UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Campbell, S., & Moore, D. A. (2024). Overprecision in the Survey of Professional Forecasters. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.92953

Organizational Behavior

Overprecision in the Survey of Professional Forecasters

Sandy Campbell[1][a], Don A. Moore[1]

[1] Management, Haas School of Business, University of California, Berkeley, CA

Keywords: overconfidence, economic forecasts, policy, cognitive bias, behavioral economics, judgment, decision making

https://doi.org/10.1525/collabra.92953

Collabra: Psychology

Vol. 10, Issue 1, 2024

Every decision depends on a forecast of its consequences. We examine the calibration of the single longest and most complete forecasting project. The Survey of Professional Forecasters has, since 1968, collected predictions of key economic indicators such as unemployment, inflation, and economic growth. Here, we test the accuracy of those forecasts (n = 16,559) and measure the degree to which they fall victim to overconfidence, both overoptimism and overprecision. We find forecasts are overly precise; forecasters report 53% confidence in the accuracy of their forecasts, but are correct only 23% of the time. By contrast, forecasts show little evidence of optimistic bias. These results have important implications for how organizations ought to make use of forecasts. Moreover, we employ novel methodology in analyzing archival data: we split our dataset into exploration and validation halves. We submitted results from the exploration half to *Collabra:Psychology*. Following editorial input, we updated our analysis plan for the validation dataset, preregistering only analyses that were consistent across different economic indicators and analytic specifications. This manuscript presents results from the full dataset, prioritizing results that were consistent in both halves of the data.

Although the study of cognitive biases has been influential (Kahneman, 2003), questions remain regarding the importance and generalizability of effects demonstrated in the research laboratory. Critics correctly note difference between convenience samples of students and the experienced professionals to whom we seek to generalize (Klein et al., 2017; Luan et al., 2019). Moreover, there are real differences between the judgments studied in the lab and high-stakes business decisions (Levitt & List, 2007). Of course, one of the main reasons that studies of cognitive bias have relied so heavily on lab tasks is that they afford accuracy benchmarks that allow researchers to identify deviations from prescriptive rationality (Kahneman & Tversky, 1996). It is rare for high-stakes economic decisions to come with benchmarks that allow researchers to assess their optimality. This paper exploits a unique dataset that allows us to score the accuracy of consequential forecasts by business professionals.

Forecasting

Since 1968, the Survey of Professional Forecasters (SPF) has collected forecasts of key economic indicators, such as unemployment, inflation, and economic growth (Croushore, 1993). The survey informs Federal Reserve's monetary policies and businesses' strategic planning. The SPF represents one of the best historical records of quantified forecasts. As such, it presents a unique opportunity to test calibration and accuracy in one of the most important forecasting records available. Moreover, given that every decision depends on a forecast, learning if and how forecasts show overconfidence offers insight into methods that could improve forecast methodologies, planning, and decisions.

Governments, businesses, and other organizations have acknowledged the foundational importance of accurate forecasts (Schoemaker & Tetlock, 2016; Silver, 2012; Tetlock & Gardner, 2015). Choices about investment, hiring, mergers, product introductions, and market entry all depend on accurate forecasts of their consequences (Gutierrez et al., 2020). Getting even a little better at forecasting the future can pay enormous dividends and help organizations achieve their goals, large and small (Paik et al., 2023). However, traditional forecasting formats exacerbate the human tendency toward overconfidence by focusing on a single point prediction. Indeed, too many organizations rely on forecasts that ask for a point prediction and maybe some estimate of the forecaster's confidence in it (Niu & Harvey, 2022). These forecasts do not have a great track record (Flyvbjerg et al., 2009; Lin & Bier, 2008). But evidence does indeed suggest that it is possible to get better at forecasting (Mellers et al., 2014). One way to help forecasters think

a Corresponding author: Sandy Campbell, Haas School of Business, University of California, Berkeley. 2220 Piedmont Avenue, Berkeley, CA 94720. Email address: sandyca@berkeley.edu

through uncertainty is by getting them to think probabilistically, and even think in probability distributions (Bruine de Bruin et al., 2000; Chang et al., 2016). Asking forecasters to report probability distributions helps them think through their uncertainty and report better-calibrated forecasts (Goldstein & Rothschild, 2014; Haran et al., 2010).

A second crucial component of debiasing includes feedback. Practice with repeated forecasts over time, accompanied by clear feedback, can help people calibrate their confidence judgments and reduce overconfidence (Arkes et al., 1987; Keren, 1987; Murphy & Winkler, 1977). On the other hand, in the absence of systematic feedback, overconfidence lacks a corrective to counter the influences of selective memory and hindsight biases (Fischhoff, 1975; Merkle, 2017). The result may be that forecasters are unaware of their biases and cannot correct them (Kluger & DeNisi, 1996).

The SPF embodies conditions that ought to reduce bias. First, because the forecast is highly consequential it ought to enhance motivation for accuracy. Critics of the research literature on overconfidence note the reliance on laboratory tasks and trivial judgments, such as trivia quizzes (Gigerenzer et al., 1991). If people believe their performance on trivia tests is unimportant, it might not stimulate as much effort as would high-stakes forecasts, like the SPF, that inform decision-making at the highest levels of the public and private sectors. Second, forecasters know they will receive feedback. In every quarter, they learn how gross domestic product, unemployment, and inflation changed in the previous quarter. The numbers for the previous quarter are printed on the survey they receive (e.g., when they are forecasting 2009 Q2, the survey displays the "correct answer" for 2009 Q1 immediately before). It is harder to get away with overconfident claims in the presence of clear feedback (Tenney et al., 2019). Feedback reduces the potential inter- and intra-personal benefits of pretending to know more than one actually does (Pulford & Colman, 1997; Schwardmann & van der Weele, 2019; Van Zant, 2021). Third, the SPF asks forecasters to report a probability distribution, thereby helping them to think through their uncertainty, explicitly estimating the probability that they might be wrong.

Forms of Overconfidence

Because overconfidence is the most significant of the cognitive biases (Kahneman, 2011), we test for its presence in the SPF. We seek to disambiguate conflicting effects in a literature that routinely uses the same word (overconfidence) to refer to different measures of distinct constructs. Some research uses the term overconfidence to describe optimistic overestimation of future performance (Amore et al., 2021). Other research uses the term overconfidence to describe exaggerated belief in one's superiority to others (Chen et al., 2018). And still other research uses the term overconfidence to refer to excessive certainty in the precision or accuracy of one's beliefs or forecasts (Owens et al., 2013). In order to distinguish them, we call these three approaches to measuring overconfidence overestimation, overplacement, and overprecision (Moore & Healy, 2008).

Overly optimistic forecasts would be a manifestation of overestimation. For instance, overestimating future economic growth or underestimating future unemployment could be considered overly optimistic. Its consequences can be profound (Beaudry & Willems, 2022; Simon & Houghton, 2003), but overestimation is far from universal (Baillon et al., 2018; Windschitl & O'Rourke, 2015). Overplacement, for its part, is the exaggerated belief in being superior to others (Camerer & Lovallo, 1999). The SPF does not afford an assessment of overplacement, since forecasters do not compare their accuracy with that of others. Overprecision, the third form of overconfidence, is the excessive faith in the accuracy of one's knowledge (Moore et al., 2015). Forecasters exhibit overprecision when they are too sure they know what is going to happen. We find consistent evidence of overprecision in economic forecasts. By contrast, evidence of optimistic overestimation is weak and inconsistent.

Prior studies have analyzed data from expert probability forecasts, including data from the U.S. SPF and the European Central Bank's macroeconomic forecasts (see, for example, Clements, 2014; Diebold et al., 1999; Engelberg et al., 2009; Giordani & Söderlind, 2003, 2006; Krüger, 2017; Lahiri & Wang, 2006). They have confirmed the usefulness of expert probability forecasts, finding that the expert consensus is generally informative about future direction of change (Kenny et al., 2014). By contrast, we focus on the calibration of individual forecasts.

Contributions and Research Innovations

We contribute to literature in at least five ways. First, we operationalize overprecision in three ways to avoid the pitfalls of assuming the nature of "true" uncertainty distributions (Kenny et al., 2015). Second, we make use of forecasters' full subjective probability distributions rather than just their 90 percent confidence intervals (Giordani & Söderlind, 2003, 2006). Third, we distinguish between overestimation and overprecision as measures of overconfidence. While the former has received more attention in the literature (Giordani & Söderlind, 2006), the latter has proven more pervasive (Casey, 2021). This, then, constitutes our key hypothesis: We predict that forecasts will claim greater certainty than their accuracy justifies. Our preregistration specifies how we plan to test this hypothesis, using three different measures of certainty, as well as the rigorous plan to only claim support for our hypothesis if all our different tests show consistent evidence of overprecision.

