Predicting Experimental Results: Who Knows What?*

Stefano DellaVigna
UC Berkeley and NBER

Devin Pope
U Chicago and NBER

This version: May 25, 2017

Abstract

We analyze how 208 academic 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, citations, academic rank, field, and contextual experience do not correlate with 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 can identify 'superforecasters' among the non-experts who outperform the experts out of sample. Seventh, these results on forecasting accuracy surprise the forecasters themselves. We present and estimate a simple model that organizes our findings.

*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, David Card, Don Moore, Philipp Strack, and Dmitry Taubinsky for their comments and suggestions. We are also grateful to Jesse Shapiro and four referees for very detailed and constructive comments. We are also grateful to the audiences at Bonn University, Frankfurt University, Harvard University, the London School of Economics, the Max Planck Institute in Bonn, MIT, Princeton University, the University of Toronto, the Wharton School, at the University of California, Berkeley, Yale University, 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, Avner Shlain, Alex Steiny, 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.

In this paper, we use data from a large experiment, and associated expert forecasts, designed to provide evidence on the accuracy of expert and non-expert forecasts 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 (forthcoming).

As part of the design, we survey 314 academics, including behavioral economists, other economists, and psychologists. We provide these 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 which include 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. Of the 314 experts contacted, 208 provided a complete set of forecasts.

We frame our analysis with a simple model of forecasts. The model allows different types of forecasters to differ in the bias and variance of their forecasts. We estimate the model via maximum likelihood. Comparing model predictions to observed data features helps make quantitative sense of our findings.

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 effort level with the actual effort level is 0.77.

Our second result is that individual experts are significantly less accurate: 96 percent of forecasters do worse than the average forecast, measuring accuracy with average absolute error across the 15 treatments. The comparison is equally striking using other measures of accuracy like mean squared error. The reason for this 'wisdom-of-crowds' effect is that averaging reduces

the noise in the individual forecasts: taking the average forecast of just 5 experts already leads to a large improvement in accuracy over individual forecasts. Our model matches this large wisdom-of-crowds effect.

So far we have treated experts as interchangeable. Asking the 'right' expert may erase most of the gains from averaging. Our third finding, though, is that none of the expertise measures improves forecasting accuracy. Full professors are, if anything, less accurate than assistant professors and similarly having more Google Scholar citations does not improve accuracy. Thus, *vertical* expertise does not appear predictive of accuracy. Our measure of *horizontal* expertise—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 improve accuracy.

Thus, various measures of expertise do not increase accuracy. Still, it is possible that academics share an understanding of incentives and behavioral forces which distinguish them from the non-experts. Our fourth finding is that, by the measure of accuracy used so far—mean absolute error and mean squared error—the undergraduate and MBA students, and especially the online forecasters are less accurate than the experts. However, undergraduates, MBAs, and even MTurk workers do as well as experts at predicting the *relative* levels of effort across the treatments. 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 MTurk sample is a stunning 0.95 (compared to 0.83 for the experts).

What explains this discrepancy? The data and estimated model show that non-experts, and especially MTurk workers, are more likely to be biased in their guess of the average effort level across the treatments. This bias worsens absolute accuracy but not accuracy in the ordering of forecasts.

Our fifth result is that measures of effort, confidence, and revealed ability can be predictive of accuracy, but with important caveats.

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 less evidence for the other samples, and no impact of forecasters clicking on the trial task, or instructions.

A measure of confidence—the number of forecasts which forecasters expect to get right within 100 points—is predictive of accuracy among PhDs, MBAs, and online workers, but less so for experts. Respondents have some, but imprecise, awareness of their own accuracy.

A third measure—accuracy in the forecast of a simple incentive-based treatment—is highly predictive of accuracy in the other conditions, especially for the non-expert samples. This measure of revealed forecasting ability predicts accuracy also when constructed using other

treatments, suggesting that there is nothing special about the incentive treatment.

Our sixth result is that it is possible to identify 'superforecasters' (Tetlock and Gardiner, 2015) among the non-experts. We do this by linear regression of forecast accuracy on measures of effort, confidence, and revealed forecasting ability, using a K-fold procedure to limit overfitting. 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. 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. Again, our model is able to reproduce these findings quantitatively.

Our seventh and final result addresses a meta-question: Did we know all of this already? We asked the experts to predict the accuracy of different groups of forecasters. The expert beliefs in this regard are systematically off target. Counterfactually, they expect highly cited experts to be more accurate, the field of experts to matter, and PhD students to be less accurate.

To what extent might these results on expertise in forecasting apply to other contexts? At least three features of our design could limit the external validity of our findings. First, the forecasting ability may differ with a task that is less artificial or for which there is a larger body of studies (e.g., the dictator game). Second, in settings with more economic detail, like pricing and supply and demand, or institutional details (e.g., health insurance), the experts could plausibly have an edge in forecasting. Third, forecasters in our setting made predictions taking just a few minutes. While researchers, managers, and policy-makers frequently take quick decisions, in other settings experts spend considerable time deliberating, conducting focus groups, or pilot studies. The expert forecasts in these cases may be more valuable. Future research can hopefully provide a more complete understanding of how expertise impacts forecasting ability.

