Predicting Experimental Results: Who Knows What?
Online Appendix

Stefano DellaVigna
UC Berkeley and NBER

Devin Pope
U Chicago and NBER

This version: May 25, 2017

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.

B Online Appendix B - Model Estimation

As mentioned in the main text, we start with the model

$$s_{i,k} - \theta_k = \eta_k^j + v_i + \sigma_i \epsilon_{i,k}$$

where here we write η_k^j instead of η_k to make explicit the fact that the fixed effects for the 15 treatments are estimated separately for each of the 5 subject groups ($j \in \{Experts,\ PhDs,\ MBAs,\ Undergraduates,\ MTurks\}$).[1]

[1] We examine the role played by allowing separate fixed effects for each subject group by also estimating several model specifications where these fixed effects are restricted to be the same for all subject groups. Columns 1 and 2 of Online Appendix Table 3 and Online Appendix Figure 14 are based on such specifications, and differences

After estimating η_k^j, we define $z_{i,k} \equiv s_{i,k} - \theta_k - \eta_k^j$ and rewrite the model as

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

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

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

The log-likelihood for this model is

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

where $\theta \equiv [v^{(1)}, v^{(2)}, \sigma^{(1)}, \sigma^{(2)}, \beta]^T$ and ϕ denotes the standard normal density. The asymptotic covariance is given by the inverse of the Fisher information, which we estimate with its sample analogue.[2]

We implement the maximum likelihood estimation of this model in R, using the package "bbmle" (which contains methods and functions for fitting ML models). To ensure the global convergence to the MLE parameters, we estimate each model about 1,000 times, taking random starting values of the estimands from uniform distributions with reasonably wide support[3], and take the estimate with the highest log likelihood.

Simulations from Estimated Model. At various points in the main text, we generate simulated data sets for the estimated model parameters to gauge how well our model fits key patterns in the data (e.g., Figure 6b). For each such exercise, we typically generate 100 simulated data sets to increase the reliability of our conclusions.

The simulation procedure is as follows. We take the distribution of forecaster characteristics x_i as given (using the empirical distribution from the data). The estimated model parameters thus give us the probability of each type, and also the bias and idiosyncratic forecast s.d.

between these results and those of our benchmark specification are discussed in the main text.

Also, in some specifications where we use forecasts on the 4-cent treatment as a covariate in the model, we drop observations corresponding to the 4-cent forecasts and estimate η_k^j on the remaining observations.

[2] In cases where we allow for more than 2 types (e.g. Column 5 in Panel A of Online Appendix Table 4), we specify the probability of types to be a multinomial logit, with a separate β for each type (other than the omitted type). In general, in a model with L types, the log-likelihood can be written as:

$$Loglik[z|\theta] = \Sigma_{i=1}^{I} \Sigma_{k=1}^{K} log\{\Sigma_{l=1}^{L}(\frac{e^{x_i^T \beta^{(l)}}}{\Sigma_{l'=1}^{L} e^{x_i^T \beta^{(l')}}}) \cdot [\frac{1}{\sigma^{(l)}} \phi(\frac{z_{i,k} - v^{(l)}}{\sigma^{(l)}})]\}$$

where we set $\beta^{(l)} \equiv 0$ for the omitted type.

[3] Note that the estimated value of the parameter need not fall within the support of random variable used to determine the starting value.

of each type. From this, we simulate the values of $\widetilde{z}_{i,k}$, and using the estimated values of $\hat{\eta}^j_k$ as well as the actual efforts in each treatment θ_k, thus recover the "simulated forecast" $\widetilde{s}_{i,k} = z_{i,k} + \theta_k + \hat{\eta}^j_k$ for each forecaster i in each treatment k. We then use this simulated dataset to recover the required variables and moments (e.g. mean absolute error, rank-order correlation, wisdom-of-crowd forecasts etc.) to compare with those from the actual data.

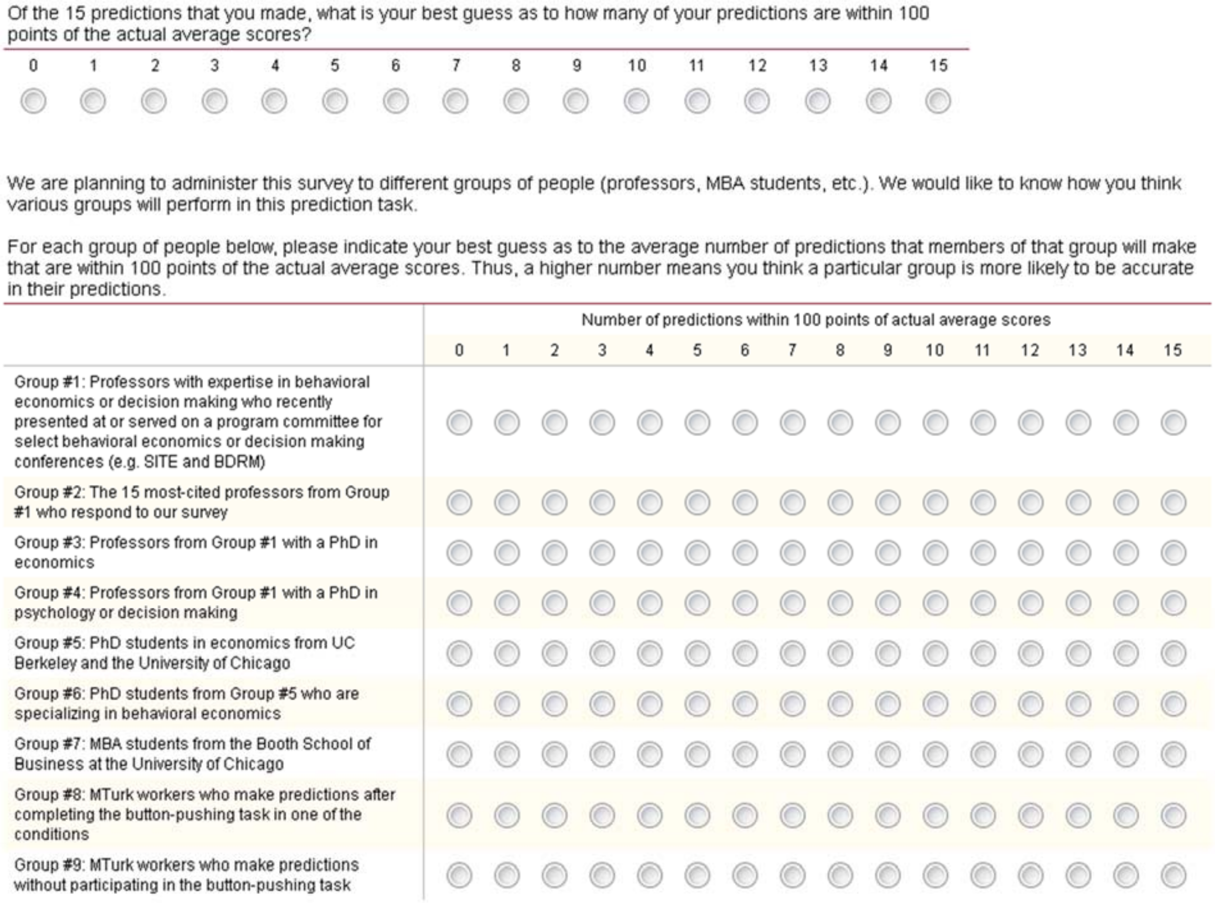
We observe that there are two sources for differences between simulated datasets for a given forecaster i. First, the simulations differ because of different draws of ϵ^k_i. Second, they also differ because, while the probability of being the "good" type for a forecaster i $p^1_i(x_i)$ is constant across simulations (due to the fact that we take x_i as given), the realized type differs across simulations.

For the simulations corresponding to superforecasters in Figures 9c-d, we define a superforecaster in a certain group as the 20% of forecasters most likely to be the "good" type in their group according to the model estimates in Column 3 of Table 3.[4]Note that since the probability of being the "good" type is $\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$ and we are keeping x_i and β fixed across simulations (taking the former directly from the data, and the latter from Column 3 of Table 3), the identities of the superforecasters across different simulations will be the same. However, their forecasts across simulations will not be the same because their realized types and $\epsilon_{i,k}$'s will differ due to randomness.

Robustness. Figure 6b and Online Appendix Figure 10b on time to completion and Figures 7c-d and Online Appendix Figure 12b on confidence are based on the same model in Column 2 of Table 3, including type indicators, the time to completion indicators, and confidence. An alternative procedure would be to produce each figure based on model estimates which include just the group indicators and the relevant variable, such as for example the time to completion for Figure 6b and Online Appendix Figures 10b. The resulting simulation plots are almost identical to the current plots in the paper.

[4] For this exercise, we define 3 groups: experts; students, i.e. the pooled sample of PhDs, MBAs and undergraduates; and MTurks.

Online 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
○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

