Expert Forecasts of Real-Effort Experiment – Pre-analysis Plan

Stefano DellaVigna, UC Berkeley & NBER

Devin Pope, University of Chicago & NBER

Introduction

We register a study intended as proof of concept for the elicitation of expert forecasts, in addition to being of interest in its own sake for behavioral interventions.

We run a field experiment on a real-effort task, randomizing subjects into 18 different treatments, each of which aims at testing the effectiveness of either standard economics levers (incentives) or of a variety of behavioral levers and nudges. In particular, we plan to examine the importance of time preferences, altruism, loss aversion, probability weighting, crowd-out of motivation, gift exchange, social comparisons, and ranking incentives, among others.

In order to gather a large enough sample (nearly 10,000 subjects), we run the experiment using subjects from Amazon Mechanical Turk. This format also allows for a transparent summary of all the different conditions to the forecasters, since the only difference between conditions consists of one or two paragraphs of text, holding everything else constant. This experiment is pre-registered as AEARCTR-0000714 ("Response of Output to Varying Incentive Structures on Amazon Turk").

We then invite experts in the areas of behavioral economics, laboratory experiments, decision-making, and standard economics to make forecasts about the amount of effort that will be demonstrated in each of the treatments. The simple set-up makes it possible to record forecasts for 15 experimental arms within a 5-15 minute survey. The analysis of these forecasts form the basis of the study we are pre-registering.

We form the list of experts starting from the list of attendees, paper participants, and program members at leading conferences of behavioral economics, experimental economics, and judgment and decision making. We also collect additional information on the experts aimed at measuring their academic rank as well as areas of expertise. This will allow us to measure the impact on forecast accuracy of both *vertical* expertise (academic rank and citations) as well as *horizontal* expertise (publications in the area).

Research Methodology

The ultimate goal of this research line is to make a case for the use of expert forecasts as a useful tool for researchers. As a case in point, we wanted to design a study with the following four *desiderata* in mind: (i) the presence of multiple treatments to forecast, making the study of expert accuracy more relevant; (ii) a large sample size in each treatment arm, guaranteeing a well-powered study; (iii) the ability to present the different treatments concisely to the forecasters; (iv) the ability to observe the experimental results in a short time frame, so as to provide feedback to the experts afterwards.

After considering a number of options, we opted for a real-effort experiment run on the Amazon Mechanical Turk (MTurk). 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 can browse the set of postings and choose to complete any HIT for the amount of money offered. MTurk has become very popular for experimental research in marketing and psychology (Paolacci and Chandler, 2014) and is also used increasingly in economics, for example for the study of preferences about redistribution (Kuziemko, Norton, Saez, Stantcheva, forthcoming).

The recruitment of subjects on Amazon Turk is easy and low-cost, making it possible to aim for a sample of 5,000 to 10,000 subjects, thus making (i) and (ii) possible. Furthermore, while in other field settings a number of variables may affect the results making the job of forecasters particularly difficult, the MTurk online platform allows for more experimental control: we can show in a concise manner the forecasters the *exact* treatment conditions as shown to the subjects, thus achieving (iii). Finally, we are able to recruit the subjects within three weeks from the start of the experiment; once the results are collected, we will reach out to the forecasters and about 2 or 3 months later we will communicate the results to the forecasters and the general public. This makes it possible to achieve also goal (iv).

While the MTurk study was conceived with the above feasibility constraints in mind, it also offers unique insights for behavioral researchers and decision-making experts in general. Over the last few years, both researchers and the general public have become more interested in behavioral economics, partly in the hope that nudges (Thaler and Sunstein, 2009) can help address phenomena such as under-saving, over-eating, excessive energy consumption, student under-performance. These interventions often attempt to address self-control problems, or lever reference dependence and loss aversion, or take advantage of social comparisons. But how effective do researchers believe these levers are?

In the real-effort experiment we introduce 18 different treatments which span many of the above behavioral levers, in addition to more standard incentives. The basic task is to press two keys alternating "A" and "B" for ten minutes. This is a real-effort task similar to those used in the literature (Amir and Ariely, 2008; Berger and Pope, 2011). It is simple to explain and does not require particular ability – and yet, given that it gets tiresome over time, participants often slow

down or stop before the end of the 10 minutes if insufficiently motivated. It is thus a good setting to study the impact of incentives and other levers to create sustained effort.

