Supplementary Material

Overprecision in the Survey of Professional Forecasters

Jump to:

FEO Analysis Plan and Results (for the Validation Data)

December 27, 2023

Timeline of events:

1. In 2019, we downloaded and split the data into two randomly selected halves: exploration and validation datasets. On September 9th, 2019, we finalized a pre-registration for the exploration dataset, specifying 48 planned tests: https://osf.io/q6x47/. In this pre-registration, we specified that we would only retain analyses and publish results that were consistent across different economic indicators and analytic specifications.
2. Between September 2019 and June 2023, we ran our pre-registered analyses on our exploration dataset. We refined and altered existing analyses, and added additional analyses in response to peer feedback across the years. The changes and additions made to the exploration pre-registration, as well as the results of all analyses conducted, appear in our Supplementary Materials, which can be found on the project OSF page (https://osf.io/q6x47/).
3. In June 2023, we submitted a manuscript containing the explorations results to *Collabra:Psychology*. Following editor and reviewer feedback, we are creating an updated pre-registration for the validation dataset. In line with our pre-registration for the exploration dataset, we retain only analyses that were consistent across different economic indicators and analytic specifications. Using our validation dataset, we will test whether the results hold up in the validation data.
    a. As noted in the manuscript submitted for review, we will supplement our validation dataset (includes data from 1968-2019 Q2) with forecasts that have been made since we last downloaded the data in 2019 (i.e., data from 2019 Q3 – 2023 Q4). We will report results from this combined dataset in the manuscript.
4. This document represents our pre-registration for the validation dataset. We split the analyses into two main groups:
    a. Pre-registered Analyses: Analyses that were consistent across different economic indicators and analytic specifications in the exploration dataset, and that we expect to replicate in the validation dataset.
    b. Exploratory Analyses: Analyses that were not consistent in the exploration dataset, but that we or reviewers nevertheless found interesting enough to justify retaining.

To aid reader comprehension, we retain changes made to the analyses pre-registered for the exploration dataset in red below. Blue text reports results for the training data set. For ease of comparison, we also retain the same subheadings (i.e., Part I, Part II, etc.). Again, when we are uncertain regarding the best way to test a question, we attempt to conduct the most reasonable variants, and only retain a finding as significant if it is statistically significant across all variations. Where appropriate, we intend to treat Forecaster ID, Year Forecast Made, and Year Being Forecast as random effects given that we are interested in generalizing beyond the existing data. Only results that emerge consistently and significantly according to the $p < .005$ standard (Benjamin et al., 2018) in both the exploration and validation samples will be retained in the manuscript.

Pre-Registered Analyses

PART I: Overprecision Analyses

1. Are forecasts overprecise?

   We employ three measures of overprecision (hit rate, Gini, and variance), and only retain a finding as significant if it is statistically significant for all three measures.

   a. Hit Rate Analysis: Peak confidence vs. hit rate. Focus on the bin(s) to which the forecaster assigns highest probability. What is the average probability? How does this compare with the rate at which they are correct? Conduct a paired t-test at the level of the forecaster comparing confidence (a probability measured for each forecast) with the hit rate.

      i. The paired t-test at the level of the forecaster comparing peak confidence with hit rate is significant; average peak confidence (0.53) is higher than average hit rate (0.23), $t = 19.741$, $df = 375$, $p < 2.2 \times 10^{-16}$.

   b. Gini Coefficient[1] Analysis: Compute a Gini coefficient for each forecast. Conduct a one-sample t-test comparing Gini coefficient, averaged within forecaster with the average Gini of realized outcomes across the entire epoch covered by the data.

      i. The t-test comparing the Gini of each forecast (ADJ_GINI), averaged within forecaster, with average Gini of the actuals across the entire epoch covered by the data (act_ADJ_GINI) is significant; the concentration of the forecasts (0.81) is higher than the concentration of the actuals (0.51), $t = 54.137$, $df = 375$, $p < 2.2 \times 10^{-16}$. This suggests that forecasts are overprecise.

   c. Variance Analysis: Compute the variance of each forecast, by computing the variance of the distribution. To do this, compute the distribution's mean, then sum the squared distance to each bin, weighted by the probability assigned to it. Conduct a one-sample t-test comparing variance, averaged within forecaster, with the average variance of realized outcomes across the entire epoch covered by the data.

      i. The t-test comparing the variance of each forecast (pred_var), averaged within forecaster, with average variance of the actuals across the entire epoch covered by the data (act_var) is significant; the average variance of the forecasts (1.34) is lower than the average variance of the actuals (5.85), $t = -53.904$, $df = 375$, $p < 2.2 \times 10^{-16}$. This suggests that forecasts are overprecise.

Using three measures of overprecision, we find that forecasters are overprecise. When the indicators are split up and *t*-tests are run for each of the indicators separately, these results hold.[2] For those interested,

[1] The Gini (1912) coefficient quantifies the concentration in a distribution (Gastwirth, 1972). Its most famous application is to economic inequality, where greater wealth concentration within a nation produces a larger Gini coefficient (Lorenz, 1905). However, its application to the concentration of a probability distribution is straightforward and has different virtues than computing its variance. For instance, a bimodal distribution could have substantial variance but still be concentrated with high probability in two locations. The Gini coefficient would reflect this concentration but variance would not.

[2] The exception to our finding of overprecision is Gini for Unemployment when you split indicators even further by yearspan (Unemp 2009Q2-2013Q4). Why? We decided against using only 5 actuals (from 2009-2013) because it's a very small sample size. For each calculation of Gini, we used the entire epoch of available data. When the binning arrangements change, more or less actuals can fall into a particular bin. 49 of the actuals fall into bin 10 (that is, unemployment < 6%) for Unemployment 2009-2013. That means that the concentration is high. The forecast data (for Gini, we're ordering by concentration) reveal that the forecasters are also pretty concentrated, but not necessarily in bin 10. Their forecasts are concentrated in the bin 2-6 range. But because it's Gini and its measuring concentration, the two Ginis end up being similar.

the results (both aggregated and split up by indicator) would hold under the Bonferroni adjustment as well (.005/48 = .0001); we did not plan a Bonferroni adjustment because our tests are not independent.

In the interests of interpretability and readability, we will only report the hit rate analyses in the manuscript. Will illustrate it with a figure that compares confidence (on the x-axis) with hit rate (on the y-axis), averaged by bin, for all forecasts.

Notes*: We combined the measures of GDP (Nominal GNP, Real GNP, and Real GDP). Tied peaks are scored proportionally (e.g., When a forecaster reports 50% confidence in each of two bins, a hit in either one of them yields a 50% hit).

Note on cleaning*: Because we use bin midpoints, which change according to the yearspan time periods specified in SPF Documentation, we created a "yearspan" variable to go along with "indicator."

