Predicting Experimental Results: Who Knows What?*

Stefano DellaVigna
UC Berkeley and NBER

Devin Pope
U Chicago and NBER

This version: August 16, 2016

Abstract

Academic experts frequently recommend policies and treatments. But how well do they anticipate the impact of different treatments? And how do their predictions compare to the predictions of non-experts? We analyze how 208 experts forecast the results of 15 treatments involving monetary and non-monetary motivators in a real-effort task. We compare these forecasts to those made by PhD students and non-experts: undergraduates, MBAs, and an online sample. We document seven main results. First, the average forecast of experts predicts quite well the experimental results. Second, there is a strong wisdom-of-crowds effect: the average forecast outperforms 96 percent of individual forecasts. Third, correlates of expertise—citations, academic rank, field, and contextual experience–do not improve forecasting accuracy. Fourth, experts as a group do better than non-experts, but not if accuracy is defined as rank ordering treatments. Fifth, measures of effort, confidence, and revealed ability are predictive of forecast accuracy to some extent, especially for non-experts. Sixth, using these measures we identify 'superforecasters' among the non-experts who outperform the experts out of sample. Seventh, we document that these results on forecasting accuracy surprise the forecasters themselves. We present a simple model that organizes several of these results and we stress the implications for the collection of forecasts of future experimental results.

*We thank Dan Benjamin, Jon de Quidt, Emir Kamenica, David Laibson, Barbara Mellers, Katie Milkman, Sendhil Mullainathan, Uri Simonsohn, Erik Snowberg, Richard Thaler, Kevin Volpp, and especially Ned Augenblick, Don Moore, Philipp Strack, and Dmitry Taubinsky for their comments and suggestions. We are also grateful to the audiences at Bonn University, Frankfurt University, the London School of Economics, the Max Planck Institute in Bonn, the Wharton School, at the University of California, Berkeley, at the 2016 JDM Preconference, the 2015 Munich Behavioral Economics Conference and at the 2016 EWEBE conference for useful comments. We thank Alden Cheng, Felix Chopra, Thomas Graeber, Johannes Hermle, Jana Hofmeier, Lukas Kiessling, Tobias Raabe, Michael Sheldon, Patricia Sun, and Brian Wheaton for excellent research assistance. We are also very appreciate of the time contributed by all the experts, as well as the PhD students, undergraduate students, MBA students, and MTurk workers who participated. We are very grateful for support from the Alfred P. Sloan Foundation (award FP061020).

1 Introduction

An economist meets a policy-maker eager to increase take-up of a program. The economist's recommendation? Change the wording of a letter. Later on, the economist advises an MBA student to emphasize a different reference price in the pricing scheme of the MBA student's company. At the end of the day, during office hours, the academic counsels a student against running a particular arm of an RCT: 'the result will be a null effect.'

Interactions such as these are regular occurrences, especially as economists are increasingly tapped for advice. A common thread runs through the three interactions: the expert advice relies on the forecast of a future research finding. In the policy-maker interaction, the expert is guessing, based on past experience, that the suggested wording will increase take-up more than other equally-expensive interventions. A similar guessing process underlies the other advice.

These interactions lead to an obvious question: How well can experts predict experimental results? The answer to this question is critical to navigate the trade-off between following expert advice or choosing broad experimentation which can be time-consuming and costly.

This naturally leads to a second group of questions: Which forms of expertise lead to more accurate forecasts? Is it having deep experience and recognition in a field (*vertical* expertise)? Or having worked on a particular topic (*horizontal* expertise)? Or is it knowing the specific setting (*contextual* expertise)? Do experts outperform non-experts? Does the answer depend on the definition of accuracy? And is it enough to poll one or two experts, or should one poll a group, even though it may be time consuming?

These questions do not have comprehensive answers, since forecasts of experimental results are not typically recorded. In the absence of this evidence, we may depend too much on informal forecasts, relying on the wrong experts, or conversely under-utilizing the experts.

In this paper, we use data from a large experiment, and associated expert forecasts, designed to provide evidence on the questions above in one particular setting. We compare the relative effectiveness of 18 treatments in a real-effort online experiment with nearly 10,000 subjects, analyzed in detail in DellaVigna and Pope (2016). The large sample size of about 550 subjects per treatment ensures precision in the estimates of the treatment effects.

As part of the design, we survey a group of 314 experts, including behavioral economists, standard economists, and psychologists. The experts are identified out of a group of participants to behavioral conferences. We provide the experts with the results of three benchmark treatments with piece-rate variation to help them calibrate how responsive participant effort was to different levels of motivation in this task. We then ask them to forecast the effort participants exerted in the other 15 conditions including monetary incentives and non-monetary behavioral motivators, such as peer comparisons, reference dependence, and social preferences. The treatments only differ in essentially one paragraph in the instructions, facilitating the comparison across treatments and thus the expert forecasts. Of the 314 experts contacted, 208

provided a complete set of forecasts. The broad selection of experts and the high response rate enables us to study how expertise influences forecasts of experimental results.

In addition to the academic experts, we also survey 147 PhD students in economics to study vertical expertise further. We also collect forecasts made by 158 undergraduate students, 160 MBA students, and 762 online workers in the sample in which we ran the experiment.

We document seven main results. First, the *average* forecast among the 208 academic experts is remarkably informative about the actual treatment effects. Across the 15 treatments, the correlation of the average forecast with the actual outcome is 0.77, and the average absolute deviation between the average forecast and the outcome is just 5 percent of the average score.

A policy-maker, a firm, or an advisee, though, will not typically have the benefit of taking the average for a large set of expert forecasts, and will typically have the opinion of one expert, or a few experts. How do individual experts do?

We document a large difference in accuracy between the average forecast and individual forecasts, our second result. The average absolute error in individual forecasts is 8 percent of the score, compared to 5 percent for the average forecast. Indeed, the average forecast outperforms 96 percent of individual forecasts. The comparison is equally striking using other measures of forecast accuracy like sum of squared errors and correlation.

What explains this large 'wisdom-of-crowds' effect? We show that the difference between average and individual accuracy does not hold in each treatment: in some treatments the majority of experts outperform the average expert. Still, there is enough idiosyncratic noise in the forecasts that, averaging over enough treatments, the mean outperforms nearly every individual expert. We also show that taking the average forecast of just 5 experts leads to a large improvement in accuracy over individual forecasts.

Thus, contacting multiple experts has first-order benefits for forecasting accuracy. Still, so far we have treated experts as interchangeable. Asking the 'right' expert may erase most of the gains from averaging. We thus consider the impact of *vertical* expertise—academic rank and citations—, *horizontal* expertise—field of expertise and having written a paper on the topic of the treatment—, and *contextual* expertise—knowledge of the experimental setting.

Our third finding is that none of these measures of expertise improve forecasting accuracy. Full professors are, if anything, less accurate than assistant professors. Similarly, having more Google Scholar citations is associated with lower accuracy. Thus, *vertical* expertise does not appear predictive of accuracy. Our measure of *horizontal* expertise, which involves a detailed coding of whether a given expert has worked on a particular topic, is orthogonal to accuracy, controlling for expert and treatment fixed effects. We also find no effect of expertise in different sub-fields, such as psychology, behavioral economics, or applied microeconomics. Finally, experience with the online sample (*contextual* expertise) does not increase accuracy.

The findings for the sample of PhD students are similar. Consistent with the null effect of vertical expertise, PhD students are at least as good as the academic experts. Furthermore,

confirming the null effect of horizontal expertise, specialization in behavioral economics does not improve the accuracy of the PhD students.

Thus, various measures of expertise do not increase accuracy. Still, it is possible that academics and academics in training (the PhD students) share an understanding of incentives and behavioral forces which distinguish them from the non-experts. We thus consider forecasts by undergraduate students, MBA students, and an online sample. These forecasters have not received much training in formal economics, though some of them arguably have more experience with incentives at work (the MBAs) and with the context (the online sample).

Do non-experts make worse forecasts? The answer, our fourth finding, depends on the definition of *worse*. By the measure of accuracy used so far—absolute error in forecasts—these groups indeed do worse. The undergraduate students are somewhat less accurate, MBA students are significantly less accurate, and online forecasters in the MTurk sample do much worse. These results are similar for the squared error measure of accuracy. When making forecasts about *magnitudes* of the experimental findings, yes, non-experts do worse than experts.

Yet, while the above measures of accuracy were the main ones we envisioned[1], they are not always the relevant ones. In our motivating examples, the policy-maker, the businessperson, and the advisee may be looking for a recommendation of the most effective treatment, or for ways to weed out the least effective ones. From this perspective, it is not as important to get the *levels* right in the forecasts, as it is to get the *order* right. We thus revisit the results using the rank-order correlation between the forecasts and the results as the measure of accuracy.[2]

Rank-order correlation does not reverse the findings on vertical, horizontal, or contextual expertise: the three forms of expertise do not help academics rank treatments better. However, this metric drastically changes the comparison between experts and non-experts: undergraduates, MBAs, and even MTurk workers rank treatments as well as the experts. Across these samples, the average individual rank-order correlation with the realized effort is about 0.4 and the wisdom-of-crowds rank-order correlation is about 0.8. In fact, the wisdom-of-crowds rank-order correlation by the online sample is a stunning 0.95 (compared to 0.83 for the experts).

How is this discrepancy possible? We show that the non-experts, and especially the online sample, are much more likely to be off in the guess of the average effort across the 15 forecasts. This offset in levels impacts the absolute error, but not necessarily the rank order. This result is consistent with psychological evidence suggesting that people struggle with absolute judgments, but are better at making relative judgments (Laming, 1984; Kahneman, Schkade, and Sunstein, 1998).

So far, we found that expertise does not help much with forecasts. The fine-grained *ex ante* measures of expertise do not increase forecasting accuracy, and experts as a group differ from

[1] In our pre-registration, we mention three measures of accuracy: absolute error, squared error, and number of correct answers within 100 points of the truth (more on this below).

[2] We deduce the ranking of treatments from the forecasts in levels.

non-experts only if the accuracy is about the levels, as opposed to the rank order, of treatments. If expertise does not help much, are there other ways, then, to discriminate among forecasters for accuracy? We consider measures of effort, confidence, and revealed ability.

Our fifth result is that such measures can be predictive of expert accuracy, but with important caveats. The predictability mostly holds among non-experts and, while generally strong for the absolute error measure, it is weak for the ordinal rank measure.

We measure effort in forecasting with the time taken for survey completion and with click-throughs to the trial task and the instructions. The evidence is mixed. For the online sample, longer time taken improves accuracy by the absolute error measure. There is much less evidence for the other samples, and the relationship flattens or flips sign using the rank-order correlation. Clicking on the trial or instruction does not have a discernible impact.

We also measure confidence: each forecaster indicated the number of forecasts which they expect to get right within 100 points. This measure is predictive of accuracy in levels among PhDs, MBAs, and online workers. Respondents, thus, are aware of their own accuracy, to some extent. Confidence is instead essentially uncorrelated with the rank-order measure of accuracy, perhaps because we elicited confidence using a cardinal, not ordinal, measure of accuracy.

We then construct a measure of revealed accuracy that captures both effort and ability. We test if accuracy in the forecast of a simple incentive-based treatment predicts accuracy in the other treatments. This variable is remarkably predictive: a 100-point increase in accuracy in the incentive treatment increases accuracy in other treatments by on average 30 points for the non-expert samples and 9 points for the expert sample. This measure remains predictive, even though less strongly so, of accuracy as measured with rank-order correlation. The measure of revealed forecasting ability predicts accuracy also when constructed using other treatments, suggesting that there is nothing special about the incentive treatment.

