# Predicting Experimental Results: Who Knows What?*

Stefano DellaVigna
UC Berkeley and NBER

Devin Pope
U Chicago and NBER

This version: April 27, 2017

**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 in future studies.

*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 the audiences at Bonn University, Frankfurt University, Harvard University, the London School of Economics, the Max Planck Institute in Bonn, MIT, 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.

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, rely on the wrong experts, or conversely under-utilize 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 314 academics, including behavioral economists, standard 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, 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 the impact of expertise on forecasts. In addition to these experts, we also survey 147 PhD students, 158 undergraduate students, 160 MBA students, and 762 workers from the online platform for 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.

A policy-maker, a firm, or an advisee, though, will typically have the opinion of just one expert, or a few experts. How do individual experts do? 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.

What explains this large 'wisdom-of-crowds' effect? Averaging removes the idiosyncratic noise in the individual forecasts. Given that in our setting the average forecast does so well, the mean outperforms nearly every individual expert. Taking the average forecast of just 5 experts already leads to a large improvement in accuracy over using individual forecasts.

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 increase 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. 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 the context (the online sample).

Are forecasts by non-experts less accurate? The answer, our fourth finding, depends on the definition of accuracy. 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.

Yet, while the above measures of accuracy were the main ones we envisioned, they are not always the relevant ones. In our motivating examples, the policy-maker, the businessperson, and the advisee may be looking for the most effective treatment, or for ways to weed out the least effective ones. From this perspective, getting the *order* of treatments right is more important than getting the *levels* right. We thus revisit the results using the rank-order

correlation between the forecasts and the experimental effort as the measure of accuracy.

Rank-order correlation does not change the findings on vertical, horizontal, or contextual expertise: the three forms of expertise do not help academics rank treatments better. However, this metric changes the comparison between experts and non-experts: undergraduates, MBAs, and even MTurk workers do as well as experts at ranking 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 online sample is a stunning 0.95 (compared to 0.83 for the experts).

What explains this discrepancy? The non-experts, and especially the online sample, are 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 relative judgments (Laming, 1984; Kahneman, Schkade, and Sunstein, 1998).

Expertise, overall, does not help much with forecast accuracy. Are there other determinants, then, of accuracy? Our fifth result is that measures of effort, confidence, and revealed ability can be predictive of accuracy, but with important caveats. The predictability mostly holds among non-experts and is stronger for absolute error than 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 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.

Thus, while ex ante proxies of expertise are not helpful in our setting, measures of effort, confidence, and especially revealed forecasting ability are generally predictive of accuracy. Can these measures help identify 'superforecasters' (Tetlock and Gardner, 2015)? We use linear regressions with a K-fold method to obtain out-of-sample predictions of accuracy.

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 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.

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. Forecasters also differ in an average bias (offset) level, across the treatments.

We estimate the model with maximum likelihood, allowing for two unobserved types. We estimate that a first, 'good' type has less bias and lower idiosyncratic variance, while a second type offers too low a forecast and has higher variance. The share of the 'good' type differs across samples: it is higher among experts and PhD students, and lower among non-experts. This simple model matches quite well the findings, reproducing the wisdom-of-crowd results, the difference between absolute error and rank order, and the superforecasting results.

To what extent might these results on expertise in forecasting apply to other contexts? At least three features of our design could affect the external validity. 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, 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.

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 multi-disciplinary 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) 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) elicits forecasts on the effect of an RCT from audiences of 4 academic presentations.[1] These studies, while providing valuable insights, do not examine the differences between different forms of expertise, or between individuals versus wisdom-of-crowds.

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.[2]

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] Our work also also relates to the literature on transparency in the social sciences (e.g., Simmons, Nelson, and Simonsohn, 2011; Vivalt, 2015).

# 2 Experiment and Survey Design

## 2.1 Real Effort Experiment

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 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).

The limited cost per subject and large available population on MTurk allow us to run several treatments, each with a large sample size. This platform also makes it possible for the experts to sample the task and to easily compare the different treatments, since the instructions for the various treatments differ essentially in only one paragraph.

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.[3] 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, 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). 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).

### 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.[4] 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*

[3]For space reasons, in Table 1 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.

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

*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] Participants who aim to minimize the sum of squared errors 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 (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 (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)

[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.

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. Notice that the respondents are kept anonymous.

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. These participants may have a better sense than academics 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. 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 (2016). 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; the blue line displays the best interpolating line which has a slope of 0.53 (s.e. 0.12). 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 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.

Thus, the average forecast 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?

The benchmark measure of accuracy for the individual expert is the absolute error in forecast by treatment, averaged across the 15 treatments.[6] 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). 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

[6] 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*.

is similar with rank-order correlation (Figure 3b), squared error and the Pearson correlation coefficient (Online Appendix Figure 1).

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, squared error, and correlation.

After clarifying the role of group size, we decompose the accuracy by treatment. Online Appendix Figures 2a-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 2d). 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.** Can a simple model make sense of these findings and organize the ones to come? 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 $\theta_k$. 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 $\theta_k$ consists of three components, each i.i.d. and independent from the other components: (i) $\eta_k \sim N(0, \sigma_\eta^2)$, the treatment effect, is a deviation for treatment $k$ that is common to all forecasters; (ii) $v_i \sim N\left(\mu, \sigma_v^2\right)$, the forecaster effect, 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 $\epsilon_{i,k} \sim N\left(0, 1\right)$ and $\sigma_i$ is independent from $\epsilon_{i,k}$, is the idiosyncratic noise component, with heterogeneous $\sigma_i$: more accurate forecasters have a lower $\sigma_i$.

We assume that the agent is unaware of the systematic bias $\mu$. 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 $(f_{i,k} - s_{i,k})^2$.

The error term $\sigma_i \varepsilon_{i,k}$ captures idiosyncratic noise in the forecasts, with some forecasters providing less noisy forecasts (lower $\sigma_i$). If $\sigma_i$ 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 $\sigma_i$ than other forecasters, there will be cross-treatment predictability: the forecasters who do well in one treatment are likely to have low $\sigma_i$, and thus do well in another treatment too.

Why do we need the additional error terms $\eta_k$ and $v_i$? A model with just the idiosyncratic error term misses two important features of the data. First, some treatments may have aggregate forecast errors, such as the very-low-pay treatment (Figure 2 and Table 1). The term $\eta_k$ 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 (as we document more later). The term $v_i$ captures an agent $i$ being more optimistic (or pessimistic) about the effect of all treatments, which we also later refer to as the bias of the forecaster.

**Estimation.** We estimate this simple model with maximum likelihood. To simplify the estimation problem, we treat the treatment effects $\eta_k$ as fixed effects instead of estimating the distribution as a random effect.[7] To estimate the $\eta_k$ fixed effects, 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, $\hat{\eta}_k = \bar{e}_k - \sum_k \bar{e}/K$.[8] Using these fixed effects, we define the residual $z_{i,k} = s_{i,k} - \theta_k - \hat{\eta}_k$ and rewrite the model as:

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

For the estimation, motivated by Heckman and Singer (1984), we allow for discrete heterogeneity in the two key parameters, $v_i$ and $\sigma_i$. 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 $(v^{(2)}, \sigma^{(2)})$, with $p^1$ denoting the share of the first type. For a given type, $z_{ik}$ is normally distributed with mean $v$ and variance $\sigma^2$. Since the types are not known, the distribution of

[7] With just 15 treatments, the distribution of the $\eta_k$ random effects would be estimated with limited precision.

[8] 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.

$z_{i,k}$ 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.[9] 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 $\phi$, 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 matrix of the estimated parameters is then given by $Var(\theta) = I(\theta) = -E[\frac{\partial^2 Loglik(\theta)}{\partial \theta \partial \theta^T}]$, where $I(\theta)$ is the information matrix, and we can estimate this quantity via the plug-in principle (i.e. using sample analogues).

We assume two types in our model specifications, defining type 1 as the one with $v$ closer to zero.[10] 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). Figure 4 displays the estimated $(\hat{v}, \hat{\sigma})$ for the two 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.

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 2 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

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

[10] If the types were in the reverse order, we would relabel $(v, \sigma)$ accordingly and reverse the sign of $\beta$.

$\sigma_i$ 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, Column (3) shows that having no heterogeneity in the average bias $v_i$ lowers the quality of the fit significantly, as the model can no more explain the bias among the non-experts. The fit is better with a 2-type model with no heterogeneity in idiosyncratic variance $\sigma_i$ (Column 2), though this model still does not do as well as the benchmark. A 1-type model with no heterogeneity (Column 4) does poorly, as it cannot capture the differences between experts and non-experts. A 3-type model (Column 5) improves the fit quantitatively 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 convergence properties. As such, we employ as benchmark the simpler 2-type model with heterogeneity in both $\sigma_i$ 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

In Section 3, we treated the 208 experts as interchangeable, and studied the implications of averaging expert forecasts versus following an individual expert. But clearly the experts in our sample differ in important ways, such as 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. 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.

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

$$a_{i,t} = \alpha + \beta X_i + \eta_t + \lambda_{o(t)} + \varepsilon_{i,t} \tag{2}$$

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 $\eta_t$, as well as fixed effects for the order $o(t) = 1, ...15$ in which the treatment is presented, to control for forecaster fatigue.[11] 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 5b 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).[12] 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 5c 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,

[11] 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.

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

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 3 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. 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 5d 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.** Columns 6 and 7 of Table 3 report the maximum-likelihood estimates of the two-type model restricted to the sample of experts, including as controls $x_i$ the expertise measures, as well as (in Column 7) the controls for effort. The results are largely similar to the ones in the reduced-form evidence: tenured professor are less likely to be of the 'good' type, and field affiliation and contextual expertise do not help much, if at all. It is interesting to note that 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 make sense, since very few experts display large systematic biases in the average forecast.

### 5.2 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 6a 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[13], 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 is to get the *order* right. We thus revisit the results using the Spearman rank-order correlation as the measure of accuracy.[14] 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 6b), 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 improves slightly the performance of the online sample (Column 4).

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 (Panel A). With respect to the rank-order measure (Panel B), though, the MTurk workers in fact do better than the experts, displaying a

[13] 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).

[14] 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.

stunning wisdom-of-crowds rank-order correlation of 0.95 (compared to 0.83 for the experts).[15]

This pattern is visible in Figure 6e, which shows how well the average forecast of the MTurkers ranks the treatments, despite being off in levels. Undergraduates and PhDs do almost as well, with MBAs doing somewhat worse than the experts (Appendix Figures 2a-c). Overall, 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 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 Figure 6e and Appendix Figures 2b-c for the average forecast, but is also displayed at the individual level in Online Appendix Figure 4a. 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 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 D in Table 2.

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.

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 5a-d,

[15]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.

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 6c and 6d, 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

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 indicative. The first measure is the time taken from initial login to the survey to survey completion.[16] 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.[17] 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 7a). 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 6a).

[16] 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).

[17] 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.

How well can the model fit this pattern? We estimate the model on the joint sample including indicators for the different groups, as well as controls for the duration taken for the survey (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 7b and Online Appendix Figure 6b 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 7a-b show no obvious difference in accuracy for individuals who do, or do not, click on such instructions.[18] In Online Appendix Table 4 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.

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 7c-d and in Columns 4 and 8 of Online Appendix Table 4 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 8a-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 8a), on the number of forecasts (out of 15) within 100 points of the actual average effort (Figure 8b), and on the rank-order correlation (Online Appendix Figure 8a). The corresponding regression results are in Online Appendix Table 5.

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

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

accuracy. Online Appendix Figure 4c 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 8c-d show that the simulated data from the model estimates including (linearly) the confidence measure (Column 3 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 experimental design: 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 (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/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 9a, 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 9a).[19]

Can the model reproduce these findings? We estimate a model adding the absolute forecast error in the 4 cent treatments (Column 4 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 9b and Online Appendix Figure 9b).

[19]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 (Online Appendix Figure 4d).

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 6 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 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 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.

Table 7 reports the results for individual accuracy (Column 1) and average accuracy for groups of 20 experts (Column 3) and 50 experts (Column 4). Panel A compares the overall group of academic experts to the optimal 20% of experts constructed using all controls, as well as the optimal 20% constructed using all controls other than the revealed-ability variable. These super-experts do not do better than the overall sample.

In the sample of PhD students, MBAs, and undergraduates (Panel B), instead, the optimal 20% of forecasters outperforms the academic experts both at the individual level (Figure 10a) and with the wisdom-of-crowds measure.[20] 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 (Column 3). Figure 10b 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 5 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 10c-d. Once again, we mirror quite closely the empirical findings.

[20] 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.

### 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 (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 11 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.[21] The experts also expect that the PhD students with expertise in behavioral economics would do better, which we do not find.[22] 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.[23]

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.

[21]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.

[22]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.

[23]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 4).

# 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.

# 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.* Vol. 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] 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] Camerer, Colin F. and Johnson, 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.

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

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

[10] Dawes, Robin M., Faust, D., & Meehl, P.E. 1989. "Clinical versus actuarial judgment." *Science*, 243, 1668-1674.

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

[12] 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.

[13] 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: 15-47.

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

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

[16] 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.

[17] 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.

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

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

[20] Heckman, James and B. Singer. 1984. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data" *Econometrica*, Vol. 52, pp. 271-320.

[21] 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.

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

[23] 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.

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

[25] Jackson, Rockoff, and Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics*, 6.

[26] 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.

[27] 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.

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

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

[30] 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.

[31] 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.

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

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

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

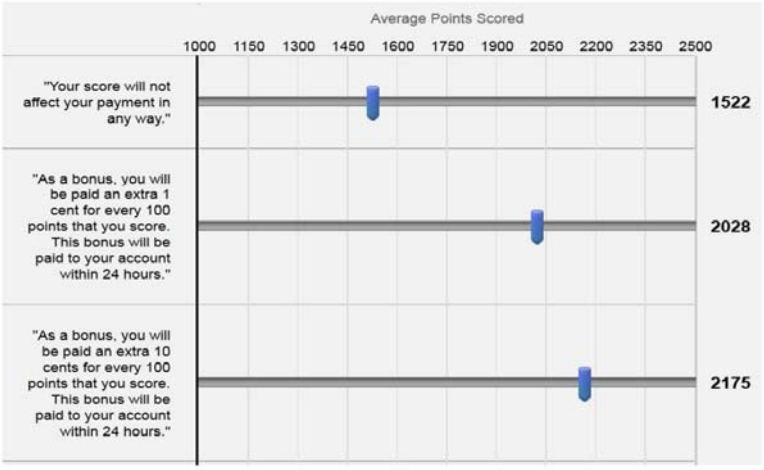
[40] Tetlock, Philip E., Dan Gardner. 2015 *Superforecasting: The Art and Science of Prediction*, Random House.

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

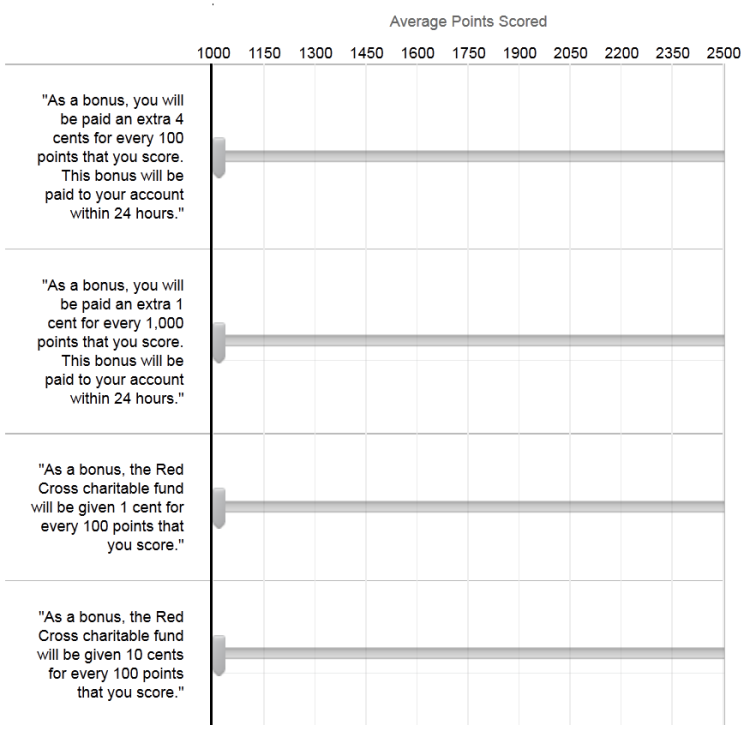
[42] Wolfers, Justin, Zitzewitz, Eric. 2004. "Prediction Markets" *The Journal of Economic Perspectives*, Vol. 18 (2), pp. 107-126.

## Figure 1. Expert Survey, Screenshots from Page 1 of Survey

**Results to Help Guide Your Predictions**
Below are the actual results from 3 of the 18 conditions. On the left, you can see the wording for each of the conditions exactly how it was shown to the MTurk participants. On the right, you will see a slider scale that indicates the average points scored for the first three conditions. The results from these three conditions can be used as a guide to help you know how effort might change with different bonuses.

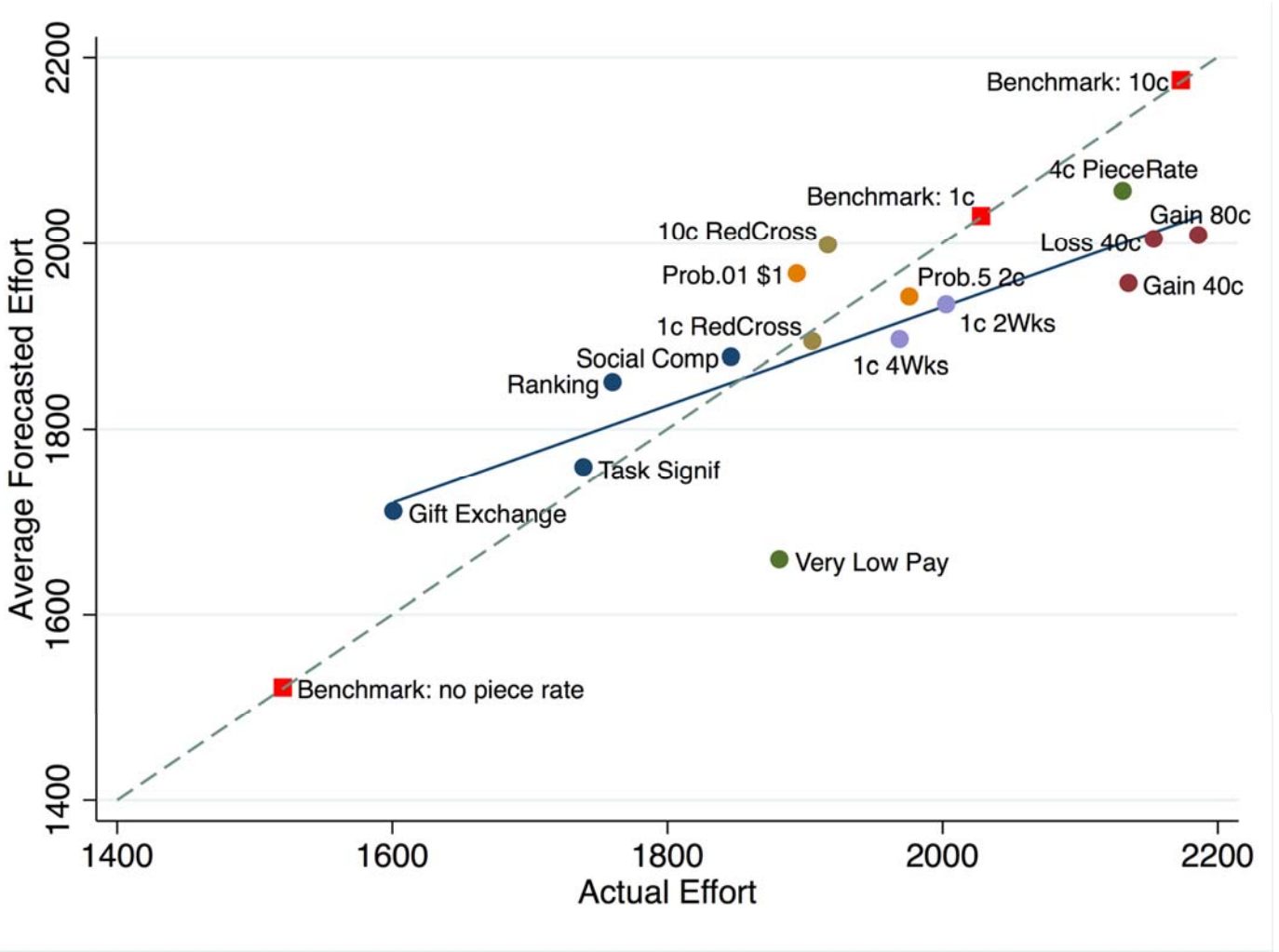



**Your Predictions**
Now we would like you to make your predictions about the average number of points scored in each of the 15 remaining conditions. For each of the conditions, we report the exact wording that the participants saw. Please use the slider scales to make your guesses.




**Notes:** Figure 1 shows screenshots reproducing portions of page 1 of the Qualtrics survey which experts used to make forecasts. The survey features first the results for 3 benchmark treatments, and then 15 sliders, one for each treatment (given that the results for 3 treatments were provided as a benchmark). For each treatment, the left side displays the treatment-specific wording which the subjects assigned to that treatment saw, and on the right side a slider which the experts can move to make a forecast.