We explore complementary findings in a companion paper (DellaVigna and Pope, forthcoming), 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.

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.

Our findings are also related to a literature on the quality of expert judgments. The literature in psychology compares expert judgments to algorithms (Meehl, 1954; Dawes, Faust,

and Meehl, 1989) and to decisions of novices. Much of this work has found that, surprisingly, experts are no more accurate than novices, even for tasks such as medical comparisons (Garb, 1989; Camerer and Johnson, 1997). Other work has shown that experience/expertise is helpful. For example, taxi drivers make better decisions over time (Haggag, McManus, and Paci, 2017) and school teachers improve steadily over the first few years of teaching (Jackson, Rockoff, and Staiger, 2014).

There is also a rich literature on forecasts of outcomes other than research results. Within psychology, the Good Judgment Project elicits forecasts by experts on national security topics (Tetlock and Gardner, 2015). We find 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).

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), and regarding the value of aggregating predictions using predictions markets (Wolfers and Zitzewitz, 2004; Snowberg, Wolfers, and Zitzewitz, 2007).

There is a much smaller literature instead on forecasts of future research results. Coffman and Niehaus (2014) report findings from a survey of 7 experts on persuasion. 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 (2016) elicits forecasts on the effect of an RCT from audiences of 4 academic presentations.[1] These studies do not examine the differences between different forms of expertise, or between individual and group forecasts.

The Science Prediction Markets (Dreber et al., 2015 and Camerer et al., 2016) present a more systematic analysis of forecasts of future experimental results. The researchers use a prediction markets and a survey to capture beliefs about the replicability of the findings of dozens of experiments in psychology and experimental economics. Like us, they find that the expert forecasts correlate with the outcome (in their case, replication of the experimental finding). These papers focus on wisdom-of-crowd forecasts, as in our first finding, and do not cover systematically the accuracy of individual experts, the impact of different forms of expertise, or differences between experts and non-experts.

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

[1] 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.

2 Experiment and Survey Design

2.1 Real Effort Experiment

We designed a 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 marketing and psychology (Paolacci and Chandler, 2014) and is also used increasingly in economics (e.g., Kuziemko, Norton, Saez, and Stantcheva, 2015). The evidence suggests that the findings of studies run on MTurk are similar to the results in more standard laboratory or field settings (Horton, Rand, and Zeckhauser, 2011; Amir, Rand, and Gal, 2012; Goodman, Cryder, and Cheema, 2013).

We pre-registered the design of the experiment on the AEA RCT Registry as AEARCTR-0000714, including pre-specifying the rules for the sample size and the inclusion in the sample. The registration also specifies the timing of the experiment and the survey. We ran the experiment first 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 access the experimental results. 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 full results only after the forecasts by the experts were collected (September 2015).

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: '*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 their treatment condition. 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.*' Table 1 reports the key content of this paragraph for all 18 treatments. 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. The final sentence on the page summarizes the condition for earning a bonus (if any) in that particular treatment. Thus, the 18 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 to redeem the earnings.

The experiment ran for three weeks in May 2015. The initial sample consists of 12,838 MTurk workers who started our task. After applying the sample restrictions detailed in DellaVigna and Pope (forthcoming), the final sample includes 9,861 subjects, about 550 per treatment. 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 (Online Appendix Table 1).

2.2 Forecaster Survey

Survey format. The survey, designed to take 15 minutes to complete, is formatted with the online platform Qualtrics and consists of two pages.[2] The first and main page 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 what the MTurkers saw: "*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.*"

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.*"[3] Participants who aim to minimize the sum of squared errors will indicate as their forecast the mean expected effort for each treatment.

[2] The survey is also pre-registered as AEARCTR-0000731.

[3] 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.

The survey then displays the mean effort in the three benchmark treatments: no piece rate, 1-cent, and 10-cent piece rate (Figure 1). 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, the treatments are presented in one of six randomized orders (the only randomization in the survey), always keeping related interventions together.

We decided ex ante the rule for the scale in the slider. To minimize the scope for confusion, 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. A research assistant checked this rule against the results, which led to a score between 1,000 and 2,500.

The second page of the survey elicits a measure of confidence in the stated forecasts. Experts indicate their best guess as to the number of forecasts that they provided that are within 100 points of the actual average effort in a treatment (Online Appendix Figure 1). 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 other groups of experts, such as 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 experts are anonymous in the data set, we use the name to match to information on each expert and to assign the prize.

Sample of Experts. We create an initial list of behavioral experts (broadly construed) consisting of: (i) 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) 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) invitees to the Russell Sage Foundation 2014 Workshop on "Behavioral Labor Economics", (v) behavioral economists in the ideas42 list, and (vi) a small number of additions. We pare down this list of over 600 people to 314 researchers, after excluding graduate students and researchers to whom neither of the authors had any connection (since we did not want to be seen as spamming researchers).