Thus, while ex ante proxies of expertise are not helpful in our setting, other measures—effort, confidence, and especially revealed forecasting ability—are generally predictive of accuracy measured with absolute error. Can these measures then help identify 'superforecasters' (Tetlock and Gardner, 2015) among the non-experts? We use simple linear regressions with an K-fold method to obtain out-of-sample predictions, and focus on absolute error since the non-experts already equal the experts according to the rank-order measure.

Our sixth result is that it is indeed possible to identify 'superforecasters'. The top 20 percent of undergraduates and PhD students identified with this procedure outperform at the individual level the sample of experts by 15 percent. The outperformance is even more striking when using the wisdom-of-crowds measure. We also identify 'superforecasters' within the MTurk sample who parallel the accuracy of academic experts. Among the academic experts, instead, there is a more limited improvement in accuracy from this procedure.

Our seventh and final result addresses a meta-question: Did we know all this already? In the spirit of the forecasting idea, we asked the experts to predict the accuracy of different

groups of forecasters. The expert beliefs in this regard are systematically off target. The experts expect high-citation experts to be significantly more accurate where, if anything, the opposite is true. They also expect a difference by the field of the forecaster and lower accuracy for PhD students, counterfactually.

These results, while just a first step, draw out implications for increasing accuracy of forecasts of research findings. Clearly, one ought to elicit forecasts from multiple people. Further, experts may not necessarily offer a more precise forecast than a well-motivated audience, and the latter sample is easier to reach. One can then screen the non-experts based on measures of effort, confidence, and accuracy on a trial question. We conjecture that more opportunities to make forecasts, and see the feedback, could lead to significant improvements in forecasting ability, and to beliefs about expertise that are more aligned with actual accuracy.

Can we make sense of our key findings with a simple model? We assume that forecasters observe a noisy signal of the truth, with some forecasters receiving more precise signals than others. The heterogeneity in informativeness is motivated by the result that forecasters who do better in one treatment also do better in other treatments.

We calibrate the model based on five moments: three variances, mean accuracy, and the cross-treatment correlation in accuracy. Our calibration implies that the non-experts on average have higher idiosyncratic noise in their signals, and also more heterogeneity, compared to experts. Due to the higher heterogeneity, some non-experts receive more precise signals than the experts (the 'super-forecasters'). We can also approximately match the differences between experts and non-experts in absolute error versus rank correlation.

We explore complementary findings in a companion paper (DellaVigna and Pope, 2016), focusing on what motivates effort and providing evidence on some leading models in behavioral economics. For each treatment, we analyze the effort choice of the subjects and the average forecast of the academic experts. The companion paper does not consider measures of accuracy of forecasts, differences in expertise, forecasts by non-experts, or beliefs about expertise.

This paper is related to several literatures that span different academic disciplines, including psychology, in addition to economics. The Good Judgment Project elicits forecasts by experts on national security topics (Tetlock and Gardner, 2015). We find in our setting significant parallels to their findings, including the fact that, while it is hard to identify good forecasters based on ex ante characteristics, it is possible to do so using measures of accuracy on a subsample of forecasts (Mellers et al., 2015).

Related to our paper is the work on wisdom of crowds. At least since Galton (1907), social scientists have been interested in cases in which the average of individual forecasts outperforms nearly all of the individual forecasters (e.g. Surowiecki, 2005). We show that the wisdom-of-crowds phenomenon does *not* apply to each treatment: in several of the treatments, the average forecast is outperformed by a majority of the forecasters. It is when considering all treatments jointly that the evidence strongly supports the wisdom of crowds.

Economics also has a rich tradition of studying prediction accuracy, including in macroeconomics and finance (e.g., Cavallo, Cruces, and Perez-Truglia, 2016; Ben-David, Graham, and Harvey, 2013). More closely related is the work on the value of aggregating predictions using predictions markets (Wolfers and Zitzewitz, 2004; Snowberg, Wolfers, and Zitzewitz, 2007).

There has been some work in economics that attempts to elicit opinions from academic experts. For example, the IGM Economic Expert panel has academic experts forecast the impact of policy issues or measures of future variables such as inflation or stock returns. On a smaller scale, several papers have elicited opinions from academics. For example, Coffman and Niehaus (2014) includes a survey of 7 experts on persuasion and Sanders, Mitchell, and Chonaire (2015) ask 25 faculty and students from two universities questions on the results of 15 select experiments run by the UK Nudge Unit. Groh, Krishnan, McKenzie, and Vishwanath (2015) elicit forecasts on the effect of an RCT from audiences of 4 academic presentations. Erev et al. (2010) ran a competition among laboratory experimenters to forecast the result of a pre-designed laboratory experiment using learning models trained on data. These efforts suggest the need for a more systematic collection of expert beliefs about research findings.[3]

We are also related to the literature on transparency in the social sciences (e.g., Simmons, Nelson, and Simonsohn, 2011; Vivalt, 2015) and in particular to recent work on replication in psychology and experimental economics, including the use of prediction markets to capture beliefs about the replicability of experimental findings (Dreber et al., 2015 and Camerer et al., 2016). We emphasize the complementarity, as our study examines differences in the informativeness of forecasts of different experts, as well as non-experts, while the Science Prediction Market examines the accuracy of a prediction market and the average in a survey of experts.

The paper proceeds as follows. After presenting the design in Section 2, in Section 3 we document the accuracy of the experts, as a group and individually. In Section 4 we present evidence on cross-sectional differences in expertise, on non-experts and 'superforecasters', and on beliefs about expertise. In Section 5 we present a simple model and in Section 6 we conclude.

2 Experiment and Survey Design

2.1 Real Effort Experiment

We summarize here the design for the experiment, with additional details in DellaVigna and Pope (2016). We designed a simple real effort task on Amazon Mechanical Turk (MTurk), varying the behavioral motivators across arms. MTurk is an online platform that allows researchers and businesses to post small tasks (referred to as HITs) that require a human to perform. Potential workers browse the postings and choose whether to complete a task for the amount offered. MTurk has become a popular platform to run experiments in market-

[3] Banerjee, Chassang, and Snowberg (2016) provide a framework on related issues of optimal experimentation.

ing and psychology (Paolacci and Chandler, 2014) and is also used increasingly in economics, such as for the study of preferences about redistribution (Kuziemko, Norton, Saez, Stantcheva, 2015). Are the results of studies run on MTurk comparable to the results in more standard laboratory or field settings? The evidence suggests that the findings are indeed qualitatively and quantitatively similar. For example, participants exhibit similar biases and overall results when playing economic games online as they do in a physical laboratory (Horton, Rand, and Zeckhauser, 2011; Amir, Rand, and Gal, 2012; Goodman, Cryder, and Cheema, 2013).

The limited cost per subject and large available population on MTurk allow us to run several treatments, each with a large sample size. Furthermore, the MTurk setting allows for a simple and transparent design: the experts can sample the task and can easily compare the different treatments, since the instructions for the various treatments differ essentially in only one paragraph. The MTurk platform also ensures a speedy data collection effort.

We pre-registered the design of the experiment on the AEA RCT Registry as AEARCTR-0000714 ("*Response of Output to Varying Incentive Structures on Amazon Turk*"). Among the pre-registered details of the experiment, we specified the rule for the sample size and the inclusion in the sample, as we detail in DellaVigna and Pope (2016).

The registration also specifies the sequencing of the experiment and the survey. We ran the experiment before seeking the forecasts in order to provide the results of three benchmark treatments to the forecasters. To ensure that there would be no leak of any results in the intervening period, we ourselves did not have access to the experimental results until after the survey collection. We designed a script that monitored the sample size as well as results in the three benchmark treatments. A research assistant ran this script and sent us daily updates so we could monitor for potential data issues. We accessed the results of the other treatments only at the end of September 2015, after the forecasts by the academic experts were collected.

The task involves alternating presses of 'a' and 'b' on a computer keyboard for 10 minutes, achieving a point for each a-b alternation, a task similar to those used in the literature (Amir and Ariely, 2008; Berger and Pope, 2011). While the task is not meaningful per se, it does have features that parallel clerical jobs: it involves repetition and it gets tiring, thus testing the motivation of the workers. It is also simple to explain to both subjects and experts.

The subjects are recruited on MTurk for a \$1 pay for participating in an '*academic study regarding performance in a simple task.*' Subjects interested in participating sign a consent form, enter their MTurk ID, and answer three demographic questions, at which point they see the instructions for the task: '*On the next page you will play a simple button-pressing task. The object of this task is to alternately press the 'a' and 'b' buttons on your keyboard as quickly as possible for 10 minutes. Every time you successfully press the 'a' and then the 'b' button, you will receive a point. Note that points will only be rewarded when you alternate button pushes: just pressing the 'a' or 'b' button without alternating between the two will not result in points. Buttons must be pressed by hand only (key-bindings or automated button-*

pushing programs/scripts cannot be used) or the task will not be approved. Feel free to score as many points as you can.' The participants then see a different final paragraph (bold and underlined) depending on the condition to which they were randomly assigned. For example, in the benchmark 10-cent treatment, the sentence reads '*As a bonus, you will be paid an extra 10 cents for every 100 points that you score. This bonus will be paid to your account within 24 hours.*' The key content of this paragraph for all 18 treatments is reported in Table 2.[4] Subjects can try the task before moving on to the real task.

As subjects press digits, the page shows a clock with a 10-minute countdown, the current points, and any earnings accumulated (depending on the condition). The final sentence on the page summarizes the condition for earning a bonus (if any) in that particular treatment. Thus, the 18 different treatments differ in only three ways: the main paragraph in the instructions explaining the condition, the one-line reminder on the task screen, and the rate at which earnings (if any) accumulate on the task screen. After the 10 minutes are over, the subjects are presented with the total points and the payout, are thanked for their participation and given a validation code which they use to redeem their earnings.

The experiment ran for three weeks in May 2015. The initial sample consists of 12,838 MTurk workers who started our experimental task. After applying the sample restrictions and dropping a subsample due to a Qualtrics software glitch, the final sample includes 9,861 subjects, about 550 per treatment. As Table 2 in DellaVigna and Pope (2016) shows, the demographics of the recruited MTurk sample matches those of the US population along gender lines, but over-represents high-education groups and younger individuals. This is consistent with previous literature documenting that MTurkers are quite representative of the population of U.S. internet users (Ipeirotis, 2009; Ross et al., 2010; Paolacci et al., 2010) on characteristics such as age, socioeconomic status, and education levels.

2.2 Forecaster Survey

Survey format. We designed the survey of experts to infer as precisely as possible the forecasts of effort in the treatments, while keeping the estimated survey duration to a maximum of 15 minutes. The survey is also pre-registered as AEARCTR-0000731.

The survey, formatted with the online survey platform Qualtrics, consists of two pages. On the first and main page, the experts read a description that introduces the task: "*We ran a large, pre-registered experiment using Amazon's Mechanical Turk (MTurk). [...] The MTurk participants [...] agreed to perform a simple task that takes 10 minutes in return for a fixed participation fee of $1.00.*" The survey then described exactly what MTurkers saw: "*You will play a simple button-pressing task. The object of this task is to alternately press the 'a' and*

[4] For space reasons, in Table 2 we omit the sentence '*The bonus will be paid to your account within 24 hours.*' The sentence does not appear in the time discounting treatments.

'b' buttons on your keyboard as quickly as possible for 10 minutes. Every time you successfully press the 'a' and then the 'b' button, you will receive a point."