Treatments. We recruit subjects offering a $1 pay for a 15-minute task (a generous payment for MTurk). In our first three treatments, we explore the power of incentives to affect behavior. We compare a condition in which we state (truthfully) that "*your score will not affect your payment in any way*" to a 1-cent piece rate condition ("*As a bonus, you will be paid an extra 1 cent for every 100 points that you score*") and a 10-cent piece rate condition ("*As a bonus, you will be paid an extra 10 cents for every 100 points that you score*"). The incentives will on average increase pay from a flat $1 payment in the first condition to about $1.20 for the second condition (assuming an average of 2,000 points) to about $3.20 for the third condition (assuming an average of 2,200 points). While it could be of interest to ask experts to forecast the effect of these incentive treatments, we decided instead to inform the experts of the average performance in these three treatments. This allows the forecasters to familiarize themselves with the performance of MTurkers and to gauge the approximate slope of the cost-of-effort function.[1]

In these first three treatments of the study pre-registered as AEARCTR-0000714 we find that average performance increases from an average of 1,522 in the first treatment to 2,028 in the 1-cent treatment and 2,175 in the 10-cent treatment, differences that are highly statistically significant given the large sample of about 550 subjects in each treatment, for a total of nearly 10,000 subjects. The fact that performance responds to incentives validates the choice of this task as one in which it is interesting to compare alternative motivations for behavior.

[1] Indeed, it is possible to estimate the level and curvature of a power cost of effort function using the productivity in the three treatments.

The remaining 15 treatments, which we ask the experts to forecast, span some of the most important behavioral features. The first two are additional incentive conditions, one which aims to test whether paying too little lowers effort as in Gneezy and Rustichini (2008) ("*As a bonus, you will be paid an extra 1 cent for every 1,000 points that you score*") and a second one that aims to test whether experts take into account appropriately the convexity of the cost of effort function ("*As a bonus, you will be paid an extra 4 cents for every 100 points that you score*").

Next, we explore the importance of social preferences and altruism in particular ("*As a bonus, the Red Cross charitable fund will be given 1 cent for every 100 points that you score.*"). We test if effort in this condition is higher than in the corresponding piece rate condition, and whether it is higher than the no-piece-rate condition. Furthermore, we test if effort responds to the amount of giving to the charity, as in a pure altruism model, or not, more consistent with a warm glow (Andreoni, 1990) or norm-type response ("*As a bonus, the Red Cross charitable fund will be given 10 cents for every 100 points that you score.*")

Next, we consider the impact of time preferences and discounting, by varying the date of delivery of the bonus ("*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*" versus "*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*"). We can compare the effort across these conditions and to the effort in the condition with immediate payment of bonus. In fact, under an assumption about the convexity of the cost of effort function, one can derive the implied impatience parameters beta and delta given the observed effort in the first three conditions. Given the increasing use of questions aimed at

measuring present-bias in surveys, it is interesting to compare the results and expert forecasts on the topic (Laibson, 1997; Frederick, Loewenstein, and O'Donoghue, 2002)[2].

We also study the role of reference dependence (Kahneman and Tversky, 1979) and in particular loss aversion. We frame a condition in both a gain frame ("*As a bonus, you will be paid an extra 40 cents if you score at least 2,000 points*") and a loss frame ("*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*"). This treatment is inspired by Hossain and List (2012) and by Fryer, Levitt, List, and Sadoff (2012) who show that framing incentives as a loss leads to increased effort by, respectively, employees and teachers. To test whether experts believe that the loss aversion coefficient is larger or smaller than 2, we also elicit forecasts for a gain condition with twice as large the incentives ("*As a bonus, you will be paid an extra 80 cents if you score at least 2,000 points*").

Next, we turn to risk aversion and in particular probability weighting (a component of prospect theory; Prelec, 1998 and Wu and Gonzalez, 1996). A first condition reads "*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.*" Notice that the expected value of the incentives is the same as in the 1-cent incentive condition. A second condition reads "*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*". The expected value of incentives is again the same, but probability

[2] An important caveat here is that present-bias in principle should apply to the utility of consumption and real effort, not to the monetary payments *per se*, since such payments can be consumed in different periods (Augenblick, Niederle, and Sprenger, forthcoming). Having said this, the elicitation of beta and delta using monetary payments is very common and our role is to compare forecasts to experimental results, not to provide ideal measures of present bias.

weighting, given the magnification of small probabilities, should lead to higher effort in the 1% treatment. This is a feature of prospect theory that is employed for example in Loewenstein, Brennan, and Volpp (2007) to enhance the effectiveness of incentives for health.