PART II: Forecast accuracy

2. Do forecasts get better over time? Presumably technological and statistical advancements lead to better models.
    a. Conduct a regression analysis at the level of the forecast with accuracy (QSR) as the dependent variable. Independent variable is the year forecast made. Include ~~fixed~~ random effects for forecaster and year being forecast. The hypothesis predicts a significant positive coefficient on year.
        i. We find a significant positive coefficient on year; with each increase in year, QSR score goes up by .13 (t = 34.59, df = 10170, $p < 2 \times 10^{-16}$).
    b. How does overprecision change as a function of time? Conduct a regression with overprecision measures as the dependent variable. Independent variable is the year forecast made. Add random effects for forecaster and year being forecast.
        i. We find that with every increase in year, average peak confidence increases by about 0.06 (t (114300) = 46.84, $p < 10^{-16}$), average variance decreases by about 0.19 (t (850.41) = -17.66, $p < 10^{-16}$), and average Gini increases by about 0.05 (t (14780) = 49.41, $p < 10^{-16}$).

3. Does accuracy go up as the forecast distance shrinks?
    a. Conduct a regression on forecast accuracy predicted by quarters distance from the moment of truth. Include ~~fixed~~ random effects for forecaster and year being forecast.
        i. As the distance to the moment of truth increases, QSR decreases by about 0.14 (t = -38.17, df = 16500, $p < 2 \times 10^{-16}$). This implies that accuracy decreases the further away from the moment of truth the forecast is.
    b. How does overprecision change as a function of forecast distance? Conduct a regression with overprecision measures as the dependent variable. Independent variable is the distance to the moment of truth. Add random effects for forecaster and year being forecast.
        i. We find that as the quarter distance to the moment of truth increases, average peak confidence decreases by about 0.08 (t (16270) = -57.12, $p < 10^{-16}$), average

variance increases by about 0.24 (t (15940) = 20.88, $p < 10^{-16}$), and average Gini decreases by about 0.05 (t (16090) = -58.08, $p < 10^{-16}$).

PART IV: Individual Level Analyses

4. To what degree are there stable individual differences between forecasters with respect to confidence? Are some forecasters more consistently over-precise than others? Are some forecasters more consistently optimistic than others?
    a. Test how much of the variance in forecaster confidence is accounted for by stable differences between forecasters. Conduct an ANOVA to test the change in R-squared when ~~fixed~~ random effects for forecaster and year forecast made are included in a regression. Perform the tests using the different measures of confidence:
        i. Overprecision
            1. Conduct the test using peak confidence
            2. Conduct the test using Gini index of the forecast
            3. Conduct the test using forecast Variance

The three ANOVAs are significant (peak confidence null vs. full mod: $\chi^2 = 5822.6$, $df = 2$, $p < 2.2 \times 10^{-16}$; Gini: $\chi^2 = 7239.5$, $df = 2$, $p < 2.2 \times 10^{-16}$; Variance: $\chi^2 = 3363.6$, $df = 2$, $p < 2.2 \times 10^{-16}$), suggesting that including random effects for forecaster and year forecast made in the regression improves the null model.

    b. Test how much of the variance in forecaster confidence is accounted for by differences between time periods (i.e., whether some time periods are more predictable than others). Conduct an ANOVA to test the change in R-squared when ~~fixed~~ random effects for year being forecast and year forecast made are included in a regression. Perform the tests using the different measures of confidence:
        i. Overprecision
            1. Conduct the test using peak confidence
            2. Conduct the test using Gini index of the forecast
            3. Conduct the test using forecast Variance

The three ANOVAs are significant (peak confidence null vs. full mod: $\chi^2 = 2015.3$, $df = 2$, $p < 2.2 \times 10^{-16}$; Gini: $\chi^2 = 1345.2$, $df = 2$, $p < 2.2 \times 10^{-16}$; Variance: $\chi^2 = 1757.5$, $df = 2$, $p < 2.2 \times 10^{-16}$), suggesting that including random effects for year being forecast and year forecast made in the regression improves the null model.

PART V: Analyses within forecasters across time

5. How does experience affect future forecasts? Does more experience forecasting (as measured by the number of prior forecasts a forecaster has made in the Survey of Professional Forecasters) improve the accuracy of forecasts, using different measures:
    a. Conduct a regression using QSR as the dependent measure. The key independent variable is the number of prior forecasts in the data prior to the year the forecast was made (e.g., a rolling "number of prior forecasts"). The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.

i. For each additional prior forecast, QSR goes up by about .0025, $t = 21.56$, $df = 2405$, $p < 2 \times 10^{-16}$.

b. Test for the effect of experience on overprecision (all three measures):
   i. Conduct a regression using peak confidence as the dependent measure. The key independent variable is the number of prior forecasts in the data. The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.
   ii. Conduct a regression using Gini index as the dependent measure. The key independent variable is the number of prior forecasts in the data. The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.
   iii. Conduct a regression using variance as the dependent measure. The key independent variable is the number of prior forecasts in the data. The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.

   With every additional prior forecast in the regressions with RE, peak confidence goes up by .001, $t = 30.93$, $df = 8426$, $p < 2 \times 10^{-16}$, Gini goes up by .0009, $t = 31.32$, $df = 8948$, $p < 2 \times 10^{-16}$, and variance goes down by .004, $t = -10.82$, $df = 1444$, $p < 2 \times 10^{-16}$. It appears that forecasting experience both increases accuracy and confidence.

PART VI: Wisdom of Crowds

6. Is the crowd wiser than the individual? If you averaged the forecasters' forecasts together (in essence making it the average of a 'crowd'), are the estimates more accurate, as measured by:
   a. Calibration in its precision: Measure forecast error as the summed squared distance, weighted by probability, between a histogram forecast and the actual outcome. Conduct a paired t-test (in which the unit of analysis is the year being forecast) comparing the average forecast error ~~for the average forecast~~ with the ~~averaged~~ error of the ~~individual~~ averaged forecasts.
      i. The average of the errors (1.097) is significantly larger than the error of the average (0.928), $t = 8.162$, $df = 115$, $p = 4.697 \times 10^{-13}$.

PART VII: Additional Analyses

7. Lazy subjects and fatigue: Determine the order of the questions on the survey. If fatigue (e.g., subjects become lazier) is part of why we observe what appears to be overprecision, then we should observe more overprecision on questions later in the survey. Compare the precision (all three measures) of forecasters who skipped GDP and answered Unemployment against forecasters who completed both using a t test of two means; the former should be less fatigued given that the fewer questions you answer, the less fatigued you should be.
   a. Our results differ from the exploration dataset, in which we found that none of the three t-tests are significant.
      i. A t-test comparing peak confidence (on unemployment) of those who completed GDP (mean = 0.55) vs. those who skipped GDP (0.53) finds support that fatigue may contribute to overprecision ($t = 3.548$, $df = 2929$, $p = 0.00039$).
      ii. A t-test comparing Gini (on unemployment) of those who completed GDP (mean = 0.83) vs. those who skipped GDP (0.81) finds support that fatigue may contribute to overprecision ($t = 3.401$, $df = 2724.1$, $p = 0.00068$).
      iii. A t-test comparing variance (on unemployment) of those who completed GDP (mean = 0.62) vs. those who skipped GDP (0.36) does not find support that fatigue may contribute to overprecision ($t = 10.821$, $df = 2889.4$, $p < 2 \times 10^{-16}$).