**Figure 2. Wisdom-of-Crowds Accuracy: Average Performance and Average Forecast by Treatment, Academic Experts**




**Notes:** Figure 2 presents the results from the 15 treatments with forecasts and the three benchmarks also reported in Table 2. Each dot indicates a treatment, with the actual (average) effort by the MTurk workers on the x-axis and the average forecast by the 208 academic experts on the y axis. The 3 benchmark treatments, for which there was no forecast, are reported with a red square. Forecasts close to the 45 degree dotted line indicate cases in which the average forecast is very close to the actual average performance. The continuous line indicates the OLS line fit across the 15 points, with estimate forecast = 876 (238) + .527 (.122) * actual.

**Figure 3. Distribution of Accuracy Measures for Individual Academic Experts versus Wisdom of Crowds: Data versus Model Fit**

**Figure 3a. Mean Absolute Error, Data**

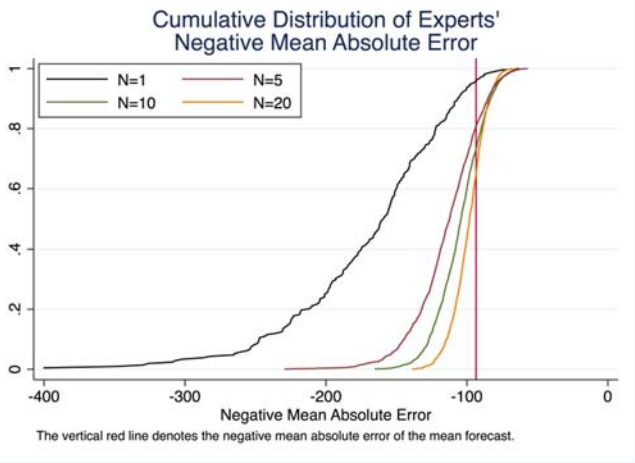



**Figure 3b. Rank-Order Correlation, Data**

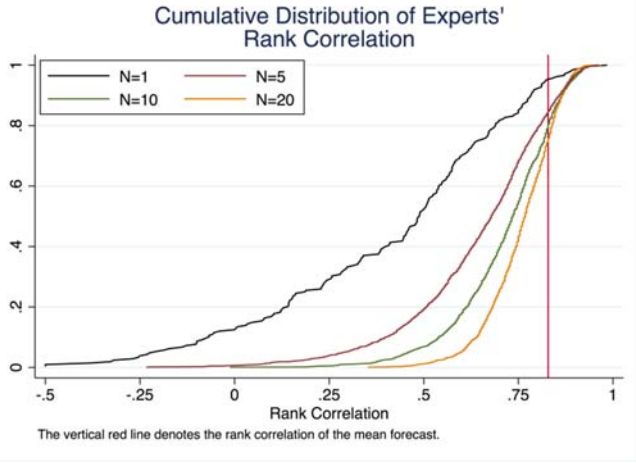



**Figure 3c. Mean Absolute Error, Model**

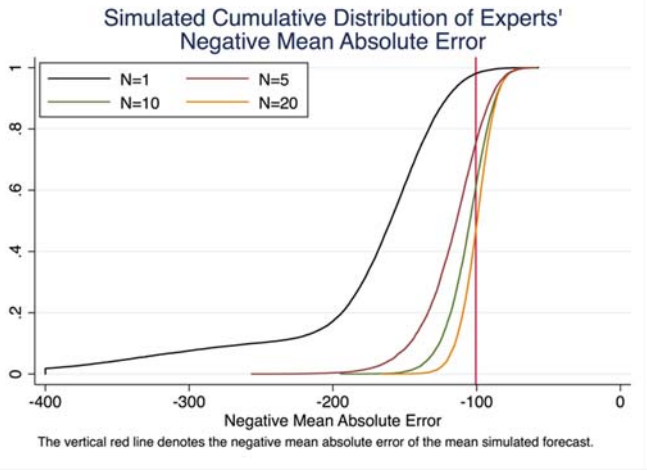
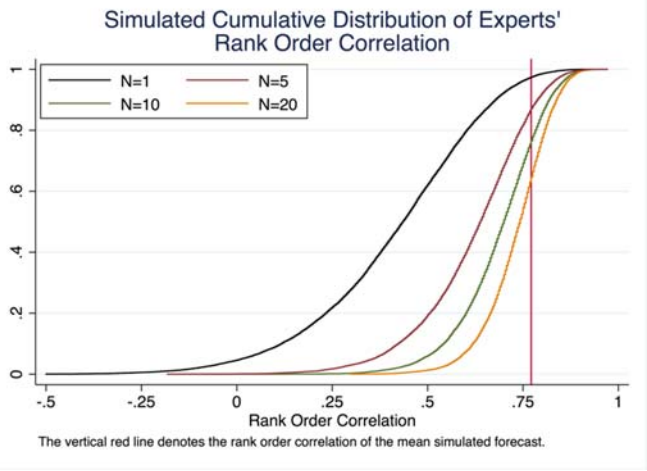
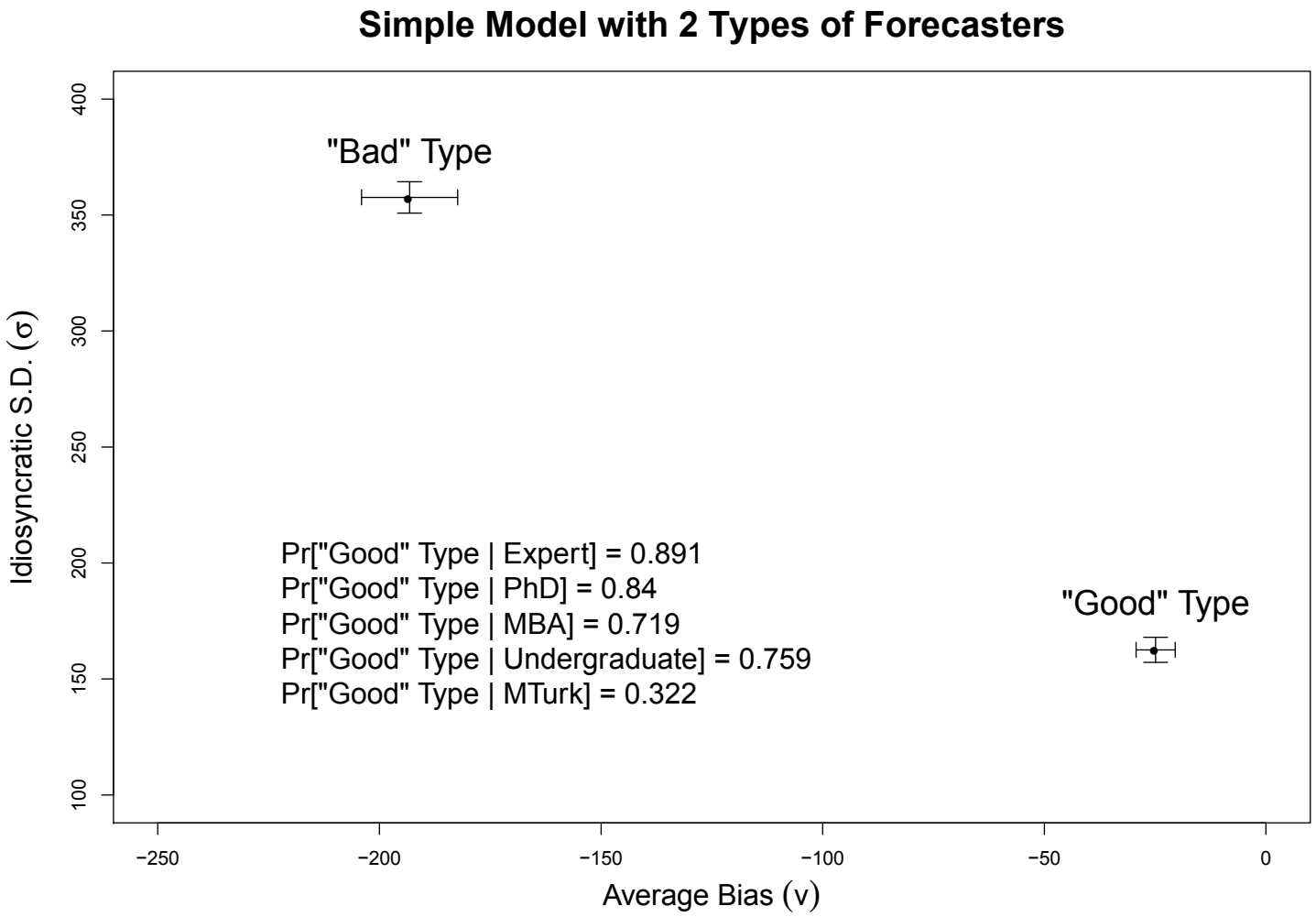



**Figure 3d. Rank-Order Correlation, Model**




**Notes:** In Figure 3a, for each of the 208 experts, we compute the absolute deviation between the forecast and the actual effort by treatment, average across the 15 treatments, take the negative, and plot the c.d.f. of this accuracy measure. The vertical red line shows the absolute error for the average, as opposed to the individual, forecast. We also form hypothetical pools of N forecasters (with N=5, 10, 20) drawn 1,500 times with replacement from the 208 experts, and for each draw take the average across the N forecasts and compute the accuracy measure. Figure 3b shows the corresponding c.d.f. for the rank-order correlation measure. Figures 3c and 3d are the model analogues of figures 3a and 3b respectively. Specifically, we simulate 100 samples of the 208 experts according to our benchmark model specification (column 1 of table 3) and use the mean absolute error and rank-order correlations from these simulations to create the c.d.f.'s. The red vertical lines in figures 3c and 3d are based on a single simulated dataset (out of the 100 generated).

**Figure 4. Maximum-Likelihood Estimates of Model of Expertise.**




**Notes:** Figure 4 plots the MLE estimates of a model with two unobserved types which differ in the average bias (v) and the idiosyncratic standard deviation (sigma). The two plotted points report the point estimates and confidence intervals from our benchmark model (Column 1 in Table 3). The probability of being the type with a smaller magnitude of bias ("good" type) is also shown in the figure.

**Figure 5. *Vertical, Horizontal,* and *Contextual* Expertise, Among Experts**

**Figure 5a. Academic Rank (Vertical Expertise)**

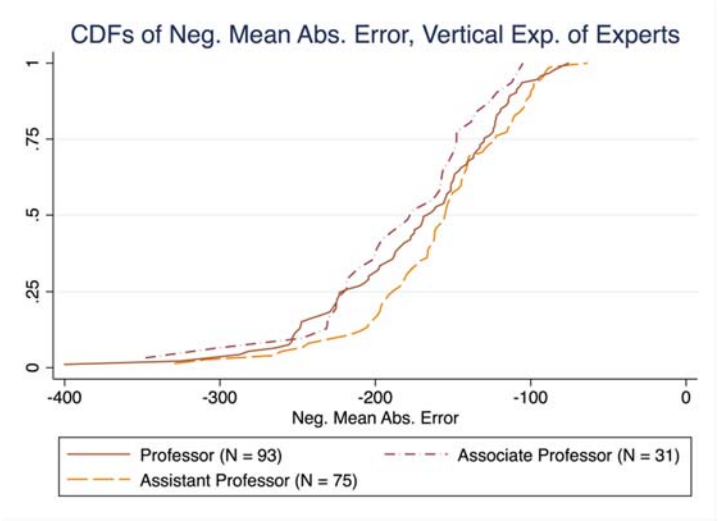



**Figure 5b. Citations (Vertical Expertise)**

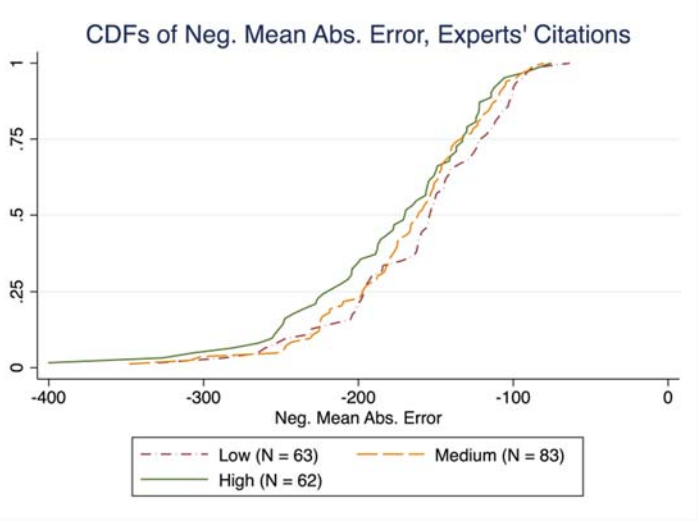



**Figure 5c. Fields (Horizontal Expertise)**

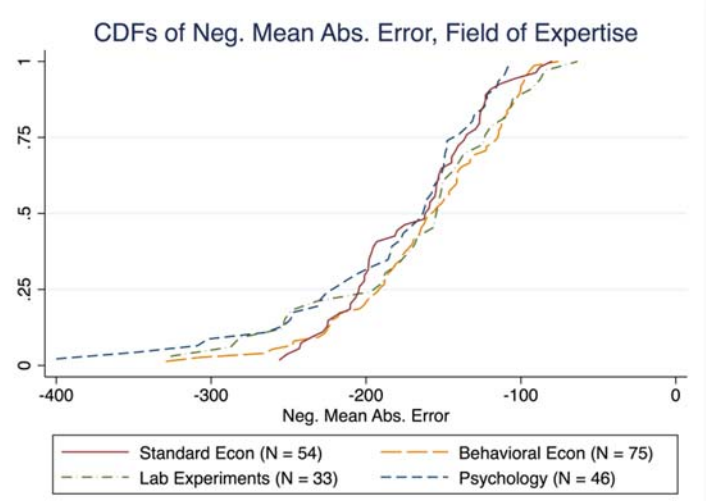
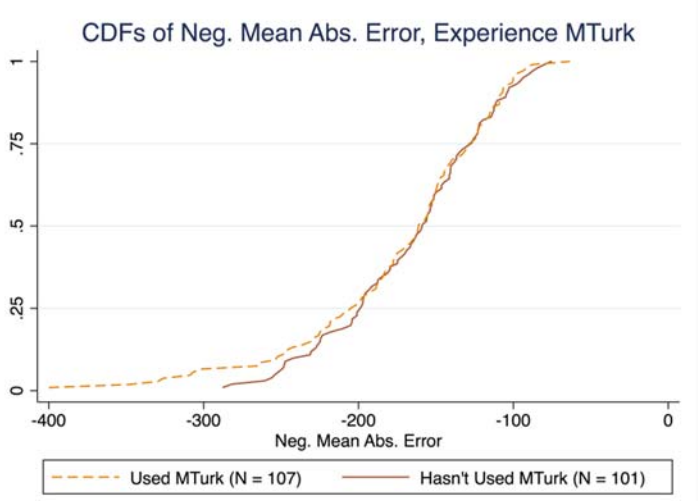



**Figure 5d. Experience with MTurk Platform (Contextual Expertise)**




**Notes:** Figure 5a presents the cumulative distribution function for the negative of the mean absolute error in forecast by the academic experts (full professors, associate professors, and assistant professors, with the "other" category omitted). Figure 5b splits the 208 academic experts into groups based on Google Scholar citations, High (Low) are the top (bottom) three deciles and Medium are the middle 4 deciles. Figure 5c splits the academic experts into four main fields based on the assessment of the authors. Figure 5d splits the academic experts based on the self-reported use of MTurk.

**Figure 6. Experts versus Non-Experts (PhDs, Undergraduates, MBAs, MTurk Workers)**

**Figure 6a. Mean Absolute Error, Data**

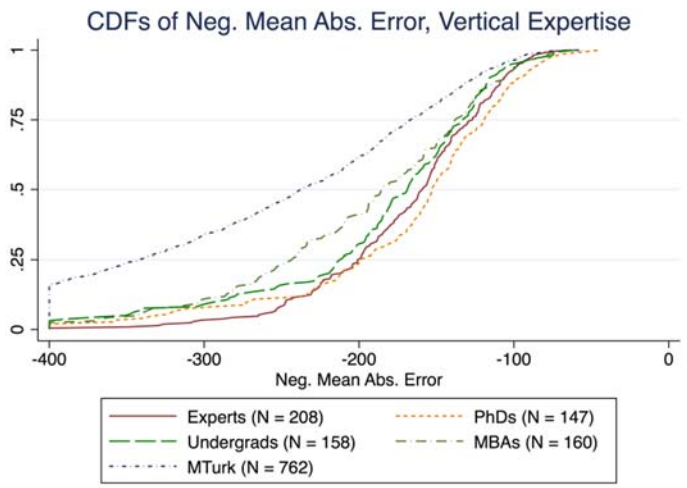



**Figure 6b. Rank-Order Correlation, Data**

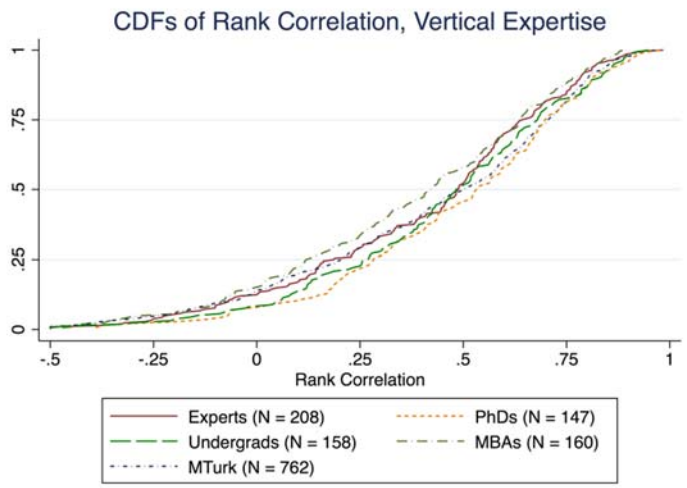



**Figure 6c. Mean Absolute Error, Model**

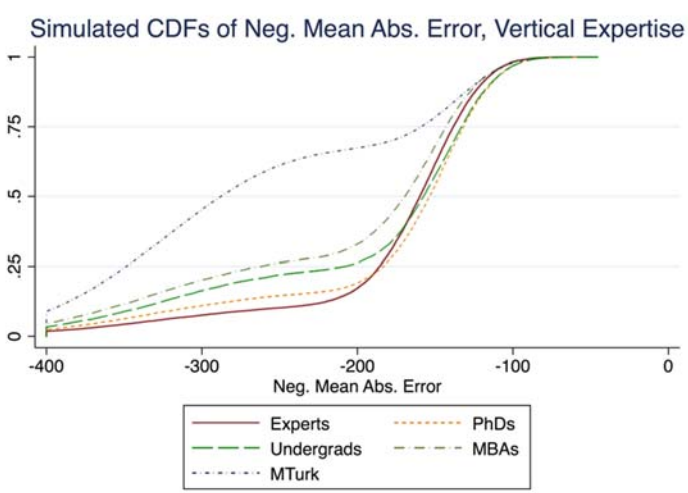
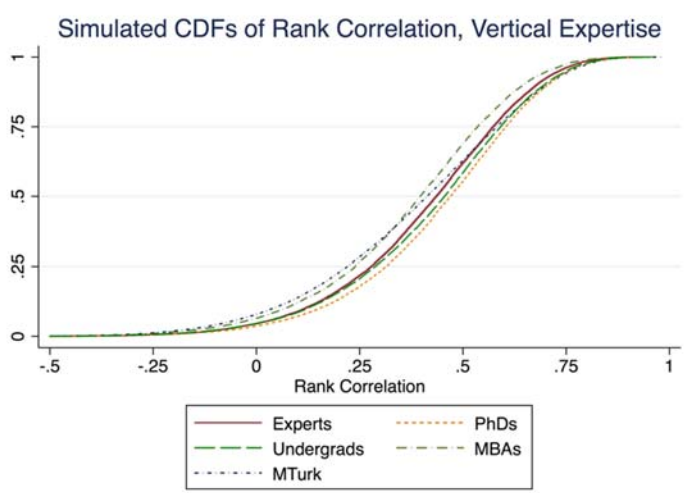
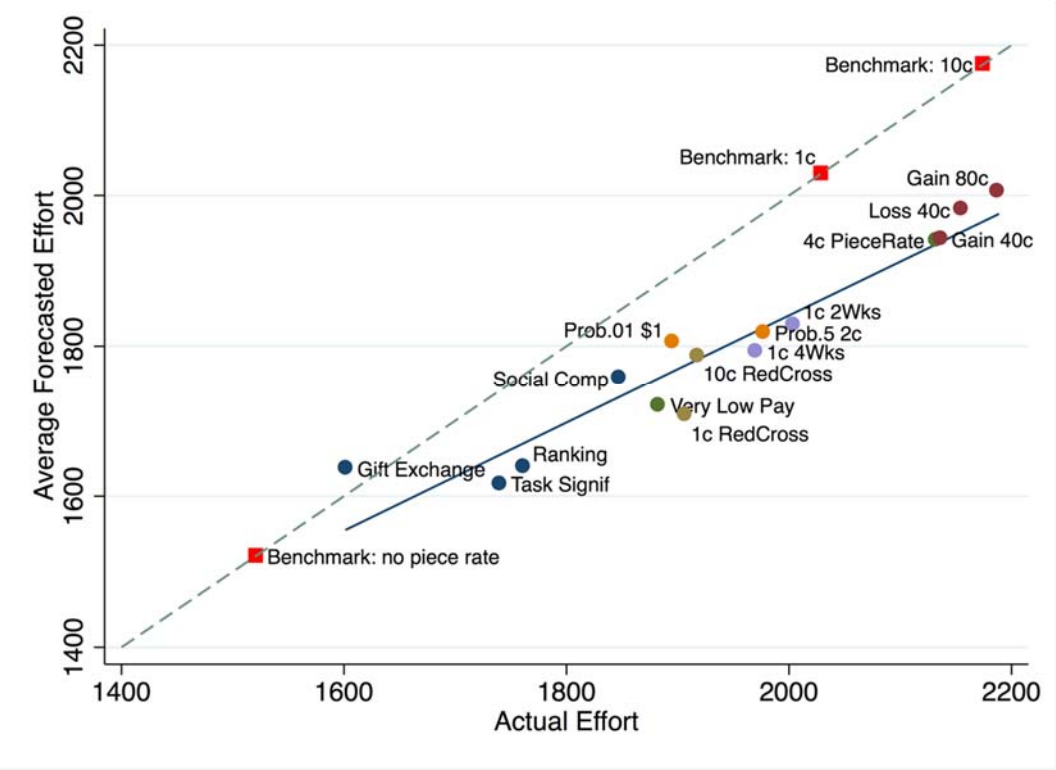



**Figure 6d. Rank-Order Correlation, Model**




**Notes:** Figures 6a-b compare the academic experts with groups of non-experts: PhD students, undergraduates, MBA students, and MTurk workers making forecasts respectively for the negative of the mean absolute error (Figure 6a) and the rank-order correlation (Figure 6b). In Figures 6c-d we show the corresponding figures from simulations for the model estimates as in Column 1 of Table 3.

