NBER WORKING PAPER SERIES

FORECASTING SOCIAL SCIENCE:
EVIDENCE FROM 100 PROJECTS

Stefano DellaVigna
Eva Vivalt

Working Paper 34493
http://www.nber.org/papers/w34493

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2025, Revised April 2026

We thank Mihai Codreanu, Anna Dreber Almenberg, Francis Diebold, Angela Duckworth, Thomas Graeber, Donald Green, Ian Krajbich, David McKenzie, Rachael Meager, Barbara Mellers, Edward Miguel, Katherine Milkman, Don Moore, Brian Nosek, Ricardo Perez- Truglia, Philip Tetlock, Mattie Toma, Severine Toussaert, Robb Willer and audiences at EIEF, FRI, Wharton, UC Berkeley, UCLA and USC. We are grateful to an outstanding team of research assistants – Leo Dai, Kevin Didi, Malek Hassouneh, Rohan Jha, and Francis Priestland. We also thank Gary Menezes, Nicholas Otis and Jo Weech for their important roles at the SSPP. We acknowledge gratefully funding from the Alfred P. Sloan Foundation, the Founders' Pledge, Open Philanthropy, and another anonymous foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Forecasting Social Science: Evidence from 100 Projects
Stefano DellaVigna and Eva Vivalt
NBER Working Paper No. 34493
November 2025, Revised April 2026
JEL No. D01, D91, O12

ABSTRACT

Forecasts about research findings affect critical scientific decisions, such as the treatments an R&D lab invests in or the statistical power of an experiment. How accurate are these forecasts, and what are the implications of potentially inaccurate beliefs? We analyze a unique data set of all 100 projects posted on the Social Science Prediction Platform from 2020 to 2024, which received 53,298 forecasts in total, including 66 projects for which we have results. We show that forecasters, on average, over-estimate treatment effects; however, the average forecast is quite predictive of the actual treatment effect. We also examine differences in accuracy. Academics have higher accuracy than non-academics, but expertise in a field does not increase accuracy. A panel of motivated repeat forecasters has higher accuracy, but this does not extend to all repeat forecasters. Confidence in the accuracy of one's forecasts is perversely associated with lower accuracy. We also document substantial cross-study correlation in accuracy and identify a group of "superforecasters". Integrating these lessons, we show how to design optimal forecasts that are appropriately shrunk and that place more weight on high-accuracy predictions. We then highlight how these optimal forecasts can be used in experimental design to increase statistical power.

Stefano DellaVigna
University of California, Berkeley
Department of Economics
and NBER
sdellavi@econ.berkeley.edu

Eva Vivalt
University of Toronto
Department of Economics
eva.vivalt@utoronto.ca

1 Introduction

Scientific research is about understanding a phenomenon, whether the impact of working from home on productivity (Bloom et al., 2024) or of nudge reminders for vaccinations (Milkman et al., 2022). In any such area, researchers build on previous findings, and scientific research can be viewed as updating prior knowledge on a research question. Yet, each paper instead tends to be seen as a study within a particular context, making it harder to examine its role in updating on a broader question. Once a paper is presented, priors are typically lost as hindsight bias (Hawkins and Hastie, 1990), making it all too easy to see a result as one that "we knew already". Losing track of priors makes it difficult to even know in which direction researchers update their beliefs in response to a paper.

There is a natural solution: collecting forecasts for research findings. If forecasts are collected before results are known, the contribution of the study becomes clearer by enabling one to compare the results to the *ex ante* forecasts. This does not guarantee that scientists will update in a Bayesian fashion, but it does make it easier for such updating to occur.[1] While this idea is simple and intuitive, the collection of forecasts of research results was rare until the creation of the Social Science Prediction Platform (SSPP) in July 2020, along the lines advocated in DellaVigna et al. (2019).[2]

From July 2020 to December 2024, the authors of 100 separate projects collected forecasts of their research results by posting them on the SSPP. These 100 projects spanned various social science fields – mostly economics, but also psychology and political science – and various methodologies – most commonly experimental studies, but also observational studies. Over these 100 projects, there were 1,482 key questions on which forecasters placed forecasts, for a total of 53,298 forecasts by 4,721 unique forecasters. This extensive data set provides a unique opportunity to examine forecast accuracy, given that for 66 of the 100 projects we are able to match forecasts with the actual research results.

Why is it important to measure the accuracy of forecasts of research findings? We highlight three contributions. First, whether forecasts of effect sizes systematically over- or under-state the impact of interventions is important because these forecasts play a key role in key scientific decisions. To give just three examples, forecasts of findings affect the choice of R&D laboratories of which treatments to implement, of a researcher of which papers to write, and of a grant agency of which projects to fund.

[1] Forecasts of research results are typically elicited as point estimates and thus generally do not capture a full prior, which is a probability distribution (Iacovone et al. (2025) is an example of a full prior elicitation). Nonetheless, some forecasts of research results do capture uncertainty (e.g., through a 10-90% range), and even forecasts of point estimates provide an indication of the priors for the research community on a topic.

[2] Some early examples of collection of forecasts of research results are Dreber et al. (2015), Groh et al. (2016), Hirshleifer et al. (2016), Camerer et al. (2018), DellaVigna and Pope (2018b), and Vivalt and Coville (2023).

If researchers systematically over-estimate the impact of interventions, they may under-power the treatment arms (DellaVigna and Linos, 2022). If they under-estimate the impact, they may inefficiently allocate too many resources to testing a particular hypothesis.

A second question is about the overall accuracy of predictions. Is the average forecast of a treatment effect informative enough about the actual result? If so, forecasts could be a useful input into the research design. Forecasts could help researchers or policymakers select which treatment arms to run given a limited sample size (or limited funding), or which outcomes to prioritize for data collection.[3]

A third question is about differences in accuracy and forecasts across groups. For example, does expertise in a particular field (or subfield) matter for forecast accuracy? Do academics have an informational advantage over non-academics? Do individuals with higher confidence in their forecasts have higher accuracy? These questions inform us of the extent to which knowledge about future scientific findings is spread among the relevant stakeholders. If different groups have different forecast accuracy, knowledge of these differences can also be used to form more informative forecasts. If so, how large are the gains from weighing these more accurate groups more heavily?

This paper draws on data from the first 100 projects posted on the SSPP to provide evidence on these three questions. This sample contains an order of magnitude more forecasts than the existing evidence, which draws on forecasts within one study, or within a smaller number of studies. Table 1 provides a list of some of these key papers. Among the important antecedents are studies on replication of experiments, showing that prediction markets, or forecasts, are quite predictive of actual replication (Dreber et al., 2015; Camerer et al., 2018). Other studies examine the prediction of the effect of multiple treatment arms within an experiment, such as within a mega-study (Milkman et al., 2021; Chu et al., 2024) or for Nudge Unit RCTs (DellaVigna and Linos, 2022). Other studies examine forecasts across a relatively small number of projects, such as Otis and De Vaan (2023) for managers and Bernard and Schoenegger (2024) for the long-term impact of development economics RCTs. Relative to these path-breaking papers, we examine a significantly larger sample of forecasts and projects and exploit unique features of the SSPP data, such as the uniform coding of confidence and expertise, as well as the linkage across projects, which allows us to provide the first evidence on the extent of cross-project accuracy in forecasts. We also provide the first calculations of the implications of using forecasts for experimental design and especially power calculations.

Our key findings are as follows. In terms of our first research question, forecasters on average

[3] As an example, Holzmeister et al. (2025) uses predictions to determine which experiments to replicate.

over-predict treatment effects, though by a smaller degree than in some of the highlighted papers in the literature (e.g., Milkman et al. 2021). The average treatment effect, normalized in standard deviation units, is 0.10 standard deviations and the average forecast is 0.18 standard deviations. The overestimation does not appear to be due to unusually low treatment effects in the SSPP sample, given that the average treatment effect of 0.10 is along the lines of the average treatment effects in papers that report results from large collections of RCTs (Vivalt, 2020; DellaVigna and Linos, 2022).

Second, do projects with higher predicted treatment effects have higher treatment effects? The answer is – largely yes. For each 0.1 higher predicted effect size in standard deviation units, the effect size is larger by 0.045. Further, there is enough information in average forecasts to justify their use for predictive purposes. This finding is in line with studies that find substantial predictive power of forecasts, such as for the replication of experiments in Camerer et al. (2016).

Third, we turn to differences in forecast accuracy, measured as the (negative of the) absolute deviation between forecasts and effect sizes (in standard deviation units). We document: (i) a sizable wisdom-of-crowds effect, with an especially steep increase in accuracy at the smallest crowd sizes, i.e., when going from just one forecast to a "crowd" of five forecasts; (ii) that academics (both PhD students and faculty) are more accurate than non-academics, but beyond that there is no further gain in accuracy from specific expertise in a subfield (e.g., development economics); (iii) a panel of largely self-selected and motivated repeat forecasters has higher accuracy, but not in a more general sample of experienced forecasters; (iv) a perverse effect of confidence, in that forecasters who are more confident than average about their forecasts for a study in fact have lower accuracy; (v) taking more time to answer a survey – a measure of effort in forecasting – is not associated with improvements in accuracy; and (vi) there is substantial cross-project accuracy, in that forecasters that are more accurate on an in-sample set of projects also do better on an out-of-sample project.

What do these results imply for optimal forecasting? We consider a researcher that aims to minimize the Mean Squared Error (MSE) in forecasting treatment effects. We show that using individual forecasts, or even wisdom-of-crowd forecasts, underperforms a constant forecast set at the mean treatment effect. However, using appropriately shrunk wisdom-of-crowd forecasts (with a shrinkage coefficient of approximately 0.4) improves accuracy by 10 percent relative to a constant-forecast benchmark. We also consider whether placing relatively more weight on forecasts by higher-accuracy forecasters can help. In doing so, a researcher faces a traditional bias-variance trade-off: higher-accuracy forecasts are more accurate on average, but represent a smaller share of the sample, and thus leads to higher variance. The optimal (out of sample) weight scales higher-accuracy forecasts by a factor of 3 and leads to a 2 percent improvement in accuracy. This improvement is larger for treatments with more forecasts, as expected given a different bias-variance trade-off.

How surprising are these results? In the spirit of our forecasting initiative, we elicit forecasts for the key findings. We also compare our results to findings in select papers in the literature which have collected forecasts within one project, and which allow for some of the comparisons we focus on – DellaVigna and Pope (2018a); Camerer et al. (2016); Milkman et al. (2021); Bessone et al. (2021); Casey et al. (2023). Several of our findings are in line with both the average forecast and with previous findings in the literature, e.g., on the (higher) accuracy of faculty and PhD students versus non-academics. For two findings – the wisdom of crowd and the impact of confidence – there is a wide discrepancy across previous studies, and the average forecast sits about mid-way compared to the previous findings. We find an impact of the wisdom of crowd on the higher end of previous estimates, and an impact of confidence more negative than previous estimates. Finally, there is no precedent for the cross-project accuracy result, and forecasters under-predict the magnitude of this finding.

In the final section, we consider the implications of these results for experimental design and specifically for power calculations and the resulting false positive report probability. For the RCTs in our sample, we back out the researchers' forecasts of effect size implied by their choice of sample size (assuming traditional 80% power calculations). We show that the experiments are significantly under-powered, with an average power of 0.44 (as opposed to 0.80), echoing the findings of Ioannidis et al. (2017). If researchers had instead used the optimal (shrunk) wisdom-of-crowd forecast to set power, they would have had a higher power of 0.65. We also show how optimal forecast can be used to determine which treatments to run and which variables to collect. As far as we are aware, this is the first evaluation of the use of forecasts for experimental design.

This paper contributes to the rapidly expanding literature on the prediction of research results with a novel, large dataset. The findings relate to previous findings in the literature, confirming some (such as on the accuracy of academics) and presenting different findings on others (such as on confidence). We also provide entirely novel evidence on cross-project accuracy. We also present the first systematic evidence on how to minimize the forecast error, including for power calculations.

The paper is also related to a vast literature on forecasts in other settings, such as about outcomes of national interest (Mellers et al., 2014; Tetlock and Gardner, 2016) or about political or economic variables (Wolfers and Zitzewitz, 2004; Ben-David et al., 2013; Weber et al., 2022; Diebold et al., 2025). Our findings relate to some key findings in this literature, such as on overconfidence, a limited information advantage of experts, and the gains in accuracy with wisdom-of-crowd forecasts. The paper also contributes to the literature on research transparency and credibility (e.g., Nosek et al. 2015; Brodeur et al. 2016; Andrews and Kasy 2019; Ferguson et al. 2023), given that forecasting can mitigate the bias against the publication of null results and inform research design (DellaVigna et al., 2019).

2 SSPP Platform and Data

SSPP Platform. The SSPP, currently funded by the Alfred P. Sloan Foundation and operational since July 2020, enables the posting of projects by Principal Investigators for the purpose of collecting predictions of research findings. Figure 1a displays a typical timeline of how coauthors of a project might collect forecasts. Authors typically post a project that either has a research design (such as a detailed pre-analysis plan) or a set of results which have not been widely presented. Before posting their project, most authors contact the SSPP team and get IRB approval for the collection of forecasts.[4] In order to collect forecasts, the authors develop a Qualtrics survey, often following exemplars and a survey guide posted on the platform. The guide emphasizes the importance of keeping the survey relatively short (5-15 minutes) and focused on "Key Questions" (KQs) which are forecasts about key findings. The median project has six key questions, e.g., about different treatments, findings for different outcome variables or for different subsamples.[5] Figure 1b displays an example of a key question.

Once the Qualtrics survey is approved by the SSPP team, it is posted on the platform and typically advertised on social media channels. The forecasts on a particular project come from a combination of users visiting the platform and SSPP panelists (see below), as well as specific individuals invited to place forecasts by the authors of a project. After the authors collect forecasts on the platform, they can download the data associated with the forecasts. They are also reminded at regular intervals to post their results on the platform to enable the measurement of the accuracy of forecasts.

Sample of Projects. From July 2020 to December 2024, the platform hosted exactly 100 different projects on a variety of social science topics, listed in Appendix Table A1. Column 1 in Table 2 displays some features of projects in this sample. While most projects are within the field of economics, 18 projects fall in psychology and 10 in political science. Within economics, the most common subfields are behavioral and experimental economics (46 projects) and development economics (34 projects).[6]

[4] The platform additionally has its own IRB approval.

[5] Roughly 74 percent of a project's KQs are author-selected, and 68 percent of projects have entirely author-selected KQs.

[6] A project can belong to multiple fields and subfields.

Figure 1c shows a word cloud of project titles, displaying the variety of topics.

An important caveat is that the SSPP sample is not a representative sample of research in economics or more broadly in the social sciences. It is more likely to feature authors that, for one reason or another, became aware of the SSPP and decided to post a project. For example, development and experimental/behavioral economics are over-represented given the early adoption by a few authors in these fields. One could also worry that papers with unexpected results – including null effects – are over-represented, as the comparison to the priors is expected to enhance the perceived learning from the project's results. While in our analysis we must take the sample as given, we document below that the average treatment effect among the projects posted is similar to that documented in large samples of RCTs. Table 2 also shows that in 26 percent of cases the authors do not know the results of the study when posting the project. As we show below, we find similar results in this subsample.

Table 2 documents other features of the SSPP projects. The majority of projects (72) elicit predictions about treatment effects, but a substantial number (43) elicit predictions about summary statistics – e.g., what percent of studies published in a set of journals in economics are about race. Across these 100 projects, there are 1,482 KQs, eliciting 53,298 forecasts by 4,721 different forecasters. The average time taken across projects, measured as the median number of minutes from survey open to survey close, is 12.4 minutes, which is within the recommended range of survey duration.

For these 100 projects, if necessary (i.e., the authors did not post the results themselves), we contacted the authors about the findings for key questions. Whenever possible, we check the results in a working paper or publication associated with the forecasts. We also code each result as capturing a treatment effect or summary statistic and renormalize the results in terms of standard deviations, as we detail below. We are ultimately able to obtain results for at least some key questions for 66 out of the 100 projects. There are two main reasons we do not have results for all projects. For the large majority of cases with missing results (28 of the 34 projects), the results are still a work in progress and findings are not yet available as of the time of writing, or we are missing a key piece of information (such as the standard deviation to renormalize). In the remaining 6 cases, even though the results are available, we cannot measure the accuracy of forecasts for a few reasons, such as the key question being forecasted being qualitative (2 projects) or otherwise not fitting our framework (4 projects).[7]

An important concern is whether the 66 projects for which results are available are representative

[7] Compared to the average project, these projects have a significantly larger set of key questions specified. They also have randomized components within some questions (e.g., of sub-questions asked) that change the question that is asked without modifying a question identifier on Qualtrics. This makes it difficult to match the key questions to results.

of the full sample. As Table 2 shows, the two sets of projects differ significantly (column 3, p-value) on only two characteristics related to the availability of results – whether a (working) paper is available for the project and whether the project was posted in the most recent year (2024). Indeed, for projects posted before 2024, results are available for 76 percent of the projects, as opposed to 39 percent for projects posted in 2024. In Columns 4-6 we document parallel patterns for the subset of projects with forecasts of treatment effects, defined more precisely below.

Coding of Standard Deviations and Treatment Effects. To compare the accuracy of forecasts across projects, it is critical to normalize magnitudes. We renormalize the key questions by the cross-sectional standard deviation in the data (of the control group in the case of an experiment). We obtain such standard deviations from the paper where possible, or from correspondence with the authors. In cases where the forecast is about a summary statistic, we also renormalize by a standard deviation. Appendix A.1 provides more details. We winsorize forecasts that are further than 1.5 standard deviations from the actual finding.[8]

For the analysis of treatment effects, we need to not only rescale in standard deviation units, but also to renorm the direction of effects where necessary. For example, consider the case of a cognitive behavioral therapy intervention (CBT) on depression. We expect CBT to decrease depression and thus to lead to a negative treatment effect on depression. We thus reverse the direction of the outcome variable, such that a higher absolute effect implies a stronger result. Similarly, we reverse the sign of the effect for studies of a police intervention on crime. As Table 2 shows, we reverse the treatment effects for 18.3 percent of key questions for a study. If we cannot determine the direction of hypothesized effects, or if there are models with conflicting predictions for the expected treatment, we do not include the key questions in the sample of normed treatments (Columns 4-6 in Table 2).[9]

Forecasters. The sample of forecasters is a combination of faculty, PhD students, and non-academics, as reported in Table 2. We ask registered users to indicate their academic status and field of expertise, but some forecasters use anonymous links to provide forecasts (as allowed by the platform), so there are forecasts for which we do not have this information. Among the forecasts with user information, 42.6 percent of forecasts are placed by PhD students and 26.2 percent by faculty. The median forecaster places predictions on just 1 study (Appendix Figure A19b).

In addition, since September 2023, the platform has a panel of forecasters. We recruited this sam-

[8]The share of forecasts winsorized is 4.0 percent and 1.8 percent in the two main samples. We also show below that the winsorization procedure is not critical to the results.

[9]This last step excludes 45 key questions, or roughly 9 percent of all treatment effects.

ple to place forecasts regularly, compensating them with $1,000 if they placed forecasts for at least 80 percent of all surveys in four consecutive quarters.[10] The sample is recruited largely from PhD students, but also includes some faculty and other researchers. The recruitment was done via an invite sent to a dozen universities and social media promotion. We implemented an initial screening of interested forecasters, ultimately yielding a sample of 92 forecasters. While the sample of active forecasters has shrunk over time, 48 were still active after one year. The panel ensured that nearly all projects posted since September 2023 received forecasts from at least 40 forecasters.

3 Forecasts of Treatment Effects

Of the 66 projects with results, 43 are about treatment effects which we can norm (in terms of direction) and renormalize in terms of standard deviations (Column 5 in Table 2). These 43 projects – 29 RCTs and 14 online or laboratory experiments – constitute the sample for this section, with 15,981 total forecasts placed on 436 key questions.[11]

Treatment Effects. Before we turn to the forecasts, we document the average treatment effect. For the sake of comparison, what is a typical treatment effect across a large set of experiments that includes unpublished results (and thus is less obviously subject to publication bias)? Vivalt (2020) provides a benchmark, spanning 307 impact evaluations in low and middle income countries, coding treatment effects in standard deviation units and reversed if necessary as is done in this paper. The average treatment effect is 0.12 standard deviations. Another benchmark is the set of 126 Nudge Unit RCTs in DellaVigna and Linos (2022), with an average treatment effect of 0.05 standard deviations.

Figure 2a displays the distribution of the treatment effects across the 437 key questions in our SSPP data set. The average treatment effect is 0.101 standard deviations, with a standard error of 0.035, and thus statistically different from 0, with a median effect size of 0.040 standard deviations. This estimate is squarely in line with the two benchmark set of findings above, and suggests that the set of experiments featured on the platform is not *prima facie* different from typical experiments.

Forecasts of Treatment Effects. For these same key questions, we present in Figure 2b the distribution of forecasts about those same treatment effects. Each observation is one of the 15,981 individual forecasts placed about one of the key questions. The average (median) forecast is 0.178 (0.100) standard deviations (s.e. of 0.008), thus larger than the average effect size documented in Figure 2a. We

[10]Since 2025, the base compensation is slated to be $400 per year, with additional incentives for accuracy.

[11]For the analysis in this Section we omit 3 key questions from one project with outlier treatment effects (respectively, the average forecast) of 0.99 (2.05), 1.83 (2.57), and 2.09 (0.71) standard deviations. We report the findings including these outliers in Appendix figures, see also Appendix A.3.

note that 11.9 percent of forecasts predict a treatment effect of exactly zero.

Thus, on average, forecasters appear to over-estimate treatment effects. It is possible, though, that this may reflect a different mixture of results in Figures 2a and 2b, since in Figure 2a one observation is a key finding, while in Figure 2b it is an individual forecast. Thus, in Figure 2c we present the distribution of the difference between each forecast and its corresponding result. We obtain a similar finding of overestimation, with an average of 0.088 standard deviations (s.e. of 0.008), close to the difference between Figures 2b and 2a. There is also evidence of overestimation for the median difference between forecast and result (0.043 standard deviations), implying that the overestimation is not a function of outliers. When considering separately the 30 RCTs versus the 14 online or lab experiments (Appendix Figures A1a-A1f), we find moderate overestimation in both samples, and we find very similar overestimation if we include the 3 outlier key questions (Appendix Figures A2a-A2c). We also find similar results whether the researchers knew the results or not, and independent of whether a working paper version of the paper existed when forecasts were collected (Appendix Figure A3).

How does this overestimation compare to findings in other papers? In some papers, such as in Milkman et al. (2021) the average forecasted treatment effect is an order of magnitude larger than the actual treatment effects. In other cases, the forecasters overpredict the treatment effect, but by a moderate amount as in Iacovone et al. (2025). In still other cases, the forecasters under-predict the effect sizes as in Groh et al. (2016). In other instances, they are on average well-calibrated (DellaVigna and Pope, 2018b). In our sample, we see all of these instances, as the distribution in Figure 2c shows. The importance of observing forecasts in a large set of diverse experiments is to estimate the frequency of these different cases. Overall, we find on average overestimation, but of moderate magnitudes.

Predictability. Our first finding thus is that on average forecasters overestimate treatment effects by about 0.08 standard deviations. A second key question is about predictability: when forecasters on average expect a larger treatment effect, is the treatment effect indeed larger? And how strong is this correlation, if any? If forecasts of treatment effects are well informed, they could be used to select which treatments to run. If, on the other hand, they are not predictive, such use is not possible.

In Figure 3 we plot, for each key question, the actual treatment effect (on the y-axis) against the average forecast (on the x-axis). To limit the role of noise, we consider the 274 key questions with at least 20 forecasts. Figure 3 displays a quite striking positive correlation, with a regression slope of 0.453 (s.e. of 0.128), and an R-squared of 0.173. That is, for each 0.1 standard deviation increase in the average forecast, the finding is about 0.045 standard deviations larger.[12] We note that the winsorization procedure centered around the results could bias the findings towards predictability, even though the share of winsorized results is small. In Appendix Figures A4a and A4b we show that the predictability is comparable if we use median forecasts (with no winsorizing) or average forecasts winsorizing at 1.5 standard deviations around 0 (as opposed to around the finding). The predictability is even larger when including the 3 outlier key questions (Appendix Figure A5) and is similar for RCTs versus online/lab experiments (Appendix Figures A6a-A6b).

How does this compare with the evidence on predictability from previous studies? The findings in individual papers are quite scattered. In some cases, as in DellaVigna and Pope (2018a), there is substantial predictability. In other settings, as in the megastudy of gyms in Milkman et al. (2021), the correlation is effectively zero. Figure 3 suggests that, while forecasters can of course get a particular treatment or even set of experimental treatments completely wrong, and while they overestimate on average, there is substantial predictable information in their average forecasts. This finding opens the door to using forecasts for the design of experiments, as we discuss below.

4 Accuracy of Forecasts

Thus far we focused on the subset of projects that are about treatment effects which can be standardized. In this section, we consider more broadly the accuracy of predictions for all key questions with results in the larger sample of 66 projects (Column 2 in Table 2).

We first compare the accuracy of individual predictions to the accuracy of the average of forecasts, to capture the strength of the wisdom of crowds. We then consider, for individual forecasts, the impact of proxies of expertise, of experience, and of confidence. We conclude with evidence on the extent of cross-project accuracy. We report the key results – the percent improvement in accuracy comparing, say, faculty to non-academics – in Figure 4a, with additional details in Appendix Figures A9a-A17b.

4.1 Average Accuracy of Individual Forecasts

Our benchmark measure of accuracy is the negative of the absolute difference between the forecast and the result, in standard deviation units. Consider an experiment with an effect size of 0.2 standard deviations. A forecaster placing a prediction of either 0.1 or 0.3 standard deviations will have an accuracy of -0.1. A forecaster placing a prediction of 0.5 standard deviations will have a lower accuracy

[12] As the figure shows, the results are very similar whether we weight by the number of responses for a key question or run an unweighted regression.

of -0.3. An alternative measure (the negative of the squared error) yields similar results.

Appendix Figures A8a and A8b display the distributions of the forecast accuracy for the key questions about (normed) treatment effects, with a mean accuracy of -0.240 standard deviations, and for the key questions about summary statistics, with a mean accuracy of -0.454 standard deviations.

4.2 Wisdom of Crowds

At least since Galton (1907), scientists have examined the value of averaging forecasts, that is, using the *wisdom of crowds*. Thus we consider how the accuracy of individual forecasts compares to the accuracy of the average of N forecasts.

We consider all key questions with at least 20 forecasts placed, covering 527 key questions over 55 projects. For each key question, we then draw, with replacement, 100 samples of N random forecasts, for N = 5, 10, or 20, and compute the average forecast in the group of N and the associated accuracy. We then average across the draws to compute the average accuracy.

Figure 4a and A9a show that the accuracy when averaging over 5 forecasts is 23.3 percent higher than the average accuracy for individual forecasts, a difference that is highly statistically significant. There are further modest improvements in accuracy moving to averages over 10 forecasts (a 27.1 percent improvement over individual forecasts) and over 20 forecasts (a 29.2 percent improvement).[13]

Thus, while the returns to averaging are large, they are clearly concave in the number of forecasts being averaged, consistent with previous findings, as in Otis and De Vaan (2023), DellaVigna and Pope (2018a) and Diebold et al. (2025). Thus collecting a sample of 10 forecasts on average already provides meaningful accuracy at a relatively low cost (as it does not require a large sample of forecasts). In Appendix Figures A10a - A10c we display the CDF of the accuracy of individual and wisdom-of-crowd predictions, split by key questions about treatment effects versus summary statistics. The improvement in accuracy due to averaging is similar in both samples.

4.3 Individual Differences in Accuracy

We now focus on differences in accuracy for individual predictions – e.g., comparing faculty to non-academics. For each comparison, we estimate OLS regressions of the measure of accuracy for a particular forecast on the relevant indicators for the characteristic – e.g., faculty status – and key question fixed effects. That is, all comparisons control for the overall level of difficulty of a particular forecast, which is important because different projects potentially draw different types of forecasters. The re-

[13] In Appendix Figure A9b we report similar results for the accuracy associated with the median forecast over N responses.

gressions cluster the standard errors at the forecaster level, to account for any correlation in accuracy for a given forecaster across questions. While each comparison is univariate, in Table 3 we display parallel results from multivariate regressions, which control for multiple determinants of accuracy. In Figure 4a we report the percent change in accuracy associated with the featured comparison – e.g., faculty versus non-academics. In Figure 4b we report the percent difference in the average forecast of treatment effects (for key questions about experimental findings), from a univariate OLS specification with key question fixed effects. A comparison of Figures 4a and 4b allows us to tie the determinants of accuracy to the determinants of (potential) overestimation of treatments.

Expertise. A natural question is whether experts are more accurate in their forecasts. In general, we rely on experts to summarize the state of science (e.g., in Panels of the National Academy) or to review the state of science (e.g., as referees), but this does not imply that they are necessarily better predictors of research findings. For example, Tetlock and Gardner (2016) documents that well-trained laypeople rival the accuracy of national security experts in predictions of security-related questions.

We consider two proxies of expertise. The first comparison is between academics and non-academics. The SSPP records the academic status of participants, which individuals can also update as part of answering a survey. We compare PhD students, faculty, and non-academics, excluding individuals without a recorded academic status. Further, we consider separately PhD students from top-10 US economics departments (about 13.6% of forecasts with information on academic status) to estimate the impact of institution.[14] The group of non-academics includes practitioners, policymakers as well as PhD holders who work outside of academia, such as in NGOs or international organizations.

As Figures 4a and A11a show, academics have a higher accuracy than non-academics, within the same questions. The difference is somewhat larger for faculty, who have a 17.1 percent higher accuracy than non-academics ($p = 0.001$), and for PhD students from top-10 institutions, who have a 19.2 percent higher accuracy than non-academics ($p < 0.001$).

To what extent is this finding due to differences in expected effect size for treatments? Given that on average forecasters over-estimate treatment effects, a group that estimates larger treatment effects will tend to have lower accuracy. Figure 4b and A11b show that indeed academics expect lower effect sizes. In particular, faculty expect 32.4 percent smaller effect sizes than non-academics.

A second type of expertise is subject matter knowledge. As mentioned above, the projects on the SSPP span various fields – mostly economics, psychology and political science – and various subfields

[14] Forecasters indicate their institution, which we complement with information from provided email addresses. We do not break down the faculty group as only 12.4% of faculty are from top-10 institutions.

– e.g., behavioral economics, experimental economics, macroeconomics, etc. For each user, we elicit their area(s) of expertise. We then measure whether there is a difference in accuracy (and in prediction of treatment effects) for projects within the area of expertise of the forecaster, that is, projects matching at least one (sub)field of expertise, compared to projects which are not.

Interestingly, Figures 4a and A12a show that there are no significant differences in accuracy when forecasters are predicting outside their field, in their field of expertise, or further in their subfield of expertise. Similarly, Figures 4b and A12b shows that there are no significant differences across these groups in the average treatment effect predicted.

Overall, any impact of expertise on accuracy is with regards to academic versus non-academic status, while we detect no difference with respect to field or subfield expertise.

Experience and Learning. In addition to expertise, accuracy could be affected by experience and learning. From this perspective, we consider the 92 forecasters that signed up to be part of the Forecaster Panel, and who placed forecasts on the majority of projects posted from September 2023 on (though some forecasters dropped out over time). These forecasters are selected on motivation and have the chance to gain experience, potentially leading to better forecasts, even though we did not provide any explicit forecasting training. Conversely, though, it is possible that some forecasters may participate out of the monetary incentive, potentially leading to lower accuracy in this group.

In Figures 4a and A13a we compare the forecasting accuracy for panelists versus non-panelists, including only projects posted after the start of the panel and, as for all other results, controlling for key question fixed effects. Panelists have 12.5 percent higher accuracy ($p = 0.009$). As Figures 4b and A13b show, while panelists tend to forecast lower treatment effects on average, this difference is not statistically significant, suggesting that this cannot fully explain the higher accuracy of panelists.