Following this introduction, the experts can experience the task by clicking on a link. They can also see the complete screenshots viewed by the MTurk workers with another click. The experts are then informed of a prize that depends on the accuracy of their forecasts. "*As added encouragement, five people who complete this survey will be chosen at random to be paid, and this payment will be based on the accuracy of each of his/her predictions. Specifically, these five individuals will each receive $1,000 - (Mean Squared Error/200), where the mean squared error is the average of the squared differences between his/her answers and the actual scores.*"[5] This reward structure is incentive compatible: participants who aim to minimize the sum of squared errors (and thus maximize their potential reward) will indicate as their forecast the mean expected effort for each treatment. We avoided a tournament payout structure (paying the top 5 performers) which could have introduced risk-taking incentives.

The survey then displays the mean effort in the three benchmark treatments: no piece rate, 1-cent, and 10-cent piece rate (see Appendix Figure 1a). The results are displayed using the same slider scale used for the other 15 treatments, except with a fixed scale. The experts then see a list of the remaining 15 treatments and create a forecast by moving the slider, or typing the forecast in a text box (though the latter method was not emphasized). The experts can scroll back up on the page to review the instructions or the results of the benchmark treatments. In order to test for fatigue, we randomize across experts the order of the treatments (the only randomization in the survey). Namely, we designate six possible orders, always keeping related interventions together, in order to minimize the burden on the experts.

We decided ex ante the rule for the scale in the slider. We wanted the slider to include, of course, the relevant values for all 18 treatments while at the same time minimizing the scope for confusion. As such, we decided against a scale between 0 and 3,500 (all possible values). Instead, we set the rule that the minimum and maximum unit would be the closest multiple of 500 that is at least 200 units away from all treatment scores. We asked the research assistant to check this rule against the results, which led to a score between 1,000 and 2,500.[6]

To summarize, in the first page of the survey the forecasters read a description of the task, have the option to sample the task and read the detailed instructions, see the results for the first three treatments and then make forecasts for the 15 other treatments.

The second page of the survey, which is designed to take only 3-5 minutes, elicits a measure of confidence in the stated forecasts (see Appendix Figure 1b). Namely, experts indicate their

[5] It is theoretically possible for the reward for accuracy to be negative for very low accuracy (the forecast errors need to exceed 400 points). This is rare in the sample and did not occur for the drawn individuals.

[6] From the email chain on 6/10/2015, email to the research assistant: "*We want to position [the bounds] at least 200 away from the lowest and highest average effort, and we want [...] min and max to be in multiples of 500*" and response: "*All of the average treatment counts are between 1,200 and 2,300*".

best guess as to the number of forecasts that they provided that are within 100 points of the actual average effort in a treatment. For example, a guess of 10 indicates a belief that the expert is likely to get 10 treatments approximately right out of 15. The experts then make a similar forecast for the average response of other groups of experts, such as the experts taken altogether and the top-15 most cited experts. Finally, the subjects indicate whether they have used MTurk subjects in their research and whether they are aware of MTurk, and finish off by indicating their name. While the identities of the experts are not revealed, we use the name to match to information on each expert and to assign the prize.[7]

Sample of Experts. We use mostly objective criteria to form a starting group of behavioral experts (broadly construed), and then contact the academics to which we have some connection (since we did not want to be seen as spamming researchers we did not know).

Our initial list comprised of: (i) all authors of papers presented at the Stanford Institute of Theoretical Economics (SITE) in Psychology and Economics and in Experimental Economics from its inception until 2014 (for all years in which the program is online); (ii) all participants of the Behavioral Economics Annual Meeting (BEAM) conferences from 2009 to 2014; (iii) individuals in the program committee and keynote speakers for the Behavioral Decision Research in Management Conference (BDRM) in 2010, 2012, and 2014; (iv) all invitees to the Russell Sage Foundation 2014 Workshop on "Behavioral Labor Economics" and (v) a list of behavioral economists compiled by ideas42. The resulting list includes experts in behavioral and experimental economics (groups (i) and (ii)), experts in decision-making and psychology (group (iii)), with a small set of additions (groups (iv) and (v)). We exclude graduate students from this list and add a small number of additional experts. We then pare down this list of over 600 people to 314 researchers to whom at least one of the two authors had some connection.

On July 10 and 11, 2015 one of the authors sent a personalized email to each of the 314 experts with subject '*[Survey on Expert Forecasts] Invitation to Participate*'. The email provided a brief introduction to the project and task and informed the expert that an email with a unique link to the survey would be forthcoming from Qualtrics. An automated reminder email was sent about two weeks later to experts who had not yet completed the survey (and had not expressed a desire to opt out from communication). Finally, one of the authors followed up with a personalized email to the non-completers.

Out of the 314 experts who were sent the survey, 213 completed it, for a participation rate of 68 percent. Out of the 213 responses, 5 had missing forecasts for at least one of the 15 treatments and are not included in the main sample. Columns 1 and 2 of Table 1 document the selection into response. Notice that the identity of the respondents is kept anonymous.

For each expert, we code four features: academic status, citations (measures of *vertical expertise*), field of expertise, and publications in an area (measures of *horizontal expertise*).

[7]The survey also has a unique identifier, providing another way to check the identity of the participant.

Searching CVs online, we code the status as Professor, Associate Professor, Assistant Professor, or Other (Post-doc and Research positions); we also record the year of PhD. For the citations, we aim to record the lifetime citation impact of a researcher using Google Scholar. For experts with a Google Scholar profile (about two thirds in our sample), we record the total citations in the profile as of April 2015. For the experts without a profile, we sum the Google Scholar citations for the 25 most cited papers by that expert. For individuals with more than 25 papers retrieved by Google Scholar, we extrapolate the additional citations for papers beyond the top 25 from citations for the 16th-25th most-cited papers.

As measures of horizontal expertise, we code field and publications in an area. For the field, we coded experts qualitatively as belonging to one of these fields: behavioral economics (including behavioral finance), applied microeconomics, economic theory, laboratory experiments, and psychology (including behavioral decision-making). As for the publications, using online CVs we code whether the individual, as far as we can tell, has written a paper on the topic of a particular treatment.[8]

Finally, on November 30, 2015, we provided personalized feedback, as we had promised. Each expert received an email from one of the authors with a personalized link to a website where they accessed a figure that included their own individual forecasts. We also randomly drew winners and distributed the prizes as promised.[9]

Other Samples. In a second round of survey collection, we also collect forecasts of a broader group: PhD students in economics, undergraduate students, MBA students, and a group of MTurk subjects recruited for the purpose.

The PhD students in our sample are in Departments of Economics at eight schools. Students at these institutions received an email from a faculty member or administrator at their school that included a brief explanation of our project and a school-specific link for those willing to participate. The participating PhD programs, the number of completed surveys, and the date of the initial request are: UC Berkeley (N=36; 7/31/2015), Chicago (N=34; 8/3/2015), Harvard (N=36; 8/4/2015), Stanford (N=5; 10/4/2015), UC San Diego (N=4; 10/7/2015), CalTech (N=7; 10/7/2015), Carnegie Mellon (N=6; 10/8/2015), and Cornell (N=19; 10/29/2015).

The first two waves of MBAs are students at the Booth School of Business at the University of Chicago who took a class in Negotiations from one of the authors: Wave 1 students (N=48,

[8]This involved some judgment calls when determining which topics counted for each treatment. For our beta-delta treatments, we include experts who wrote a paper about beta-delta or about time preferences more broadly. For the charitable donation treatments, we included papers about charitable giving or social preferences. Lastly, we separately categorized experts as having worked in the area of reference dependence and/or probability weighting rather than bunching together anyone who has worked on prospect theory into one category. For example, if an expert had just one paper about loss aversion, this expert would have horizontal expertise for the reference dependent framing treatments, but not for the probability weighting treatments.

[9]Since the survey included also other participants—PhDs, undergraduates, and MBAs—two of the prizes went to the experts. The prizes for the MTurk forecasters differ and are described below.

7/31/2015) took a class in Winter 2015 and Wave 2 students (N=60, 2/26/2016) took a class in Winter 2016. A third wave includes MBA students at Berkeley Haas (N=52, 4/7/2016).

The undergraduates are students at the University of Chicago and UC Berkeley who took at least an introductory class in economics: Wave 1 from Berkeley (N=36, 10/26/2015), Wave 2 from Berkeley (N=30, 11/17/2015), and Wave 3 from Chicago (N=92, 11/12/2015).

All of these participants saw the same survey (with the exception of demographic questions at the end of the survey) as the academic experts, and were incentivized in the same manner.

On 10/4/2016, we recruited MTurk workers (who were not involved in the initial experiment) to do a 10-minute task and take a 10-15 minute survey for a $1.50 fixed payment. These participants obviously have direct experience with working on MTurk and may have a better sense than academics or others about the priorities and interests of the MTurk population.

Half of the subjects (N = 269) were randomly assigned to an 'experienced' condition and did the 10-minute button-pressing task (in a randomly-assigned treatment) just like the MTurkers in our initial experiment before completing the forecasting survey. The other half of the subjects (N=235) were randomly assigned to an 'inexperienced' condition and did an unrelated 10-minute filler task (make a list of economic blogs) before completing the survey. Workers in both samples were told that they would be entered into a lottery and 5 of them would randomly win a prize based on the accuracy of their forecasts equal to $100 – Mean Squared Error/2,000. Thus, if their forecasts were off by 100 points in each treatment, they would receive $95 and if they were off by 300 points in each treatment, they would receive $55.

On 2/12/2016 we recruited an additional sample of MTurk workers (N= 258) who were not involved with any of the previous MTurk tasks. Like the 'experienced' MTurk sample above, they first participated in the 10-minute button-pressing task and then took the forecasting survey. For this sample, however, we made especially salient the value of trying hard when making their forecasts.[10] We also changed the incentives such that all participants were paid based on the accuracy of their forecasts (as opposed to being entered into a lottery). Specifically, each participant was told they would receive $5 – Mean Squared Error/20,000. Thus, if their forecasts were off by 100 points in each treatment, they would receive $4.50 and if they were off by 300 points in each treatment, they would receive $.50.

3 Accuracy of Expert Forecasts: Average and Individual

How does the average effort in the 15 experimental arms compare to the forecasts of the 208 academic experts? Table 2 lists the treatments, summarized by the category of the treatment

[10] At the top of the survey portion of the task, we wrote "*Important: you have the potential to increase your earnings for this HIT substantially by doing well on this survey. So please take your time and think hard about your answers.*" We concluded the survey instructions with one final note of encouragement: "*As you can see, if you are accurate you have the potential to substantially increase your earnings for this HIT.*"

(Column 1), the wording used (Column 2), and the sample size (Column 3). The table also reports the average effort in the treatment (Column 4) and the average forecast for that treatment by the 208 experts (Column 5), reproduced from DellaVigna and Pope (2016).

Figure 1 displays in graphical format the evidence on the accuracy of the average forecast. Each of the 18 points in the scatter plot represents a treatment, with the x axis indicating the average effort exerted in the treatment (Column 4 in Table 2) and the y axis indicating the average forecast by the experts (Column 5 of Table 2). The treatments are color-coded to group together treatments based on similar motivators. The benchmark treatments (three red squares) are on the 45 degree line since there was no forecast for those treatments.

Figure 1 shows our first main result: the experts, taken altogether, do a remarkable job of forecasting the average effort. The correlation between the forecasts and the actual effort is 0.77; the blue line displays the best interpolating line which has a slope of 0.527 (s.e. 0.122). Measured otherwise, there is only one treatment for which the distance between the average forecast and the average effort is larger than 200 points, the very-low-pay treatment. Across all 15 treatments, the average absolute error (Column 6 of Table 2) averages just 94 points, or 5 percent of the average effort across the treatments. In particular, the average expert forecast ranks in the correct order all the six treatments with no private monetary incentives: gift exchange, the psychology-based treatments, and the charitable-giving treatments.