Next, we consider a treatment in the spirit of the gift exchange literature (Akerlof, 1982; Fehr, Kirchsteiger, and Riedl, 1992; Gneezy and List, 2006). Namely, we examine whether an increase in pay, unconditional on performance, increases effort: "*In appreciation to you for performing this task, you will be paid a bonus of 40 cents.*"

The last three treatments are drawn mostly from the psychology and decision-making research. The first is a simple version of social comparisons (Cialdini et al., 2007) and anchoring ("*In a previous version of this task, many participants were able to score more than 2,000 points*"). Cialdini's research shows that under some conditions social comparisons can be effective.[3]

The second treatment takes advantage of the competitive effects of ranking performance ("*After you play, we will show you how well you did relative to other participants who have previously done this task*"). In some experiments, the comparison to others, even with no extra incentives, leads to higher effort (Bandiera, Barankay, and Rasul, 2013; Barankay, 2012).

The final manipulation regards task significance ("*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.*"). This treatment builds on Grant (2008) and follow-up research which documents how employees work harder at a task if they are told about its importance.

[3] Notice that the question is framed to ensure no deception of the respondents. Indeed, many subjects in the pilot were able to score above 2,000 points.

Data. We have collected the results of this 18-arm pre-registered field experiment, for which we have IRB approval at the University of Chicago and at UC Berkeley, following the procedures that we pre-registered as AEARCTR-0000714 ("Response of Output to Varying Incentive Structures on Amazon Turk"). In particular, we decided to run the MTurk experiment *before* we survey the forecasters. This choice allows us to provide the experts with information on the average productivity in the three baseline and piece rate treatments which serve as anchors for the forecast. At the same time, this raises at least in principle issues of contamination of forecasts if, perhaps unintentionally, some of the experts came to hear of the results. To address this issue, we have specified a procedure such that even the PIs will not see the results of the experiments prior to the forecasting stage. Namely, the PIs have written a simple do file that creates key summary statistics (across all treatments) as well as the more detailed information about the three piece rate treatments. This do file does not output any information on the 15 treatments to be forecasted. An RA for the project in Chicago then runs the do file about once a day during the period of data collection to ensure there are no problems. This ensures enough monitoring on data collection while enabling the researchers to be blind with respect to the key results. Thus, the PIs themselves will be able to record a set of forecasts.[4] The data collection for this study was completed in mid June 2015.

This do file also specifies the exact sample of mTurk workers who are used in the final sample. We did several sample restrictions in order to produce the final sample. Most of these sample restrictions were anticipated (e.g. people who failed to complete the survey) and some were not (e.g. a technical glitch in the software that affected some participants). Importantly, these sample

[4] We should point out though that the PIs have seen the results of a relatively small pilot with about 500 subjects designed to make sure all the procedures work correctly.

restrictions were done prior to the PIs seeing any of the results for the 15 treatments to be forecasted. Below we list each of the sample restrictions and the impact that it had on our final sample size.

The initial sample consists of 12,838 MTurk workers who started our experimental task. Of these, 721 were dropped because they experienced technical problems with the survey. This technical problem occurred over a several-hour period when the software program Qualtrics moved to a new server. Individuals during this time period experienced a malfunctioning of the counter that kept track of their scores. Next, 48 workers were dropped for scoring above 4,000 points. During a small pilot, we determined that scoring more than 4,000 points was physically impossible for the majority of our workers, and thus we worried that any score of 4,000 would be due to using a cheat (e.g. a key-binding program). Also, 1,543 workers were dropped because they failed to complete the experiment (for example, many participants only filled out the demographics portion of the experiment and were never assigned a treatment). Next, 364 workers were dropped because they stopped the task and logged in again. We stated in the instructions to the workers that they could not stop the task and log in again. This restriction was put into place so as to discourage workers who may want to log in and obtain a different treatment. In addition, 187 workers were dropped because their HIT was not approved for some reason (e.g. they did not have a valide MTurk ID). Finally, 114 observations were dropped because they never did a single button press. We were concerned that these participants may have experienced a technical malfunction or that their results were simply not recorded for some reason. After these sample restrictions, we are left with 9,861 completed tasks with valid results.