While it is unlikely that all forecasters are equally diligent (Glas & Hartmann, 2018), we find mixed evidence that the overprecision we document is attributable to forecaster fatigue or laziness. In the exploration data, we compare precision of forecasters who did not complete initial sections of the survey with those who completed both initial and later sections, and find no evidence that forecasts become more overprecise when forecasters are fatigued as they work through the SPF questionnaire, t (2124.1) = 0.597, p = 0.551. However, as noted above, we do find some evidence of fatigue in the validation data, for the peak confidence and Gini measures. In line with our pre-registration, we relegate these analyses to the supplement.

8. Round numbers: Laziness and fatigue may increase the use of round numbers. Uncertainty should decrease precision; fatigue might plausibly increase precision. Create a new variable that assigns a point for every bin probability that ends in a 5, and two points for every bin probability that ends in a 0 (do not include bins with the number 0), and test whether the rows with more round numbers are more precise. Correlate the column with the new variable (higher value = more round numbers) against the column with peak confidence.
    a. The correlation is -0.30, t = -40.482, df = 16557, $p < 2.2 \times 10^{-16}$, suggesting that the more round numbers they used, the less their precision was, implying uncertainty (e.g., uncertainty leads to a wider distribution and more round numbers).

9. Does the increase in accuracy with experience come about as a result of "bad" forecasters dropping out? For each forecast, include an indicator variable for whether it's the forecaster's last forecast recorded in the data. Then, use a logistic regression to test whether the rolling average prior forecast accuracy predicts dropout.
    a. Our results differ from the exploration dataset, in which we found that coefficient on rolling average prior forecast accuracy was not significant (z = -1.054, p = .292), suggesting that the increase in accuracy with experience does not come about as a result of "bad" forecasters dropping out.
        i. The logistic regression model was fitted to predict the binary outcome last_forecast using prev_qsr as the predictor. The coefficient on rolling average prior forecast accuracy (prev_qsr) is negative and significant (z = -1.1849, p = 5.06×10^{5}). For every one-unit increase in prev_qsr, the odds of last_forecast being 1 are multiplied by $e^{-1.1849}$ (i.e., the odss of last_forecast being 1 decrease by a factor of approximately 0.306). This multiplication factor is less than 1, indicating that the odds of a forecaster dropping out (i.e., their last_forecast being 1) decrease as their rolling average prior forecast accuracy increases (i.e., with increasing prev_qsr).

Exploratory Analyses

The following analyses are ones that yielded inconsistent or null results in the exploration dataset, but that audiences consistently inquire on. Thus, we plan to run these as exploratory analyses and include the results in our supplementary materials.

1. What is the effect of being right on precision?
    a. ~~Is overprecision tempered with by being wrong?~~ What is the effect of being right on overprecision?
        i. The effect of being ~~wrong~~ right on subsequent precision: Measure forecaster error ~~as the summed squared distance (weighted by probability) between the forecast and the actual~~ as the average of prior QSR scores. Conduct a regression predicting precision using peak confidence as the dependent variable. The independent variable is forecast error in a forecaster's prior forecast; ~~fixed~~ random effects for forecaster and year being forecast.
        ii. Conduct the above analysis using Gini index instead of peak confidence.
        iii. Conduct the above analysis using variance instead of Gini.

    The coefficient on prev_qsr is not significant in any of the models (peak confidence model t = -0.869, df = 7310.65, p = 0.385; Gini model t = -1.514, df = 6250, p = 0.13; variance model t = -1.593, df = 4775.95, p = 0.111) the average of prior QSR scores does not significantly predict any of the three measures of overprecision.

    b. The effect of being right on precision: Measure rightness as the rolling average probability attached to the actual outcome. Add ~~fixed~~ random effects for forecaster and year being forecast.
        i. Conduct the analysis with peak confidence
        ii. Conduct the analysis with Gini.
        iii. Conduct the analysis with variance.

    When the probability attached to the actual outcome increased, peak confidence increased by .15, t = 6.07, df = 6832, p = 1.35 x 10^{-9}, Gini increased by .08, t = 4.987, df = 6135, p = 6.31 x 10^{-7}, and variance decreased by .12, t = -6.764, df = 5604.90, p = 1.48 x 10^{-11}.

2. Are forecasters from the financial services industry different in their precision[3]?
    a. Conduct a regression at the level of the forecast, with ~~fixed~~ random effects for forecaster and year being forecast, in which peak confidence serves as the dependent measure and forecaster's industry (financial services or other) serves as the independent variable.
    b. Conduct the analysis above, except with forecast Gini as the dependent measure.
    c. Conduct the analysis above, using forecast variance as the dependent measure.

    Our results differ from the exploration dataset, in which all regressions showed that forecasters from the financial services industry were not significantly different in their precision, across all indicators and all measures.

[3] Note*: The validation pre-registration contains a typo – the authors should have taken out optimism from the question "Are forecasters from the financial services industry different in their precision or optimism?"

Forecasters from the financial services industry are not significantly different in their average peak confidence from those in non-financial services industry (t =-1.47, df = 1727, p = 0.14), but are significantly different in their average Gini (t = -2.649, df = 3024, p = 0. 0.00812) and average variance (t = -2.88, df = 748.09, p = 0.00409). Specifically, forecasters in the financial services industry have a lower average Gini (0.017 less than intercept, or non-financial services average Gini), and a lower average variance (0.203 less than intercept, or non-financial services average variance).

Because of the inconsistencies in results for different operationalizations of precision, we do not believe the data present a clear conclusion.

3. Fewer bins are less taxing: Find indicators that have a different number of bins, and compare peak confidence using a t test of two means.

   Within indicators, there are survey years during which RGDP has 10 bins, and ones in which RGDP has 11 bins. A t-test reveals that forecasters on average report significantly higher peak confidence with fewer bins (0.546 vs. 0.481), t = 12.862, df = 5222.1, $p < 2.2 \times 10^{-16}$.

   RGNP has 6 bins, and the average peak confidence is 0.665. NGNP has 15 bins, and the average peak confidence is 0.521. We compare forecasters' peak confidence when there are 10 bins (mean = 0.546) to when there are 6 (mean = 0.665) and find forecasters on average report significantly higher peak confidence with fewer bins (t = -16.991, df = 1953.9, $p < 2.2 \times 10^{-16}$).

   We compare forecasters' peak confidence when there are 11 bins (mean = 0.481) to when there are 6 (mean = 0.665) and find that forecasters on average report significantly higher peak confidence with fewer bins (t = -26.94, df = 1839.3, $p < 2.2 \times 10^{-16}$).

   We compare forecasters' peak confidence when there are 15 bins (mean = 0.521) to when there are 11 (mean = 0.481) and find that forecasters on average report significantly higher peak confidence with more bins (t = -6.515, df = 2422.4, $p = 8.793 \times 10^{-11}$).

   We compare forecasters' peak confidence when there are 15 bins (mean = 0.521) to when there are 10 (mean = 0.546) and find that forecasters on average report significantly higher peak confidence with fewer bins (t = 3.943, df = 2558.1, $p = 8.276 \times 10^{-5}$).

   This evidence is sufficiently mixed that it leaves us skeptical that fewer bins are less taxing and lead to less overprecision.