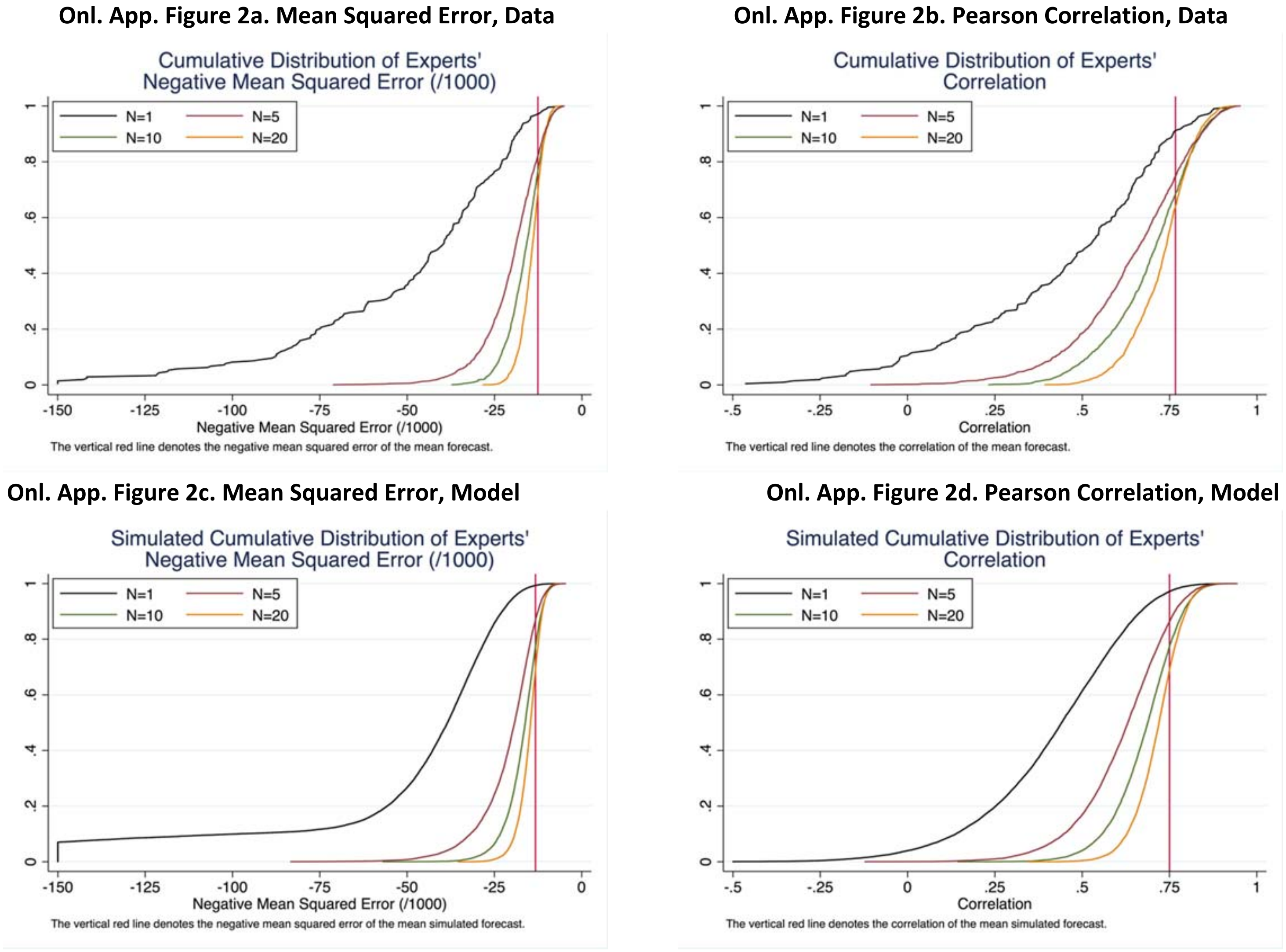
We are planning to administer this survey to different groups of people (professors, MBA students, etc.). We would like to know how you think various groups will perform in this prediction task.

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

Online Appendix Figure 2. Individual Expert Accuracy versus Aggregate (Wisdom-of-Crowds) Accuracy, Additional Accuracy Measures



Onl. App. Figure 2a. Mean Squared Error, Data

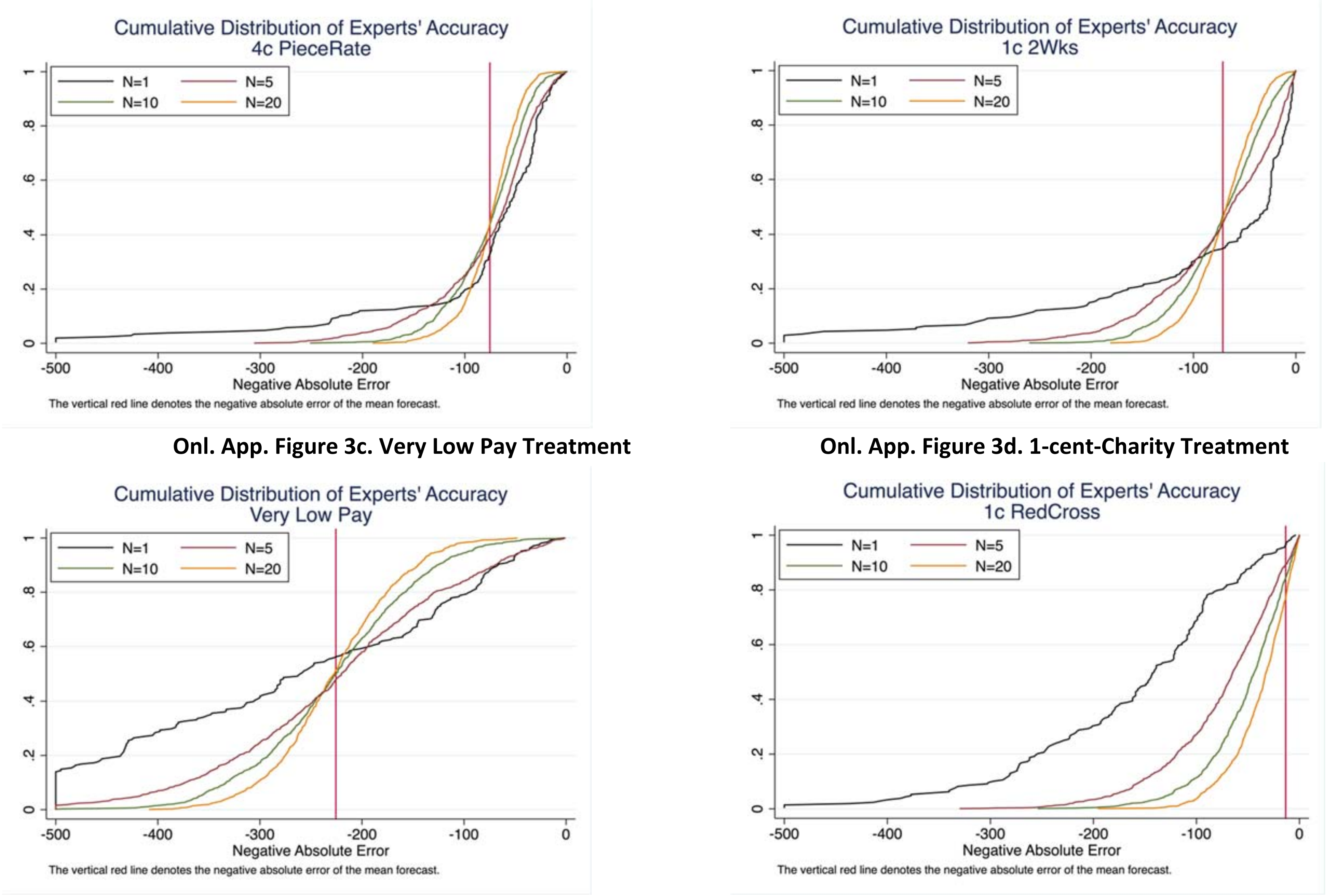
Onl. App. Figure 2b. Pearson Correlation, Data

Onl. App. Figure 2c. Mean Squared Error, Model

Onl. App. Figure 2d. Pearson Correlation, Model

Notes: Online Appendix Figure 2 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 3. Individual Expert Accuracy versus Aggregate (Wisdom-of-Crowds) Accuracy, Representative Treatments



Onl. App. Figure 3a. 4-cent Piece Rate Treatment

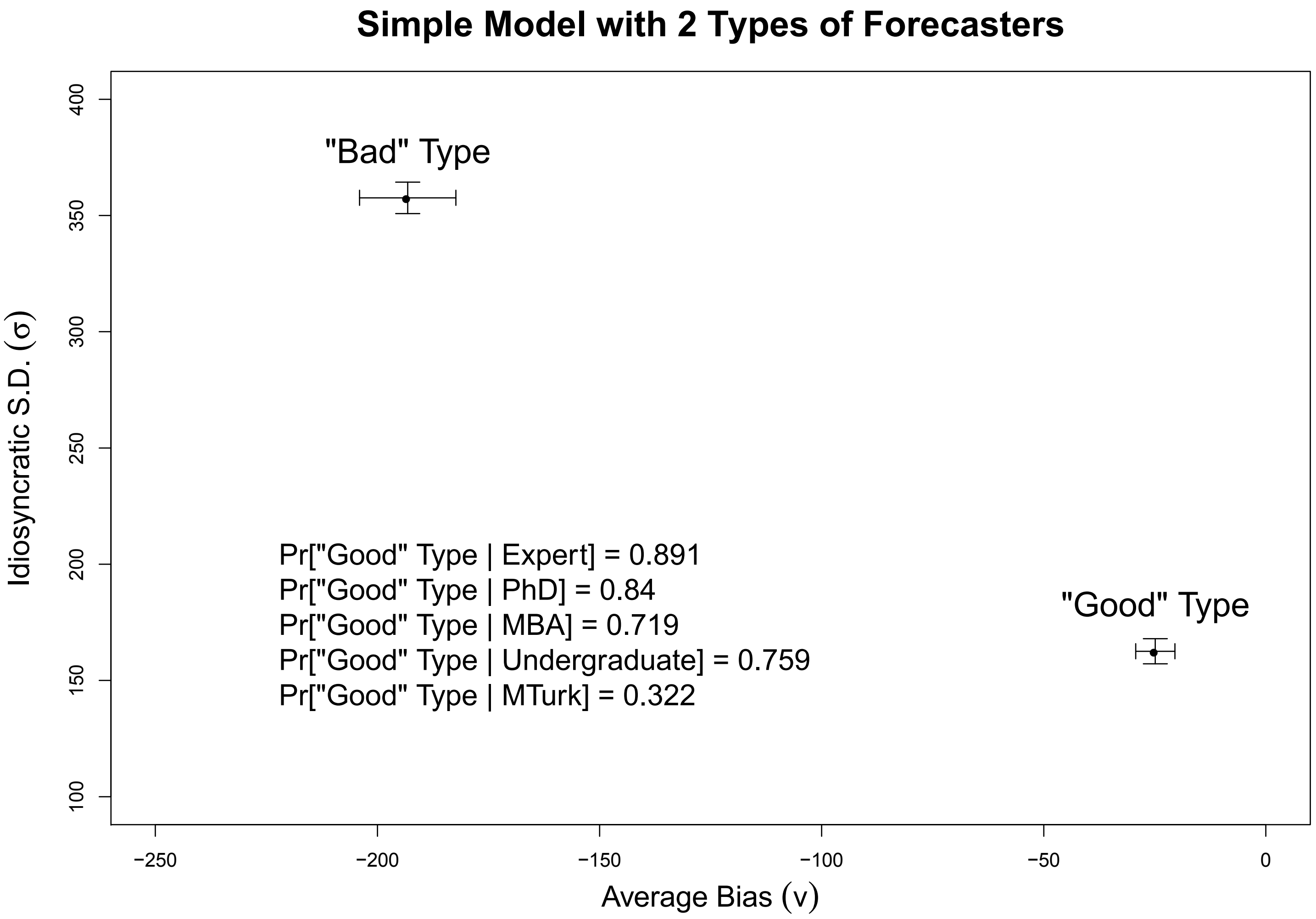
Onl. App. Figure 3b. 1-cent-in-2-weeks Treatment