Contacting Experts. With the results collected and the three baseline treatment results stored, we will contact via email a group of 300+ experts to ask for their forecasts. The email will provide the link to a conveniently formatted Qualtrics survey where the experts find an explanation for the survey as well as the results of the first three treatments. The experts are then invited to forecast the average effort in the remaining 15 treatments using a convenient slider scale (we reproduce a copy of the survey in the Appendix). We also elicit the confidence of the experts, and provide an incentive pay of up to $1,000 to five selected forecasters as incentive for accuracy.

We determined the group of experts as follows. We collected the list of all authors of papers presented at the Stanford Institute of Theoretical Economics in Psychology and Economics and in Experimental Economics from its inception until 2014 (for all years in which the program is online). We combine this list with participants of the Behavioral Annual Meeting (BEAM) conferences from 2009 to 2014, and with the program committee and keynote speakers for the Behavioral Decision Research in Management Conference (BDRM) in 2010, 2012, and 2014. We also include a list of invites to the Russell Sage Foundation 2014 Workshop on "Behavioral Labor Economics" and a list of behavioral economists compiled by ideas42. In addition, we include researchers with at least 5 highly-cited papers (at least 100 Google Scholar citations) in relevant keywords and a small number of additional experts that would have been missed by the criteria above. This starting list provides a long list of over 600 people. Since we did not want to be seen as spamming researchers, we further pared down the list to a little over 300 researchers to whom at least one of the two PIs had some connection.

We notify each of these researchers via email inviting them to make forecasts on the results of a real-effort experiment and provide them a link to a unique version of the survey. We then store the

results of their forecasts for those that respond, and send a reminder email to those who have not responded within one week.[5]

In a second round of survey collection, we also collect the results of forecasts of a broader group, including PhD students in economics, MBA students at the Booth School of Business at the University of Chicago, and a group of MTurk subjects that are recruited for the purpose.

Given the experimental treatments and the expert forecasts, we can then analyze the *overall* accuracy of forecasters. As an example using simulated data, Figure 1a displays a scatterplot of average performance versus the forecast for a very well-calibrated forecaster: the average results and the forecasts are highly correlated. Figure 1b instead displays an example of noisier forecasts.

