitioners opt for enlarged confidence bands when they are less likely to obtain news-worthy results? The simple answer is that in the long run the reported forecasts will be more credible and the uncertainty measures better calibrated. As aptly put by Draper (1995) in his concluding remark, "which is worse - widening the bands now or missing the truth later?".

References

Amisano, G. and J. Geweke (2017). Prediction using several macroeconomic models. *Review of Economics and Statistics* 99, 912-925.

Andrade, P., R. Crump, S. Eusepi and E. Moench (2016). Fundamental disagreement. *Journal of Monetary Economics* 83, 106-128.

Ang, A., G. Bekaert and M. Wei (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics* 54, 1163-1212.

Baltagi, B.H., G. Bresson and A. Pirotte (2006). Joint LM test for heteroskedasticity in a one way error component model. *Journal of Econometrics* 134, 401-417.

Baltagi, B.H., B.C. Jung and S.H. Song (2010). Testing for heteroskedasticity and serial correlation in a random effects panel data model. *Journal of Econometrics* 154, 122-124.

Bates, J.M. and C.W.J. Granger (1969). The combination of forecasts. *Operational Research Quarterly* 20, 451-468.

Boero, G., J. Smith and K.F. Wallis (2008). Uncertainty and disagreement in economic prediction: the Bank of England survey of external forecasters. *Economic Journal* 118, 1107-1127.

Brown, M.B. and A.B. Forsythe (1974). Robust tests for equality of variances. *Journal of the American Statistical Association* 69, 364-367.

Bruine de Bruin, W., C.F. Manski, G. Topa, and W. van der Klaauw (2011). Measuring consumer uncertainty about future inflation. *Journal of Applied Econometrics* 26, 454-478.

Buckland, S.T., K.P. Burnham, and N.H. Augustin (1997). Model selection: an integral part of inference. *Biometrics* 53, 603-618.

Bunn, D. (1985). Statistical efficiency in the linear combination of forecasts. *International Journal of Forecasting* 1, 151-163.

Campbell, J.Y., M. Lettau, B.G. Malkiel and Y. Xu (2001). Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. *Journal of Finance* 56, 1-43.

Capistrán, C. and A. Timmermann (2009). Forecast combination with entry and exit of experts. *Journal of Business & Economic Statistics* 27, 429-440.

Carroll, C.D. (2003). Macroeconomic expectations of households and professional forecasters. *Quarterly Journal of Economics* 118, 269-298.

Chen, W.W. and R.S. Deo (2004). Power transformations to induce normality and their applications. *Journal of the Royal Statistical Society B* 66, 117-130.

Clark, T.E., M.W. McCracken, and E. Mertens (2020). Modeling time-varying uncertainty of multiple-horizon forecast errors. *Review of Economics and Statistics* 102, 17-33.

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

Clemen, R. and R. Winkler (1986). Combining economic forecasts. *Journal of Business and Economic Statistics* 4, 39-46.

Clements, M.P. (2014). Forecast uncertainty - ex ante and ex post: U.S. inflation and output growth. *Journal of Business & Economic Statistics* 32, 206-216.

Clements, M.P. and A.B. Galvao (2017). Model and survey estimates of the term structure of US macroeconomic uncertainty. *International Journal of Forecasting* 33, 591-604.

Coibion, O. and Y. Gorodnichenko (2015). Is the Phillips curve alive and well after all? Inflation expectations and the missing disinflation. *American Economic Journal: Macroeconomics* 7, 197-232.

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics* 30, 109-126.

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

Davies, A. and K. Lahiri (1999). Re-examining the rational expectations hypothesis using panel data on multiperiod forecasts. In Hsiao, C., K. Lahiri, L-F Lee, and H.M. Pesaran (eds.), *Analysis of Panels and Limited Dependent Variable Models*, Cambridge University Press, 226-254.

Diebold, F.X. and J.A. Lopez (1996). Forecast evaluation and combination. In Maddala, G.S. and C.R. Rao (eds.), *Handbook of Statistics*, volume 14, Amsterdam: North-Holland, 241-268.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society B* 57, 45-97.

Engelberg, J., C.F. Manski and J. Williams (2011). Assessing the temporal variation of macroeconomic forecasts by a panel of changing composition. *Journal of Applied Econometrics* 26, 1059-1078.

Gaglianone, W.P. and L.R. Lima (2012). Constructing density forecasts from quantile regressions. *Journal of Money, Credit and Banking* 44, 1589-1607.

Gastwirth, J.L., G.R. Gel and W. Miao (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Sciences* 24, 343-360.

Genre, V., G. Kenny, A. Meyler and A. Timmermann (2013). Combining expert forecasts: can anything beat the simple average? *International Journal of Forecasting* 29, 108-121.