Is their higher accuracy the result of experience or selection? To examine this question, we broaden the analysis to all forecasters who place forecasts on at least 10 projects and examine their accuracy. We consider separately their first five forecasts, to study whether the frequent forecasters have different accuracy even in their early interactions, and then their later forecasts (from the 10th project onwards), to capture the impact of learning.

Figures 4a and A14a show that these frequent forecasters do not have higher accuracy than other forecasters, either initially or in their later forecasts. Thus the higher accuracy for panelists comes from a particular selection into that group, as opposed to from the experience of being a frequent forecaster.[15] To examine this further, in Table 4 we examine which features predict being a member of the panel (the dependent variable). Panel members are less likely to have extreme, winsorized forecasts, suggesting more conservative forecasts, and they also take more time in placing their forecasts.[16]

Confidence. An important question is whether individuals are knowledgeable about their accuracy. That is, when they are more confident of the accuracy of their forecasts, are their forecasts closer to the actual results? This is the natural expectation to the extent that individuals have a correct model of their knowledge. In some cases, though, predictors may have a wrong model of the world so that more confident individuals are in fact less correct, as in Kruger and Dunning (1999); Enke et al. (2023).

The SSPP platform has a suggested wording for questions eliciting forecaster confidence for a study: "*How confident are you in your predictions for this study? If you are confident, it means that you believe your predictions are very accurate.*" The respondents use a 5-point Likert scale to indicate their confidence in their predictions. In analyzing confidence, we include the projects using this question ($N = 50$) as well as a few other projects ($N = 16$) which have their own confidence question. We code the confidence level to be below-median for a project, at the median, or above-median.

Figures 4a and A15a show that the group with below-median confidence and median confidence have about the same accuracy. But, strikingly, the group with higher confidence has a 16.2 percent *lower* accuracy ($p = 0.004$) than the group with below-median confidence. This result contrasts with the typical finding in the confidence literature that higher stated confidence is associated with higher accuracy (e.g., Campbell and Moore 2024).

How does higher confidence lead to lower accuracy? Figures 4b and A15b provide at least a partial explanation. Higher confidence is associated with 54.4 percent higher predicted treatment effects. Given that on average forecasters overestimate the treatment effects, this leads to lower accuracy.

In Figure A16a we replicate these results for the surveys that use the 5-point Likert scale wording. The decrease in accuracy is driven by the responses indicating "very high" or "extreme" confidence in the accuracy of the forecasts. Within the confidence literature, we are aware of one study that finds a parallel result of lower accuracy for the top group, Moore et al. (2017).

How much does overestimation of treatment effects account for the perverse impact of confidence on accuracy? We address this with a decomposition in light of a model. As detailed in Appendix B, we assume, as in DellaVigna and Pope (2018a), that the forecasters receive a normally-distributed signal

[15] Our screening of potential forecasters for the panel was very light-touch, merely considering whether applicants were indeed researchers. Therefore, this selection into the panel is primarily self-selection.

[16] They are, mechanically, also more likely to be field or subfield experts, given that they are less likely to have missing field information.

about the experimental result, which they use to place their forecast. Forecasters differ in two ways: (i) the average extent of overestimation of treatment effects (v), and (ii) the noisiness of their signal around the actual treatment effect (σ). We estimate the parameters by maximum likelihood, allowing these parameters to differ for low-confidence versus high-confidence forecasts. Consistent with the findings above, high-confidence predictions are associated with a higher degree of overestimation of treatment effects, $v_L = 0.076$ versus $v_H = 0.137$. In addition, high-confidence forecasts are also associated with noisier signals $\sigma_L = 0.416$ versus $\sigma_H = 0.473$. In a decomposition, we attribute 30 percent of the difference in absolute accuracy to overestimation, with the rest to a less precise signal. Thus, high-confidence predictions are not just too optimistic, but also associated with a less precise signal. This suggests that higher confidence is indicative of overprecision (Moore and Healy, 2008).

As an additional piece of evidence, in Column 2 of Table 4 we regress an indicator for forecasts in the top third of confidence on various characteristics. Overall there is limited predictability, other than high confidence being positively predicted by extreme forecasts. In Section 4.4 below, we examine whether the perverse effect of confidence is due to individuals with persistent high confidence, or if it reflects within-person variation in confidence across projects.

Time Taken. A final comparison is with respect to a (noisy) proxy of effort: the number of minutes taken by a forecaster from opening a survey to completing it. We perform a median split by response time within a particular survey, after excluding completion times exceeding the minimum of 60 minutes or two standard deviations from average completion time for that project. Figures 4a and A17a show that the two groups have similar accuracy.

4.4 Multivariate Comparisons and Within-Forecaster Determinants

To what extent are the results above related? Differences in confidence, for example, may account for the difference in the accuracy of panelists, or of academics. In Column 1 of Table 3, we estimate an OLS regression of absolute accuracy on all the covariates (as well as key question fixed effects). For covariates that are not defined for a particular project – e.g., confidence for projects that did not include a question on confidence – we include an indicator variable for the missing variable so as to present results over the whole sample. We replicate all the univariate results, suggesting that the patterns found correspond to independent determinants of accuracy: faculty and PhD students from top-10 programs have higher accuracy, as do panelists, while forecasts with higher confidence have lower accuracy; also similarly, field experts and frequent forecasters are no more precise.

In the next specification (Column 2), we add forecaster fixed effects. This allows us to decompose the confidence impact into a personal attribute versus a question-specific effect: is the finding due to individuals who tend to be more confident all the time, or does it apply to cases in which individuals are more confident compared to their typical level? Once we control for forecaster fixed effects, the impact of high confidence is *positively* associated with accuracy. When individuals are more confident relative to their typical average, they are, if anything, slightly more accurate (though not significantly so). Thus, the perverse effect of confidence comes from persistent overconfidence. The specification with fixed effects does not change the results for the impact of field or subfield expertise.

Next, in Columns 3 and 4 we show that using an alternative measure of accuracy, the (negative of the) squared deviation between the forecast and the findings, we find similar results.

Finally, in Columns 5 and 6 we show that the results for the forecasts of treatment effects (for the subsample of experimental key questions) replicate the univariate findings – academics tend to predict somewhat lower treatment effects and conversely for forecasts with higher confidence.

4.5 Cross-Project Predictability

A key outstanding question for the use of forecasts is whether there is an individual skill component, that is, whether some forecasters are persistently more accurate. The previous results identifying accuracy differences by confidence and expertise suggest that there may be such a component, but do not examine it directly. In the context of forecasting questions of national security interest, Tetlock and Gardner (2016) finds substantial cross-question predictability and labels the high-accuracy predictors "Superforecasters". Until now, to the best of our knowledge, no one has been able to study this for research forecasts, since doing so requires a linkage across projects, which only the SSPP provides.

To study the extent of predictability, we test whether a forecaster's *relative* accuracy in one set of studies predicts their *relative* accuracy in a held-out study. Consider forecasters that have provided forecasts for at least five projects with results available, and for key questions with at least 20 forecasts, yielding a sample of 124 forecasters. We examine the accuracy in project j as a function of the average accuracy in 4 other randomly sampled projects. Since projects differ substantially on accuracy, to reduce noise we measure relative accuracy compared to other forecasters by considering demeaned accuracy in a project. We repeat this procedure for all projects j that a given forecaster has in the sample. We report additional details about this procedure in Appendix C.

Figure 5 presents a bin scatter plot comparing the relative accuracy for the leave-out project j

on the y-axis versus the average relative accuracy for the 4 in-sample projects on the x-axis. The figure also includes a regression line estimated from the underlying data. Without any cross-project predictability, the regression slope should be zero. Perfect predictability (which is highly implausible given that the estimates of accuracy are noisy) would imply a coefficient of 1. We estimate a coefficient of 0.469 (s.e. of 0.057), indicating an economically and statistically significant degree of cross-project predictability. Since one may worry about some mechanical correlation, we also report a placebo regression which has the same values on the x-axis, but replaces the prediction for project j with the prediction by another forecaster, also for project j. The placebo line is flat. We conclude that there is indeed substantial cross-project predictability.

We can also display the accuracy of prediction for a leave-out project, splitting into terciles of accuracy over the in-sample predictions. As Appendix Figure A18a shows, the group in the top third of accuracy has a 33.3 percent higher accuracy (out-of-sample) than the group in the bottom third. Appendix Figure A18b documents that the groups with high (out-of-sample) accuracy are less likely to overestimate treatment effects, contributing to higher accuracy.

Thus, as in Tetlock and Gardner (2016), there are "superforecasters" with reliably higher accuracy. Who are these superforecasters? In the sample of 124 forecasters for whom we can apply the procedure, we indicate with a superforecaster indicator any forecast with an in-sample accuracy score greater than 0. In Table 4 we show that high-confidence or a winsorized forecast are negatively predictive of a high-accuracy forecast, consistent with the previous findings.

5 Maximizing the Accuracy of Forecasts

To use forecasts for experimental design or power calculations, maximizing the accuracy is critical. Building on the evidence above, we evaluate the accuracy of various forecasts in terms of the Mean Squared Error (MSE). For each of the 44 RCTs with results, we leave study i out and use all other studies to create a forecasting model, and we then estimate its accuracy in study i. We measure the accuracy across all key questions in a study, assigning equal weight to each key question.

Table 5 shows the average MSE under a number of alternative benchmarks. Row 1 presents the MSE if we use one randomly drawn forecast. In comparison, as Row 2 shows, using the average (wisdom-of-crowd) forecast leads to a 57.8 percent improvement in accuracy, as measured by the MSE. This improvement is not surprising given that individual forecasts are very noisy predictors.

As we showed above, the average forecast, while predictive of treatment effects, on average overstates the treatment effect, leading to lower accuracy. In Row 3, we present the MSE associated with the wisdom-of-crowd forecast, recentered so that on average across studies it matches the average actual treatment effect. The recentering improved the MSE by 8.3 percent.

How well do wisdom-of-crowd forecasts perform relative to a constant (well calibrated) forecast? In Row 4 we compute the MSE where for each study i we use the average treatment effect over all other studies. Strikingly, the MSE is 8.9 percent lower than the MSE associated with the re-centered wisdom-of-crowd forecast and 16.9 percent lower than the MSE associated with the baseline wisdom-of-crowd forecast. This result would seem to indicate that forecasts, even re-centered, do not carry enough predictive information, which contrasts with the finding of a R^2 of 0.173 in Figure 3.

The solution for the puzzle is that the optimal forecast needs to be shrunk, as in Figure 3 which indicates a weight of about 0.4 on the wisdom-of-crowd forecast. In row 5, we compute an optimal predictor. Formally, for study i we take all other studies and regress the results on the average forecast and take the regression slope and intercept, which we use to make a prediction for study i. Averaging over all studies, we find an MSE of 0.0468, 10.4 percent lower than the constant forecast. One can thus sizably improve the accuracy by using forecasts, but it requires appropriate shrinkage.

To highlight the role of the number of forecasts used, in row 6 we compute the appropriately shrunk forecast but using only 10 forecasts for each study. The MSE is very close to the one in Row 5, indicating that doing an appropriate shrinkage is more important than collecting many forecasts.

Next, we consider whether one can improve the accuracy by giving more weight to predictors with higher *ex ante* accuracy. For each study i, we use all other studies to estimate the predictors of the MSE as in Column 3 of Table 5 and create a predicted rank of forecasters in study i. Here we face a bias-variance trade off. If we restrict ourselves to forecasts by higher-accuracy predictors (individuals with lower confidence, academics, and panelists), we obtain a forecast with higher accuracy, but at the cost of a smaller sample size. Row 8 shows that, when we use only forecasts in the top 50 percent of predicted accuracy, the accuracy is barely (0.6 percent) higher compared to the wisdom-of-crowd forecast (Row 7): the improvement in bias is only slightly higher than the increase in variance.

To find the optimal bias-variance trade off, we estimate the optimal weight α on forecasts in the top half of predicted accuracy. Across all studies and leaving study i out, we find the weight α that minimizes the MSE and use that weight for study i. The average optimal α equals 3, implying that forecasts in the top half gets 3 times as much weight. As Row 9 shows, this specification leads to a 1.8 percent accuracy improvement relative to the forecast in row 8 and a 2.5 percent improvement relative to the unweighted wisdom-of-crowd forecast (Row 7).

In the final step, we ask how much the optimal weighing of higher-accuracy forecasts matters once we incorporate optimal shrinking. Compared to the baseline shrinking MSE (reproduced in Row 10), in Row 11 we display the MSE which incorporates both shrinking and optimal weighing of the top 50 percent of forecasts. Formally, for each study i we use all other studies to find the optimal α, allowing for a different shrinkage for each α. The estimate that combines both weighting and shrinking improves the MSE by 2.1 percent relative to just shrinking (Row 10).

Further, in Row 12 we allow for a separate weight for forecasters in the predicted top 25 percent of accuracy (estimated to be optimally over-weighted by a factor of 2.5) and in the bottom 25 percent of accuracy (estimated to be optimally underweighted by a factor of 0.38). The resulting forecasts further lower the MSE by 2.6 percent. Overall, incorporating *ex ante* differences in accuracy increases the predicted accuracy by 4.7 percent compared to just using shrinkage (Row 10). Figure 7 shows a parallel plot to Figure 3 predicting treatment effects, but using the optimal forecast as in Row 12. The figure shows the shrinkage of the forecasts, and the tighter predictive accuracy.

In Appendix Table A4 we document similar results if we weight the key questions by the number of forecasts. In such a sample, which over-represents key questions with more forecasts, the accuracy improvement associated with using forecasters in the top group is even larger, as expected given that the variance increase associated with using the top forecasts will be lower.

Overall, forecasts can play an important role in predicting research results, provided one applies an appropriate shrinkage adjustment. Incorporating information on cross-sectional differences further improves accuracy. In Section 7, we consider the role of this in power calculations.

6 Comparison to Forecasts and Literature

Which of these results are consistent with previous findings in the literature from (smaller) samples of forecasts of research results? Which results are anticipated by forecasters? In this section, we compare our main findings to parallel results from the literature. We also compare the results to 52 forecasts from a survey run on the SSPP platform between July 2025 and August 2025.

We focus on the results in Section 4 on absolute forecast accuracy and specifically: (i) wisdom-of-crowds forecasts with 5 forecasts relative to one forecast; (ii) wisdom-of-crowds forecasts with 20 forecasts relative to one forecast; (iii) faculty members' forecasts versus non-academics' forecasts; (iv) PhD students' forecasts versus non-academics' forecasts; (v) subfield experts' forecasts versus non-experts' forecasts; (vi) forecasts by panelists versus non-panelists; (vii) forecasts by frequent forecasters versus infrequent forecasters; (viii) forecasts of projects in which the forecaster has high versus low confidence; (ix) forecasts by superforecasters versus forecasts by forecasters with low cross-project accuracy; and (x) forecasts in the top half of time taken compared to bottom half of time taken.

We compare to findings in a few select papers which have collected forecasts and which allow for some of the same comparisons we focus on – DellaVigna and Pope (2018a); Camerer et al. (2016); Milkman et al. (2021); Bessone et al. (2021); Casey et al. (2023). These papers provide a benchmark for the estimates of the wisdom-of-crowds effect, expertise, the role of confidence, and time taken, though they do not have estimates for the impact of panel membership, experience, or "superforecasters".[17] For these and all other topics, we compare to the average forecast we collected on the SSPP.

In Figure 6 we present the actual results in percent change on the x-axis, and on the y-axis we present both the average prediction from the 52 forecasters, as well as the featured findings in the literature. Accurate predictions will lie close to the 45 degree line. Several of our findings are in line with both the average forecast and with previous findings, e.g., on the (higher) accuracy of faculty and PhD students. There is a substantial discrepancy across previous studies in terms of the magnitude of the gain associated with the wisdom of crowds, from 0 to 50 percent, and the average forecast in our SSPP survey sits about mid-way among the previous literature. We find an impact of the wisdom of crowds on the higher end of previous estimates. The impact of confidence is also quite heterogeneous across previous studies, with two studies finding a positive effect and one study with a negative effect of confidence on accuracy, with the average forecast wedged in between these previous estimates. Our estimate is actually more negative than all three previous estimates.

We then turn to the findings for which there is no precedent (known to us) in the literature. The forecasters under-predict the extent of the cross-project accuracy. They instead predict near perfectly the accuracy advantage of panelists, but they over-predict the impact of experience.

We can speculate about the result on overconfidence. Why is confidence sometimes positively predictive of accuracy and other times negatively predictive? In Camerer et al. (2018) and DellaVigna and Pope (2018a), higher confidence is associated with higher accuracy, while in Casey et al. (2023) it is negatively correlated with accuracy. While many factors could be at play, a key difference could be that in DellaVigna and Pope (2018a) the predictors were informed of a few key average treatment effects and had to predict the treatment effects for 15 other interventions; thus the predictions largely

[17] Appendix E contains details relating to the comparison with the literature.

amounted to ranking the effectiveness of interventions. In comparison, for most projects, as in Casey et al. (2023), the forecast is largely about the level of the treatment effect. A conjecture is that forecasters are better calibrated about the prediction of which treatment is more effective, as opposed to predicting the level of treatment effects. As Enke et al. (2023) show, there are tasks where higher confidence is associated with higher accuracy and others where the opposite is true. It seems that making predictions about the level of treatment effects is an example of the Dunning-Kruger effect (Kruger and Dunning, 1999) – unsophisticated forecasters are unaware of their lack of knowledge.

Finally, we asked forecasters to predict the average treatment effect and the average prediction about the treatment effect. Forecasters are well calibrated with respect to the average treatment effect (0.102 versus 0.100) and also in conjecturing that forecasters on average over-predict (0.163 versus 0.178).[18] All in all, forecasters have a good understanding of treatment effect findings and predictions.

7 Implications for Experimental Design

In this section, we highlight how forecasts can be used for experimental design, focusing on power calculations and the decision of which treatments to run. Consider a researcher that determines the sample size of an RCT by doing a statistical power calculation aiming for 80% power given an effect size. Traditionally, researchers use an effect size from estimates in the literature or just a guess. We argue that using forecasts for the effect size would improve the experimental design.

For our sample, we use the 43 experiments with results and in particular the 418 (out of 437 total) key questions for which we could find the sample sizes. As a first step we infer the effect size researchers presumably used in their power calculations (provided they did one). Specifically, given the sample size, for each key question we computed the implied effect size in standard deviations, and we take the median across key questions for a given study. As Column 1 in Table 6 shows, the average sample size for a treatment cell is 2,330 subjects, implying a forecasted effect size of 0.190 standard deviations. In comparison, the average treatment effect is 0.095 standard deviations, and the actual average statistical power is 0.440, well below the envisioned 0.80 power. We thus replicate the finding of Ioannidis et al. (2017) that on average experiments in economics are underpowered.[19]

[18]We randomized whether we informed predictors of the average treatment effect found in of RCTs found in two large studies cited above (Vivalt, 2020; DellaVigna and Linos, 2022). The group treated with this information predicted treatment effects of 0.085 and forecasts of 0.133, while the control group predicted treatment effects of 0.120 and forecasts of 0.194. Thus, while the treatment lowered the predicted treatment effects in both levels and forecasts, in both groups the predictors are aware that forecasters will tend to overestimate treatment effects.

[19]It is possible that researchers do not truly envision an effect size of 0.190 on average and instead face constraints in determining sample size. Our general argument – that researchers can use forecasts to optimally allocate sample – will still go through.

In Column 2, we consider the statistical power if researchers use for the effect size the maximum of 0.05 and the optimal forecast, as in Row 12 of Table 5. The reason for bounding the effect size at 0.05 is that it would typically be prohibitively costly to appropriately power a study to detect an effect size of, say, 0.01 standard deviations. The average forecasted effect size, thus bounded, is 0.104, resulting in an average sample size of 5,116 and a statistical power of 0.647. Thus, using forecasts leads to a substantially higher statistical power, though at the cost of the cost of doubling the sample size.

The power calculations in Column 2 make use of the forecasts to determine the sample size, but not in determining which treatments to run. In Column 3, instead, we assume that researchers only run treatments and consider outcome variables where the optimal forecast is above 0.05 standard deviations, as treatments with forecasts below the threshold would require very large sample sizes to be adequately powered. We acknowledge that the 0.05 threshold is arbitrary, as the optimal decision will depend on the cost of collecting additional observations, relative to the return. Nonetheless, this simple rule illustrates the possible benefits of using forecasts to decides which treatments to run. In this subsample, the predicted effect size is 0.115 standard deviations, and the implied sample size is 3,667, only 55 percent larger than the actual sample size used in the experiments. The resulting statistical power is 0.632, a sizable improvement compared to the one in the data, at only a small additional cost of extra sample. This improvement is related to the fact that optimal forecasts are significantly related to the actual effect size and can be effectively used to make treatment decisions.

Column 4 shows the statistical power for the key questions that the optimal forecast would induce researchers to drop (as they lie below the 0.05 sd threshold). These treatments do indeed have a lower treatment effect of 0.040, compared to 0.115 for the treatments that the optimal forecasts would suggest running. Indeed, the studies in this subsample are badly underpowered, with a power of 0.250. Thus the optimal forecast can help weed out treatments and outcomes that are very likely underpowered.

While much remains to be explored, this section shows that forecasts can be used to at least partially address the problem that on average experiments are statistically underpowered.

8 Conclusion

In this paper, we considered a unique data set of 53,298 forecasts of 1,482 key questions across 100 projects, each of which collected predictions from academics and non-academics about the findings of their paper. Authors generally use the collected forecasts to compare their findings to the average (or median) forecast collected to assess the novelty of the results in their paper.

The linkage across projects and forecasters allows us to address several questions across a broad set of studies: How accurate are forecasts? Are they predictive of treatment effects? How much does accuracy vary across forecasters – would it be useful to focus on subsets of forecasters with more accurate forecasts? How can one use forecasts for experimental design?

In short, the forecasts contain a significant amount of information that could be used for study design. Although the forecasts overestimated treatment effects on average, they were predictive of the actual treatment effect. We also detect clear differences in forecast accuracy. Academics have higher accuracy than non-academics, but expertise in a field or subfield does not increase accuracy. Confidence in one's forecasts is perversely associated with lower accuracy. Finally, our repeat forecaster panelists had higher accuracy than other repeat forecasters who did not participate in the panel. We show how to integrate these findings to minimize the MSE of treatment effect prediction, using appropriately shrunk average forecasts which overweight predictions by ex-ante more accurate forecasters.

Some of our findings line up with previous studies, such as the wisdom-of-crowds effect, while other findings differ from previous ones, such as the perverse impact of forecaster confidence. Finally, other findings are entirely novel, given that previous studies could not examine cross-study accuracy (which we find to be substantial), or did not explore the implications for experimental design.

There are some important caveats to these results. First, the selection of projects on the SSPP is certainly not random. As the platform expands beyond the (relatively) early adopters, it will be useful to examine if differences in the selection of studies change the patterns above. Second, at present, there is no protocol for training and improving the accuracy of forecasters (for example, the ones with high confidence). It will be interesting to study whether training protocols help with accuracy. Third, forecasts do not currently benefit from AI tools that are likely to be of help in the future (Hewitt et al., 2024). It will be interesting to measure the performance of these tools as they develop.

With these caveats in mind, we hope that the evidence in this paper will help further the use of research forecasts, including expanding their use in new directions, such as for study design.

Table 1: Select Existing Evidence on Forecasts of Research Results

	Predicted Outcome		N_s	N_u	N_f	Forecaster Attributes	Cross-study Accuracy	Confidence	Wisdom of Crowds
	(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)
The SSPP	**Exp. effect & Sum. stat. (SD)**	**With results** **All**	**66** **100**	**2,369** **4,721**	**34,952** **53,298**	✓	✓	✓	✓
Camerer et al., 2018	Rep. prob. (cont. [0,1])		21	211	N/A (Pred. market)				(Market vs ind.)
DellaVigna and Linos, 2022	Exp. effect (ppt)		14	237	3*237	✓			
Otis and DeVaan, 2023	Exp. effect (SD)		6	681	4,528				
Bernard and Schoenegger, 2024	Exp. effect (SD & more)		7	91 Academics	≈ 4,600	✓		✓	✓
Milkman et al., 2022	Exp. effect (ppt)		1	24 Scientists	528	(Experts vs laypeople)			

Notes: Table 1 provides a list of some key papers with evidence on forecasts of research results, comparing their sample of predicted outcomes, projects/studies (N_s), forecasters (N_u), and forecasts (N_f) to that of the SSPP. Counts displayed for the SSPP are for all projects and forecasts for which results could be collected, while those in parentheses include the full set of forecasts across all 100 projects. The scope of analysis of each paper is also compared to that of the SSPP (columns 5-8).

Table 2: Summary Statistics for Main Analysis Samples

	All Projects	(1) w/ Results	Diff. (1) - (2)	Normed Treat	(4) w/ Results	Diff. (4) - (5)
	(1)	(2)	(3)	(4)	(5)	(6)
Posted in 2024 (%)	26.0	15.2	$p = 0.10$	19.6	11.6	$p = 0.30$
Number of Key Questions	14.8	10.9	$p = 0.52$	9.5	10.1	$p = 0.78$
Median Time Taken (Min.)	12.4	13.0	$p = 0.61$	13.5	13.9	$p = 0.83$
Is Economics (%)	85.0	84.8	$p = 0.98$	88.2	86.0	$p = 0.75$
Is Dev. Economics (%)	34.0	34.8	$p = 0.91$	52.9	46.5	$p = 0.54$
Is Beh./Exp. Economics (%)	46.0	40.9	$p = 0.52$	43.1	41.9	$p = 0.90$
Has Paper (%)	66.0	86.4	$p < 0.01$	72.5	83.7	$p = 0.20$
Results Known by Authors (%)	74.0	78.8	$p = 0.48$	78.4	81.4	$p = 0.72$
Has Treatment (%)	72.0	77.3	$p = 0.45$	100.0	100.0	-
Is Reversed (%)	-	-	-	18.3	17.6	$p = 0.92$
Faculty Forecasts (%)	26.2	26.5	$p = 0.90$	25.7	25.9	$p = 0.97$
PhD Students Forecasts (%)	42.6	42.5	$p = 0.99$	44.5	44.7	$p = 0.97$
Top-10 PhD Students Forecasts (%)	14.3	13.6	$p = 0.64$	14.0	13.7	$p = 0.88$
Panelists Forecasts (%)	24.2	20.3	$p = 0.47$	22.7	18.7	$p = 0.58$
N - Projects	100	66	-	51	43	-
N - Key Questions	1,482	722	-	483	436	-
N - Forecasters	4,721	2,369	-	1,571	1,419	-
N - Responses	7,882	4,250	-	3,070	2,566	-
N - Forecasts	53,298	34,952	-	18,750	15,981	-

Notes: Table 2 provides a list of summary statistics for the sample of all projects (column 1), all projects with results (column 2), all projects of normed treatment effects (column 4), and all projects of normed treatment effects with results (column 5). Aggregating at the project-level, the following summary statistics are provided: the percent of projects posted in 2024, the average number of key questions per project, the median time taken (in minutes) to complete a response, the percent of economics projects and within economics the percent of development and behavioral/experimental economics projects, the percent of projects with an associated (working) paper, the percent of projects with at least one treatment key question, the percent of key questions that are reversed, the percent of forecasts from faculty, PhD students, PhD students from top-10 programs, and panelists. Columns 3 and 6 provide the p-value for a two-tailed t-test between Columns 1 and 2 and between 4 and 5, respectively.

Table 3: Determinants of Absolute, Squared Forecast Accuracy & Forecasts of Treatments

	Absolute Forecast Accuracy		Squared Forecast Accuracy		Forecasts of Treatments	
	(1)	(2)	(3)	(4)	(5)	(6)
PhD Students	0.054**		0.076**		−0.038*	
	(0.023)		(0.032)		(0.021)	
PhD Students Top-10	0.040**		0.049**		−0.030*	
	(0.017)		(0.023)		(0.018)	
Faculty	0.073***		0.097***		−0.051**	
	(0.023)		(0.032)		(0.022)	
Field Experts	−0.016	−0.051	−0.034	−0.103	0.014	−0.014
	(0.020)	(0.038)	(0.027)	(0.063)	(0.021)	(0.048)
Subfield Experts	−0.008	−0.042	−0.024	−0.091	0.023	−0.021
	(0.020)	(0.039)	(0.028)	(0.065)	(0.022)	(0.048)
Time Taken High	0.005	−0.006	0.008	−0.010	0.018*	0.021
	(0.008)	(0.012)	(0.011)	(0.018)	(0.009)	(0.014)
Freq. Forecasters Resp. ≤ 5	−0.011	−0.019	−0.018	−0.025	0.030	0.028
	(0.016)	(0.019)	(0.021)	(0.029)	(0.019)	(0.031)
Freq. Forecasters Resp. > 10	−0.022	0.060*	−0.039	0.097	0.024	−0.090*
	(0.023)	(0.036)	(0.033)	(0.060)	(0.029)	(0.050)
Panelists	0.061***		0.096***		−0.018	
	(0.021)		(0.031)		(0.024)	
Confidence Median	−0.003	0.014	−0.005	0.020	0.035***	0.030*
	(0.012)	(0.015)	(0.017)	(0.022)	(0.013)	(0.017)
Confidence High	−0.050***	0.021	−0.069***	0.031	0.069***	0.016
	(0.016)	(0.017)	(0.023)	(0.026)	(0.018)	(0.019)
Key Question Fixed Effects	✓	✓	✓	✓	✓	✓
Forecaster Fixed Effects		✓		✓		✓
Normed Treatments					✓	
Mean Dependent Variable	-0.354	-0.354	-0.280	-0.280	0.186	0.186
% Forecasts Winsorized	4.69	4.69	4.69	4.69	1.80	1.80
R^2	0.409	0.551	0.383	0.515	0.246	0.440
N - Projects	66	66	66	66	51	51
N - Key Questions	722	722	722	722	483	483
N - Forecasters	2,369	2,369	2,369	2,369	1,571	1,571
N - Responses	4,250	4,250	4,250	4,250	3,070	3,070
N - Forecasts	34,952	34,952	34,952	34,952	18,750	18,750

Notes: Table 3 provides estimates of multivariate OLS regressions on the determinants of absolute forecast accuracy (columns 1-2), squared forecast accuracy (columns 3-4), and normed forecasts of treatments (columns 5-6). All regressions include covariates for: academic status (Non-Academics (omitted), PhD Students, PhD students from top-10 PhD programs, Faculty), field expertise (Non-Experts (omitted), Field Experts, Subfield Experts), time taken (Low, i.e. below-median within a project (omitted), High, i.e. above-median within a project), frequent forecasters (Less than 5 responses (omitted), More than 10 responses and first 5, More than 10 responses and beyond first 10), panel membership (Non-Panelists (omitted), Panelists), and confidence (Low, i.e. below-median within a project (omitted), Median, High, i.e. above-median within a project). Note that, for a given forecaster, their academic and panelist status can be updated from one survey to another, but given limited variation, columns 2, 4, and 6 – which contain forecaster fixed effects – do not display the estimates for those variables. Missing levels, e.g. for a project without a confidence question specified, are accounted for with missing indicators (omitted). Key question fixed effects are included in all regressions, while forecaster fixed effects are included in the even numbered columns. Standard errors (in parentheses) are clustered at the forecaster level. Significance levels are indicated by * = 10%, ** = 5%, and *** = 1%.