**Figure 6e. Wisdom-of-Crowds Accuracy: Average Performance and Average Forecast by Treatment, MTurk**




**Notes:** Figure 6e presents the parallel data in Figure 2 (wisdom-of-crowd forecasts) for MTurk forecasters.

**Figure 7. Accuracy and Effort in Taking Task**

**Figure 7a. Time Taken in Completing the Survey, Data**

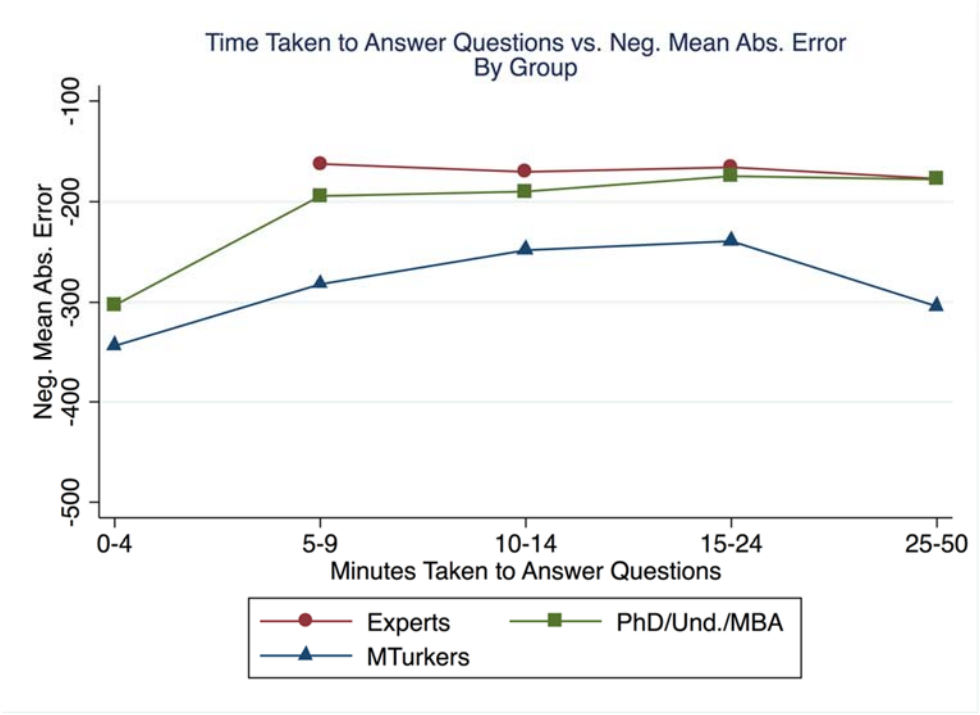
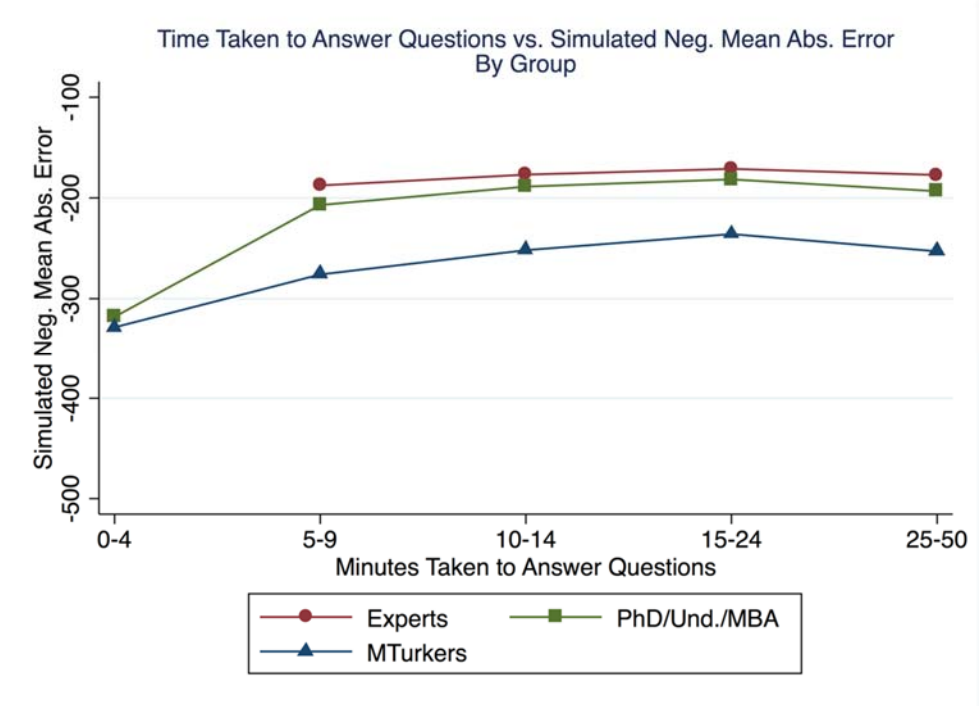



**Figure 7b. Time Taken in Completing the Survey, Model**




**Notes:** Figures 7a plots the accuracy for three groups of forecasters (academic experts; undergraduate, MBA, and PhD students; and MTurkers) as a function of how long they took to complete the survey. Specifically, the figures plot the average accuracy by minutes of the time taken for survey completion. In this Figure and in subsequent Figures 8 and 9 (and Online Appendix Figures 4, 6, 8, and 9) we only plot cells with at least 3 observations within a group. Figure 7b presents the corresponding figure from simulations for the model estimates as in Column 2 of Table 3. This MLE specification forces the same effect of time taken for the different groups.

**.Figures 8. Accuracy and Confidence in One's Own Expertise, by Confidence Level (0 to 15)**

**Figure 8a. Mean Absolute Error, Data**

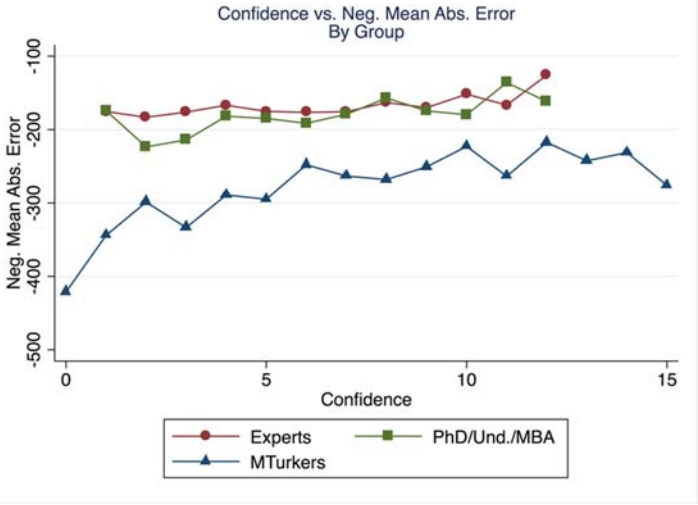



**Figure 8b. Number Correct out of 15, Data**

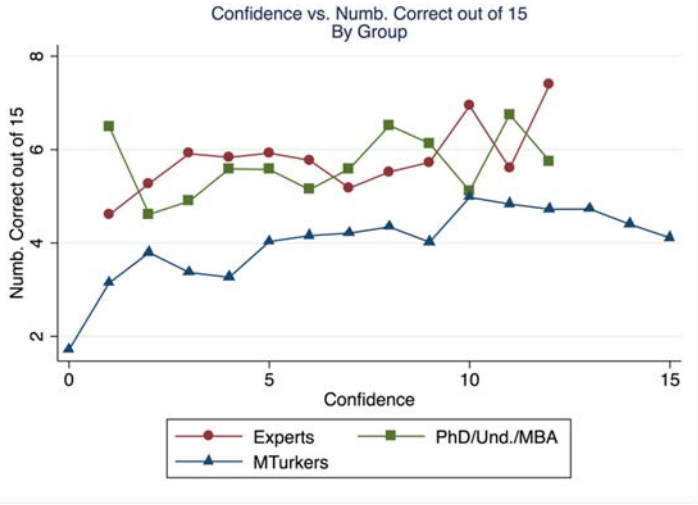



**Figure 8c. Mean Absolute Error, Model**

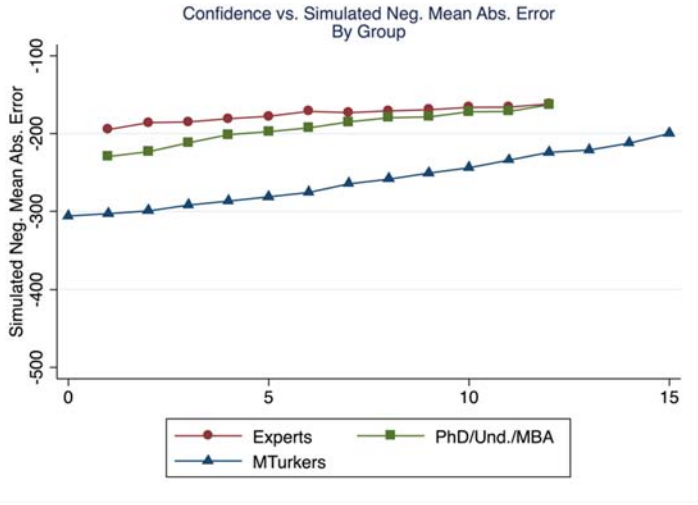
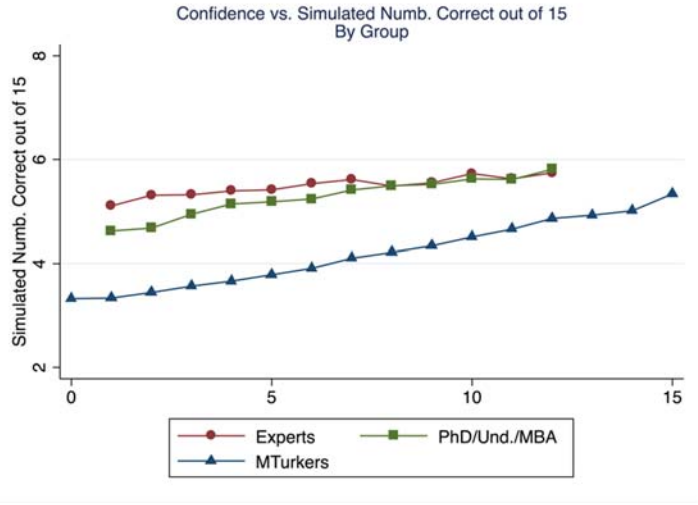



**Figure 8d. Number Correct out of 15, Model**




**Notes:** Figures 8a-b plots the average accuracy for three groups of forecasters (academic experts, undergraduate/MBA/PhD students, and MTurkers) by how confident the respondent felt about the accuracy. In particular, each survey respondent indicated how many out of 15 forecasts he or she made were going to be accurate up to 100 points relative to the truth. Figures 8c-d present the corresponding figures from 100 simulations for the model estimates as in Column 3 of Table 3.

**Figures 9. Accuracy and Revealed Expertise (Forecasting of 4c Piece Rate), by Decile**

**Figure 9a. Deciles in Accuracy of Forecasting the 4c Piece Rate Treatment, Data**

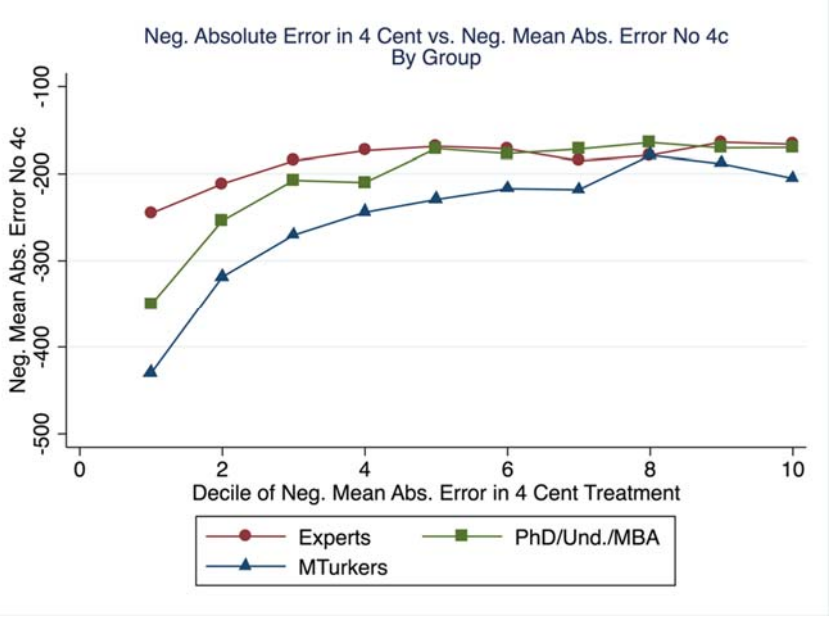
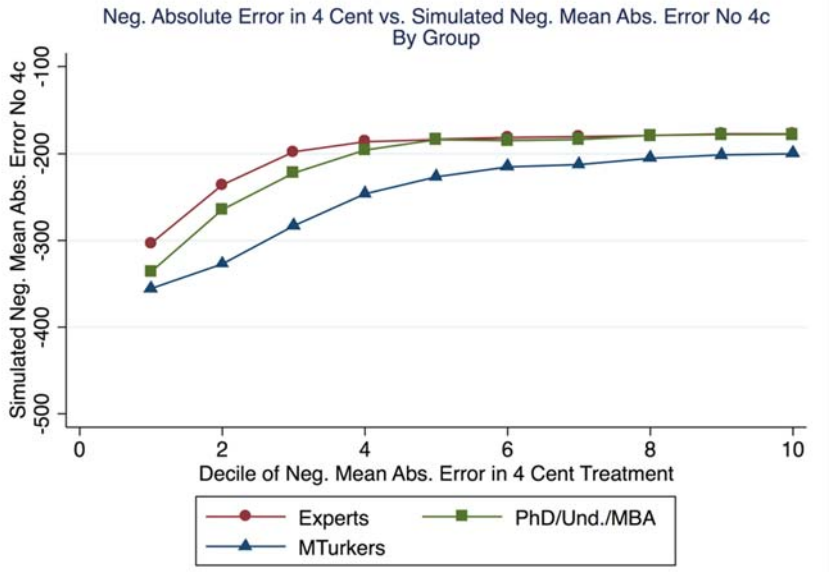



**Figure 9b. Deciles in Accuracy of Forecasting the 4c Piece Rate Treatment, Model**




**Notes:** Figures 9a plots the average accuracy for three groups of forecasters (academic experts, undergraduate/MBA/ PhD students, and MTurkers) by decile of a revealed-accuracy measure (the decile thresholds are computed using all three groups). Namely, we take the absolute distance between the forecast and the actual effort for the 4-cent piece rate treatment, a treatment for which the forecast should not involve behavioral factors. For these plots the accuracy measure is computed excluding the 4-cent treatment. Figure 9b presents the corresponding figures from 100 simulations for the model estimates as in Column 4 of Table 3.

**Figure 10. Superforecasters: Selecting Non-Experts to Match Accuracy of Experts**

**Figure 10a. Individual Accuracy, Data**

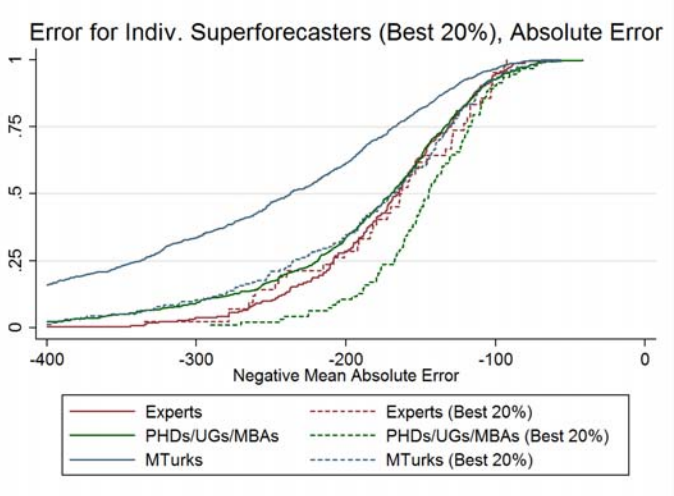



**Figure 10b. Wisdom-of-Crowds Accuracy (20 Forecasters), Data**

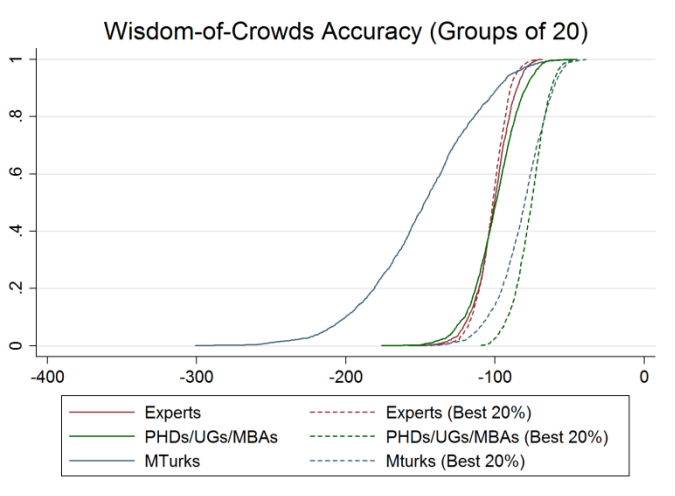



**Figure 10c. Individual Accuracy, Model**

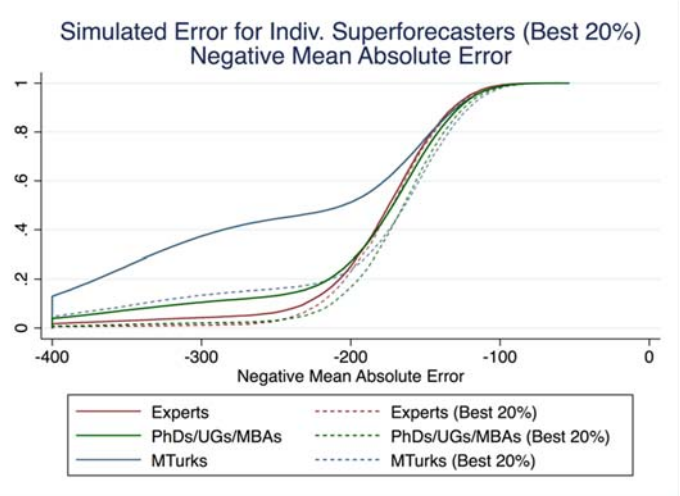
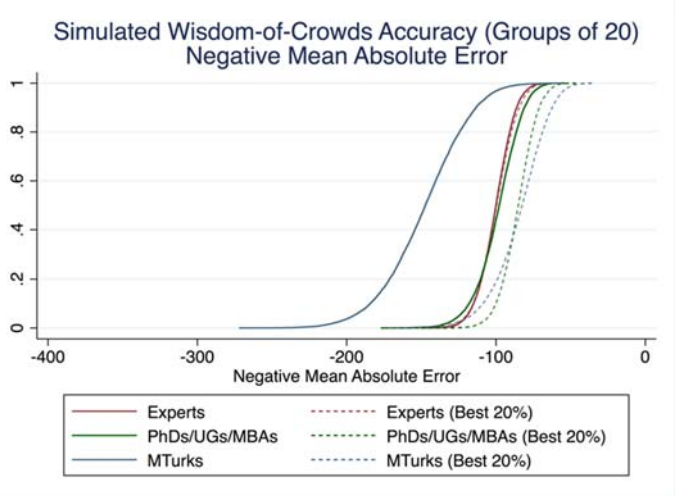
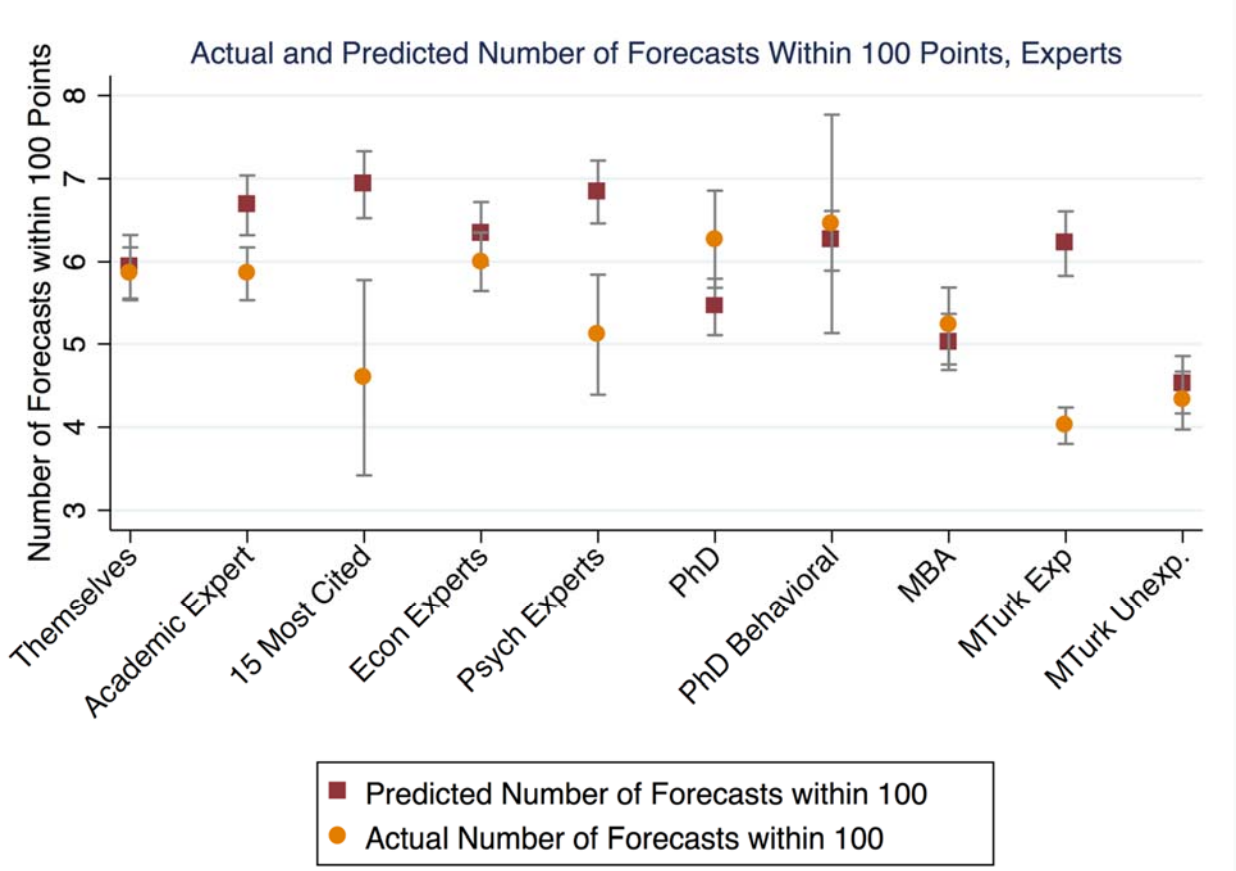