Geweke J. and G. Amisano (2014). Analysis of variance for Bayesian inference. *Econometric Reviews*, 33, 270-288.

Giordani, P. and P. Söderlind (2003). Inflation forecast uncertainty. *European Economic Review* 47, 1037-1059.

Granger, C.W.J. (1989). Combining forecasts: twenty years later. *Journal of Forecasting* 8, 167-173.

Granger, C.W.J. and Y. Jeon (2004). Thick modeling. *Economic Modelling* 21, 323-343.

Gupta, S. and P. Wilton (1987). Combination of forecasts: an extension. *Management Science* 33, 356-372.

Hall, P. and C.C. Heyde (1980). *Martingale Limit Theory and its Application*. Academic Press, New York.

Hansen, B.E. (2008). Least squares forecast averaging. *Journal of Econometrics* 146, 342-350.

Hendry, D.F. and M.P. Clements (2004). Pooling of forecasts. *Econometrics Journal* 7, 1-31.

Hsiao, C. and M.H. Pesaran (2008). Random coefficient panel data models. In Matyas, L. and P. Sevestre (eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*. Springer, 3rd edition.

Iachine, I., H.C. Peterson, and K.O. Kyvik (2010). Robust tests for the equality of variances for clustered data. *Journal of Statistical Computation and Simulation* 80, 365-377.

Issler, J.V. and L.R. Lima (2009). A panel data approach to economic forecasting: the bias-corrected average forecast. *Journal of Econometrics* 152, 153-164.

Jo, S. and R. Sekkel (2019). Macroeconomic uncertainty through the lens of professional forecasters. *Journal of Business & Economic Statistics* 37, 436-446.

Jurado, K., S.C. Ludvigson and S. Ng (2015). Measuring uncertainty. *American Economic Review* 105, 1177-1216.

Lahiri, K. and X. Sheng (2008). Evolution of forecast disagreement in a Bayesian learning model. *Journal of Econometrics* 144, 325-340.

Lahiri, K. and X. Sheng (2010). Measuring forecast uncertainty by disagreement: the missing link. *Journal of Applied Econometrics* 25, 514-538.

Lahiri, K., H. Peng and X. Sheng (2020). Measuring Uncertainty of a Combined Forecast and Some Tests for Forecaster Heterogeneity. CESifo Working Paper No. 8810.

Lahiri, K., H. Peng and Y. Zhao (2017). On-line learning and forecast combination in unbalanced panels. *Econometric Reviews* 36, 257-288.

Lahiri, K. and Y. Zhao (2016). Determinants of consumer sentiment over business cycles: evidence from the US surveys of consumers. *Journal of Business Cycle Research* 12, 187-215.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Non Experimental Data*, John Wiley and Sons, Inc.

Levene, H. (1960). Robust tests for equality of variances. In Olkin, I. (ed.), *Contributions to Probability and Statistics* 278-292. Stanford University Press.

Little, R.J.A. and D.B. Rubin (2002). *Statistical Analysis with Missing Data*, 2nd Edition.

Makridakis, S. (1989). Why combining works? *International Journal of Forecasting* 5, 601-603.

Manski, C.F. (2011). Interpreting and combining heterogeneous survey forecasts. In Clements, M.P. and D.F. Hendry (eds.), *Oxford Handbook of Economic Forecasting* 457-472, Oxford University Press.

Newbold, P. and D.I. Harvey (2001). Forecast combination and encompassing. In Clements, M.P. and D.F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell, Oxford.

Ozturk, E. and X.S. Sheng (2018). Measuring global and country-specific uncertainty. *Journal of International Money and Finance* 88, 276-295.

Palm, F.C. and A. Zellner (1992). To combine or not to combine? Issues of combining forecasts. *Journal of Forecasting* 11, 687-701.

Patton, A.J. and A. Timmermann (2010). Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion. *Journal of Monetary Economics* 57, 803-820.

Patton, A.J. and A. Timmermann (2011). Predictability of output growth and inflation: a multi-horizon survey approach. *Journal of Business & Economic Statistics* 29, 397-410.

Pesaran, M.H. (1987). The limits to rational expectations. Basil Blackwell, Oxford.

Pesaran, M.H. and M. Weale (2006). Survey expectations. In Elliott, G., C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting* 1, 715-776, Elsevier.

Pesaran, M.H. and T. Yamagata (2008). Testing slope homogeneity in large panels. *Journal of Econometrics* 142, 50-93.

Phillips, P.C.B. and H.R. Moon (1999). Linear regression limit theory for nonstationary panel data. *Econometrica* 67, 1057-1111.

Reifschneider, D. and P. Tulip (2019). Gauging the uncertainty of the economic outlook using historical forecasting errors: The Federal Reserve's approach. *International Journal of Forecasting* 35, 1564-1582.