Table 4: Determinants of Panelists, High Confidence & High Accuracy

	Panelists	High Confidence	High Accuracy
	(1)	(2)	(3)
PhD Students	0.063	−0.083	0.064
	(0.064)	(0.074)	(0.066)
PhD Students Top-10	0.076	−0.047	−0.030
	(0.058)	(0.071)	(0.075)
Faculty	−0.072	−0.100	0.078
	(0.077)	(0.072)	(0.070)
Field Experts	0.440***	−0.104**	0.059
	(0.073)	(0.050)	(0.059)
Subfield Experts	0.344***	0.038	0.077
	(0.074)	(0.054)	(0.053)
Time Taken High	0.066**	0.054**	0.008
	(0.029)	(0.024)	(0.040)
Confidence Median	0.014		−0.064
	(0.036)		(0.044)
Confidence High	−0.021		−0.196***
	(0.046)		(0.058)
Freq. Forecasters Resp. ≤ 5		0.128**	0.155
		(0.051)	(0.136)
Freq. Forecasters Resp. > 10		0.037	0.073
		(0.069)	(0.137)
Panelists		−0.089	0.227***
		(0.059)	(0.048)
Forecast Winsorized	−0.089**	0.129***	−0.135***
	(0.036)	(0.038)	(0.032)
Key Question Fixed Effects	✓	✓	✓
Forecaster Fixed Effects			
Mean Dependent Variable	0.634	0.360	0.313
R^2	0.350	0.091	0.126
N - Projects	36	66	53
N - Key Questions	310	1,201	483
N - Forecasters	901	2,175	124
N - Responses	2,906	4,214	1,592
N - Forecasts	21,816	37,366	14,025

Notes: Table 4 estimates multivariate OLS regressions on the determinants of being a panelist (Column 1), high confidence (column 2) and high forecast accuracy (column 3). High-forecast accuracy is taken as the top-third of forecasts in in-sample accuracy levels (as defined in Figure 5). High-confidence is a response indicating above-median confidence in a particular project. Column 3 considers forecasts for all projects with a minimum of 20 responses, and for all forecasters with a minimum of 5 responses with results (identical to Figure 5), while column 2 considers forecasts for projects that have a confidence question specified. All regressions include controls as in Table 3 as well as key question fixed effects. Standard errors (in parentheses) are clustered at the forecaster level. Significance levels are indicated by * = 10%, ** = 5%, and *** = 1%.

Table 5: Maximizing the Predicted Accuracy of Forecasts

Row	Forecasting Method	MSE	Percentage Difference
1	Individual forecast	0.1483	-
2	Average forecast (wisdom of the crowd)	0.0626	57.81%
3	Demeaned average forecast (OOS) Avg. OOS constant: 0.088	0.0574	8.30%
4	Constant forecast (average of other studies)	0.0522	8.93%
5	Shrunk average forecast Avg. OOS β: 0.404, intercept: 0.028	0.0468	10.38%
6	Shrunk average forecast (N=10) Avg. OOS β: 0.322, intercept: 0.042	0.0471	-0.50%
7	Average forecast (wisdom of the crowd)	0.0626	-
8	Average of top 50% forecasters (OOS-ranked)	0.0622	0.62%
9	Optimally weighted forecast (avg OOS α=3, top 50%)	0.0610	1.81%
10	Shrunk average forecast Avg. OOS β: 0.404, intercept: 0.028	0.0468	-
11	Shrunk optimally weighted forecast (top 50%) Optimal weight α: 3.22, Avg. OOS β: 0.399, intercept: 0.030	0.0458	2.11%
12	Shrunk optimally weighted forecast (top/bottom 25%) Optimal weights OOS α: 2.50, γ: 0.38, Avg. OOS β: 0.358, intercept: 0.037	0.0446	2.60%

Notes: Table 5 presents the out-of-sample prediction accuracy, measured by Mean Squared Error (MSE), of different forecasts. Row 1 evaluates the accuracy using one (randomly drawn) forecast. Row 2 calculates the accuracy of the average forecast per key question (the wisdom of the crowd). Row 3 adjusts this average forecast by subtracting the mean forecast bias observed across all other out-of-sample studies. Row 4 evaluates a constant forecast equal to the simple average of the realized results from all other out-of-sample studies. Row 5 applies an out-of-sample shrinkage adjustment, transforming the average forecast using coefficients estimated from a regression of the treatment effect on the average forecast for all other studies. Row 6 applies this same shrinkage procedure but simulates the average forecast using bootstrapped sub-samples of N=10 responses. Row 7 repeats the simple average forecast from Row 2. Row 8 calculates the average forecast using only respondents ranked in the top 50% for predicted out-of-sample accuracy. The top 50% is estimated via respondent characteristics from all other out-of-sample studies with a regression as in Table 3, Column 3. Row 9 computes an optimally weighted average where the top 50% of forecasters receive an out-of-sample optimized weight (α), while the rest receive a weight of 1. Row 10 repeats the baseline shrunk average forecast from Row 5. Row 11 applies the out-of-sample shrinkage regression adjustment to the optimally weighted forecast from Row 9. Row 12 applies the out-of-sample shrinkage adjustment to a weighted average that assigns independent optimal weights to both the top 25% (α) and bottom 25% (γ) of forecasters.

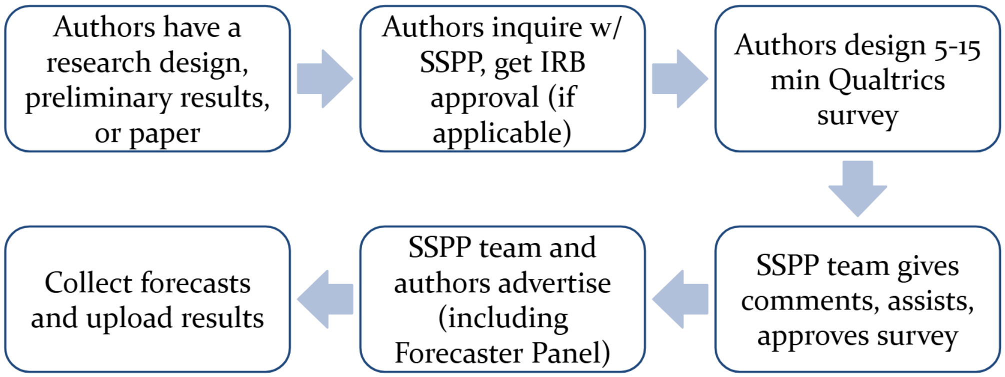
Table 6: Power Calculations Under Different Scenarios

Metric	Study Results (1)	Optimal Scenarios: Using Optimal Forecasts (2)	Optimal Scenarios: Treatments Run Under Optimal Forecasts (3)	Optimal Scenarios: Treatments Not Run Under Optimal Forecasts (4)
Effect Sizes				
Treatment Effect	0.095	0.095	0.115	0.040
Forecasted Effect	0.190	0.104	0.116	0.031
Sample Size				
Actual Sample Size (N)	2330	2330	2371	1515
Sample Size Under Optimal Forecasts (N)	-	5116	3667	-
Sample Power				
Actual Power	0.440	0.440	0.475	0.250
Power Under Optimal Forecasts	-	0.647	0.632	-
Sample Counts				
Number of Questions (KQs)	418	418	326	92
Number of Projects	43	43	42	19

Notes: Table 6 presents statistics related to various power calculations for 43 experimental projects. For each of the 43 projects, we compute the median value across key questions in a project, and then average across projects. Column 1 presents the results of the data, including the actual power. We computed the forecasted effect given the observed sample size, assuming the authors did a traditional 80 percent power calculation. Column 2 uses as forecasted effect the maximum between the optimal forecast, as in Row 12 in Table 5, and 0.05 (given that powering a study to reject an alternative of 0.01 standard deviations, for example, is prohibitively costly in terms of sample). Column 3 restricts the sample to key questions where the optimal forecast is above 0.05 standard deviations. Column 4 considers the converse sample, that is, key question with the optimal forecast below 0.05 standard deviations. The forecasted sample size under optimal forecasts presents the sample size needed to achieve a power $(1 - \beta)$ of 80%. The forecasted power under optimal forecasts presents the power given the effect size implied by the forecasted treatment effect and the actual treatment effect. All results assume a type 1 error rate (α) of 5%. Project sample sizes are winsorized at the 95th percentile. Three projects elicited forecasts for long-run and short-run outcomes. The results for the long-run outcomes were provided to us, but we weren't able to identify the long-run Ns. In these three cases, we assume a 70% attrition rate. For the remaining projects, we assume no differential attrition between treatment and control arms, taking the ratio of treatment to control samples at baseline, and extrapolating to endline using sample sizes reported in tables. Additionally, for two projects, we are unable to find the sample size for the full set of key questions, instead utilizing the subset of key questions we are able to identify.

Figure 1: SSPP Project-Posting Process, Key Question Example & Word Cloud of Project Titles

(a) Timeline of the SSPP Project-Posting Process



(b) Example of a Key Question

Please predict the percent change in the number of votes received by clean and criminal major party candidates in treated villages, where voters received information about the criminal charges (or lack thereof) of each major party candidate.

For example, if you enter 25, you are predicting that that treated villages have an increase of 25% of votes relative to control villages (not percentage points).

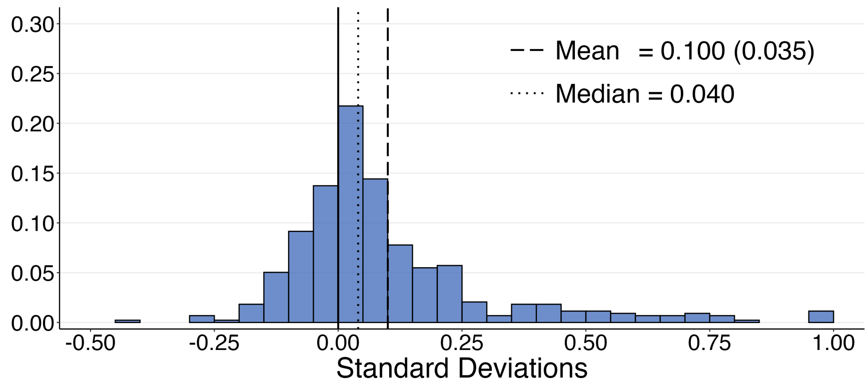
	Number of votes in control villages	Percent change in treated villages	Implied number of votes in treatment villages
Clean candidate votes	268 votes	10 %	295 votes
Criminal candidate votes	252 votes	-5 %	239 votes
Total major party votes	520 votes	2.7%	534 votes

(c) Word Cloud of SSPP Project Titles

Notes: Figure 1a presents a flowchart of the typical timeline for posting a project on the SSPP. Figure 1b presents an example of a key question from a survey. Figure 1c presents a word cloud of the project titles of all 100 projects posted on the SSPP in the 2020-24 period.

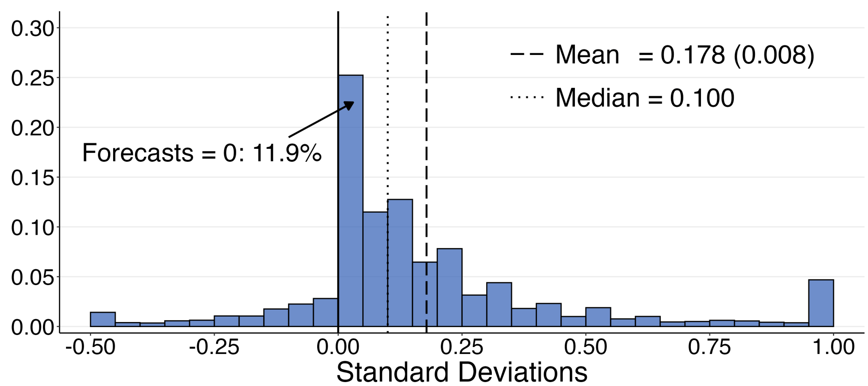
Figure 2: Distributions of Realized and Forecasted Treatment Effects

(a) Realized Treatment Effects



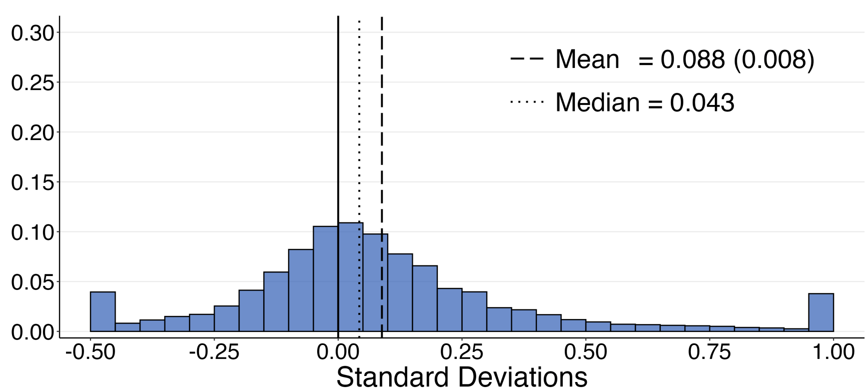
Sample: $N_p = 44$, $N_q = 437$, $N_u = 1{,}420$, $N_r = 2{,}568$, $N_f = 15{,}983$

(b) Forecasts of Treatment Effects



Sample: $N_p = 44$, $N_q = 437$, $N_u = 1{,}420$, $N_r = 2{,}568$, $N_f = 15{,}983$

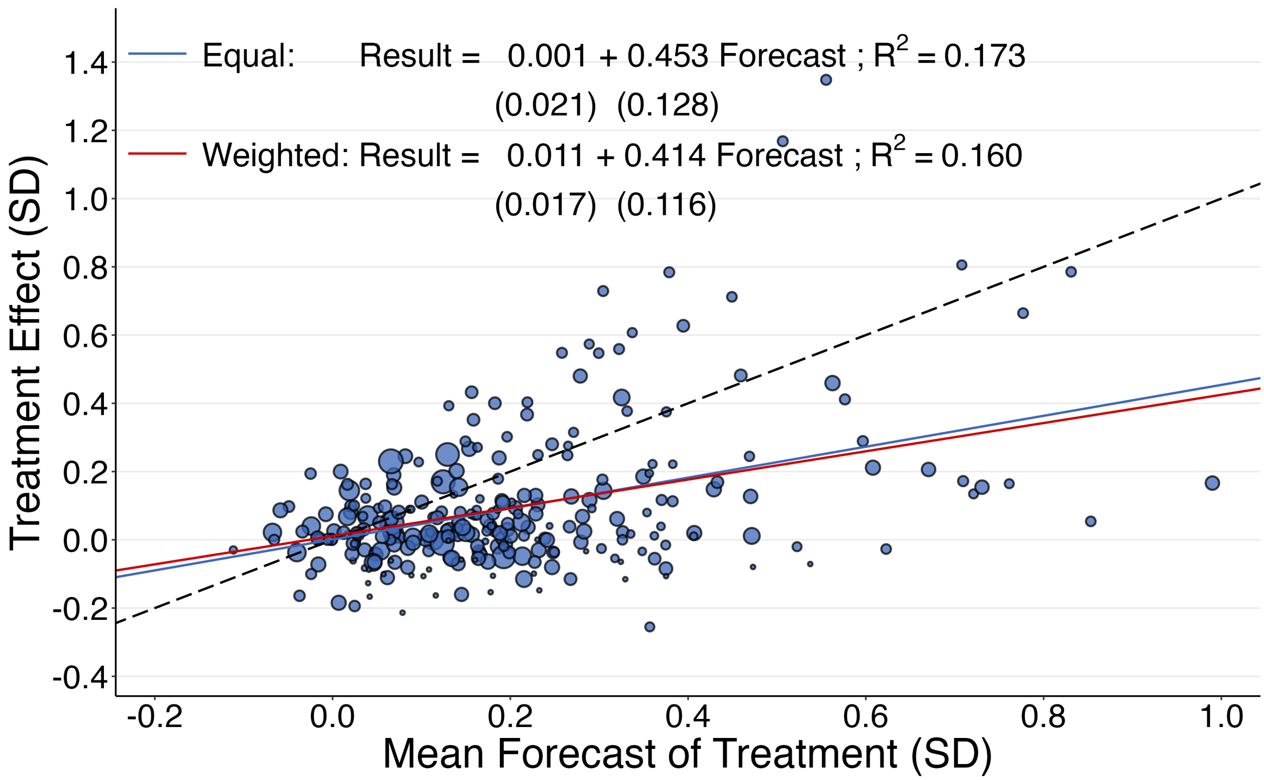
(c) Overestimation of Treatment Effects



Sample: $N_p = 44$, $N_q = 437$, $N_u = 1{,}420$, $N_r = 2{,}568$, $N_f = 15{,}983$

Notes: Figure 2a displays the distribution of realized treatment effects across all normed key questions for which results could be collected. Figure 2b displays the distribution of the normed forecasts of these treatment effects at the individual forecast level. Figure 2c displays the distribution of the difference between each forecast of a key question and the corresponding realized result. Values below -0.5 or above 1.0 standard deviations are displayed at, respectively, -0.5 and 1.0. Across all figures, means are indicated by dashed lines, medians by vertical dotted lines, while the solid lines indicate the 0 standard deviations mark. Standard errors, clustered at the key question level, are provided in parentheses where relevant. A breakdown of the full sample is provided in the footer of each figure.

Figure 3: Predictability of Results from Average Forecasts

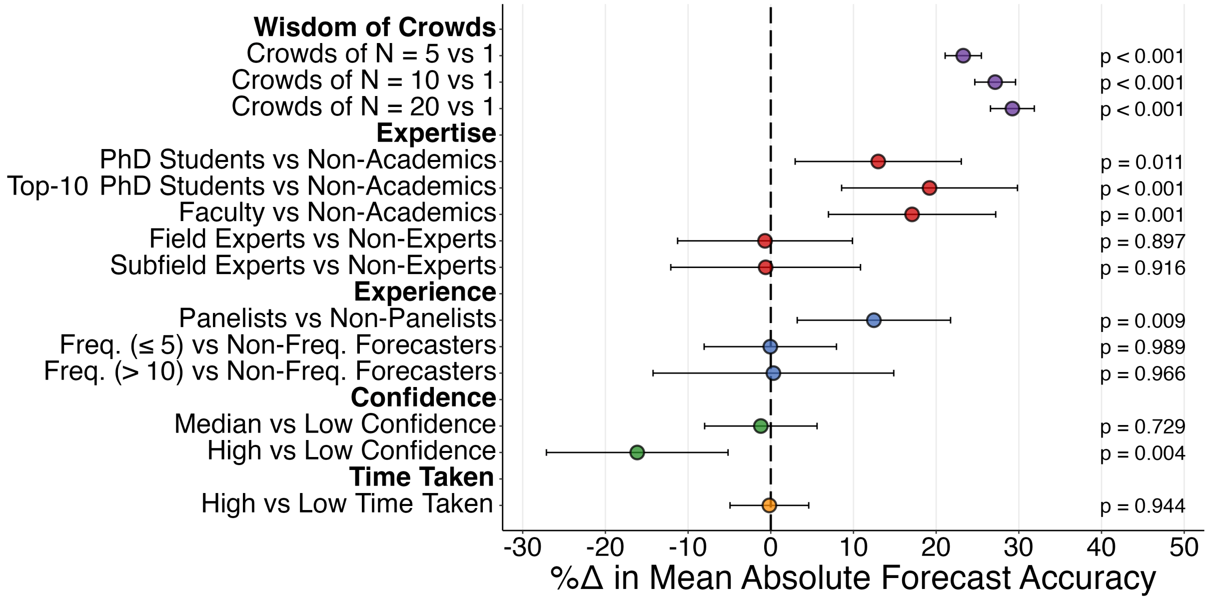


Sample: N_p = 37, N_q = 274, N_u = 1,388, N_r = 2,486, N_f = 14,743

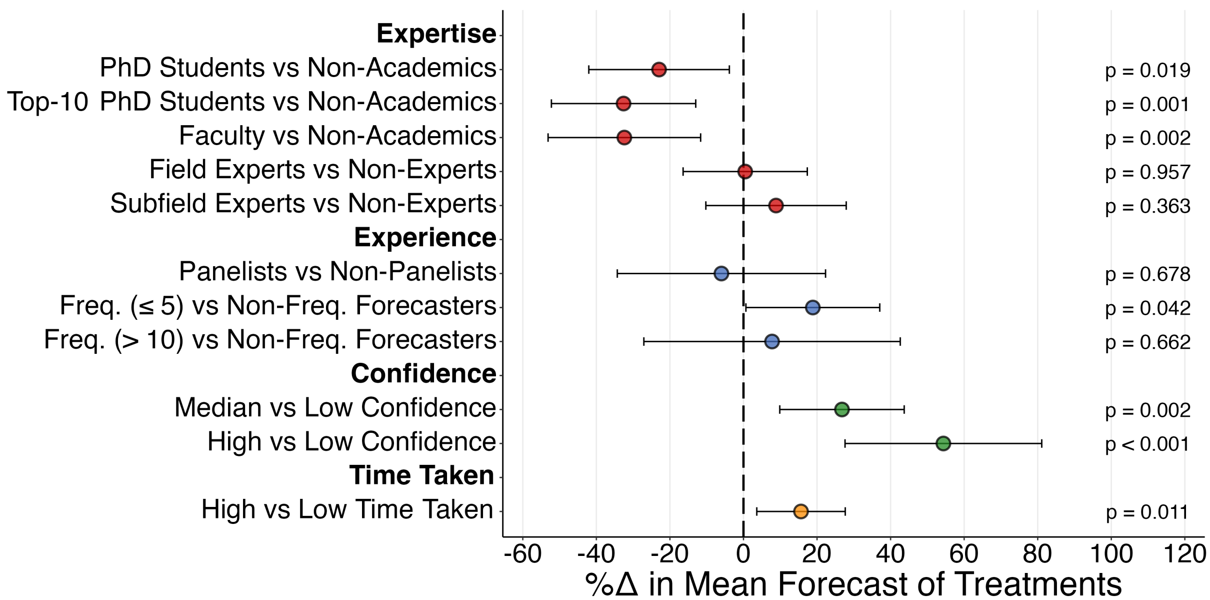
Notes: Figure 3 presents a scatter plot of all normed key questions of treatment effects with a minimum of 20 responses, with the average normed forecast on the x-axis, and the actual normed treatment effect on the y-axis. Points are scaled to represent the number of responses collected by each key question. The solid line displays the unweighted fitted regression provided in the top-left of the figure, with standard errors, clustered at the project level, provided in parentheses. Included in the regression are 4 key questions lying outside the plot axes ranges, with average forecasts and treatment effects of: (-0.01,-0.41), (-0.23,0.23), (1.0,0.26), (0.56,1.35), respectively. Excluded from the figure are 3 outlier key questions, as described in the text; Figure A5 shows the results including the outliers. The dashed line displays the 45-degree line. A breakdown of the full sample is provided in the footer of the figure.

Figure 4: Summary of Cross-Sectional Accuracy and Forecasts of Treatments Comparisons

(a) Absolute Forecast Accuracy

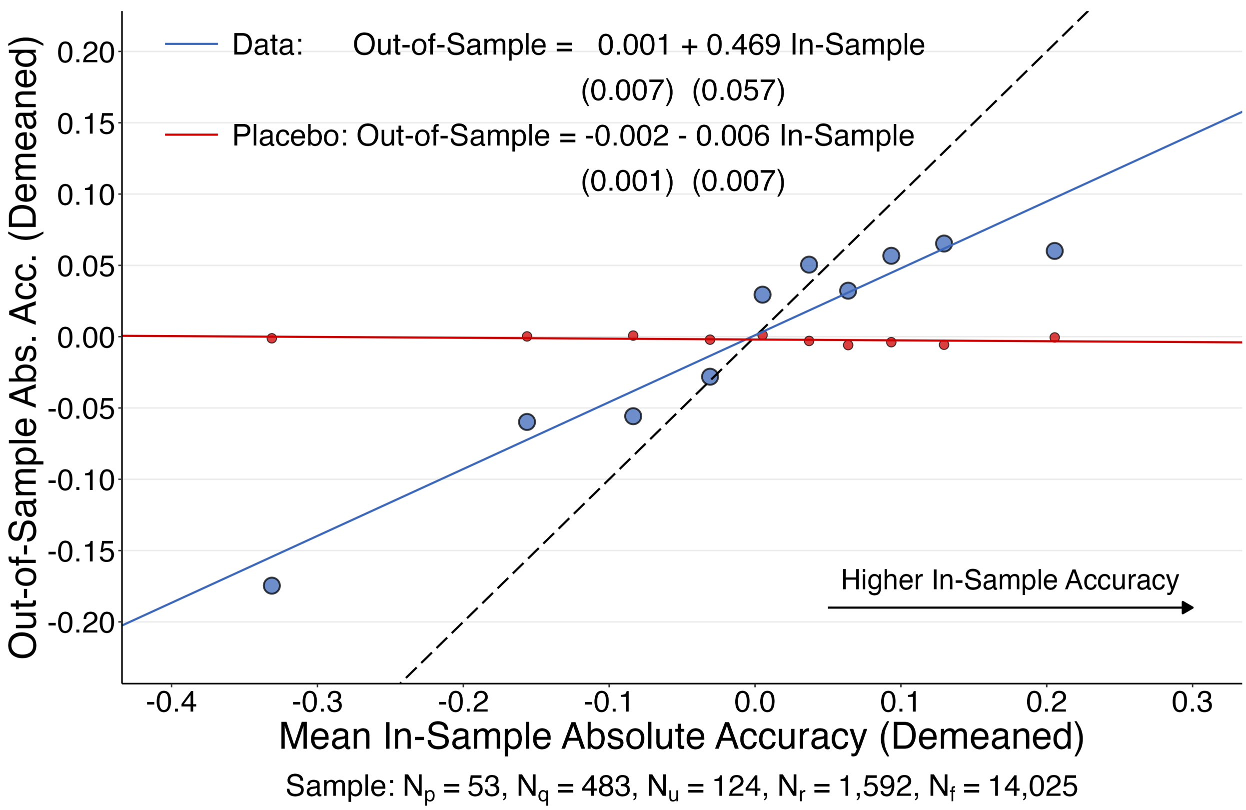


(b) Forecasts of Treatments



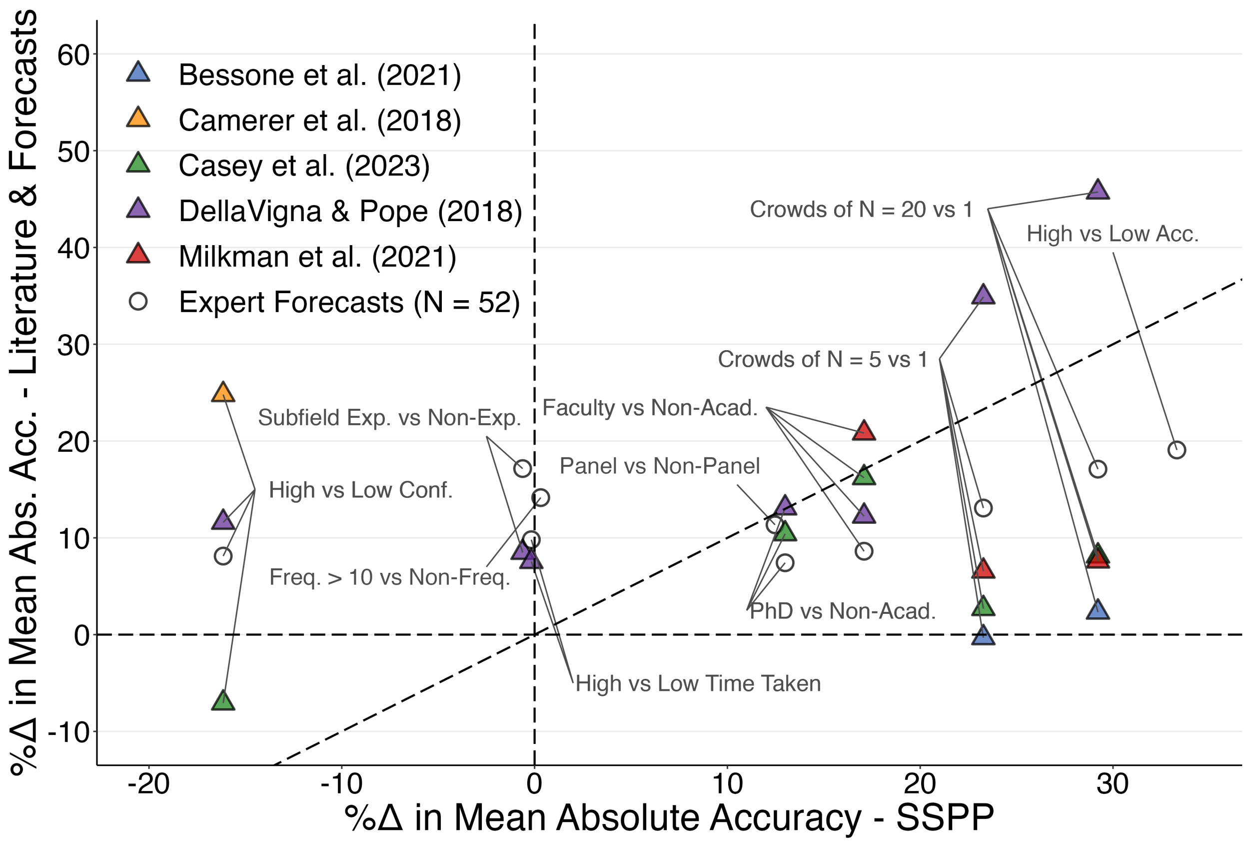
Notes: Figure 4a presents a point plot of the percent changes in average accuracy for each cross-sectional comparison of interest relative to a control group, specifically: wisdom of crowds simulated groups of 5, 10 and 20 relative to individual forecasts, PhD students, PhD students from top-10 departments, and faculty members to non-academics, field and subfield experts to non-experts, panelists to non-panelists, frequent forecasters (first 5 and responses beyond the first 10) to non-frequent forecasters (less than 5 responses), median and high (above-median) confidence to low (below-median) confidence, and high (above-median) time taken to low (below-median) time taken. Estimates are from a set of univariate regression specifications of absolute forecast accuracy on the highlighted dimension, including key question fixed effects, with the point estimates then translated into percent changes relative to the control group. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-values in each row correspond to that of a two-tailed t-test between the group in question and the control group. Figure 4b provides analogous results for the percent changes in the average normed forecast of treatments. A breakdown of the full sample, and a specific count of the number of forecasts from each group, is provided in the footer and labels of the corresponding barplots spanning Appendix Figures A9a and A11b-A17b.