4. Compute a crowd forecast and test for calibration on the three measures of overprecision. Average all forecast probabilities for a given indicator, year forecast made, year being forecast, and bin arrangement to get the crowd forecast.[4]

[4] Note*: The validation pre-registration contains an error – this analysis should actually be in the section above ("pre-registered" not "exploratory" analyses), as we discuss it in the paper.

Our results differ from the exploration dataset for peak confidence, where we found evidence of underprecision. The average peak confidence is 41.9% and the average hit rate is 45.7%, t = -4.3208, df = 187, $p < 2.2 \times 10^{-16}$.

> In the crowd forecast, we do not find evidence of underprecision using our peak confidence measure. The average peak confidence is 42.01% and the average hit rate is 42.98%, t = -1.219, df = 241, $p < 0.224$. We find evidence of overprecision using our variance (average variance of forecasts = 1.41, average variance of actuals = 3.73, t = -24.13, df = 241, $p < 2.2 \times 10^{-16}$) and Gini measures (average Gini of forecasts = 0.72, average Gini of actuals = 0.53, t = 32.376, df = 241, $p < 2.2 \times 10^{-16)}$. This may be due to how the variance and Gini of the actuals are calculated.

FEO Analysis Plan and Results (for the Training Data)

This document outlines analyses of data from the Survey of Professional Forecasters. We previously split the data into two randomly selected halves, blocking to obtain proportional representation for all years forecasts made, but randomly splitting within years forecasts made. On September 9th, 2019, we finalized a document specifying 48 planned tests for evidence of overconfidence in the Survey of Professional Forecasters and posted that plan here: https://osf.io/q6x47/.

We expected to update our analysis plans after conducting these 48 tests on the training half of the data. This document presents the results of those tests as well as our updated analysis plan. Planned analyses appear in black. Changes to the original document appear in red below. Blue text reports results for the training data set. When we are uncertain regarding the best way to test a question, we attempt to conduct the most reasonable variants, and only retain a finding as significant if it is statistically significant across all variations. Where appropriate, we intend to treat Forecaster ID, Year Forecast Made, and Year Being Forecast as random effects given that we are interested in generalizing beyond the existing data.

We will conduct the 55 planned tests listed in this document on the second half of the data, the "validation" data set. If results emerge consistently and significantly according to the $p < .005$ standard (Benjamin et al., 2018) in both the exploration and validation samples, we plan to write them up and attempt to publish them. Results that did not make it into the main manuscript due to constraints on length or narrative relevance are reported here.

Note for reviewers: This document can also be found on our OSF page for this project, in the case that the red and blue colors that we use do not show up on submissions processed by the journal.

PART I: Overprecision Analyses

10. Are forecasts overprecise?

We employ three measures of overprecision (hit rate, Gini, and variance), and only retain a finding as significant if it is statistically significant for all three measures.

a. Hit Rate Analysis: Peak confidence vs. hit rate. Focus on the bin(s) to which the forecaster assigns highest probability. What is the average probability? How does this compare with the rate at which they are correct? Conduct a paired t-test at the level of the forecaster comparing confidence (a probability measured for each forecast) with the hit rate.

   i. The paired t-test at the level of the forecaster comparing peak confidence with hit rate is also significant; average peak confidence (0.53) is higher than average hit rate (0.28), $t = 17.488$, $df = 369$, $p < 2.2 \times 10^{-16}$.

b. Gini Coefficient[5] Analysis: Compute a Gini coefficient for each forecast. Conduct a one-sample t-test comparing Gini coefficient, averaged within forecaster with the average Gini of realized outcomes across the entire epoch covered by the data.

   i. The t-test comparing the Gini of each forecast (ADJ_GINI), averaged within forecaster, with average Gini of the actuals across the entire epoch covered by the data (act_ADJ_GINI) is significant; the concentration of the forecasts (0.82) is higher than the concentration of the actuals (0.51), $t = 54.398$, $df = 369$, $p < 2.2 \times 10^{-16}$. This suggests that forecasts are overprecise.

c. Variance Analysis: Compute the variance of each forecast, by computing the variance of the distribution. To do this, compute the distribution's mean, then sum the squared distance to each bin, weighted by the probability assigned to it. Conduct a one-sample t-test comparing variance, averaged within forecaster, with the average variance of realized outcomes across the entire epoch covered by the data.

   i. The t-test comparing the variance of each forecast (pred_var), averaged within forecaster, with average variance of the actuals across the entire epoch covered by the data (act_var) is significant; the average variance of the forecasts (1.27) is lower than the average variance of the actuals (5.85), $t = -52.921$, $df = 369$, $p < 2.2 \times 10^{-16}$. This suggests that forecasts are overprecise.

Using three measures of overprecision, we find that forecasters are overprecise. When the indicators are split up and *t*-tests are run for each of the indicators separately, these results hold.[6] For those interested,

[5] The Gini (1912) coefficient quantifies the concentration in a distribution (Gastwirth, 1972). Its most famous application is to economic inequality, where greater wealth concentration within a nation produces a larger Gini coefficient (Lorenz, 1905). However, its application to the concentration of a probability distribution is straightforward and has different virtues than computing its variance. For instance, a bimodal distribution could have substantial variance but still be concentrated with high probability in two locations. The Gini coefficient would reflect this concentration but variance would not.

[6] The exception to our finding of overprecision is Gini for Unemployment when you split indicators even further by yearspan (Unemp 2009Q2-2013Q4). Why? We decided against using only 5 actuals (from 2009-2013) because it's a very small sample size. For each calculation of Gini, we used the entire epoch of available data. When the binning arrangements change, more or less actuals can fall into a particular bin. 47 of the actuals fall into bin 10 (that is, unemployment < 6%) for Unemployment 2009-2013. That means that the concentration is high. The forecast data (for Gini, we're ordering by concentration) reveal that the forecasters are also pretty concentrated, but not necessarily in bin 10. Their forecasts are concentrated in the bin 2-6 range. But because it's Gini and its measuring concentration, the two Ginis end up being similar.

the results (both aggregated and split up by indicator) would hold under the Bonferroni adjustment as well (.005/48 = .0001); we did not plan a Bonferroni adjustment because our tests are not independent.

In the interests of interpretability and readability, we will only report the hit rate analyses in the manuscript. Will illustrate it with a figure that compares confidence (on the x-axis) with hit rate (on the y-axis), averaged by bin, for all forecasts.

Notes*: We combined the measures of GDP (Nominal GNP, Real GNP, and Real GDP). Tied peaks are scored proportionally (e.g., When a forecaster reports 50% confidence in each of two bins, a hit in either one of them yields a 50% hit).
Note about cleaning*: Because we use bin midpoints, we created a "yearspan" variable to go along with "indicator."

11. Situational influences on forecasts: Is there a correlation between overprecision and macroeconomic trends? Does forecast precision decrease following economic downturns?
    a. Conduct a regression at the level of the forecast. Peak confidence is the dependent measure. The key independent variable is a dummy variable for whether the prior year (given that recessions are recognized after the fact) to the year forecast made was a recession year. Include random effects for forecaster.
        i. When the prior year was a recession year, peak confidence decreased by about 1.8% (t = -5.319, df = 1.227 x 10^4, p = 1.06 x 10^{-7}); this suggests that forecasters were less precise following recession years.
    b. Conduct the above regression, using Gini as the dependent measure.
        i. When the prior year was a recession year, Gini decreased by about .01 (t = -4.824, df = 1.223 x 10^4, p = 1.42 x 10^{-6}). Concentration in forecast answers decreased, implying forecasters were less precise following recession years.
    c. Conduct the above regression, using variance as the dependent measure.
        i. When the prior year was a recession year, variance went down, but not significantly (t = -0.051, df = 1.199 x 10^4, p = 0.96).