On July 10 and 11, 2015 we sent a personalized contact email to each of the 314 experts, followed by an automated reminder email 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, we 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 Appendix Table 1 document the selection into response.

For each expert, we code four features. As measures of *vertical* expertise we code (i) the academic status from online CVs (Professor, Associate Professor, Assistant Professor, or Other) and (ii) the lifetime citations of a researcher using Google Scholar (as of April 2015). As measures of *horizontal* expertise, we code (iii) the main field of expertise (behavioral economics, applied microeconomics, economic theory, laboratory experiments, and psychology), and (iv) whether the expert has written a paper on the topic of a particular treatment.

In November 2015 we provided personalized feedback to each expert in the form of an email with a personalized link to a figure that included their own individual forecasts. We also randomly drew winners and distributed the prizes as promised.

Other Samples. We also collect forecasts from a broader group: PhD students in economics, undergraduate students, MBA students, and MTurk subjects recruited for the purpose.

The PhD students are from the Departments of Economics at eight schools: UC Berkeley (N=36), Chicago (N=34), Harvard (N=36), Stanford (N=5), UC San Diego (N=4), CalTech (N=7), Carnegie Mellon (N=6), and Cornell (N=19). The MBA students are at the Booth School of Business (N=108) and at Berkeley Haas (N=52). The undergraduate students are at the University of Chicago (N=92) and UC Berkeley (N=66). 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.

We also 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. 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 (making a list of economic blogs) before completing the survey. Both groups were informed that 5 of the workers would randomly win a prize based on the accuracy of their forecasts equal to $100 – Mean Squared Error/2,000. An additional sample of MTurk workers (N= 258) did the same task as the 'experienced' MTurk sample above, but with higher emphasis on the returns to forecasting accuracy: each participant was told they would receive $5 – Mean Squared Error/20,000.

3 Accuracy of Expert Forecasts: Average and Individual

How does the average effort by treatment compare to the expert forecasts? Table 1 lists the treatments, summarized by category (Column 1), wording (Column 2), and sample size

(Column 3). The table also reports for each treatment the average effort (Column 4) and the average forecast by the 208 experts (Column 5), reproduced from DellaVigna and Pope (forthcoming). We display this information in Figure 2, where each of the 18 points represents a treatment, with the average effort on the x axis and the average expert forecast on the y axis. The color-coding groups 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 2 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; 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 1) 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.

Turning to individual experts' performance, our benchmark measure of accuracy is the absolute error in forecast by treatment, averaged across the 15 treatments. We also construct a measure of rank-order correlation between the 15 forecasts and the treatments.

Figure 3a displays the cumulative distribution function of the absolute error for the 208 experts (labeled '*N=1*'), compared to the wisdom-of-crowds error (vertical red line). In this figure and throughout the paper, we show results for the negative of the absolute error, so as to display a measure of *accuracy*. The figure shows that 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 average forecast (169 points vs. 93 points, Columns 1 and 2 in Table 2). This finding is known as 'wisdom of crowds': the average over a crowd outperforms most individuals in the crowd. This finding is similar with rank-order correlation (Figure 3b), squared error and the Pearson correlation coefficient (Online Appendix Figures 2a-b).

How many experts does it take to achieve a level of accuracy similar to the one for the group average? Figures 3a-b also plot the counterfactual accuracy of forecasts averaged over smaller groups of N experts, with $N = 5, 10, 20$. Namely, we bootstrap 1,500 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 tail of high-error forecasts and achieve an average absolute error rate of 114, down from 169 (Column 4 in Table 2). 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 2). The pattern is very similar with rank-order correlation (Figure 3b), squared error, and correlation (Online Appendix Figures 2a-b).

After clarifying the role of group size, we decompose the accuracy by treatment. Online

Appendix Figures 3a-b display two treatments in which the majority of forecasters outperform the average forecast, showing that the wisdom-of-crowds pattern does not apply in each treatment. In other treatments, though, the wisdom-of-crowds forecast is spot on (e.g., Online Appendix Figure 3d). Columns 7 and 8 of Table 1 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 Model and Estimation

Model. We model agent i making forecasts about the results in treatments $k = 1, ..., K$. Let $\theta = (\theta_1, \ldots, \theta_K)$ 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 $f_{i,k}$ 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:

$$s_{i,k} = \theta_k + \eta_k \ + v_i + \sigma_i \epsilon_{i,k} \, . \tag{1}$$

The deviation of the signal $s_{i,k}$ 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) $v_i \sim N\left(\mu, \sigma_v^2\right)$ is a deviation for forecaster i that is common across all treatments (with a possible bias term if $\mu \neq 0$); (iii) $\sigma_i \epsilon_{i,k}$, with $\epsilon_{i,k} \sim N\left(0, 1\right)$ independently of σ_i, is an idiosyncratic noise term.

We assume that the agent is unaware of the systematic bias μ. Given this and the uninformative prior, the signal $s_{i,k}$ is an agent's best estimate (that is, $f_{i,k} = s_{i,k}$), given that it minimizes the (subjective) expected loss $(f_{i,k} - \theta_k)^2$.