Onl. App. Figure 3c. Very Low Pay Treatment

Onl. App. Figure 3d. 1-cent-Charity Treatment

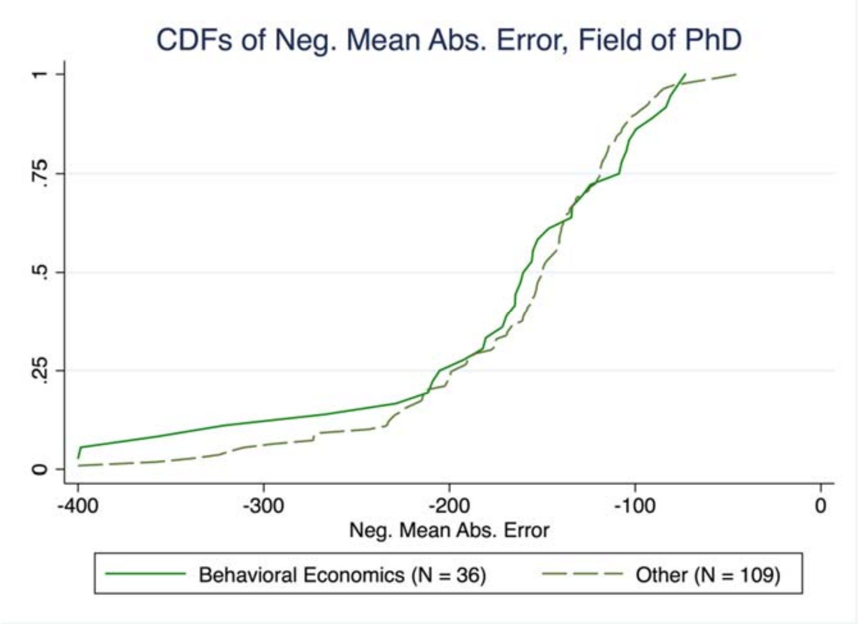
Notes: The figure presents the same information as in Figure 3a for four treatments using the negative of the absolute mean error as accuracy measure. Graphs are censored at -500.

Online Appendix Figure 4. Maximum-Likelihood Estimates of Model of Expertise.



Notes: Online Appendix 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.

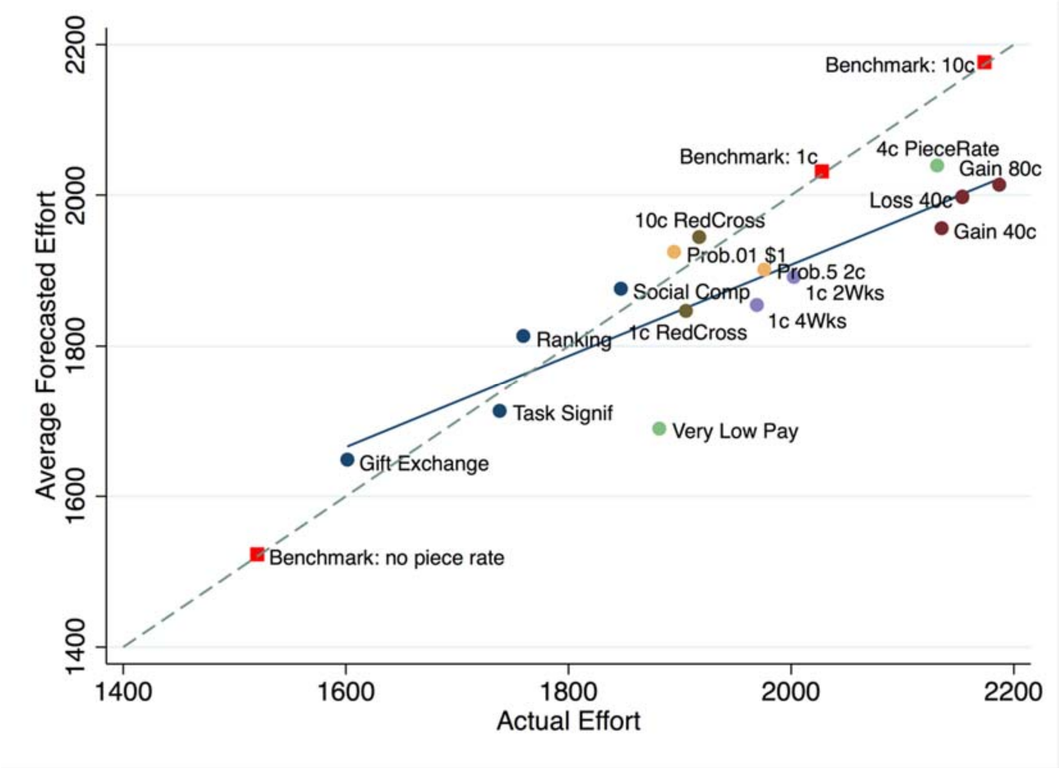
Online Appendix Figure 5. Horizontal Expertise for PhD Students



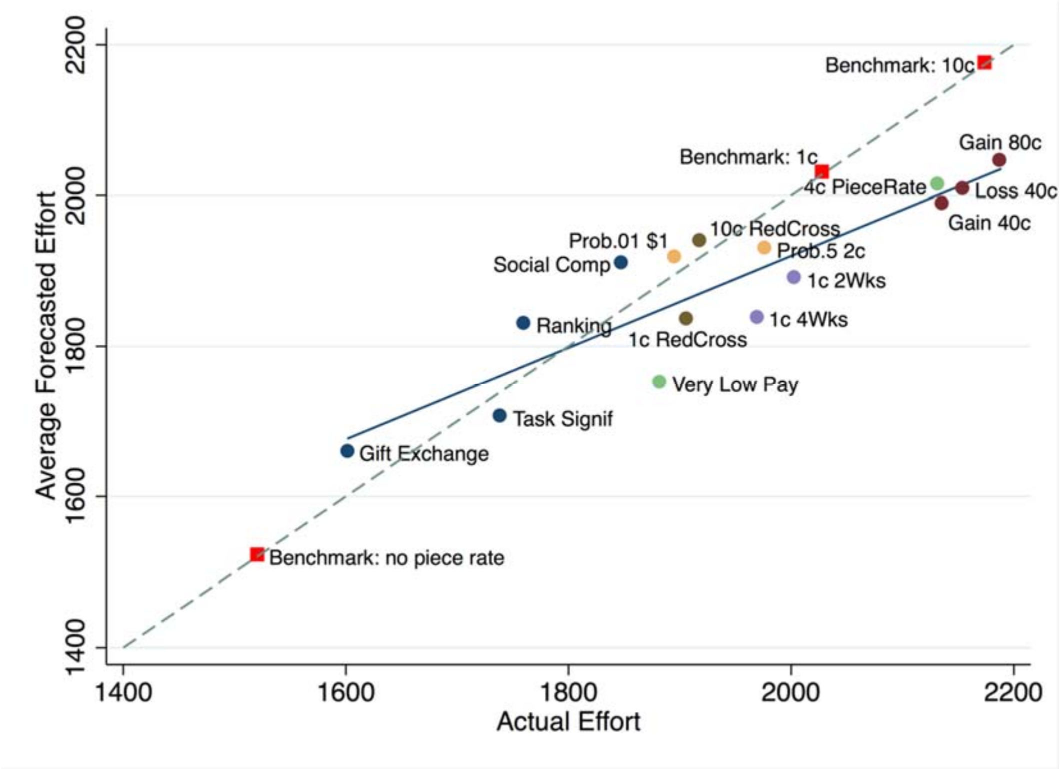
Notes: Online Appendix Figure 5 presents the c.d.f. for the negative of the mean absolute error for the PhD students participating depending on whether the (self-reported) field of specialization is Behavioral Economics or other.