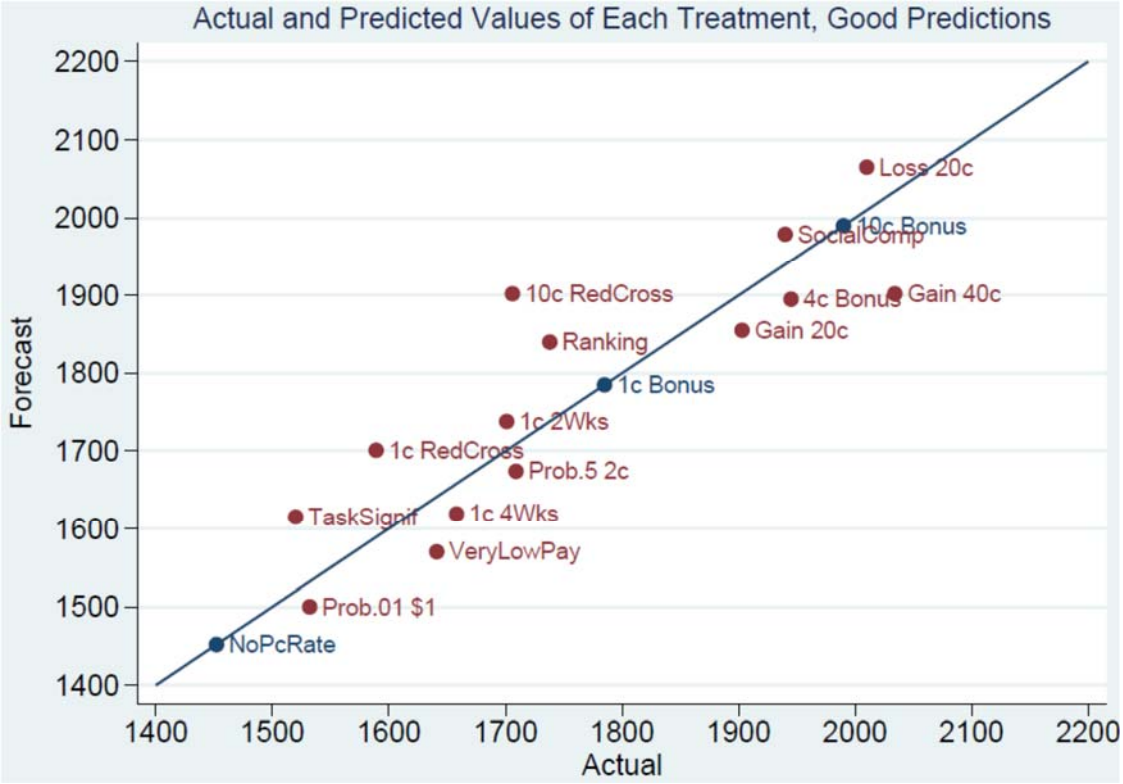

Figure 1a. Hypothetical Example, high forecasting ability.

[5] The Qualtrics survey sent to the experts is identical for all respondents other than for the order of the treatments, which is randomized in groups to control for possible order effects.

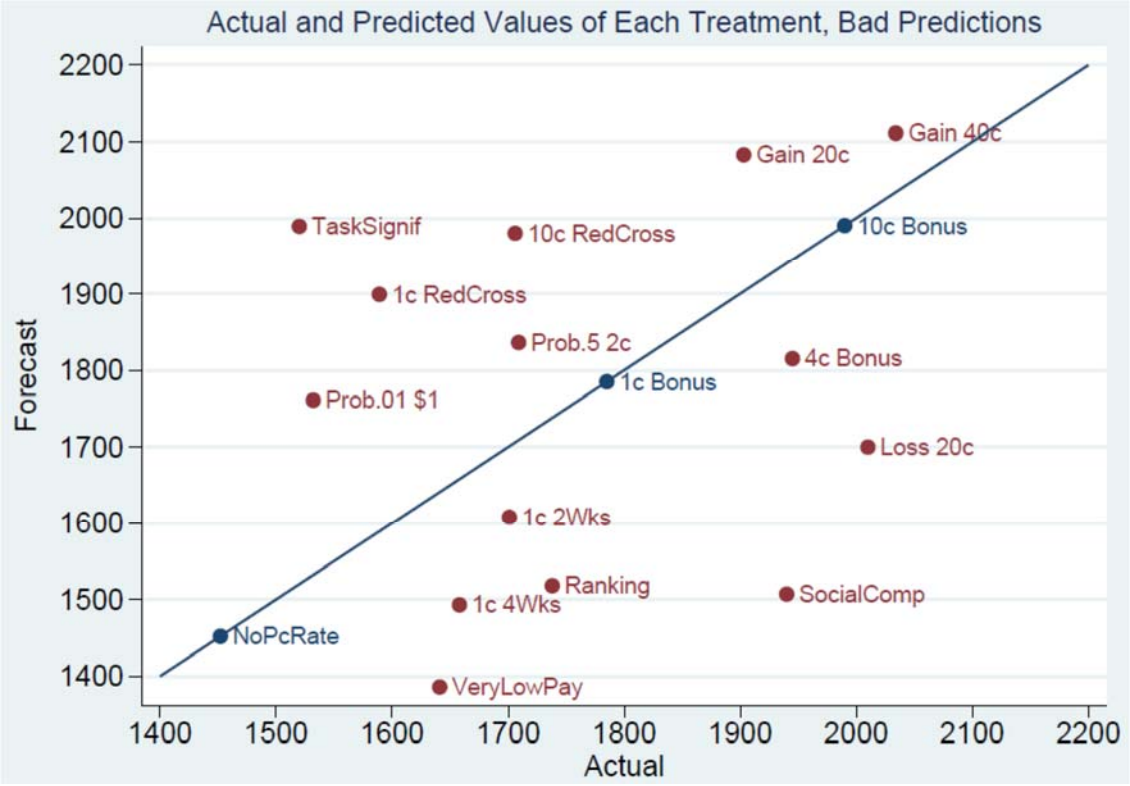

Figure 1b. Hypothetical Example, low forecasting ability.