The error term $\sigma_i \varepsilon_{i,k}$ captures idiosyncratic noise in the forecasts, with some forecasters providing less noisy forecasts (lower σ_i). 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. 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.

The term η_k allows for differences in the mean forecast across treatments, potentially capturing an incorrect common reading of the literature (or of the context) for a particular treatment, or an unusual experimental finding. The term v_i captures an agent i being more optimistic (or pessimistic) about the effect of all treatments, which we also refer to as the forecaster bias.

Estimation. To simplify the estimation problem, we treat η_k as fixed effects instead of estimating the distribution as a random effect.[4] To estimate the treatment-level effects η_k, notice from (1) that the expected forecast error in treatment k equals $E[s_{i,k} - \theta_k] = \eta_k + E[v_i]$. Thus, to estimate $\hat{\eta}_k$, we first compute the average forecast error for treatment k, $\bar{e}_k = \sum_i (f_{i,k} - \theta_k)/I$, and then we demean it to take out the $E[v_i]$ component. Thus, $\hat{\eta}_k = \bar{e}_k - \sum_k \bar{e}/K$. We estimate separate treatment-level effects η_k for each group of forecasters (academics, PhD students, undergraduates, MBAs, and MTurks.)[5] Using these fixed effects, we define the residual $z_{i,k} = s_{i,k} - \theta_k - \hat{\eta}_k$ and rewrite the model as:

$$z_{i,k} = v_i + \sigma_i \epsilon_{i,k}.$$

We estimate this transformed model with maximum likelihood. Motivated by Heckman and Singer (1984), we allow for discrete heterogeneity in the two key parameters, v_i and σ_i. For our benchmark estimates, we assume that there are 2 (unobservable) types of forecasters: type 1 with $(v^{(1)}, \sigma^{(1)})$, and type 2 with $(v^{(2)}, \sigma^{(2)})$. Since the types are not known, the distribution of $z_{i,k}$ for a given forecaster is described by a mixture of normals. The observables x_i (such as indicators for the group of experts versus the non-experts) predict the likelihood of type 1:

$$p_i^1(x_i) \equiv Pr((v_i, \sigma_i) = \left(v^{(1)}, \sigma^{(1)}\right)) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}.$$

The likelihood takes a convenient form.[6] Let $\theta \equiv [v^{(1)}, v^{(2)}, \sigma^{(1)}, \sigma^{(2)}, \beta]^T$ denote the vector of parameters to estimate. Denoting the standard normal density as ϕ, the likelihood is:

$$Lik[z|\theta] = \Pi_{i=1}^I \Pi_{k=1}^K \{(\frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}) \cdot [\frac{1}{\sigma^{(1)}} \phi(\frac{z_{i,k} - v^{(1)}}{\sigma^{(1)}})] + (\frac{1}{1 + e^{x_i^T \beta}}) \cdot [\frac{1}{\sigma^{(2)}} \phi(\frac{z_{i,k} - v^{(2)}}{\sigma^{(2)}})]\}.$$

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

We assume two types in our model specifications, defining type 1 as the one with v closer to zero. Column 1 in Table 3 reports the benchmark estimate, using the data for all groups of forecasters, and including as control variables x_i just the indicators for the 4 groups (plus the omitted category). Online Appendix Figure 4 displays the estimated $(\hat{v}, \hat{\sigma})$ for the two

[4] Above, we assume $\eta_k \sim N(0, \sigma_\eta^2)$ to ensure that the optimal forecast is the signal $s_{i,k}$. Instead of estimating σ_η^2, we use the realized $\hat{\eta}_k$ as fixed effects, simplifying the estimation considerably.

[5] To operationalize this, we regress the demeaned forecast errors on the complete set of treatment dummies, so that the estimated fixed effects have mean zero by construction. We then construct $z_{i,k}$ by summing the residuals from this regression and the mean forecast error. In order to capture differences in these treatment fixed effects across different groups of forecasters, we estimate this regression separately for each group of forecasters (faculty, PhDs, MBAs, undergraduates and Mturkers), demeaning the forecast error using the group-specific means.

[6] More generally, allowing for more types, the probability of types is distributed multinomial logit, with separate β's for each type (except for the omitted type).

types. The first type has a small estimated average bias $\hat{v}^{(1)} = -24.9$, and a relatively small idiosyncratic standard deviation $\hat{\sigma}^{(1)} = 162.6$. The second type instead has a large average forecast bias $\hat{v}^{(2)} = -193.2$, and an idiosyncratic standard deviation which is more than twice as high, $\hat{\sigma}^{(2)} = 357.6$. Both sets of estimates are highly precise. Thus, the first type can be interpreted roughly as the "good" type, since the forecasts are closer to the truth and have lower variance on average.

Online Appendix Figure 4 reports the share of the two types that are implied by the estimated coefficients on the types, $\hat{\beta}$. For the experts, the share of the good type is $p^1 = e^{-0.74+2.84} / \left(1 + e^{-0.74+2.84}\right) = 0.89$, and similarly for the PhD students. The share of type-1 forecasters is lower for MBAs and undergraduates, and is only 0.32 for the MTurk sample, matching the fact that a sizable share of MTurk forecasters forecast too low an effort across the different treatments.