Online Appendix Figures 6a-d. Wisdom-of-Crowds Accuracy, Other Groups

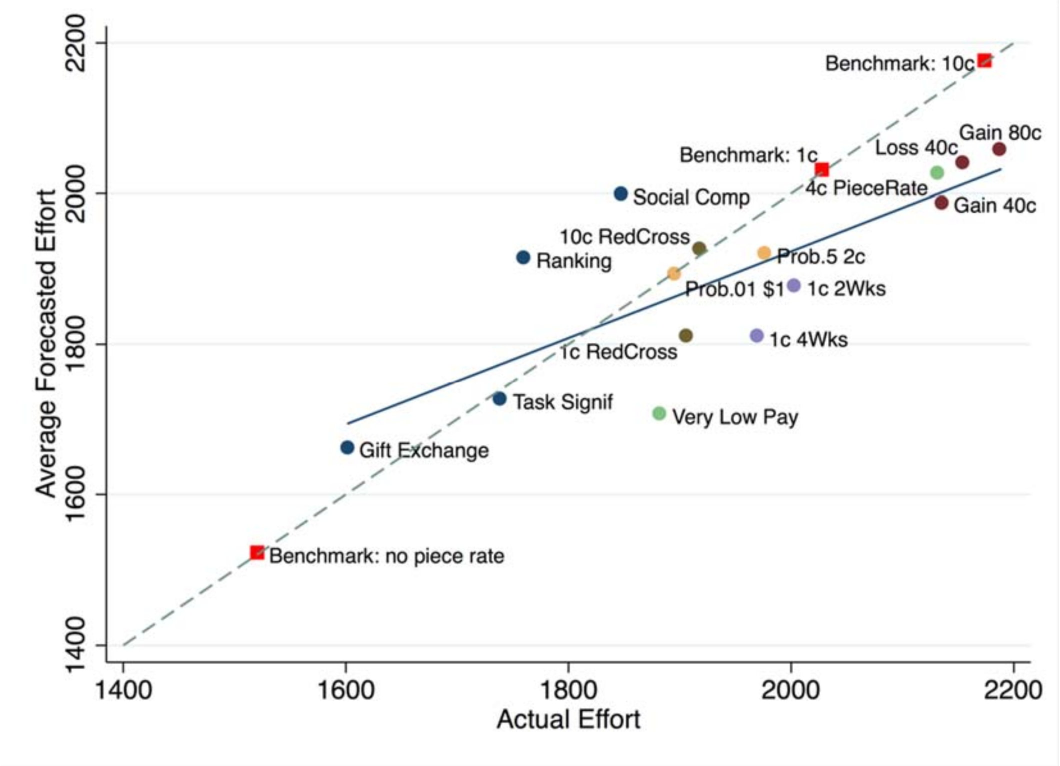
Appendix Figure 2a. PhD Students



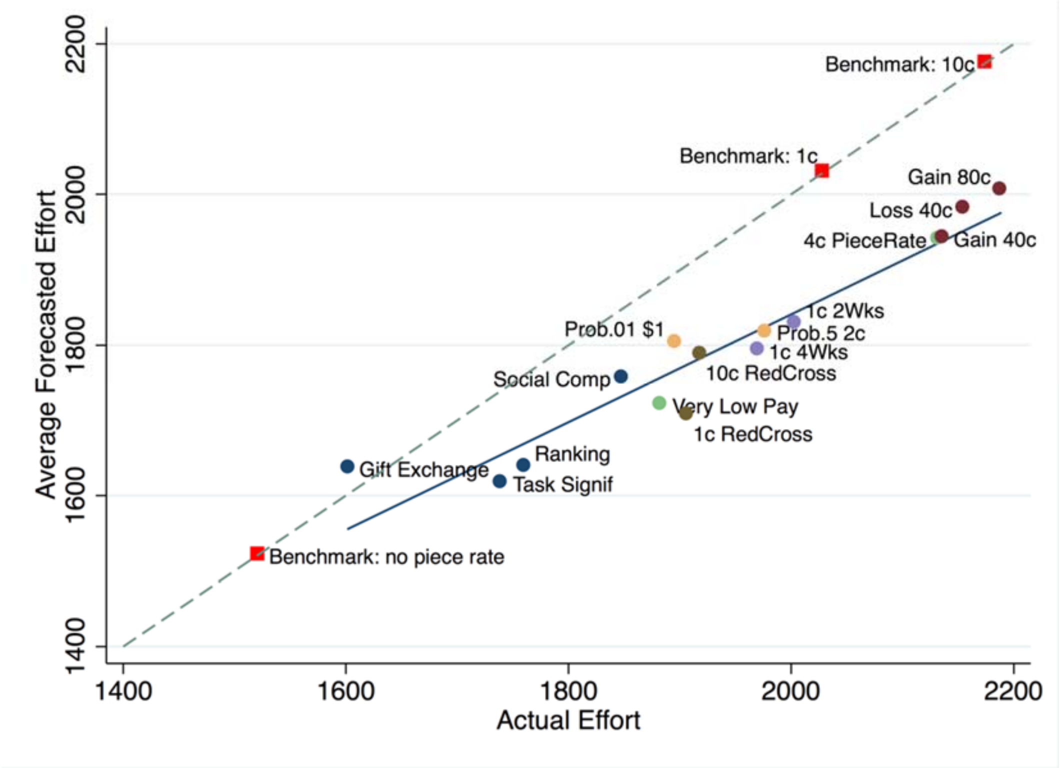
Online Appendix Figure 6b. Undergraduate Students



Online Appendix Figure 6c. MBA Students

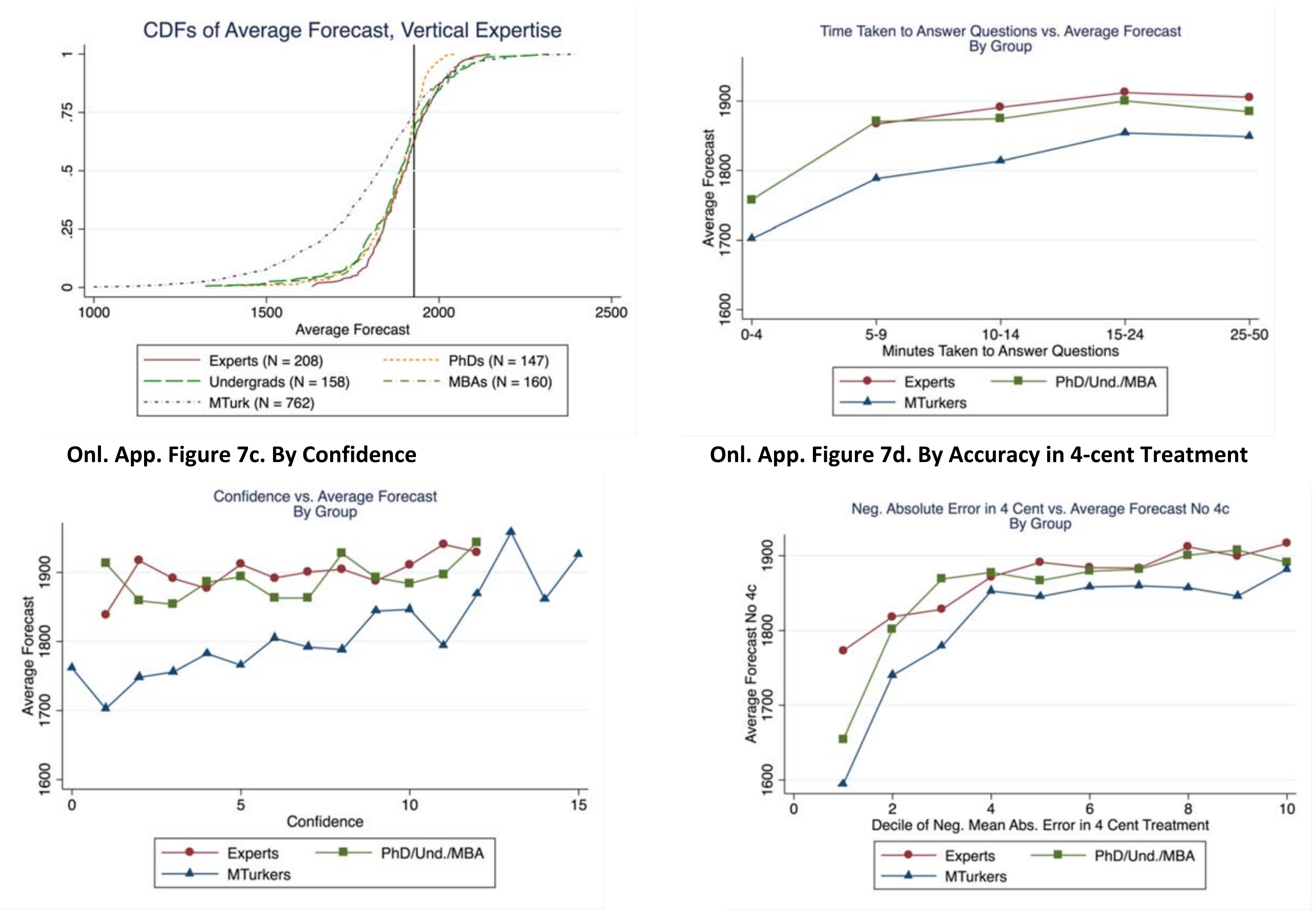


Online Appendix Figure 6d. MTurk



Notes: These figures present the parallel evidence to Figure 2 for the other samples of forecasters

Online Appendix Figures 7a-d. Average Forecast Across All 15 Treatments, Key Findings



Onl. App. Figure 7a. Distribution of Average Forecast

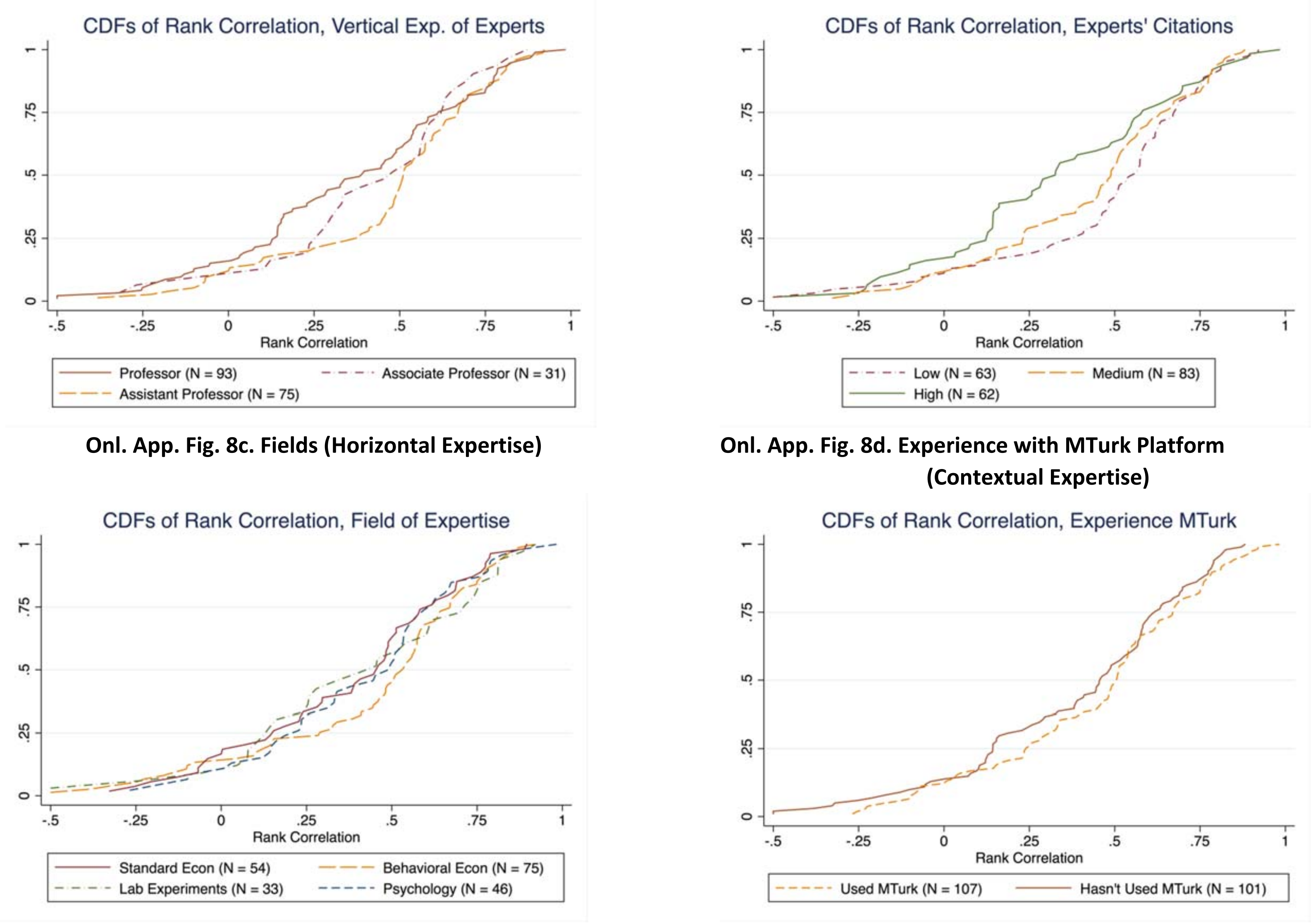
Onl. App. Figure 7b. By Time to Completion

Onl. App. Figure 7c. By Confidence

Onl. App. Figure 7d. By Accuracy in 4-cent Treatment

Notes: These figures present evidence on the average forecast across the 15 treatments. Online Appendix Figure 7a 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 7b-d show that the average forecast increases in the time taken to do the survey (Figure 7b), in the confidence (Figure 7c), and in the accuracy of forecast of the 4c treatment (Figure 7d).