Our fourth contribution rests in our comprehensive analysis of the data—we broaden our analysis from a narrow focus on just inflation or gross domestic product growth by using multiple indicators from the SPF. Furthermore, we distinguish between overestimation and overprecision across these indicators, operationalize both constructs in various ways, and explore forecaster experience, fatigue, and inattentiveness as potential causes of our results.

Our fifth contribution is methodological: we employ a split-sample approach on archival data that allows for preregistered confirmatory hypothesis testing. We preregis-

tered our analyses, along with a plan to split the data into two parts: exploration and validation. The split sample approach advances research practice by drawing on insights from preanalysis plans and replications (Christensen et al., 2019; Fafchamps & Labonne, 2017; Miguel et al., 2014). It offers the strengths of both registered reports (Chambers et al., 2015) and completed reports, since it affords both the opportunity to refine the analytical approach by working with the data but also a reduction in concerns about overfitting or selective reporting thanks to replication in the hold-out sample. Given a sufficiently large total sample size, researchers can split their datasets into "training" and "testing" samples – in our case, we randomly selected half our data, stratified on the year the forecast was made, the quarter, the year being forecast, and the indicator, and dubbed our two samples "exploration" and "validation" for ease of interpretation. The first author cleaned and split the data, and set aside the confirmation set; neither author touched the confirmation dataset until undertaking the confirmatory tests specified in our June 2023 preregistration.

Split samples are sometimes used to refine a model and estimate its predictive performance. Cross validation is a statistical technique that splits the data into training and testing sets (Shao, 1993). The technique can employ multiple iterations of cross-validation, averaged over many different splits of the data. The split-sample method we employ, by contrast, is less focused on optimizing a model's predictive performance and instead offers the potential for confirmatory hypothesis testing in the hold-out sample. This approach affords both (1) the opportunity to work with a complex dataset to establish appropriate statistical specifications for hypothesis testing and (2) the possibility to conduct confirmatory tests on the hold-out sample that replicate the original results. Little organizational research employs cross-validation, but we hope that will change.

The appeal of having a dataset for exploration is clear: The greatest constraint of preanalysis plans is that they limit the potential learning associated with exploration (Collins et al., 2021). Exploration is frequently useful in better understanding one's data. However, exploratory data analysis can inflate the risk of overfitting and false-positive results (Hofman et al., 2021; Tenney et al., 2021). The split sample approach allows researchers to conduct exploratory analyses on part of the data, then register a confirmatory analysis plan documenting the hypotheses they plan to validate in the hold-out sample. While fields such as machine learning frequently use split sample approaches (Cawley & Talbot, 2010), to our knowledge, the approach has seen limited use in the behavioral sciences. This may be because split-sample techniques require large sample sizes (Anderson & Magruder, 2017). Nevertheless, we believe this method should see more widespread use by researchers who use large archival datasets, and we demonstrate a potential implementation in this paper.

Our research timeline:

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 preregistration for the exploration dataset, specifying 48 planned tests: https://osf.io/q6x47/.
2. Between September 2019 and June 2023, we ran our preregistered 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 preregistration, 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/sj5kr/).
3. In June 2023, we submitted a manuscript containing the initial results to *Collabra:Psychology*. Following editor and reviewer feedback, we created an updated preregistration for the validation dataset (https://osf.io/dtuzv). We retained only analyses that were consistent across different economic indicators and analytic specifications. We tested whether the results held in the validation data. Note that we supplemented our validation dataset (which contained data from 1968-2019 Q2) with forecasts that had been made since we last downloaded the data in 2019 (i.e., data from 2019 Q3 – 2023 Q4). We report results from this combined dataset here.

Our approach has a few key features worth highlighting. First, in presenting our project to peers and audiences, and in the manuscript submitted for review, we reported results from the exploration half of the data (12,359 forecasts). This left open the opportunity to incorporate feedback prior to preregistering analyses for the validation dataset. Second, in the interest of maintaining a high standard with respect to replicability, we highlight in this final manuscript only those results that hold in both the exploration and validation samples, and clearly distinguish when they do not (e.g., results that were inconsistent in the exploration dataset, that we therefore did not test in the validation dataset). While we acknowledge consistently null results that could be potentially interesting, we hesitate to make any inferences from these null results.

Third, in the interest of transparency, we preregistered both our exploratory analyses and our validation analyses. As noted in our timeline of events, we detailed in our supplementary materials how we updated our plan as we explored the data and revised our opinions about the best way to test our hypotheses. We amended our analyses based on what we learned, and submitted this updated plan and draft for review. Following editor and reviewer feedback, we updated this document further, preregistered a plan for the validation dataset, and conducted these preregistered analyses on the validation half of the data set, supplemented with the most recent forecasting data. Our preregistrations, data, and analysis code are available at https://osf.io/sj5kr/.

Method

The Survey of Professional Forecasters (SPF) is the oldest continuous forecasting survey in the United States. The American Statistical Association and the National Bureau

of Economic Research started the SPF in 1968. The Federal Reserve Bank of Philadelphia took over the survey in 1990. Actual releases, documentation, mean and median forecasts of all the respondents, as well as the individual responses from each (anonymous) forecaster are available online. The SPF's documentation explains all variable definitions and transformations, the survey's timing, files, and changes over time. The documentation makes it easy for any reader to understand exactly what the survey's results are and how to interpret them. It is constantly being updated to reflect new information about the survey as it evolves, and as such, is the best reference for an up-to-date account on the SPF. We tested forecast accuracy using the best revised outcome data from the Saint Louis Federal Reserve (https://fred.stlouisfed.org/). The Philadelphia Fed has also done many of their own analyses on the data; the one most relevant is Stark (2010), who quantifies the extent to which the data revisions matter and looks at forecaster performance relative to that of simple benchmark time-series models, operationalized using root-mean-square error.

Understanding the Survey of Professional Forecasters

The Philadelphia Federal Reserve provides sample surveys (https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/form-examples) showing question wording and forecasting items (see also Croushore and Stark (2019) for a table on the latter). Viewing the survey from 1981 Q2 and the one from 2014 Q1 will show that the formatting does not change much over time, but the variables, the number of bins, and the ranges covered by the bins have been amended various times; again, these amendments can be found in the SPF's documentation. The rationale for changes to the survey are often quite simple, as explained in Croushore and Stark (2019). For example, the addition of real gross national product and its components in the third quarter of 1981 was to allow analysts to better assess the strength of broad economic conditions.

For a summary on the survey's structure and variables, its participants, and a history on the survey and its uses by researchers, there is no better report than that by Croushore and Stark (2019). Dean Croushore revived the survey back in the 1990's, and Tom Stark is the co-creator and current assistant director of the center that runs the survey. Here, we note a few key points: The survey goes out once each quarter, immediately after the U.S. Bureau of Economic Analysis (BEA) releases numbers on the previous quarter's gross domestic product (GDP). Forecasters receive information on the previous quarter's outcomes when making their forecast, and thus get feedback on how they did on their prior quarter's forecast. Forecasters are given just over a week to send in their forecasts, and their forecasts are compiled and released to the press and the public immediately thereafter.

Croushore and Stark (2019) define a professional forecaster as "a person for whom forecasting is a major component of their job." Survey respondents include those who work at forecasting firms and make forecasts for their external clients, banks or other financial institutions that generate forecasts for their internal and external clients, chief economists for industry trade groups and manufacturers, and academics who study optimal forecast methods. Croushore and Stark (2019) maintain forecasters' anonymity, citing evidence that, when they are identifiable, some forecasters seek publicity by providing extreme forecasts to stand out from the pack (Lamont, 2002; Laster et al., 1999). They note that over the years, they have faced some pressure from academics and other data users to release the names of the forecasters with their projections, but have stood firmly by their decision because of concerns about how the forecasts might be affected.

This paper's Supplemental Online Materials (SOM) detail how we obtained, organized, and analyzed the data: https://osf.io/sj5kr/. Where appropriate, we treated Forecaster ID, Year Forecast Made, and Year Being Forecast as random effects given that we are interested in generalizing beyond the existing data. We planned to count effects as worthy of publication only if we replicated them in both halves of the data. Where we were uncertain regarding the most suitable analysis for testing a particular hypothesis, we conducted multiple tests. When they produced inconsistent results, it undermined our faith in the durability of that claim. Therefore, we prioritize results that are consistent and statistically significant across specifications, indicators, and across both halves of the data. Seeking to reduce the risk of false positives, we set an alpha of .005 for statistical significance (Benjamin et al., 2018). We note that we explored many interesting questions, but given our stringent criteria, many of our exploratory analyses did not make it into the main body of this manuscript. For example, we asked whether there is a correlation between overprecision and macroeconomic trends (e.g., whether forecast precision might decrease after say, recession years). We found inconsistent evidence across various specifications, and thus this analysis and its results were relegated to the SOM (e.g., see pages 5-7). We invite interested readers with further questions to explore the SOM, as many of the answers will lie there.