Thus, an average of forecasts across many experts does a remarkable job forecasting. But a policy-maker, a firm, or an advisee will not typically be able to obtain forecasts for a large number of experts. How accurate, then, is the forecast of an individual expert?

Figure 2 and Table 3 provide information on the accuracy of individual forecasts using several measures. For the benchmark measure (absolute error, Panel A in Table 3 and Figure 2a), we compute the absolute error in forecast by treatment, and average across the 15 treatments.[11] We construct similarly the measure of squared error (Panel B in Table 3 and Figure 2b). We also compute the correlation (Panel C in Table 3 and Figure 2c) and rank-order correlation between the 15 forecasts and the treatments (Panel D in Table 3 and Figure 2d).

Figure 2a displays the cumulative distribution function of the absolute error for the 208 experts. The figure also displays the wisdom-of-crowds error (vertical red line), as well as two benchmarks for accuracy of prediction: random forecasts between 1,000 and 2,500 (dotted blue line) and random forecasts between 1,500 and 2,200 (vertical blue line).[12]

The figure shows that the accuracy of individual experts is substantially worse than the accuracy of the average forecast: 96 percent of experts have a lower accuracy than the average expert, and the average individual absolute error is 81 percent larger than the error of the

[11] In this figure and throughout the paper, we show results for the negative of the absolute error and the negative of the squared error, so as to display a measure of *accuracy*.

[12] For the random benchmarks, we draw 10,000 random forecasts from a uniform distribution in the specified range and compute the average error over the 10,000 draws.

average forecast (169 points vs. 93 points, Columns 1 and 2 in Table 3). This finding is known as 'wisdom of crowds': the average over a crowd outperforms most individuals in the crowd.

At the same time, there is clearly information in the individual forecasts: the large majority of experts are more accurate than one would predict based on random choice (blue lines).

Figures 2b, 2c, and 2d, and Panels B, C,and D of Table 3, show the findings with the three alternative measures of accuracy. The results are parallel: the large majority of experts do not do as well as the average expert, but they outperform random choice.

What explains the large wisdom-of-crowds effect? In particular, how many experts does it take to achieve a level of accuracy similar to the one for the group average?

In Figures 3a-d we plot once again the distribution of the individual and wisdom-of-crowds accuracy, but we also plot the counterfactual accuracy of forecasts averaged over smaller groups of N experts, with $N = 5, 10, 20$. Namely, we bootstrap 10,000 groups of N experts with replacement from the pool, and compute for each treatment the accuracy of the average forecast across the N forecasts. As Figure 3a shows, averaging over 5 forecasts is enough to eliminate the right tail of high-error forecasts and achieve an average absolute error rate of 114, down from 169 (Column 4 in Table 3). With 20 experts, the average absolute error, 99 points, is nearly indistinguishable from the one with the full sample (93 points) (Column 5 in Table 3). The pattern is very similar with squared error, correlation, and rank-order correlation.

After clarifying the role of group size, we now decompose the accuracy by treatment. Figures 4a-b display two treatments in which the majority of individual forecasters outperform the average forecast. Thus, the wisdom-of-crowds pattern does not apply in each treatment. In the treatment with the largest deviation of the mean forecast from the actual (very-low-pay, Figure 4c), more than 40 percent of experts do better than the wisdom-of-crowds estimate. There are then, however, treatments in which the wisdom-of-crowds is spot on (Figure 4d), and the large majority of experts do worse. Columns 7 and 8 of Table 2 present the expert accuracy by treatment. Across treatments, 37 percent of subjects do better than the average.

The critical point is that, while several experts do better than the wisdom-of-crowds in an individual treatment, it is not typically the *same* experts who do well, since the errors in forecast have a limited correlation across treatments. The wisdom-of-crowd estimate outperforms individual experts by doing reasonably well throughout. We return to this point below.

4 Determinants of Forecast Accuracy

4.1 Measures of Expertise

So far we have treated the 208 experts as interchangeable, and studied the implication of averaging expert forecasts versus following an individual expert. But clearly the experts in our sample differ in important ways. For example, the experts differ in *vertical* expertise—

academic rank and citations—, *horizontal* expertise—field of expertise and having a paper on the topic of the treatment—, and *contextual* expertise—knowledge of the experimental context.

These dimensions may be important determinants of the ability to forecast future research findings. We may thus be able to identify the 'right' experts within the overall group who have individual accuracy comparable to the accuracy of average forecasts.

We focus this section on our benchmark measure of accuracy: the (negative of) the absolute error rate; the results are very similar using the (negative of) squared error instead. We return later to the results for an ordinal measure of accuracy, the rank-order correlation.

Vertical Expertise. The first dimension of expertise which we consider is the vertical recognition within a field. Full professors have a recognition and prerogatives, like tenure, that most associate professors do not have, a difference *a fortiori* from assistant professors. In Figure 5a, we plot the distribution of the absolute error variable (averaged across the 15 treatments) by academic rank of the experts. Surprisingly, assistant professors are more accurate, if anything, than associate and full professors with respect to either accuracy measure.

Figure 5a presents a further test of the vertical expertise hypothesis: to the extent that depth of expertise matters, there should be a difference with respect to PhD students. Yet, PhD students, if anything, do better than the associate and full professors in their forecasts.

Table 4 provides regression-based evidence on expertise, specified as follows:

$$a_{i,t} = \alpha + \beta X_i + \eta_t + \lambda_{o(t)} + \varepsilon_{i,t} \tag{1}$$

An observation is a forecaster-treatment combination, and the dependent variable is a measure of accuracy $a_{i,t}$ for forecaster i and treatment t, such as the negative of the absolute error in forecast. The key regressors are the expertise variables X_i. The regression also includes treatment fixed effects η_t, as well as fixed effects for the order $o(t) = 1, ...15$ in which the treatment is presented, to control for forecaster fatigue.[13] The standard errors are clustered at the forecaster level to allow for correlation in errors across multiple forecasts by an individual.

Column 1 confirms the graphical findings on academic rank: associate and full professors have a higher error rate in forecasts than assistant professors (the omitted category), and PhD students are comparable to assistant professors.

Academic rank is of course an imperfect measure of vertical expertise. A measure that more directly captures the prominence of a researcher is the cumulative citation impact, which we measure with Google Scholar citations. Citations, among other features, are very strong predictors of salaries among economists (Hilmer, Hilmer, and Ransom, 2015). Figure 5b presents a split of the expert sample into thirds based on citations. The split has some overlap with the academic rank, but there is plenty of independent variation. The evidence suggests a perverse

[13]The term $o(t)$ is identified because there are six possible orders of presentations of treatments. We find no evidence of a trend of accuracy over the 15 forecasts, and the results are essentially identical if we remove the treatment and order fixed effects.

effect of citations: the least-cited third of experts has the highest forecasting accuracy. Column 2 of Table 4 corroborates this finding: Google Scholar citations, in logs to reduce the skewness, has a statistically significant negative effect on citations.

Thus, there is no evidence that vertical expertise improves the forecasting accuracy and some evidence to the contrary. One interpretation of the latter result is that prominent experts have a very high value of time and thus put less time and effort into the survey. While the regression above does not control for measures of effort, we show below that two measures of effort do not predict accuracy for experts; furthermore, high-rank and high-citation experts do not appear to be taking the survey faster or less carefully.

Horizontal Expertise. Experts differ not only vertically on prominence, but also horizontally in the topics in which they have expertise. Among the 'horizontal' features we consider, one is the main field of expertise. For each of the 312 experts sent a survey, we code a primary field: behavioral economics (including behavioral finance), applied microeconomics, economic theory, laboratory experiments, and psychology (including behavioral decision-making).[14] It is not obvious a priori which way field would affect the results, but we thought that behavioral economists may have an edge compared to standard economists given the emphasis on behavioral factors in the experiment. Further, given the emphasis on quantitative forecasts, it was possible that psychologists may be at a disadvantage.

Figure 6a displays the results graphically, and Column 3 in Table 4 presents the regression-based evidence. The differences between the groups, if any, are small. There is no statistically significant evidence that behavioral economists outperform standard economists, and only suggestive evidence of an advantage over psychologists.

While the evidence so far has considered vertical expertise and field of expertise separately, in Column 4 we include all the variables jointly: the point estimates remain relatively similar, but none of the variables is statistically significant.

Next, we turn to a more direct test of horizontal expertise. We code for each expert whether he or she has written a paper on a topic that is covered by the treatment at hand, and create an indicator variable for the match of treatment t with the expertise of expert i. For example, an expert with a paper on present-bias but no paper on social preferences is coded as an expert for the treatments with delayed pay, but not for the treatments on charitable giving. We also code whether the author has written a highly influential paper (by our assessment) on a topic.

In the specification testing for horizontal expertise (Column 5 of Table 4), we add expert fixed effects since we are identifying expertise for a given expert. (The regressions already include treatment fixed effects.) The results indicate a null effect of horizontal expertise: if anything, having written a paper lowers the accuracy (albeit not significantly). The confidence intervals are tight enough that we can reject that horizontal expertise increases accuracy by 8

[14] The coding is admittedly subjective, but at least was done before the data analysis.

points, just 5 percent of the average absolute error. The effect of writing an influential paper is also not significantly different from zero, though, not surprisingly, it is less precisely estimated.

As a final measure of horizontal expertise we test whether PhD students who self-report specializing in behavioral economics have higher accuracy. Figure 6b and Column 6 of Table 4 show that the variable has no discernible impact.

Contextual Expertise. So far, we have focused on academic versions of expertise: academic rank, citations, expertise in a field, and having written a paper on a topic. Knowledge of the setting, which we label *contextual expertise*, may play a more important role. Thus, we elicit from the experts their knowledge of the MTurk sample.

The survey respondents self-report whether they are aware of MTurk and whether they have used MTurk for one of their studies. Among the experts, all but 3 report having heard of MTurk, but the experts are equally split in terms of having used it. Thus, in Figure 7 we compare the accuracy of the two sub-samples of experts. The experts are indistinguishable with respect to absolute forecast error, as Column 7 of Table 4 also shows.

4.2 Non-Experts

Thus, various measures of expertise do not increase accuracy. Still, it is possible that academics and academics in training (the PhD students) share an understanding of incentives and behavioral forces which distinguish them from the non-experts. We thus compare their forecasts to forecasts by undergraduate students, MBA students, and an online sample. These forecasters have not received much training in formal economics, though some of them arguably have more experience with incentives at work (the MBAs) and with the context (the online sample).

Do non-experts make worse forecasts? Figure 8a compares the distribution of absolute error in forecasting for experts and non-experts. The figure provides evidence of a difference between experts and non-experts. The undergraduate students are somewhat less accurate, MBA students are significantly less accurate, and online forecasters in the MTurk sample do much worse. Column 1 in Table 5 shows that the difference in accuracy between the samples is statistically significant. Furthermore, Column 2 in Table 5 shows that the differences in accuracy between experts and non-experts replicate using, as a measure of accuracy, squared, instead of absolute, errors. Thus, when making forecasts about magnitudes of the experimental findings, yes, non-experts do worse than experts.