We test whether forecast precision decreases following economic downturns, adding forecaster random effects, and do not find a result that holds across all three measures of overprecision. As stated at the beginning of this document, we therefore will not retain this as a publishable result.

Note on cleaning*: We used data from FRED to determine recession years. The data file can be found in the FRED Data folder on OSF. 0 counted as NO, 0.25-1.0 counted as YES.

PART II: Forecast accuracy

12. Score forecasts using the quadratic scoring rule *(14)*. That is: $1+2r-\sum p^2$ where r is the probability assigned to the right answer, and $\sum p^2$ is the sum of all squared probabilities (for all answers).
    a. Conduct an analysis comparing actual forecasts with simply predicting the base rate. For each forecast, compare (via paired t-test) the QSR for the actual forecast with QSR score had the forecaster forecasted the distribution of all rolling prior outcomes covered by the data. If this test is significant, conduct a regression on the difference in QSR scores for

the actual and the base rate forecasts, including random effects for forecaster and year being forecast. A systematic difference would show up as a significant intercept in the regression results.

i. The paired-test is significant (forecast QSR is better than base rate QSR, t = -31.929, df = 12358, $p < 2.2 \times 10^{-16}$), but the regression with random effects is not (t = 0.272, df = 50.958, p = 0.787). This suggests that adding Forecast ID and Year Being Forecast to the regression improves the model (i.e., the two variables have predictive power that explain the difference).

b. Conduct the same analysis as above, but replace the base rate with a rolling 5-year average. That is, the forecaster predicted the distribution in outcomes that occurred over the prior five years.

i. The paired t-test is significant (forecast QSR is better than rolling prior five year QSR, t = -23.614, df = 12358, $p < 2.2 \times 10^{-16}$), but the regression with RE is not (t = 0.821, df = 50.289, p = 0.415).

c. Conduct the same analysis as above, but use a repetition of last year. That is, the forecaster predicted with 100% confidence that last year's outcome would repeat itself.

i. The paired t-test is significant (forecast QSR is better than repetition of last year QSR, t = -84.088, df = 12358, $p < 2.2 \times 10^{-16}$), and the regression with RE is significant (t = 5.988, df = 49.763, $p = 2.32 \times 10^{-7}$). This suggests that forecaster QSR is significantly better than using the heuristic of simply predicting with 100% confidence that last year's outcome would repeat itself.

d. Conduct the same analysis as above, but use a uniform distribution. That is, the forecaster predicted that all the bins provided by in the survey are equally likely.

i. The paired t-test is significant (forecast QSR is better than uniform distribution QSR, t = -22.049, df = 12358, $p < 2.2 \times 10^{-16}$), but the regression with RE is not significant (t = 0.537, df = 51.419, p = 0.594).

Overall Finding: We find that all four paired t-tests are significant. As pre-registered, we run regressions on the difference score and look for a significant intercept. Only one of four models with random effects for forecaster and year being forecast find that forecast QSR is better than the heuristic QSR. Given that the raw means aren't far apart, adding forecaster ID and year being forecast to the regression can explain the variance. (The mean difference is largest for forecast QSR vs last year rep QSR (1.19 vs. 0.47); the other heuristic QSRs are around 1.)

As stated at the beginning of this document, we will report the results of the regressions with random effects, and will retain a significant finding only if it is significant for both the paired t-test and model with the random effects.

13. Do forecasts get better over time? Presumably technological and statistical advancements lead to better models.

a. Conduct a regression analysis at the level of the forecast with accuracy (QSR) as the dependent variable. Independent variable is the year forecast made. Include ~~fixed~~ random effects for forecaster and year being forecast. The hypothesis predicts a significant positive coefficient on year.

b. How does overprecision change as a function of time? Conduct a regression with overprecision measures as the dependent variable. Independent variable is the year forecast made. Add random effects for forecaster and year being forecast.
    i. We find a significant positive coefficient on year; with each increase in year, QSR score goes up by .11 (t = 24.69, df = 4149, $p < 2 \times 10^{-16}$).
    ii. We find that with every increase in year, average peak confidence increases by about 0.06 (t (1.028×10^4) = 39.78, $p < 10^{-16}$), average variance decreases by about 0.30 (t (7.801×10^3) = -33.50, $p < 10^{-16}$), and average Gini increases by about 0.05 (t (1.069×10^4) = 42.02, $p < 10^{-16}$).

14. Does accuracy go up as the forecast distance shrinks?
    a. Conduct a regression on forecast accuracy predicted by quarters distance from the moment of truth. Include ~~fixed~~ random effects for forecaster and year being forecast.
    b. How does overprecision change as a function of forecast distance? Conduct a regression with overprecision measures as the dependent variable. Independent variable is the distance to the moment of truth. Add random effects for forecaster and year being forecast.
        i. As the distance to the moment of truth increases, QSR decreases by about 0.12 (t = -26.55, df = 1.233×10^4, $p < 2 \times 10^{-16}$). This implies that accuracy decreases the further away from the moment of truth the forecast is.
        ii. We find that as the quarter distance to the moment of truth increases, average peak confidence decreases by about 0.07 (t (1.207×10^4) = -49.84, $p < 10^{-16}$), average variance increases by about 0.32 (t (1.189×10^4) = 37.20, $p < 10^{-16}$), and average Gini decreases by about 0.05 (t (1.178×10^4) = -50.49, $p < 10^{-16}$).

PART III: Overestimation Analyses

Conduct the following analyses on GDP and Unemployment. We omit inflation (CPI and PCE), as it is less clear how to define optimism on this measure. Is higher inflation good (for debtors)? Or is higher inflation bad (for lenders)?

15. Are forecasts overly optimistic or pessimistic?
    a. Optimism Point Prediction Analyses: Are forecasters' point predictions systematically optimistic or pessimistic?
        i. Conduct a paired t-test comparing point-prediction forecasts with outcomes. If this test is significant, conduct a regression on the difference between forecast and outcome, including ~~fixed~~ random effects for forecaster and year being forecast. A systematic difference would show up as a significant intercept in the regression results.

Unemployment

The paired t-test is significant, suggesting that forecasters are systematically pessimistic (the mean point prediction is 6.22 and actuals mean is 6.13, t = 10.432, df = 6889, $p < 2.2 \times 10^{-16}$); they think unemployment will be higher than it actually is. However, the regression with RE for forecaster and year being forecast is not significant, suggesting that forecasters are not systematically optimistic or pessimistic (t = .539, df = 51.761, p = 0.592).

GDP

The paired t-test is significant, suggesting that forecasters are systematically optimistic (mean point prediction is 2.64 and actuals mean is 2.36, $t = 14.716$, $df = 4265$, $p < 2.2 \times 10^{-16}$); they think GDP growth will be higher than it is. The regression with forecaster and year being forecast as random effects is not significant ($t = 0.727$, $df = 22.20$, $p = 0.475$). Consistent with our pre-registered criteria, we will not retain this result as significant.