Participation in the SPF is limited to those with professional forecasting experience. The SPF asks forecasters to report both point predictions and histograms for forecasts of economic growth, unemployment, and inflation (both Consumer Price Index and Personal Consumption Expenditures). For each of these indicators, the survey breaks the range of possible outcomes into a set of mutually exclusive and exhaustive bins, and forecasters indicate the probability that the outcome will fall into each bin. For example, forecasters predicting growth in Gross Domestic Product (GDP) from 2005 to 2006 reported the probability that GDP would grow by 6 percent or more, 5.0 to 5.9 percent, 4.0 to 4.9 percent, 3.0 to 3.9 percent, 2.0 to 2.9 percent, 1.0 to 1.9 percent, 0.0 to 0.9 percent, -1.0 to -0.1 percent, -2.0 to -1.1 percent, or decrease by more than 2 percent. We note one potential limitation of histograms: bins require bin boundaries, which could influence the responses. We address this limitation with supplemental analyses that ask whether the number of bins influences our outcomes.

These histograms enable us to estimate overprecision better than do confidence intervals (Ben-David et al., 2013) for at least three important reasons. First, histogram responses are generally more accurate and less biased than are confidence intervals (Haran et al., 2010). Evidence suggests that people make a number of systematic errors when specifying confidence intervals (Soll & Klayman, 2004; Teigen & Jørgensen, 2005). Indeed, those errors are severe enough that it is worth questioning the degree to which people are even able to faithfully report percentiles from a subjective probability distribution, as confidence intervals require (Hoffrage, 2004; Moore et al., 2015). Second, interpreting histograms requires researchers to make fewer assumptions about the distribution of probability within a reported confidence interval. Third, histograms provide richer information than do confidence intervals. This richness affords multiple useful tests of overprecision. We employ three: peak confidence, variance, and Gini.

Overprecision Measure 1: Peak Confidence

Peak confidence focuses on the bin to which each forecaster assigned the greatest probability and compares this confidence, averaged across forecasters, to the rate at which they were correct—that is, the percentage of the time the truth landed in the focal bin. We refer to this maximum probability as a forecaster's peak confidence. A forecaster scored a 'hit' if they assigned their peak confidence to the correct range. We divided hits whenever the maximum confidence was tied between multiple bins. So, for example, if a forecaster indicated that they believed there was a 50% chance the outcome would land in each of two bins, they scored 50% of a hit if the outcome did indeed land in one of them. The advantage of this measure is its simplicity and interpretability: it provides our headline result that confidence (53%) exceeds accuracy (23%). Its shortcoming, that it ignores the confidence assigned to other bins, is remedied by our other two measures, variance and Gini.

Overprecision Measure 2: Variance

To measure the variance of each forecast, we computed the distribution's mean, then summed the squared distance to each bin, weighted by the probability assigned to it. Higher variance would imply less precision, as probabilities are spread out more across bins. We calculated the variance of each forecast and compared that with the variance of the actuals. The variance of the actuals was computed as the variance of realized outcomes across the entire epoch covered by the data. The SOM (page 2) details this calculation (see also 231226_FEO-SPF Validation Analyses.R). Variance nicely captures the spread of a unimodal distribution; however, a bimodal forecast could be fairly concentrated, in the sense that all the probability is concentrated in just two bins, but score high on variance. Our third measure addresses this concern.

Overprecision Measure 3: Gini

The Gini coefficient computes the concentration of the distribution across bins, much as a nation's Gini index captures the concentration of wealth across individuals (Gini, 1912; Lorenz, 1905). To compute a Gini coefficient, we order the bins in a forecast by the probability assigned them, from most to least. The most dispersed distribution assigns equal probabilities to all bins, and receives a Gini coefficient of 0. The most concentrated distribution assigns 100% probability to one bin, and receives a Gini index of 1. Figure 1's left panel shows these two cumulative distributions and the shaded region between them. Panel B shows a distribution assigning 50% probability to each of two bins, with a Gini index of .875. That is, the shaded blue region (labeled A) below the red cumulative curve represents 87.5% of the entire triangular region. The Gini coefficient is the proportion of the area under the curve:

$$G = \frac{A}{A + B}$$

To understand the difference between variance and Gini, consider a bimodal distribution with high variance (since most of the probability mass is far from the distribution's mean) but concentrated with high probability in two locations. The Gini coefficient would reflect this concentration, but variance would not. Higher Gini scores reflect greater concentration, and a Gini coefficient of 1 would result from a forecast that assigned 100% probability to one bin. The Gini of the actuals was computed as the Gini of realized outcomes across the entire epoch covered by the data. The SOM provides more details on our Gini calculations (see page 2), and the exact calculation can be found in the R files on OSF (see 231226_FEO-SPF Validation Analyses.R). In practice, we should not expect to see vast differences in measuring Gini versus variance, as forecasters typically report unimodal distributions. We measure both to capture whether there are any differences, with the expectation that the distributions should appear relatively similar, and both will consistently show evidence of forecaster overprecision.

We employ three different measures because each one captures a different aspect of precision in judgment and it was not clear *ex ante* which of the three was the right one for our analysis. Variance has historically been one of the most commonly used measures in prior literature, which has evaluated probabilistic forecasts by comparing these against the distribution of observed outcomes (Casey, 2021; Giordani & Söderlind, 2003, 2006; Kenny et al., 2015). This approach has its limitations, as the underlying assumption is that forecasters are attempting to map subjective uncertainty to realized distributions for a given time period. It is additionally problematic that the time period typically equates to a small sample size, and the choice of period could be unrepresentative. This concern applies to our variance and Gini measures, but not peak confidence.

Thus, our preregistered analysis plan stipulated a conservative approach in which all three measures must show a consistent effect before we conclude it is reliable. We report our main result with all three measures, but for simplicity and ease of interpretation, focus on peak confidence. Note

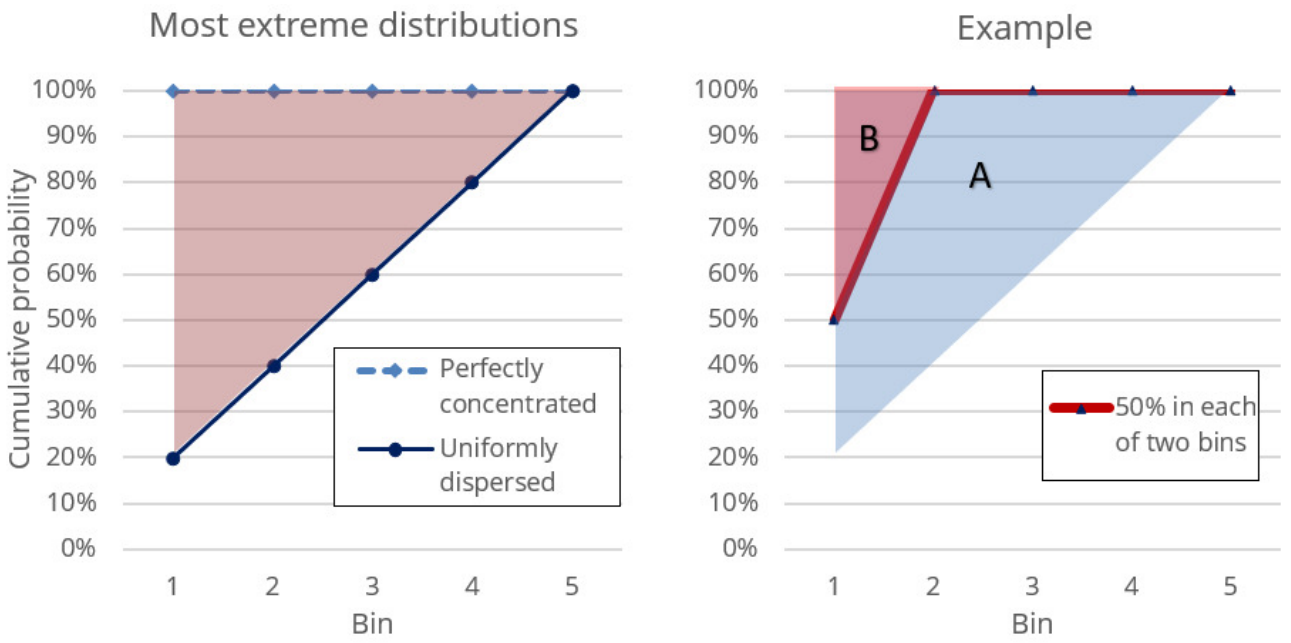

Figure 1. Example 5-bin distributions scored according to Gini index. The left panel shows the shaded area between the maximally dispersed and concentrated distributions. The percentage of that region below the actual cumulative distribution is the Gini coefficient.

that where the total probabilities, summed over bins within one reported distribution, do not sum to 100%, we normalize to 100% by dividing each cell by the total.