Online Appendix Figures 8a-d. Key Findings on Vertical, Horizontal, and Contextual Expertise, Rank-Order Correlation



Onl. App. Fig. 8a. Academic Rank (Vertical Expertise)

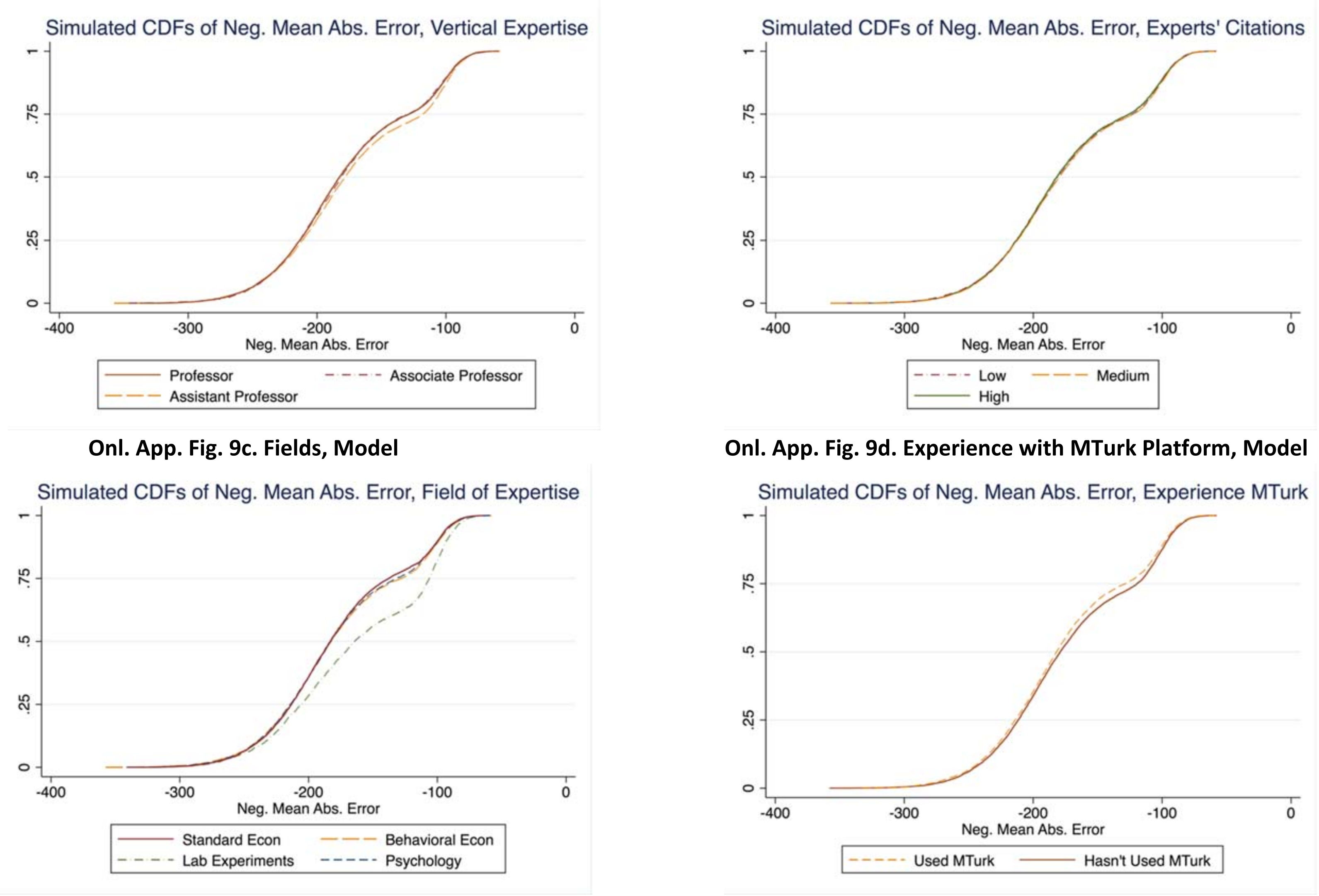
Onl. App. Fig. 8b. Citations (Vertical Expertise)

Onl. App. Fig. 8c. Fields (Horizontal Expertise)

Onl. App. Fig. 8d. 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 Figures 9a-d. Key Findings on Vertical, Horizontal, and Contextual Expertise, Model Results


Onl. App. Fig. 9a. Academic Rank, Model

Onl. App. Fig. 9b. Citations, Model

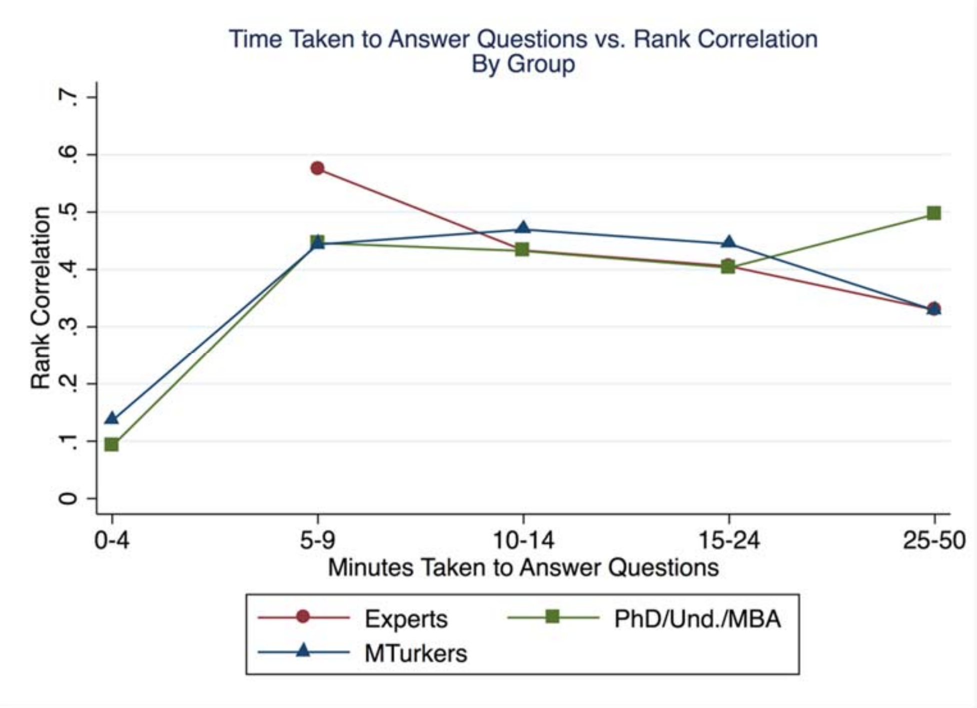
Onl. App. Fig. 9c. Fields, Model

Onl. App. Fig. 9d. Experience with MTurk Platform, Model

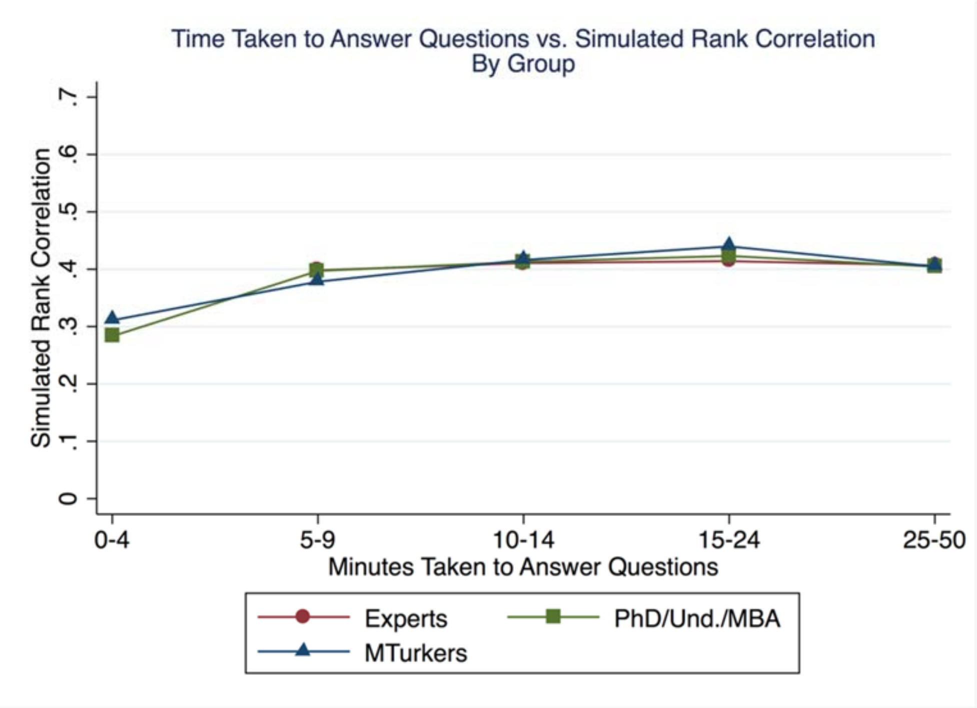
Notes: These figures present the key results on vertical, horizontal, and contextual expertise using simulated data from the model estimates for the academic experts (Column 4, Online Appendix Table 3).