Figure 5: Cross-Project Predictability



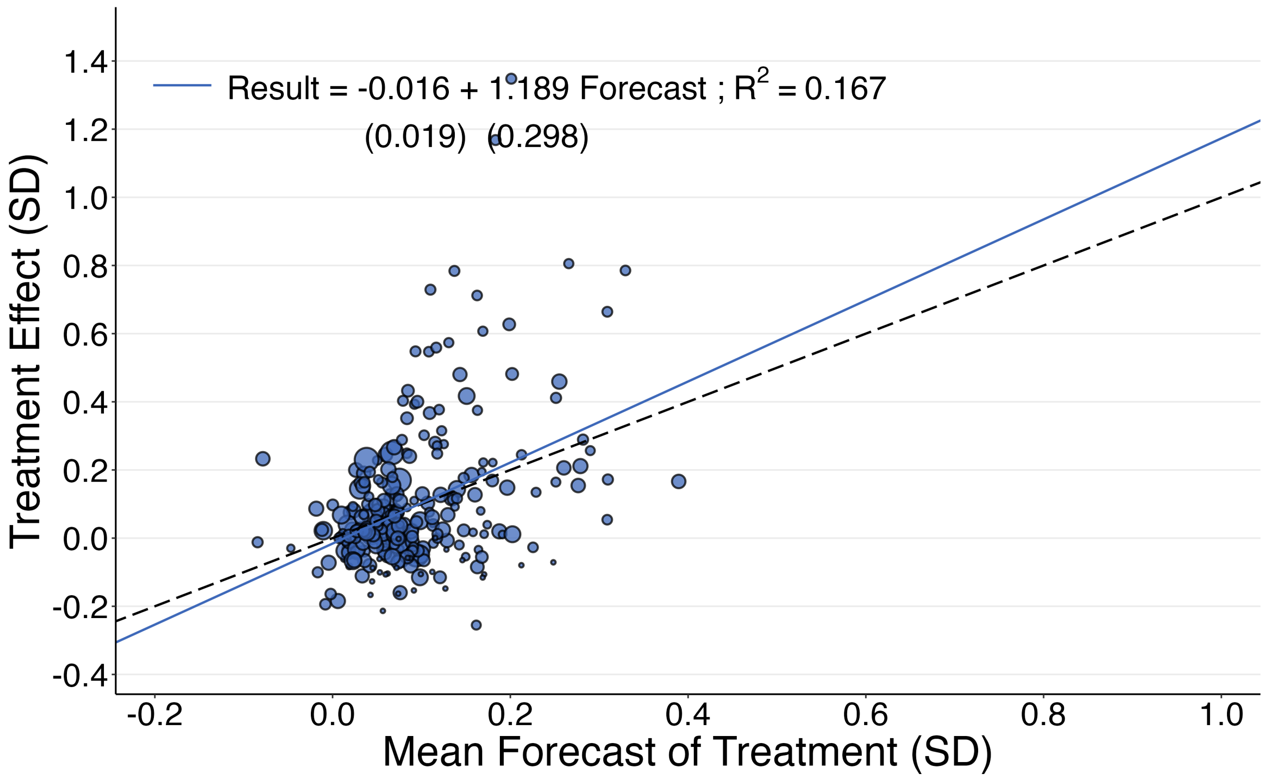
Notes: Figure 5 presents a binscatter plot, across all forecasters with a minimum of 5 responses across projects with results available and that have at least 20 responses. For each forecaster and each project we plot the demeaned absolute accuracy for this out-of-sample project on the y-axis, compared to, on the x-axis, the average demeaned absolute accuracy across 4 other in-sample projects, drawn without replacement. This process is repeated for all projects for which a forecaster has recorded a response. The decile bins of these underlying points are displayed in blue. A placebo version is also implemented, plotting the demeaned absolute accuracy of the out-of-sample project for some other randomly selected forecaster on the y-axis instead, repeating this 100 times. The decile bins of these underlying placebo points are displayed in red. The solid lines display the fitted regressions on the underlying, unbinned, data (non-placebo and placebo) provided in the top-left of the figure, with standard errors, clustered at the forecaster level, in parentheses. The dashed line displays the 45-degree line. A breakdown of the full sample is provided in the footer of the figure.

Figure 6: Comparison of Main Results to Existing Literature and Expert Forecasts



Notes: Figure 6 presents a scatter plot of 10 main results on the differences in absolute forecast accuracy, with, on the x-axis, the observed results in percent change (relative to the omitted group), compared to, on the y-axis, the parallel findings in 5 papers in the literature, as detailed in Appendix E, as well as the average prediction from $N = 52$ expert forecasters. Specifically, the 10 main results are: (i) wisdom-of-crowd accuracy with 5 forecasts relative to one forecast; (ii) wisdom-of-crowd accuracy with 20 forecasts relative to one forecast; (iii) members' forecasts versus non-academics' forecasts; (iv) PhD students' forecasts versus non-academics' forecasts; (v) subfield experts' forecasts versus non-experts' forecasts; (vi) forecasts by panelists versus non-panelists; (vii) forecasts by frequent forecasters versus infrequent forecasters (viii) forecasts of projects in which the forecaster has high versus low confidence; (ix) forecasts by superforecasters versus forecasts by forecasters with low cross-project accuracy; and (x) forecasts in the top half of time taken compared to bottom half of time taken. The dashed lines display the 45-degree line as well as the 0 percent change marks.

Figure 7: Predictability of Results from Shrunk Average Forecasts



Sample: $N_p = 37$, $N_q = 274$, $N_u = 1{,}388$, $N_r = 2{,}486$, $N_f = 14{,}743$

Notes: Figure 7 presents a scatter plot of all normed key questions of treatment effects with a minimum of 20 responses, with the out-of-sample shrunk average normed forecast on the x-axis, and the actual normed treatment effect on the y-axis. The shrunk average normed forecast applies a weighted out-of-sample shrinkage adjustment, transforming the average forecast using coefficients estimated from regressing realized results on forecasts in all other studies while increasing/decreasing the weighting on forecasters with highest/lowest predictive ability (see Appendix D). Points are scaled to represent the number of responses collected by each key question. The solid line displays the unweighted fitted regression provided in the top-left of the figure, with standard errors, clustered at the project level, provided in parentheses. Included in the regression are 4 key questions lying outside the plot axes ranges, with average forecasts and treatment effects of: (-0.01,-0.41), (-0.23,0.23), (1.0,0.26), (0.56,1.35), respectively. Excluded from the figure are 3 outlier key questions, as described in the text; Figure A7 shows the results including the outliers. The dashed line displays the 45-degree line. A breakdown of the full sample is provided in the footer of the figure.

References

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–2794.

Bedoya, G., Y. Belyakova, A. Coville, T. Escande, M. Isaqzadeh, and A. Ndiaye (2024). The Enduring Impacts of a Big Push during Multiple Crises: Experimental Evidence from Afghanistan. *World Bank Policy Research Working Paper Series*.

Ben-David, I., J. R. Graham, and C. R. Harvey (2013). Managerial Miscalibration. *The Quarterly Journal of Economics 128*(4), 1547–1584.

Bernard, D. R. and P. Schoenegger (2024). Forecasting Long-Run Causal Effects. *Available at SSRN 4702393*.

Bessone, P., G. Rao, F. Schilbach, H. Schofield, and M. Toma (2021). The Economic Consequences of Increasing Sleep Among the Urban Poor. *The Quarterly Journal of Economics 136*(3), 1887–1941.

Bloom, N., R. Han, and J. Liang (2024). Hybrid Working from Home Improves Retention Without Damaging Performance. *Nature 630*(8018), 920–925.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science 351*(6280), 1433–1436.

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, et al. (2018). Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015. *Nature Human Behaviour 2*(9), 637–644.

Campbell, S. and D. A. Moore (2024). Overprecision in the survey of professional forecasters. *Collabra: Psychology 10*(1), 92953.

Casey, K., R. Glennerster, E. Miguel, and M. Voors (2023). Long-Run Effects of Aid: Forecasts and Evidence from Sierra Leone. *The Economic Journal 133*(652), 1348–1370.

Chu, J. Y., J. G. Voelkel, M. N. Stagnaro, S. Kang, J. N. Druckman, D. G. Rand, and R. Willer (2024). Academics Are More Specific, and Practitioners More Sensitive, in Forecasting Interventions to Strengthen Democratic Attitudes. *Proceedings of the National Academy of Sciences 121*(3), e2307008121.

DellaVigna, S. and E. Linos (2022). RCTs to Scale: Comprehensive Evidence from Two Nudge Units. *Econometrica 90*(1), 81–116.

DellaVigna, S. and D. Pope (2018a). Predicting Experimental Results: Who Knows What? *Journal of Political Economy 126*(6), 2410–2456.

DellaVigna, S. and D. Pope (2018b). What Motivates Effort? Evidence and Expert Forecasts. *The Review of Economic Studies 85*(2), 1029–1069.

DellaVigna, S., D. Pope, and E. Vivalt (2019). Predict Science to Improve Science. *Science 366*(6464), 428–429.

Diebold, F. X., A. Mora, and M. Shin (2025). On the wisdom of crowds (of economists). *arXiv preprint arXiv:2503.09287*.

Dreber, A., T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson (2015). Using Prediction Markets to Estimate the Reproducibility of Scientific Research. *Proceedings of the National Academy of Sciences 112*(50), 15343–15347.

Enke, B., T. Graeber, and R. Oprea (2023). Confidence, Self-Selection, and Bias in the Aggregate. *American Economic Review 113*(7), 1933–1966.

Ferguson, J., R. Littman, G. Christensen, E. L. Paluck, N. Swanson, Z. Wang, E. Miguel, D. Birke, and J.-H. Pezzuto (2023). Survey of Open Science Practices and Attitudes in the Social Sciences. *Nature Communications 14*(1), 5401.

Galton, F. (1907). Vox Populi. *Nature 75*(1949), 450–451.

Groh, M., N. Krishnan, D. McKenzie, and T. Vishwanath (2016). The Impact of Soft Skills Training on Female Youth Employment: Evidence from a Randomized Experiment in Jordan. *IZA Journal of Labor & Development 5*, 1–23.

Hawkins, S. A. and R. Hastie (1990). Hindsight: Biased Judgments of past Events After the Outcomes Are Known. *Psychological Bulletin 107*(3), 311.

Hewitt, L., A. Ashokkumar, I. Ghezae, and R. Willer (2024). Predicting Results of Social Science Experiments Using Large Language Models. *Preprint*.

Hirshleifer, S., D. McKenzie, R. Almeida, and C. Ridao-Cano (2016). The Impact of Vocational Training for the Unemployed: Experimental Evidence from Turkey. *The Economic Journal 126*(597), 2115–2146.

Holzmeister, F., M. Johannesson, C. F. Camerer, Y. Chen, T.-H. Ho, S. Hoogeveen, J. Huber, N. Imai, T. Imai, L. Jin, et al. (2025). Examining the Replicability of Online Experiments Selected by a Decision Market. *Nature Human Behaviour 9*(2), 316–330.

Iacovone, L., D. McKenzie, and R. Meager (2025). Bayesian impact evaluation with informative priors: An application to a colombian management and export improvement program. *Econometrica 93*(5), 1915–1935.

Ioannidis, J. P., T. D. Stanley, and H. Doucouliagos (2017). The power of bias in economics research. *Economic Journal 127*(605), F236–F265.

Kruger, J. and D. Dunning (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology 77*(6), 1121.

Mellers, B., L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, D. Moore, P. Atanasov, S. A. Swift, et al. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science 25*(5), 1106–1115.

Milkman, K. L., L. Gandhi, M. S. Patel, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, T. W. Lee, J. Rothschild, J. E. Bogard, et al. (2022). A 680,000-Person Megastudy of Nudges to Encourage Vaccination in Pharmacies. *Proceedings of the National Academy of Sciences 119*(6), e2115126119.

Milkman, K. L., D. Gromet, H. Ho, J. S. Kay, T. W. Lee, P. Pandiloski, Y. Park, A. Rai, M. Bazerman, J. Beshears, et al. (2021). Megastudies Improve the Impact of Applied Behavioural Science. *Nature 600*(7889), 478–483.

Moore, D. A. and P. J. Healy (2008). The Trouble with Overconfidence. *Psychological Review 115*(2), 502.

Moore, D. A., S. A. Swift, A. Minster, B. Mellers, L. Ungar, P. Tetlock, H. H. Yang, and E. R. Tenney (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science 63*(11), 3552–3565.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, et al. (2015). Promoting an Open Research Culture. *Science 348*(6242), 1422–1425.

Otis, N. and M. De Vaan (2023). Evaluating Managerial Expectations. *Available at SSRN 4419744*.

Tetlock, P. E. and D. Gardner (2016). *Superforecasting: The Art and Science of Prediction*. Random House.

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association 18*(6), 3045–3089.

Vivalt, E. and A. Coville (2023). How Do Policymakers Update Their Beliefs? *Journal of Development Economics 165*, 103121.

Weber, M., F. d'Acunto, Y. Gorodnichenko, and O. Coibion (2022). The Subjective Inflation Expectations of Households and Firms: Measurement, Determinants, and Implications. *Journal of Economic Perspectives 36*(3), 157–184.

Wolfers, J. and E. Zitzewitz (2004). Prediction Markets. *Journal of Economic Perspectives 18*(2), 107–126.

Forecasting Social Science: Evidence from 100 Projects

Appendix

Stefano DellaVigna Eva Vivalt

A Data Details

A.1 Types of Coded Standard Deviations

For each key question, k, we code a standard deviation, σ_k, such that forecasts, $f_{i,k}$, and results, θ_k, can be expressed in standard deviation units, i.e., $\tilde{f}_{i,k} = \frac{f_{i,k}}{\sigma_k}$ and $\tilde{\theta}_k = \frac{\theta_k}{\sigma_k}$. Key questions can generally be standardized in four broad ways, depending on the subject of the key question – treatment effect or summary statistic – and the units in which forecasts are elicited, e.g., standard deviations, percentage points, dollars, etc. Specifically:

(i) For key questions eliciting predictions in standard deviations, let $\sigma_k = 1$, i.e., no additional standardization is necessary. 35% of key questions are of this type.

(ii) For key questions eliciting predictions for a binary outcome in percent or percentage point units, we code the standard deviation of the outcome variable as the population standard deviation of a Bernoulli random variable. That is, for a key question with a result of θ_k%, the standard deviation is given by $\sigma_k = \sqrt{\theta_k(1-\theta_k)}$. For example, consider the effect of a cash transfer on employment. If the control group had an employment rate of 80%, the key question's forecasts and results would be standardized by $\sqrt{0.8(1-0.8)} = 0.4$. 36.1% of key questions are of this type.

(iii) For key questions eliciting predictions of a treatment effect in percent or percentage points where the forecast is the percentage difference relative to a specific value (e.g., control or baseline), we follow a two-step approach in standardizing. First, we convert the percent or percentage point to the underlying unit of the outcome variable by multiplying by the relative value, e.g., the average of the control group. Second, we divide by the standard deviation of the outcome variable. For example, consider the percentage point difference in revenue for a treatment group following an entrepreneurship training program relative to a control group. First, we multiply by the average revenue of the control group at baseline to obtain the difference in revenue in dollar terms. We then divide by the standard deviation of revenue of the control group at baseline. 5.6% of key questions are of this type.

(iv) For key questions eliciting predictions in units other than standard deviations, percentage points or percents, e.g., dollars or Likert scale questions, we standardize using the standard deviation of the outcome variable. That is, for a key question about an outcome Y, the standard deviation is given by $\sigma_k = \sigma_Y$. For key questions associated with an experiment, specifically treatment effects, we prioritize taking baseline values of the control group, and for key questions of summary statistics, we prioritize taking pooled values (i.e., values of the entire sample). 18.2% of key questions are of this type.

A few key questions do not perfectly fall under one of these four categories. Standardizing is done on a case-by-case basis for these questions, which comprise 5% of key questions.

A.2 Norming of Treatment Effects

As discussed in the text, we define, when possible the direction in which the model predicts the treatment effect to go and reverse the direction of the effect in a fraction of cases. For example, in a project on social norms to encourage water saving the treatment effect on water consumption is reversed, as the presumption is that social norms would decrease water consumption. In some cases, we do not norm the treatment effect, as there are opposing models. These key questions are not used for the analysis of normed treatments. An example is the forecast of the impact of Cash Transfers on labor supply. A traditional model implies that that the labor supply would be reduced. An alternative model holds that there would be no such effect, or even positive effects.

A.3 Treatment Effect Outliers

In a small number of cases, measuring the effect size of a treatment in standard deviation units leads to magnitudes that may be seen as mis-leading in their size. Consider for example the case of a treatment that encourages a behavior that is very rare in the control group, implying that the standard deviation in the control group is very small. Even a modestly-sized effect of the treatment would loom very large when measured in units of standard deviations of the control group. More generally, whenever the take-up of some outcome is very small in the control group, effect sizes can loom large as measured in standard deviation. For three key questions in one project, (Bedoya et al., 2024), we encounter precisely this issue. In two key questions, the treatment effect is measured in terms of number of cows, which is a very sparse outcome in control, with an average of 0.1 cows (SD of 0.376). In the remaining key question, the treatment targets savings, which is rare among the control group, with 4.9% of the control sample having any savings (average savings of $3.80 USD, SD of $39.80). Given this, it is not surprising that the treatment effects and the forecasts are outliers: they have treatment effects (respectively, the average forecast) of 0.99 (2.05), 1.83 (2.57), and 2.09 (0.71) standard deviations. We exclude these 3 key questions from the main results on forecasts of normed treatment effects, but show the results including them in Appendix Figures A2b and A5.

A.4 Academic Status

We code the academic status of forecasters using a combination of information from the profile on the SSPP platform, as well as information that users may have indicated in a particular survey. For survey with information coming from both sources, we take the highest degree of academic status across them. To code whether a student is in a top-10 department, we use a combination of institution (if indicated) and email address. We coded as top-10 institutions (in alphabetical order) Berkeley, Chicago, Columbia, Harvard, Michigan, MIT, Northwestern, Princeton, Stanford, and Yale. We include postdoctoral researchers in the Faculty category.

A.5 Confidence Question

In 50 of the 100 studies, the wording of the confidence question is taken from the suggested format on the platform: "*How confident are you in your predictions for this study? If you are confident, it means that you believe your predictions are very accurate.*" The respondents use a 5-point Likert scale to indicate their confidence in their predictions. Appendix Figure A19a shows the distribution of the confidence question for these cases. In 16 other studies, the researchers use a differently worded confidence question, but we can map it into our framework by determining which responses indicate higher, versus lower, confidence relative to the median confidence level.

B Decomposition of Effect of Overconfidence on Accuracy

In this section we estimate a simple model of forecasts of treatment effects, aiming to decompose the observed differences in accuracy for high-confidence versus low-confidence forecasts. The model builds on DellaVigna and Pope (2018a). We model agent i making forecasts about the results in treatments $k = 1, ..., K$. Let $\theta = (\theta_1, ..., \theta_K)$ be the outcome (unknown to the agent) in the K treatments. We assume that the agent aims to minimize the squared distance between their forecast $f_{i,k}$ and the result

θ_k. We assume that agents start with a noninformative prior and that agent i, with $i = 1, ..., I$, draws a signal $s_{i,k}$ about the outcome of treatment k:

$$s_{i,k} = \theta_k + \eta_k + o + \sigma\epsilon_{i,k} \tag{1}$$

The deviation of the signal $s_{i,k}$ from the truth θ_k consists of three components, each of which is independent of the others:

(i) $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$ is a deviation for treatment k that is common to all forecasters, with mean zero. To simplify the estimation problem, we treat η_k as fixed effects instead of estimating the distribution as a random effect.

(ii) o is a deviation that is common across all treatments, denoting the (possible) overprediction (for positive o) of treatment effects.

(iii) $\sigma\epsilon_{i,k}$, with $\epsilon_{i,k} \sim \mathcal{N}(0,1)$ independent of scalar σ, is an idiosyncratic noise term. A larger σ denotes a noisier signal of the treatment effect.

Summing all components together, we find:

$$s_{i,k} \sim \mathcal{N}(\theta_k + o, \sigma_\eta^2 + \sigma^2) \tag{2}$$

We assume that each agent i is unaware that their prediction for project k has a systemic bias o. Given a noninformative prior, the signal $s_{i,k}$ is an agent's best estimate (i.e., $s_{i,k} = f_{i,k}$), since it minimizes the (subjective) expected loss $(f_{i,k} - \theta_k)^2$.

To estimate the treatment-level effects η_k, notice that the expected forecast error in treatment k equals $\mathbb{E}[f_{i,k} - \theta_k] = \eta_k + o$. Thus, to estimate $\hat{\eta}_k$, we first compute the average forecast error for treatment k, $\bar{e}_k = \Sigma_i\big((f_{i,k} - \theta_k)/I\big)$, and then we demean it. Thus, $\hat{\eta}_k = \bar{e}_k - \Sigma_k\big(\bar{e}_k/K\big)$. We define the residual $z_{i,k} = f_{i,k} - \theta_k - \hat{\eta}_k$ and rewrite the model as:

$$z_{i,k} = o + \sigma\epsilon_{i,k} \tag{3}$$

We estimate this transformed model with maximum likelihood. We allow for heterogeneity in the two key parameters, o and σ, with respect to observable differences in confidence. We assume four types of forecasts: forecasts with low (i.e., below median) confidence, median confidence, high (i.e., above median) confidence, and missing confidence. These types derive their residuals from separate distributions, i.e., low-confidence forecasts have the parameters $(o^{(l)}, \sigma^{(l)})$, median-confidence forecasts have the parameters $(o^{(Me)}, \sigma^{(Me)})$, high-confidence forecasts have $(o^{(h)}, \sigma^{(h)})$, and missing-confidence forecasts have the parameters $(o^{(Mi)}, \sigma^{(Mi)})$.

The likelihood takes a convenient form. Let $\Theta \equiv [o^{(l)}, o^{(Me)}, o^{(h)}, o^{(Mi)}, \sigma^{(l)}, \sigma^{(Me)}, \sigma^{(h)}, \sigma^{(Mi)}]^T$ denote the vector of parameters to estimate. Using $\mathbb{1}(l_{i,k})$, $\mathbb{1}(m_{i,k})$, $\mathbb{1}(h_{i,k})$ as indicators for agent i having low, median and high confidence in their forecasts on project k respectively, and denoting the standard normal density as ϕ, the likelihood is

$$\begin{aligned} \text{Lik}[z|\Theta] = \prod_{i=1}^{I}\prod_{k=1}^{K}\Big\{\mathbb{1}(l_{i,k})\cdot\Big[\frac{1}{\sigma^{(l)}}\phi\Big(\frac{z_{i,k}-o^{(l)}}{\sigma^{(l)}}\Big)\Big] + \mathbb{1}(m_{i,k})\cdot\Big[\frac{1}{\sigma^{(Me)}}\phi\Big(\frac{z_{i,k}-o^{(Me)}}{\sigma^{(Me)}}\Big)\Big] + \\ \mathbb{1}(h_{i,k})\cdot\Big[\frac{1}{\sigma^{(h)}}\phi\Big(\frac{z_{i,k}-o^{(h)}}{\sigma^{(h)}}\Big)\Big] + \big(1-\mathbb{1}(l_{i,k})-\mathbb{1}(m_{i,k})-\mathbb{1}(h_{i,k})\big)\cdot\Big[\frac{1}{\sigma^{(Mi)}}\phi\Big(\frac{z_{i,k}-o^{(Mi)}}{\sigma^{(Mi)}}\Big)\Big]\Big\} \end{aligned} \tag{4}$$

The asymptotic covariance is given by the inverse of the Fisher information, which we estimate with its sample analogue.

Table A2 reports our results. Relative to low confidence forecasters, high confidence forecasters both overpredict more on average ($o^{(h)} = 0.137$ versus $o^{(l)} = 0.076$) and have higher variance in their predictions ($\sigma^{(h)} = 0.473$ versus $\sigma^{(l)} = 0.416$). The differences in both parameters across the two groups are statistically significant.

Thus, the lower accuracy associated with high-confidence forecasts is due to both a higher degree of overestimation of treatment effects, and noisier signals. How much does each factor contribute to the difference in accuracy of forecasts? Table A3 reports the mean absolute forecast accuracy for the various confidence levels. In Column 1 we report the empirical estimates (for predictions of treatment effects for which we have results on accuracy) while in Column 2 we report the average from simulations drawn for the estimated parameters. As the table shows, the simulated data somewhat overstate the absolute error across the groups, but it reproduces well the difference in accuracy of about 0.05 standard deviations between the high-confidence and the low-confidence group. In Column 3, we present a simulation in which we aim to decompose how much of the effect is due to variation in the prediction of the signal. Namely, we hold the value of o to o^l, while allowing σ to vary across groups. This estimate is able to reproduce around 70% of the total predicted difference in accuracy between the high-confidence and the low-confidence groups. Column 4 shows a complementary simulation which instead holds σ while varying o; this simulation is able to reproduce 30% of the total predicted difference. Thus, we attribute 70 percent of the difference in accuracy between low- and high- confidence forecasters to a less precise signal, with the rest due to overestimation.

C Cross-Project Predictability Procedure

Consider the set of all key questions, $\mathcal{K}$, for which results could be collected and for which at least 20 forecasts have been recorded, and the corresponding set of projects of these questions, $\mathcal{P}$. Consider further the set of all forecasters, $\mathcal{I}$, with a minimum of 5 responses to projects in $\mathcal{P}$. We are interested in cross-project accuracy, i.e., whether, for a specific forecaster, accuracy in a set of projects predicts accuracy in a leave-out project. Some projects may be more difficult to forecast than others, so we will need to demean the forecasts to examine this question. The following steps detail the underlying procedure:

(i) Consider the full set of projects for which forecaster i has recorded a response, i.e., $p \in \mathcal{P}_i$, and select one project as the out-of-sample project, $\{p_1\} = \mathcal{P}_i^{\text{Out}}$. Similarly, randomly select (without replacement) four other projects as the in-sample set of projects, $\{p_2, p_3, p_4, p_5\} = \mathcal{P}_i^{\text{In}}$

(ii) Calculate forecaster i's absolute forecast accuracy (negative absolute forecast error) $a_{ikp} = -\left|f_{ikp} - \theta_k\right|$, where f_{ikp} is the standardized forecast and θ_k is the corresponding standardized result. Demean this absolute forecast accuracy for all key questions k in the selected projects $p \in \mathcal{P}_i^{\text{Out}} \cup \mathcal{P}_i^{\text{In}}$:

$$\hat{a}_{ikp} = a_{ikp} - \frac{1}{\left|\mathcal{I}_p\right|} \sum_{i \in \mathcal{I}_p} a_{ikp} \tag{1}$$

(iii) Given that a single project, $p \in \mathcal{P}_i$, could contain multiple key questions, $k \in \mathcal{K}^p$, we must first collapse the response of forecaster i for project p to a single point. For forecaster i, average over all the demeaned absolute accuracies associated with the key questions in project p, $\mathcal{K}_p$. Specifically, define the average demeaned absolute forecast accuracy for forecaster i in project p as:

$$\bar{a}_{ip} = \frac{1}{\left|\mathcal{K}_p\right|} \sum_{k \in \mathcal{K}_p} \hat{a}_{ikp} \tag{2}$$

$\bar{a}_{ip} > 0$ indicates that forecaster i outperforms other forecasters on average, while $\bar{a}_{i,p} < 0$ indicates that the forecaster is less accurate than others in project p.

(iv) The set of points that make up the underlying data in the binscatter plot displayed in Figure 5 are then defined by:

$$x_{ip_1} = \frac{1}{\left|\mathcal{P}_i^{\text{In}}\right|} \sum_{p \in \mathcal{P}_i^{\text{In}}} \bar{a}_{ip} \tag{3}$$

$$y_{ip_1} = \bar{a}_{ip_1} \tag{4}$$

where (x_{ip_1} represents the average demeaned absolute forecast error of the in-sample set of projects (on the x-axis of the figure) and y_{ip_1}) represents the demeaned absolute forecast error for the out-of-sample project (on the y-axis of the figure).

(v) Repeat steps (i) - (iv) for all projects $p \in \mathcal{P}_i$, and for all forecasters $i \in \mathcal{I}$. Each forecaster i should have exactly $|\mathcal{P}_i|$ underlying points.

A placebo version of the process described above can also be implemented, where instead of taking $y_{ip_1} = \bar{a}_{ip_1}$, we take the demeaned absolute forecast error for the out-of-sample project p_1 from any other forecaster $i' \in \mathcal{I}^{p_1}$ where $i' \neq i$, i.e., $y_{ip_1} = \bar{a}_{i'p_1}$, keeping x_{ip_1} unchanged. To be a good placebo, we want:

$$\text{Cov}\left(\frac{1}{\left|\mathcal{P}_i^{\text{In}}\right|} \sum_{p \in \mathcal{P}_i^{\text{In}}} \bar{a}_{ip}, \bar{a}_{i'p_1}\right) = 0 \tag{5}$$

Expanding this covariance leads to a combination of cross-terms across projects and forecasters. By construction, given that $p_1 \notin \mathcal{P}_i^{\text{In}}$, there is no project-specific correlation between these terms, and any mechanical correlation can be reduced to forecaster-specific cross-terms between $\bar{a}_{ip} = \frac{1}{|\mathcal{K}_p|}\sum_{k \in \mathcal{K}_{p_1}} \hat{a}_{ikp}$ and $\bar{a}_{i'p_1} = \frac{1}{|\mathcal{K}_{p_1}|}\sum_{k \in \mathcal{K}_{p_1}} \hat{a}_{i'kp_1}$.

Notice that the forecast by forecaster i for project p_1 can be found in the demeaning of $a_{i'kp_1}$, and similarly other forecasts by forecaster i' could be possibly found across $\mathcal{P}_i^{\text{In}}$. Further, other forecasters $j \notin \{i, i'\}$ could have recorded forecasts for both the out-of-sample project and one of the in-sample projects, leading to non-zero cross-terms in the covariance due to some underlying forecaster-specific idiosyncrasy common across all forecasts by forecaster j. As such, we require that the set of forecasts with which we demean the out-of-sample project, and the set of forecasts with which we demean the in-sample set of projects, belong to mutually exclusive sets of forecasters.

Formally, consider the set of all forecasters across the set of in-sample projects, $\mathcal{I}^{\text{In}} = \mathcal{I}_{p_2} \cup \mathcal{I}_{p_3} \cup \mathcal{I}_{p_4} \cup \mathcal{I}_{p_5}$, and similarly the set of all forecasters across the out-of-sample project, $\mathcal{I}^{\text{Out}} = \mathcal{I}_{p_1}$. If we demean the out-of-sample project by a set of forecasts that belong to forecasters $i \in A \subseteq \mathcal{I}^{\text{Out}}$, we must then demean the in-sample projects by a set of forecasts belonging to forecasters $i \in B \subseteq \mathcal{I}^{\text{In}}$

such that $A \cap B = \emptyset$. In practice, this is implemented by finding the intersection of $\mathcal{I}^{\text{In}}$ and $\mathcal{I}^{\text{Out}}$ at each iteration of the procedure, and randomly allocating half of the forecasters to either the out-of-sample, A, or in-sample, B, demeaning set, and omitting their forecasts in the parallel demeaning. The average (median) demeaning set consists of 54.87 (44) forecasts and no demeaning set is smaller than 9 forecasts.

D Forecasting Aggregation Methods and Out-of-Sample Accuracy Analysis

To evaluate the out-of-sample prediction accuracy of various forecasting methods seen in Tables 5 and A4, we aggregate the forecast outcomes at the key question level. Consider the set of all key questions $\mathcal{K}$, partitioned into project-specific subsets $\mathcal{K}_p$ for each project $p \in \mathcal{P}$. For a given key question $k \in \mathcal{K}_p$, let f_{ik} denote the standardized forecast elicited from forecaster $i \in \mathcal{I}_k$, and let θ_k be the realized, standardized true outcome.

Our primary evaluation metric is the Mean Squared Error (MSE), computed over all key questions. For Table A4, the weighted mean of squared error is used - weighting by number of forecasts on a question. For any target project p, estimated parameters are trained on the leave-one-project-out sample, denoted $\mathcal{P}_{-p} = \mathcal{P} \setminus \{p\}$. The baseline specifications—reported sequentially by their respective rows in the main accuracy table—are constructed as follows:

D.1 Naive Aggregation Methods

Row 1: Individual Forecast. This represents the most granular baseline, wherein the predicted value is simply the individual forecast itself, $\hat{\theta}_{ik} = f_{ik}$. The overall performance is evaluated by computing the average squared error across all individual forecasts.

Row 2: Average Forecast (Wisdom of the Crowd). This method applies a simple arithmetic mean across all forecasters for a given key question, serving as a conventional "wisdom of the crowd" baseline. The aggregated prediction is given by: $\hat{\theta}_k = \bar{f}_k = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} f_{ik}$

Row 3: Demeaned Average Forecast (Out-of-Sample). For a target project p, we first calculate the global mean forecast error exclusively across all other projects $p' \in \mathcal{P}_{-p}$. Let this out-of-sample bias be $\bar{B}_{-p}$:

$$\bar{B}_{-p} = \frac{\sum_{p' \neq p} \sum_{k' \in \mathcal{K}_{p'}} \sum_{i \in \mathcal{I}_{k'}} (f_{ik'} - \theta_{k'})}{\sum_{p' \neq p} \sum_{k' \in \mathcal{K}_{p'}} |\mathcal{I}_{k'}|} \tag{6}$$

The point forecast for a key question $k \in \mathcal{K}_p$ is then computed by taking the simple average of the elicited forecasts and subtracting this estimated out-of-sample bias:

$$\hat{\theta}_k = \bar{f}_k - \bar{B}_{-p} \tag{7}$$

Row 4: Constant Forecast (Average of Other Projects). For any target project p, the predicted outcome for its constituent key questions is the arithmetic mean of all realized true values in the out-of-sample set of projects:

$$\hat{\theta}_k = \frac{1}{|\mathcal{K}_{-p}|} \sum_{k' \in \mathcal{K}_{-p}} \theta_{k'} \tag{8}$$

where $|\mathcal{K}_{-p}|$ is the total number of key questions outside of project p.