Measuring Forecaster Optimism and Accuracy

We examined optimistic overconfidence among forecasters by testing whether forecasts imply a brighter economic future than actually occurs. For example, if a forecaster predicted the unemployment rate would be 5% in 2008 and the actual unemployment rate was 2%, the original forecast would be pessimistic. On the other hand, since greater GDP growth represents a strong economy, if a forecaster predicted 10% GDP growth in 2008 and GDP growth was actually 5%, the 10% forecast would qualify as optimistic. We omit inflation from the optimism analyses because it is less clear how optimism would manifest itself; the economy overall benefits from some moderate inflation, but forecasters may disagree about how much. Inferring optimism is further complicated by the fact that low inflation is good for savers but high inflation is good for debtors.[1] We analyze optimism using both forecasters' actual point predictions, which they provide in the survey, and the mean of the distribution represented by their histogram forecasts.

We also consider the relationship between overprecision and accuracy. We score the accuracy of forecasts using the quadratic scoring rule (Selten, 1998). The QSR ranges from 0 to 2, and higher scores reflect more accurate forecasts; it is the inverse of the Brier (1950) score.

Sample and Hold-Out

The data include 32,462 forecasts by 443 different forecasters from 1968 through 2019 Q2.

After we cleaned and split data, 16,217 individual forecasts remained in the exploration set. We dropped 2,814 empty data rows. We excluded 1,532 forecasts lacking outcomes, because they were forecasted sufficiently far into the future that outcome data were not yet available, making it impossible for us to assess their accuracy. After exclusions, 12,359 forecasts from 370 different forecasters remain for our exploratory analyses.

The validation dataset we set aside in 2019 consists of 16,245 individual forecasts. We supplement the validation dataset with forecasts that have been made since we last downloaded the data in 2019; this includes 7,944 forecasts from 2019 Q3 through 2023 Q4. We dropped 5,141 empty data rows. We excluded 2,489 forecasts lacking outcomes, because they were forecasted sufficiently far into the future that outcome data were not yet available, making it impossible for us to assess their accuracy. After exclusions, 16,559 forecasts from 396 different forecasters remain for our validation analyses.

As preregistered, we highlight in this final manuscript only those results that hold in both the exploration and validation samples, and will clearly distinguish when they do not (e.g., results that were inconsistent in the exploration dataset, that we therefore did not run in the validation dataset).

1 We acknowledge that high economic growth and low unemployment are not necessarily good for everyone, but they are more consistently beneficial across the economy.

Results

Are forecasters overprecise?

We employ three measures of overprecision. Peak confidence measures the bin(s) to which the forecaster assigns highest probability. We compute the average reported probability, and compare that against the rate at which the forecasters are correct. Specifically, we conducted a paired samples *t*-test at the level of the forecaster comparing confidence with the hit rate. This test compared, for each forecaster, reported confidence that they knew what would happen with the rate at which they were correct. Across all indicators (GDP growth, inflation, and unemployment), confidence (M = 53%) significantly exceeds hit rate (M = 23%), t (375) = 19.74, p < .001. This result is robust to different approaches to measuring overprecision. Our second overprecision measure finds the average variance of the forecasts (1.34) is lower than the average variance of the actuals (5.85), t (375) = -53.90, p < .001. Similarly, our third approach computes the Gini index and finds that the concentration of the forecasts (0.81) is higher than the concentration of the actuals (0.51), t (375) = 54.137, p < .001, suggesting that forecasts are overprecise. In principle, it would be possible to conduct this analysis separately for different indicators, we are wary of conducting subgroup analyses in the absence of a strong theoretical motivation to expect differences. Without a theory to guide our analyses, we increase the risk of capitalizing on chance and identifying patterns that are not robust to replication.

Figure 2 graphs confidence against accuracy for all bins. Bins assigned 100% probability are correct about 66% of the time, and bins assigned 50% probability are correct about 34% of the time. Of course, the corollary must be that some bins assigned low probabilities are correct more often than forecasters predict; the forecaster who reports 100% confidence in the incorrect bin must have also assigned 0% confidence to the correct bin. Table 1 groups bins by the level of confidence assigned them, and provides frequency counts for each confidence level, as well as mean reported confidence and hit rate pooled across all variables and all forecasters.

Table 1. Confidence and hit rate across all bins in validation data, grouped by percent confidence level.

Confidence Level	Frequency	Confidence	Hit
0 – 9%	119748	0.008	0.035
10 – 19%	17202	0.123	0.142
20 – 29%	16026	0.243	0.212
30 – 39%	3323	0.348	0.278
40 – 49%	5215	0.417	0.295
50 – 59%	5497	0.537	0.342
60 – 69%	1456	0.680	0.462
70 – 79%	487	0.750	0.454
80 – 89%	1016	0.815	0.519
90 – 99%	745	0.925	0.636
100%	411	1.000	0.659

Are forecasters overly optimistic about the future?

In our exploration dataset, we examined optimistic overconfidence among forecasters by testing whether forecasts imply a brighter economic future than actually occurs. We used both forecasters' actual point predictions and the mean of the distribution represented by their histogram forecasts. Across the various specifications, we found inconsistent results, as enumerated in the supplemental online materials (see pages 5-6). For instance, only some of our analytic specifications found pessimistic forecasts of unemployment; a paired samples *t*-test comparing point-prediction forecasts with outcomes found that mean forecasts (6.22%) exceeded the realized unemployment rate (6.13%), t (6889) = 10.43, p < .001. However, some of our analytic specifications found optimistic forecasts of GDP growth, with growth forecasts higher (2.64%) than realized growth rates (2.36%), t (4265) = 14.72, p < .001. Results that are inconsistent across indicators and across analytic specifications undermine our faith in any general conclusion about forecasters' propensity toward optimism or pessimism. Our inconsistent results mirror results in the literature on overestimation: Evidence does not suggest a general propensity toward overestimation or optimism across the board, despite numerous documented localized deviations from perfect calibration (Cooper et al., 1988; Norem & Cantor, 1986). In line with our preregistered plan, we do not conduct tests on optimism in our validation dataset.

Forecast Accuracy

Our next analyses consider forecast accuracy. Naturally, accuracy increases as the distance shrinks between the time the forecast is made and the moment of truth when the outcome is realized. We conducted a regression on forecast accuracy, predicted by temporal distance from the moment of truth (as measured in quarters), with random effects for forecaster and year being forecast. We find that as the distance to the moment of truth increases, accuracy, as measured by quadratic scoring rule (Selten, 1998) or QSR, decreases by about 0.14 per quarter, t (16500) = -38.17, p < .001. This implies that indeed, forecast accuracy decreases with time. Forecasts that are four years out are less accurate that forecasts made a quarter from the moment of truth, as Figure 3 shows. We also test whether peak confidence changes with distance to the moment of truth, again adding random effects for forecaster and year being forecast. We find that as the distance to the moment of truth increases, average peak confidence decreases by about 0.08 per quarter, t (16270) = -57.12, p < .001. Together, these results are consistent with Clements (2014, 2018), who finds a tendency towards overconfidence among individual forecasters only at longer horizons (e.g., in excess of a year).

