Date: December 27, 2023

Manuscript Title: Overprecision in the Survey of Professional Forecasters

Dear Editor:

Thank you for your invitation to revise and resubmit our manuscript *Overprecision in the Survey of Professional Forecasters*, for publication in *Collabra: Psychology*. In the document below, we attempt to address issues raised and describe how we revised the manuscript.

We appreciate your time and consideration. We look forward to hearing from you.

Sincerely,

[redacted]

Notes

Guidance from editor and reviewer appear below in the Times New Roman font.

Our responses are interspersed in bolded numbered paragraphs in the Calibri font.

Additions/changes to the manuscript in response to editor/reviewer feedbcak are italicized in the Calibiri font. We attempt to add verbatim passages to make the review process easier, but note that smaller changes, or changes that were not made directly in response to comments may not be captured in this letter.

Editor

Based on the very positive reaction of the first reviewer, which is consistent with my own independent evaluation, I strongly encourage you to submit a revised version of the manuscript, after adding the results pertaining to the hold-out validation sample.

I fully agree with the reviewer's claim that the manuscript describes solid research, which makes a valuable and reliable contribution. As such it is indeed a good fit and suitable for publication in Collabra: Psychology.

> **We appreciate this positive assessment of our paper, and are grateful for the opportunity to resubmit.**

Regarding guidance about how to deal with the hold-out sample - Conforming with the journal's mission and its focus on scientific rigor, transparency, and openness, my only requisites are that you: (a) explain and justify the rationale for the hold-out sample methodology you choose to endorse, and (b) pre-register your plan.

> **We thank you for this invitation. Our approach allowed us to employ an archival dataset but also discipline our analysis using a preregistered analysis plan. Moreover, by holding out part of our sample until after peer review, our approach afforded some of the benefits of a registered report, in which reviewers can have input into the analysis plan. The revised manuscript highlights these contributions in its closing paragraph.**
>
> > *"(pp 25-26) 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."

We have since added a pre-registration for the validation dataset (https://osf.io/dtuzv), and a timeline of events to aid with clarity, transparency, and reader comprehension.

(pp 7) "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/q6x47/).*
3. *In June 2023, we submitted a manuscript containing the explorations results to Collabra:Psychology. Following editor and reviewer feedback, we created an updated preregistration for the validation dataset (https://osf.io/dtuzv). In line with our preregistration for the exploration dataset, 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 in the manuscript."*

The reviewer also provides a few suggestions about possible ways to improve the readability and potential influence of this work. Although novelty and potential impact are not considered important criteria for publication in Collabra, and thus addressing these comments is not mandatory, you might want to consider some of these suggestions. For example, in the Methods, perhaps better clarify what each of

the over-precision measures independently captures, and in the Discussion, perhaps elaborate on future directions incited by the current results.

We thank you for your feedback. The revised manuscript clarifies the strengths and weaknesses of our various overprecision measures in the methods section, and includes a section on open questions in the Discussion.

(pp 11-14) "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. …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. … 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 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. …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."

"(pp 23-24) 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)?"

Reviewer 1

Thank you for the opportunity to review this paper. In my opinion, this is a nice solid paper that would be a valuable, reliable contribution to the literature, even though the contribution on the margin may not be huge. This paper does not tectonically shift my understanding of overconfidence, but it sharpens the resolution of my beliefs about overconfidence and increases my confidence in their veracity. In my opinion the paper is quite polished – it reads as though it has had close editing and revision. So in my opinion the paper is quite close.

> **We appreciate your positive assessment of the manuscript and thank you for your review and feedback.**

I would like to make a suggestion to the authors that I think would improve the impact and engagement with the paper. I do not view these as deal-breakers but I feel fairly strongly that they would be wise to consider.

This paper leaves the topic feeling fairly closed. It is quite matter-of-fact and does very little hand-holding for the reader regarding what they should find interesting here and what new questions this work raises. I really like the methodological solidity and the to-the-point nature of the paper, but I submit that the authors have abandoned the notion of reader engagement slightly too much. I'll try to explain.

We want the reader to be excited about this work – to learn from it and want to reference and build on it. In some ways, this paper currently reads like an intellectual dead-end. (I'm being a bit overly dramatic to make the point.) Even as-is, I think the paper could make impact through being a good role-model in terms of transparency and approach. But I would recommend that the authors ask themselves, what does the reader come out of this paper wanting to explore next? What strings are offered here that readers will want to pull on in future work?

The authors could offer some generative insights, in addition to the button-ed up results they present, which reflect back on previously studied phenomena. If they choose to do this, the authors can and should be very explicit about what are exploratory/suggestive results, as opposed to the critical pre-registered results. But I think adding some excitement about ways to build on this would benefit the paper quite a bit. The paper has a lot of closure, it would be nice for there to be a sense of opening in some respects as well.

> **We thank you for this helpful suggestion. The revised manuscript highlights a few provocative results that we think raise questions worthy of future investigation. We have added a couple of paragraphs to the paper's general discussion that consider some of the open questions and research opportunities highlighted by our results.**
>
> **Additionally, in our supplement, we explicitly state when results from the exploration data were not replicated in the validation data. We do not draw conclusions from conflicting results, and instead leave interpretation open to readers, who might be inspired to explore these questions further.**
>
> > *"(pp 23-24) 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)? …"

Additionally, the three measures of overprecision are interesting. The authors could potentially get a little more payoff from helping the reader understand why one might be more interesting than another. Or put differently, there is probably a lens by which each of them would be the ideal/interesting way of looking at it. Explaining that 'lens' for each of the 3 might be nice for the reader. E.g., "This measure is useful for answering an empirical question such as, …?"

We like this idea. The revised manuscript offers stronger opinions on the virtues and shortcomings of the various approaches to the measurement of overprecision, particularly in the Methods section.

(pp 11-14) "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. …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. … 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 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. …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."

Again, I really value the matter-of-factness of the paper and am not advocating for belaboring simple points in the front-end. But adding a bit of researcher interpretation might be nice. What should we make of these patterns? What questions do they raise? Given these findings, what are natural next questions to ask?

We appreciate the invitation to be more opinionated; the revision attempts to do so.

Minor
I like the middle of pg 9. The clarity and transparency is admirable and appreciated.

We appreciate your positive note on clarity and transparency.

I am a little unclear on whether this paper does justice to the hard-core decision science research on forecasting. It feels a little stuck between psychology and that crowd and clarifying that positioning might help a little bit. Although the authors already seem to be trying in this respect, so I'm not sure what the appropriate actionable recommendation there is.

Thank you for your feedback. We attempt to address your comment in our Discussion section. In summary, there is a growing literature on forecasting, and our paper attempts to incorporate some of the most useful insights from that literature. In particular, our analyses reflect the state of the art with respect to scoring forecast accuracy using incentive-compatible quadratic scoring rules as well as the aggregation of individual forecasts to capitalize on the wisdom of the crowd.