While the accuracy of the average forecaster is of interest, a main focus of our design is the ability to estimate the forecasting ability for different types of experts. A first relevant dimension of expertise is *vertical*. For each of the experts contacted, we obtain measures of their academic rank: PhD student, Assistant Professor, Associate Professor, and Professor--as well as the number of years since graduation. Further, as a measure of academic achievements, we collect information on citations from Google Scholar for each expert. Using these variables, we can then examine whether more prominent authors do better in their forecasts than less prominent authors. Notice that the result is not obvious: while *vertical* expertise is likely to help, it is possible that more prominent experts produce lower quality forecasts for example because of more demands on their time. As part of the survey design, we can measure this margin in part by recording the time taken

to undertake the survey as well as whether they click on the practice task and on the detailed instructions.

In addition to *vertical* expertise, we also consider the impact of *horizontal* expertise, defined as knowledge about the field at hand. In particular, we aim to study whether expertise in a particular area, such as reference dependence, makes one a better forecaster for treatments involving that phenomenon. To make such horizontal comparisons, we coded publications of the experts in a number of key topics[6], allowing us to test whether, holding constant a given vertical prominence, a given researcher does better in the treatments ideally suited to him or her. Also along a horizontal dimension, we aim to compare accuracy for expertise in different areas, such as behavioral economics, lab experiments, economic theory, and so on.

In addition to the comparison between and within experts in accuracy, we can also compare the accuracy of *individual* forecasts versus the accuracy of *group* forecasts. Namely, we can test for wisdom of the crowd effects: the forecasts of non-experts, while not as accurate individually as the forecasts of experts, may be comparable to the expert forecasts when averaged. This type of result for example appears in some data sets for inflation forecasts.

Finally, we are interested in testing for *overconfidence* of the experts about the accuracy of their forecasts. We ask forecasters to predict how many of their 15 forecasts will fall within 10% of the actual average effort in the respective treatment. We can then compare this prediction to the actual number. Are experts overconfident about the precision of their forecasts like many lay people? Are higher experts, in the vertical sense, more, or less, overconfident? We also ask a second group of confidence-related questions. We ask experts "*for each group of people below, please indicate*

[6] We use Google Scholar and search for prominent publications using keywords related to the experimental conditions, such as "reference dependence", "loss aversion", "altruism", etc. We aim to record papers with around 100 citations or higher.

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." We then list different groups of expert participants, such as experts with economics PhDs versus experts with a PhD in psychology or decision-making, or MTurk workers. We can thus test the beliefs of experts about the expertise of others.

While the above tests are largely focused on tests of the forecasting ability of experts, this rich data set of expert forecasts also allows us to collect information on the priors of experts regarding some key behavioral phenomena.[7] Independent of the results, it will be interesting to see which treatments experts expect to be most effective at increasing effort, and which ones least effective. We are also interested in the *variance* of the forecasts: which are the treatments for which the experts largely agree in their forecasts, and which are the treatments for which they are divided? For example, it is quite possible that experts would be quite divided on the crowd-out of incentives, while they may agree on the impact of loss aversion. Or it may quite be the opposite – we have no way to tell given that no such evidence exists to this date. Given that the treatments cover a large swath of behavioral economics, the elicitation of such priors will be of interest.

We should also add that this main focus of the study – on comparing accuracy of forecasts across researchers and by topic, as well as the elicitation of the variance of priors -- is not possible using prediction markets which aggregate the priors to the ones of the "marginal investors". While that is a useful number, we argue that understanding the distribution of beliefs and its determinants is an important endeavor, in addition to capturing an average forecast.

[7] We should clarify that, to keep the survey format as simple and intuitive as possible, we do not recover Bayesian priors given that we do not ask for a confidence interval in the predictions. This Bayesian angle is one that we hope to cover in follow-up research.

Appendix – Bibliography

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

Andreoni, James. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving", *Economic Journal*. Vol. 100(401) (1990), pp. 464-477.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. "No Margin, No Mission? A Field Experiment on Incentives for Pro-Social Tasks." *Journal of Public Economics* Vol. 120 (2014): 1-17.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. "Working over time: Dynamic inconsistency in real effort tasks" *Quarterly Journal of Economics*, forthcoming.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. "Team Incentives: Evidence from a Firm Level Experiment." *Journal of the European Economic Association* Vol. 11, No. 5 (2013): 1079-1114.

Barankay, Iwan. "Rank Incentives: Evidence from a Randomized Workplace Experiment." Working Paper (2012).

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

Bertrand, Marianne and Sendhil Mullainathan. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* Vol. 94, No. 4 (2004): 991-1013.