Do forecasts improve over time? We conducted a regression at the level of the forecast on accuracy, as measured by the QSR. The independent variable is the year the forecast

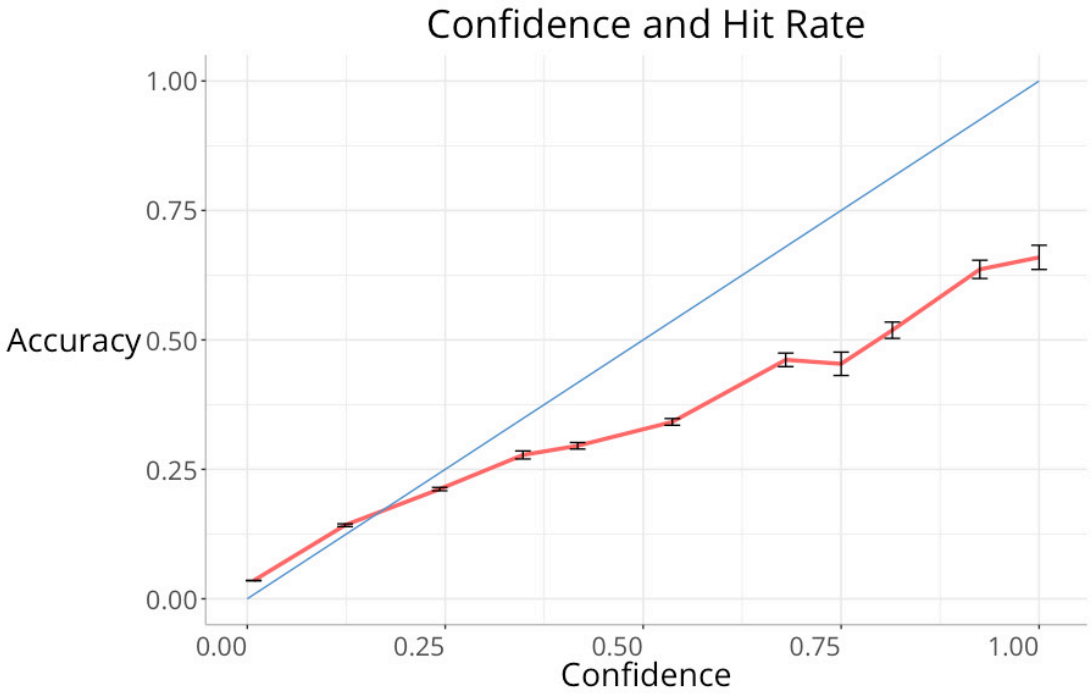

Figure 2. Confidence and hit rate, conditional on probability assigned to each bin for all bins.

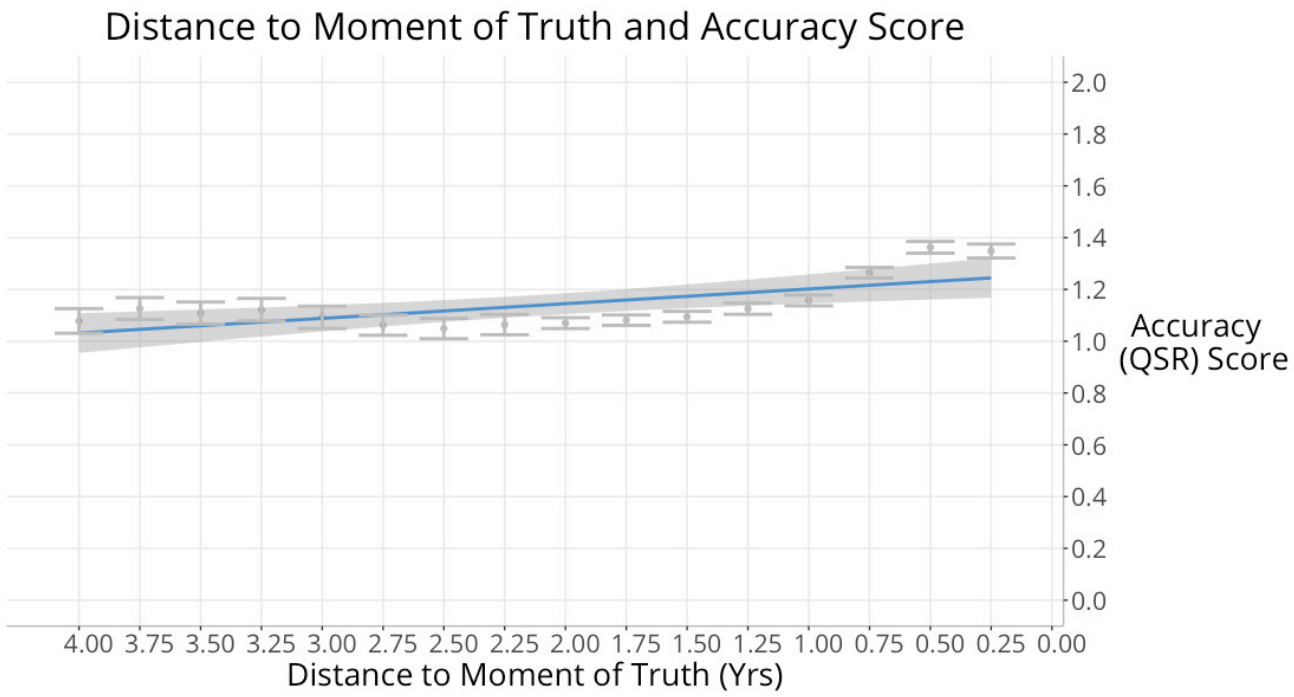

Figure 3. Accuracy (QSR) score as moment of truth approaches.

was made and we included random effects for forecaster and year being forecast. We find that over time, forecast accuracy improves. Each passing year increases the mean QSR score by .13, t (10170) = 34.59, p < .001. Forecasts made when the survey began in 1968 scored an average of 0.67 on accuracy. By 2016, mean accuracy had grown to 1.40 out of a maximum possible score of 2. This implies that over time, forecasters have become more accurate their predictions. Table 2 summarizes the means and standard deviations for all three measures of overconfidence, as well as the average forecast distance to the moment of truth, and the average number of forecasts per forecaster.

We also consider how peak confidence changes as a function of the year the forecast was made, again adding random effects for forecaster and year being forecast. We find that over time, average peak confidence increases by about 0.06 per quarter, t (114300) = 46.84, p < .001.

Table 2. Summary Statistics, averaged across all forecasts

	Mean	Standard Deviation
Peak Confidence	0.523	0.193
Hit rate	0.365	0.465
Gini	0.807	0.129
Variance	0.852	1.406
Distance to moment of truth (in years)	1.112	0.917
No. of Forecasts per forecaster	44.040	83.831

Forecaster Experience

Having found that forecasters are overprecise but that forecasts get better, both over time and closer to the moment of truth, we asked how experience affects forecasts. We tested the relationship between accuracy and number of prior forecasts in a regression that included fixed effects for forecaster and year being forecast. The results show that for each additional prior forecast, accuracy increases by about .0025, t (2405) = 21.56, p < .001. However, precision also increases. Specifically, with every additional prior forecast, peak confidence goes up by .001, t (8426) = 30.93, p < .001. Lamont (2002) argued that this could result from increased risk-taking by experienced forecasters. While this is possible, it would contradict the general tendency for risk seeking to decline with age (MacCrimmon & Wehrung, 1990). Instead, it is consistent with evidence suggesting that precision in judgment grows with experience, leaving overprecision intact even as accuracy increases (Gaba et al., 2022; McKenzie et al., 2008).

Is the Crowd Wiser than the Individual?

We tested whether the crowd is wiser than the individual, measuring forecast error as the summed squared distance, weighted by probability, between a histogram forecast and the actual outcome. We then compared the average forecast error with the error of the average forecasts with a paired t-test. We find that error of the average is better than the average of the errors—that the crowd is wiser than the individual. The average of the errors (1.097) is significantly larger than the error of the average (0.928), t (115) = 8.16, p < .001. Put simply, the data suggests forecasters are collectively more accurate than they are individually.

This raises the question of whether the crowd is better calibrated than the individuals within it. We compute a crowd forecast by averaging all forecast probabilities across individuals for each indicator, year forecast made, year being forecast, and bin arrangement. We find that the overprecision we observe in individual forecasts is attenuated in the aggregated crowd forecast. In the crowd forecast, average peak confidence (42.01%) does not differ from the average hit rate (42.98%), t(241) = -1.22, p < 0.224. However, the results for our Gini and variance overprecision measures show that even aggregate forecasts are significantly more precise than accurate. This may be due to the way in which the Gini and variance for the actuals are calculated – there is no distribution of actual outcomes for each year; there is only a single number, the actual outcome for a given indicator for a given year. Our analysis compared the variance and Gini of a forecaster's distribution of probabilities to the average Gini and variance of realized outcomes across the entire epoch covered by the data.

Benchmarking Against Prior Literature

How do our results compare with other published findings of overprecision in economic forecasts? Ben-David et al. (2013) report a 36% hit rate for 80% confidence intervals, a hit rate comparable to that of other studies employing confidence intervals (Alpert & Raiffa, 1982; Cesarini et al., 2006; Glaser et al., 2013). The SPF does not ask for 80% confidence intervals, but we can infer them. We computed a mean and variance of each reported distribution. Assuming a normal distribution allows us to estimate an 80% confidence interval. The SPF's forecasts achieve a hit rate of 57.91% inside these 80% confidence intervals. Inside 90% confidence intervals, the hit rate is 66.41%. We believe that question format, not sample selection, is the most likely reason for the better calibration of the forecasts we analyze. Indeed, the forecasters who take part in the SPF may well be the same chief financial officers that responded to the Ben-David et al. (2013) survey. Instead, evidence suggests that the histogram elicitation employed by the SPF helps respondents think through the full range of possible outcomes and therefore elicits better-calibrated reports than do confidence intervals (Goldstein & Rothschild, 2014; Haran et al., 2010; Langnickel & Zeisberger, 2016).

