# 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 $\eta_k^j$ instead of $\eta_k$ 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 $\eta_k^j$, 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 $\sigma$. 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 $\phi$ 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 $\eta_k^j$ 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 $\beta$ 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 $\hat{\eta}_k^j$ as well as the actual efforts in each treatment $\theta_k$, thus recover the "simulated forecast" $\widetilde{s}_{i,k} = z_{i,k} + \theta_k + \hat{\eta}_k^j$ 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 $\epsilon_i^k$. Second, they also differ because, while the probability of being the "good" type for a forecaster $i$ $p_i^1(x_i)$ 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 $\beta$ 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.