How well does this simple model match the facts? In Figures 3c-d we display evidence for the experts using simulated data for the model estimates in Column 1 of Table 3. The model fits quite well the distribution of individual accuracy, the wisdom-of-crowd accuracy, and the speed of convergence when using draws of 5, 10, or 20 simulated forecasters.

Online Appendix Table 4 displays the fit of this model (reproduced in Column 1) for several key moments, such as the individual absolute error, the wisdom-of-crowd error, the rank order correlation, and the correlation across treatments in the absolute error. The table also displays the estimates, and quality of fit, of alternative models: 2-type models with heterogeneity only in σ_i or only in v_i, a 1-type model and a 3-type model (using the same variables x_i as the predictors of type). Among the 2-type models, having no heterogeneity in the average bias v_i lowers the quality of the fit significantly, as the model no longer explains the bias among the non-experts. The fit is better with a 2-type model with no heterogeneity in idiosyncratic variance σ_i, though this model still does not do as well as the benchmark. A 1-type model with no heterogeneity does poorly, as it cannot capture the differences between experts and non-experts. A 3-type model improves the fit as it can reproduce a larger bias in forecast among some of the non-experts. However, it does not much improve the qualitative fit of the moments (Panel B) and it has much worse numerical convergence properties. As such, we employ as benchmark the simpler 2-type model with heterogeneity in both σ_i and v_i, and we return to it below to display how closely this model mirrors additional empirical findings.

5 Determinants of Forecast Accuracy

5.1 Measures of Expertise

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 4a, 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.

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

$$a_{i,k} = \alpha + \beta X_{i,k} + \zeta_k + \lambda_{o(i,k)} + \varepsilon_{i,k} \tag{2}$$

An observation is a forecaster-treatment combination, and the dependent variable is a measure of accuracy $a_{i,k}$ for forecaster i and treatment k, such as the negative of the absolute error in forecast. The key regressors are the expertise variables $X_{i,k}$. The regression also includes treatment fixed effects ζ_k, as well as fixed effects for the order $o(i,k) = 1, ...15$ in which the treatment is presented, to control for forecaster fatigue. The term $o(i,k)$ is identified because there are six possible orders of presentations of treatments. 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).

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 4b presents a split of the expert sample into three groups based on citations. The split has some overlap with the academic rank, but there is plenty of independent variation. The evidence suggests a perverse effect of citations: the least-cited group of experts has the highest forecasting accuracy.

Thus, there is no evidence that vertical expertise improves the forecasting accuracy and some evidence to the contrary. One interpretation of this result is that prominent experts have a very high value of time and thus put less time and effort into the survey. In Columns 2 and 4 we add controls for effort, discussed in detail in a later section. Adding these controls does not change the point estimates at all. This is not surprising, since 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). The coding is admittedly subjective, but at least was done before the data analysis. 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 4c displays the results: the differences between the groups, if any, are small. Controlling for citations and academic rank (Column 3 of Table 4) and further controlling for effort (Column 4), there is similarly no evidence of differences by field of expertise.

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. In this specification (Column 5), 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 9 points, just 5 percent of the average absolute error.

As a final measure of horizontal expertise we test whether PhD students who self-report specializing in behavioral economics have higher accuracy. Online Appendix Figure 5 shows 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.

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 4d we compare the accuracy of the two sub-samples of experts. The experts are indistinguishable with respect to absolute forecast error, as Columns 3 and 4 of Table 4 also show.

Model. Column 4 of Table 3 reports the maximum-likelihood estimates of the two-type model restricted to the sample of experts, including as controls x_i the expertise measures and the controls for effort. The results largely match the ones in the reduced-form evidence: tenured professors are less likely to be of the 'good' type, and field affiliation and contextual expertise do not help much, if at all (see also Online Appendix Figures 9a-d). In this specification with just the experts, the estimated parameters for the two types indicate more limited heterogeneity between the two types, especially in the bias term v: this makes sense, since very few experts display large systematic biases in the average forecast.

5.2 Non-Experts

Do non-experts make worse forecasts than experts? Figure 5a shows that the distribution of absolute error is quite different for 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. In this specification, we also split the MTurk sample by a (self-reported) measure of education. The MTurkers with a college degree have a higher accuracy, though still lower than the one of undergraduates or MBAs. In Column 2, we show that controlling for measures of effort reduces the differences in accuracy between the groups, but the difference between the experts on the one hand and the MBAs and MTurk forecasters remains substantial. Thus, when making forecasts about magnitudes of the experimental findings, experts are indeed more accurate than non-experts.

Yet, while the above measures of accuracy were the main ones we envisioned for this study and the ones we specified in the pre-registration, they are not always the relevant ones. Policy-makers 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 is to get the *order* right. We thus revisit the results using the Spearman rank-order correlation as the measure of accuracy.[7] We correlate the ranking of the 15 treatments implied by the forecasts with the ranking implied by the actual average MTurk effort.