**Figure 10d. Wisdom-of-Crowds Accuracy (20 Forecasters), Model**




**Notes:** Figures 10a-b compare, for each of three groups of forecasters (academic experts, undergraduate/PhD/MBA students, and MTurkers), the accuracy of the overall group versus the accuracy of the top 20% (the "superforecasters") according to the regression in Table 6. To compute the superforecasters, we use a 10-fold method to ensure no in-sample overfitting. Figure 10a plots the distribution of the individual-level accuracy, while Figure 10b plots the wisdom-of-crowds accuracy for groups of sample size 20, using 1,500 bootstraps. Figures 10c-d presents the corresponding figures from simulations for the model estimates.

**Figure 11. Beliefs about Expertise**




**Notes:** Figure 11 compares the average accuracy of a group with the forecasted accuracy for that group by the 208 academic experts. Namely, the red squares report the average forecast of the number of correct answers (within 100 points of the truth) out of 15. The forecast is averaged across the academic experts making the forecast. The yellow circle represents the actual accuracy (number of correct answers within 100 points of the truth) for that same group. For example, for the 15-most cited experts, this takes the top-15 experts in citations and compares the average of their individual accuracy. Notice that the sample slightly differs from the overall sample to be consistent with the question asked. For MBAs we only include Chicago MBAs and for PhDs we only include Berkeley and Chicago PhDs since the question mentioned only those groups (see Appendix Figure 1b).

**Table 1. Findings by Treatment: Effort in Experiment and Expert Forecasts**

| Category | Treatment | N | Mean Effort (s.e.) | Mean Forecast | Absolute Error, Mean Forecast | Error, Indiv. Forecast (Mean and s.d.) | Percent Experts Outperforming Mean |
|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Piece Rate | "Your score will not affect your payment in any way." | 540 | 1521 (31.23) | | Benchmark | | |
| | "As a bonus, you will be paid an extra **1 cent** for every **100 points** that you score." | 558 | 2029 (27.47) | | Benchmark | | |
| | "As a bonus, you will be paid an extra **10 cents** for every **100 points** that you score." | 566 | 2175 (24.28) | | Benchmark | | |
| | "As a bonus, you will be paid an extra **4 cents** for every **100 points** that you score." | 562 | 2132 (26.42) | 2057 | 75 | 88.34 (111.78) | 67.31 |
| Pay Enough or Don't Pay | "As a bonus, you will be paid an extra **1 cent** for every **1,000 points** that you score." | 538 | 1883 (28.61) | 1657 | 226 | 284.97 (195.37) | 44.23 |
| Social Preferences: Charity | "As a bonus, the Red Cross charitable fund will be given **1 cent** for every 100 points that you score." | 554 | 1907 (26.85) | 1894 | 13 | 164.37 (117.97) | 3.85 |
| | "As a bonus, the Red Cross charitable fund will be given **10 cents** for every 100 points that you score." | 549 | 1918 (25.93) | 1997 | 79 | 182.1 (107.68) | 16.85 |
| Social Preferences: Gift Exchange | "In appreciation to you for performing this task, you will be paid a **bonus** of **40 cents**. Your score will not affect your payment in any way." | 545 | 1602 (29.77) | 1709 | 107 | 164.16 (165.6) | 53.85 |
| Discounting | "As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **two weeks** from today." | 544 | 2004 (27.38) | 1933 | 71 | 92.2 (129.4) | 65.38 |
| | "As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **four weeks** from today." | 550 | 1970 (28.68) | 1895 | 75 | 114.67 (137.22) | 57.21 |
| Gains versus Losses | "As a bonus, you will be paid an **extra 40 cents** if you score at least 2,000 points." | 545 | 2136 (24.66) | 1955 | 181 | 186.42 (142.7) | 62.02 |
| | "As a bonus, you will be paid an **extra 40 cents**. However, you will **lose this bonus** (it will not be placed in your account) unless you score at least 2,000 points. " | 532 | 2155 (23.09) | 2002 | 153 | 167.06 (126.28) | 57.21 |
| | "As a bonus, you will be paid an **extra 80 cents** if you score at least 2,000 points." | 532 | 2188 (22.99) | 2007 | 181 | 188 (121.38) | 53.37 |
| Risk Aversion and Probability Weighting | "As a bonus, you will have a **1% chance** of being paid an **extra $1** for every 100 points that you score. One out of every 100 participants who perform this task will be randomly chosen to be paid this reward." | 555 | 1896 (28.44) | 1967 | 71 | 222.37 (139.87) | 12.5 |
| | "As a bonus, you will have a **50% chance** of being paid an **extra 2 cents** for every 100 points that you score. One out of two participants who perform this task will be randomly chosen to be paid this reward." | 568 | 1977 (24.73) | 1941 | 36 | 131.48 (126.66) | 20.19 |
| Social Comparisons | "Your score will not affect your payment in any way. In a previous version of this task, **many participants** were able to score more than **2,000 points.**" | 526 | 1848 (32.14) | 1877 | 29 | 177.63 (114.22) | 6.73 |
| Ranking | "Your score will not affect your payment in any way. After you play, we will show you **how well you did relative** to other participants who have previously done this task." | 543 | 1761 (30.63) | 1850 | 89 | 196.21 (155.38) | 29.81 |
| Task Significance | "Your score will not affect your payment in any way. We are interested in how fast people choose to press digits and we would like you to do your very best. So **please try as hard as you can**." | 554 | 1740 (28.76) | 1757 | 17 | 181.3 (142.24) | 4.81 |
| **Average Across the 15 (Non-Benchmark) Treatments** | | | 1941 | 1900 | 94 | 169.42 | 37.02 |

**Notes:** The Table lists the 18 treatments in the Mturk experiment. The treatments differ just in one paragraph explaining the task and in the vizualization of the points earned. Column (2) reports the key part of the wording of the paragraph. For brevity, we omit from the description the sentence "This bonus will be paid to your account within 24 hours" which applies to all treatments with incentives other than in the Time Preference ones where the payment is delayed. Notice that the bolding is for the benefit of the reader of the Table. In the actual description to the MTurk workers, the whole paragraph was bolded and underlined. Column (1) reports the conceptual grouping of the treaments, Columns (3) and (4) report the number of MTurk subjects in that treatment and the mean number of points, with the standard errors. Column (5) reports the mean forecast among the 208 experts of the points in that treatment. Columns (1)-(5) are reproduced from DellaVigna and Pope (2016). Column (6) reports the absolute error between the average effort and the average expert forecast (the wisdom-of-crowds measure), while Column (7) reports the average and the standard error of the absolute error in forecast for the *individual* expert. Finally, Column (8) reports the share of individual expert forecasts with a lower error than the wisdom-of-crowds average forecast.

**Table 2. Accuracy of Forecasts by Group of Forecasters versus Random Guesses**

| | Average Accuracy (and s.d.) of *Individual* Forecasts | Accuracy of *Mean* Forecast (Wisdom of Crowds) | % Forecasters Doing Better Than Mean Forecast | Wisdom of Crowds: Accuracy Using Average of Simulated Group of Forecasters, Mean (and s.d.) | |
|---|---|---|---|---|---|
| | | | | Group of 5 | Group of 20 |
| | (1) | (2) | (3) | (4) | (5) |
| **Panel A. Mean Absolute Error** | | | | | |
| *Groups* | | | | | |
| Academic Experts (N=208) | 169.42 (56.24) | 93.48 | 4.33 | 113.92 (23.59) | 98.7 (11.79) |
| PhD Students (N=147) | 167.78 (74.26) | 91.65 | 8.16 | 113.47 (31.29) | 97.93 (14.5) |
| Undergraduates (N=158) | 187.84 (86.25) | 87.86 | 3.16 | 116.03 (35.65) | 94.26 (17.66) |
| MBA Students (N=160) | 198.17 (86.31) | 100.72 | 7.50 | 129.4 (34.84) | 110.69 (17.61) |
| Mturk Workers (N=762) | 271.57 (144.90) | 146.93 | 17.85 | 170.32 (65.03) | 150.35 (39.54) |
| *Benchmark for Comparison* | | | | | |
| Random Guess in 1000-2500 | 416.14 | | | | |
| Random Guess in 1500-2200 | 223.93 | | | | |
| **Panel B. Rank-Order Correlation Between Actual Effort and Forecasts** | | | | | |
| *Groups* | | | | | |
| Academic Experts (N=208) | 0.41 (0.32) | 0.83 | 4.81 | 0.65 (0.19) | 0.76 (0.09) |
| PhD Students (N=147) | 0.48 (0.30) | 0.86 | 6.80 | 0.69 (0.19) | 0.80 (0.09) |
| Undergraduates (N=158) | 0.45 (0.31) | 0.87 | 5.06 | 0.68 (0.17) | 0.81 (0.08) |
| MBA Students (N=160) | 0.37 (0.33) | 0.71 | 17.50 | 0.56 (0.21) | 0.67 (0.11) |
| Mturk Workers (N=762) | 0.42 (0.35) | 0.95 | 0.26 | 0.68 (0.21) | 0.87 (0.07) |
| *Benchmark for Comparison* | | | | | |
| Random Guess in 1000-2500 | 0.00 | | | | |
| Random Guess in 1500-2200 | 0.00 | | | | |
| **Panel C. Mean Squared Error** | | | | | |
| *Groups* | | | | | |
| Academic Experts (N=208) | 49822 (34169) | 12606 | 2.88 | 20081 (8312) | 14430 (3213) |
| PhD Students (N=147) | 50775 (47835) | 11980 | 6.12 | 19651 (10929) | 13918 (4129) |
| Undergraduates (N=158) | 60271 (61306) | 9769 | 2.53 | 20104 (12548) | 12207 (4574) |
| MBA Students (N=160) | 69855 (63412) | 13334 | 3.75 | 24763 (12825) | 16199 (4930) |
| Mturk Workers (N=762) | 128801 (130559) | 23660 | 9.71 | 43232 (30803) | 28749 (14062) |
| *Benchmark for Comparison* | | | | | |
| Random Guess in 1000-2500 | 249294 | | | | |
| Random Guess in 1500-2200 | 75097 | | | | |
| **Panel D. Correlation Between Actual Effort and Forecasts** | | | | | |
| *Groups* | | | | | |
| Academic Experts (N=208) | 0.45 (0.29) | 0.77 | 9.13 | 0.64 (0.17) | 0.73 (0.09) |
| PhD Students (N=147) | 0.51 (0.28) | 0.86 | 4.76 | 0.72 (0.16) | 0.82 (0.07) |
| Undergraduates (N=158) | 0.49 (0.30) | 0.89 | 3.80 | 0.72 (0.16) | 0.84 (0.07) |
| MBA Students (N=160) | 0.42 (0.32) | 0.77 | 13.13 | 0.61 (0.19) | 0.72 (0.09) |
| Mturk Workers (N=762) | 0.43 (0.35) | 0.95 | 0.00 | 0.69 (0.19) | 0.88 (0.06) |
| *Benchmark for Comparison* | | | | | |
| Random Guess in 1000-2500 | 0.00 | | | | |
| Random Guess in 1500-2200 | 0.00 | | | | |

**Notes:** The Table reports evidence on the accuracy of forecasts made by the five groups of forecasters: academic experts, PhD students, undergraduates, MBA students, and MTurk workers. Panel A presents the results for the benchmark measure (mean absolute error), Panel B on the rank-order correlation between actual average effort and the forecast, Panel C presents the results on mean squared error, and Panel D on the corresponding correlation. Within each Panel and for reach group, the table reports the average individual accuracy across the forecasters in the group (Column 1) versus the accuracy of the average forecast in the group (Column 2). The difference is often referred to as "wisdom of crowds". Column 3 displays the percent of individuals in the group with an accuracy higher than the wisdom-of-crowd accuracy (Column 2). In Columns 4 and 5 we present counterfactuals on how much the distribution of accuracy would shift if instead of considering individual forecasts (Column 1) we considered the accuracy of average forecasts made by groups of 5 (Column 4) or 20 (Column 5). Random guesses are from a uniform distribution in (1000, 2500) and (1500, 2200), respectively.

## Table 3. Maximum-Likelihood Estimate of Model

| Sample: | All Forecasters | | | | | Experts Only | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| ***Estimated Parameters for the 2 Types*** | | | | | | | |
| **$v^{(1)}$ (Average Bias, Type 1)** | -24.89 | -23.44 | -23.87 | -22.68 | -21.04 | 17.80 | 18.50 |
| | (2.25) | (2.19) | (2.19) | (2.12) | (2.09) | (4.29) | (4.25) |
| **$v^{(2)}$ (Average Bias, Type 2)** | -193.19 | -200.22 | -199.10 | -247.66 | -248.33 | -60.54 | -60.73 |
| | (5.53) | (5.57) | (5.63) | (7.04) | (6.90) | (4.96) | (4.96) |
| **$\sigma^{(1)}$ (Idiosyncratic s.d., Type 1)** | 162.58 | 165.42 | 164.73 | 187.72 | 187.13 | 59.19 | 59.28 |
| | (2.74) | (2.61) | (2.64) | (2.15) | (2.11) | (4.38) | (4.34) |
| **$\sigma^{(2)}$ (Idiosyncratic s.d., Type 2)** | 357.59 | 358.07 | 358.48 | 368.27 | 366.10 | 216.15 | 216.09 |
| | (3.46) | (3.36) | (3.42) | (3.78) | (3.71) | (3.82) | (3.81) |
| ***Predictors of Forecasters Being of Type 1, Logit Coefficients*** | | | | | | | |
| **Constant** | -0.74 | -0.88 | -1.87 | 1.51 | 0.46 | -1.15 | -1.94 |
| | (0.07) | (0.08) | (0.12) | (0.08) | (0.14) | (0.25) | (0.47) |
| **Indicator for Expert** | 2.84 | 2.52 | 3.12 | 2.45 | 2.42 | | |
| | (0.14) | (0.16) | (0.15) | (0.24) | (0.25) | | |
| **Indicator for PhD** | 2.40 | 1.99 | 2.57 | 1.84 | 1.69 | | |
| | (0.13) | (0.13) | (0.14) | (0.20) | (0.21) | | |
| **Indicator for MBA** | 1.69 | 1.36 | 1.91 | 1.20 | 1.05 | | |
| | (0.10) | (0.11) | (0.11) | (0.16) | (0.17) | | |
| **Indicator for Undergraduate** | 1.89 | 1.90 | 2.08 | 2.07 | 2.23 | | |
| | (0.11) | (0.12) | (0.12) | (0.18) | (0.20) | | |
| **Response Time: 0-4 mins** | | -3.01 | | | -0.70 | | |
| | | (0.83) | | | (0.21) | | |
| **Response Time: 10-14 mins** | | 0.56 | | | 0.38 | | 0.80 |
| | | (0.09) | | | (0.11) | | (0.40) |
| **Response Time: 15-24 mins** | | 0.90 | | | 0.66 | | 0.72 |
| | | (0.10) | | | (0.13) | | (0.40) |
| **Response Time: 25+ mins** | | 0.46 | | | 0.41 | | 0.82 |
| | | (0.12) | | | (0.19) | | (0.41) |
| **Predicted # Forecasts within 100 pts** | | | 0.17 | | 0.12 | | |
| | | | (0.01) | | (0.02) | | |
| **100 x Negative 4-Cent Error** | | | | 0.74 | 0.71 | | |
| | | | | (0.03) | (0.03) | | |
| **Indicator for Associate Professor** | | | | | | -0.50 | -0.48 |
| | | | | | | (0.28) | (0.28) |
| **Indicator for Professor** | | | | | | -0.65 | -0.68 |
| | | | | | | (0.28) | (0.29) |
| **Indicator for Other Rank** | | | | | | 0.02 | -0.01 |
| | | | | | | (0.41) | (0.41) |
| **Decile of Google Scholar Citations** | | | | | | 0.07 | 0.08 |
| | | | | | | (0.05) | (0.05) |
| **Indicator for Field: Applied Micro** | | | | | | -0.08 | -0.02 |
| | | | | | | (0.25) | (0.25) |
| **Indicator for Field: Theory** | | | | | | -0.01 | 0.14 |
| | | | | | | (0.36) | (0.37) |
| **Indicator for Field: Lab** | | | | | | 0.72 | 0.71 |
| | | | | | | (0.22) | (0.23) |
| **Indicator for Field: Psychology** | | | | | | 0.06 | 0.20 |
| | | | | | | (0.26) | (0.27) |
| **Indicator for having used Mturk** | | | | | | -0.22 | -0.24 |
| | | | | | | (0.19) | (0.19) |
| ***N*** | 21,525 | 21,525 | 21,525 | 20,090 | 20,090 | 3120 | 3120 |
| ***Log-likelihood*** | -150,184 | -150,058 | -150,061 | -139,694 | -139,616 | -20,729 | -20,726 |

**Notes:** The table reports the MLE estimation results for the discrete heterogeneity model described in the paper. All models in the table allow for two types of forecasters, where type 1 has a smaller magnitude of average bias. The sample of columns 1 through 5 include all forecasts, except when accuracy of the forecast on the 4-cent treatment is used as a predictor of type, in which case forecasts on the 4-cent treatment are omitted. In column 1, only indicators for subject groups (with MTurks as the omitted category) are used as predictors of types. In columns 2, 3 and 4, response time, a measure of the forecasters' confidence in their own forecasts, and accuracy of the forecast on the 4-cent treatment are respectively added to the subject group indicators as predictors of type in the model. In column 5, all the aforementioned variables are used as predictors of forecaster type. The sample for columns 6 and 7 are restricted to academic experts. In column 6, only measures of the horizontal and vertical expertise of experts are used as predictors of type, and response time is added as a predictor for column 7 (no experts took less than 5 minutes to respond, hence the omission of the indicator for response time being less than 5 minutes in column 7). All specifications with a measure of forecaster confidence also include an indicator for missing confidence measure, which is not shown in this table.