Yet, while the above measures of accuracy were the main ones we envisioned for this study[15], they are not always the relevant ones. Policymakers or businesspersons may simply be looking for a recommendation of the most effective treatment, or for ways to weed out the least effective ones. From this perspective, it is not as important to get the *levels* right in the forecasts, as it

[15]In our pre-registration, we mention three measures of accuracy: absolute error, squared error, and number of correct answers within 100 points of the truth (more on this below).

is to get the *order* right. We thus revisit the results using rank-order correlation as the measure of accuracy.[16] We correlate the ranking of the 15 treatments implied by the forecasts with the ranking implied by the actual average MTurk effort.

As Figure 8b shows, the rank-order correlation drastically changes the comparison with the non-experts. By the rank accuracy measure, undergraduates, MBAs, and even MTurk workers do about as well as the experts. Across all these samples, the average individual rank-order correlation with the realized effort is about 0.4 (Column 1 of Panel C, Table 3).

We present regression-based evidence using the specification

$$a_i = \alpha + \beta X_i + \varepsilon_i.$$

Notice that the rank-order correlation measure a_i is defined at the level of forecaster i, as opposed to at the treatment-forecaster level. Column 3 of Table 5 shows that there is no statistically significant difference in accuracy across the groups according to this measure.

This evidence so far regards measures of accuracy for individual forecasters. What about wisdom-of-crowds measures? When considering the accuracy of the mean forecast (Column 3 of Table 3), MBA students and especially MTurk workers display worse accuracy than experts with respect to both absolute error (Panel A) and squared error (Panel B). With respect to the rank-order measure (Panel C), though, the MTurk workers in fact do better than the experts, displaying a stunning wisdom-of-crowds rank-order correlation of 0.95 (compared to 0.83 for the experts). This pattern is also visible in Appendix Figure 2d, which shows just how well the average forecast of the MTurkers ranks the treatments, despite being off in levels. With the wisdom-of-crowds measure, MBAs do somewhat worse than experts on rank-order correlation, though still at a high correlation of 0.71 (see also Appendix Figure 2c). Overall, though, the wisdom-of-crowds results parallel the findings for individual accuracy.

What explains this discrepancy between the measures of accuracy in levels and the rank-based one? The difference occurs because non-experts, and especially the online sample, create informed forecasts for treatments, but often center them on an incorrect guess for the average effort across the 15 forecasts. In our particular setting, the non-experts expect too low a level of effort on average. This pattern is visible in Appendix Figures 2b-d for the average forecast, but is also displayed at the individual level in Appendix Figure 3a. A full quarter of MTurk workers forecast an average effort across the 15 treatments that is 200 points or more below the average actual effort (indicated by the red line). The other groups of non-experts—MBAs and undergraduates—also tend to display low forecasts, though not as much as the MTurk workers. In comparison, essentially none of the experts is off by so many points in the forecasts.

To further document whether an offset in level is a reason for the discrepancy, we explore the simple correlation between the individual forecasts and the average results. The correlation

[16] We thank seminar audiences and especially Katy Milkman for the suggestion to use rank-order correlation as an additional measure of accuracy.

measure is based on levels, as opposed to ranks, but it does not measure whether the level of effort is matched. As such, if non-experts mainly differ from experts in a level offset, they should be similar to experts according to simple correlation. Column 4 in Table 5 and Panel D in Table 3 show that this is indeed the case.

Thus, non-experts, while at a disadvantage to experts in forecasting the absolute level of accuracy, do as well in ranking the performance of the treatments. This is consistent with psychological evidence suggesting that people struggle with absolute judgments, but are better at making relative judgments. Miller (1962) argues that memory constraints lead humans to heavily rely on relative judgments as a heuristic in many settings. Laming (1984) further argues that people will be especially prone to make relative (as opposed to absolute) judgments when making magnitude estimations for a string of assignments. Difficulties in making absolute, versus relative, judgments matter for environmental and legal settings (e.g., Kahneman, Schkade, and Sunstein, 1998). Thus, it is not overly surprising that non-experts do better in providing a rank order, as opposed to an absolute measure of accuracy.

A striking factor about this result is that non-experts do as well as experts on ranking treatments despite spending significantly less effort on the task as measured by time spent and click-through on instruction. As Table 1 shows, undergraduates and MTurk workers (though not MBAs) spend less time on the survey than experts, and all three non-expert samples are much less likely than experts to click on the trial task or on the detailed instructions. We analyze further the effort measures in the next section.

Finally, one may wonder if the rank-order correlation changes the results in the previous section on vertical, horizontal, and contextual expertise of experts. In Appendix Figures 4a-c, we show that this is not the case.

4.3 Other Correlates of Accuracy

So far, we found that expertise does not help much with forecasts. The fine-grained *ex ante* measures of expertise do not increase forecasting accuracy, and experts as a group differ from non-experts only if the accuracy is about the levels, as opposed to the rank order, of treatments. If expertise does not help much, are there other ways, then, to discriminate among forecasters for accuracy? We consider measures of effort, confidence, and revealed ability.

Effort. A key variable that is likely to impact the quality of the forecasts is the effort put into the survey. While effort is unobservable, we collect two proxies that are likely to be quite indicative. The first measure is the time taken from initial login to the Qualtrics survey to survey completion. We cap this measure at 50 minutes, about the 90th percentile among experts, since participants who took very long were likely multi-tasking, or even returned to the survey hours or days later. The average time taken is 21 minutes among the experts, the PhD students and the MBA students, and lower in the other samples (Table 1).

Second, we keep track if the forecasters clicked on the practice link to try the task, and whether they clicked on the full experimental instructions. There is substantial heterogeneity, with 44 percent of experts and 48 percent of PhDs clicking on the practice task, but only 11 and 16 percent among undergraduates and MBA students, and 0 percent of the MTurk workers. The click rates on the instructions follow parallel trends but are about half the size.

Within each major group of forecasters—experts; undergraduate, PhD, and MBA students pooled; and MTurk workers—we display the average accuracy by decile of time taken, where the decile thresholds are formed on the joint sample.

Figure 9a and Columns 1-3 of Table 6 show the impact of time spent on the negative of the absolute forecasting error for the three groups of forecasters. The patterns differ significantly by sample. For the MTurk sample, there is a clear positive relationship between time spent and accuracy. The relationship is inverse U-shaped instead for the students and the experts. In these two samples, accuracy generally increases quite monotonically until the 4th or 5th decile, but then flattens or declines. In part, this may reflect the fact that individuals in the top deciles may well have left the browser window open for the task, and returned to it later; thus, a longer duration does not necessarily indicate more effort in doing the task.

With respect to the rank-order correlation (Figure 9b and Columns 5-7 of Table 6), there is less evidence of a relationship. While forecasters in the bottom decile—those taking 5 minutes or less—do worse, there is no obvious pattern for the other deciles, and in fact among the experts the forecasters taking longer do significantly worse on rank-order correlation.

We then turn to a second measure of effort in taking the task: whether the forecasters clicked on the trial task or on the full instructions for the task. Doing either, presumably, indicates higher effort. Figures 9c-d and Columns 1-2 and 5-6 of Table 6 show no obvious difference in accuracy for individuals who do, or do not, click on such instructions.[17]

In Table 6 we also report the effect of a further proxy of effort: the delay in days from when the invitation was sent out to when it was taken. Presumably individuals that are more enthusiastic are likely to do the survey sooner and with more effort. This variable has no obvious effect.

Overall, this evidence points to a mixed role played by effort in forecasting, other than at the very left tail (short durations). Yet, we cannot tell why some people appear to exert more effort than others. Are they more motivated? Do they have more free time? Are they just multi-tasking and thus taking longer?

In Figures 9e-f and in Columns 4 and 8 of Table 6 we present an attempt to instrument for effort. We run a third group of 250 MTurkers with increased incentives for accuracy in forecasting. Namely, we pay *each* participant in the survey a sum up to \$5 for accuracy, computed as \$5-MSE/20,000. This payment is higher than the promise to randomly pay two

[17]We do not display the coefficient on clickthrough for the MTurk sample, since no one in this sample clicked on the additional material.

of the MTurk workers in the other sample an accuracy bonus up to $100. In addition, we made the reward for accuracy more salient in the survey (see Section 2). The higher incentives appear to have no impact on forecasting accuracy, suggesting that, at least for the sample of MTurk workers, moral hazard in survey taking does not appear to play a major role.

Confidence. We also consider a measure of confidence, to test whether respondents appear to be aware of their own accuracy. On the second page of the survey, each forecaster indicated the number of forecasts (out of 15) which they expected to get within 100 points of the correct answer. As discussed below, each forecaster also indicated the number of forecasts that they expected other groups of forecasters to get right.

Figures 10a-c report the average accuracy for the same three groups—academic experts; PhD, undergraduate, and MBA students; and MTurk workers—as a function of each of the confidence levels from 0 to 15. We document the impact on absolute error (Figure 10a), on the number of forecasts (out of 15) within 100 points of the actual average effort (Figure 10b), and on the rank-order correlation (Figure 10c).

The confidence level is clearly predictive of accuracy with respect to both absolute error and the number of correct answers. This is especially true for MTurk workers, but also holds for the other groups. The relationship, though, is much flatter with respect to the rank-order measure. Appendix Figure 3c shows how the two findings co-exist: higher confidence increases the average forecast across all 15 treatments, which is too low for forecasters with low confidence. Thus, higher confidence removes this average bias in forecasting and thus improves the accuracy according to absolute error, but does not improve the ordering of treatments.

The regression results in Table 7 confirm the graphical findings. An increase in the expected number of correct answers of 5 (out of 15) reduces the average absolute error by 25 points for the student sample and by 44 points for the MTurk sample, a highly significant relationship. For experts instead, a similar increase in accuracy has a smaller (and not statistically significant) impact. The table also documents a similar pattern of effects on whether the forecast falls within 100 points of the actual effort (Columns 4-6), but displays a more limited relationship of confidence with rank-order correlation (significant only for the student sample).

Revealed Accuracy. Our third measure aims to capture an ability to make forecasts which may not be reflected in the (coarse) effort measures nor in the measure of confidence. In particular, if there are differences in forecasting skill, forecasters who are more accurate in one treatment are also likely to be more accurate in other treatments. We thus examine the correlation of accuracy across treatments, avoiding extrapolation across very similar treatments: the effort in these treatments will presumably be correlated, inducing a correlation in accuracy.

To start with, we consider a unique treatment within the design of the experiment: a 4-cent piece-rate incentive. Before making any forecasts, the forecasters were informed of the average effort in three treatments with varying piece rate: (i) no piece rate, (ii) piece rate of 1 cent per 100 points, and (iii) piece rate of 10 cents per 100 points. One of the 15 treatments which they

then predict is one with a piece rate of 4 cents per 100 points. Based on just the effort in the three benchmark treatments, as we show in DellaVigna and Pope (2016), it is possible to predict the effort in the 4 cent treatment accurately. We thus take the absolute deviation between the forecast and realized effort for the 4-cent treatment as a measure of 'revealed accuracy', presumably capturing the ability to work mentally through a simple model. None of the other treatments have this simple piece-rate property, so it is unlikely that there is a mechanical correlation between the prediction for the 4-cent treatment and the other treatments.

In Figures 11a-b, we plot the average accuracy for the three usual groups of forecasters, as a function of deciles in the accuracy of forecasting the 4-cent treatment. (We omit the 4-cent treatment in constructing the accuracy measures on the y axis, which thus refer to the other 14 treatments.) The correlation is striking: forecasters who do better in forecasting the 4c treatment also do better in the other treatments. The association is particularly strong in the MTurk sample. Indeed, for the top deciles there is almost no difference in accuracy between the MTurk sample and the sample of experts and students, bridging what is instead a large gap in accuracy of over 100 points for the bottom deciles.