Row 5: Shrunk Average Forecast. This method applies a linear shrinkage to the simple average forecast to correct for potential attenuation bias or overconfidence. For a target project p, we regress the realized truths on the mean forecasts strictly across the training set of out-of-sample projects $\mathcal{P}_{-p}$:

$$\theta_{k'} = \gamma_0 + \gamma_1 \bar{f}_{k'} + \nu_{k'} \quad \text{for } k' \in \mathcal{K}_{-p} \tag{9}$$

Let $\hat{\gamma}_0$ and $\hat{\gamma}_1$ be the out-of-sample intercept and slope estimates, respectively. The predicted outcome for a key question $k \in \mathcal{K}_p$ is then constructed by scaling the standard "wisdom of the crowd" average using these parameters:

$$\hat{\theta}_k = \hat{\gamma}_0 + \hat{\gamma}_1 \bar{f}_k \tag{10}$$

D.2 Generation of Out-of-Sample Predicted Errors

Following the methodology of within-forecaster determinants found in Section 4.4, we generate an out-of-sample predicted error ranking. For each project p, we train a fixed-effects linear regression model exclusively on the sample of projects $\mathcal{P}_{-p}$. The dependent variable is the squared forecast error, defined as $e_{ik}^2 = (f_{ik} - \theta_k)^2$.

We project e_{ik}^2 onto a vector of forecaster and response-level covariates, denoted X_{ik}. The characteristic space X_{ik} is made of the determinants used in Section 4.4.[2]

[2] PhD Students, PhD Students at Top-10 Institutions, Faculty, Field Experts, Subfield Experts, Time Taken High, Confidence Median, Confidence High, Frequent Forecasters ($\leq$ 5 responses), Frequent Forecasters ($>$ 10 responses), and Panelists.

The structural equation takes the form:

$$e_{ik}^2 = X_{ik}\beta + \alpha_k + \epsilon_{ik} \tag{11}$$

where α_k accounts for the key question fixed effects. We estimate $\hat{\beta}_{-p}$ via ordinary least squares on $\mathcal{P}_{-p}$. Note the α_k fixed effects for out-of-sample can not be estimated, but can be assumed to be constant since we are forming a ranking *within* a single question. The expected squared error for forecast i on an out-of-sample question $k \in \mathcal{K}_p$ is thus generated by:

$$\hat{e}_{ik}^2 = X_{ik}\hat{\beta}_{-p} \tag{12}$$

Forecasters are subsequently ranked within each key question according to $\hat{e}_{ik}^2$, such that a smaller predicted error yields a higher rank. Let $r_{ik} \in (0, 1]$ denote the resulting percentile rank of forecaster i for question k.

D.3 Weighting Top Forecasts

Row 8: Average of Top 50% Forecasters (Out-of-sample ranked). This method calculates a simple arithmetic mean exclusively across the subset of forecasters predicted to be in the top 50% for a given key question. Utilizing the out-of-sample predicted ranks r_{ik} generated in Section D.2, the aggregated prediction is given by:

$$\hat{\theta}_k = \frac{1}{|\{i \in \mathcal{I}_k \mid r_{ik} \leq 0.50\}|} \sum_{i \in \mathcal{I}_k,\, r_{ik} \leq 0.50} f_{ik} \tag{13}$$

Row 9: Optimally Weighted Forecast for Top 50%. Consider a weighted average giving additional weight to predicted top 50%. Let τ denote a predefined fractional cutoff (e.g., $\tau = 0.50$ for the top 50% of forecasters). We use an up-weighting parameter $\alpha \geq 1$ assigned to forecasters whose predicted out-of-sample rank falls within the specified threshold. The weight applied to forecaster i is defined as:

$$w_{ik}(\alpha) = \begin{cases} \alpha & \text{if } r_{ik} \leq \tau \\ 1 & \text{otherwise} \end{cases} \tag{14}$$

This generates an α-weighted consensus forecast for question k:

$$\bar{f}_k(\alpha) = \frac{\sum_{i \in \mathcal{I}_k} w_{ik}(\alpha) f_{ik}}{\sum_{i \in \mathcal{I}_k} w_{ik}(\alpha)} \tag{15}$$

We perform a grid search over a discrete set of candidate parameters $\mathcal{A}$ to minimize the MSE strictly over the training set $\mathcal{P}_{-p}$. The minimization procedure is defined as:

$$\alpha^*_{-p} = \arg\min_{\alpha \in \mathcal{A}} \sum_{p' \neq p} \sum_{k' \in \mathcal{K}_{p'}} \left(\bar{f}_{k'}(\alpha) - \theta_{k'}\right)^2 \tag{16}$$

The final target prediction applies this optimal out-of-sample weight parameter: $\hat{\theta}_k = \bar{f}_k(\alpha^*_{-p})$.

Rows 11 & 12: Shrunk Optimally Weighted Forecasts This method combines the optimal weighting framework with linear shrinkage. For any candidate α, we compute the weighted mean forecast $\bar{f}_{k'}(\alpha)$ on the out-of-sample training set $\mathcal{P}_{-p}$. We then fit a univariate regression mapping these weighted means to the realized truths:

$$\theta_{k'} = \gamma_0 + \gamma_1 \bar{f}_{k'}(\alpha) + \nu_{k'} \quad \text{for } k' \in \mathcal{K}_{-p} \tag{17}$$

Let $\hat{\gamma}_0(\alpha)$ and $\hat{\gamma}_1(\alpha)$ be the estimated intercept and slope (the linear shrinkage parameter), respectively. The optimal weight α^*_{-p} is derived by minimizing the out-of-sample Mean Squared Error of the resulting shrunk predictions across the parameter grid:

$$\alpha^*_{-p} = \arg\min_{\alpha \in \mathcal{A}} \sum_{p' \neq p} \sum_{k' \in \mathcal{K}_{p'}} \left(\hat{\gamma}_0(\alpha) + \hat{\gamma}_1(\alpha)\bar{f}_{k'}(\alpha) - \theta_{k'}\right)^2 \tag{18}$$

For a question k in the target test project p, we compute the optimally weighted forecast $\bar{f}_k(\alpha^*_{-p})$ and scale it using the out-of-sample optimal shrinkage estimates to produce the final predicted point outcome:

$$\hat{\theta}_k = \hat{\gamma}_0(\alpha^*_{-p}) + \hat{\gamma}_1(\alpha^*_{-p})\bar{f}_k(\alpha^*_{-p}) \tag{19}$$

E Comparison to Previous Literature

Figure 6 presents a comparison between the main results from the SSPP and 18 analogous findings drawn from five prior studies. This section outlines those studies and explains how their data compares with the SSPP.

All of the earlier studies elicit forecasts for a single experiment only; none ask forecasters to predict outcomes across multiple studies. As a result, we are unable to directly compare our findings on cross-project accuracy, forecaster experience, or panelists with prior research.

Data on the relative accuracy of faculty/PhD students versus non-academics are available in Casey et al. (2023), DellaVigna and Pope (2018a), and Milkman et al. (2021). On the platform, non-academic forecasters include policy professionals, think tank researchers, and other relevant stakeholders. This is most similar to Casey et al. (2023), where non-academic forecasters include policy makers in Sierra Leone and the OECD, as well as undergraduate students in Sierra Leone. Non-academic forecasters are composed of undergraduate and MBA students in DellaVigna and Pope (2018a), and they are composed of laypeople and practitioners in Milkman et al. (2021).

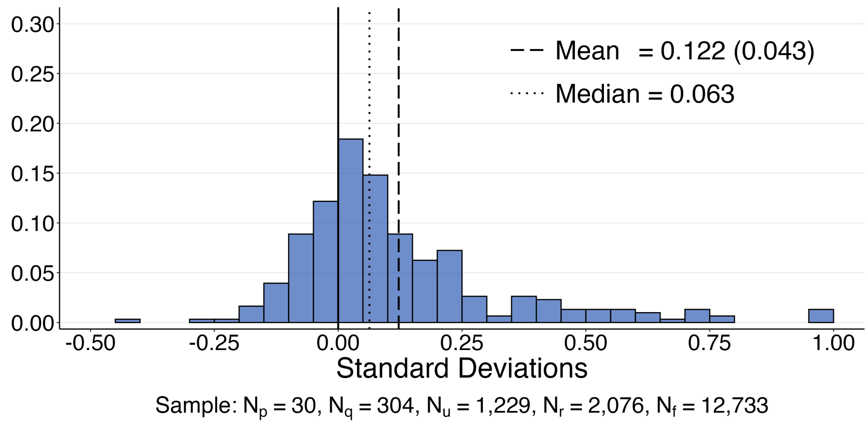
Self-reported forecaster confidence data are available in Casey et al. (2023), DellaVigna and Pope (2018a), and Camerer et al. (2016). For each of these studies, we compare forecasters in the top and bottom terciles of confidence within a project, mirroring the approach in our main analysis.

We are able to construct wisdom-of-crowds predictions (with replacement) for group sizes of $N = 5$ and $N = 20$ using data from Bessone et al. (2021), Casey et al. (2023), DellaVigna and Pope (2018a), and Milkman et al. (2021). We cannot perform a similar analysis for Camerer et al. (2016) because the outcomes are binary (i.e., a paper either replicates or does not). As a result, forecast errors are strictly positive or negative for each forecaster (depending on whether the paper replicated or not), and this the average *individual* absolute error and the absolute error for the *wisdom-of-crowd* forecast coincide.

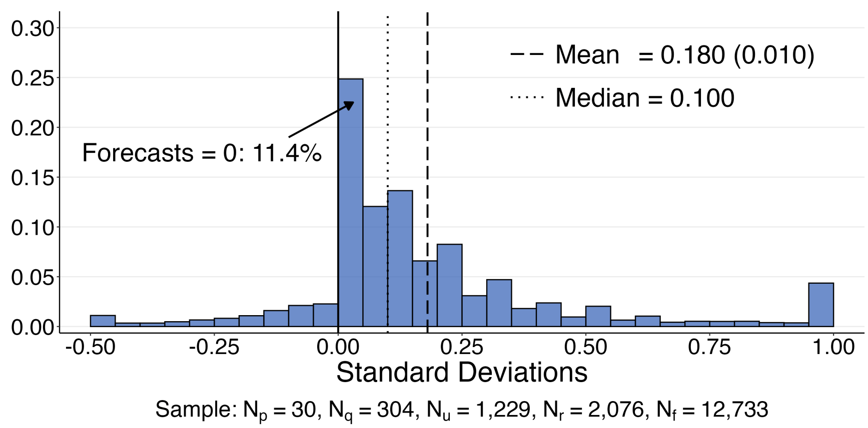
Finally, DellaVigna and Pope (2018a) includes additional data on forecast accuracy by subfield expertise (i.e., forecasts made by faculty who specialize in behavioral economics/lab experiments/psychology versus undergraduate/MBA students), as well as by the amount of time taken to produce forecasts (below median versus above median).

Figure A1: Distributions of Realized and Forecasted Treatment Effects by Project Type

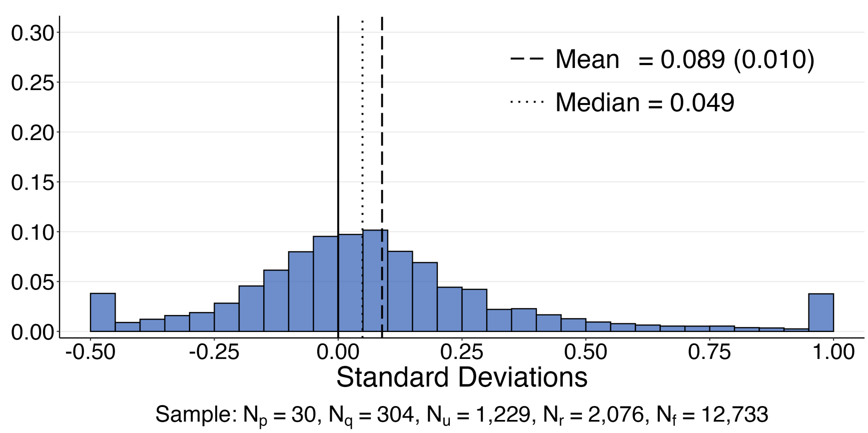
(a) Realized Treatment Effects – RCTs



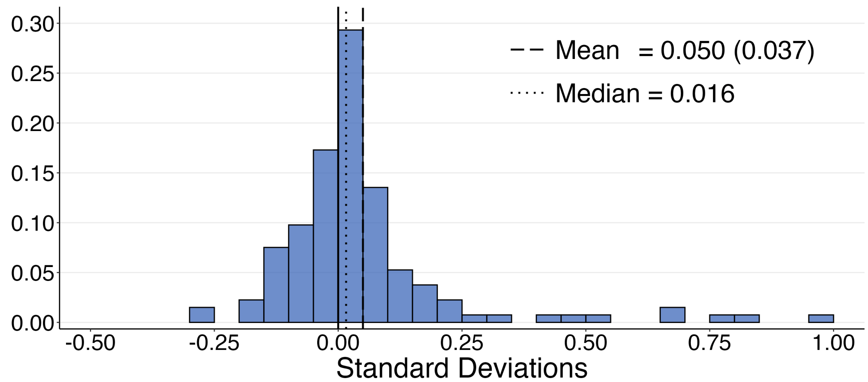
(b) Forecasts of Treatment Effects – RCTs



(c) Overestimation of Treatment Effects – RCTs

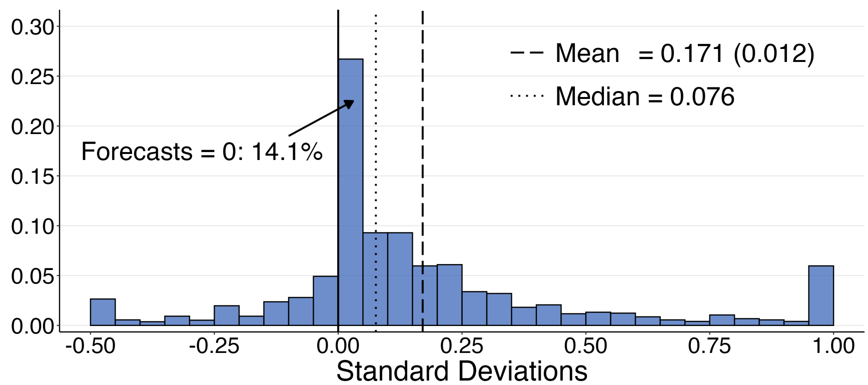


(d) Realized Treatment Effects – Non-RCTs



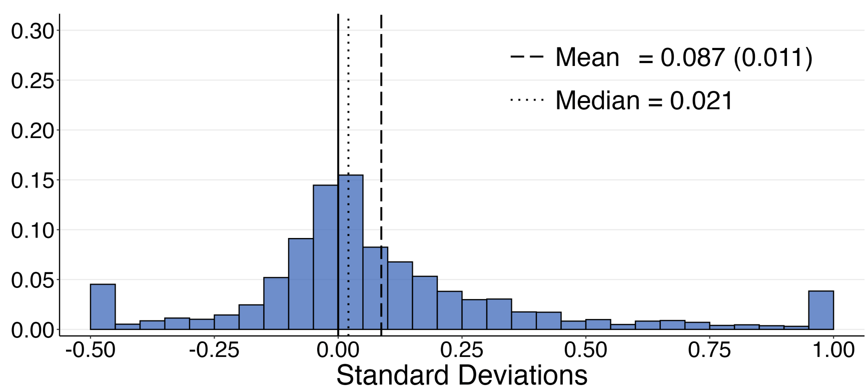
Sample: $N_p = 14$, $N_q = 133$, $N_u = 350$, $N_r = 492$, $N_f = 3{,}250$

(e) Forecasts of Treatment Effects – Non-RCTs



Sample: $N_p = 14$, $N_q = 133$, $N_u = 350$, $N_r = 492$, $N_f = 3{,}250$

(f) Overestimation of Treatment Effects – Non-RCTs

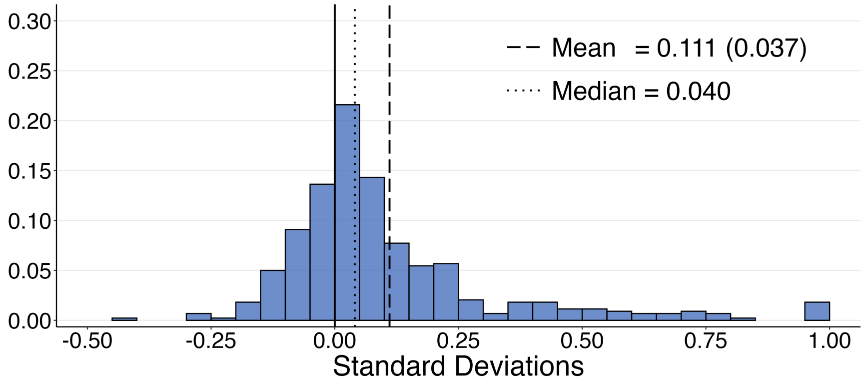


Sample: $N_p = 14$, $N_q = 133$, $N_u = 350$, $N_r = 492$, $N_f = 3{,}250$

Notes: Appendix Figure A1a displays the distribution of realized treatment effects across all normed key questions of RCT projects for which results could be collected. Appendix Figure A1b displays the distribution of the normed forecasts of these treatment effects at the individual forecast level. Appendix Figure A1c displays the distribution of the difference between each forecast of a key question and the corresponding realized result. Appendix Figures A1d-A1f display analogous results for non-RCT projects. Across all figures, means are indicated by dashed lines, medians by vertical dotted lines, while the solid lines indicate the 0 standard deviations mark. Standard errors, clustered at the key question level, are provided in parentheses where relevant. A breakdown of the full sample is provided in the footer of each figure.

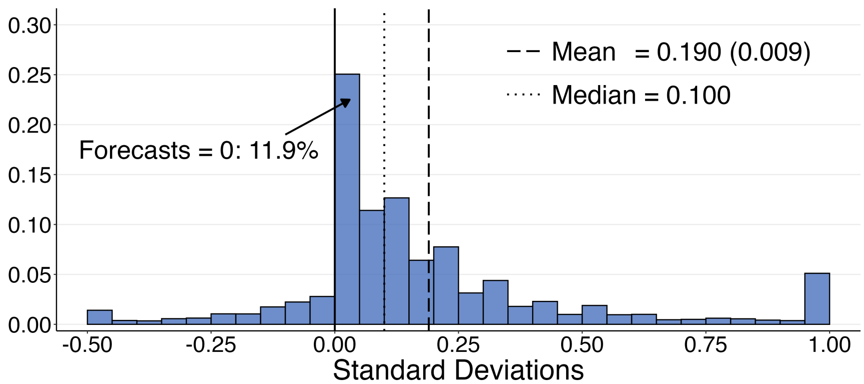
Figure A2: Distributions of Realized and Forecasted Treatment Effects – With Outliers

(a) Realized Treatment Effects



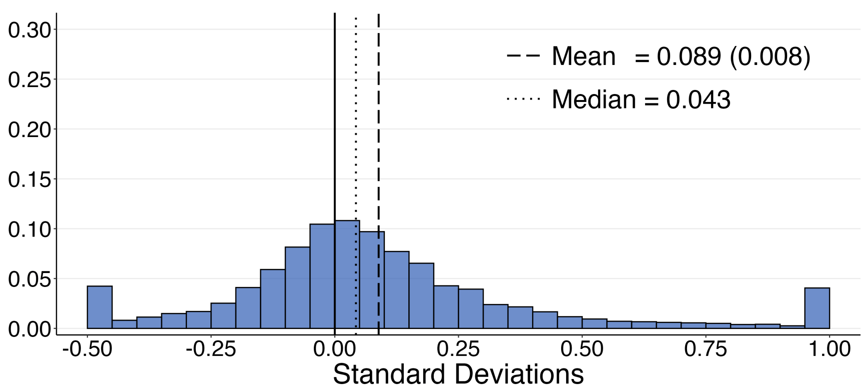
Sample: N_p = 44, N_q = 440, N_u = 1,420, N_r = 2,568, N_f = 16,100

(b) Forecasts of Treatment Effects



Sample: N_p = 44, N_q = 440, N_u = 1,420, N_r = 2,568, N_f = 16,100

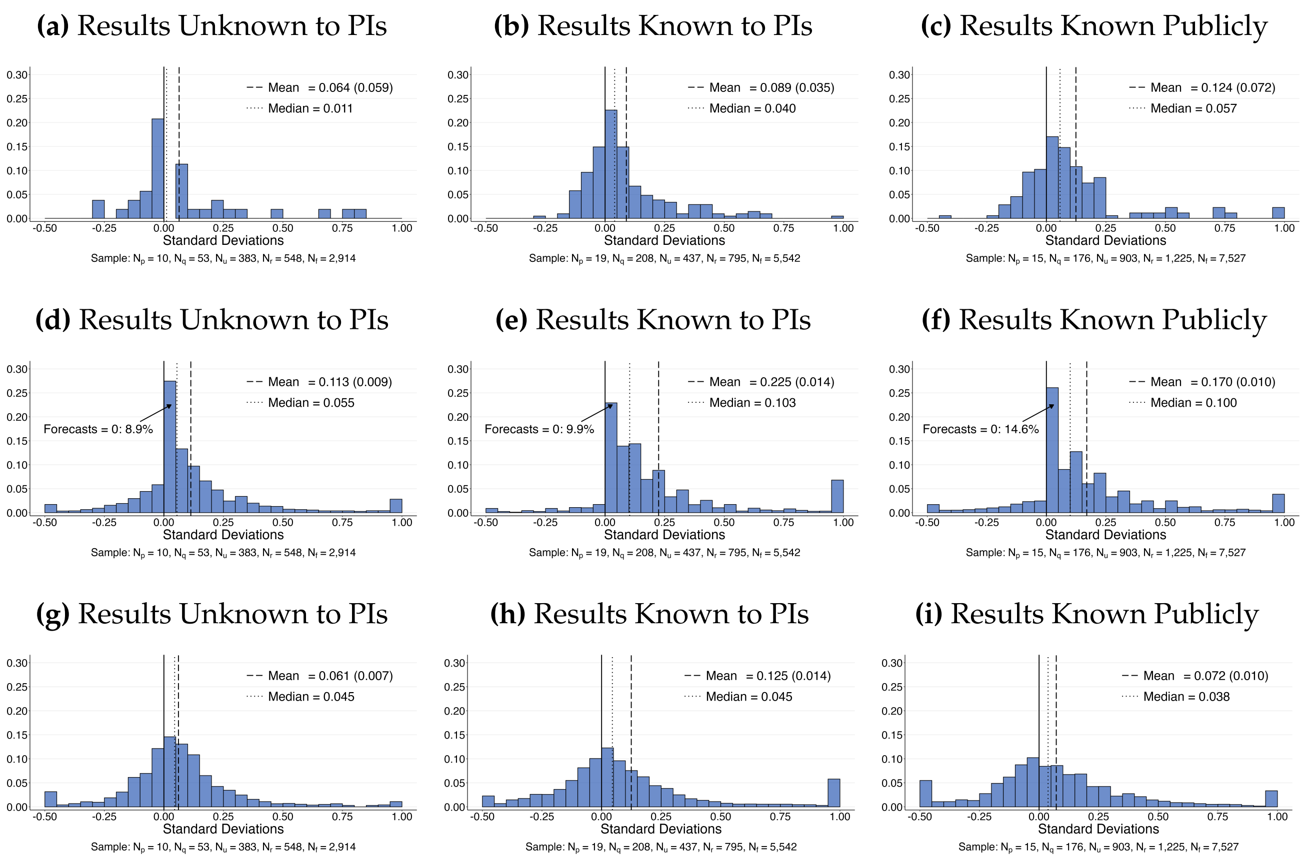
(c) Overestimation of Treatment Effects



Sample: N_p = 44, N_q = 440, N_u = 1,420, N_r = 2,568, N_f = 16,100

Notes: Appendix Figure A1a displays the distribution of realized treatment effects across all normed key questions for which results could be collected, including 3 outlier key questions, as detailed in Appendix A.3. Appendix Figure A1b displays the distribution of the normed forecasts of these treatment effects at the individual forecast level. Appendix Figure A1c displays the distribution of the difference between each forecast of a key question and the corresponding realized result. Across all figures, means are indicated by dashed lines, medians by vertical dotted lines, while the solid lines indicate the 0 standard deviations mark. Standard errors, clustered at the key question level, are provided in parentheses where relevant. A breakdown of the full sample is provided in the footer of each figure.

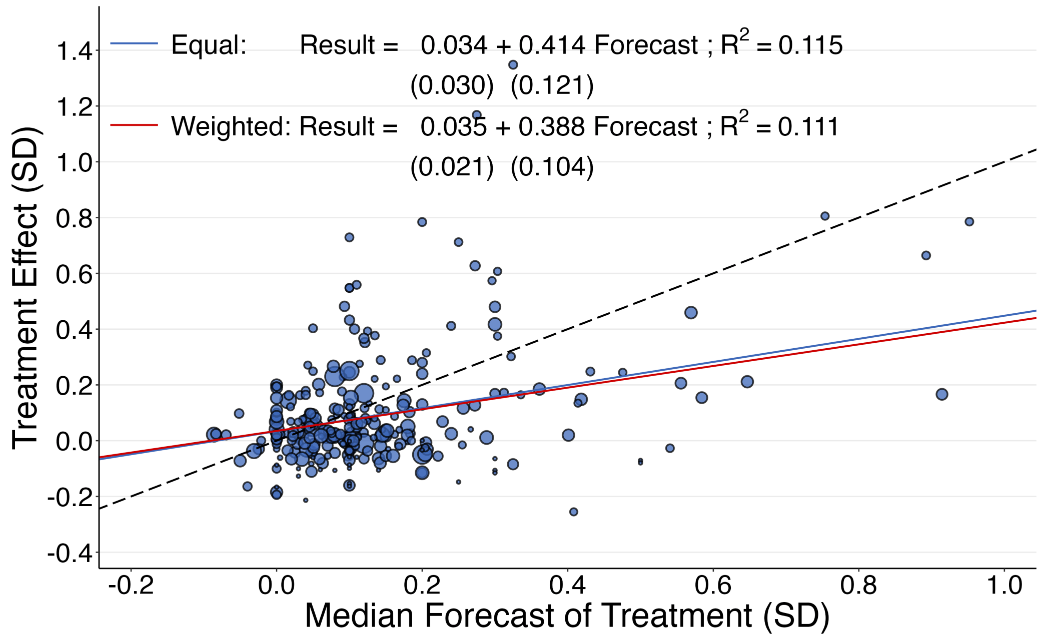
Figure A3: Distributions of Realized and Forecasted Treatment Effects by Result Knowledge



Notes: Across all normed key questions of RCT projects for which results could be collected, appendix Figure A3 displays the distribution of realized treatment effects, forecasts, and the difference betweeen each forecast and the resulting treatment effect. These have then been categorised into three columns, to show the distributions for the following cases: (first column) before the survey was ran the results were unknown to the Principal Investigators (PIs) and public or only baseline data had been collected; (second column) the results were known to the PIs but not the public; (third column) the results were publicly known or data had been collected.

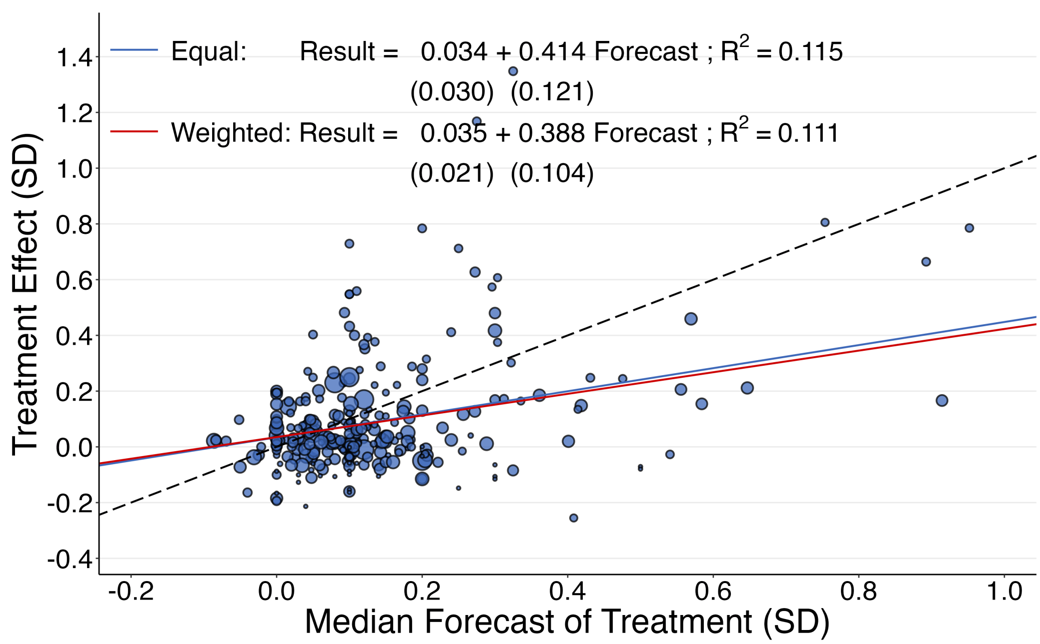
Figure A4: Predictability of Results from Forecasts Varying Aggregation & Winsorization

(a) Median Forecast & No Winsorization



Sample: $N_p = 37$, $N_q = 274$, $N_u = 1{,}388$, $N_r = 2{,}486$, $N_f = 14{,}743$

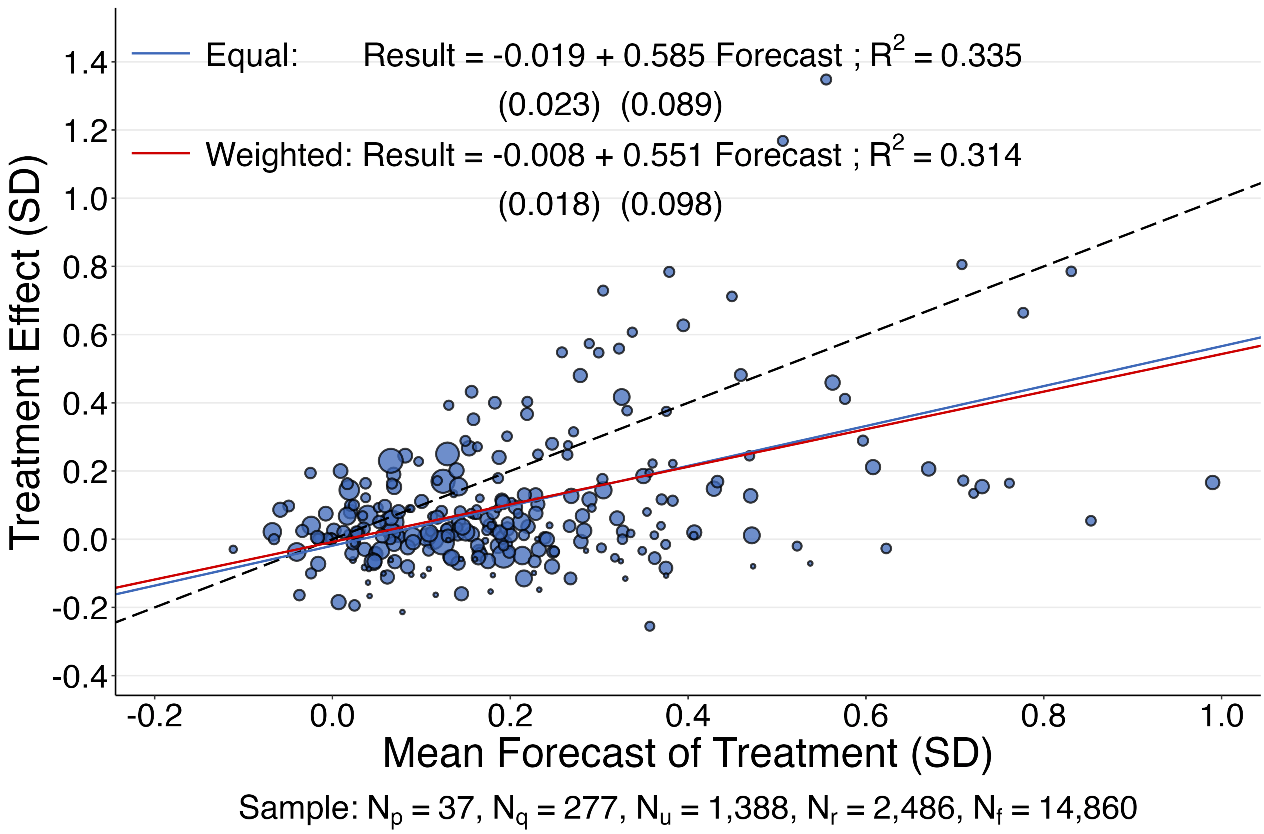
(b) Mean Forecast & 1.5 SD Around 0 Winsorization



Sample: $N_p = 37$, $N_q = 274$, $N_u = 1{,}388$, $N_r = 2{,}486$, $N_f = 14{,}743$

Notes: Appendix Figure A4a presents a scatter plot of all normed key questions of treatment effects with a minimum of 20 responses, with the median unwinsorized normed forecast on the x-axis, and the actual normed treatment effect on the y-axis. Points are scaled to represent the number of responses collected by each key question. The solid line displays the unweighted fitted regression provided in the top-left of the figure, with standard errors, clustered at the project level, provided in parentheses. The dashed line displays the 45-degree line. Appendix Figure A4b displays an analogous result but with the average normed forecast on the x-axis, winsorizing at 1.5 standard deviations around 0. A breakdown of the full sample is provided in the footer of each figure.