Several prior studies share a potential weakness with ours: adverse selection. That is, because individuals choose whether to take part or which forecasts to provide, it is possible that they choose to report those beliefs of which they are most confident and therefore most likely to be overconfident. Therefore, it is worth comparing the size of the effect we report with that of laboratory experiments in which adverse selection is unlikely to play an important role. For example, Soll and Klayman (2004) report that 80% confidence intervals contain the right answer 48% of the time. We find that the overprecision we document is, if anything, smaller in size.

Discussion

We intended to analyze expert forecasts for signs of overconfidence. The results reveal that forecasters are overprecise in their forecasts of the economy. That is, they are, on average, too sure their forecasts will prove accurate. However, we find little evidence of forecaster optimism. We do find that forecasts get better over time, and closer to the moment of truth. We also find that experience improves accuracy, but that it simultaneously increases overprecision. We note that the crowd is wiser than the individual forecaster. Our findings document overprecision in expert forecasts, adding to literature on overprecision and show-

ing it to be widespread and enduring, even in the field, and even in highly consequential forecasts (Campbell & Moore, 2022; Moore & Flynn, 2008).

The results we present contradict the theory of ecological rationality, which holds that professional experience ought to minimize decision biases (Klein et al., 2017; Luan et al., 2019). The study of naturalistic decision making focuses on professionals operating within their domains of expertise. Some of this research simply takes as given that expert decisions represent the gold standard for performance and ecological rationality (Zsambok & Klein, 1997). The problem is that when expert decisions each apply to unique instances without obvious normative standards, it is difficult to assess expert bias. Our results offer a unique opportunity to assess the accuracy of a large set of expert forecasts. The results suggest that these expert forecasts fall victim to overprecision in judgment and that these experienced professionals are too sure of themselves.

This excessive certainty in forecasts can make organizational decision makers too sure of themselves too willing to draw unjustifiable conclusions (Barber & Odean, 2000). They will too readily take risky actions, too confidently introduce risky products, ignore useful advice, and too rarely consider why they might be wrong (Minson & Chen, 2021; Simon & Houghton, 2003). Understanding how overprecision influences forecasts may be helpful for interpreting those forecasts for using them to make decisions that depend on future states of the world. Our findings suggest that forecasters are consistently overprecise. One consequence may be that organizations are too slow to act when their confident predictions prove incorrect, such as the U.S. Federal Reserve Bank's delayed action to respond to inflation (Konchitchki et al., 2023; Moss, 2022).

Open Questions

Our results highlight several questions ripe for future investigation. One set of questions considers the different ways to elicit and assess forecasts. Which of our three approaches to the measurement of overprecision is the best? Which one most accurately captures forecasters' confidence? If reporting a subjective probability distribution using a histogram helps forecasters report better-calibrated and more accurate forecasts, why is that? Is it because it forces people to explicitly consider why they might be wrong (Moore, 2023)? Or is it because forecasters infer useful guidance from the bin boundaries designated by the question asker?

Another set of research questions considers differences between forecasters. Are some forecasters more biased than others? Some research has sought to identify better forecasters (Mellers et al., 2015) or the psychological traits that predict well-calibrated confidence judgments (Binnendyk & Pennycook, 2023; Lawson et al., 2023). Other research questions whether there are indeed durable traits that predict forecast overprecision (Li et al., 2023; Moore & Dev, 2017). What is the best way to train forecasters to improve their ability to question their own assumptions, think probabilistically, and incorporate others' input (Chang et al., 2016; Mellers et al., 2014)?

Given our results, organizational decision makers may want to discount forecasters' confidence when devising policy for an uncertain future. This may involve adjusting downward the confidence implied by forecasts or the confidence communicated to decision makers, or relying more on the aggregated crowd forecast distribution. Existing literature on measuring and adjusting for overconfidence in individual forecasts, for example, by exploiting heterogeneities in the data and applying different scoring rules (e.g., Schanbacher, 2014), may benefit from the findings and nuanced considerations posed in our research. New methods that combine individual forecasts to establish a crowd consensus forecasts, or on select-crowd strategies, are another promising approach (Mannes et al., 2014).

Policy makers, for their part, may seek out flexible and adaptive policies designed to respond automatically to changing economic conditions, such as unemployment insurance that increases government spending and fiscal stimulus when economic growth slows without the need for additional legislative action. In addition, organizations might want to consider training forecasters to increase awareness of their vulnerability to overprecision (Moore et al., 2017). Recent methods that examine the predictability of forecast errors are compatible with this idea – Bordalo *et al*. (2020), for example, demonstrate how individual forecasts tend to overreact, while consensus forecasts tend to underreact relative to full information and rational expectations. Krüger (2017) suggests ensemble methods, which use a collection of point forecasts to construct a forecast distribution, may hold even more promise than histogram-based forecast distributions.

The Survey of Professional Forecasters represents, in many ways, the ideal conditions for accurate and unbiased forecasts. Nevertheless, overprecision persists, despite the clarity of feedback, high stakes, professional training, and ample experience. This is consistent with prior literature suggesting that overprecision in judgment is relatively ubiquitous and difficult to debias (Grubb, 2015; Moore et al., 2015; Soll et al., 2016). Given this, we expect that overprecision is likely to be present in other important forecasts, including forecasts of geopolitical events and business performance. Improving calibration between judgment and reality in these sorts of high-stakes forecasts could have enormous collective benefits for wiser investing, planning, and policy.

Methodological Contributions

We hope that some of the methodological innovations we employ might be of use in future research. In the past, research using archival field data has routinely violated the methodological assumptions of inferential statistics: That researchers plan their analytical specifications and statistical tests a priori. Unfortunately, conducting exploratory tests using the same data employed for confirmatory hypothesis tests runs the risk of inflating the false-positive rate by capitalizing on chance. Our approach seeks to avoid this problem by splitting the data into exploration and validation subsets.

We demonstrate one possible implementation in this paper, but note that there are many variations. For example, the most common partition in computer vision looks closer to 70/30, where 70% of the data is used for training and 30% is used for validation. Depending on the goal of the research, 80/10/10 partitions are also common, where the training set is used to fit the model's parameters (i.e., the model learns to make predictions or classifications), the validation set is used to adjust the model's hyperparameters and make decisions about the model architecture (i.e., the model is tuned and evaluated), and the testing set is used to test the model (i.e., to provide an unbiased evaluation of the final model's performance). As one reviewer pointed out, a nefarious researcher interested in archival data could in theory split the data, but run analyses on both sets. There is no perfect safeguard against this. That same reviewer noted that if data are still incoming (like with the SPF), "future" data could be used as a final testing set for interested readers or reviewers.

The split-sample approach even holds some of the benefits of registered reports, wherein journal reviewers can weigh in on the best analytical approaches and research designs prior to validation. This approach may allow social sciences, such as sociology or economics, that rely heavily on field data, to benefit from the innovations of open science. As we have endeavored to demonstrate in this paper, there are cases where results from the exploration set do not replicate in the validation set. This illustrates the utility of the split-sample methodology, ensuring that we do not capitalize on chance.

Overall, our split-sample methodology offers a practical solution to the challenges of using archival field data in social science research. By adhering to the principles of open science and providing a clear distinction between exploratory and confirmatory analyses, we aim to enhance the credibility and reproducibility of research findings. The variations in data partitioning, as well as the potential for future data to serve as an additional testing set, further demonstrate the flexibility and robustness of this approach. Our paper contributes to a growing body of literature that seeks to improve research practices, and we encourage other researchers to consider these methodological innovations in their own work.

Contributions

Contributed to conception and design: S.C, D.A.M
Contributed to acquisition of data: S.C.
Contributed to analysis and interpretation of data: S.C.
Drafted and/or revised the article: S.C., D.A.M
Approved submitted version for publication: S.C, D.A.M

Acknowledgements

Thanks to Amelia Dev, Sydney Mayes, and Shreya Agarwal for capable research assistance. We are grateful to Itzhak Ben-David, Jack Soll, Dan Stone, and to seminar participants at UC Berkeley for helpful feedback.

Competing Interests

We have no competing interests to disclose. D.A.M. is an Editor at Collabra. He was not involved in the review process of this article.

Data Accessibility Statement

All data, including all associated protocols, code, and materials, are available online (https://osf.io/sj5kr/).