Note on cleaning*: The point prediction datasets are different from the datasets that contain the full SPDs. For unemployment point predictions, we have 6910 observations after taking out rows for which there's no actual and no unemployment point prediction. For GDP, it's 4266 observations.

b. Optimism and pessimism in probability forecasts (histogram):
    i. Compute a weighted point prediction for each histogram forecast, treating each bin as its midpoint and weighting it by its assigned probability. Treat end bins as bounded by the historical max and min. Conduct a t-test comparing this weighted point-prediction forecast with outcomes. If this test is significant, conduct a regression on the difference between forecast and outcome, including ~~fixed~~ random effects for forecaster and year being forecast. A systematic difference would show up as a significant intercept in the regression results.

Unemployment

The paired t-test is significant, suggesting that forecasters' point predictions are systematically pessimistic (implied point prediction mean is 6.93 and actuals mean is 6.38, $t = 36.347$, $df = 2138$, $p < 2.2 \times 10^{-16}$). The regression with random effects for forecaster and year being forecast is significant ($t = 3.858$, $df = 10.558$, $p = 0.003$), and the intercept is positive, suggesting that forecasters are systematically pessimistic.

GDP

The paired t-test is significant, suggesting that forecasters' implied point predictions (i.e., histogram) are systematically pessimistic (implied point prediction mean is 2.49 and actuals mean is 2.58, $t = -5.18$, $df = 5691$, $p = 2.295 \times 10^{-7}$). This is not consistent with the previous result for GDP point prediction (paired t-test showed optimism for GDP point predictions); although the difference is modest, this inconsistency undermines our faith in the veracity of the result. The regression with forecaster and YBF as random effects is not significant ($t = -1.345$, $df = 36.659$, $p = 0.187$). As such, we will not treat this result as significant; forecasters' GDP implied point predictions are not systematically optimistic or pessimistic.

Note on cleaning*: We have 2139 observations after taking out rows for which there's no actual and no implied unemployment point prediction. For GDP, it's 5692 observations.

16. Do point predictions show more optimism than histogram distributions do?
    a. Conduct a paired t-test comparing point-predictions and weighted point prediction forecasts. If this test is significant, conduct a regression on the difference between them, including ~~fixed~~ random effects for forecaster and year being forecast. A systematic difference would show up as a significant intercept in the regression results.

Unemployment

The paired *t*-test is significant, suggesting that forecasters are more optimistic in their unemployment point predictions (mean 6.78) then their implied point prediction's (mean 6.89), $t = 10.489$, $df = 992$, $p < 2.2 \times 10^{-16}$. The model with forecaster and year being forecast included as random effects is significant at $p = .0003$; forecasters are more optimistic in their unemployment point predictions than in their implied point predictions (N=993).

GDP

The paired t-test is significant, suggesting that forecasters are more optimistic in their GDP point predictions (mean 2.65) than in their GDP implied point predictions (mean 2.52), $t = -13.6$, $df = 1970$, $p < 2.2 \times 10^{-16}$. The regression with RE is significant, $t = -4.981$, $df = 35.309$, $p = 1.66 \times 10^{-5}$, but suggests the opposite result – that forecasters are more optimistic in their GDP implied point predictions (e.g., the reverse of the paired t-test).

Note on cleaning*: Because we can only include forecasters who answered both, we have significantly less point prediction data. We have 993 observations for unemployment and 1971 for GDP.

17. Are forecasters more pessimistic during ~~economic booms~~ economic recessions?
    a. Conduct a regression at the level of the forecast. The difference between forecast point prediction and actual outcome is the dependent variable. The key independent variable is a dummy variable for whether the prior year was a recession year. Include ~~fixed~~ random effects for forecaster.
    b. Repeat the above analysis, but using weighted point prediction.

Unemployment point prediction (PP) and Implied PP

(a): The regression with (pp-actual) ~ recession + RE is significant, $t = -30.26$, $df = 6687$, $p < 2 \times 10^{-16}$. The intercept is positive, suggesting that forecasters think unemployment will be higher than the actual on average, implying pessimism. The coefficient on RECESSIONYES is negative, suggesting that during recession years, this difference goes down, implying that they get less pessimistic.

(b): The regression with (implied pp-actual) ~ recession + RE is significant, $t = -15.64$, $df = 2123.86$, $p < 2 \times 10^{-16}$; the intercept is positive, suggesting pessimism, and the coefficient on RECESSIONYES is negative, suggesting that when the prior year was a recession year, this difference goes down, implying that forecasters get less pessimistic.

GDP PP and Implied PP

(a): The regression with (pp-actual) ~ recession + RE is significant, $t = 19.66$, $df = 4188.16$, $p < 2 \times 10^{-16}$. The intercept is negative, suggesting that forecasters were, on average, pessimistic. The coefficient on RECESSIONYES is positive, suggesting that when the prior year was a recession year, the difference decreases and forecasters become less pessimistic on average.

(b): The regression with (implied pp-actual) ~ recession + RE is significant, $t = 13.82$, $df = 5687.31$, $p < 2 \times 10^{-16}$; the intercept is negative, suggesting that forecasters are pessimistic on average; the coefficient on RECESSIONYES is positive, suggesting that when the prior year was a recession year, forecasters were less pessimistic.

Because of the inconsistencies in our tests of optimism, we do not believe the data present a clear or interpretable conclusion. Therefore, we have elected to omit these analyses from the manuscript.

Note*: We conduct the analysis with year forecast made (as opposed to prior year) as the recession year; results do not change.

PART IV: Individual Level Analyses

18. To what degree are there stable individual differences between forecasters with respect to confidence? Are some forecasters more consistently over-precise than others? Are some forecasters more consistently optimistic than others?
    a. Test how much of the variance in forecaster confidence is accounted for by stable differences between forecasters. Conduct an ANOVA to test the change in R-squared when ~~fixed~~ random effects for forecaster and year forecast made are included in a regression. Perform the tests using the different measures of confidence:
        i. Overprecision
            1. Conduct the test using peak confidence
            2. Conduct the test using Gini index of the forecast
            3. Conduct the test using forecast Variance
        ii. Overestimation
            1. Conduct the tests using the difference between point prediction and actual
            2. Conduct the test using the difference between weighted point prediction and actual

The three ANOVAs for overprecision are significant (peak confidence null vs. full mod: $\chi^2 = 4948.4$, $df = 2$, $p < 2.2 \times 10^{-16}$; Gini: $\chi^2 = 6402.1$, $df = 2$, $p < 2.2 \times 10^{-16}$; Variance: $\chi^2 = 5064$, $df = 2$, $p < 2.2 \times 10^{-16}$), suggesting that including random effects for forecaster and year forecast made in the regression improves the null model.