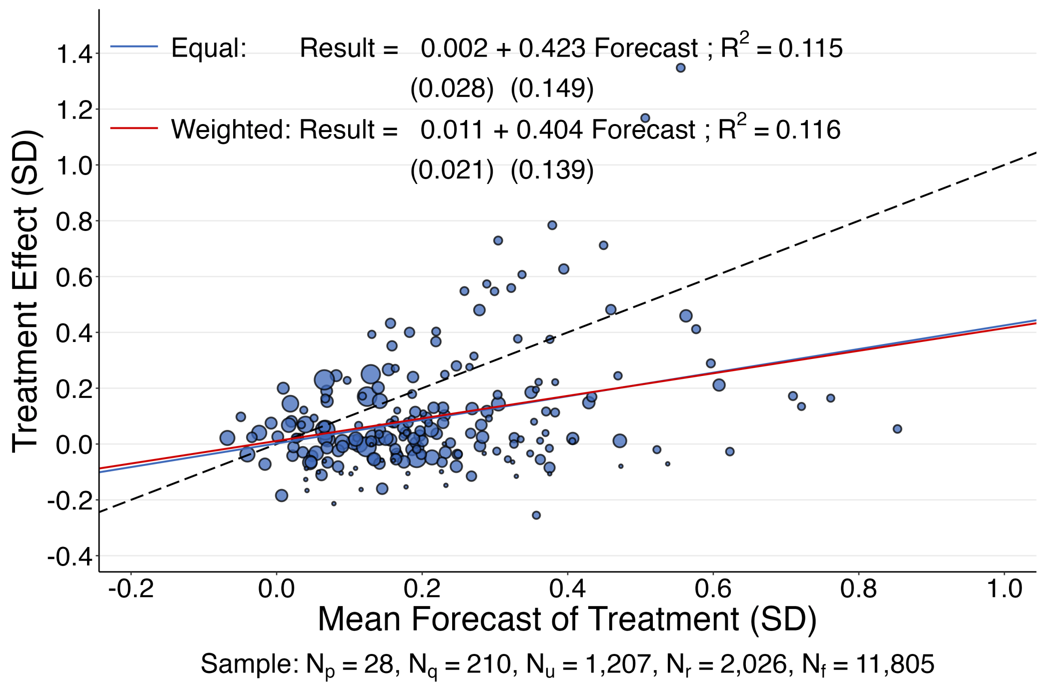
Figure A5: Predictability of Results from Average Forecasts – With Outliers



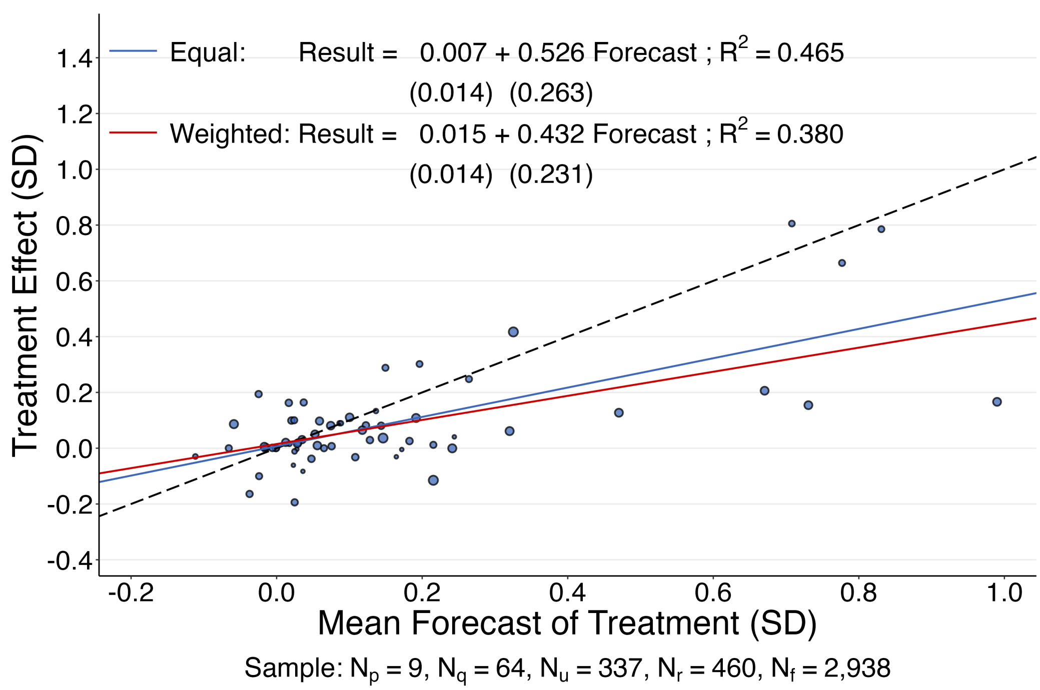
Notes: Appendix Figure A5 presents a scatter plot of all normed key questions of treatment effects with a minimum of 20 responses, with the average normed forecast on the x-axis, and the actual normed treatment effect on the y-axis. The difference compared to Figure 3 is that we include 3 outlier key questions (see Appendix A.3). Points are scaled to represent the number of responses collected by each key question. The solid line displays the unweighted fitted regression provided in the top-left of the figure, with standard errors, clustered at the project level, provided in parentheses. Included in the regression are 4 key questions lying outside the plot axes ranges, with average forecasts and treatment effects of: (-0.01,-0.41), (-0.23,0.23), (1.0,0.26), (0.56,1.35), respectively. The dashed line displays the 45-degree line. A breakdown of the full sample is provided in the footer of the figure.

Figure A6: Predictability of Results from Forecasts by Project Type

(a) RCTs

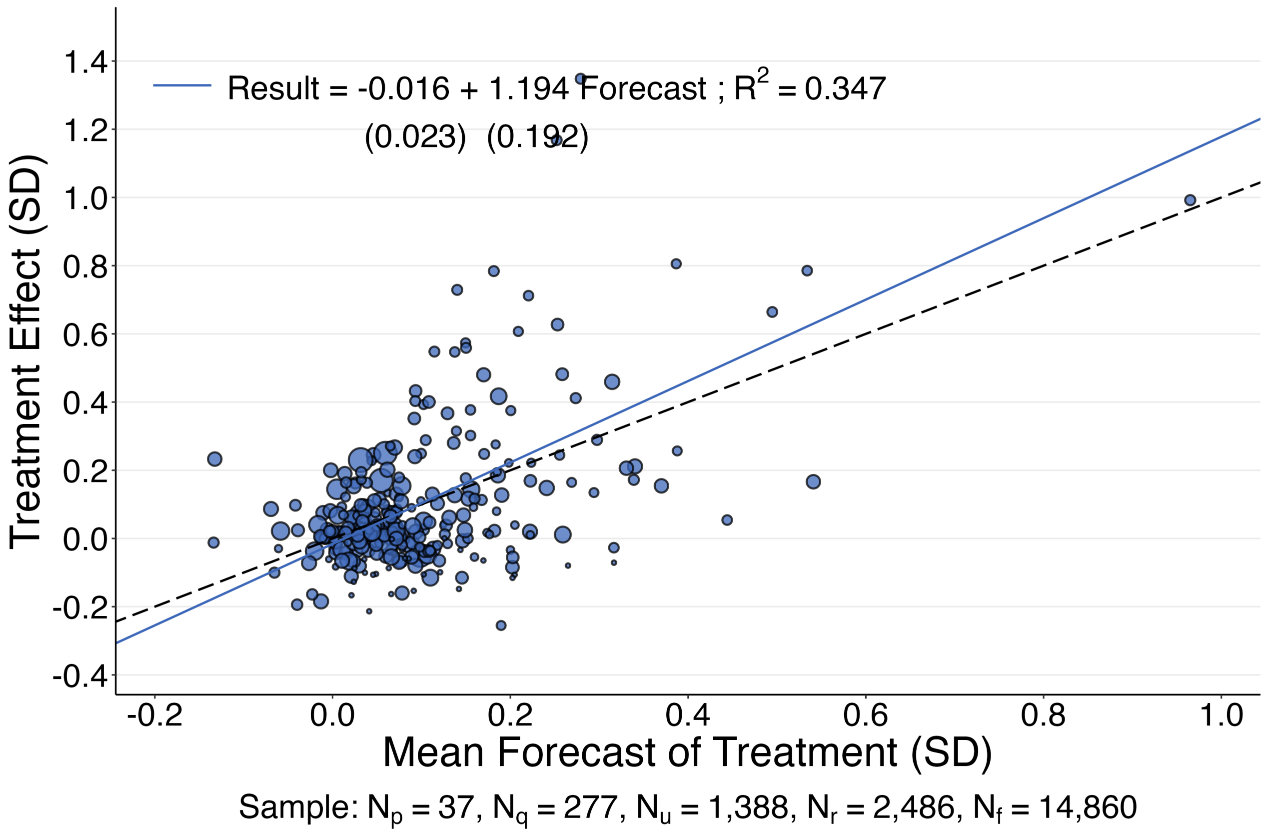


(b) Non-RCTs



Notes: Appendix Figure A6a presents a scatter plot of all normed key questions of treatment effects from RCT projects with a minimum of 20 responses, with the average normed forecast on the x-axis, and the actual normed treatment effect on the y-axis. Points are scaled to represent the number of responses collected by each key question. The solid line displays the unweighted fitted regression provided in the top-left of the figure, with standard errors, clustered at the project level, provided in parentheses. The dashed line displays the 45-degree line. Appendix Figure A6b displays an analogous result for key questions from non-RCT projects. A breakdown of the full sample is provided in the footer of each figure.

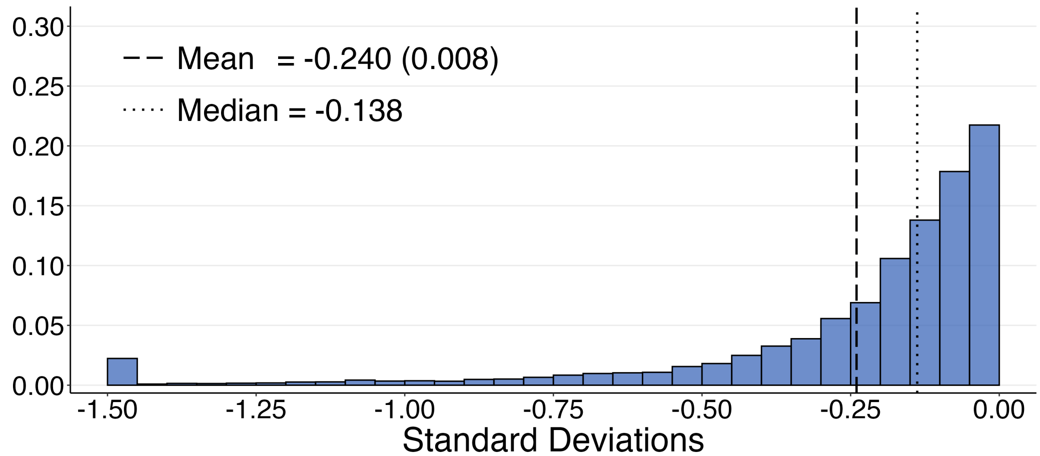
Figure A7: Predictability of Results from Shrunk Weighted Average Forecasts – With Outliers



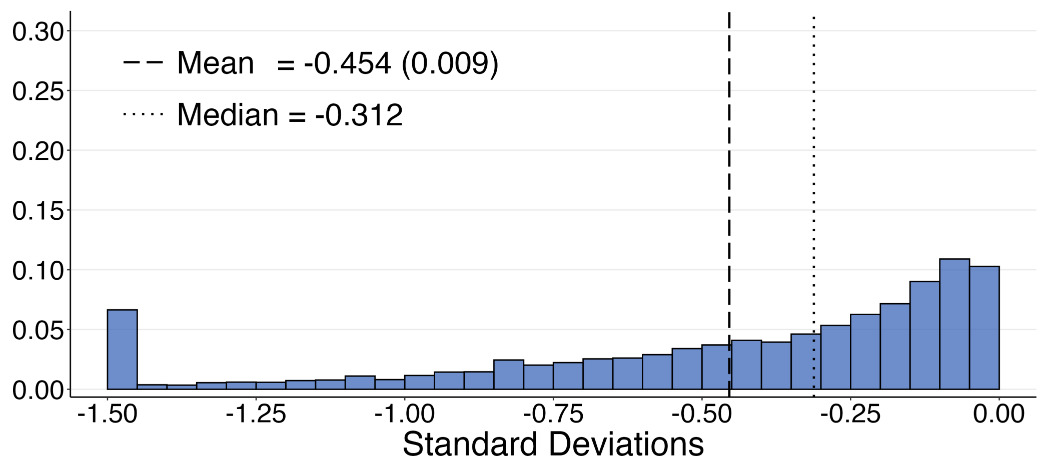
Notes: Appendix Figure A7 presents a scatter plot of all normed key questions of treatment effects with a minimum of 20 responses, with the out-of-sample shrunk average normed forecast on the x-axis, and the actual normed treatment effect on the y-axis. The shrunk average normed forecast applies a weighted out-of-sample shrinkage adjustment, transforming the average forecast using coefficients estimated from regressing realized results on forecasts in all other studies while increasing/decreasing the weighting on forecasters with highest/lowest predictive ability (see Appendix D). The difference compared to Figure 7 is that we include 3 outlier key questions (see Appendix A.3). Points are scaled to represent the number of responses collected by each key question. The solid line displays the unweighted fitted regression provided in the top-left of the figure, with standard errors, clustered at the project level, provided in parentheses. Included in the regression are 4 key questions lying outside the plot axes ranges, with average forecasts and treatment effects of: (-0.01,-0.41), (-0.23,0.23), (1.0,0.26), (0.56,1.35), respectively. The dashed line displays the 45-degree line. A breakdown of the full sample is provided in the footer of the figure.

Figure A8: Distribution of Absolute Forecast Accuracy by Key Question Subject

(a) Normed Treatment Effects



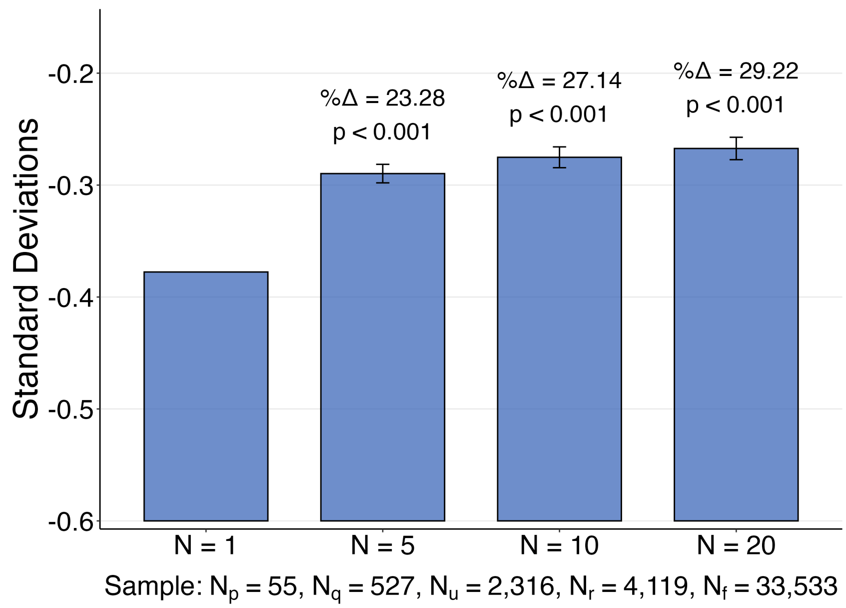
(b) Summary Statistics



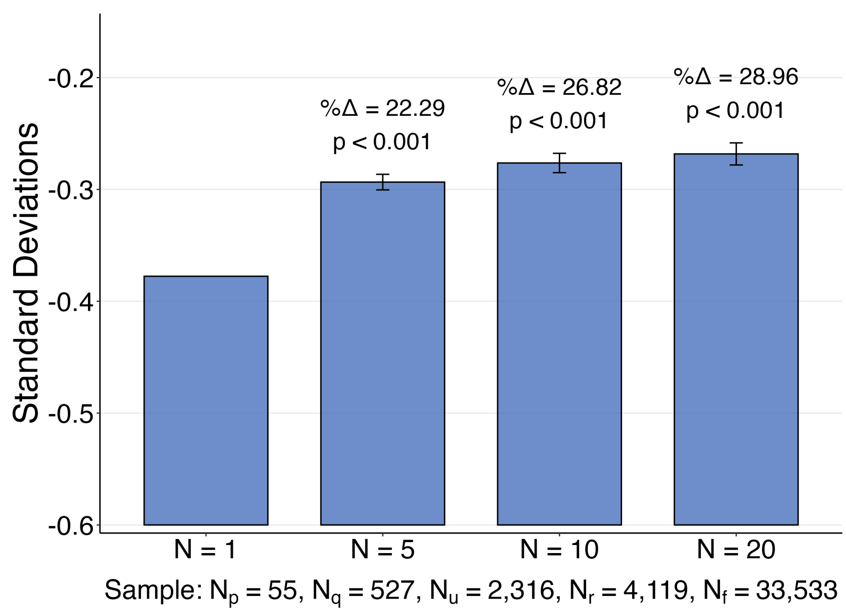
Notes: Appendix Figure A8a displays the distribution of the negative of the absolute difference between the forecasted and realized result (i.e. absolute forecast accuracy) for all key questions of normed treatment effects for which results could be collected. Appendix Figure A8b displays an analogous distribution but for key questions of summary statistics. Across both figures, means are indicated by dashed lines and medians by vertical dotted lines. Standard errors, clustered at the key question level, are provided in parentheses where relevant. A breakdown of the full sample is provided in the footer of each figure.

Figure A9: Accuracy of the Mean & Median Forecasts of Simulated Groups of Forecasters

(a) Mean Forecast



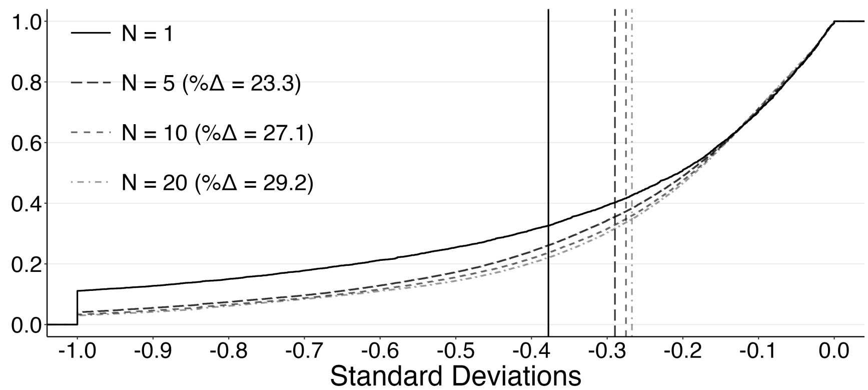
(b) Median Forecast



Notes: Appendix Figure A9a displays the difference in the mean absolute forecast accuracy across simulated groups of size $N = 1, 5, 10, 20$. For each key question with a minimum of 20 responses, and for which results could be collected, N forecasts are sampled, with replacement, calculating the median forecast and the associated absolute accuracy, repeating this procedure 100 times. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the key question level, while the displayed p-value above each bar corresponds to that of a two-tailed t-test between individual forecasts ($N = 1$) and each of the three group sizes. The percent change from individual forecasts to each of the other three group sizes is also displayed above each bar. Appendix Figure A9b provides analogous results for the difference in the median absolute forecast across the simulated groups. A breakdown of the full sample is provided in the footer of each figure.

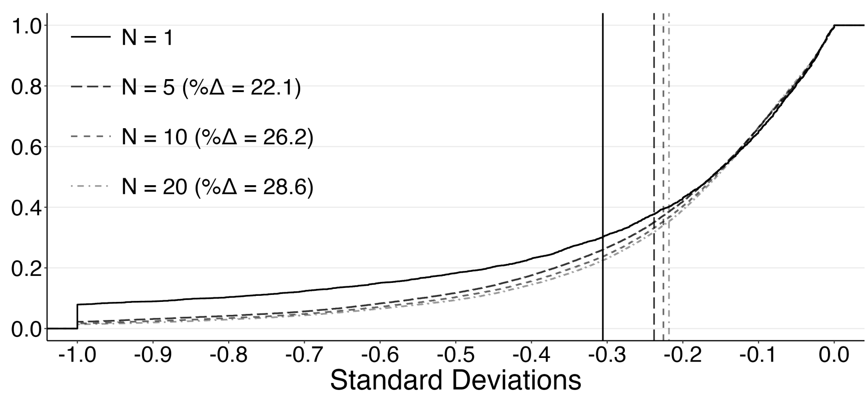
Figure A10: Distribution of Accuracy by Simulated Groups of Forecasters

(a) Pooled



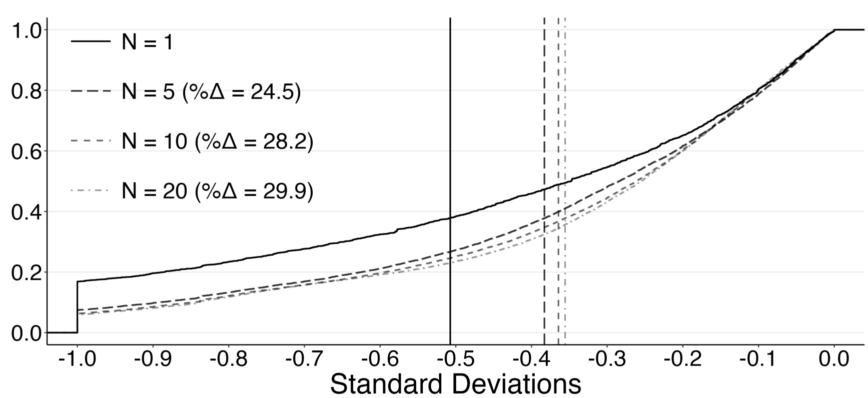
Sample: $N_p = 55$, $N_q = 527$, $N_u = 2{,}316$, $N_r = 4{,}119$, $N_f = 33{,}533$

(b) Treatment Effects



Sample: $N_p = 42$, $N_q = 339$, $N_u = 1{,}461$, $N_r = 2{,}779$, $N_f = 18{,}902$

(c) Summary Statistics

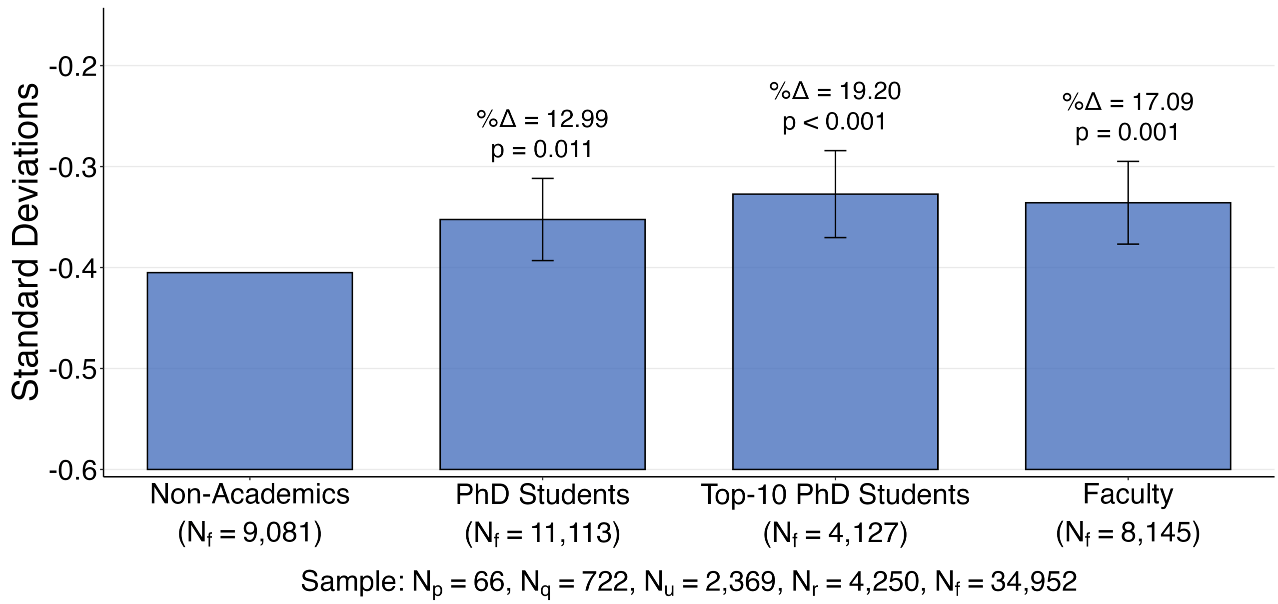


Sample: $N_p = 23$, $N_q = 188$, $N_u = 1{,}428$, $N_r = 2{,}202$, $N_f = 14{,}631$

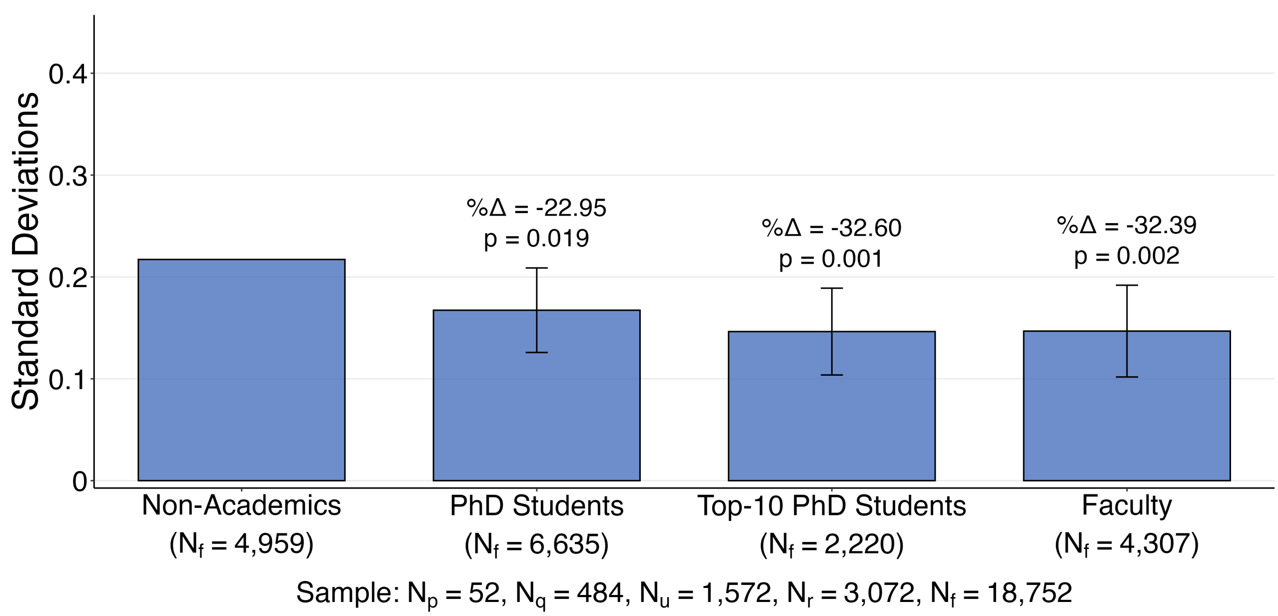
Notes: Appendix Figure A10a displays the distribution of absolute forecast accuracy across simulated groups of size $N = 1, 5, 10, 20$. For each key question with a minimum of 20 responses, and for which results could be collected, N forecasts are sampled, with replacement, calculating the average forecast and the associated absolute accuracy, repeating this procedure 100 times. Means, across all repetitions, are indicated by vertical lines. The percent change from the average of individual forecasts to each of the other three group sizes is displayed in parentheses next to each group label. Appendix Figures A10b and A10c display analogous distributions for key questions of treatment effects and summary statistics separately. A breakdown of the full sample is provided in the footer of each figure.