Camerer, Colin. "Behavioral Economics: Reunifying Psychology and Economics." *Proceedings of the National Academy of Sciences of the United States of America* Vol. 96, No. 19 (1999): 10575-10577.

Cialdini, Robert M., et al. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science* Vol. 18, No. 5 (2007).

Deci, Edward L. "Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology* Vol. 18, No. 1 (1971): 105-115.

DellaVigna, Stefano. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature* Vol. 47, No. 2 (2009): 315-372.

Erev, Ido, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan M. Herzog, Robin Hau, Ralph Hertwig, Terrance Stewart, Robert West, and Christiane Lebiere, "A Choice Prediction Competition:

Choices from Experience and from Descirption" *Journal of Behavioral Decision Making,* 23 (2010): 15-47.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature* Vol. 40, No. 2 (2002): 351-401.

Fryer, Roland Jr., Stephen D. Levitt, John A. List, and Sally Sadoff. "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." NBER Working Paper No. 18237 (2012).

Gneezy, Uri and Aldo Rustichini. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics* Vol. 115, No. 3 (2000): 791-810.

Grant, Adam M. "The Significance of Task Significance: Job Performance Effects, Relational Mechanisms, and Boundary Conditions." *Journal of Applied Psychology* Vol. 93, No. 1 (2008): 108-124.

Haggag, Kareem, Devin Pope, and Justin Sydnor. "Projection Bias and Commitment Devices." Mimeo, 2014.

Hossain, Tanjim and List, John A. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Science* Vol. 58, No. 12 (2012): 2151-2167.

Kahneman, Daniel and Amos Tversky. “Prospect Theory: An Analysis of Decision Under Risk.” *Econometrica* Vol. 47, No.2 (1979): 263-292.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. “How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments.” *American Economic Review* (forthcoming).

Laibson, David. “Golden Eggs and Hyperbolic Discounting.” *Quarterly Journal of Economics* Vol. 112, No. 2 (1997): 443-477.

Loewenstein, George, Troyen Brennan, and Kevin G. Volpp. “Asymmetric Paternalism to Improve Health Behaviors.” *Journal of the American Medical Association* Vol. 298, No. 20 (2007): 2415-2417.

Milgram, Stanley. “Behavioral Study of Obedience.” *The Journal of Abnormal and Social Psychology* Vol. 64, No. 4 (1963): 371-378.

Mullainathan, Sendhil and Richard H. Thaler. “Behavioral Economics.” In *International Encyclopedia of the Social & Behavioral Sciences* (2001): 1094-1100.

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

Pennisi, Elizabeth. “A Low Number Wins the GeneSweep Pool.” *Science* Vol. 300, No. 5625 (2003): 1484.

Prelec, Drazen. “The Probability Weighting Function.” *Econometrica* Vol. 66, No. 3 (1998): 497-527.

Rabin, Matthew. “Psychology and Economics.” *Journal of Economic Literature* Vol. 36, No. 1 (1998): 11-46.

Tetlock, Philip E. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press (2005).

Tetlock, Philip E., et al. “The Good Judgement Project: A Large Scale Test of Different Methods of Combining Expert Predictions.” AAAI Technical Report FS-12-06 (2012).

Thaler, Richard H. and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New York: Penguin Group (2009).

Wu, George and Richard Gonzalez. “Curvature of the Probability Weighting Function.” *Management Science* Vol. 42, No. 12 (1996): 1676-1690.

1 Appendix A: Simple Model

Basic Incentive Treatments. The participant in the real-effort experiment maximizes the return from effort e (the button presses) net of the cost of effort. To draw the parallel to the experiment, assume that e denotes the number of A-B presses in hundreds. We assume that for each effort unit e the individual receives an intrinsic reward, s, and the piece-rate p. (Notice that for $s = 0$ effort would equal zero in the no-piece rate treatment) We assume a cost of effort function $c(e)$ which satisfies $c'(0) = 0$, $c'(e) > 0$ and $c''(e) > 0$ for all $e > 0$. Thus, assuming risk-neutrality, an individual solves

$$\max_{e \geq 0} (s+p)e - c(e),$$

leading to

$$e^* = c'^{-1}(s+p).$$

A useful special case is the power cost function $c(e) = ke^{1+\gamma}/(1+\gamma)$. This parametrization is characterized by a constant elasticity of effort $1/\gamma$ with respect to the value of effort. Under this assumption, we obtain

$$e^* = \left(\frac{s+p}{k}\right)^{1/\gamma}. \tag{1}$$

Optimal effort e^* is increasing in the piece rate p and in the intrinsic motivation s and decreasing in the cost parameter k. Notice that (1) has three unknowns, p, κ and γ. Given that the experts are informed of the observed effort for three different piece rates p, it is in principle possible to back out on estimate for the three parameters.