Figure 11b shows that the pattern is different for the rank-order correlation measure of accuracy. There is still evidence of an increase in accuracy for higher deciles for the MTurk workers and for the students, but the evidence is less strong. As Appendix Figure 3d shows, part of the reason is that forecasters with higher revealed accuracy produce forecasts with on average a higher (and thus more correct) forecast, thus improving accuracy according to the absolute error measure, but not by the rank order measure.

Table 8 displays this evidence in a regression setting. We include in the regression also all the other control variables: vertical expertise and field of the experts (just for the expert regression in Columns 1 and 4), time taken to complete the survey and the confidence level. For the absolute error measure of accuracy (Columns 1-3), even after controlling for these variables, the 4-cent variable has remarkable explanatory power. In fact, it is the only variable that consistently predicts accuracy for the academic experts (Column 1). An increase of 100 points in the accuracy of the 4-cent prediction increases the accuracy in the other treatments by an average of 9.5 points for the experts. The predictability of accuracy is two to three times larger in all the other samples: students (23.8 points for each 100 points), and MTurks (31.4 points for each 100 points). We experimented with non-linear specifications in the 4-cent error term, but a linear specification captures the effect of the variable well. The table also shows that introducing the revealed-accuracy control generally reduces the load on the other predictors of accuracy, though confidence remains a significant predictor.

Table 8 also shows that there is a relationship, if more muted, for the rank-order correlation variable (Columns 4-6) for both students and MTurk workers. For these two groups, an increase of 100 points in accuracy for the 4-cent treatment increases the rank-order correlation by 0.02, a 5 percent effect relative to a mean correlation of 0.4.

Taking this one step further, it is natural to ask whether there is something special about the 4-cent treatment when it comes to capturing 'revealed accuracy'. Would the results differ if one constructed a variable based on one group of treatments, and then used it to predict accuracy in the forecasts of other treatments, excluding any treatments that are mechanically related? That is what we do in Table 9. In column 3, for example, we use the average accuracy in forecasts of the two charity treatments to predict accuracy in the other treatments. We report the results for the academic experts (Panel A) and for the other samples (Panel B).

Remarkably, almost all measures are helpful to predict accuracy in other treatments. The point estimates are not exactly comparable across columns because the different columns omit different treatments, but nonetheless the predictability hovers around 5-15 units for the experts and 20-40 units for the other samples. This result has an important implication. It does not appear that the critical component is accuracy in forecasting a model-driven incentive (which is a specific skill for the 4-cent treatment), but rather a general ability to form forecasts.

4.4 Superforecasters

As we have seen in Section 4.2, non-experts do as well as experts with respect to ranking treatments, but not with regards to measures of accuracy in levels, such as the negative of the absolute error rate. Thus, if one aims to obtain forecasts with the lowest absolute error rate, forecasts by academic experts are preferable. Yet, academic experts are busy professionals that are harder to reach than other samples such as students or online samples. Is there a way to use the latter, more available samples, and yet match the accuracy of the expert sample?

In the context of the Good Judgment Project, Mellers et al. (2015) and Tetlock and Gardner (2015) phrase a similar question as one of finding 'superforecasters'. Is it possible to find non-experts (in their setting individuals who do not have access to classified information) who nonetheless predict outcomes of national security as well as, or better than, the experts? Mellers et al. (2015) and Tetlock and Gardner (2015) find that it is possible to do so using the previous track record of forecasters.

In our context, to identify superforecasters we use the variables examined so far: measures of expertise, effort, confidence, and revealed accuracy. As Section 4.3 shows, especially the latter measure, which is in spirit of using the track record of a forecaster, is predictive of forecasting accuracy. We thus take the same specification as in Table 8, with all these control variables, and for each sample of experts we predict accuracy. To avoid in-sample data mining, we use a 10-fold method to obtain out-of-sample predictions. For each subgroup, we randomly split the forecasters into 10 equal-sized groups. We leave out the first tenth, estimate the model with the remaining nine tenths of the data, and predict accuracy in the left-out tenth. Then we rotate the same procedure with the next tenth of the data until we covered all the observations. Within each group, we select the top percentile in predicted accuracy.

Table 10 reports the results for both individual accuracy (Column 1) and wisdom-of-crowds average (Column 3). Panel A reports the results for the academic experts, comparing the overall group of experts to the optimal 20% and optimal 10% of experts constructed using all controls, as well as the optimal 20% constructed using all controls other than the revealed-ability variable. The table shows that among the experts we are unable to select a subset of super-experts who will do better than the overall group. The accuracy measure for the full sample of experts and for the top 20% (constructed with all controls) overlap (Figure 12a).

The results differ for the other groups of forecasters. In the sample of PhD students, MBAs, and undergraduates (Panel B), the optimal 20% of forecasters outperforms significantly the academic experts both at the individual level (Figure 12a) and with the wisdom-of-crowds measure.[18] Indeed, the wisdom-of-crowds absolute error for the top 20% in this group is as low as 73 points, compared to 95 points for the average expert, a difference that is statistically significant (Column 4).[19] Figure 12b displays the results for the wisdom-of-crowds measure for bootstrapped samples of 20 forecasters.

The results are equally striking for the MTurk workers. While on average MTurk workers have a much higher individual absolute error than experts (272 points on average versus 175 points), picking the top 20% of MTurkers nearly closes the gap for individual accuracy. Further, when using the wisdom-of-crowds measure, the selected MTurk forecasters actually *outperform* the academic experts, achieving an accuracy of 73, compared to 95 for the experts, a difference that is statistically significant. The revealed-ability variable plays an important role: the prediction without it does not achieve the same accuracy.

Thus, especially if it is possible to observe the track record, even with a very short history (in this case we use just one forecast), it is possible to identify subsamples of non-expert forecasters with accuracy that matches or surpasses the accuracy of expert samples.

4.5 Beliefs about Expertise

Our seventh and final result addresses a meta-question: Did we know all this already? Perhaps there was a shared understanding of these main issues, that for example vertical and horizontal expertise do not matter for the quality of forecasting.

In the spirit of the forecasting idea, on the second page of the survey we elicited the expected accuracy for different groups of forecasters (Appendix Figure 1b). In order to compare the responses to the data while keeping the forecasts simple, we asked for the expected number of treatments that an individual from a particular group would guess within 100 points of

[18] While omitting the 4-cent revealed-ability variable decreases the ability to identify superforecasters, the top 20% group selected using the other variables (effort and confidence) already outperforms the experts.

[19] To compute whether the wisdom-of-crowd accuracy of sample at hand is statistically significantly different from the one of the overall sample of experts, we bootstrap the sample 1,000 time. At each bootstrap we redraw both the experts and the non-experts, determining a new group of superforcasters.

the truth. For example, the forecasters guess the average number of correct answers for the academic experts participating in the survey. Next, they guess the average number of correct answers for the 15-most cited academics participating in the survey. The differences between the two guesses is a measure of belief about the impact of vertical expertise.

Figure 13 plots the beliefs of the 208 experts compared with the actual accuracy for the specified group of forecasters. The first cell indicates that the experts are on average accurate about themselves, expecting to get about 6 forecasts 'correct', in line with the realization. Furthermore, as the second cell shows, the experts expect other academics to do on average somewhat better than them, at 6.7 correct forecasts. Thus, this sample of experts does not display evidence of overconfidence (Healy and Moore, 2008), possibly because the experts were being particularly cautious not to fall into such a trap.

The key cells are the next ones, on the expected accuracy for other groups. The experts expect the 15 most-cited experts to be somewhat more accurate when the opposite is true. They also expect experts with a psychology PhD to be more accurate where, once again, the data points if anything in the other direction. They also expect that PhD students would be significantly less accurate, whereas the PhD students match the experts in accuracy.[20] The experts also expect that the PhD students with expertise in behavioral economics would do better, which we do not find.[21] The experts correctly anticipate that MBA students and MTurk workers would do worse. However, they think that having experienced the task among the MTurkers would raise noticeably the accuracy, counterfactually.

Overall, the beliefs about the determinants of expertise are systematically off target. This is understandable given the lack of previous evidence on the accuracy of research forecasts.

5 Model and Calibration

We presented a set of findings about forecasts of research results. Can a simple model make sense of the key findings? We model agent i making forecasts about the results in treatments $k = 1, ..., K$. Let $\theta = (\theta_1, \dots, \theta_K)$ be the outcome (unknown to the agent) in the K treatments. Given the incentives in the survey, the agent aims to minimize the squared distance between the forecast f_k^i and the result θ_k. We assume that agents start with a non-informative prior and that agent i, with $i = 1, ..., I$, draws a signal s_k^i about the outcome of treatment k:

$$s_k^i = \theta_k + \eta_k + v_i + \sigma_i \epsilon_k^i \,. \tag{2}$$

[20] For the PhD students we report the actual accuracy including only Univeristy of Chicago and UC Berkeley PhDs, since the survey refers only to these two groups. The results are similar (and more precisely estimated) if we use all PhD students to compute the actual accuracy.

[21] We did not elicit forecasts about undergraduate students since we had not decided yet whether to contact a sample of undergraduates at the time the survey launched.

The deviation of the signal s_k^i from the truth θ_k consists of three components, each i.i.d. and independent from the other components: (i) $\eta_k \sim N(0, \sigma_\eta^2)$ is a deviation for treatment k that is common to all forecasters; (ii) $v_i \sim N(\mu, \sigma_v^2)$ is a deviation for forecaster i that is common across all treatments (with a possible bias term if $\mu \neq 0$); (iii) $\sigma_i \epsilon_k^i$, with $\epsilon_k^i \sim N(0,1)$, is the idiosyncratic noise component, with heterogeneous σ_i: more accurate forecasters are characterized by a lower σ_i. We assume that σ_i is independent from ϵ_k^i and that the idiosyncratic variance σ_i^2 follows an inverse gamma distribution: $\sigma_i^2 \sim IG(\alpha, \beta)$.

We assume that the agent is unaware of the systematic bias μ. Given this and the uninformative prior, the signal s_k^i is an agent's best estimate (that is, $f_k^i = s_k^i$), given that it minimizes the (subjective) expected squared distance between the forecast and the result in treatment k.

The error term ε_k^i captures idiosyncratic noise in the forecasts. Importantly, the forecasters differ in the extent of idiosyncratic noise, with some experts providing less noisy forecasts (lower σ_i). This heterogeneity has implications for the correlation of errors across treatments. If σ_i is very similar across forecasters, the absolute error in one treatment will have little predictability for the absolute error in another treatment for the same person, as the error in forecast arises from noise that is similar across all forecasters. If some forecasters, instead, have significantly lower σ_i than other forecasters, there will be cross-treatment predictability: the forecasters who do well in one treatment are likely to have low σ_i, and thus do well in another treatment too. Thus, heterogeneity in σ_i can capture the results on revealed forecasting ability.

Why do we need the additional error terms η_k and v_i? A model with just the idiosyncratic error term misses two important features of the data. First, some treatments appear harder to forecast than others, as Table 2 shows. Given the large sample size of forecasters, these cross-treatment differences are unlikely to be due to idiosyncratic error. The term η_k allows for such differences, potentially capturing an incorrect common reading of the literature (or of the context) for a particular treatment, or an unusual experimental finding.

Second, forecasters differ in the average forecast across all 15 treatments, again more than one would expect based on idiosyncratic noise. Appendix Figure 3a shows that in particular non-experts tend to under-forecast effort, with a large heterogeneity. The term v_i captures an agent i being more optimistic (or pessimistic) about the effect of all treatments.