Figure A11: Accuracy & Forecast Magnitude by Expertise – Academic Status

(a) Absolute Forecast Accuracy



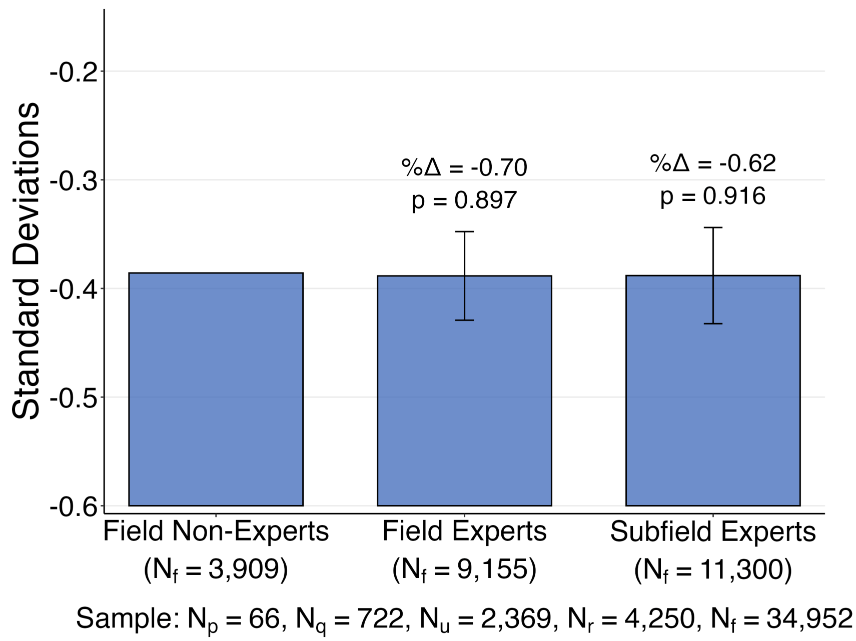
(b) Forecasts of Treatments



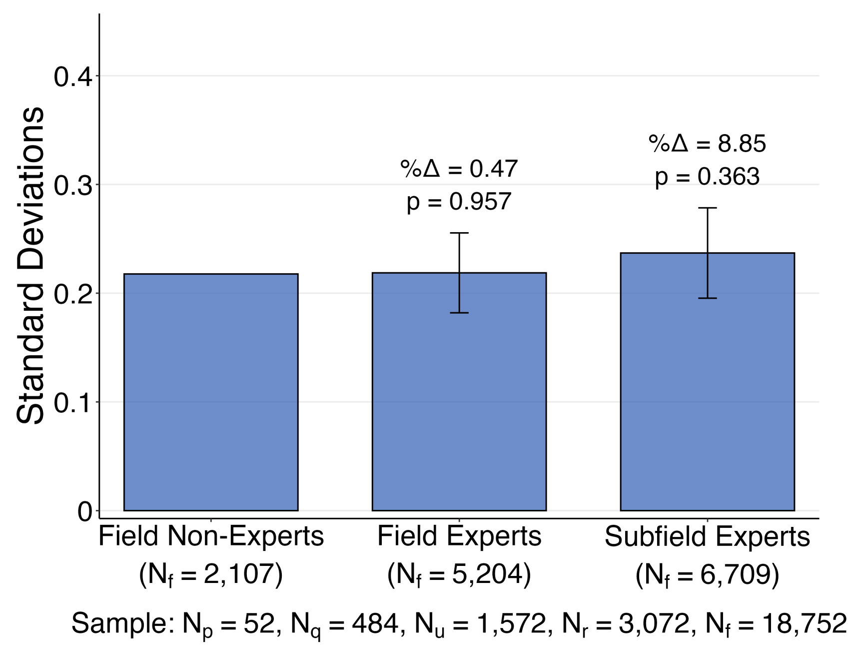
Notes: Appendix Figure A11a displays the difference in the average absolute forecast accuracy across individuals by academic status, specifically: non-academics, PhD students, PhD students from a top-10 PhD program, and faculty members. Individuals without a recorded academic status are excluded. Similarly, Appendix Figure A11b displays the difference in the average normed forecasts of treatments across each group. Estimates for each figure are from a univariate regression specification which includes key question fixed effects. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-value above each bar corresponds to that of a two-tailed t-test between non-academics and each of the two other groups. The percent change from non-academics to each of the other two groups is also displayed above each bar. A breakdown of the full sample is provided in the footer of each figure, with a specific count of the number of forecasts from each group provided below each label in parentheses.

Figure A12: Accuracy & Forecast Magnitude by Expertise – Area of Knowledge

(a) Absolute Forecast Accuracy



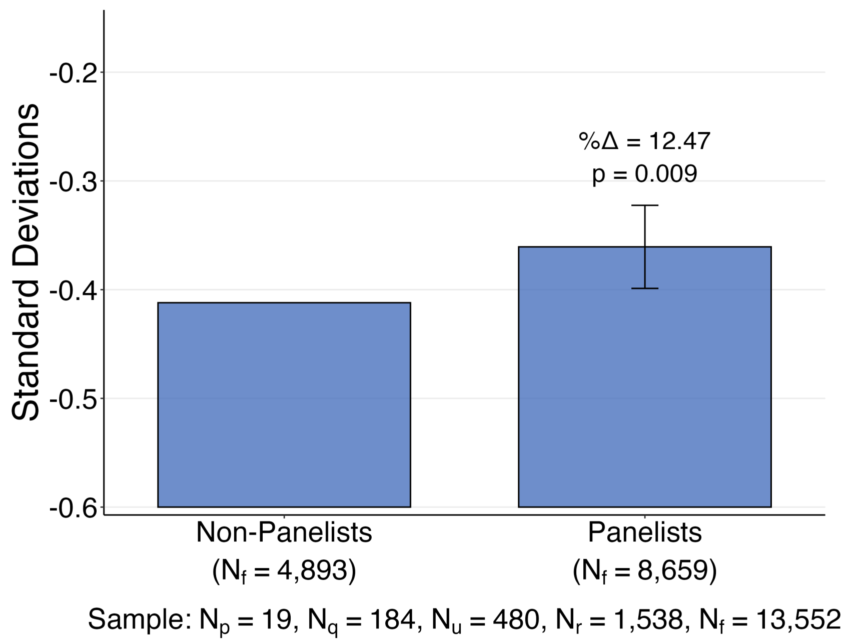
(b) Forecasts of Treatments



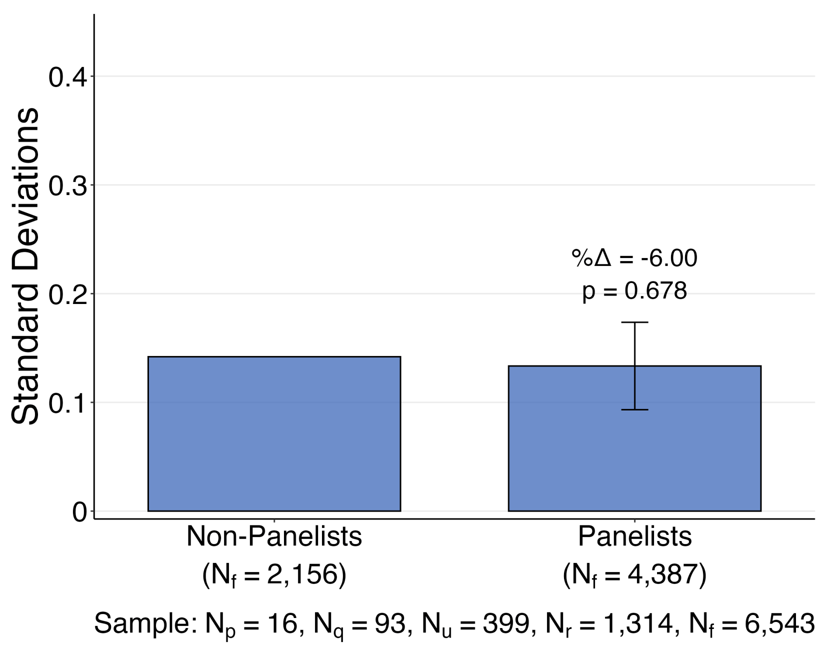
Notes: Appendix Figure A12a displays the difference in the average absolute forecast accuracy across academics by area of knowledge, specifically: field non-experts, field experts (subfield non-experts), and subfield experts. Individuals without a recorded field and subfield are excluded. Similarly, Appendix Figure A12b displays the difference in the average normed forecasts of treatments across each group. Estimates for each figure are from a univariate regression specification which includes key question fixed effects. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-value above each bar corresponds to that of a two-tailed t-test between field non-experts and each of the two other groups. The percent change from field non-experts to each of the other two groups is also displayed above each bar. A breakdown of the full sample is provided in the footer of each figure, with a specific count of the number of forecasts from each group provided below each label in parentheses.

Figure A13: Accuracy & Forecast Magnitude by Panel Membership

(a) Absolute Forecast Accuracy



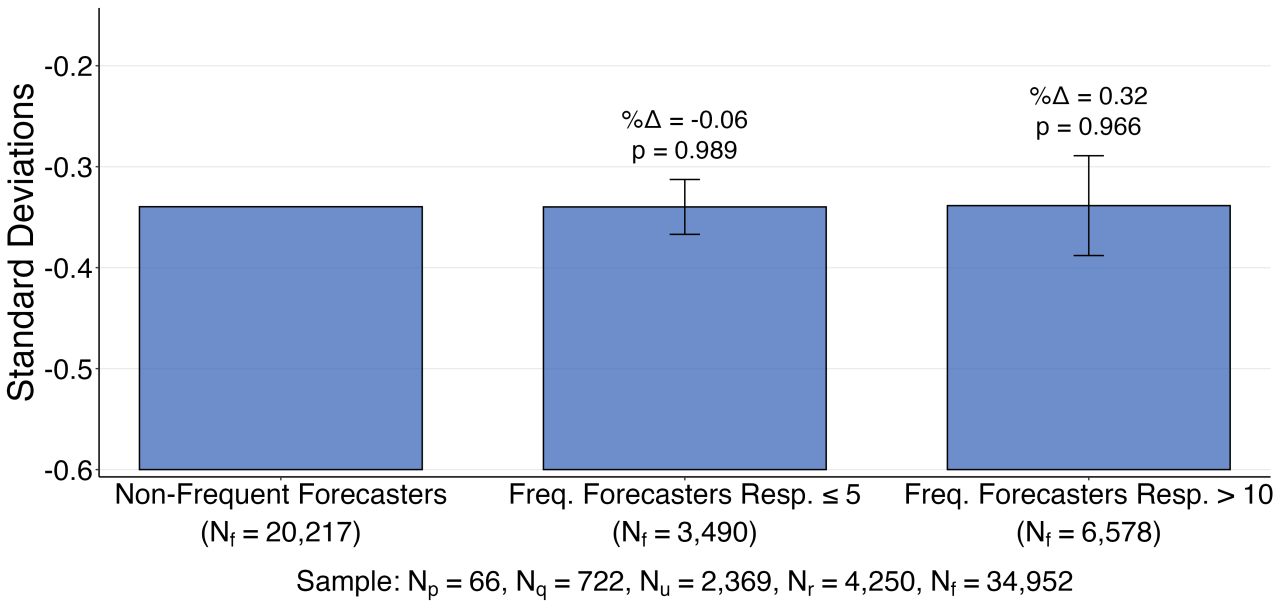
(b) Forecasts of Treatments



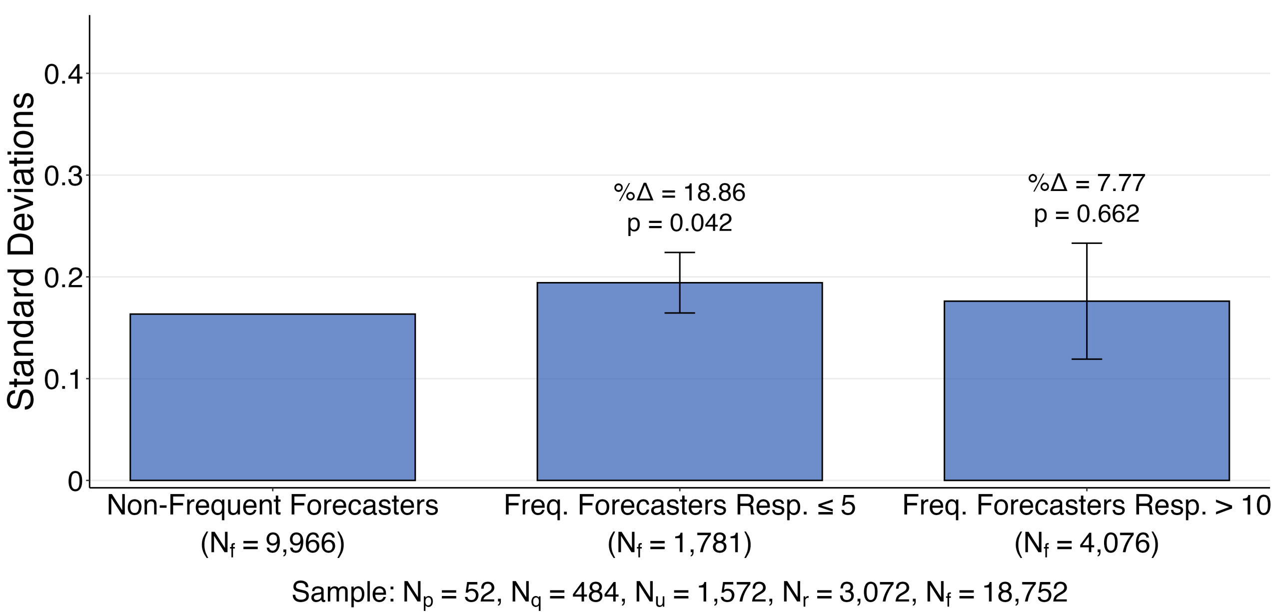
Notes: Appendix Figure A13a displays the difference in the average absolute forecast accuracy between non-panelists and panelists for all projects posted in September 2023 and beyond. Similarly, Appendix Figure A13b displays the difference in the average normed forecasts of treatments between those two groups. Estimates for each figure are from a univariate regression specification which includes key question fixed effects. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-value above each bar corresponds to that of a two-tailed t-test between non-panelists and panelists. The percent change from non-panelists to panelists is also displayed above each bar. A breakdown of the full sample is provided in the footer of each figure, with a specific count of the number of forecasts from each group provided below each label in parentheses.

Figure A14: Accuracy & Forecast Magnitude by Experience & Learning

(a) Absolute Forecast Accuracy



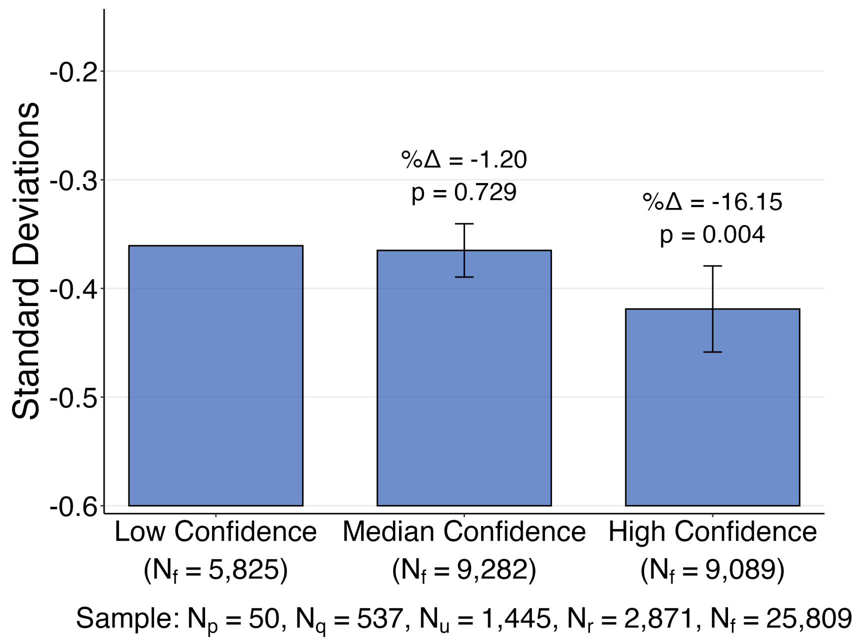
(b) Forecasts of Treatments



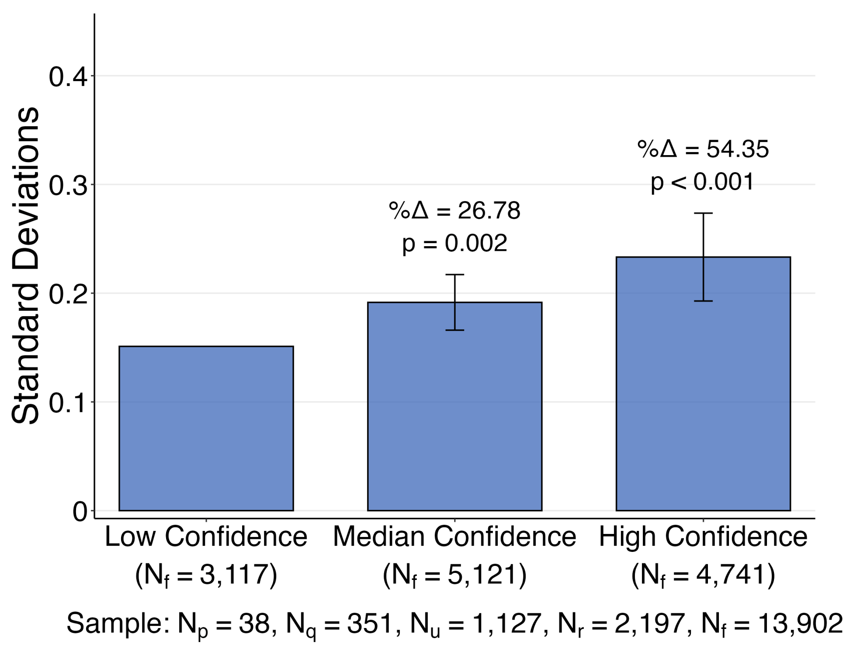
Notes: Appendix Figure A14a displays the difference in the average absolute forecast accuracy across forecasters by experience over time, specifically: non-frequent forecasters (less than 5 responses), frequent forecasters' (10 or more responses) first 5 responses, and frequent forecasters' responses beyond the first 10. Similarly, Appendix Figure A14b displays the difference in the average normed forecasts of treatments across each group. Estimates for each figure are from a univariate regression specification which includes key question fixed effects. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-value above each bar corresponds to that of a two-tailed t-test between non-frequent forecasters and each of the two other groups. The percent change from non-frequent forecasters to each of the other two groups is also displayed above each bar. A breakdown of the full sample is provided in the footer of each figure, with a specific count of the number of forecasts from each group provided below each label in parentheses.

Figure A15: Accuracy & Forecast Magnitude by Confidence

(a) Absolute Forecast Accuracy



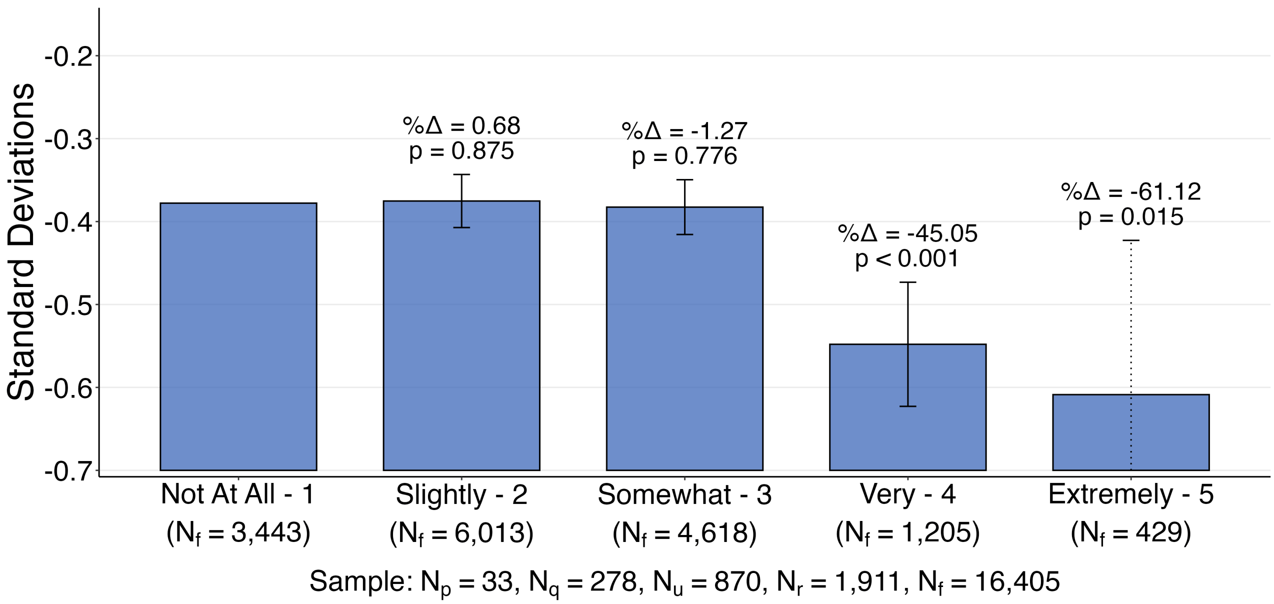
(b) Forecasts of Treatments



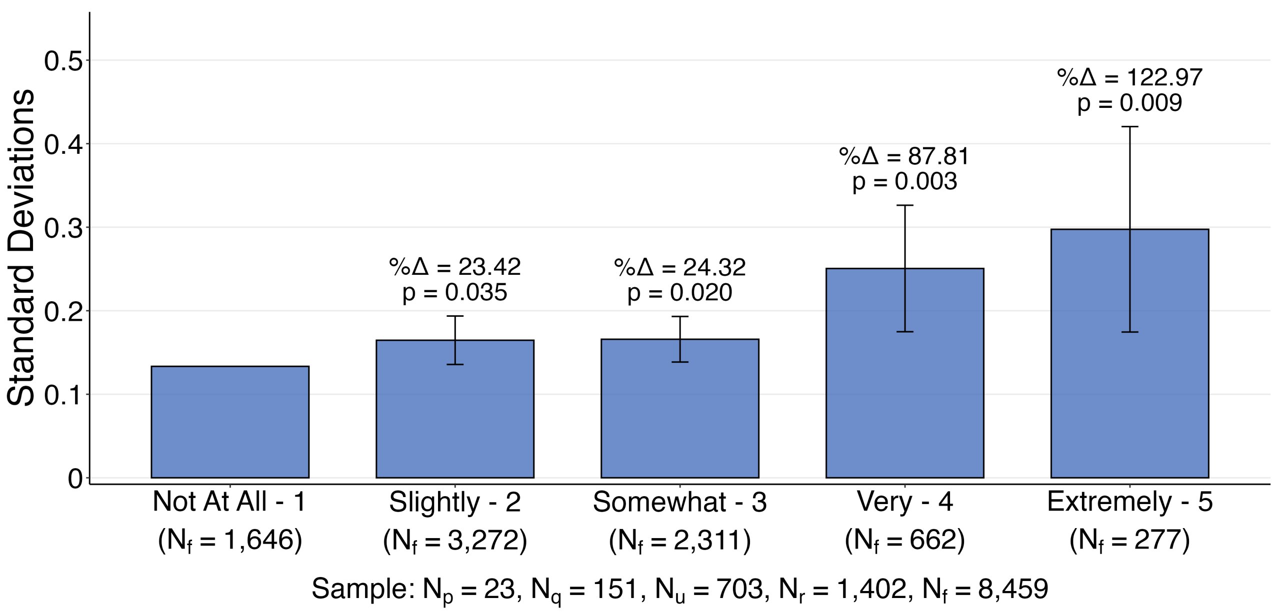
Notes: Appendix Figure A15a displays the difference in the average absolute forecast accuracy across individuals by confidence level, specifically, for all projects with a confidence question, and within each key question, those with: low (below-median), median, and high (above-median) confidence (individuals without a recorded confidence level are excluded). Similarly, Appendix Figure A15b displays the difference in the average normed forecasts of treatments across each group. Estimates for each figure are from a univariate regression specification which includes key question fixed effects. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-value above each bar corresponds to that of a two-tailed t-test between low confidence responses and each of the two other groups. The percent change from low confidence responses to each of the other two groups is also displayed above each bar. A breakdown of the full sample is provided in the footer of each figure, with a specific count of the number of forecasts from each group provided below each label in parentheses.

Figure A16: Accuracy & Forecast Magnitude by Reported Confidence

(a) Absolute Forecast Accuracy



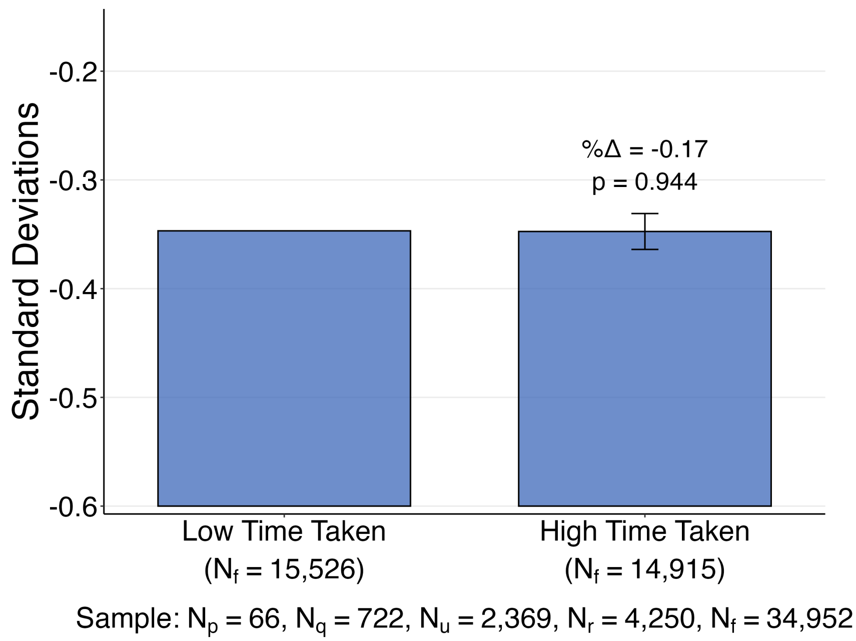
(b) Forecasts of Treatments



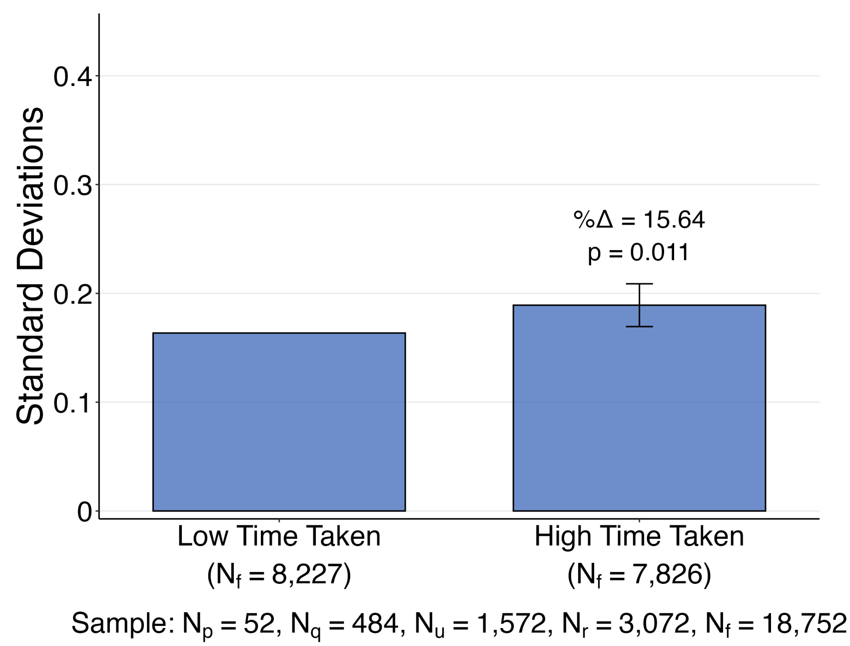
Notes: Appendix Figure A16a displays the difference in the average absolute forecast accuracy across individuals by reported confidence level, specifically, for all projects with a confidence question utilising the platform suggested categories, which are listed on the x-axis above. Individuals without a recorded confidence level are excluded. Similarly, Appendix Figure A16b displays the difference in the average normed forecasts of treatments across each group. Estimates for each figure are from a univariate regression specification which includes key question fixed effects. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-value above each bar corresponds to that of a two-tailed t-test between "Not At All" reported confidence and the other groups. The percent change from "Not At All" confident to the other groups is also displayed above each bar. A breakdown of the full sample is provided in the footer of each figure, with a specific count of the number of forecasts from each group provided below each label in parentheses. The dotted line confidence bound indicates that it is truncated and goes below the x-axis.

Figure A17: Accuracy & Forecast Magnitude by Time Taken

(a) Absolute Forecast Accuracy



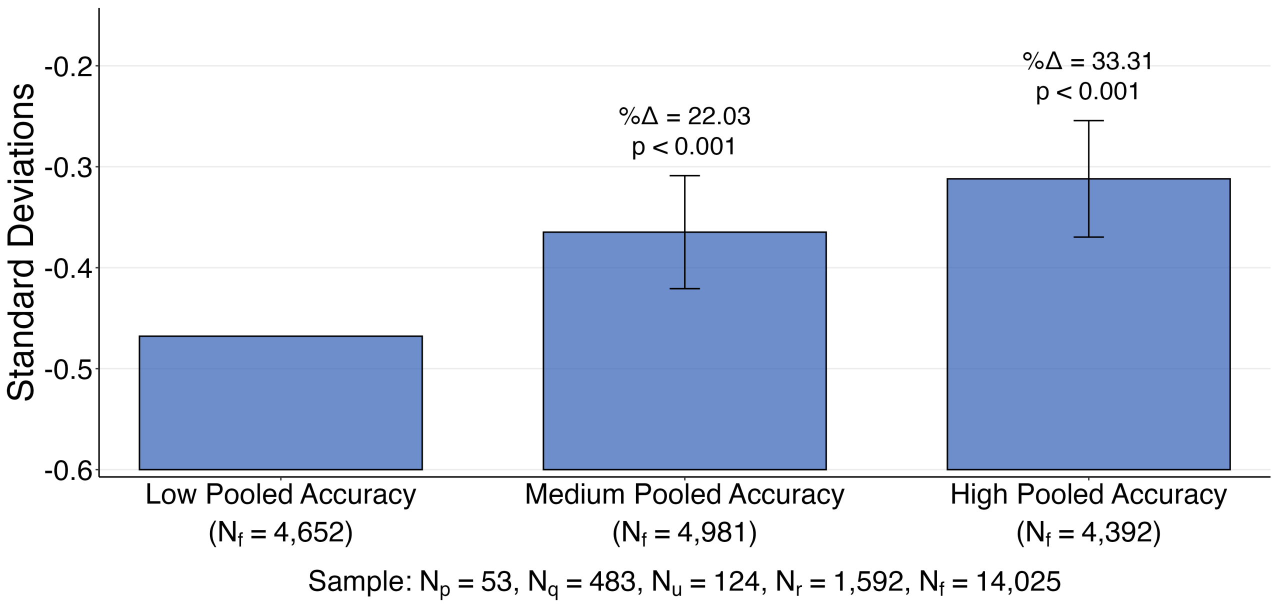
(b) Forecasts of Treatments



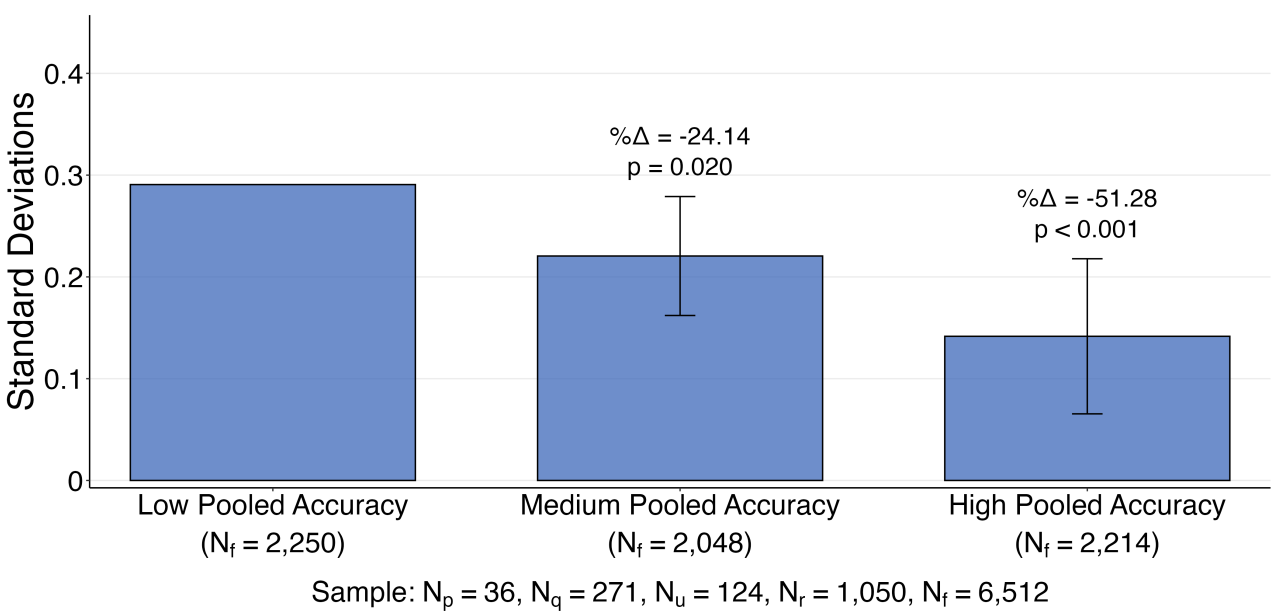
Notes: Appendix Figure A17a displays the difference in the average absolute forecast accuracy across individuals by the time taken to complete a response, specifically, within each project, those with: low (below-median), and high (above-median) time taken (individuals that spend more than 45 minutes are excluded). Similarly, Appendix Figure A17b displays the difference in the average normed forecasts of treatments between those two groups. Estimates for each figure are pulled from a univariate regression specification which includes key question fixed effects. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-value above each bar corresponds to that of a two-tailed t-test between the low and high time taken groups. The percent change from the low time taken set of responses to the high time taken is also displayed above each bar. A breakdown of the full sample is provided in the footer of each figure, with a specific count of the number of forecasts from each group provided below each label in parentheses.