Online Appendix Figure 10. Accuracy and Effort in Taking Task, Rank-Order Correlation

Onl. App. Figure 10a. Time Taken in Completing the Survey, Data

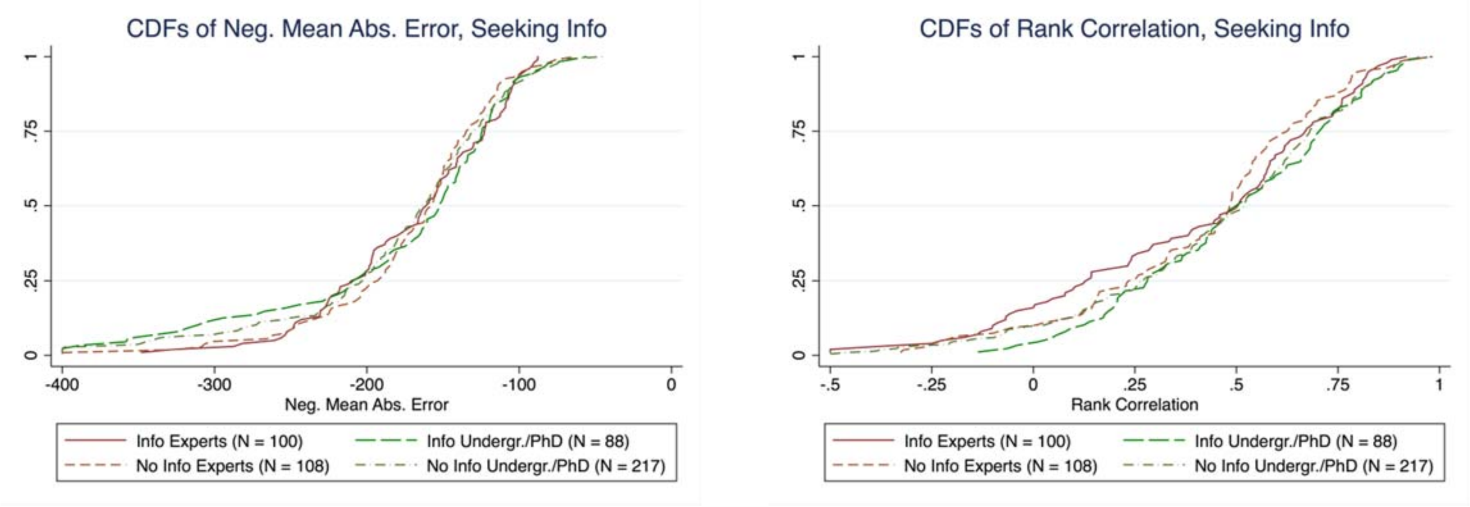


Onl. App. Figure 10b. Time Taken in Completing the Survey, Model

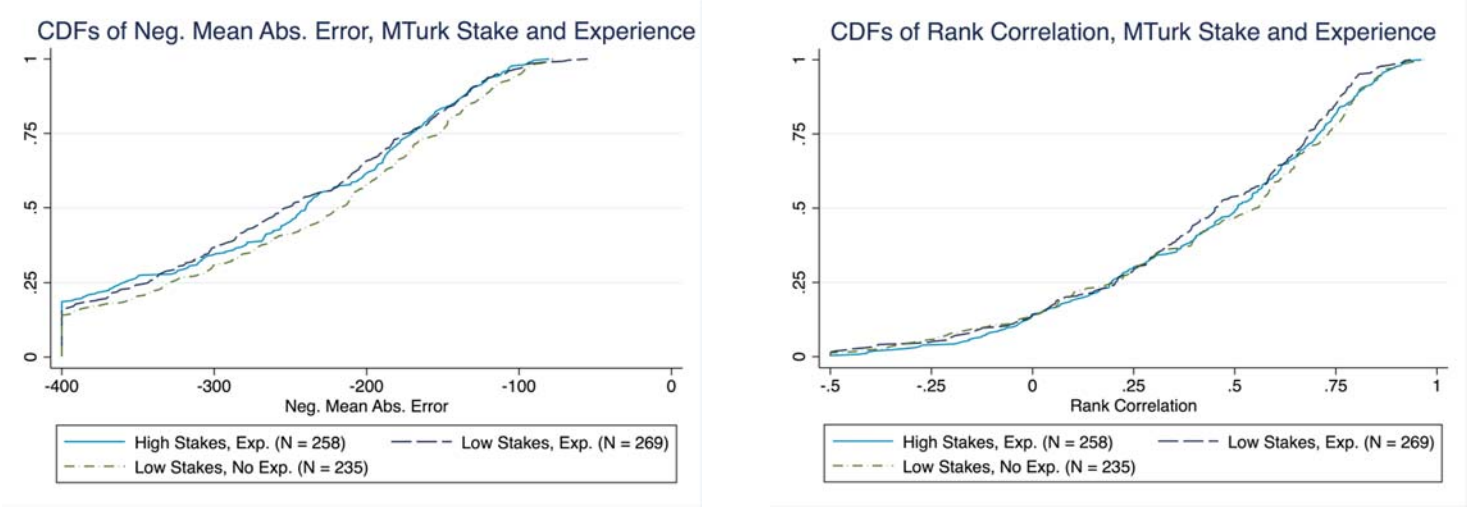


Notes: Online Appendix Figure 10a plots the accuracy for three groups of forecasters (academic experts; undergraduate, MBA, and PhD students; and MTurkers) as a function of how long they took to complete the survey. Specifically, the figures plot the average accuracy by minutes of time taken for survey completion. Online Appendix Figure 10b presents the corresponding figure from simulations for the model estimates as in Column 2 of Table 3. Specifically, we simulate the estimated model 100 times, taking as given (from the data) the empirical distribution of forecasters' characteristics that are used to predict forecaster "type", and average over the 100 simulations.

Online Appendix Figures 11a-b. Expert Checked Task or Full Instructions



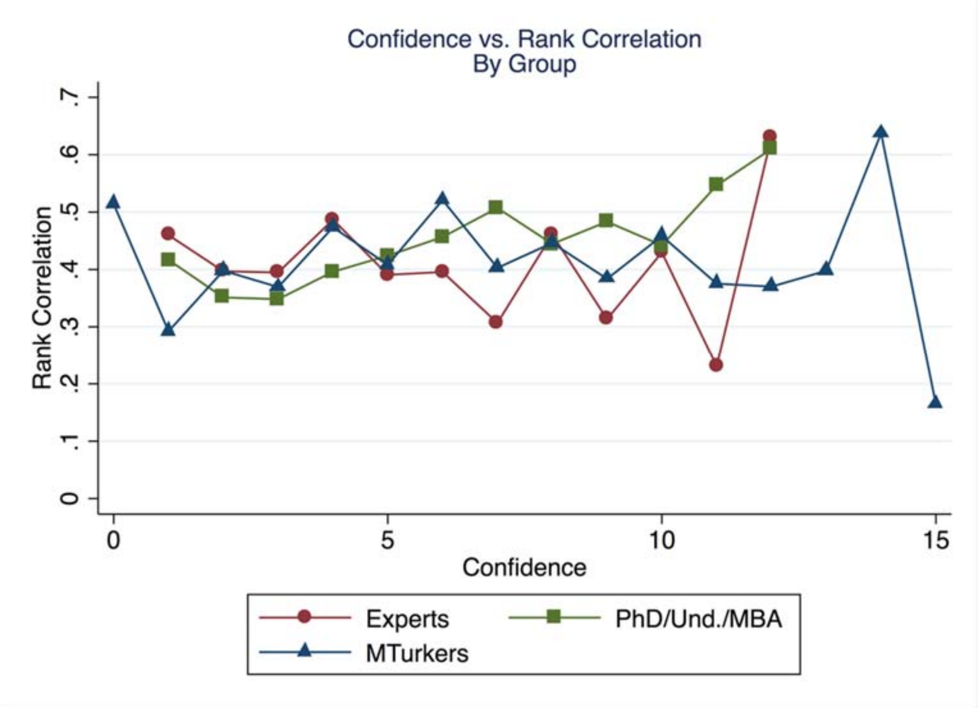
Online Appendix Figures 11c-d. Effect of Stake Size and Experience on Motivation, MTurk Sample



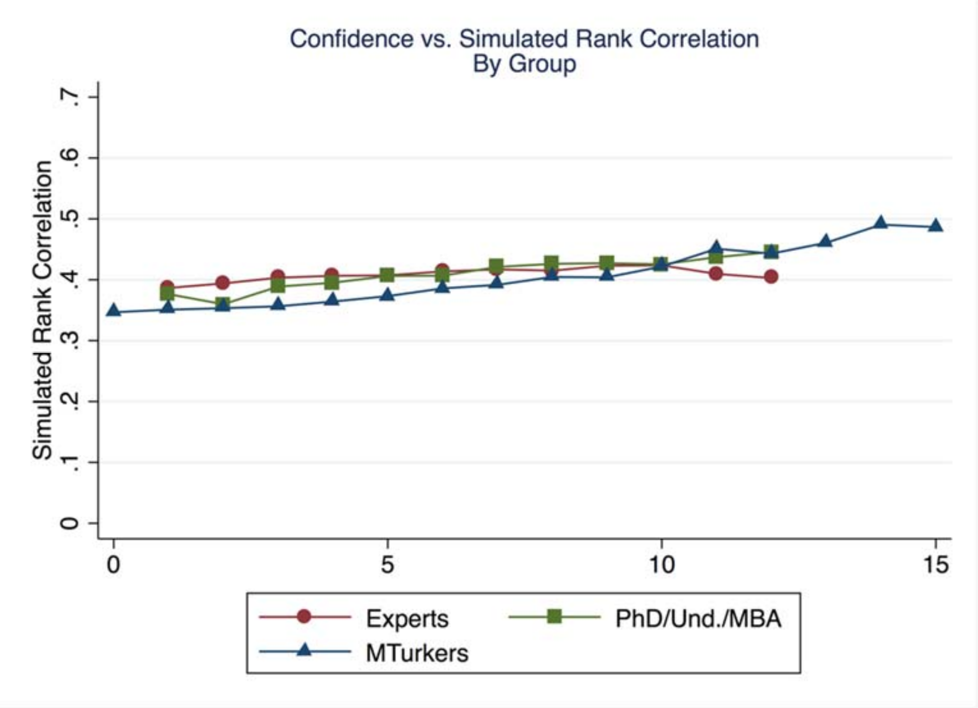
Notes: Online Appendix Figures 11a-b split two of the groups into whether they clicked on a link for a trial of the task or the link for additional instructions. (The MTurk group is excluded because no one in the group clicked on the link). Online Appendix Figures 11c-d compare three MTurk subgroups 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 12. Accuracy and Confidence in Taking Task, Rank-Order Correlation