## Table 4. Impact of *Vertical*, *Horizontal*, and *Contextual* Expertise on Forecast Accuracy

| Dep. Var. (Measure of Accuracy): | (Negative of) Absolute Forecast Error in Treatment *t* by | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| ***Measures of Vertical Expertise (Omitted: Assistant Professor)*** | | | | | |
| **Associate Professor** | -23.86** | -23.78** | -18.40 | -17.15 | |
| | (11.48) | (11.66) | (13.12) | (13.32) | |
| **Full Professor** | -16.03* | -16.81* | -10.61 | -11.94 | |
| | (8.61) | (8.85) | (14.58) | (14.44) | |
| **Other (Post-Doc or Research Scientist)** | 16.08 | 18.85 | 12.73 | 15.73 | |
| | (12.20) | (12.23) | (12.36) | (12.37) | |
| **Decile Google Scholar Citations** | | | -1.14 | -1.03 | |
| | | | (2.34) | (2.29) | |
| ***Main Field of Expertise (Omitted: Behavioral Economics)*** | | | | | |
| **Applied Microeconomics** | | | -4.14 | -4.63 | |
| | | | (9.32) | (9.39) | |
| **Economic Theory** | | | -12.18 | -18.01 | |
| | | | (13.93) | (14.22) | |
| **Laboratory Experiments** | | | -1.71 | -3.32 | |
| | | | (12.23) | (12.43) | |
| **Psychology or Behavioral Decision-Making** | | | -10.84 | -15.18 | |
| | | | (12.98) | (13.57) | |
| ***Measure of Contextual Expertise*** | | | | | |
| **Has Used Mturk in Own Research (Self-Reported)** | | | -6.30 | -6.92 | |
| | | | (8.37) | (8.36) | |
| ***Measures of Horizontal Expertise*** | | | | | |
| **Expert *i* Has Written Paper on Topic of Treatment *t*** | | | | | -7.09 |
| | | | | | (8.33) |
| **Fixed Effects for Forecaster *i*:** | | | | | X |
| **Effort Controls: Survey Completion Time, Click on Practice Task, Click on Instructions, and Delay Start:** | | X | | X | |
| **Controls:** | Fixed Effects for Treatment and for Order of Treatments | | | | |
| **Sample:** | Academic Experts | | | | |
| ***N*** | 3120 | 3120 | 3120 | 3120 | 3120 |
| ***R Squared*** | 0.119 | 0.120 | 0.121 | 0.123 | 0.263 |

**Notes:** The table reports the result of OLS regressions of measures of forecast accuracy on expertise measures. The dependent variable is the (negative of) the absolute forecast error and an observation in the regression is a forecaster-treatment combination, with each forecaster providing forecasts for 15 treatments. Column (5) includes as horizontal measure of expertise an indicator for whether the expert has written a paper on the topic of the relevant treatment. This specification also includes fixed effects for the expert i (unlike the other columns). Columns (3) and (4) use as control variables the decile of Google Scholar citations for the researcher, main field of expertise, and an indicator for whether the researcher has used MTurk. Columns (2) and (4) include as controls time to survey completion, whether the forecaster clicked on practice or the instructions and how many days the forecaster delayed starting the survey. All specifications include fixed effects for the order in which the expert encountered a treatment (to control for fatigue) and fixed effects for the treatment. Standard errors are clustered by individual.

* significant at 10%; ** significant at 5%; *** significant at 1%

## Table 5. Experts versus Non-Experts

| Dep. Var. (Measure of Accuracy): | (Neg.) Absolute Forecast Error in Treat. *t* by Forec. *i* | | Rank-Order Correlation for Forecaster *i* | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| ***Indicator for Group (Omitted Category: Academic Experts)*** | | | | |
| **PhD Students** | 1.64 | 0.42 | 0.071** | 0.073** |
| | (7.25) | (7.17) | (0.033) | (0.033) |
| **Undergraduate Students** | -18.42** | -11.56 | 0.037 | 0.042 |
| | (7.88) | (7.99) | (0.033) | (0.034) |
| **MBA Students** | -28.76*** | -30.41*** | -0.040 | -0.033 |
| | (7.84) | (9.30) | (0.034) | (0.041) |
| **Mturk Workers (College Degree)** | -88.78*** | -74.96*** | 0.030 | 0.043 |
| | (7.81) | (8.70) | (0.028) | (0.032) |
| **Mturk Workers (No College Degree)** | -117.47*** | -105.60*** | -0.014 | -0.003 |
| | (8.98) | (9.77) | (0.029) | (0.032) |
| **Control for Survey Time, Click Practice, Click Instructions, and Missing Click:** | | X | | X |
| **Fixed Effects:** | Fixed Effects for Treatment and for Order of Treatments | | | |
| **Sample:** | Academic Experts, PhD Students, Undergraduate Students, MBA Students, Mturk Workers | | | |
| ***N*** | 21525 | 21525 | 1435 | 1435 |
| ***R Squared*** | 0.071 | 0.082 | 0.009 | 0.055 |

**Notes:** The table reports the result of OLS regressions of measures of forecast accuracy on other forms of expertise. In Columns (1)-(2) the dependent variable is the (negative of) the absolute forecast error and an observation in the regression is a forecaster-treatment combination. In Columns (3)-(4), the dependent variable is the rank-order correlation between forecast and actual effort across the treatments, and each observation is a forecaster i. Columns (1)-(2) include fixed effects for the order in which the expert encountered a treatment (to control for fatigue) and fixed effects for the treatment. Columns (2) and (4) include as controls time to survey completion and whether the forecaster clicked on practice or the instructions. Standard errors are clustered by individual.

* significant at 10%; ** significant at 5%; *** significant at 1%

## Table 6. Impact of Revealed Accuracy, Effort, and Motivation

| Dep. Var. (Measure of Accuracy): | (Negative of) Absolute Forecast Error in Treatment *t* by Forecaster *i* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| ***Measures of Revealed Accuracy*** | | | |
| **(Negative of) Absolute Error in Forecast of 4-cent Piece Rate Treatment / 100** | 9.60** | 23.90*** | 31.14*** |
| | (3.74) | (3.08) | (1.89) |
| ***Controls for Time to Completion (Omitted 5-9 minutes)*** | | | |
| **Survey Completion Time 0-4 Minutes** | . | -36.43 | -17.36 |
| | | (41.35) | (16.81) |
| **Survey Completion Time 10-14 Minutes** | -15.05 | -11.67 | 19.38** |
| | (11.97) | (10.52) | (9.46) |
| **Survey Completion Time 15-24 Minutes** | -13.49 | -4.10 | 20.83* |
| | (13.53) | (9.46) | (11.99) |
| **Survey Completion Time 25+ Minutes** | -29.53** | 1.72 | -10.31 |
| | (12.89) | (10.16) | (22.20) |
| ***Control for Confidence*** | | | |
| **Number of Own Answers Expected Within 100 Points of Actual** | 0.50 | 3.78*** | 5.44*** |
| | (1.47) | (1.21) | (1.37) |
| ***Measures of Attention to Instructions*** | | | |
| **Clicked on Practice Task** | -3.29 | -8.02 | |
| | (8.42) | (9.57) | |
| **Clicked on Full Instructions** | 3.84 | -23.43 | |
| | (10.49) | (16.48) | |
| ***Mturk Education*** | | | |
| **College Degree** | | | 12.24 |
| | | | (8.24) |
| **Fixed Effects:** | Fixed Effects for Treatment 1-14 and for Order 1-14 of Treatments | | |
| **Sample Indicators Interacted with Fixed Effects:** | | X | |
| **Indicator for Missing Confidence Variable:** | X | X | X |
| **Indicator for Missing Click:** | | X | |
| **Controls for Expertise:** | X | | |
| **Sample:** | Academic Experts | PhDs, Undergr., MBAs | Mturk Workers |
| ***N*** | 2912 | 6510 | 10668 |
| ***R Squared*** | 0.115 | 0.124 | 0.164 |

**Notes:** The table reports the result of OLS regressions of forecast accuracy on measures of revealed forecasting accuracy. The dependent variable is the (negative of) the absolute forecast error and an observation in the regression is a forecaster-treatment combination, with each forecaster providing forecasts for 14 treatments. These regressions examine whether being more accurate in the forecast of a (non-behavioral) treatment increases the accuracy of forecasts in other treatments as well. The regressions also includes an indicator for missing confidence, as well as the other listed variables. The specification in Column (1) also includes controls for rank, decile of citations, and for field of expertise of the academic experts. The regressions also include fixed effects for the order in which the expert encountered a treatment (to control for fatigue) and fixed effects for the treatment. Standard errors are clustered by individual.

* significant at 10%; ** significant at 5%; *** significant at 1%

## Table 7. Accuracy of Optimal Forecasters

| | Individual Accuracy | | Wisdom-of-Crowds | |
|---|---|---|---|---|
| | Average Accuracy (and s.d.) of *Individual* Forecasts | Difference in Accuracy Relative to All Academic Experts | Accuracy (and s.d. of bootstrap) of *Mean* Forecast of group of *N* forecasters | |
| | | | N = 20 | N = 50 |
| | (1) | (2) | (3) | (4) |
| **Mean Absolute Error** | | | | |
| *Panel A. Academic Experts* | | | | |
| All Academic Experts (N=208) | 175.21 | | 101.18 | 96.63 |
| | (58.37) | | (12.73) | (7.95) |
| Optimal 20% (4ct Control) (N=42) | 173.14 | -2.07 | 102.07 | 97.59 |
| | (60.54) | (8.21) | (10.62) | (6.56) |
| Optimal 20% (No 4ct Control) (N=42) | 175.06 | -0.15 | 102.6 | 98.12 |
| | (58.66) | (8.02) | (10.56) | (6.53) |
| *Panel B. PhD/Undergraduates/MBA* | | | | |
| All PhD/UG/MBA (N=465) | 188.89 | 13.68** | 100.05 | 95.5 |
| | (83.25) | (5.59) | (16.42) | (10.11) |
| Optimal 20% (4ct Control) (N=93) | 147.97 | -27.24*** | 76.34 | 72.9 |
| | (42.26) | (5.95) | (11.28) | (7.14) |
| Optimal 20% (No 4ct Control) (N=93) | 166.21 | -9.00 | 87.38 | 83.84 |
| | (67.40) | (8.05) | (13.28) | (8.42) |
| *Panel C. Mturks* | | | | |
| All Mturks (N=762) | 272.02 | 96.81*** | 148.62 | 145.36 |
| | (143.23) | (6.58) | (39.31) | (25.37) |
| Optimal 20% (4ct Control) (N=152) | 189.15 | 13.94* | 81.73 | 76.76 |
| | (82.04) | (7.78) | (17.26) | (11.65) |
| Optimal 20% (No 4ct Control) (N=152) | 224.55 | 49.34*** | 108.21 | 102.37 |
| | (128.52) | (11.16) | (27.79) | (18.29) |

**Notes:** The table reports the absolute error at both the individual and windom-of-crowds level for different groups, including "superforecasters". Panel A depicts the academic experts, Panel B the students, and Panel C the Mturk workers. Within each panel, we consider the overall group and two subsamples of optimal forecasters. The subsamples are generated with a regression as in Table 6, determining with a 10-fold method the 20% predicted optimal forecasters out of sample. The last group of optimal forecasters is generated not using the revealed-accuracy variable based on the forecast for the 4-cent treatment. In Column (1) we report the average individual accuracy for the groups, and in parentheses are the standard deviations of the average individual absolute errors not including the 4-cent treatment. In Column (2) we test for differences relative to the sample of all 208 academic experts. In Columns (3) and (4) we present wisdom-of-crowd average group-level accuracy for each of the groups. We sample 1500 groups of 20 (column 3) and 50 (column 4) at each row, and compute the absolute error for the average forecast in the group - first averaging over the group, and then across treatments. In parantheses are the SD of the bootstrapped average absolute errors.

* significant at 10%; ** significant at 5%; *** significant at 1%

## Appendix Figure 1. Expert Survey, Screenshot from Page 2 of Survey

Of the 15 predictions that you made, what is your best guess as to how many of your predictions are within 100 points of the actual average scores?

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

We are planning to administer this survey to different groups of people (professors, MBA students, etc.). We would like to know how you think various groups will perform in this prediction task.

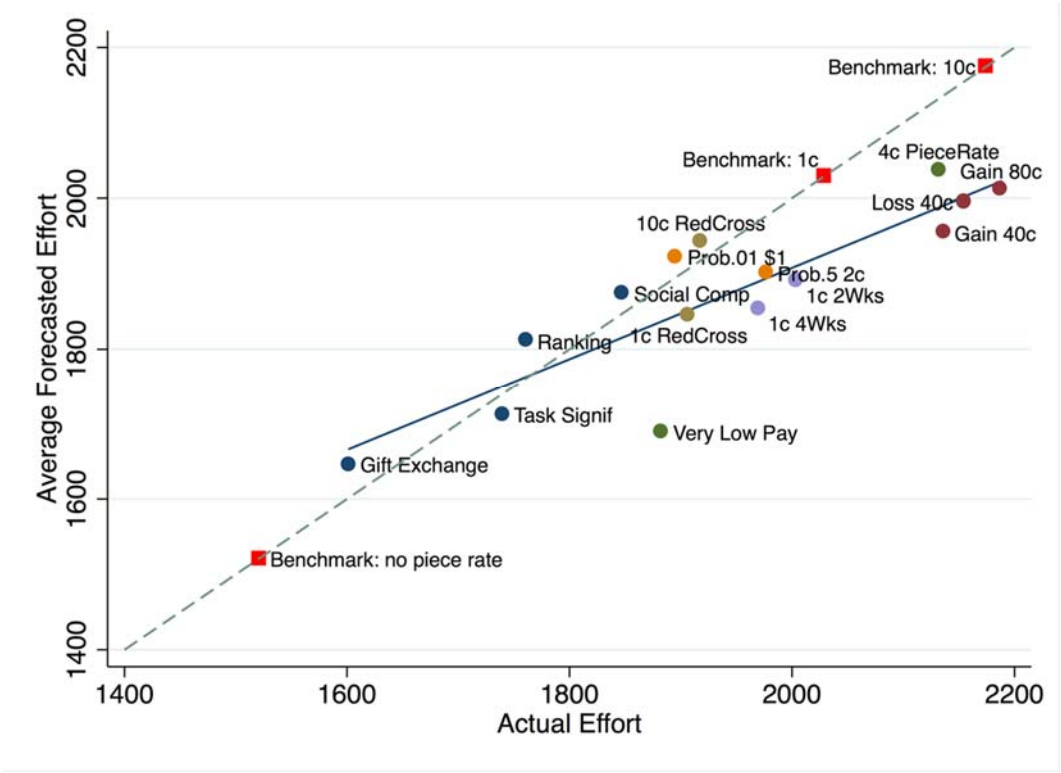
For each group of people below, please indicate your best guess as to the average number of predictions that members of that group will make that are within 100 points of the actual average scores. Thus, a higher number means you think a particular group is more likely to be accurate in their predictions.

| | Number of predictions within 100 points of actual average scores | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Group #1: Professors with expertise in behavioral economics or decision making who recently presented at or served on a program committee for select behavioral economics or decision making conferences (e.g. SITE and BDRM) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Group #2: The 15 most-cited professors from Group #1 who respond to our survey | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Group #3: Professors from Group #1 with a PhD in economics | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Group #4: Professors from Group #1 with a PhD in psychology or decision making | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Group #5: PhD students in economics from UC Berkeley and the University of Chicago | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Group #6: PhD students from Group #5 who are specializing in behavioral economics | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Group #7: MBA students from the Booth School of Business at the University of Chicago | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Group #8: MTurk workers who make predictions after completing the button-pushing task in one of the conditions | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Group #9: MTurk workers who make predictions without participating in the button-pushing task | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

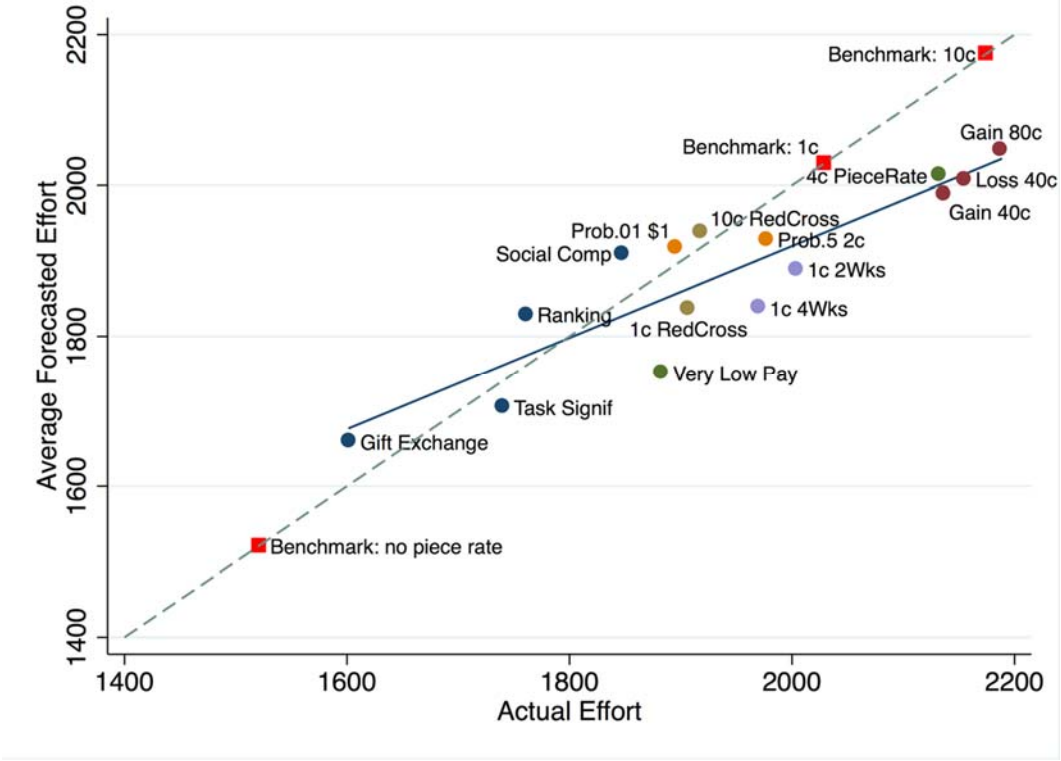
**Notes:** Appendix Figure 1 shows a screenshot reproducing portions of page 2 of the Qualtrics survey which experts used to make forecasts.

**Appendix Figures 2a-c. Wisdom-of-Crowds Accuracy, Other Groups**

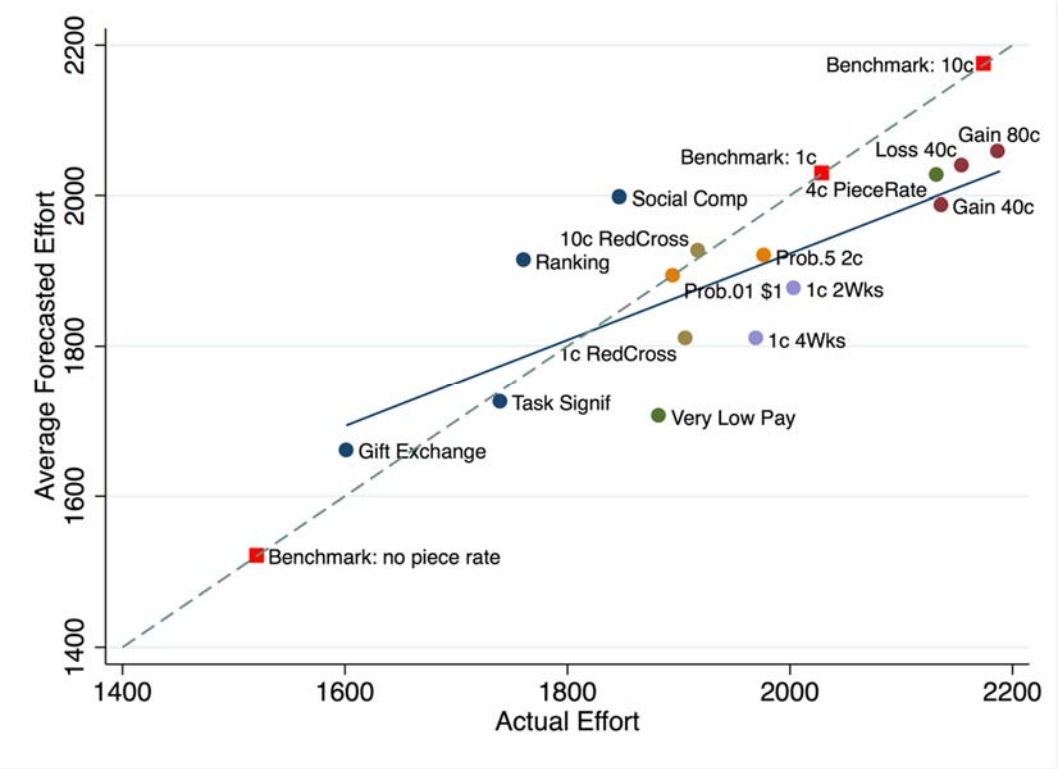
**Appendix Figure 2a. PhD Students**




**Appendix Figure 2b. Undergraduate Students**



**Appendix Figure 2c. MBA Students**




**Notes:** These figures present the parallel evidence to Figure 2 and Figure 6e for the other samples of forecasters.

## Appendix Table 1. Summary Statistics, All Groups of Forecasters

| | Academic Experts, Invited to Participate | Academic Experts, Completed Survey | PhD Students | Undergraduate Students | MBA Students | Mturk Workers |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| **Academic Rank (Academic Experts)** | | | | | | |
| Assistant Professor | 0.26 | 0.36 | | | | |
| Associate Professor | 0.15 | 0.15 | | | | |
| Professor | 0.55 | 0.45 | | | | |
| Other | 0.04 | 0.04 | | | | |
| **Citations (Academic Experts)** | | | | | | |
| Google Scholar Citations | 7742 | 6326 | | | | |
| **Primary Field (Academic Experts)** | | | | | | |
| Behavioral Econ. | 0.30 | 0.36 | | | | |
| Applied Micro | 0.17 | 0.19 | | | | |
| Economic Theory | 0.09 | 0.07 | | | | |
| Econ. Lab Exper. | 0.17 | 0.16 | | | | |
| Social Psych or Decision Making | 0.26 | 0.22 | | | | |
| Field Behavioral Econ. (PhDs) | | | 0.24 | | | |
| **Heard of Mturk** | | 0.98 | 0.73 | 0.25 | 0.31 | . |
| **Used Mturk** | | 0.51 | 0.17 | 0.03 | 0.02 | . |
| **Minutes Spent (capped at 50)** | | 21.21 | 21.46 | 16.06 | 21.86 | 10.09 |
| **Clicked Practice Task** | | 0.44 | 0.48 | 0.11 | 0.12 | 0.00 |
| **Clicked Instructions** | | 0.22 | 0.18 | 0.01 | 0.04 | 0.00 |
| **Days Waited Till Survey Completion** | | 11.36 | 3.90 | 2.99 | 2.47 | 0.00 |
| **Confidence (Expected No. Own Forecasts Within 100 Pts. of Actual)** | | 5.77 | 6.53 | 6.32 | 5.66 | 6.81 |
| **Absolute Error in 4c Treatment** | | 88.34 | 103.89 | 162.80 | 125.57 | 265.22 |
| **Observations** | 312 | 208 | 147 | 158 | 160 | 762 |

**Notes:** The table presents summary statistics for the samples used in the survey: the academic experts (Columns 1 and 2), the PhD students (Column 3), the undergraduate students (Column 4), the MBA students (Column 5), and the Mturk workers (Column 6). Columns 1 and 2 compare characteristics of the overall sample of academic experts contacted (Column 1) versus the characteristics of the experts that completed the forecast survey (Column 2).