The rank-order correlation drastically changes the comparison with the non-experts. By the rank accuracy measure (Figure 5b), undergraduates, MBAs, and even MTurk workers do about as well as the experts (and PhD students do better). Across these samples, the average individual rank-order correlation with the realized effort is around 0.4 (Table 2, Panel B).

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 (and PhD students have significantly higher accuracy than the experts according to this measure).

This result is striking because non-experts spend significantly less effort on the task as measured by time spent and click-through on instruction (Appendix Table 1). Controlling for these effort measures therefore improves slightly the relative performance of the online sample (Column 4 of Table 5).

This evidence so far concerns the accuracy of individual forecasters. With respect to the wisdom-of-crowds measures, MBA students and especially MTurk workers display worse accuracy than experts with respect to absolute error (Table 2, Panel A). With respect to the rank-order measure (Table 2, Panel B), 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

[7] We deduce the ranking of treatments from the forecasts in levels. We thank seminar audiences and especially Katy Milkman for the suggestion to use rank-order correlation as an additional measure of accuracy.

0.83 for the experts).[8]

What explains the 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 choose too low a level of effort on average, perhaps because the sliders (which they had to move) were centered on the left. This pattern is visible in Online Appendix Figures 6b-d for the average forecast, and at the individual level in Online Appendix Figure 7a. 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 vertical black 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.

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.[9] This is consistent with psychological evidence suggesting that people struggle with absolute judgments, but are better at making relative judgments (Miller, 1962, Laming, 1984, 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.

One may also wonder if the rank-order correlation changes the results in the previous section on vertical, horizontal, and contextual expertise of experts. In Online Appendix Figures 8a-d, we show that this is not the case.

Model. Can the model make sense of the difference between the absolute error measure and the rank-order correlation? Figures 5c and 5d, generated using the parameter estimates in Column 1 of Table 3, show that we reproduce quite closely the observed patterns in the data. Not surprisingly, the two-type model produces more bimodality than observed in the data for the MTurk sample, but otherwise the qualitative patterns are quite close.

5.3 Other Correlates of Accuracy

Effort. A key variable that is likely to impact the quality of the forecasts is the effort put into

[8] One might wonder whether this higher correlation is due to the larger sample size for MTurks. To get at this question, we randomly draw 10,000 samples of 208 MTurks with replacement repeatedly and calculate the rank-order correlation for each draw. The average rank-order correlation is 0.940, suggesting that the higher rank-order correlation is not due to the larger sample size for the MTurk forecasters.

[9] To further document whether the forecaster bias is a reason for the discrepancy, we explore the Pearson correlation between the individual forecasts and the average results. The correlation measure is based on levels, as opposed to ranks, but it does not measure whether the level of effort is matched. If non-experts mainly differ from experts in a level offset, they should be similar to experts according to simple correlation, as indeed shown in Panel B in Online Appendix Table 2.

the survey. While effort is unobservable, we collect two proxies that are likely to be indicative. The first measure is the time taken from initial login to the survey to survey completion.[10] We cap this measure at 50 minutes, about the 90th percentile among experts, since participants who took very long (sometimes returning to the survey after hours or days) might have been multi-tasking. The average time taken is 21 minutes among the experts, the PhD students and the MBA students, and lower in the other samples (Appendix 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, 12, and 0 percent among undergraduates, MBAs, and MTurk workers respectively.[11] 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 (mean absolute error) as a function of time taken (Figure 6a). Forecasters taking less than 5 minutes do significantly worse in both the student and online sample (no expert falls in this category). More surprisingly, there is not much difference in accuracy between forecasters taking 5-9 minutes and forecasters taking longer, both among the experts and among the students (though in the online sample, the group taking 10-14 minutes does better than the group taking 5-9 minutes). There is some evidence of decline for individuals taking longer than 25 minutes, likely due to multi-tasking. There is a similar pattern with rank-order correlation (Online Appendix Figure 10a).

How well can the model fit this pattern? We estimate the model on the joint sample including indicators for the different groups, controls for the duration taken for the survey, as well as controls for confidence introduced below (Column 2 of Table 3). The model restricts the coefficients on completion time to be the same for all three groups, so it is not obvious that the model predictions will match patterns in the data closely. Nonetheless, Figure 6b and Online Appendix Figure 10b show that the simulated data based on the model estimates reproduce quite well the patterns in the data.

We then turn to the 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. Online Appendix Figures 11a-b show no obvious difference in accuracy for individuals who do, or do not, click on such instructions. In Online Appendix Table 6 we 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. It seems plausible that individuals who are more enthusiastic about the survey complete it sooner and with more effort. This variable has no obvious effect.

[10] It is possible that, to the opposite, longer time taken denotes lower skill. This is less likely an interpretation for respondents taking a very short time (e.g., less than 5 minutes).

[11] For 37% of MBAs, we believe the links to click on practice and instructions malfunctioned during the survey, leading to no recorded clicks. In regressions, we include an indicator for missing click data.