The same holds for overestimation – including random effects for FID and YFM significantly improves the null model for both pp and implied pp and for both GDP (point prediction null vs. full mod: $\chi^2 = 4576.1$, $df = 2$, $p < 2.2 \times 10^{-16}$; implied point prediction: $\chi^2 = 4591.7$, $df = 2$, $p < 2.2 \times 10^{-16}$) and Unemployment (point prediction null vs. full mod: $\chi^2 = 4043.2$, $df = 2$, $p < 2.2 \times 10^{-16}$; implied point prediction: $\chi^2 = 628.88$, $df = 2$, $p < 2.2 \times 10^{-16}$)

    b. Test how much of the variance in forecaster confidence is accounted for by differences between time periods (i.e., whether some time periods are more predictable than others). Conduct an ANOVA to test the change in R-squared when ~~fixed~~ random effects for year being forecast and year forecast made are included in a regression. Perform the tests using the different measures of confidence:
        i. Overprecision
            1. Conduct the test using peak confidence
            2. Conduct the test using Gini index of the forecast

3. Conduct the test using forecast Variance

ii. Overestimation

1. Conduct the tests using the difference between point prediction and actual
2. Conduct the test using the difference between weighted point prediction and actual

The three ANOVAs for overprecision are significant (peak confidence null vs. full mod: $\chi^2 = 1366.5$, $df = 2$, $p < 2.2 \times 10^{-16}$; Gini: $\chi^2 = 832.4$, $df = 2$, $p < 2.2 \times 10^{-16}$; Variance: $\chi^2 = 1805$, $df = 2$, $p < 2.2 \times 10^{-16}$), suggesting that including random effects for year being forecast and year forecast made in the regression improves the null model.

The same holds for overestimation – including random effects for YBF and YFM significantly improves the null model for both pp and implied pp, and for both GDP (point prediction null vs. full mod: $\chi^2 = 7720.7$, $df = 2$, $p < 2.2 \times 10^{-16}$; implied point prediction: $\chi^2 = 1647.5$, $df = 2$, $p < 2.2 \times 10^{-16}$) and Unemployment point prediction null vs. full mod: $\chi^2 = 6424.2$, $df = 2$, $p < 2.2 \times 10^{-16}$; implied point prediction: $\chi^2 = 7237.7$, $df = 2$, $p < 2.2 \times 10^{-16}$).

PART V: Analysis within forecasters across time

19. How does experience affect future forecasts? Does more experience forecasting (as measured by the number of prior forecasts a forecaster has made in the Survey of Professional Forecasters) improve the accuracy of forecasts, using different measures:
    a. Conduct a regression using QSR as the dependent measure. The key independent variable is the number of prior forecasts in the data prior to the year the forecast was made (e.g., a rolling "number of prior forecasts"). The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.

For each additional prior forecast, QSR goes up by about .0025, $t = 15.08$, $df = 1422$, $p < 2 \times 10^{-16}$.

    b. Test for the effect of experience on overprecision (all three measures):
        i. Conduct a regression using peak confidence as the dependent measure. The key independent variable is the number of prior forecasts in the data. The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.
        ii. Conduct a regression using Gini index as the dependent measure. The key independent variable is the number of prior forecasts in the data. The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.
        iii. Conduct a regression using variance as the dependent measure. The key independent variable is the number of prior forecasts in the data. The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.

With every additional prior forecast in the regressions with RE, peak confidence goes up by .002, $t = 25.80$, $df = 6721$, $p < 2 \times 10^{-16}$, Gini goes up by .001, $t = 27.92$, $df = 7551$, $p < 2 \times 10^{-16}$, and variance goes down by .009, $t = -22.07$, $df = 2973$, $p < 2 \times 10^{-16}$. It appears that forecasting experience both increases accuracy and confidence.

20. Test for the effect of experience on optimism:
    a. Conduct a regression using the difference between forecast (point prediction) and actual as the dependent measure. The key independent variable is the number of prior forecasts in the data. The analysis includes ~~fixed~~ random effects for forecaster and year being forecast.
    b. Conduct the above analysis, but using weighted point prediction.

Unemployment
(a): The regression with random effects yields a directionally positive intercept, implying that the forecasters are pessimistic on average, but it is not significant (t = 3.236, df = 10.219, p = 0.009). The coefficient on the number of prior forecasts is negative suggesting that the difference between pp and actual decreases; with every additional prior forecast, forecasters become less pessimistic (t = -9.276, df = 828.64, $p < 2 \times 10^{-16}$).
(b) The regression with random effects yields a positive intercept, implying that the forecasters are pessimistic on average (t = 3.606, df = 10.038, p = 0.005). The coefficient on the number of prior forecasts is negative, suggestion that the difference between implied pp and actual decreases; with every additional prior forecast, forecasters become less pessimistic (t = -11.966, df = 889.668, $p < 2 \times 10^{-16}$).
GDP
(a): The regression with random effects yields a non significant intercept (t = .900, df = 22.635, p = 0.378) and coefficient (t = -1.604, df = 172.23, p = 0.11).
(b): The result for implied pp is consistent with the above, with a non significant intercept (t = 22.62, df = 0.263, p = 0.795) and coefficient (t = -1.055, df = 156.68, p = 0.293).

Note*: For all implied pp/histogram, we're using the smaller dataset – only using observations for which there's a match in pp (e.g., the forecaster answered both the full distribution (from which we pull an implied pp and the pp).

Inconsistencies between the results of different tests of optimism muddy the consistency or clarity of any conclusion, making us reluctant to make strong claims. We omit discussion of these analyses from the manuscript.

~~21. Is overprecision tempered with by being wrong?~~ What is the effect of being right on overprecision?
    a. The effect of being ~~wrong~~ right on subsequent precision: Measure forecaster error ~~as the summed squared distance (weighted by probability) between the forecast and the actual~~ as the average of prior QSR scores. Conduct a regression predicting precision using peak confidence as the dependent variable. The independent variable is forecast error in a forecaster's prior forecast; ~~fixed~~ random effects for forecaster and year being forecast.
    b. Conduct the above analysis using Gini index instead of peak confidence.
    c. Conduct the above analysis using variance instead of Gini.

The coefficient on prev_qsr is not significant in any of the models (peak confidence model t = 1.662, df = 7012, p = 0.097; Gini model t = 1.405, df = 6523, p = 0.16; variance model t = -2.094, df = 8911.72, p =

0.036) the average of prior QSR scores does not significantly predict any of the three measures of overprecision.

22. The effect of being right on precision: Measure rightness as the rolling average probability attached to the actual outcome. Add ~~fixed~~ random effects for forecaster and year being forecast.
    a. Conduct the analysis with peak confidence
    b. Conduct the analysis with Gini.
    c. Conduct the analysis with variance.

When the probability attached to the actual outcome increased, peak confidence increased by .19, t = 7.708, df = 6612, p = 1.46 x 10^{-14}, Gini increased by .11, t = 6.799, df = 6317, p = 1.15 x 10^{-11}, and variance decreased by .73, t = -5.449, df = 9007, p = 5.2 x 10^{-8}.

23. Do forecasters who start out being more precise (all three measures) tend to lessen their degree of precision over time? Repeat three analyses above replacing forecaster error with forecaster's average forecast precision in prior forecasts, using all three measures:
    a. Conduct the analysis with peak confidence.
    b. Conduct the analysis with Gini.
    c. Conduct the analysis with variance.