Submitted: June 13, 2023 PST, Accepted: January 14, 2024 PST

References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). Cambridge University Press. https://doi.org/10.1017/cbo9780511809477.022

Amore, M. D., Garofalo, O., & Martin-Sanchez, V. (2021). Failing to learn from failure: How optimism impedes entrepreneurial innovation. *Organization Science*, *32*(4), 940–964. https://doi.org/10.1287/orsc.2020.1359

Anderson, M., & Magruder, J. (2017). *Split-sample strategies for avoiding false discoveries*. National Bureau of Economic Research. https://doi.org/10.3386/w23544

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, *39*(1), 133–144. https://doi.org/10.1016/0749-5978(87)90049-5

Baillon, A., Bleichrodt, H., Keskin, U., l'Haridon, O., & Li, C. (2018). The Effect of Learning on Ambiguity Attitudes. *Management Science*, *64*(5), 2181–2198. https://doi.org/10.1287/mnsc.2016.2700

Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance*, *55*(2), 773–806. https://doi.org/10.1111/0022-1082.00226

Beaudry, P., & Willems, T. (2022). On the Macroeconomic Consequences of Over-Optimism. *American Economic Journal: Macroeconomics*, *14*(1), 38–59. https://doi.org/10.1257/mac.20190332

Ben-David, I., Graham, J. R., & Harvey, C. R. (2013). Managerial miscalibration. *Quarterly Journal of Economics*, *128*(4), 1547–1584. https://doi.org/10.1093/qje/qjt023

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., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Binnendyk, J., & Pennycook, G. (2023). Individual differences in overconfidence: A new measurement approach. *PsyArXiv*. https://doi.org/10.31234/osf.io/ugb3s

Bordalo, P., Gennaioli, N., Ma, Y., & Shleifer, A. (2020). Overreaction in macroeconomic expectations. *American Economic Review*, *110*(9), 2748–2782. https://doi.org/10.1257/aer.20181219

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078

Bruine de Bruin, W., Fischhoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty-fifty chance." *Organizational Behavior and Human Decision Processes*, *81*(1), 115–131. https://doi.org/10.1006/obhd.1999.2868

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American Economic Review*, *89*(1), 306–318. https://doi.org/10.1257/aer.89.1.306

Campbell, S., & Moore, D. (2022). High stakes overconfidence. *Academy of Management Proceedings*, *2022*(1). https://doi.org/10.5465/ambpp.2022.17329abstract

Casey, E. (2021). Are professional forecasters overconfident? *International Journal of Forecasting*, *37*(2), 716–732. https://doi.org/10.1016/j.ijforecast.2020.09.002

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107.

Cesarini, D., Sandewall, Ö., & Johannesson, M. (2006). Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior & Organization*, *61*(3), 453–470. https://doi.org/10.1016/j.jebo.2004.10.010

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2. https://doi.org/10.1016/j.cortex.2015.03.022

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, *11*(5), 509–526. https://doi.org/10.1017/s1930297500004599

Chen, J. S., Croson, D. C., Elfenbein, D. W., & Posen, H. E. (2018). The Impact of Learning and Overconfidence on Entrepreneurial Entry and Exit. *Organization Science*, *29*(6), 989–1009. https://doi.org/10.1287/orsc.2018.1225

Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research*. University of California Press.

Clements, M. P. (2014). Forecast uncertainty—Ex ante and ex post: US inflation and output growth. *Journal of Business & Economic Statistics*, *32*(2), 206–216. https://doi.org/10.1080/07350015.2013.859618

Clements, M. P. (2018). Are macroeconomic density forecasts informative? *International Journal of Forecasting*, *34*(2), 181–198. https://doi.org/10.1016/j.ijforecast.2017.10.004

Collins, H. K., Whillans, A. V., & John, L. K. (2021). Joy and rigor in behavioral science. *Organizational Behavior and Human Decision Processes*, *164*, 179–191. https://doi.org/10.1016/j.obhdp.2021.03.002

Cooper, A. C., Woo, C. Y., & Dunkelberg, W. C. (1988). Entrepreneurs' perceived chances for success. *Journal of Business Venturing*, *3*(2), 97–108. https://doi.org/10.1016/0883-9026(88)90020-1

Croushore, D. (1993). Introducing: The survey of professional forecasters. *Business Review-Federal Reserve Bank of Philadelphia*, *6*, 3–13.

Croushore, D., & Stark, T. (2019). Fifty years of the survey of professional forecasters. *Economic Insights*, *4*(4), 1–11.

Diebold, F. X., Tay, A. S., & Wallis, K. F. (1999). Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters. In *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive WJ Granger* (pp. 76–90). Oxford University Press. https://doi.org/10.1093/oso/9780198296836.003.0003

Engelberg, J., Manski, C. F., & Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business & Economic Statistics*, *27*(1), 30–41. https://doi.org/10.1198/jbes.2009.0003

Fafchamps, M., & Labonne, J. (2017). Using split samples to improve inference on causal effects. *Political Analysis*, *25*(4), 465–482. https://doi.org/10.1017/pan.2017.22

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288–299.

Flyvbjerg, B., Garbuio, M., & Lovallo, D. (2009). Delusion and deception in large infrastructure projects. *California Management Review*, *51*(2), 170–194. https://doi.org/10.2307/41166485

Gaba, V., Lee, S., Meyer-Doyle, P., & Zhao-Ding, A. (2022). *Prior Experience of Managers and Maladaptive Responses to Performance Feedback: Evidence from Mutual Funds*. Organization Science.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. https://doi.org/10.1037/0033-295x.98.4.506

Gini, C. (1912). Variabilità e mutabilità. In E. Pizetti & T. Salvemini (Eds.), *Memorie Di Metodologica Statistica* (Reprint). Libreria Eredi Virgilio Veschi.

Giordani, P., & Söderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, *47*(6), 1037–1059. https://doi.org/10.1016/s0014-2921(02)00236-2

Giordani, P., & Söderlind, P. (2006). Is there evidence of pessimism and doubt in subjective distributions? Implications for the equity premium puzzle. *Journal of Economic Dynamics and Control*, *30*(6), 1027–1043. https://doi.org/10.1016/j.jedc.2005.05.001

Glaser, M., Langer, T., & Weber, M. (2013). True Overconfidence in Interval Estimates: Evidence Based on a New Measure of Miscalibration. *Journal of Behavioral Decision Making*, *26*(5), 405–417. https://doi.org/10.1002/bdm.1773

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*(1), 1–14. https://doi.org/10.1037/e513702014-109

Grubb, M. D. (2015). Overconfident consumers in the marketplace. *Journal of Economic Perspectives*, *29*(4), 9–36. https://doi.org/10.1257/jep.29.4.9

Gutierrez, C., Åstebro, T., & Obloj, T. (2020). The impact of overconfidence and ambiguity attitude on market entry. *Organization Science*, *31*(2), 308–329. https://doi.org/10.1287/orsc.2019.1300

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, *5*(7), 467–476. https://doi.org/10.1037/e615882011-200

Hoffrage, U. (2004). Overconfidence. In R. F. Pohl (Ed.), *Cognitive illusions: Fallacies and biases in thinking, judgment, and memory* (pp. 235–254). Psychology Press.

Hofman, J. M., Goldstein, D. G., Sen, S., Poursabzi-Sangdeh, F., Allen, J., Dong, L. L., Fried, B., Gaur, H., Hoq, A., Mbazor, E., Moreira, N., Muso, C., Rapp, E., & Terrero, R. (2021). Expanding the scope of reproducibility research through data analysis replications. *Organizational Behavior and Human Decision Processes*, *164*, 192–202. https://doi.org/10.1016/j.obhdp.2020.11.003

Kahneman, Daniel. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Kahneman, Daniel. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Kahneman, Daniel, & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*(3), 582–591. https://doi.org/10.1037/0033-295x.103.3.582

Kenny, G., Kostka, T., & Masera, F. (2014). How informative are the subjective density forecasts of macroeconomists? *Journal of Forecasting*, *33*(3), 163–185. https://doi.org/10.1002/for.2281

Kenny, G., Kostka, T., & Masera, F. (2015). Density characteristics and density forecast performance: A panel analysis. *Empirical Economics*, *48*(3), 1203–1231. https://doi.org/10.1007/s00181-014-0815-9

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes*, *39*(1), 98–114. https://doi.org/10.1016/0749-5978(87)90047-1

Klein, G., Shneiderman, B., Hoffman, R. R., & Ford, K. M. (2017). Why Expertise Matters: A Response to the Challenges. *IEEE Intelligent Systems*, *32*(6), 67–73. https://doi.org/10.1109/mis.2017.4531230