# A Online Appendix A - Survey Details

We decided ex ante the rule for the scale of 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. (It is physically very hard to obtain scores above 3,500.) 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. From the email chain on 6/10/2015, we emailed 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 we received the response: "*All of the average treatment counts are between 1,200 and 2,300*".

**Experts.** 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.

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*). 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 the 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 (and extrapolate additional citations for papers beyond the top 25 from citations for the 16th-25th most-cited papers on Google Scholar).

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.

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.

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. Since the survey included 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.

**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, 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.

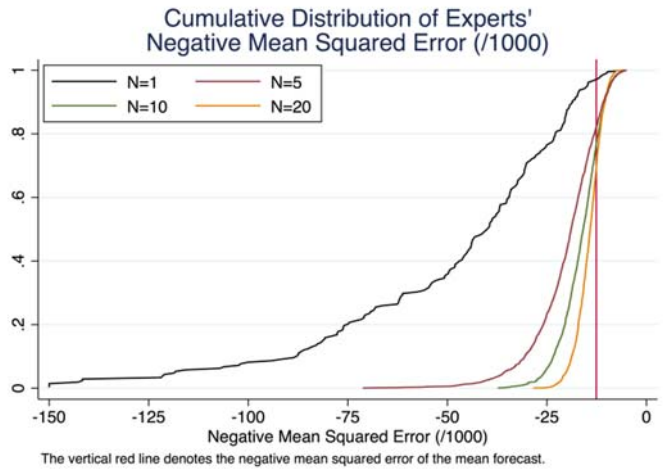
On 10/4/2016, we recruited MTurk workers (who were not involved in the initial experiment) to do a 10-minute task and take a 10-15 minute survey for a $1.50 fixed payment. These participants obviously have direct experience with working on MTurk and may have a better sense than academics or others about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

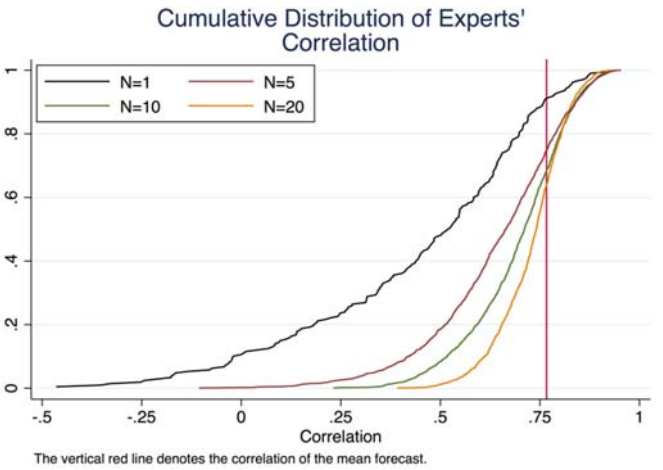
On 2/12/2016 we recruited an additional sample of MTurk workers (N= 258) who were not involved with any of the previous MTurk tasks. Like the 'experienced' MTurk sample above, they first participated in the 10-minute button-pressing task and then took the forecasting survey. For this sample, however, we made especially salient the value of trying hard when making their forecasts. 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 $0.50.

**Online Appendix Figure 1. Individual Expert Accuracy versus Aggregate (Wisdom-of-Crowds) Accuracy, Additional Accuracy Measures**

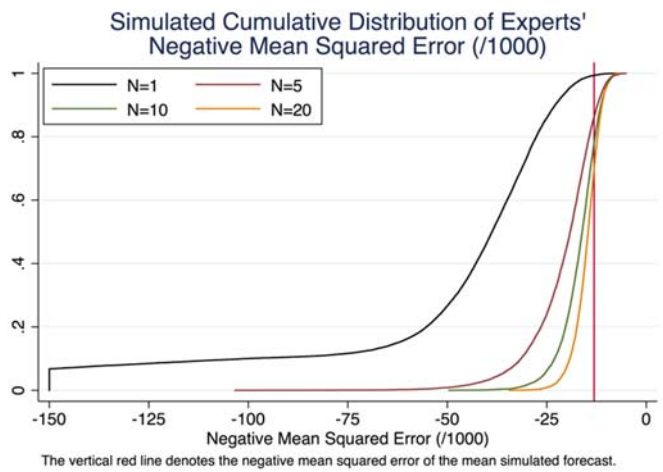
**Onl. App. Figure 1a. Mean Squared Error, Data**



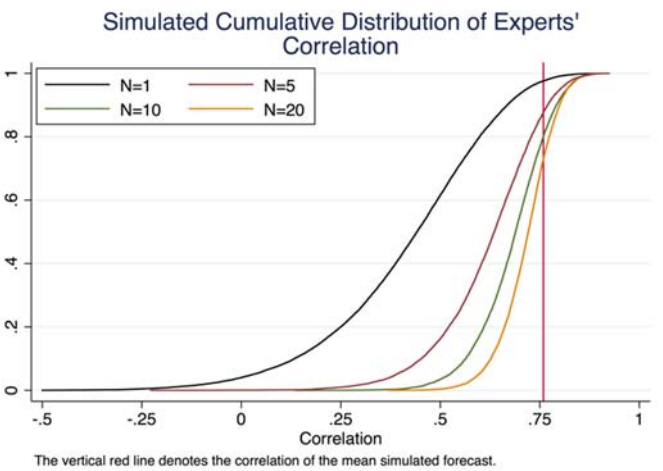

**Onl. App. Figure 1b. Pearson Correlation, Data**




**Onl. App. Figure 1c. Mean Squared Error, Model**



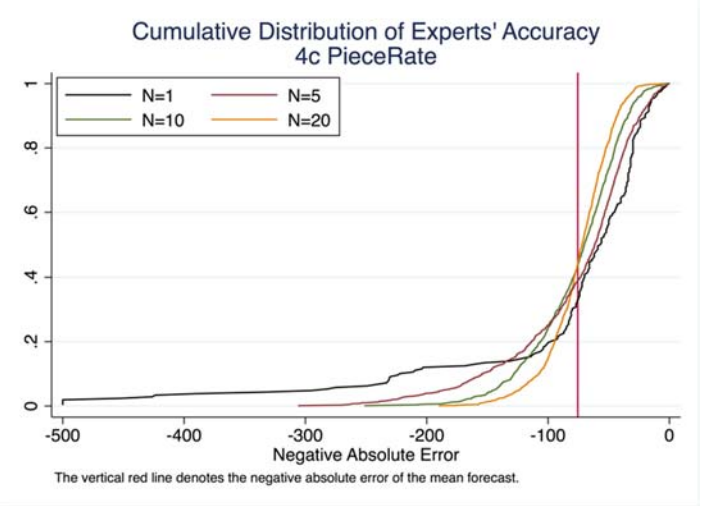

**Onl. App. Figure 1d. Pearson Correlation, Model**



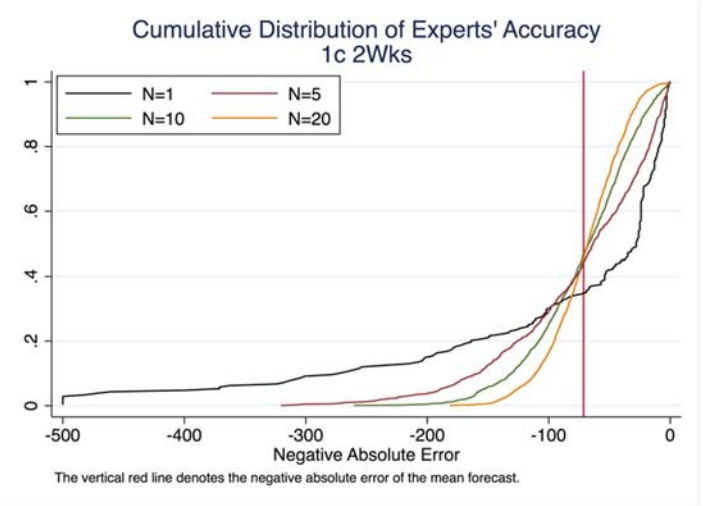

**Notes:** Online Appendix Figure 1 presents the same information as in Figure 3 in the text, but for different measures of forecaster accuracy: the (negative of) the mean squared error, and the Pearson correlation between the forecast and the treatment results.

**Online Appendix Figure 2. Individual Expert Accuracy versus Aggregate (Wisdom-of-Crowds) Accuracy, Representative Treatments**

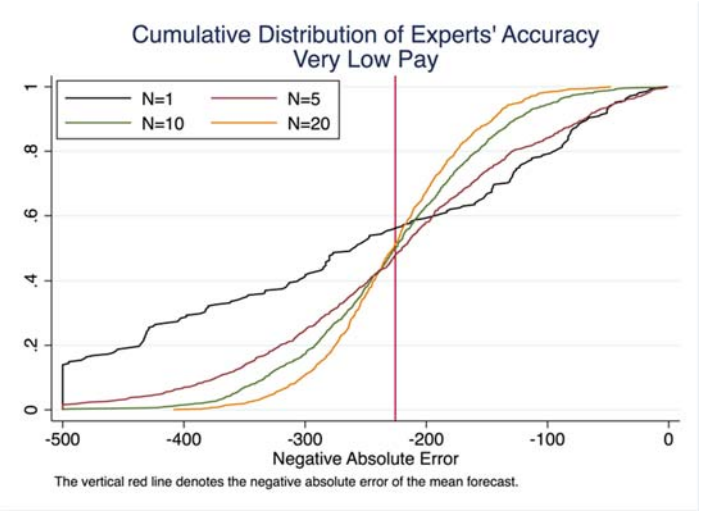
**Onl. App. Figure 2a. 4-cent Piece Rate Treatment**



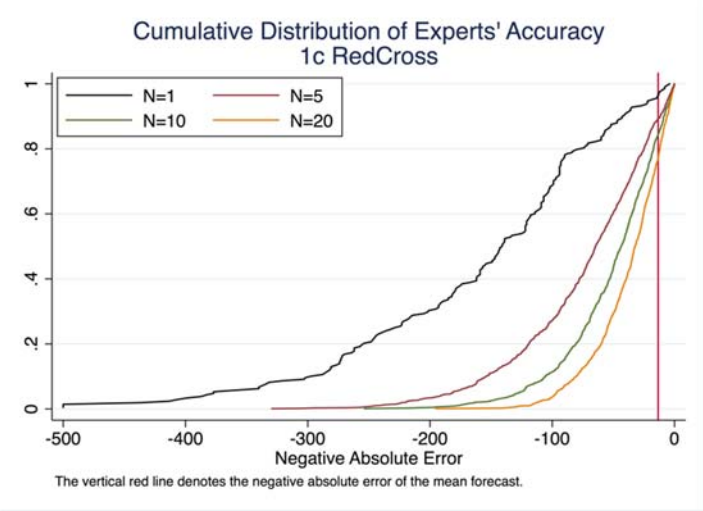

**Onl. App. Figure 2b. 1-cent-in-2-weeks Treatment**




**Onl. App. Figure 2c. Very Low Pay Treatment**

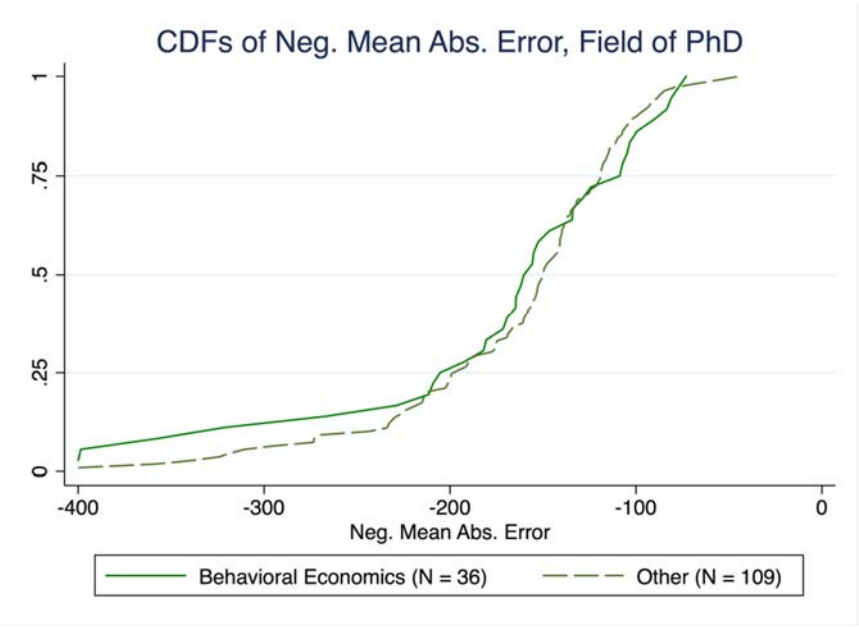



**Onl. App. Figure 2d. 1-cent-Charity Treatment**




**Notes:** The figure presents the same information as in Figure 3a for four treatments using the negative of the absolute mean error as accuracy measure. Graphs are censored at -500.

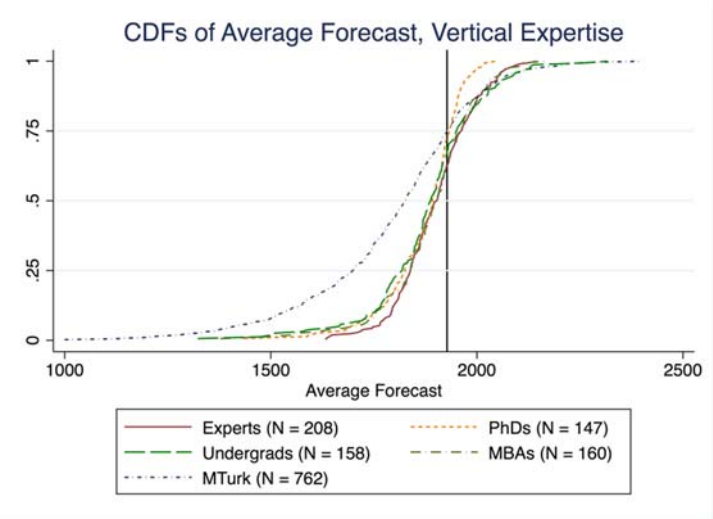
**Online Appendix Figure 3. Horizontal Expertise for PhD Students**



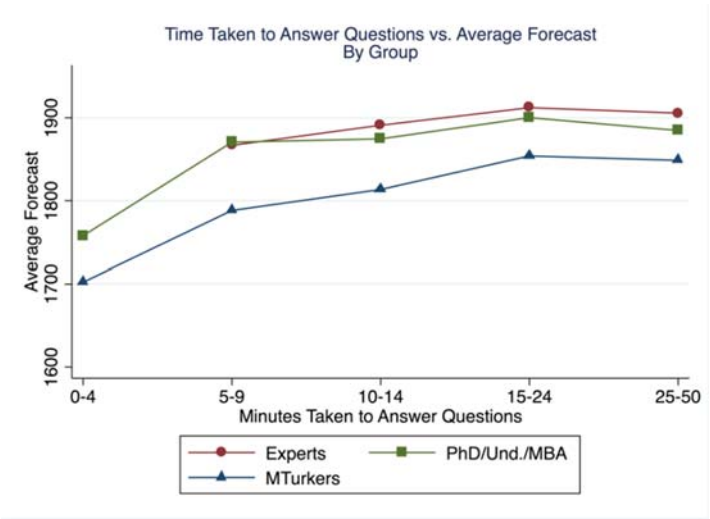

**Notes:** Online Appendix Figure 3 presents the c.d.f. for the negative of the mean absolute error for the PhD students participating depending on whether the (self-reported) field of specialization is Behavioral Economics or other.

**Online Appendix Figures 4a-d. Average Forecast Across All 15 Treatments, Key Findings**

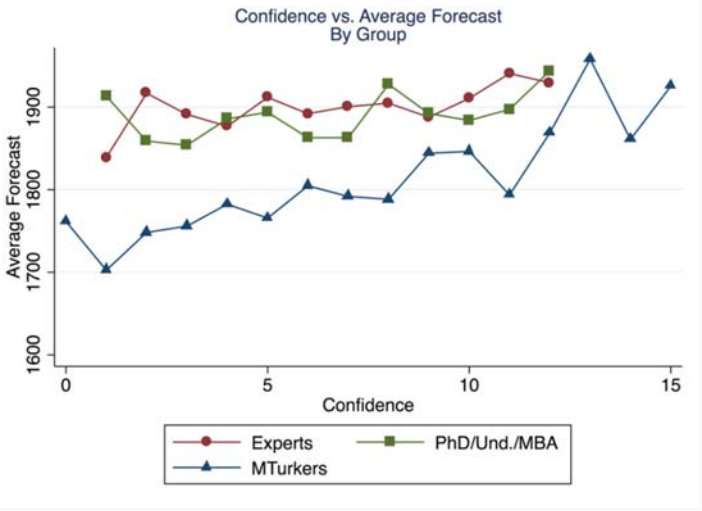
**Onl. App. Figure 4a. Distribution of Average Forecast**



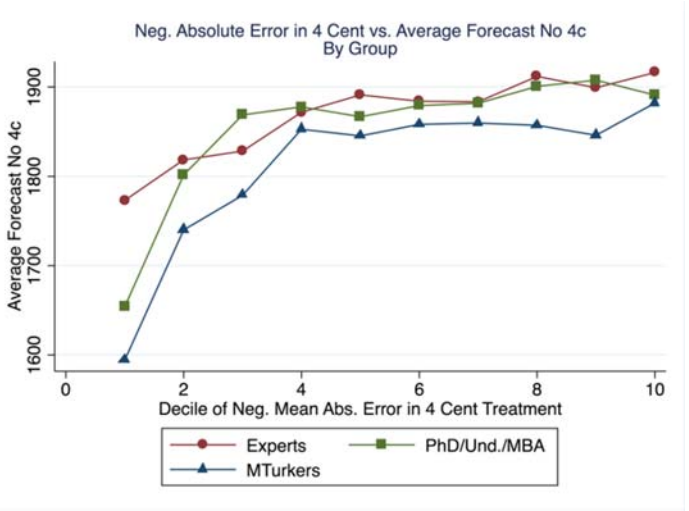

**Onl. App. Figure 4b. By Time to Completion**




**Onl. App. Figure 4c. By Confidence**



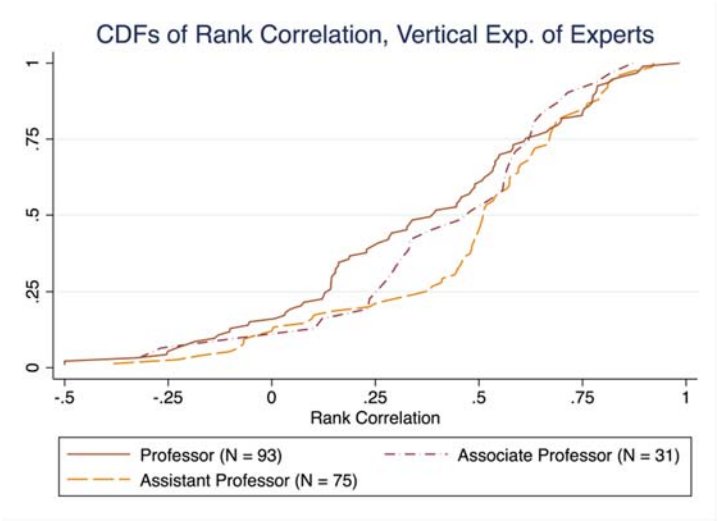

**. App. Figure 4d. By Accuracy in 4-cent Treatment**



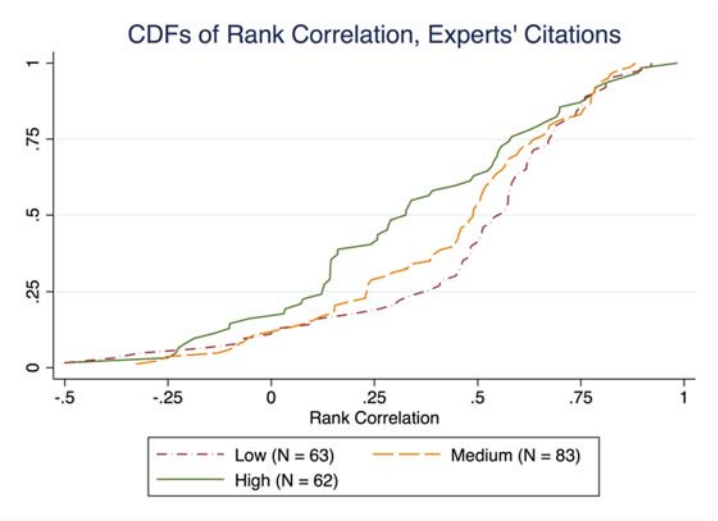

**Notes:** These figures present evidence on the average forecast across the 15 treatments. Online Appendix Figure 4a shows that MTurkers are much more likely to have offered a low forecast relative to the average actual effort (vertical black line). Online Appendix Figures 4b-d show that the average forecast increases in the time taken to do the survey (Figure 4b), in the confidence (Figure 4c), and in the accuracy of forecast of the 4c treatment (Figure 4d).