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?

In Online Appendix Figures 11c-d and in Columns 4 and 8 of Online Appendix Table 6 we present an attempt to exogenously induce higher forecasting effort. We recruit a group of 250 MTurkers with increased incentives for accuracy in forecasting. Namely, we pay *each* survey participant a sum up to \$5 for accuracy, computed as \$5-MSE/20,000. This payment is higher than the promise to randomly pay two of the MTurk workers in the other sample an accuracy bonus up to \$100. In addition, we made the reward for accuracy more salient (see Section 2). The higher incentives had 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 examine 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. Figures 7a-b report the average accuracy for the three groups—experts, students, and MTurk workers—as a function of the confidence level from 0 to 15. We document the impact on absolute error (Figure 7a), on the number of forecasts (out of 15) within 100 points of the actual average effort (Figure 7b), and on the rank-order correlation (Online Appendix Figure 12a). The corresponding regression results are in Online Appendix Table 7.

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, perhaps because we elicited confidence using a cardinal, not ordinal, measure of accuracy. Online Appendix Figure 7c 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.

Figures 7c-d show that the simulated data from the model estimates including (linearly) the confidence measure (Column 2 in Table 3) provides a good fit to the data.

Revealed Accuracy. If there are differences in forecasting skill, forecasters who are more accurate in one treatment are likely to be more accurate in other treatments as well. We thus examine the correlation of accuracy across treatments, avoiding extrapolation across very similar treatments: the result in these treatments will presumably be correlated, inducing a mechanical correlation in accuracy.

To start, we consider a unique treatment within the experiment: the 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

has 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 (forthcoming), it is possible to predict the effort in the 4 cent treatment accurately. We 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/willingness to perform a simple calibration mentally. 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 Figure 8a, we plot the average accuracy for the three groups of forecasters as a function of deciles in the accuracy of forecasting the 4-cent treatment, omitting the 4-cent treatment in constructing the accuracy measures for related plots. The correlation is strong: 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 a large gap in accuracy of over 100 points for the bottom deciles. This correlation between accuracy in the 4-cent treatment and accuracy in other treatments is more muted with rank-order correlation (Online Appendix Figure 13a).

Can the model reproduce these findings? We estimate a model adding the absolute forecast error in the 4 cent treatments (Column 3 in Table 3), obviously excluding the 4-cent treatment from the observations. The simulations using the point estimates once again reproduce quite well the observed patterns (Figure 8b and Online Appendix Figure 13b).

Table 6 displays the regression-based evidence, including all the controls: vertical expertise and field of the experts (just for the expert regression in Column 1), time to survey completion and the confidence level. Even with these controls, the 4-cent variable has substantial explanatory power: an increase of 100 points in the accuracy of the 4-cent prediction increases the accuracy in the other treatments by an average of 9.6 points for the experts (Column 1), 23.9 points for the students (Column 2) and 31.1 points for the MTurks (Column 3). We experimented with non-linear specifications in the 4-cent accuracy, but a linear specification captures the effect of the variable well. Introducing the revealed-accuracy control generally reduces the load on the other variables, though confidence remains a significant predictor.

Next, we examine whether there is something special about the 4-cent treatment when it comes to capturing 'revealed accuracy'. In Online Appendix Table 8 we constructed an accuracy variable based on one group of treatments, and use it to predict accuracy in the forecasts of other treatments. Interestingly, almost all measures are helpful to predict accuracy in other treatments (omitting treatments that are variations of the variable used for 'revealed accuracy'). 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. Thus, the critical component is not 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.

5.4 Superforecasters

As we have seen in Section 5.2, 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 match the accuracy of the expert sample using non-experts (who tend to be more available)?

In our context, to identify 'superforecasters' (Mellers et al., 2015 and Tetlock and Gardner, 2015) we use the variables examined so far: measures of expertise, effort, confidence, and revealed accuracy. As Section 5.3 shows, the revealed accuracy measure (which is in spirit of using the track record of a forecaster) is especially predictive of forecasting accuracy. We thus take the same specification as in Table 6, with all these control variables, and for each sample 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.

Online Appendix Table 9 reports the results for individual accuracy (Column 1) and average accuracy for groups of 20 experts (Column 3) and 50 experts (Column 4). The optimal 20% of experts constructed using all controls does not do better than the overall sample of experts. In the sample of PhD students, MBAs, and undergraduates, instead, the optimal 20% of forecasters outperforms the academic experts both at the individual level (Figure 9a) and with the wisdom-of-crowds measure (Figure 9b). Indeed, the wisdom-of-crowds absolute error for the top 20% in this group is as low as 76 points for groups of 20 forecasters, compared to 101 points for the average expert. Figure 9b displays the results for the wisdom-of-crowds measure for bootstrapped samples of 20 forecasters.

The results are equally striking for the online sample. 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 *outperform* the academic experts, achieving an accuracy of 81, compared to 101 for the experts. 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. Furthermore, forecasts by the non-expert samples are much cheaper and easier to obtain: one can easily sample a couple hundred online forecasters and then extract the 'superforecasters'. In comparison, getting even a dozen expert forecasts on a systematic basis may be hard.