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Konchitchki, Y., Moore, D., & Zhang, B. (2023). Predictable Errors in Monetary Policy Communications and Decisions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4592687

Krüger, F. (2017). Survey-based forecast distributions for Euro Area growth and inflation: Ensembles versus histograms. *Empirical Economics*, *53*(1), 235–246. https://doi.org/10.1007/s00181-017-1228-3

Lahiri, K., & Wang, J. G. (2006). Subjective probability forecasts for recessions. *Business Economics*, *41*(2), 26–37. https://doi.org/10.2145/20060204

Lamont, O. A. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior & Organization*, *48*(3), 265–280. https://doi.org/10.1016/s0167-2681(01)00219-0

Langnickel, F., & Zeisberger, S. (2016). Do we measure overconfidence? A closer look at the interval production task. *Journal of Economic Behavior & Organization*, *128*, 121–133. https://doi.org/10.1016/j.jebo.2016.04.019

Laster, D., Bennett, P., & Geoum, I. S. (1999). Rational bias in macroeconomic forecasts. *The Quarterly Journal of Economics*, *114*(1), 293–318. https://doi.org/10.1162/003355399555918

Lawson, A., Larrick, R. P., & Soll, J. B. (2023). Forms of Overconfidence: Reconciling Divergent Levels with Consistent Individual Differences. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4558486

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments tell us about the real world? *Journal of Economic Perspectives*.

Li, S., Hale, R., & Moore, D. A. (2023). *Is overconfidence an individual difference?* [Unpublished Manuscript].

Lin, S.-W., & Bier, V. M. (2008). A study of expert overconfidence. *Reliability Engineering & System Safety*, *93*(5), 711–721. https://doi.org/10.1016/j.ress.2007.03.014

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*(70), 209–219. https://doi.org/10.1080/15225437.1905.10503443

Luan, S., Reb, J., & Gigerenzer, G. (2019). Ecological Rationality: Fast-and-Frugal Heuristics for Managerial Decision Making under Uncertainty. *Academy of Management Journal*, *62*(6), 1735–1759. https://doi.org/10.5465/amj.2018.0172

MacCrimmon, K. R., & Wehrung, D. A. (1990). Characteristics of Risk Taking Executives. *Management Science*, *36*(4), 422–435. https://doi.org/10.1287/mnsc.36.4.422

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. https://doi.org/10.1037/a0036677

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior & Human Decision Processes*, *107*(2), 179–191. https://doi.org/10.1016/j.obhdp.2008.02.007

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Merkle, C. (2017). Financial overconfidence over time: Foresight, hindsight, and insight of investors. *Journal of Banking & Finance*, *84*, 68–87. https://doi.org/10.1016/j.jbankfin.2017.07.009

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L. D., Nosek, B. A., Peterson, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31.

Minson, J. A., & Chen, F. S. (2021). Receptiveness to Opposing Views: Conceptualization and Integrative Review. *Personality and Social Psychology Review*, *26*(2), 93–111. https://doi.org/10.1177/10888683211061037

Moore, D. A. (2023). Overprecision is a property of thinking systems. *Psychological Review*, *130*(5), 1339–1350. https://doi.org/10.1037/rev0000370

Moore, D. A., & Dev, A. S. (2017). Overconfidence. In V. Zeigler-Hill & T. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences*. Springer. https://doi.org/10.1007/978-3-319-28099-8_1157-1

Moore, D. A., & Flynn, F. J. (2008). The case for behavioral decision research in organizational behavior. *Annals of the Academy of Management*, *2*(1), 399–431. https://doi.org/10.5465/19416520802211636

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295x.115.2.502

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552–3565. https://doi.org/10.1287/mnsc.2016.2525

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In G. Wu & G. Keren (Eds.), *Handbook of Judgment and Decision Making* (pp. 182–209). Wiley. https://doi.org/10.1002/9781118468333.ch6

Moss, D. (2022, July 31). *Blaming Inflation on Central Banks? We Enabled Them*. Bloomberg.Com. https://www.bloomberg.com/opinion/articles/2022-07-31/central-banks-believed-too-much-in-their-inflation-fighting-powers-we-let-them

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest*, *2*, 2–9.

Niu, X., & Harvey, N. (2022). Point, interval, and density forecasts: Differences in bias, judgment noise, and overall accuracy. *Futures & Foresight Science*, *4*(3–4). https://doi.org/10.1002/ffo2.124

Norem, J. K., & Cantor, N. (1986). Defensive pessimism: Harnessing anxiety as motivation. *Journal of Personality and Social Psychology*, *51*(6), 1208–1217. https://doi.org/10.1037/0022-3514.51.6.1208

Owens, B. P., Johnson, M. D., & Mitchell, T. R. (2013). Expressed humility in organizations: Implications for performance, teams, and leadership. *Organization Science*, *24*(5), 1517–1538. https://doi.org/10.1287/orsc.1120.0795

Paik, E. T., Pollock, T. G., Boivie, S., Lange, D., & Lee, P. M. (2023). A Star Is Born: The Relationship Between Performance and Achieving Status Through Certification Contests in the Context of Equity Analysts. *Organization Science*, *34*(1), 75–99. https://doi.org/10.1287/orsc.2021.1563

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences*, *23*(1), 125–133. https://doi.org/10.1016/s0191-8869(97)00028-7

Schanbacher, P. (2014). Measuring and adjusting for overconfidence. *Decisions in Economics and Finance*, *37*(2), 423–452. https://doi.org/10.1007/s10203-013-0153-y

Schoemaker, P. J. H., & Tetlock, P. E. (2016). Superforecasting: How to upgrade your company's judgment. *Harvard Business Review*, *May*.

Schwardmann, P., & van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, *3*(10), 1055–1061. https://doi.org/10.1038/s41562-019-0666-7

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, *1*(1), 43–61. https://doi.org/10.1023/a:1009957816843

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494. https://doi.org/10.1080/01621459.1993.10476299

Silver, N. (2012). *The signal and the noise: Why so many predictions fail—But some don't*. Penguin Press.

Simon, M., & Houghton, S. M. (2003). The relationship between overconfidence and the introduction of risky products: Evidence from a field study. *Academy of Management Journal*, *46*(2), 139–149. https://doi.org/10.2307/30040610

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299–314.

Soll, J. B., Milkman, K. L., & Payne, J. W. (2016). A user's guide to debiasing. In G. Wu & G. Keren (Eds.), *Handbook of Judgment and Decision Making*. Wiley.

Stark, T. (2010, May 28). *Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters*. Federal Reserve Bank of Philadelphia. https://www.philadelphiafed.org/-/media/frbp/assets/economy/reports/research-rap/2010/realistic-evaluation-of-real-time-forecasts.pdf?la=en&hash=B4170CB97E35B102BA963CDAF926D19D

Teigen, K. H., & Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology*, *19*(4), 455–475. https://doi.org/10.1002/acp.1085

Tenney, E. R., Costa, E., Allard, A., & Vazire, S. (2021). Open science and reform practices in organizational behavior research over time (2011 to 2019). *Organizational Behavior and Human Decision Processes*, *162*, 218–223. https://doi.org/10.1016/j.obhdp.2020.10.015

Tenney, E. R., Meikle, N. L., Hunsaker, D., Anderson, C., & Moore, D. A. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, *116*(3), 396–415. https://doi.org/10.1037/pspi0000150

Tetlock, P. E., & Gardner, D. (2015). Superforecasting: The art and science of prediction. *Signal*.

Van Zant, A. B. (2021). Strategically overconfident (to a fault): How self-promotion motivates advisor confidence. *Journal of Applied Psychology*, *107*(1), 109–129. https://doi.org/10.1037/apl0000879

Windschitl, P. D., & O'Rourke, J. L. (2015). Optimism biases: Types and causes. In G. Keren & G. Wu (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 431–455). https://doi.org/10.1002/9781118468333.ch15

Zsambok, C. E., & Klein, G. A. (1997). *Naturalistic Decision Making*. Lawrence Erlbaum Associates.

Supplementary Materials

Supplemental Material

Download: https://collabra.scholasticahq.com/article/92953-overprecision-in-the-survey-of-professional-forecasters/attachment/194530.pdf?auth_token=H6KZcU9487b8rhOlhrcE

Peer Review History

Download: https://collabra.scholasticahq.com/article/92953-overprecision-in-the-survey-of-professional-forecasters/attachment/194531.docx?auth_token=H6KZcU9487b8rhOlhrcE

Response Letter

Download: https://collabra.scholasticahq.com/article/92953-overprecision-in-the-survey-of-professional-forecasters/attachment/194532.pdf?auth_token=H6KZcU9487b8rhOlhrcE