**Online Appendix Figures 5a-d. Key Findings on Vertical, Horizontal, and Contextual Expertise, Rank-Order Correlation**

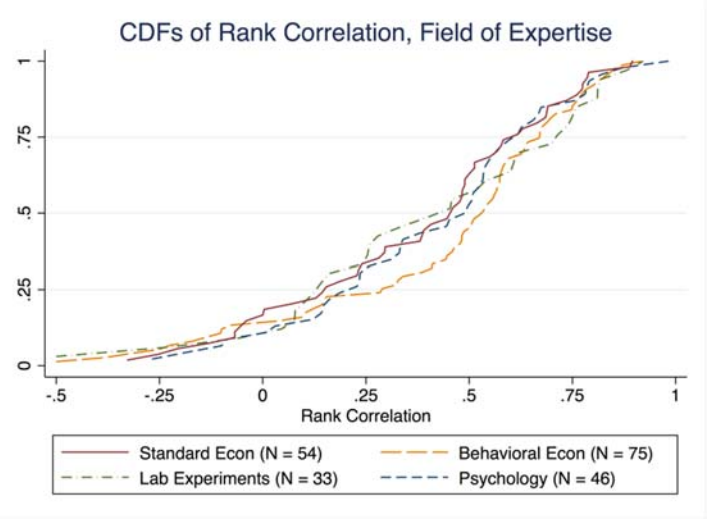
**Appendix 5a. Academic Rank (Vertical Expertise)**



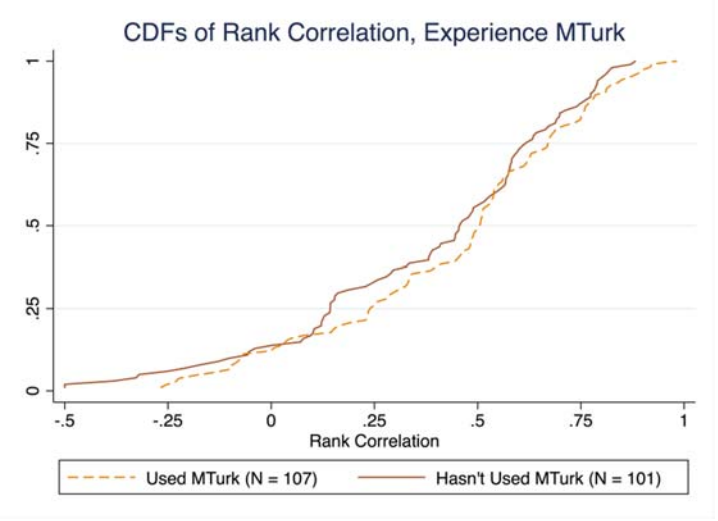

**Appendix Figure 5b. Citations (Vertical Expertise)**




**Appendix Figure 5c. Fields (Horizontal Expertise)**



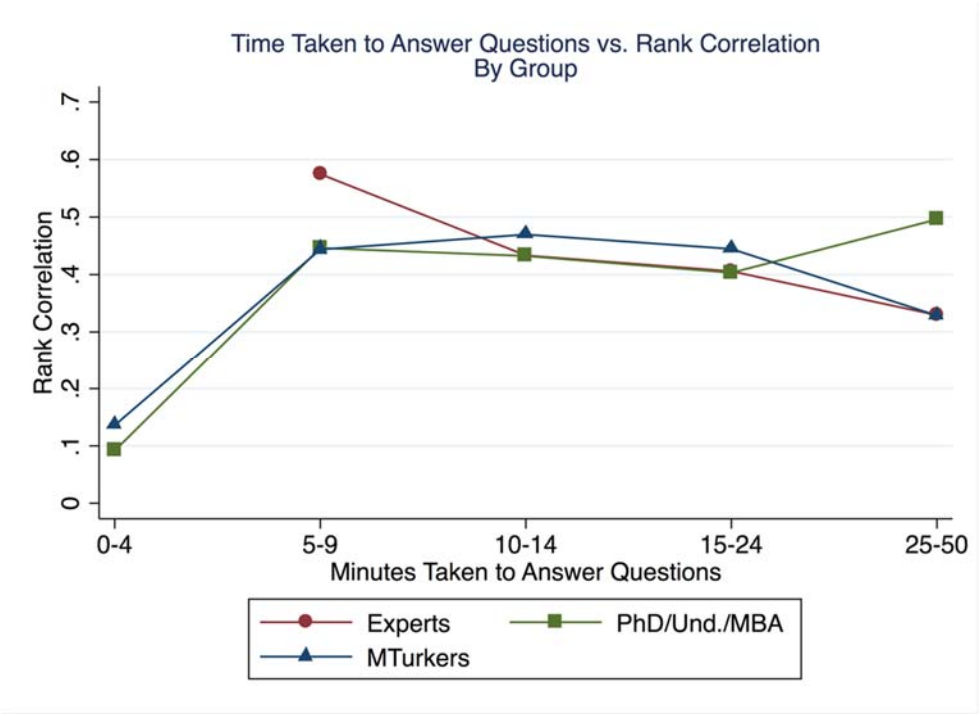

**Appendix Figure 5d. Experience with MTurk Platform (Contextual Expertise)**



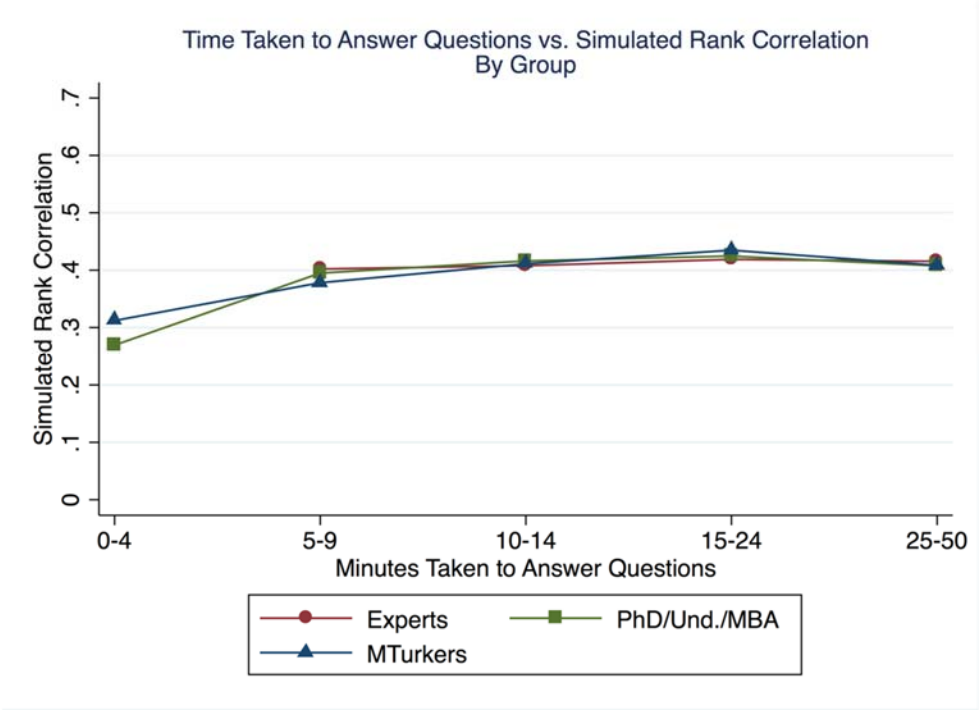

**Notes:** These figures replicate key results on vertical, horizontal, and contextual expertise using the rank-order correlation measure.

**Online Appendix Figure 6. Accuracy and Effort in Taking Task, Rank-Order Correlation**

**Onl. App. Figure 6a. Time Taken in Completing the Survey, Data**

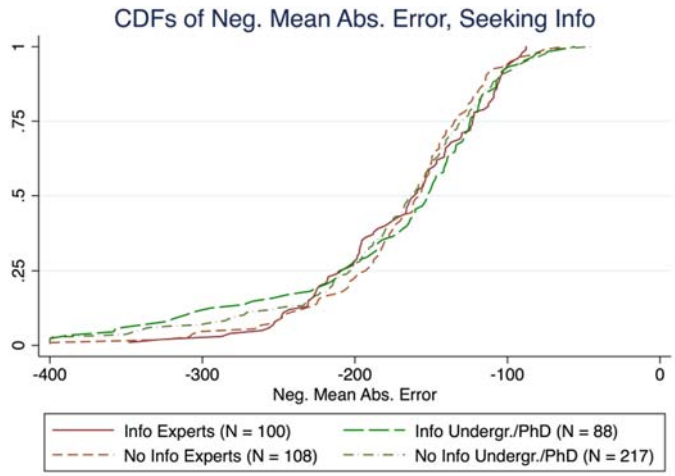


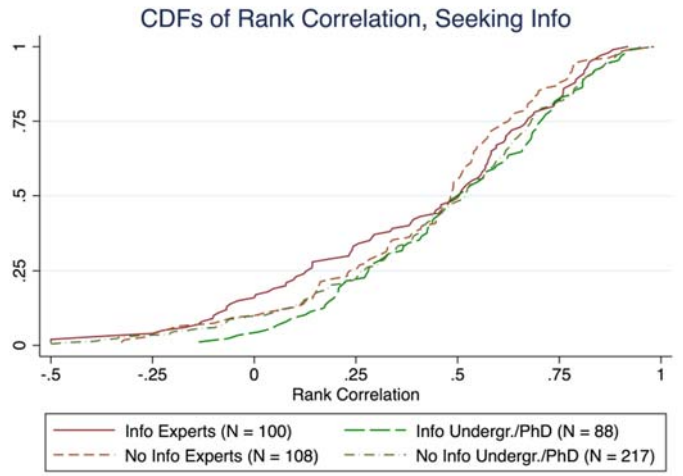

**Onl. App. Figure 6a. Time Taken in Completing the Survey, Model**



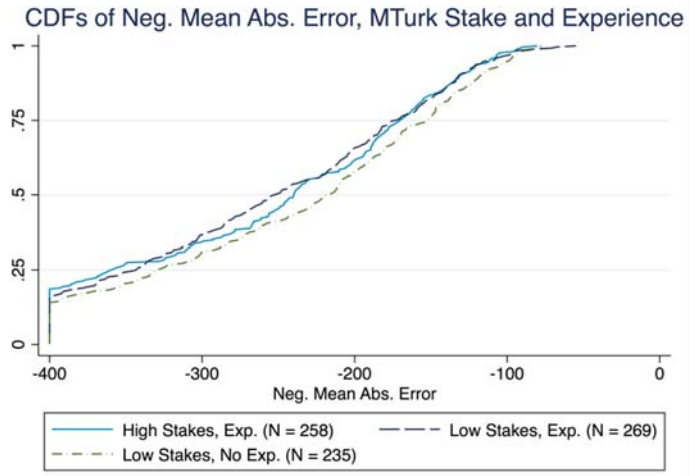

**Notes:** Online Appendix Figure 6a plots the accuracy for three groups of forecasters (academic experts; undergraduate, MBA, and PhD students; and MTurkers) as a function of how long they took to complete the survey. Specifically, the figures plot the average accuracy by minutes of time taken for survey completion. Onl. App. Figure 6b plots the corresponding figure for simulated data at the model estimates.

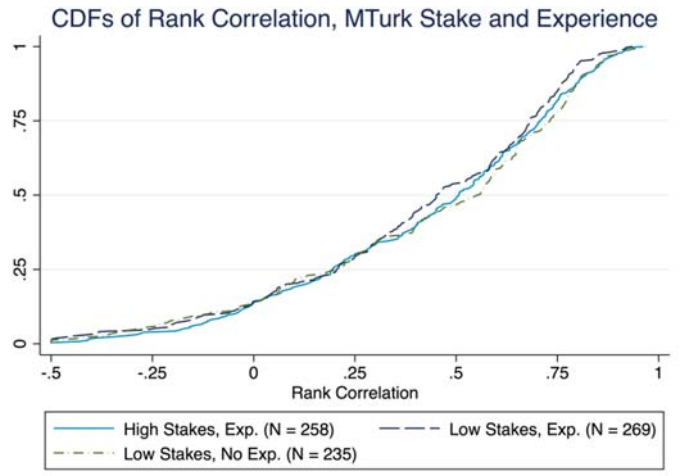
**Online Appendix Figures 7a-b. Expert Checked Task or Full Instructions**




**Online Appendix Figures 7c-d. Effect of Stake Size and Experience on Motivation, MTurk Sample**

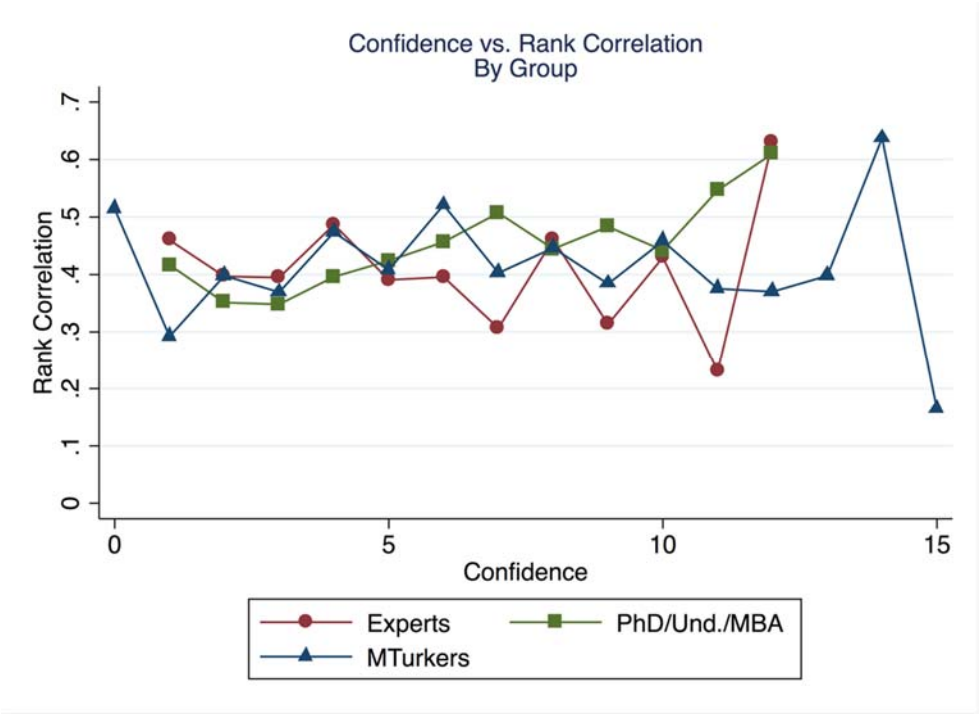


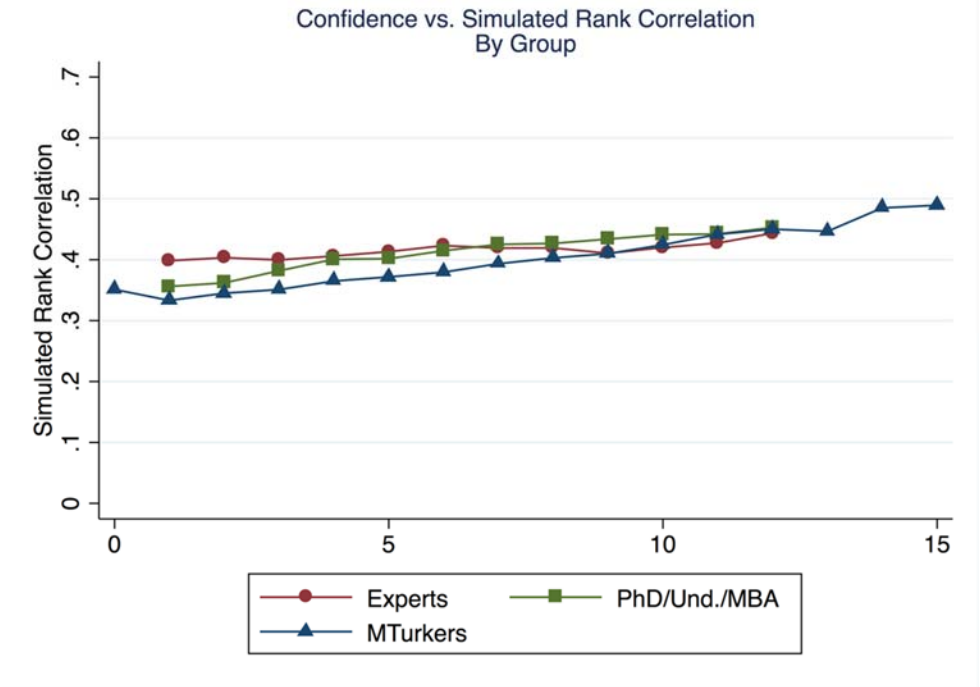

**Notes:** Online Appendix Figures 7a-b split two of the groups into whether they clicked on a link for a trial of the task or the link for additional instructions. (The MTurk group is excluded because no one in the group clicked on the link). Online Appendix Figures 7c-d compare three MTurk subgroups who differ in the incentives for survey accuracy and experience with the task. The low-stake group is informed that 5 out of the responses would be eligible for up to $100 for accuracy. The high-stake group is informed that each respondent will receive up to $5 for accuracy of the survey responses. Experienced groups experienced the task before making forecasts.

**Online Appendix Figure 8. Accuracy and Confidence in Taking Task, Rank-Order Correlation**

**Onl. App. Figure 8a. Confidence, Data**



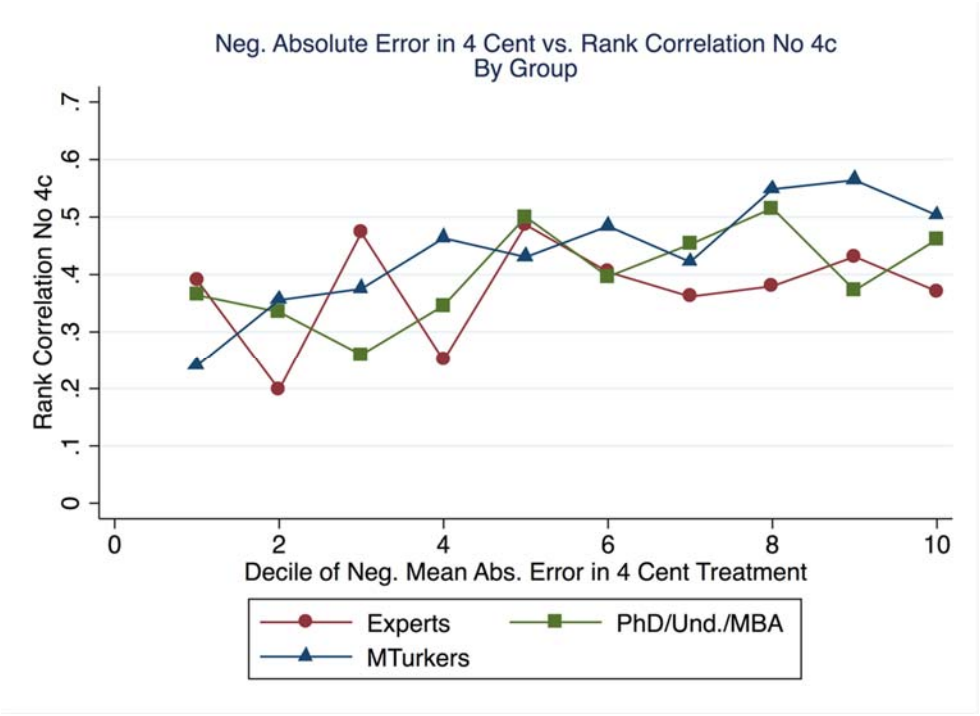

**Onl. App. Figure 8a. Confidence, Model**



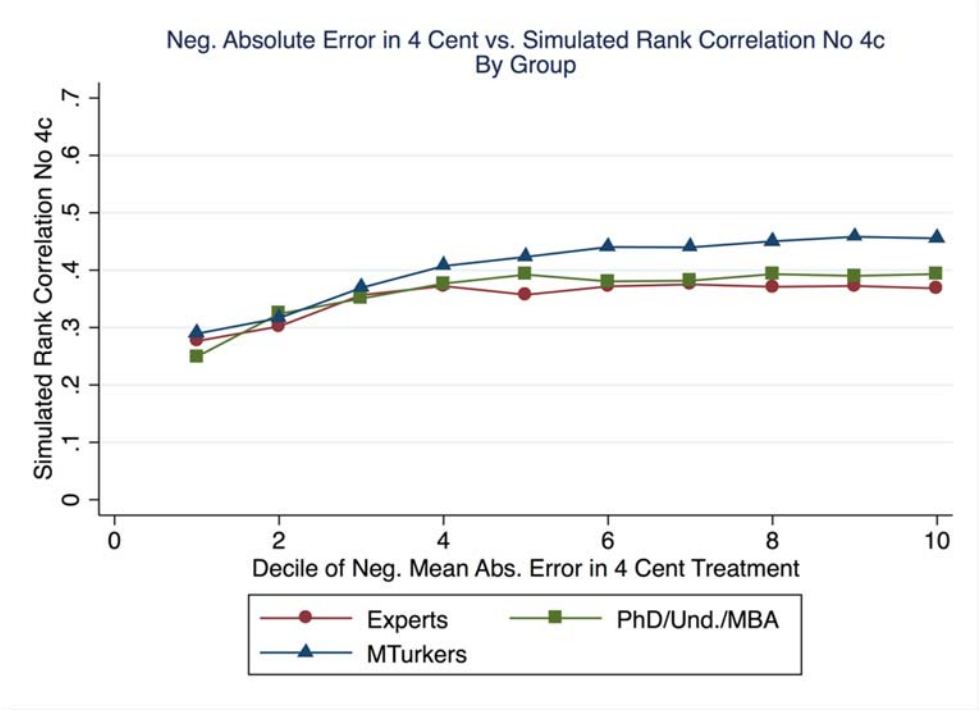

**Notes:** Online Appendix Figure 8a plots the average accuracy for three groups of forecasters (academic experts, undergraduate/MBA/PhD students, and MTurkers) by how confident the respondent felt about the accuracy. In particular, each survey respondent indicated how many out of 15 forecasts he or she made were going to be accurate up to 100 points relative to the truth. Onl. App. Figure 8b plot the corresponding figure from simulated data at the model estimate.