When prior precision goes up by one unit, peak confidence goes up by .56, t = 23.85, df = 712.89, $p < 2$ x 10^{-16}, Gini by .60, t = 27.54, df = 630, $p < 2$ x 10^{-16}, and variance goes up by .36, t = 17.59, df = 1293, $p < 2$ x 10^{-16}. In accordance with what we stated at the beginning of this document, the inconsistencies undermine our faith in this result.

24. Are forecasters from the financial services industry different in their precision or optimism?
    a. Conduct a regression at the level of the forecast, with ~~fixed~~ random effects for forecaster and year being forecast, in which peak confidence serves as the dependent measure and forecaster's industry (financial services or other) serves as the independent variable.
    b. Conduct the analysis above, except with forecast Gini as the dependent measure.
    c. Conduct the analysis above, using forecast variance as the dependent measure.
    d. Conduct the analysis above, using forecast optimism as the dependent measure (where optimism is the difference between the point prediction and the actual outcome).

All regressions show that forecasters from the financial services industry are not significantly different in their precision (finan_serv coefficient on peak confidence model t = 0.60, df = 1027, p = 0.549; Gini model t = -0.464, df = 1887, p = 0.643; variance model t = 1.047, df = 1585, p = 0.295) or optimism (finan_serv coefficient on unemployment optimism model t = -1.084, df = 66.257, p = 0.282; GDP optimism model t = -0.999, df = 144.237, p = 0.320), across all indicators and all measures.

PART VI: Wisdom of Crowds

25. Is the crowd wiser than the individual? If you averaged the forecasters' forecasts together (in essence making it the average of a 'crowd'), are the estimates more accurate, as measured by:

a. QSR: Compute accuracy (QSR) for each forecast and for the crowd forecast. Use a paired t-test to ask which is larger.
    i. The crowd's accuracy (as measured by QSR) is 1.33 and is higher than the average accuracy for individual forecasts, which is 1.18, $t = -32.851$, $df = 93$, $p < 2.2 \times 10^{-16}$.
b. Calibration in its precision: Measure forecast error as the summed squared distance, weighted by probability, between a histogram forecast and the actual outcome. Conduct a paired t-test (in which the unit of analysis is the year being forecast) comparing the average forecast error ~~for the average forecast~~ with the ~~averaged~~ error of the ~~individual~~ averaged forecasts.
    i. The average of the errors (0.97) is significantly larger than the error of the average (0.78), $t = 7.392$, $df = 93$, $p = 6.2 \times 10^{-11}$.

Note*: One of our reviewers recommended dropping our wisdom of crowds analysis, noting that "in assessing whether the crowd is wise, the average accuracy is compared in two ways against the accuracy of the average. The latter is more accurate, which is said to support crowd wisdom. That's true, but the accuracy of the average *must* exceed the average accuracy. It is a theorem (related to the Spearman Brown Prophecy formula), but time to time it appears as an empirical analysis in papers." As such, we decided to drop 16a from the validation analyses, but keep 16b.

PART VII: Additional Analyses

26. Lazy subjects and fatigue: Determine the order of the questions on the survey. If fatigue (e.g., subjects become lazier) is part of why we observe what appears to be overprecision, then we should observe more overprecision on questions later in the survey. Compare the precision (all three measures) of forecasters who skipped GDP and answered Unemployment against forecasters who completed both using a t test of two means; the former should be less fatigued given that the fewer questions you answer, the less fatigued you should be.

We find that none of these three *t*-tests are significant (t-test comparing the peak confidence of those who filled GDP vs. those who skipped it ($t = 0.597$, $df = 2124.1$, $p = 0.551$); the Gini ($t = 0.981$, $df = 2124.1$, $p = 0.327$); the variance ($t = -0.915$, $df = 2124.1$, $p = 0.36$)), providing suggestive evidence that fatigue does not explain why we observe overprecision in the data.

27. Fewer bins are less taxing: Find indicators that have a different number of bins, and compare peak confidence using a t test of two means.

Within indicators, there are survey years during which RGDP has 10 bins, and ones in which RGDP has 11 bins. A *t*-test reveals that forecasters are on average significantly more overprecise with fewer bins (.55 vs. .49), $t = 11.152$, $df = 4596.7$, $p < 2.2 \times 10^{-16}$.

RGNP has 6 bins, and the average peak confidence is .66. NGNP has 15 bins, and the average peak confidence is .515. We compare forecasters' peak confidence when there are 10 bins to when there are 6 and find that forecasters are more overprecise when there are fewer bins ($t = 145.92$, $df = 2421$, $p < 2.2 \times 10^{-16}$). We compare forecasters' peak confidence when there are 11 bins to when there are 6 and again find that forecasters are more overprecise when there are fewer bins ($t = 121.44$, $df = 2228$, $p < 2.2 \times 10^{-16}$). We compare forecasters' peak confidence when there are 15 bins to when there are 11 and find that forecasters are more overprecise when there are more bins ($t = -3.856$, $df = 2676$, $p = .0001$). This

evidence is sufficiently mixed that it leaves us skeptical that fewer bins are less taxing and lead to less overprecision.

28. Round numbers: Laziness and fatigue may increase the use of round numbers. Uncertainty should decrease precision; fatigue might plausibly increase precision. Create a new variable that assigns a point for every bin probability that ends in a 5, and two points for every bin probability that ends in a 0 (do not include bins with the number 0), and test whether the rows with more round numbers are more precise. Correlate the column with the new variable (higher value = more round numbers) against the column with peak confidence.

The correlation is -0.27, t = -31.581, df = 12357, $p < 2.2 \times 10^{-16}$, suggesting that the more round numbers they used, the less their precision was, implying uncertainty (e.g., uncertainty leads to a wider distribution and more round numbers).

29. Does the increase in accuracy with experience come about as a result of "bad" forecasters dropping out? For each forecast, include an indicator variable for whether it's the forecaster's last forecast recorded in the data. Then, use a logistic regression to test whether the rolling average prior forecast accuracy predicts dropout.

The coefficient on rolling average prior forecast accuracy is not significant (z = -1.054, p = .292), suggesting that the increase in accuracy with experience does not come about as a result of "bad" forecasters dropping out.

30. Compute a crowd forecast and test for calibration on the three measures of overprecision. Average all forecast probabilities for a given indicator, year forecast made, year being forecast, and bin arrangement to get the crowd forecast.

We find evidence of underprecision using our peak confidence measure. The average peak confidence is 41.9% and the average hit rate is 45.7%, t = -4.3208, df = 187, $p < 2.2 \times 10^{-16}$. We find evidence of overprecision using our variance (average variance of forecasts = 1.35, average variance of actuals = 3.79, t = -25.307, df = 187, $p < 2.2 \times 10^{-16}$) and Gini measures (average Gini of forecasts = 0.72, average Gini of actuals = 0.53, t = 27.364, df = 187, $p < 2.2 \times 10^{-16)}$. This may be due to how the variance and Gini of the actuals are calculated.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., & Camerer, C. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics*, *54*(3), 306–316.

Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie Di Metodologica Statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*.

Glas, A., & Hartmann, M. (2018). Overconfidence versus rounding in survey-based density forecasts. *Available at SSRN 3202810*.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209–219.