We now document that this simple model can make sense of several qualitative features of the data. We calibrate the five model parameters: σ_η^2, μ, σ_v^2, α, and β. To tie down these parameters, we use three variances, a measure of average bias, as well as the between-treatment correlation in absolute error for one forecaster. We then use the calibrated model to check how well we match some key features in the data. We do the calibration separately for the sample of 208 experts and for the other samples (students and MTurks).

As Panel A of Table 11 shows, the first moment is the variance of the forecast error, $s_k^i - \theta_k$ which, as we show in the Appendix, equals $V(s_k^i - \theta_k) = \sigma_\eta^2 + \sigma_v^2 + E(\sigma_i^2)$. Second, we consider the variance of the wisdom-of-crowds error, obtained by averaging s_k^i across all forecasters i,

$\bar{s}_k = \Sigma_i s_k^i/I$: $V\left(\bar{s}_k - \theta_k\right) = \sigma_\eta^2 + \left[\sigma_v^2 + E\left(\sigma_i^2\right)\right]/I$. Intuitively, the only part of the variance that does not shrink is the treatment-specific variance. Third, we consider the variance of the average error for forecaster i across all treatments k, $\bar{s}^i = \Sigma_k s_k^i/K$: $V\left(\bar{s}^i - \Sigma_k \theta_k/K\right) = \sigma_v^2 + \left[\sigma_\eta^2 + E\left(\sigma_i^2\right)\right]/K$. The overall variance and the treatment-specific variance are shrunk by averaging, but the person-specific variance σ_v^2 is not. These three expressions allow one to back out σ_v^2, σ_η^2, and $E\left(\sigma_i^2\right)$. To tie down the average bias in forecast μ, we use the overall average error, $\sum_{i,k}\left(s_k^i - \theta_k\right)/(I * K)$. Finally, to identify α and β, the parameters determining the distribution of σ_i, we use the correlation in errors across treatments.

As Panel A shows, the experts, compared to the non-experts, have a significantly lower variance $V\left(s_k^i - \theta_k\right)$ and also lower variance of the average error $V\left(\bar{s}^i - \theta_k\right)$, as Appendix Figure 3a documents. The experts have instead a *higher* variance of the wisdom-of-crowd error compared to non-experts, as one can see comparing Figure 1 (for experts) to Appendix Figures 2a-d (non-experts). These figures also indicate that, while for experts there is only a negligible bias μ on average, there is a sizeable bias for non-experts. The final moment is the correlation of the absolute error across treatments, which we take from Column 1 in Table 9.

Panel B displays the implied calibrated values of the parameters. The experts have a higher calibrated variance σ_η^2 of the treatment-specific error than non-experts, but a lower variance σ_v^2 of the forecaster-specific error, as well as a lower average idiosyncratic variance $E\left(\sigma_i^2\right)$.

To identify the heterogeneity in σ_i^2, we use the cross-treatment correlation in absolute error for a forecaster. Figure 14a plots the implied correlation in absolute error across treatments as we increase the heterogeneity in the variance σ_i^2, *holding constant* the average variance $E\left(\sigma_i^2\right)$ at the calibrated value.[22] For low values of the (log) standard deviation of σ_i^2 (on the x axis), the implied correlation of errors is quite low: if individuals are off in one treatment, they are not much more likely to be off in another treatment (other than because of the realized v_i term). For high values of the (log) standard deviation of σ_i^2, instead, some forecasters are much better than others in making forecasts. In this case, absolute error in one treatment will be more informative of the error in another treatment. The observed correlation (.09 for experts and .29 for non-experts, Column 1 of Table 9) pins down approximately the two parameters of the inverse gamma distribution for experts and non-experts and thus the distribution of σ_i^2.[23] As Figure 14b shows, non-experts have on average higher idiosyncratic variance and more heterogeneity. Given the higher heterogeneity, there are more 'superforecasters' (agents with low σ_i) among the non-experts.

[22]Each point reports the average correlation from 1,000 simulated samples for those parameter values. Each sample has the same number of individuals (208 for the experts and 1,227 for the non-experts) and the same number of treatments (15) as in the data. Within a sample, we correlated absolute error in each of 14 treatments on absolute error on the 15th treatment (held constant) for each person. This mirrors the regressions in Table 8 and 9. Each calibration varies α and β so as to keep $E\left(\sigma_i^2\right)$ constant, but vary the variance $Var\left(\sigma_i^2\right)$.

[23]The sample of non-experts achieves an asymptote of correlation of 0.27, we thus pick a point with high enough standard deviation. Picking a higher or lower point in the range leads to very similar results.

In Panel C of Table 11 we examine whether this simple model can match some key features of the forecasting data. We simulate 1,000 draws for the calibrated parameters, each draw with the same number of forecasters and the same number of treatments as in the sample (15). We display the average of the statistic examined over the 1,000 draws.

Comparing the data (Columns 1 and 3) to the calibrated values (Columns 2 and 4), the model matches remarkably well the average individual absolute error and the average wisdom-of-crowd error for both experts and non-experts. It also reproduces very closely the error with a group of 5 forecasters, implying that the model reproduces the speed of convergence in the error due to aggregation of opinions. Figure 14c displays these patterns visually, comparing the c.d.f. of the absolute error for the two groups, in the data versus the calibrations.

We can also use the calibrated model to benchmark the 'superforecasting' results. We select the 20% with the lowest absolute error in a fixed treatment, and examine the absolute error in the other treatments. This selection criterion mirrors the results in Tables 8 and 9, in which the absolute error in one treatment is the strongest predictor in determining the sample of 'superforecasters' (Table 10). Within the experts, the forecasters chosen in this way display similar individual and wisdom-of-crowds absolute error as in the sample of all experts. Within the non-experts, instead, the superforecasters outperform the overall group of non-experts both individually and as a group. These results match the findings in the data, and reflect the wider dispersion of both forecaster errors and idiosyncratic error among the non-experts.

Next, we turn to the rank-order correlation in forecasts, a measure that reverses a key result: non-experts are as good as experts in rank-ordering treatments, and are in fact better when using a wisdom-of-crowds measure. Can our simple calibration match this fact?

The calibration indeed predicts that experts will have a similar rank-order correlation as non-experts, though it overstates the level of the individual-level correlation for both groups (about 0.6 compared to 0.4 in the data). The calibration matches remarkably well the wisdom-of-crowds rank-order correlation, reproducing not just the qualitative features, but also the magnitudes in the data: about 0.8 for experts and 0.9 for non-experts.

We also check whether the model matches a different measure of strength of the wisdom of crowds: the share of forecasters that does better than the wisdom-of-crowds forecasts. In this respect, the calibrations match quite well the features in the data.

Finally, we consider a key question raised initially: does the model need all the components? In Appendix Table 1, we replicate the calibrations turning off in turn each error term component. Calibrations without the forecaster-specific error (and bias $\mu = 0$) lead to unrealistic wisdom-of-crowds accuracy for the non-experts (Columns 3 and 4). Calibrations without the treatment-specific error instead lead to unrealistic wisdom-of-the-crowds accuracy for the experts (Columns 5 and 6). Finally, assuming no heterogeneity in the idiosyncratic error (constant σ), we cannot match the correlation of errors across treatments (Columns 7 and 8).

Overall, this model is able to reproduce several stylized features of the data, including in-

dividual accuracy versus the wisdom of crowds, performance of 'superforecasters', differences between experts and non-experts, and differences between absolute error and rank-order correlation. We should, however, be clear that this simple model should be seen just as a starting point to understand how forecasters form their beliefs about future research findings.

6 Conclusion

When it comes to forecasting future research results, *who* knows *what*? We have attempted to provide systematic evidence within one particular setting, taking advantage of forecasts by a large sample of experts and of non-experts regarding 15 different experimental treatments.

Within this context, forecasts carry a surprising amount of information, especially if the forecasts are aggregated to form a wisdom-of-crowds forecast. This information, however, does not reside with traditional experts. Forecasters with higher vertical, horizontal, or contextual expertise do not make more accurate forecasts. Furthermore, forecasts by academic experts are more informative than forecasts by non-experts only if a measure of accuracy in 'levels' is used. If forecasts are used just to rank treatments, non-experts, including even an easy-to-recruit online sample, do just as well as experts. Thus, the answer to the *who* part of the question above is intertwined with the answer to the *what* part.

Even if one restricts oneself to the accuracy in 'levels' (absolute error and squared error), one can select non-experts with accuracy meeting, or exceeding, that of the experts. Therefore, the information about future experimental results is more widely distributed than one may have thought. We presented also a simple model to organize the evidence on expertise.

The current results, while just a first step, already draw out a number of implications for increasing accuracy of research forecasts. Clearly, asking for multiple opinions has high returns. Further, traditional experts may not necessarily offer a more precise forecast than a well-motivated audience, and the latter is easier to reach. One can then screen the non-experts based on measures of effort, confidence, and accuracy on a trial question.

The results stress what we hope is a message from this paper. As academics we know so little about the accuracy of expert forecasts that we appear to hold incorrect beliefs about expertise and are not well calibrated in our accuracy. We conjecture that more opportunities to make forecasts, and receive feedback, could lead to significant improvements. We hope that this paper will be followed by other studies examining forecast accuracy.

References

[1] Amir, On, and Dan Ariely. "Resting on Laurels: The Effects of Discrete Progress Markers as Subgoals on Task Performance and Preferences." *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Vol. 34(5) (2008), 1158-1171.

[2] Amir, Ofra, David G. Rand, and Ya'akov K. Gal, 2012. "Economic games on the Internet: The effect of $1 stakes." *PLoS ONE*, 7(2), e31461.

[3] Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg. 2016. Forthcoming. "Decision Theoretic Approaches to Experiment Design and External Validity", *Handbook of Field Experiments.*

[4] Ben-David, Itzhak, John Graham, Cam Harvey, 2013, "Managerial Miscalibration", *Quarterly Journal of Economics* 128 (4), 1547–1584.

[5] Berger, Jonah, and Devin Pope. "Can Losing Lead to Winning." *Management Science* Vol. 57(5) (2011), 817-827.

[6] Camerer, Colin et al.. 2016. "Evaluating Replicability of Laboratory Experiments in Economics" *Science*, 10.1126.

[7] Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia. 2016 "Inflation Expectations, Learning and Supermarket Prices: Evidence from Survey Experiments" Working paper.

[8] Coffman, Lucas and Paul Niehaus. 2014. "Pathways of Persuasion" Working paper.

[9] DellaVigna, Stefano and Devin Pope. 2016. "What Motivates Effort? Evidence and Expert Forecasts" NBER Working paper w22193.

[10] Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. "Using prediction markets to estimate the reproducibility of scientific research", *PNAS,* Vol. 112 no. 50, 15343–15347.

[11] Erev, Ido, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan M. Herzog, Robin Hau, Ralph Hertwig, Terrance Stewart, Robert West, and Christiane Lebiere, "A Choice Prediction Competition: Choices from Experience and from Description." *Journal of Behavioral Decision Making*, 23 (2010): 15-47.

[12] Galton, Francis. 1907. "Vox Populi " *Nature*, No. 1949, Vol. 75, 450-451.

[13] Goodman, Joseph K., Cynthia E. Cryder, and Amar Cheema. 2013. "Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples." *Journal of Behavioral Decision Making*, 26, 213-224.

[14] Groh, Matthew, Nandini Krishnan, David McKenzie, Tara Vishwanath. 2015. "The Impact of Soft Skill Training on Female Youth Employment: Evidence from a Randomized Experiment in Jordan" Working paper.