Onl. App. Figure 12a. Confidence, Data



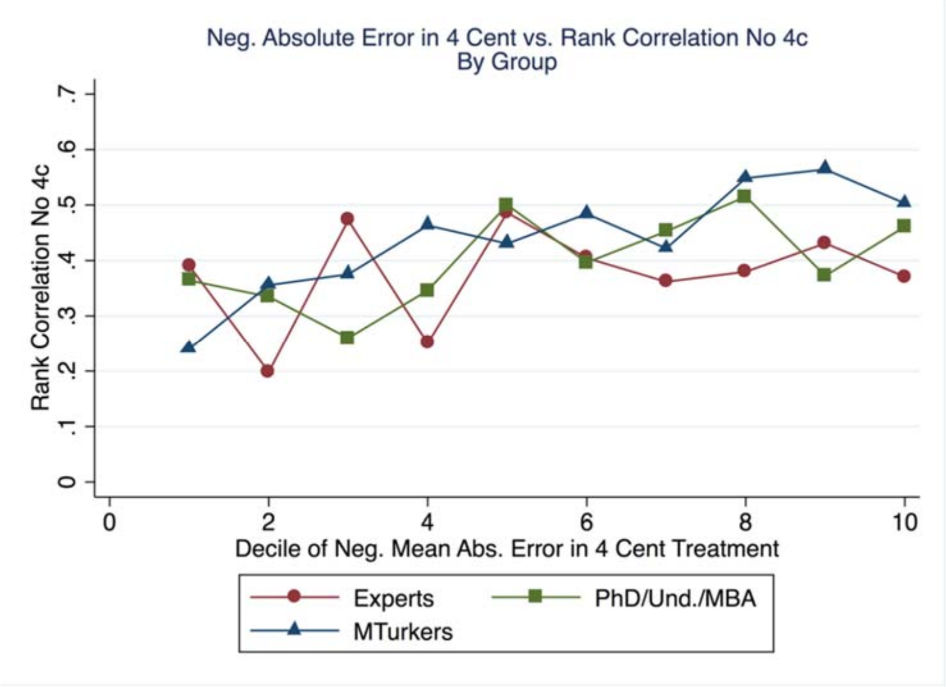
Onl. App. Figure 12b. Confidence, Model



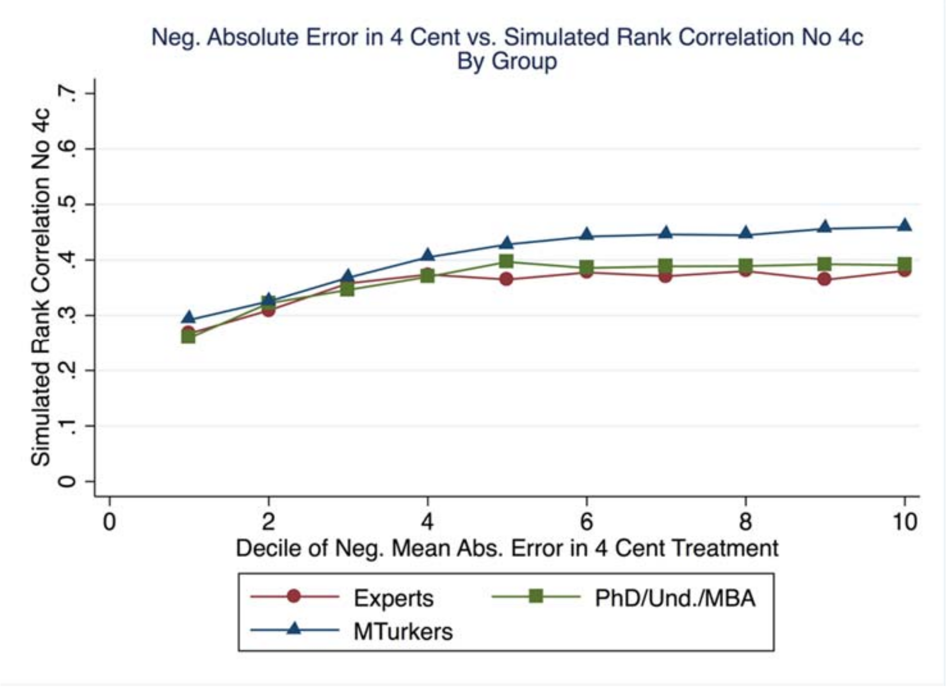
Notes: Online Appendix Figure 12a plots the average accuracy for three groups of forecasters (academic experts, undergraduate/MBA/PhD students, and MTurkers) by how confident the respondent felt about the accuracy. In particular, each survey respondent indicated how many out of 15 forecasts he or she made were going to be accurate up to 100 points relative to the truth. Online Appendix Figure 12b presents the corresponding figure from simulations for the model estimates as in Column 2 of Table 3. Specifically, we simulate the estimated model 100 times, taking as given (from the data) the empirical distribution of forecasters' characteristics that are used to predict forecaster "type", and average over the 100 simulations.

Online Appendix Figure 13. Accuracy and Revealed Accuracy, Rank-Order Correlation

Onl. App. Figure 13a. Accuracy in 4-cent Treatment, Data

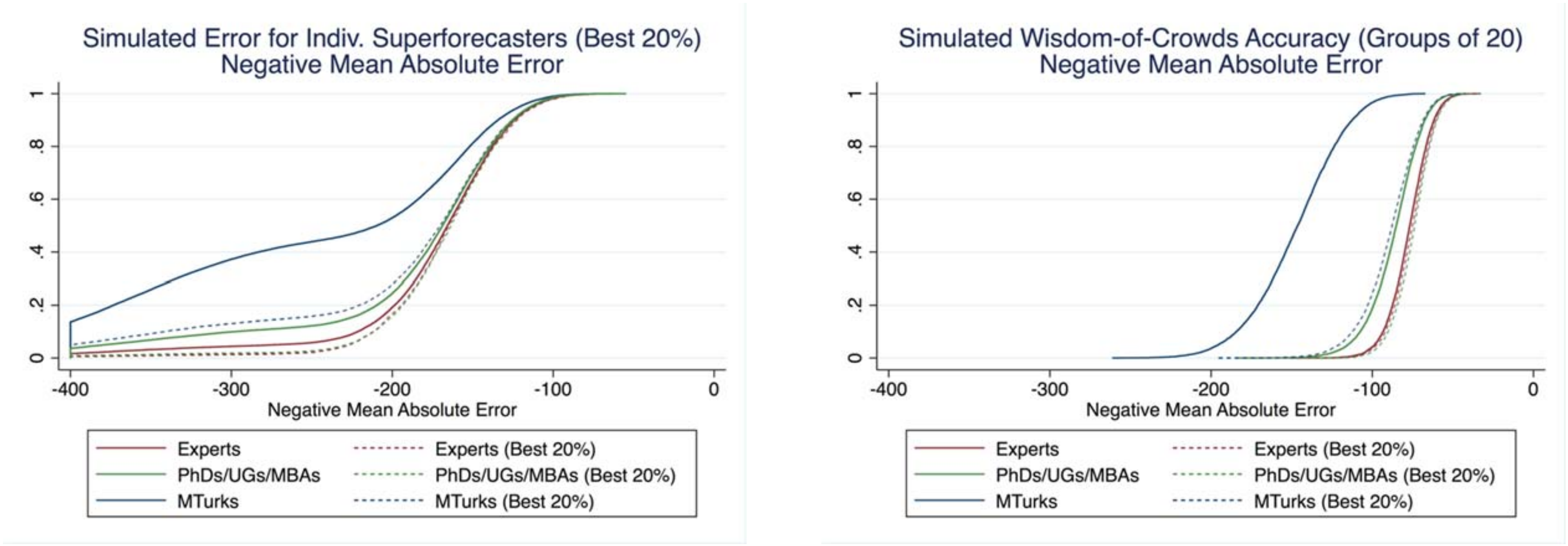


Onl. App. Figure 13b. Accuracy in 4-cent Treatment, Model



Notes: Online Appendix Figure 13a 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. Online Appendix Figure 13b presents the corresponding figure from simulations for the model estimates as in Column 3 of Table 3. Specifically, we simulate the estimated model 100 times, taking as given (from the data) the empirical distribution of forecasters' characteristics that are used to predict forecaster "type", and average over the 100 simulations.

Online Appendix Figure 14. Superforecasters: Selecting Non-Experts to Match Accuracy of Experts, Treatment Effects Same for All Groups

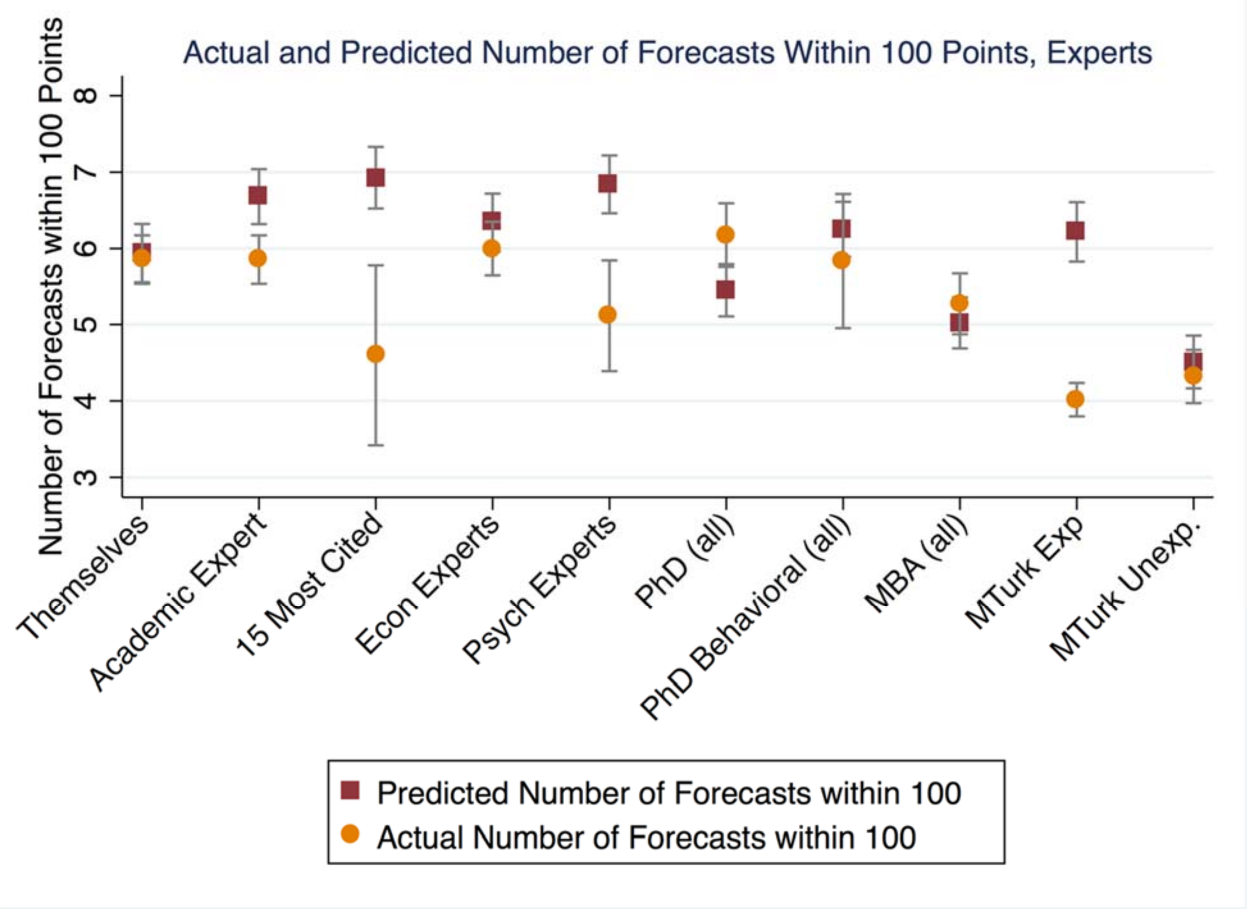


Onl. App. Figure 14a. Individual Accuracy, Model

Onl. App. Figure 14a. Wisdom-of-Crowds Accuracy (20 Forecasters), Model

Notes: Online Appendix Figures 14a-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") simulated from the model estimates as in Column 2 of Online Appendix Table 3, which forces treatment effects to be the same across all groups of forecasters. Specifically, we simulate the estimated model 100 times, taking as given (from the data) the empirical distribution of forecasters' characteristics that are used to predict forecaster "type". The superforecasters for the simulated datasets are defined as the top 20% of forecasters within each group in terms of probability of being the "good" type. Online Appendix Figure 14a plots the distribution of the individual-level accuracy, while Online Appendix Figure 14b plots the wisdom-of-crowds accuracy for groups of sample size 20.