Rossi, B. and T. Sekhposyan (2015). Macroeconomic uncertainty indices based on nowcast and forecast error distributions. *American Economic Review* 105, 650-655.

Smith, J. and K.F. Wallis (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics* 71, 331-355.

Souleles, N.S. (2004). Expectations, heterogeneous forecast errors, and consumption: micro evidence from the Michigan consumer sentiment surveys. *Journal of Money, Credit and Banking* 36, 39-72.

Timmermann, A. (2006). Forecast combinations. In Elliott, G., C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*. Elsevier, 135-196.

Wei, X. and Y. Yang (2012). Robust forecast combinations. *Journal of Econometrics* 166, 224-236.

Yucel, R.M. (2011). Inference by multiple imputation under random coefficients and random covariances model. *Statistical Modelling* 11, 351-370.

Table 1: Size of Z_{nT}^{bsc} test

		$\sigma_\varepsilon^2 = 0.05$			$\sigma_\varepsilon^2 = 0.25$			$\sigma_\varepsilon^2 = 1.25$		
		n=20	n=60	n=120	n=20	n=60	n=120	n=20	n=60	n=120
DGP I	T=20	0.065	0.049	0.043	0.067	0.046	0.047	0.066	0.047	0.041
	T=60	0.075	0.052	0.054	0.074	0.051	0.051	0.076	0.053	0.051
	T=120	0.078	0.059	0.051	0.080	0.057	0.050	0.074	0.058	0.054
DGP II	T=20	0.136	0.061	0.055	0.137	0.065	0.054	0.138	0.061	0.052
	T=60	0.143	0.067	0.056	0.147	0.070	0.055	0.137	0.067	0.058
	T=120	0.148	0.073	0.062	0.138	0.066	0.060	0.142	0.070	0.056

Note: Rejection rates of Z_{nT}^{bsc} test under $H_0 : \sigma_{\varepsilon i}^2 = \sigma_\varepsilon^2$ for all i at the 5% nominal level based on two-sided $N(0,1)$ test and 5000 replications. We consider all combinations of $T = 20, 60, 120$ and $n = 20, 60, 120$. Data are generated according to $e_{it} = \lambda_t + \varepsilon_{it}$. λ_t is generated as a moving average process of order one such that $\lambda_t = \xi_t - 0.5\xi_{t-1}$, with $\xi_t \overset{iid}{\sim} U(-1, 1)$. ε_{it} are randomly generated from either a normal distribution $N(0, \sigma_{\varepsilon i}^2)$ under DGP 1, or a uniform distribution $U(-\sqrt{3}\sigma_{\varepsilon i}, \sqrt{3}\sigma_{\varepsilon i})$ under DGP 2. To assess the size of our tests, we let $\sigma_{\varepsilon i}^2 = \sigma_\varepsilon^2$ for all i and set $\sigma_\varepsilon^2 = 0.05, 0.25$, and 1.25, respectively.

Table 2: Power of Z_{nT}^{bsc} test

			$r=0.3$			$r=0.5$			$r=0.7$		
			n=20	n=60	n=120	n=20	n=60	n=120	n=20	n=60	n=120
DGP I	$p=0.3$	T=20	0.16	0.25	0.42	0.26	0.50	0.77	0.37	0.73	0.95
		T=60	0.55	0.92	1.00	0.85	1.00	1.00	0.96	1.00	1.00
		T=120	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$p=0.5$	T=20	0.43	0.83	0.99	0.74	0.99	1.00	0.91	1.00	1.00
		T=60	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		T=120	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$p=0.7$	T=20	0.82	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
		T=60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		T=120	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DGP II	$p=0.3$	T=20	0.50	0.79	0.97	0.75	0.98	1.00	0.90	1.00	1.00
		T=60	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		T=120	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$p=0.5$	T=20	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		T=60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		T=120	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$p=0.7$	T=20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		T=60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		T=120	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: See Table 1. Under DGP I, $\varepsilon_{it} \sim N(0, \sigma^2_{\varepsilon_i})$; under DGP II, $\varepsilon_{it} \sim U(-\sqrt{3}\sigma_{\varepsilon_i}, \sqrt{3}\sigma_{\varepsilon_i})$, where $\sigma^2_{\varepsilon} = 0.05$. r measures the percentage of $\sigma^2_{\varepsilon_i}$ that differ from σ^2_{ε}. p measures the magnitude of the deviation of $\sigma^2_{\varepsilon_i}$ from σ^2_{ε} on the whole.