**Online Appendix Figure 9. Accuracy and Revealed Accuracy, Rank-Order Correlation**

**Onl. App. Figure 9a. Accuracy in 4-cent Treatment, Data**




**Onl. App. Figure 9b. Accuracy in 4-cent Treatment, Model**




**Notes:** Online Appendix Figure 9a plots the average accuracy for three groups of forecasters (academic experts, undergraduate/MBA/ PhD students, and MTurkers) by decile of a revealed-accuracy measure (the decile thresholds are computed using all three groups). Namely, we take the absolute distance between the forecast and the actual effort for the 4-cent piece rate treatment, a treatment for which the forecast should not involve behavioral factors. For these plots the accuracy measure is computed excluding the 4-cent treatment. Onl. App. Figure 9b plots the corresponding evidence from simulations at the model estimate.

## Online Appendix Table 1. Summary Statistics, Mturk

| | Mean | US Census |
|---|---|---|
| | (1) | (2) |
| Button Presses | 1936 | |
| Time to complete survey (minutes) | 12.90 | |
| US IP Address Location | 0.85 | |
| India IP Address Location | 0.12 | |
| Female | 0.54 | 0.52 |
| Education | | |
| High School or Less | 0.09 | 0.44 |
| Some College | 0.36 | 0.28 |
| Bachelor's Degree or more | 0.55 | 0.28 |
| Age | | |
| 18-24 years old | 0.21 | 0.13 |
| 25-30 years old | 0.30 | 0.10 |
| 31-40 years old | 0.27 | 0.17 |
| 41-50 years old | 0.12 | 0.18 |
| 51-64 years old | 0.08 | 0.25 |
| Older than 65 | 0.01 | 0.17 |
| Observations | 9861 | |

**Notes:** Column (1) of Online Appendix Table 1 lists summary statistics for the final sample of Amazon Turk survey participants (after screening out ineligible subjects). Column (2) lists, where available, comparable demographic information from the US Census.

**Online Appendix Table 2. Model Estimates, Fit and Robustness**

| | | Benchmark Estimates | No Heterog. in Idiosyncratic Std. Dev. | No Heterog. in Forecast Bias | One Type (No Heterogen.) | Three Types |
|---|---|---|---|---|---|---|
| | | (1) | (2) | (3) | (4) | (5) |
| **Panel A. Model Estimates** | | | | | | |
| $v^{(1)}$ (Average Bias, Type 1) | | -24.9 (2.3) | -33.6 (1.9) | -69.1 (1.7) | -100.6 (1.9) | -2.22 (3.1) |
| $v^{(2)}$ (Average Bias, Type 2) | | -193.2 (5.5) | -600.8 (6.8) | | | -68.1 (3.9) |
| $v^{(3)}$ (Average Bias, Type 3) | | | | | | -705.6 (12.5) |
| $\sigma^{(1)}$ (Idiosyncratic s.d., Type 1) | | 162.6 (2.7) | 213.4 (1.3) | 169.6 (2.6) | 281.2 (1.4) | 83.2 (5.7) |
| $\sigma^{(2)}$ (Idiosyncratic s.d., Type 2) | | 357.6 (3.5) | | 374.2 (4.2) | | 252.0 (3.7) |
| $\sigma^{(3)}$ (Idiosyncratic s.d., Type 3) | | | | | | 174.8 (6.8) |
| Log Likelihood | | -150,184 | -150,388 | -150,701 | -151,924 | -149,986 |
| **Panel B. Moments Implied by Model Estimates** | **Data** | **Estimates** | **Estimates** | **Estimates** | **Estimates** | **Estimates** |
| | **Experts** | **Mturks** | **Experts** | **Mturks** | **Experts** | **Mturks** | **Experts** | **Mturks** | **Experts** | **Mturks** | **Experts** | **Mturks** |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Average Individual Absolute Error | *169.4* | *271.6* | 175.3 | 267.7 | 193.6 | 264.3 | 182.1 | 262.3 | 252.0 | 243.1 | 180.0 | 269.8 |
| Average Absolute Error with 5 Forecasters | *113.6* | *174.4* | 116.2 | 169.4 | 119.2 | 172.6 | 123.1 | 137.9 | 154.0 | 139.7 | 117.7 | 172.7 |
| Average Absolute Error with 10 Forecasters | *104.3* | *155.3* | 105.2 | 153.9 | 106.1 | 160.8 | 112.2 | 111.8 | 136.4 | 123.1 | 106.0 | 160.1 |
| Average Absolute Error with 20 Forecasters | *99.0* | *151.7* | 99.9 | 147.5 | 99.9 | 156.4 | 107.1 | 97.2 | 126.9 | 114.6 | 100.1 | 153.7 |
| Wisdom-of-Crowds Absolute Error | *93.5* | *146.9* | 94.1 | 143.7 | 93.1 | 152.4 | 100.9 | 83.5 | 115.0 | 110.9 | 94.0 | 150.1 |
| *Rank-Order Correlation:* | | | | | | | | | | | | |
| Average Individual Rank-Order Correlation | *0.41* | *0.42* | 0.41 | 0.39 | 0.36 | 0.47 | 0.40 | 0.37 | 0.28 | 0.37 | 0.43 | 0.44 |
| Wisdom-of-Crowds Rank-Order Correlation | *0.83* | *0.95* | 0.81 | 0.95 | 0.81 | 0.95 | 0.81 | 0.94 | 0.80 | 0.95 | 0.81 | 0.95 |
| *Percent Individual Forecasters Outperforming* | | | | | | | | | | | | |
| Wisdom-of-Crowds Absolute Error | *4.3* | *17.8* | 1.0 | 18.7 | 0.2 | 18.8 | 0.8 | 0.1 | 0.1 | 0.1 | 7.6 | 9.3 |
| Wisdom-of-Crowds Rank-Order Correlation | *4.8* | *0.3* | 1.3 | 0.0 | 0.9 | 0.0 | 1.2 | 0.0 | 0.6 | 0.0 | 3.6 | 0.1 |
| *Cross-Treatment Correlation of Absolute Error* | | | | | | | | | | | | |
| Avg. Regression Correlation of Abs. Errors | *0.09* | *0.33* | 0.17 | 0.15 | 0.05 | 0.53 | 0.10 | 0.10 | 0.00 | 0.00 | 0.17 | 0.50 |

**Notes:** This table examines the robustness of the benchmark discrete heterogeneity model (specifically, column 1 of table 3), presenting estimates from several variants of this model and examining the goodness-of-fit by comparing key moments computed using model simulations to moments from the data. Panel A reports the estimated types for the various model specifications. In columns 1-3, only indicators of subject group are used as predictors of type, but the idiosyncractic s.d. and average bias of the forecasters are restricted to be constant across the two types respectively in columns 2 and 3. Column 4 presents the results for a model with only one type of forecaster whereas column 5 shows the results for a specification with 3 types of forecasters (with idiosyncractic s.d. and average bias allowed to vary for each type of forecaster and only indicators for subject groups used to predict types). The logit coefficients are not shown in this table due to space constraints. Panel B reports moments from simulated data corresponding to the various model specifications in the respective columns of panel A and compares them to moments from the actual data. The moments are computed separately for the 208 academic experts and 762 MTurks for maximum contrast, even though simulations are based on the full sample which also includes PhDs, MBAs and undergraduates. Reported moments for the simulated data are averages over 100 simulations. Within each simulation, we sample 5/10/20 forecasters at random with replacement 100 times to compute the average absolute error with 5/10/20 forecasters for that particular simulation. We do so 1,000 times for the same moments in the actual data for this table, since we cannot average over many realizations of the data as we do with the simulations.

## Online Appendix Table 3. Impact of *Vertical*, *Horizontal*, and *Contextual* Expertise on Forecast Accuracy

| **Dep. Var. (Measure of Accuracy):** | **Rank-Order Correlation for Forecaster *i*** | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| ***Measures of Vertical Expertise (Omitted: Assistant Professor)*** | | | | |
| **Associate Professor** | -0.05 | -0.06 | -0.02 | -0.02 |
| | (0.06) | (0.07) | (0.07) | (0.08) |
| **Full Professor** | -0.11** | -0.12** | -0.04 | -0.05 |
| | (0.05) | (0.05) | (0.09) | (0.09) |
| **Other (Post-Doc or Research Scientist)** | 0.10 | 0.15** | 0.10* | 0.14** |
| | (0.06) | (0.07) | (0.06) | (0.07) |
| **Decile Google Scholar Citations** | | | -0.01 | -0.01 |
| | | | (0.01) | (0.01) |
| ***Main Field of Expertise (Omitted: Behavioral Economics)*** | | | | |
| **Applied Microeconomics** | | | -0.05 | -0.05 |
| | | | (0.07) | (0.06) |
| **Economic Theory** | | | -0.03 | -0.09 |
| | | | (0.07) | (0.07) |
| **Laboratory Experiments** | | | -0.01 | -0.03 |
| | | | (0.07) | (0.07) |
| **Psychology or Behavioral Decision-Making** | | | -0.00 | -0.04 |
| | | | (0.07) | (0.07) |
| ***Measure of Contextual Expertise*** | | | | |
| **Has Used Mturk in Own Research (Self-Reported)** | | | 0.05 | 0.03 |
| | | | (0.05) | (0.05) |
| **Effort Controls: Survey Completion Time, Click on Practice Task, Click on Instructions, and Delay Start:** | | X | | X |
| **Sample:** | Academic Experts | | | |
| ***N*** | 208 | 208 | 208 | 208 |
| ***R Squared*** | 0.035 | 0.112 | 0.047 | 0.122 |

**Notes:** The table reports the result of OLS regressions of measures of forecast accuracy on expertise measures. The dependent variable is the rank-order correlation between forecast and actual effort across the treatments, and each observation is a forecaster i. Columns (3) and (4) use as control variables the decile of Google Scholar citations for the researcher, main field of expertise, and an indicator for whether the researcher has used Murk. Columns (2) and (4) include controls time to survey completion, whether the forecaster clicked on practice or the instructions, and how many days the forecaster delayed starting the survey. Standard errors are clustered by individual.

* significant at 10%; ** significant at 5%; *** significant at 1%

## Online Appendix Table 4. Impact of Effort and Motivation on Forecast Accuracy

| Dep. Var. (Measure of Accuracy): | (Negative of) Absolute Forecast Error in Treatment *t* by Forecaster *i* | | | | Rank-Order Correlation between Forecasts and Effort by Forecaster *i* | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| ***Time to Completion (Omitted 5-9 minutes)*** | | | | | | | | |
| **Survey Completion Time 0-4 Minutes** | . | -112.22** (52.13) | -61.26*** (20.83) | | . | -0.374*** (0.132) | -0.308*** (0.046) | |
| **Survey Completion Time 10-14 Minutes** | -11.14 (11.18) | 6.04 (12.79) | 33.80*** (12.21) | | -0.152** (0.071) | -0.008 (0.047) | 0.026 (0.028) | |
| **Survey Completion Time 15-24 Minutes** | -10.27 (12.06) | 22.13* (12.02) | 42.82*** (14.48) | | -0.196*** (0.066) | -0.035 (0.044) | 0.001 (0.037) | |
| **Survey Completion Time 25+ Minutes** | -23.63** (11.52) | 21.20 (12.88) | -22.05 (33.67) | | -0.292*** (0.071) | 0.070 (0.047) | -0.115 (0.100) | |
| ***Measures of Attention to Instructions*** | | | | | | | | |
| **Clicked on Practice Task** | -3.14 (8.43) | -3.36 (9.96) | . | | -0.068 (0.052) | 0.031 (0.039) | . | |
| **Clicked on Full Instructions** | 1.13 (10.41) | -29.64* (16.74) | . | | 0.104* (0.058) | -0.134** (0.061) | . | |
| ***Delay in Survey Completion*** | | | | | | | | |
| **Days Waited to Take Survey (Since Invitation)** | -0.08 (0.25) | -0.03 (0.87) | . | | 0.000 (0.001) | 0.000 (0.002) | . | |
| ***Mturk Incentives and Experience*** | | | | | | | | |
| **Higher Incentives (up to $5) for Forecast Accuracy** | | | | -6.27 (13.24) | | | | 0.029 (0.030) |
| **Experienced the Task** | | | | -23.86** (11.98) | | | | -0.026 (0.032) |
| **Controls for Expertise:** | X | | | | X | | | |
| **Control for Missing Click:** | | X | | | | X | | |
| **Fixed Effects:** | Fixed Effects for Treatment 1-15 and for Order 1-15 of Treatments | | | | | | | |
| **Sample Indicators Interacted with Fixed Effects:** | | X | | | | | | |
| **Indicators for Samples:** | | | | | | X | | |
| **Sample:** | Academic Experts | PhDs, Undergr., MBAs | Mturk Workers | | Academic Experts | PhDs, Undergr., MBAs | Mturk Workers | |
| ***N*** | 3120 | 6975 | 11430 | 11430 | 208 | 463 | 762 | 762 |
| ***R Squared*** | 0.123 | 0.071 | 0.032 | 0.020 | 0.120 | 0.068 | 0.067 | 0.001 |

**Notes:** The table reports the result of OLS regressions of measures of forecast accuracy on measures of effort and motivation. In Columns (1)-(4) the dependent variable is the (negative of) the absolute forecast error and an observation in the regression is a forecaster-treatment combination, with each forecaster providing forecasts for 15 treatments. In Columns (5)-(8), the dependent variable is the rank-order correlation between forecast and actual effort across the 15 treatments, and each observation is a forecaster i. The specification in Columns (1) and (5) include controls for rank and for field of expertise of the academic expert. The time of survey completion is measured between the logged opening time and the logged submission time. Each forecaster has the option to click and open a practice task and/or to click or open the PDF with full instructions. Indicators for either are measures of forecaster effort. A further measure of motivation is the delay in days between when the forecasters were invited and when the survey was completed. In Columns (4) and (8) we compare MTurk workers with baseline incentives for forecast accuracy and with heightened incentives and those who have experienced the task. Columns (1)-(4) include fixed effects for the order in which the expert encountered a treatment (to control for fatigue) and fixed effects for the treatment. Standard errors are clustered by individual.

* significant at 10%; ** significant at 5%; *** significant at 1%

s

## Online Appendix Table 5. Impact of Confidence on Forecast Accuracy

| Dep. Var. (Measure of Accuracy): | (Negative of) Absolute Forecast Error in Treatment *t* by Forecaster *i* | | | Forecast Within 100 Points of Actual Effort in Treatment *t* for Forecaster *i* | | | Rank-Order Correlation between Forecasts and Effort by Forecaster *i* | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| ***Measures of Confidence*** | | | | | | | | | |
| **Number of Own Forecasts Expected To Be Within 100 Points of Actual (Out of 15)** | 1.57 | 5.03*** | 8.78*** | 0.001 | 0.007** | 0.009*** | -0.007 | 0.018*** | -0.002 |
| | (1.39) | (1.35) | (1.77) | (0.004) | (0.003) | (0.002) | (0.009) | (0.005) | (0.004) |
| **Fixed Effects:** | Fixed Effects for Treatment 1-15 and for Order 1-15 of Treatments | | | | | | | | |
| **Sample Indictators Interacted with Fixed Effects:** | | X | | | X | | | | |
| **Indicators for Sample:** | | | | | | | | X | |
| **Indicator for Missing Confidence Variable:** | X | X | X | X | X | X | X | X | X |
| **Controls for Time to Completion:** | X | X | X | X | X | X | X | X | X |
| **Controls for Expertise:** | X | | | X | | | X | | |
| **Sample:** | Academic Experts | PhDs, Undergr., MBAs | Mturk Workers | Academic Experts | PhDs, Undergr., MBAs | Mturk Workers | Academic Experts | PhDs, Undergr., MBAs | Mturk Workers |
| ***N*** | 3120 | 6975 | 11430 | 3120 | 6975 | 11430 | 208 | 465 | 762 |
| ***R Squared*** | 0.124 | 0.078 | 0.045 | 0.173 | 0.107 | 0.042 | 0.129 | 0.088 | 0.068 |

**Notes:** The table reports the result of OLS regressions of measures of forecast accuracy on measures of confidence. In Columns (1)-(3) the dependent variable is the (negative of) the absolute forecast error and in Columns (4)-(6) the dependent variable is an indicator for whether the forecast falls within 100 points of the actual average effort in the treatment. In these columns, an observation in the regression is a forecaster-treatment combination, with each forecaster providing forecasts for 15 treatments. In Columns (7)-(9), the dependent variable is the rank-order correlation between forecast and actual effort across the 15 treatments, and each observation is a forecaster i. The measure of confidence is the forecast by the participant of the number of treatments that he/she expects to get within 100 points of the actual one. This variable varies from 0 (no confidence) to 15 (confidence in perfect forecast). All columns include the controls for time of completion used in Table 6, as well as an indicator for the few observations in which the confidence variable is missing (in which case the confidence variable itself is seto to zero). The specifications in Columns (1), (4), and (7) also includes controls for rank and for field of expertise of the academic experts. Columns (1) to (6) include fixed effects for the order in which the expert encountered a treatment (to control for fatigue) and fixed effects for the treatment. Standard errors are clustered by individual.

* significant at 10%; ** significant at 5%; *** significant at 1%

## Online Appendix Table 6. Impact of Revealed Accuracy by Groups of Treatments

| Dep. Var. (Measure of Accuracy): | (Negative of) Absolute Forecast Error in Treatment *t* by Forecaster *i* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Group of Treatments Omitted:** | **4-cent Piece Rate** | **Pay Enough** | **Charity** | **Gift Exchange** | **Discounting** | **Gains vs. Losses** | **Prob. Weighting** | **Psychology Treatments** |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| ***Panel A. Forecasts by Academic Experts*** | | | | | | | | |
| **(Negative of) Absolute Error in Forecast in Relevant Treatments / 100** | 9.57** | 7.55*** | 18.66*** | 3.84 | 8.51** | -3.91 | 17.95*** | 9.84*** |
| | (3.73) | (2.09) | (3.54) | (3.06) | (3.68) | (2.90) | (4.50) | (3.60) |
| **Fixed Effects:** | Fixed Effects for Treatment 1-15 and for Order 1-15 of Treatments | | | | | | | |
| **Controls for Expertise, Confidence and Time to Completion:** | X | X | X | X | X | X | X | X |
| **Sample:** | Academic Experts | | | | | | | |
| ***N*** | 2912 | 2912 | 2704 | 2912 | 2704 | 2496 | 2704 | 2496 |
| ***R Squared*** | 0.115 | 0.102 | 0.149 | 0.137 | 0.112 | 0.150 | 0.137 | 0.153 |
| ***Panel B. Forecasts by PhDs, Undergrads, MBAs, Mturks*** | | | | | | | | |
| **(Negative of) Absolute Error in Forecast in Relevant Treatments / 100** | 29.81*** | 28.32*** | 39.13*** | 17.20*** | 34.19*** | 28.83*** | 39.90*** | 43.33*** |
| | (1.66) | (1.77) | (1.98) | (2.04) | (1.76) | (2.50) | (2.21) | (2.81) |
| **Fixed Effects:** | Fixed Effects for Treatment 1-15 and for Order 1-15 of Treatments, interacted with the Sample indicators | | | | | | | |
| **Controls for Confidence and Time to Completion:** | X | X | X | X | X | X | X | X |
| **Sample:** | PhDs, Undergraduates, MBAs, Mturkers | | | | | | | |
| ***N*** | 17178 | 17178 | 15951 | 17178 | 15951 | 14724 | 15951 | 14724 |
| ***R Squared*** | 0.181 | 0.171 | 0.195 | 0.114 | 0.200 | 0.144 | 0.197 | 0.181 |

**Notes:** The table reports the result of OLS regressions of forecast accuracy on measures of revealed forecasting accuracy in other treatments. Each column reports the regression of forecaster accuracy as a function of accuracy in the identified treatments (leaving those treatments outside the sample). Thus, for example, in Column (2) we examine whether accuracy in forecasting the pay-enough-or-don't-pay-at-all treatment increases accuracy in forecast for the other treatments. Panel A reports the results for the sample of academic experts, while Panel B reports the results for the sample of PhD students, undergraduates, MBAs, and MTurkers. The regressions include the same controls for confidence and time to completion as in Table 8. The specification in Panel A also includes controls for rank and for field of expertise of the academic experts. All columns include fixed effects for the order in which the expert encountered a treatment (to control for fatigue) and fixed effects for the treatment. Standard errors are clustered by individual.

* significant at 10%; ** significant at 5%; *** significant at 1%