Online Appendix Figure 15. Beliefs about Expertise, All PhDs and MBAs



Notes: Online Appendix Figure 15 compares the average accuracy of a group with the forecasted accuracy for that group by the 208 academic experts, as in Figure 10. Actual accuracy is calculated using all PhDs and MBAs surveyed.

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. Accuracy of Forecasts, Squared Error and Pearson Correlation

	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 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 B. Pearson 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 on mean squared error, and Panel B on the Pearson 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.

Online Appendix Table 3. Estimate of Model, Additional Specifications

Sample and Specification:	All Forecasters, Treatment Effects Equal for All Groups		Experts Only
	(1)	(2)	(3)
Estimated Parameters for the 2 Types			
$v^{(1)}$ (Average Bias, Type 1)	-25.12	-19.63	17.80
	(2.24)	(2.14)	(4.29)
$v^{(2)}$ (Average Bias, Type 2)	-200.54	-257.03	-60.54
	(6.10)	(7.25)	(4.96)
$\sigma^{(1)}$ (Idiosyncratic s.d., Type 1)	170.70	192.70	59.19
	(2.69)	(2.09)	(4.38)
$\sigma^{(2)}$ (Idiosyncratic s.d., Type 2)	361.54	365.68	216.15
	(3.47)	(3.74)	(3.82)
Predictors of Forecasters Being of Type 1, Logit Coefficients			
Constant	-0.65	0.52	-1.15
	(0.07)	(0.14)	(0.25)
Indicator for Expert	2.82	2.26	
	(0.15)	(0.24)	
Indicator for PhD	2.47	1.64	
	(0.14)	(0.21)	
Indicator for MBA	1.66	1.03	
	(0.11)	(0.17)	
Indicator for Undergraduate	1.96	2.27	
	(0.11)	(0.21)	
Response Time: 0-4 mins		-0.64	
		(0.21)	
Response Time: 10-14 mins		0.38	
		(0.11)	
Response Time: 15-24 mins		0.64	
		(0.13)	
Response Time: 25+ mins		0.49	
		(0.20)	
Predicted # Forecasts within 100 pts		0.12	
		(0.02)	
100 x Negative 4-Cent Error		0.71	
		(0.03)	
Indicator for Associate Professor			-0.50
			(0.28)
Indicator for Professor			-0.65
			(0.28)
Indicator for Other Rank			0.02
			(0.41)
Decile of Google Scholar Citations			0.07
			(0.05)
Indicator for Field: Applied Micro			-0.08
			(0.25)
Indicator for Field: Theory			-0.01
			(0.36)
Indicator for Field: Lab			0.72
			(0.22)
Indicator for Field: Psychology			0.06
			(0.26)
Indicator for having used Mturk			-0.22
			(0.19)
N	21,525	20,090	3,120
Log-likelihood	-150,460	-139,880	-20,729

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 and 2 include forecasts by all forecasters, except that forecasts on the 4-cent treatment are omitted in column 2 since accuracy of the forecast on the 4-cent treatment is used as a predictor of type. The fixed effects for forecasts on different treatments are restricted to be the same across all subject groups in columns 1 and 2, whereas it is allowed to vary by subject group in column 3. Only subject group indicators are used as predictors of type in column 1. In column 2, response time, a measure of the forecasters' confidence in their own forecasts, and accuracy of the forecast on the 4-cent treatment are added to the subject group indicators as predictors of type in the model. The sample for column 3 is restricted to academic experts and

Online Appendix Table 4. 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 (2.2)		-33.6 (1.9)		-69.1 (1.7)		-100.6 (1.9)		-2.5 (3.1)	
$v^{(2)}$ (Average Bias, Type 2)			-193.2 (5.5)		-600.8 (6.8)						-68.3 (3.9)	
$v^{(3)}$ (Average Bias, Type 3)											-705.7 (12.6)	
$\sigma^{(1)}$ (Idiosyncratic s.d., Type 1)			162.6 (2.7)		213.4 (1.3)		169.6 (2.6)		281.2 (1.4)		83.9 (5.9)	
$\sigma^{(2)}$ (Idiosyncratic s.d., Type 2)			357.6 (3.5)				374.2 (4.2)				252.4 (3.8)	
$\sigma^{(3)}$ (Idiosyncratic s.d., Type 3)											174.6 (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*	175.1	267.9	193.3	264.3	181.8	262.4	252.3	243.2	179.7	270.0
Average Absolute Error with 5 Forecasters	*113.7*	*173.3*	115.9	169.5	118.6	172.6	122.7	138.2	154.6	139.7	117.5	173.0
Average Absolute Error with 10 Forecasters	*103.8*	*156.5*	105.1	154.0	105.8	160.8	112.1	112.0	136.8	123.1	105.9	160.0
Average Absolute Error with 20 Forecasters	*98.3*	*150.4*	99.2	147.1	98.9	156.2	106.4	97.0	127.3	115.3	99.5	153.4
Wisdom-of-Crowds Absolute Error	*93.5*	*146.9*	93.7	144.0	92.4	152.6	101.0	83.7	115.2	110.9	93.7	150.2
Rank-Order Correlation:												
Average Individual Rank-Order Correlation	*0.41*	*0.42*	0.42	0.39	0.36	0.47	0.41	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	0.95	0.80	0.95	0.81	0.95
Percent Individual Forecasters Outperforming												
Wisdom-of-Crowds Absolute Error	*4.3*	*17.8*	1.0	18.9	0.1	19.0	1.0	0.1	0.1	0.1	7.5	9.4
Wisdom-of-Crowds Rank-Order Correlation	*4.8*	*0.3*	1.4	0.0	0.8	0.0	1.3	0.0	0.5	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	0.09	0.10	0.00	0.00	0.17	0.51

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 5. 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)
	0.10	0.15**	0.10*	0.14**
Other (Post-Doc or Research Scientist)	(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-			-0.00	-0.04
Making			(0.07)	(0.07)
Measure of Contextual Expertise				
Has Used Mturk in Own Research			0.05	0.03
(Self-Reported)			(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 6. 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	.	-112.22**	-61.26***		.	-0.374***	-0.308***	
0-4 Minutes		(52.13)	(20.83)			(0.132)	(0.046)	
Survey Completion Time	-11.14	6.04	33.80***		-0.152**	-0.008	0.026	
10-14 Minutes	(11.18)	(12.79)	(12.21)		(0.071)	(0.047)	(0.028)	
Survey Completion Time	-10.27	22.13*	42.82***		-0.196***	-0.035	0.001	
15-24 Minutes	(12.06)	(12.02)	(14.48)		(0.066)	(0.044)	(0.037)	
Survey Completion Time	-23.63**	21.20	-22.05		-0.292***	0.070	-0.115	
25+ Minutes	(11.52)	(12.88)	(33.67)		(0.071)	(0.047)	(0.100)	
Measures of Attention to Instructions								
Clicked on Practice Task	-3.14	-3.36	.		-0.068	0.031	.	
	(8.43)	(9.96)			(0.052)	(0.039)		
Clicked on Full Instructions	1.13	-29.64*	.		0.104*	-0.134**	.	
	(10.41)	(16.74)			(0.058)	(0.061)		
Delay in Survey Completion								
Days Waited to Take Survey	-0.08	-0.03	.		0.000	0.000	.	
(Since Invitation)	(0.25)	(0.87)			(0.001)	(0.002)		
Mturk Incentives and Experience								
Higher Incentives (up to $5)				-6.27				0.029
for Forecast Accuracy				(13.24)				(0.030)
Experienced the Task				-23.86**				-0.026
				(11.98)				(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%

Online Appendix Table 7. 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 Indicators 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 8. 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**	**Discounti ng**	**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** (3.73)	7.55*** (2.09)	18.66*** (3.54)	3.84 (3.06)	8.51** (3.68)	-3.91 (2.90)	17.95*** (4.50)	9.84*** (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*** (1.66)	28.32*** (1.77)	39.13*** (1.98)	17.20*** (2.04)	34.19*** (1.76)	28.83*** (2.50)	39.90*** (2.21)	43.33*** (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%

Online Appendix Table 9. 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		100.72	96.97
	(58.37)		(12.09)	(7.95)
Optimal 20% (4ct Control) (N=42)	173.14	-2.07	102.07	97.59
	(60.54)	(8.21)	(10.70)	(6.29)
Optimal 20% (No 4ct Control) (N=42)	175.06	-0.15	102.48	98.14
	(58.66)	(8.02)	(10.81)	(6.61)
Panel B. PhD/Undergraduates/MBA				
All PhD/UG/MBA (N=465)	188.89	13.68**	100.66	95.56
	(83.25)	(5.59)	(16.29)	(10.02)
Optimal 20% (4ct Control) (N=93)	147.97	-27.24***	76.5	73.24
	(42.26)	(5.95)	(10.77)	(6.70)
Optimal 20% (No 4ct Control) (N=93)	166.21	-9.00	87.78	83.94
	(67.40)	(8.05)	(13.30)	(8.35)
Panel C. Mturks				
All Mturks (N=762)	272.02	96.81***	147.8	144.39
	(143.23)	(6.58)	(39.73)	(26.11)
Optimal 20% (4ct Control) (N=152)	189.15	13.94*	81.2	76.69
	(82.04)	(7.78)	(17.7)	(12.04)
Optimal 20% (No 4ct Control) (N=152)	224.55	49.34***	107.9	102.29
	(128.52)	(11.16)	(28.32)	(18.81)

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%