Table 3: Measures of historical uncertainty in SPF inflation forecasts

Horizon	$RMSE_{AF}$	$RMSE_{RT}$	$RMSE_{LPS}$	Z^{o}_{nT}	Z^{bsc}_{nT}
1-quarter ahead	0.830	1.158	1.167	4.340***	4.278***
2-quarter ahead	0.923	1.135	1.140	3.857***	3.888***
3-quarter ahead	0.977	1.187	1.196	3.786***	3.971***
4-quarter ahead	1.002	1.215	1.226	3.564***	3.776***
5-quarter ahead	1.064	1.295	1.305	2.196**	2.258**

Note: $RMSE_{AF}$ is the conventional uncertainty measure in equation (7), $RMSE_{RT}$ is the Reifschneider and Tulip (2019)'s uncertainty measure in equation (8) and $RMSE_{LPS}$ is our suggested uncertainty measure in equation (6). In testing the null hypothesis that $RMSE_{RT}$ is the same as $RMSE_{LPS}$, the original test statistic Z^{o}_{nT} is defined in Theorem 2, and Z^{bsc}_{nT} is the bias and skewness corrected test statistic as defined in Theorem 3. The actual inflation rate for 1991-2017 is taken from the first quarterly release of Federal Reserve Bank of Philadelphia "real-time" data set. The inflation forecasts are taken from the Survey of Professional Forecasters from 1991:Q1 until 2017:Q3. *** and ** indicate significance at the 1% and 5% level, respectively.

Table 4: Measures of historical uncertainty in SPF output growth forecasts

Horizon	$RMSE_{AF}$	$RMSE_{RT}$	$RMSE_{LPS}$	Z^{o}_{nT}	Z^{bsc}_{nT}
1-quarter ahead	1.402	1.638	1.642	3.558***	3.511***
2-quarter ahead	1.738	1.920	1.922	3.744***	3.716***
3-quarter ahead	1.913	2.080	2.082	2.612***	2.591***
4-quarter ahead	2.055	2.245	2.249	-0.107	-0.003
5-quarter ahead	2.104	2.269	2.272	0.566	0.671

Note: See Table 3. The actual output growth rate for 1991-2017 is taken from the first quarterly release of Federal Reserve Bank of Philadelphia "real-time" data set. The output growth forecasts used in this study are taken from the Survey of Professional Forecasters from 1991:Q1 until 2017:Q3. *** indicates significance at the 1% level.

Table 5: Measures of historical uncertainty in MSC inflation forecasts from 104 cohorts

Survey period	$RMSE_{AF}$	$RMSE_{RT}$	$RMSE_{LPS}$	Z^{o}_{nT}	Z^{bsc}_{nT}
1979Q4-1989Q4	1.39	3.07	3.33	8.27***	6.55***
1990Q1-1999Q4	1.28	2.51	2.92	6.08***	5.14***
2000Q1-2009Q4	2.13	2.66	2.72	20.93***	12.81***
2010Q1-2017Q4	2.31	2.70	2.78	8.54***	6.70***
Whole sample	1.79	2.82	2.96	13.62***	9.52***

Note: $RMSE_{AF}$ is the conventional uncertainty measure in equation (7), $RMSE_{RT}$ is the Reifschneider and Tulip (2019)'s uncertainty measure in equation (8) and $RMSE_{LPS}$ is our suggested uncertainty measure in equation (6). In testing the null hypothesis that $RMSE_{RT}$ is the same as $RMSE_{LPS}$, the original test statistic Z^{o}_{nT} is defined in Theorem 2, and Z^{bsc}_{nT} is the bias and skewness corrected test statistic as defined in Theorem 3. The inflation forecasts of households are taken from University of Michigan Survey of Consumers (MSC) forecast of price changes over the next 12 months. The actual inflation rate is calculated as the annual percentage change in the Consumer Price Index for All Urban Consumers. The pseudo-balanced panel includes 104 forecasters by dividing the survey participants into 104 cohorts by their age/gender/income from the fourth quarter of 1979 through the fourth quarter of 2017. *** indicates significance at the 1% level.

Table 6: Measures of historical uncertainty in MSC inflation forecasts from 24 cohorts

Survey period	$RMSE_{AF}$	$RMSE_{RT}$	$RMSE_{LPS}$	Z^{o}_{nT}	Z^{bsc}_{nT}
1979Q4-1989Q4	1.41	2.08	2.16	10.44***	6.96***
1990Q1-1999Q4	1.34	1.71	1.89	15.68***	8.98***
2000Q1-2009Q4	2.15	2.30	2.33	6.51***	5.07***
2010Q1-2017Q4	2.30	2.38	2.44	10.52***	6.97***
Whole sample	1.81	2.13	2.20	22.45***	11.08***

Note: See Table 5. The pseudo-balanced panel includes 24 forecasters by dividing the survey participants into 24 cohorts by their age/income from the fourth quarter of 1979 through the fourth quarter of 2017. *** indicates significance at the 1% level.