Figure A18: Accuracy & Forecast Magnitude by In-Sample Accuracy

(a) Absolute Forecast Accuracy



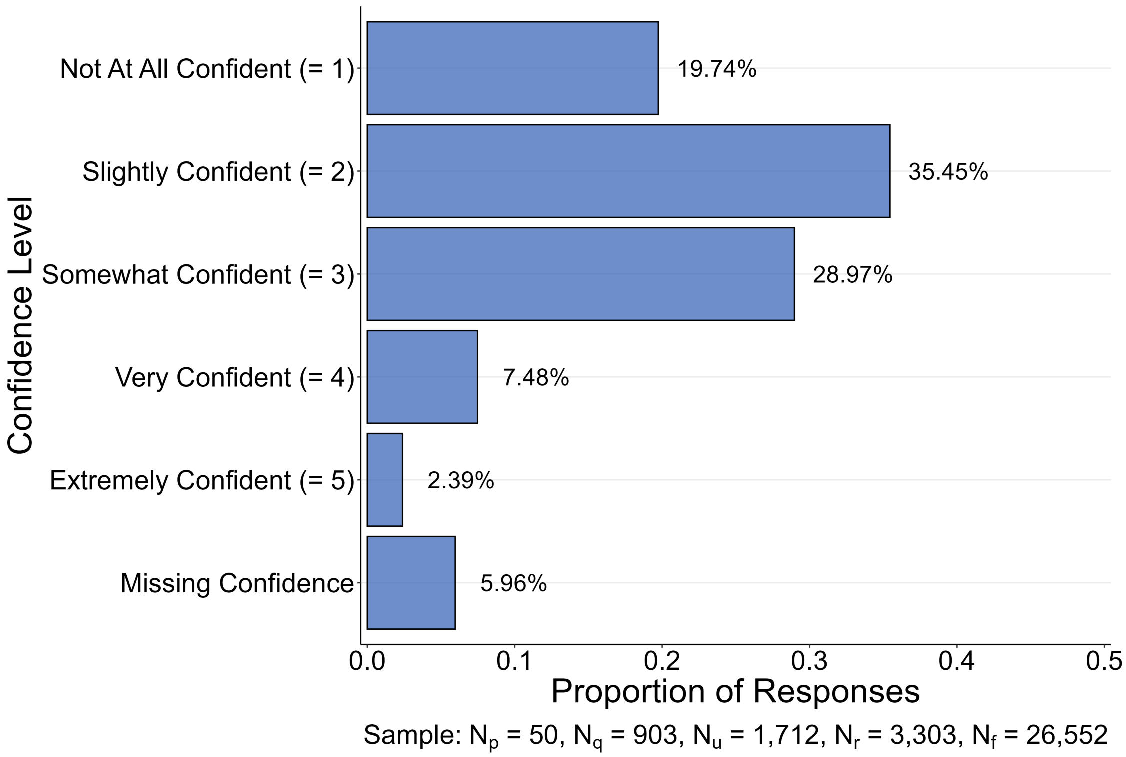
(b) Forecasts of Treatments



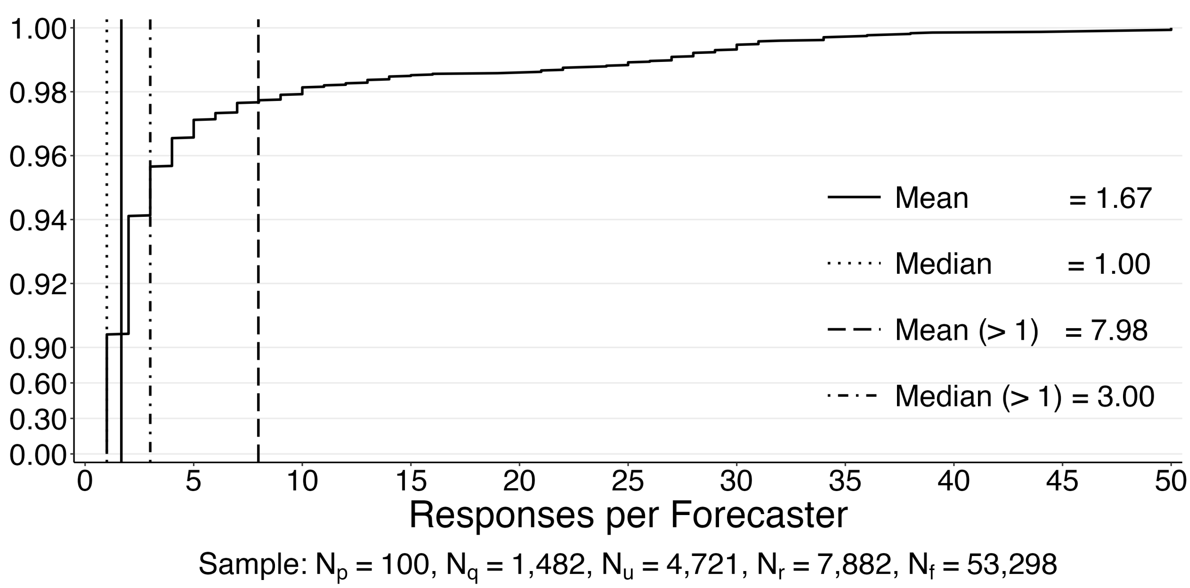
Notes: Appendix Figure A18a displays the difference in the average absolute forecast accuracy across forecasters by in-sample accuracy levels (as defined in Figure 5), specifically, for all projects with a minimum of 20 responses, and for all forecasters with a minimum of 5 responses with results, those with: low (bottom-third), medium (middle-third), and high (top-third) in-sample accuracy. Similarly, Appendix Figure A18b displays the difference in the average normed forecasts of treatments across each group. Estimates for each figure are from a univariate regression specification which includes key question fixed effects. Error bars indicate 95% confidence intervals constructed using standard errors clustered at the forecaster level, while the displayed p-value corresponds to that of a two-tailed t-test between the low accuracy group and each of the two other groups. The percent change from the low in-sample accuracy group to each of the other two groups is also displayed. A breakdown of the full sample is provided in the footer of each figure, with a specific count of the number of forecasts from each group provided below each label in parentheses.

Figure A19: Distribution of Confidence and Responses Per Forecaster

(a) Distribution of Confidence Per Unique Response



(b) Distribution of Unique Responses Per Forecaster



Notes: Appendix Figure A19a displays the distribution of forecaster-reported confidence levels across all $N = 50$ projects that utilize the SSPP co nfidence question. Appendix Figure A19b displays the distribution of the number of unique responses per forecaster. Vertical lines indicate the mean, median, conditional and unconditional on having a more than one response.

Table A1: Summary Statistics for the $N = 100$ Projects on the SSPP (2020-2024)

	Project		Subject		Dates	Counts			Forecaster Types				Analysis Samples		
i	Title	Coauthors	Field	RCT	Open	N_q	N_r	N_f	$\%_{panel}$	$\%_{prof}$	$\%_{phd}$	$\%_{non}$	$\%_{s1}$	$\%_{s2}$	$\%_{s3}$
1	Irregular Migration	Bah et al.	Eco	✓	07-20	27	108	2,799	0.00	0.15	0.44	0.40	0.33	0.33	1.00
2	Policy-Makers	Coville, Vivalt	Eco		07-20	16	229	1,796	0.00	0.11	0.38	0.50	0.00	0.00	1.00
3	Depression	Bhat et al.	Eco	✓	07-20	7	209	1,372	0.00	0.14	0.53	0.32	0.86	0.86	1.00
4	Supreme Court	Levy Paluck et al.	Psy		07-20	16	106	1,620	0.00	0.08	0.49	0.44	0.00	0.00	0.00
5	Race	Advani et al.	Eco		08-20	2	325	645	0.00	0.21	0.27	0.51	0.00	0.00	1.00
6	Gender Credit	Almås et al.	Eco		08-20	7	88	571	0.00	0.19	0.58	0.23	0.00	0.00	0.00
7	Shelter Refugee	Leone et al.	Eco	✓	08-20	14	151	977	0.00	0.18	0.45	0.38	0.86	0.86	1.00
8	Social Cues	Munger et al.	Pol		09-20	8	40	293	0.00	0.06	0.66	0.29	0.75	0.75	0.75
9	Public Spending	Celhay et al.	Eco	✓	11-20	18	48	846	0.00	0.02	0.78	0.20	0.22	0.00	0.00
10	US UCT	Hauser et al.	Eco, Psy	✓	11-20	24	23	552	0.00	0.05	0.59	0.37	1.00	1.00	1.00
11	Binary Choices	Chapman et al.	Eco		11-20	3	110	326	0.00	0.25	0.50	0.25	0.00	0.00	0.67
12	COVID-19 Mozambique	Allen IV et al.	Eco	✓	11-20	5	74	363	0.00	0.20	0.50	0.30	1.00	1.00	1.00
13	Hedonic Adaptation	Mrkva	Psy		12-20	1	13	13	0.00	0.00	0.50	0.50	0.00	0.00	0.00
$\sum_i = 13$	Total - 2020	-	-	-	-	148	1,524	12,173	0.00	0.14	0.48	0.39	0.45 [0.54]	0.42 [0.46]	0.70 [0.69]
14	German Farmers	Rommel et al.	Eco		01-21	5	9	45	0.00	0.00	0.78	0.22	0.80	0.80	1.00
15	Business Practices	Gertler et al.	Eco	✓	01-21	5	63	309	0.00	0.20	0.54	0.25	1.00	1.00	1.00
16	HIV Mozambique	Mahumane et al.	Eco	✓	01-21	5	46	230	0.00	0.19	0.44	0.37	1.00	0.00	0.00
17	Comparing Overpreci-sion	Moore, Hale	Psy		02-21	1	20	20	0.00	0.20	0.66	0.14	0.00	0.00	1.00
18	Nudge Tournament	Dimant et al.	Psy		02-21	72	5	256	0.00	0.00	0.50	0.50	0.78	0.78	1.00
19	Prosocial Behavior	Kang et al.	Eco, Psy		04-21	2	39	75	0.00	0.12	0.65	0.24	0.00	0.00	0.00

Table A1: Summary Statistics for the $N = 100$ Projects on the SSPP (2020-2024) *(continued)*

	Project		Subject		Dates	Counts			Forecaster Types				Analysis Samples		
i	Title	Coauthors	Field	RCT	Open	N_q	N_r	N_f	$\%_{panel}$	$\%_{prof}$	$\%_{phd}$	$\%_{non}$	$\%_{s1}$	$\%_{s2}$	$\%_{s3}$
20	Crowdsourcing COVID-19	Golden et al.	Pol		04-21	475	95	1,229	0.00	0.41	0.41	0.17	0.00	0.00	0.00
21	CBT Liberia	Blattman et al.	Eco, Psy	✓	05-21	6	66	390	0.00	0.28	0.26	0.46	1.00	1.00	1.00
22	Vaccination Uptake	Campos-Mercade et al.	Eco		06-21	10	60	575	0.00	0.28	0.41	0.31	1.00	1.00	1.00
23	Health Workers	Deserranno et al.	Eco	✓	07-21	3	34	102	0.00	0.41	0.41	0.19	1.00	1.00	1.00
24	Support Taxation	Giaccobasso et al.	Eco		07-21	8	71	540	0.00	0.08	0.20	0.72	0.50	0.50	0.50
25	Customer Discrimination	Chan	Eco		07-21	4	13	52	0.00	0.00	0.75	0.25	1.00	1.00	1.00
26	Political Pressure	Acemoglu et al.	Eco		07-21	2	24	47	0.00	0.19	0.57	0.25	0.00	0.00	0.00
27	Affective Polarization	Bauer et al.	Eco		08-21	8	37	274	0.00	0.06	0.61	0.33	0.00	0.00	0.62
28	Sexual Harassment	Dahl, Knepper	Eco		09-21	3	19	53	0.00	0.07	0.47	0.47	0.00	0.00	1.00
29	Confidence Institutions	Enke et al.	Eco		10-21	-	139	139	-	-	-	-	-	-	-
30	Systemic Discrimination	Kline et al.	Eco		11-21	-	811	811	-	-	-	-	-	-	-
$\sum_i = 17$	Total - 2021	-	-	-	-	609	1,551	5,147	0.00	0.22	0.44	0.35	0.16 [0.53]	0.15 [0.47]	0.19 [0.65]
31	Unclaimed Funds	Vivalt	Eco	✓	02-22	2	88	176	0.00	0.25	0.40	0.34	0.50	0.50	1.00
32	Prolific Workers	Gandhi et al.	Psy		02-22	-	444	444	-	-	-	-	-	-	-
33	Human Capital	Deshpande, Dizon-Ross	Eco	✓	02-22	1	71	71	0.00	0.98	0.00	0.01	1.00	1.00	1.00
34	LR France	Bernard et al.	Eco	✓	03-22	30	63	970	0.00	0.18	0.44	0.38	0.80	0.80	0.80
35	LR Kenya	Bernard et al.	Eco	✓	03-22	54	26	652	0.00	0.11	0.39	0.50	0.89	0.89	0.89
36	LR Uganda	Bernard et al.	Eco	✓	03-22	44	45	1,014	0.00	0.17	0.45	0.37	0.82	0.82	0.82
37	LR Liberia	Bernard	Eco	✓	03-22	20	4	80	0.00	0.00	0.50	0.50	0.80	0.80	0.90
38	Adoption Nudges	DellaVigna et al.	Eco		03-22	35	147	2,346	0.00	0.16	0.13	0.71	0.00	0.00	1.00

Table A1: Summary Statistics for the $N = 100$ Projects on the SSPP (2020-2024) *(continued)*

	Project		Subject		Dates	Counts			Forecaster Types				Analysis Samples		
i	Title	Coauthors	Field	RCT	Open	N_q	N_r	N_f	$\%_{panel}$	$\%_{prof}$	$\%_{phd}$	$\%_{non}$	$\%_{s1}$	$\%_{s2}$	$\%_{s3}$
39	Causal UCT	Bartik et al.	Eco, Pol	✓	04-22	24	218	2,750	0.37	0.47	0.22	0.30	0.42	0.42	1.00
40	Salary Benchmarking	Cullen et al.	Eco		05-22	5	15	59	0.00	0.37	0.09	0.54	0.00	0.00	0.20
41	Exam Scores	Liu, Wang	Eco		05-22	2	6	12	0.00	0.00	0.33	0.66	0.00	0.00	1.00
42	Role Models	Asanov et al.	Eco	✓	06-22	18	54	945	0.00	0.37	0.26	0.36	0.44	0.00	0.00
43	Feedback University	Kinne, Rehwinkel	Eco		06-22	1	3	3	0.00	0.00	1.00	0.00	1.00	1.00	1.00
44	Panel Attrition	Dillon et al.	Eco	✓	07-22	8	53	406	0.00	0.14	0.51	0.35	0.25	0.25	1.00
45	Representation Learning	Li, Roy	Eco		07-22	2	47	94	0.00	0.20	0.40	0.40	0.00	0.00	0.00
46	Research Ethics	Velez, Da In Lee	Pol		10-22	16	63	932	0.00	0.68	0.18	0.14	0.00	0.00	1.00
47	Personal Initiative	Thomas et al.	Psy, Eco	✓	10-22	6	28	168	0.00	0.11	0.37	0.52	1.00	1.00	1.00
48	Carbon Pricing	Innocenti, Fang	Eco		10-22	9	26	222	0.00	0.05	0.62	0.33	1.00	1.00	1.00
49	Stock Price	Ba, Whitefield	Pol, Eco, Soc		10-22	6	23	133	0.00	0.11	0.33	0.56	0.00	0.00	1.00
50	Craigslist Negotiation	Kirgios	Psy		10-22	15	47	660	0.00	0.14	0.39	0.46	1.00	1.00	1.00
51	Non-binary Identity	Kirgios	Psy		11-22	3	27	81	0.00	0.04	0.50	0.46	1.00	1.00	1.00
52	Debt Moratorium	Fiorin et al.	Eco	✓	12-22	3	32	96	0.00	0.20	0.30	0.50	1.00	1.00	1.00
$\sum_i = 22$	Total - 2022	-	-	-	-	304	1,530	12,314	0.05	0.35	0.30	0.35	0.60 [0.68]	0.58 [0.64]	0.85 [0.86]
53	LR Spain	Bernard et al.	Eco	✓	02-23	9	49	441	0.00	0.30	0.33	0.37	0.78	0.78	0.78
54	LR Togo	Bernard et al.	Eco	✓	02-23	20	33	532	0.00	0.13	0.39	0.48	0.80	0.80	0.80
55	LR Afghanistan	Bernard et al.	Eco	✓	02-23	15	39	585	0.00	0.22	0.43	0.36	0.47	0.47	0.67
56	Belief Updating	Aina et al.	Eco		02-23	1	36	36	0.00	0.49	0.32	0.20	0.00	0.00	0.00
57	LR Sierra Leone	Bernard, Karing	Eco	✓	03-23	15	30	412	0.00	0.21	0.37	0.41	0.80	0.80	0.80
58	Migration Mentoring	Franzl et al.	Eco	✓	03-23	2	73	142	0.00	0.15	0.33	0.52	0.00	0.00	0.00

Table A1: Summary Statistics for the $N = 100$ Projects on the SSPP (2020-2024) *(continued)*

	Project		Subject		Dates	Counts			Forecaster Types				Analysis Samples		
i	Title	Coauthors	Field	RCT	Open	N_q	N_r	N_f	$\%_{panel}$	$\%_{prof}$	$\%_{phd}$	$\%_{non}$	$\%_{s1}$	$\%_{s2}$	$\%_{s3}$
59	Depth Reasoning	Salant, Spenkuch	Eco		03-23	6	43	245	0.00	0.08	0.54	0.38	0.00	0.00	1.00
60	Recession Mortality	Finkelstein et al.	Eco		03-23	5	6	23	0.00	0.00	0.50	0.50	0.20	0.20	0.20
61	Sorting	Bryan et al.	Eco		04-23	10	10	100	0.00	0.20	0.30	0.50	0.00	0.00	0.40
62	Managers' Beliefs	Chaurey et al.	Eco	✓	05-23	9	45	390	0.00	0.24	0.36	0.40	0.00	0.00	0.00
63	Developing Agriculture	Burlig et al.	Eco	✓	06-23	5	37	169	0.00	0.27	0.24	0.49	1.00	1.00	1.00
64	Subscription Nudges	Mrkva et al.	Eco, Psy		06-23	27	45	732	0.00	0.15	0.41	0.44	0.00	0.00	0.81
65	Job Loss	Abebe et al.	Eco	✓	07-23	15	73	1,049	0.55	0.28	0.37	0.35	0.67	0.67	1.00
66	Satisfying Preferences	Arrieta, Bolte	Eco		07-23	5	60	300	0.42	0.17	0.51	0.33	0.00	0.00	1.00
67	Sustainable Transport	Andor et al.	Eco	✓	08-23	8	142	786	0.00	0.23	0.31	0.46	0.00	0.00	0.00
68	SMS-based Training	Zia Mehmood	Eco	✓	09-23	9	74	650	0.51	0.10	0.46	0.43	0.67	0.67	0.89
69	Correlating Overpreci-sion	Moore et al.	Psy		10-23	6	34	160	0.76	0.00	0.00	1.00	0.00	0.00	1.00
70	Default Interest	Castellanos et al.	Eco	✓	10-23	1	73	73	0.77	0.23	0.37	0.40	0.00	0.00	1.00
71	Girl Empow-erment	Yu et al.	Pol, Eco	✓	11-23	12	93	1,110	0.78	0.18	0.42	0.41	0.08	0.08	1.00
72	Salary Equalization	George, Mattsson	Eco		11-23	6	88	518	0.86	0.20	0.41	0.39	0.50	0.50	0.50
73	Clean Air	Mattsson et al.	Eco		11-23	2	86	172	0.83	0.18	0.43	0.39	0.00	0.00	0.00
74	Job Market	Chen et al.	Eco	✓	12-23	3	72	216	0.89	0.21	0.43	0.36	1.00	1.00	1.00
$\sum_i = 22$	Total - 2023	-	-	-	-	191	1,241	8,841	0.38	0.20	0.39	0.41	0.37 [0.50]	0.37 [0.50]	0.71 [0.77]
75	Self-Persuasion	Zhang, Rand	Psy		02-24	8	64	505	0.88	0.19	0.48	0.34	0.50	0.50	1.00
76	Policy Adoption	Rey-Biel et al.	Eco, Pol	✓	03-24	3	80	240	0.72	0.21	0.39	0.40	0.00	0.00	1.00

Table A1: Summary Statistics for the $N = 100$ Projects on the SSPP (2020-2024) *(continued)*

	Project		Subject		Dates	Counts			Forecaster Types				Analysis Samples		
i	Title	Coauthors	Field	RCT	Open	N_q	N_r	N_f	$\%_{panel}$	$\%_{prof}$	$\%_{phd}$	$\%_{non}$	$\%_{s1}$	$\%_{s2}$	$\%_{s3}$
77	Equivalence Testing	Fitzgerald	Eco		03-24	3	68	182	0.84	0.22	0.42	0.35	0.00	0.00	0.67
78	Endowment Effect	Camerer et al.	Eco, Psy		03-24	10	132	1,308	0.48	0.35	0.32	0.32	0.00	0.00	1.00
79	Noncompetes	Starr et al.	Eco		04-24	12	75	875	0.81	0.16	0.42	0.41	0.00	0.00	1.00
80	Ghana Tuberculosis	Duch	Eco	✓	04-24	4	57	228	0.86	0.18	0.44	0.39	0.50	0.50	0.75
81	Returns Information	Allen IV et al.	Eco	✓	04-24	1	99	99	0.61	0.17	0.41	0.42	0.00	0.00	0.00
82	Psychosocial Training	Beltramo et al.	Eco	✓	06-24	8	92	736	0.67	0.18	0.41	0.40	1.00	0.00	0.00
83	Women Political	Young	Pol	✓	07-24	10	19	174	0.68	0.21	0.53	0.27	0.00	0.00	0.00
84	Biodiversity Conservation	Anna et al.	Eco		07-24	14	68	945	0.79	0.16	0.37	0.47	0.00	0.00	0.00
85	Retail Nudges	Mrkva	Psy, Oth		07-24	2	84	168	0.69	0.21	0.46	0.32	0.00	0.00	0.00
86	Air Quality	Gazze et al.	Eco	✓	07-24	1	68	68	0.81	0.19	0.43	0.38	0.00	0.00	0.00
87	Financial Education	Kaiser et al.	Eco	✓	07-24	36	64	2,123	0.81	0.16	0.42	0.43	0.33	0.33	1.00
88	Water Nudge	Espinosa-Goded et al.	Eco	✓	07-24	3	100	199	0.60	0.13	0.40	0.46	1.00	1.00	1.00
89	Teacher Training	Brunetti et al.	Eco	✓	07-24	4	69	274	0.81	0.20	0.38	0.43	1.00	0.00	0.00
90	Wage Inequality	Page et al.	Eco		07-24	45	78	2,232	0.76	0.15	0.41	0.44	0.20	0.00	0.00
91	Female Entrepreneurship	Asiedu et al.	Eco	✓	08-24	10	61	610	0.87	0.16	0.46	0.37	1.00	1.00	1.00
92	Psychological Support	Park et al.	Eco	✓	09-24	1	69	69	0.70	0.19	0.41	0.39	1.00	0.00	0.00
93	Mass Reproducibility	Brodeur et al.	Eco, Oth		09-24	5	261	1,305	0.16	0.21	0.27	0.52	0.00	0.00	0.00
94	CCC UCT	Abdul-Razzak et al.	Eco	✓	09-24	4	78	312	0.60	0.23	0.33	0.44	0.00	0.00	0.00
95	Publication Patterns	Hoces de la Guardia et al.	Eco		09-24	2	71	142	0.55	0.20	0.32	0.48	0.00	0.00	0.00
96	Peer Punishment	Alsobay et al.	Eco, Psy, Soc		09-24	20	52	909	0.69	0.26	0.37	0.37	0.00	0.00	1.00

Table A1: Summary Statistics for the $N = 100$ Projects on the SSPP (2020-2024) *(continued)*

	Project		Subject		Dates	Counts			Forecaster Types				Analysis Samples		
i	Title	Coauthors	Field	RCT	Open	N_q	N_r	N_f	$\%_{panel}$	$\%_{prof}$	$\%_{phd}$	$\%_{non}$	$\%_{s1}$	$\%_{s2}$	$\%_{s3}$
97	Evidence Transmission	Shaukat et al.	Eco		11-24	9	52	462	0.69	0.16	0.42	0.42	0.00	0.00	0.00
98	Commitment Demand	Suchy, Toussaert	Eco		11-24	3	75	225	0.52	0.21	0.27	0.51	0.00	0.00	0.00
99	Fiscal Events	Bergeron et al.	Eco		12-24	1	52	52	0.73	0.16	0.45	0.39	0.00	0.00	0.00
100	Digital Empowerment	Joshi et al.	Eco	✓	12-24	8	47	376	0.79	0.17	0.43	0.40	1.00	0.00	0.00
$\sum_i = 26$	Total - 2024	-	-	-	-	227	2,035	14,818	0.63	0.20	0.38	0.42	0.27 [0.38]	0.14 [0.19]	0.47 [0.38]
-	Total	-	-	-	-	1,479	7,881	53,293	0.23	0.21	0.39	0.39	0.32 [0.52]	0.29 [0.44]	0.49 [0.66]

Notes: Table A1 lists the 100 projects posted on the SSPP in the 2020-2024 period, sorted in order of date published. Each project is identified by a shortened title and the list of coauthors associated with the project, with further information on the main fields of the project, whether it is an RCT, and the month-year in which the project was posted. Counts are provided of the number of unique key questions associated with the project (N_q), the number of unique responses to the project (N_r), both complete and incomplete, and the number of unique forecasts (N_f) to the project's key questions, roughly equal to the product of the previous two numbers. Next, the share of responses from SSPP panelists ($\%_{panel}$), and, of the individuals with a recorded academic status, the share of responses from faculty members ($\%_{prof}$), PhD students ($\%_{phd}$), and non-academics ($\%_{non}$). Finally, the share of key questions in each of the 3 main analysis samples outlined in Table 2, specifically: normed treatment effects ($\%_{s1}$; column 4 of Table 2), normed treatment effects with results ($\%_{s2}$; column 5 of Table 2), and all key questions with results ($\%_{s3}$; column 2 of Table 2). Aggregates by each year, and across the entire sample are also provided, with the numbers in brackets below each of the analysis sample columns providing analogous percentages for the share of projects with at least one key question belonging in the respective analysis sample (e.g. 66% overall for the sample of projects with results).

Table A2: Maximum–likelihood estimates $\hat{\Theta}$

	Confidence Level			
	Low	Median	High	Missing
$\hat{o}$ (Over-Prediction)				
Estimate	0.076	0.093	0.137	0.109
Std. Error	0.005	0.004	0.005	0.003
$\hat{\sigma}$ (Noisiness of Signal)				
Estimate	0.416	0.381	0.473	0.351
Std. Error	0.004	0.003	0.004	0.002

Table A3: Simulated Results

Confidence	Empirical	Simulated	Sim., Holding $o = o^{(l)}$	Sim., Holding $\sigma = \sigma^{(l)}$
Low	-0.361	-0.456	-0.456	-0.456
Median	-0.365	-0.436	-0.432	-0.458
High	-0.419	-0.502	-0.491	-0.468
Missing	-0.443	-0.417	-0.410	-0.458

Notes: All estimates are for mean absolute forecast accuracy. Column 1 reports the results in the data, column 2 simulates our model with various draws of $\epsilon_{i,k} \sim \mathcal{N}(0,1)$, columns 3 and 4 also simulate our model but use the low confidence estimates for o and σ respectively.

Table A4: Maximizing the Predicted Accuracy of Forecasts Weighted by Number of Responses

Row	Forecasting Method	MSE	Percentage Difference
1	Individual forecast	0.1218	-
2	Average forecast (wisdom of the crowd)	0.0502	58.76%
3	Demeaned average forecast (OOS) Avg. OOS constant: 0.088	0.0433	13.72%
4	Constant forecast (average of other studies)	0.0382	11.83%
5	Shrunk average forecast Avg. OOS β: 0.404, intercept: 0.028	0.0330	13.53%
6	Shrunk average forecast (N=10) Avg. OOS β: 0.317, intercept: 0.034	0.0328	0.88%
7	Average forecast (wisdom of the crowd)	0.0502	-
8	Average of top 50% forecasters (OOS-ranked)	0.0448	10.85%
9	Optimally weighted forecast (avg OOS α=50, top 50%) (at grid max)	0.0449	-0.16%
10	Shrunk average forecast Avg. OOS β: 0.404, intercept: 0.028	0.0330	-
11	Shrunk optimally weighted forecast (top 50%) Optimal weight α: 24.68, Avg. OOS β: 0.406, intercept: 0.029	0.0314	4.89%
12	Shrunk optimally weighted forecast (top/bottom 25%) Optimal weights OOS α: 15.02, γ: 0.40, Avg. OOS β: 0.341, intercept: 0.040	0.0306	2.63%

Notes: Table 5 presents the out-of-sample prediction accuracy, measured by Mean Squared Error (MSE) weighted by number of forecasts on each question, of different forecasts. Row 1 evaluates the accuracy using one (randomly drawn) forecast. Row 2 calculates the accuracy of the average forecast per key question (the wisdom of the crowd). Row 3 adjusts this average forecast by subtracting the mean forecast bias observed across all other out-of-sample studies. Row 4 evaluates a constant forecast equal to the simple average of the realized results from all other out-of-sample studies. Row 5 applies an out-of-sample shrinkage adjustment, transforming the average forecast using coefficients estimated from a regression of the treatment effect on the average forecast for all other studies. Row 6 applies this same shrinkage procedure but simulates the average forecast using bootstrapped sub-samples of N=10 responses. Row 7 repeats the simple average forecast from Row 2. Row 8 calculates the average forecast using only respondents ranked in the top 50% for predicted out-of-sample accuracy. The top 50% is estimated via respondent characteristics from all other out-of-sample studies with a regression as in Table 3, Column 3. Row 9 computes an optimally weighted average where the top 50% of forecasters receive an out-of-sample optimized weight (α), while the rest receive a weight of 1. Row 10 repeats the baseline shrunk average forecast from Row 5. Row 11 applies the out-of-sample shrinkage regression adjustment to the optimally weighted forecast from Row 9. Row 12 applies the out-of-sample shrinkage adjustment to a weighted average that assigns independent optimal weights to both the top 25% (α) and bottom 25% (γ) of forecasters.