[15] Moore, Don A.; Healy, Paul J. 2008. **"**The trouble with overconfidence." *Psychological Review*, Vol 115(2), 502-517.

[16] Hilmer, Christiana E., Michael J. Hilmer, and Michael R. Ransom. 2015. "Fame and the Fortune of Academic Economists: How the Market Rewards Influential Research in Economics." *Southern Economic Journal*, Vol. 82(2), pp. 430–452.

[17] Horton, John J. and Chilton, Lydia B. 2010. "The Labor Economics of Paid Crowdsourcing" *Proceedings of the 11th ACM Conference on Electronic Commerce.*

[18] Horton, John J., David Rand, and Richard Zeckhauser. 2011. "The online laboratory: conducting experiments in a real labor market" *Experimental Economics*, Vol. 14(3), pp 399-425.

[19] Ipeirotis, Panagiotis G. "Analyzing the Amazon Mechanical Turk Marketplace. 2010. " *XRDS: Crossroads, The ACM Magazine for Students* Vol. 17, No. 2: 16-21.

[20] Kahneman, Daniel, David Schkade, Cass Sunstein. 1998. "Shared Outrage and Erratic Awards: The Psychology of Punitive Damages" *Journal of Risk and Uncertainty*, Vol. 16(1), pp 49–86.

[21] Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015. "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105(4): 1478-1508.

[22] Laming, Donald. 1984. "The relativity of 'absolute' judgments." *British Journal of Mathematical and Statistical Psychology*, 37(2), 152-183.

[23] Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, Philip Tetlock. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," *Perspectives on Psychological Science* May 2015 vol. 10 no. 3 267-281.

[24] Miller, George A. 1956. "The magical number seven, plus or minus two: some limits on our capacity for processing information." *Psychological Review*, 63(2), 81-97.

[25] Open Science Collaboration. (2015). "Estimating the reproducibility of psychological science." *Science*, 349(6251)

[26] Paolacci, Gabriele. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgement and Decision Making* Vol. 5, No. 5: 411-419.

[27] Paolacci, Gabriele, and Jesse Chandler. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* Vol 23(3), 184-188.

[28] Ross, Joel, et al. 2010. "Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk." In CHI '10 Extended Abstracts on Human Factors in Computing Systems: 2863-2872.

[29] Sanders, Michael, Freddie Mitchell, and Aisling Ni Chonaire. 2015. "Just Common Sense? How well do experts and lay-people do at predicting the findings of Behavioural Science Experiments" Working paper.

[30] Joseph P. Simmons, Leif D. Nelson and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant", *Psychological Science*, Vol. 22(11), pp. 1359-1366.

[31] Snowberg, Erik, Justin Wolfers and Erik Zitzewitz. 2007. "Partisan Impacts on the Economy: Evidence from Prediction Markets and Close Elections." *Quarterly Journal of Economics* 122, 2, 807-829.

[32] Surowiecki, James. 2005. *The Wisdom of Crowds.* Knopf Doubleday Publishing.

[33] Tetlock, Philip E., Dan Gardner. 2015 *Superforecasting: The Art and Science of Prediction*, Random House.

[34] Vivalt, Eva. 2016. "How Much Can We Generalize from Impact Evaluations?" Working paper.

[35] Wolfers, Justin, Zitzewitz, Eric. 2004. "Prediction Markets" *The Journal of Economic Perspectives*, Vol. 18 (2), pp. 107-126.

A Appendix A - Model and Calibration Appendix

We present here the derivation of the three variances used in the calibration. Notice first that

$$\begin{aligned} Var\left(\sigma_i\varepsilon_k^i\right) &= E(\sigma_i^2\left(\varepsilon_k^i\right)^2) - [E(\sigma_i^2\varepsilon_k^i)]^2 = E\left(\sigma_i^2\right)E(\left(\varepsilon_k^i\right)^2) - E\left(\sigma_i^2\right)^2 E\left(\varepsilon_k^i\right)^2 \\ &= E\left(\sigma_i^2\right) * 1 - E\left(\sigma_i^2\right)^2 * 0 = E\left(\sigma_i^2\right). \end{aligned}$$

The cross-sectional variance satisfies

$$Var\left(s_k^i - \theta_k\right) = Var\left(\eta_k\right) + Var\left(v_i\right) + Var\left(\sigma_i\varepsilon_k^i\right) = \sigma_\eta^2 + \sigma_v^2 + E\left(\sigma_i^2\right).$$

The wisdom-of-crowds variance equals

$$\begin{aligned} Var(\bar{s}_k - \theta_k) &= \frac{1}{I^2}\left[\sum_{i=1}^{I} Var(\eta_k + v_i + \sigma_i\varepsilon_k^i) + 2\sum_{i<j} Cov(\eta_k + v_i + \sigma_i\varepsilon_k^i, \eta_k + v_j + \sigma_j\varepsilon_k^j)\right] \\ &= \frac{1}{I^2}[I\left(\sigma_\eta^2 + \sigma_\nu^2 + E\left(\sigma_i^2\right)\right) + 2\sum_{i<j}\underbrace{Var(\eta_k)}_{\sigma_\eta^2} + \underbrace{Cov(\eta_k, v_j + \sigma_j\varepsilon_k^j)}_{0} \\ &+ \underbrace{Cov(\eta_k, v_i + \sigma_i\varepsilon_k^i)}_{0} + \underbrace{Cov(v_i + \sigma_i\varepsilon_k^i, v_j + \sigma_j\varepsilon_k^j)}_{0})] \\ &= \frac{1}{I^2}\left[I\left(\sigma_\eta^2 + \sigma_\nu^2 + E\left(\sigma_i^2\right)\right) + I(I-1)\sigma_\eta^2\right] = \sigma_\eta^2 + \frac{1}{I}(\sigma_\nu^2 + E\left(\sigma_i^2\right)). \end{aligned}$$

The average-bias variance equals

$$\begin{aligned} Var\left(\overline{s^i - \theta}\right) &= \frac{1}{K^2}\left[\sum_{k=1}^{K} Var\left(\eta_k + v_i + \sigma^i\varepsilon_k^i\right) + 2\sum_{k<l}\underbrace{Cov\left(\eta_k + v_i + \sigma^i\varepsilon_k^i, \eta_l + v_i + \sigma^i\varepsilon_l^i\right)}_{\sigma_\nu^2}\right] \\ &= \frac{1}{K^2}\left[K\left(\sigma_\eta^2 + \sigma_\nu^2 + E\left(\sigma_i^2\right)\right) + K\left(K-1\right)\sigma_\nu^2\right] = \sigma_\nu^2 + \frac{1}{K}\left(\sigma_\eta^2 + E\left(\sigma_i^2\right)\right). \end{aligned}$$

Given these expressions, we can solve for $E\left(\sigma_i^2\right)$, $\hat{\sigma}_\nu^2$ and $\hat{\sigma}_\eta^2$:

$$\begin{aligned} \hat{\sigma}_\eta^2 &= \frac{1}{I-1}\left(IVar(\bar{s}_k - \theta_k) - Var\left(s_k^i - \theta_k\right)\right) \\ \hat{\sigma}_\nu^2 &= \frac{1}{K-1}\left(KVar\left(\overline{s_i - \theta}\right) - Var\left(s_k^i - \theta_k\right)\right) \\ E\left(\sigma_i^2\right) &= Var\left(s_k^i - \theta_k\right) - \hat{\sigma}_\eta^2 - \hat{\sigma}_\nu^2. \end{aligned}$$

Finally, for the distribution of σ_i^2, which we assume to be inverse gamma distributed $IG(\alpha, \beta)$, from standard properties we obtain

$$\alpha = 2 + \frac{E\left[\sigma_i^2\right]}{Var\left[\sigma_i^2\right]}, \tag{3}$$

$$\beta = E\left[\sigma_i^2\right]\left(1 + \frac{E\left[\sigma_i^2\right]}{Var\left[\sigma_i^2\right]}\right). \tag{4}$$

Given an implied $E\left[\sigma_i^2\right]$ from the calibration, we can vary $Var\left[\sigma_i^2\right]$ through the implied values of α and β to match the correlation of absolute error across treatments.

Calibration. For the calibration in Table 11, we take the moments in Panel A from the data: the overall variance in error, the variance of the wisdom-of-crowd error, the variance of the average error in forecast, and the average bias. For each of the three variances, we report the square root (the standard deviation). For the correlation of absolute error across treatments, we take the coefficients in Column (1) of Table 9, appropriately divided by 100 (given that in Table 9 the regressor was divided by 100): 0.09 for experts and 0.29 for non-experts. Notice that Table 9 is not exactly the correlation of absolute error for treatment i on absolute error for treatment j on two grounds. First, we estimate a regression and not a simple correlation. Second, there are additional controls in the regression. However, the additional controls have a limited impact and thus the regression coefficient is close to a simple correlation coefficient, assuming that the variance of the dependent variable is similar to the variance of the independent variable, which on average it is.

As we explain above, the three variances identify σ_η^2, σ_v^2, and $E\left(\sigma_i^2\right)$. In Panel B we report the square root of these terms. For the third term, notice that we are thus reporting $\sqrt{E\left(\sigma_i^2\right)}$. To identify the α and β parameters for the distribution of σ_i^2, we use the correlation of absolute errors across treatments described above. Specifically, as expressions (3) and (4) make clear, for any given $E\left[\sigma_i^2\right]$ (which we take at the estimated value in Panel B), an assumed value of $Var\left[\sigma_i^2\right]$ pins down α and β. In Figure 14b, we vary $Var\left[\sigma_i^2\right]$ keeping constant $E\left[\sigma_i^2\right]$, generating combinations of α and β. For each of these combinations, we simulate 1,000 draws. Each sample has the same number of individuals (208 for the experts and 1,227 for the non-experts) and the same number of treatments (15) as in the data. Within a sample, we correlated absolute error in each of 14 treatments on absolute error on the 15th treatment (held constant) for each person. This mirrors the regressions in Table 8 and 9. For each point, the y axis reports the average estimated regression slope. As Figure 14b shows, the larger the variance in σ_i^2, the more accuracy in one treatment predicts accuracy in another treatment, as expected. Furthermore, the predictability is higher for non-experts than for experts for any level of $Var\left[\sigma_i^2\right]$ because non-experts have higher variance in v_i, and draws of v_i induce correlation across treatments. From Figure 14b, we set the values of α and β. We should note that the correlation of errors for non-experts comes close, but does not quite reach 0.29 in the figure; we thus pick a high value subject to not violating the constraint $\alpha \geq 2$. The exact value chosen is immaterial to the results in Panel C.

Having determined the values of all five parameters, we report in Panel C the results of simulations of populations with the appropriate parameters. More precisely, we do 1,000 draws of simulated populations, each of which with 15 treatments and as many subjects as appropriate (that is, 208 for the experts and 1,227 for non-experts). Each draw simulates a realization of our survey if the underlying parameters were the hypothesized ones. Within each draw, we compute the relevant statistic, and we report the mean across the 1,000 draws.

The statistics are computed as in the paper. For 'superforecasters', we fix one treatment (the equivalent of the 4-cent treatment) and pick the 20% of subjects with the lowest realized absolute error. We then compute the absolute error for these subjects, averaging across the remaining 14 treatments. Notice that for most of the statistics, the draw of θ_k does not matter, as all is a function of the error $s_k^i - \theta_k$. However, for the rank-order correlation the realized θ_k matter, and we take those from the data. (For example, the closer the θ_k are to each other, the worse the rank-order correlation will be, all else constant).