We provide a model-based parallel to this result. We estimate a model similar to the one in Table 6, with all controls, in Column 3 of Table 3. Using simulations from data sets drawn for the estimated parameters, we evaluate the accuracy of superforecasters (defined as forecasters in the top 20% of the probability of being the "good type") in Figures 9c-d. Once again, we mirror quite closely the empirical findings.

For these results, an important role is played by the fact that the different groups (such as academics versus MTurks) have different estimated treatment-level effects $\hat{\eta}_k$. Online Appendix Figures 14a-b show that, if we force the treatment-level effects $\hat{\eta}_k$ to be the same across groups, the model does not match the finding that the 'superforecaster' students and MTurks do better than the experts. An important component of the model fit is the larger idiosyncratic treatment-level error $\hat{\eta}_k$ for the experts in treatments such as the very-low-pay treatment.

5.5 Beliefs about Expertise

Our seventh and final result addresses a meta-question: Did we know all of this already? Perhaps it was expected that, for example, vertical and horizontal expertise would not matter for the quality of forecasting in our task.

On the second page of the survey we elicited the expected accuracy for different groups of forecasters (Online Appendix Figure 1). Specifically, we asked for the expected number of treatments that an individual from a particular group would guess within 100 points of 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 10 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. 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 overplacement (Healy and Moore, 2008).

Next, we consider 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 expect experts with a psychology PhD to be more accurate where the data points if anything in the other direction. They expect that PhD students would be significantly less accurate, counterfactually.

The experts also expect that the PhD students with expertise in behavioral economics would do better, which we do not find.[12] The experts do 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.[13]

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.

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 experts in the traditional sense. 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 present several 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 attempt to identify superforecasters among the non-experts using measures of effort, confidence, and accuracy on a trial question.

The results stress what we hope is a message from this paper. As academic economists 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.

[12] 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.

[13] The group of MTurk workers who first experience the task has an absolute error that is 24 points higher than the group which did not experience the task before making the forecasts (Online Appendix Table 6).

References

[1] Amir, On, and Dan Ariely. 2008. "Resting on Laurels: The Effects of Discrete Progress Markers as Subgoals on Task Performance and Preferences." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34 (5): 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] Ben-David, Itzhak, John Graham, and Cam Harvey. 2013. "Managerial Miscalibration." *Quarterly Journal of Economics* 128 (4): 1547-1584.

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

[5] Camerer, Colin F., et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351 (6280): 1433-1436.

[6] Camerer, Colin F. and Johnson, and Eric J. 1997. "The process-performance paradox in expert judgment: how can experts know so much and predict so badly." *Research on judgment and decision making: currents connections, and controversies*, 342.

[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] Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical versus actuarial judgment." *Science* 243 (4899): 1668-1674.

[10] DellaVigna, Stefano and Devin Pope. Forthcoming. "What Motivates Effort? Evidence and Expert Forecasts" *Review of Economic Studies*.

[11] 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* 112 (50): 15343-15347.

[12] Erev, Ido, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan M. Herzog, Robin Hau, Ralph Hertwig, Terrance Stewart, Robert West, and Christiane Lebiere. 2010. "A Choice Prediction Competition: Choices from Experience and from Description." *Journal of Behavioral Decision Making* 23 (1): 15-47.

[13] Galton, Francis. 1907. "Vox Populi." *Nature* 75 (7): 450-451.

[14] Garb, H.N. 1989. "Clinical judgment, clinical training, and professional experience." *Psychological Bulletin* 105: 387-396.

[15] 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 (3): 213-224.

[16] Groh, Matthew, Nandini Krishnan, David McKenzie, and Tara Vishwanath. 2016. "The Impact of Soft Skill Training on Female Youth Employment: Evidence from a Randomized Experiment in Jordan." *IZA Journal of Labor and Development,* 5 (9): 1-23.

[17] Haggag, Kareem, Brian McManus, and Giovanni Paci. 2017. "Learning by Driving: Productivity Improvements by New York City Taxi Drivers." *American Economic Journal: Applied Economics* 9 (1): 70-95.

[18] Heckman, James, and Burton Singer. 1984. "A method for minimizing the impact of distributional assumptions in econometric models for duration data." *Econometrica* 52 (2): 271-320.

[19] 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* 82 (2): 430-452.

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

[21] Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher effects and teacher-related policies." *Annual Review of Economics* 6 (1): 801-825.

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

[23] 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.

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

[25] Meehl, Paul E. 1954. "Clinical versus statistical prediction: a theoretical analysis and a review of the evidence." Minneapolis: University of Minnesota Press.

[26] Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions." *Perspectives on Psychological Science* 10 (3): 267-281.

[27] 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.

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

[29] Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." *Science* 349 (6251): aac4716.

[30] Paolacci, Gabriele, and Jesse Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23 (3): 184-188.

[31] 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.

[32] 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.

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

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

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