Pay-Enough or Don't Pay at All Treatment. In the treatment with $p = .001$, the low pay may crowd out the subjects' motivation, which we can model as a decrease in s.

Charities Treatments. For the charitable giving treatments, we assume pure altruism, with altruism α toward the charity receiving payment p_{ch} for each unit of effort e. The optimal effort then is

$$e^*_{ch} = \left(\frac{s+\alpha p_{ch}}{k}\right)^{1/\gamma}. \tag{2}$$

Thus, it is possible to infer the implied approximate α comparing the observed e^*_{ch} and the corresponding e^*, given (1) and (2).

Discounting Treatments. Similarly, we model the treatments with delayed payment of the bonus with a present-bias model:

$$e^*_t = \left(\frac{s+\beta\delta^t p}{k}\right)^{1/\gamma},$$

where β is the short-run impatience factor and δ is the long-run discounting factor. (Notice that in principle discounting should apply to consumption, not monetary payments, but we

apply it along lines commonly used in the literature). Once again, by comparing e_t^* in the discounting treatments to e^* in the piece rate treatments it is possible to approximately back out $\beta\delta^t$.

Probabilistic Treatments. In the stochastic treatments, the subjects earn piece rate p_{st} with probability P, and no piece rate otherwise. The parameters p_{st} and P are chosen such that $p_{st} * P = \$0.01$, the piece rate in the first incentive treatment. Specifically, we run the combinations $(p_{st}, P) = (\$0.02, 0.5), (\$1, 0.01)$. The utility maximization is

$$\max_{e\geq 0} se + \pi\left(P\right) u\left(p\right) e - c\left(e\right),$$

where is $u\left(p\right)$ is the (possibly concave) utility of payment, and $\pi\left(P\right)$ is the probability weighting. The number of button-presses is given by

$$e^* = \left(\frac{s + \pi(P)u\left(p\right)}{k}\right)^{1/\gamma}.$$

With probability weighting, given $\pi(0.01) \gg 0.01$ while $\pi(0.5) \approx 0.5$ we expect that, for $u\left(p\right)$ approximately linear, effort will be higher in the condition with a 0.01 probability of a \$1 piece rate. Conversely, with no probability weighting and concave utility, effort will be higher in the condition with a 0.5 probability of a \$0.02 piece rate.

Gain-Loss Treatment. For the gain treatment with threshold T, subjects can earn a payment G (equal to \$0.40 or \$0.80) if they exceed a target performance T. Following the Koszegi-Rabin gain-loss notation (but with backward-looking reference points), the decision-maker maximizes

$$\max_{e\geq 0} se + \mathbf{1}_{\{e\geq T\}}G + \eta\left(\mathbf{1}_{\{e\geq T\}}G - 0\right) - c\left(e\right). \tag{3}$$

The first term, $se + \mathbf{1}_{\{e\geq T\}}G$, captures the 'consumption' utility, while the second term, $\eta\left(\mathbf{1}_{\{e\geq T\}}G - 0\right)$, captures the gain utility relative to the reference point of no bonus; the parameter η denotes the weight on gain utility, which is often parametrized at 1. In the loss treatment, the decision-maker is assumed to take performance T as the reference point and thus maximizes

$$\begin{aligned} &\max_{e\geq 0} se + \mathbf{1}_{\{e\geq T\}}G + \eta\lambda\left(0 - \mathbf{1}_{\{e<T\}}G\right) - c\left(e\right) \text{ or} \\ &\max_{e\geq 0} se + \mathbf{1}_{\{e\geq T\}}G + \eta\lambda\mathbf{1}_{\{e\geq T\}}G - \eta\lambda G - c\left(e\right) \end{aligned} \tag{4}$$

Notice that conditions (3) and (4) would lead to the same solution for $\lambda = 1$, but with $\lambda > 1$ effort is higher in the loss treatment for a fixed G. Notice also that the gain condition for $G = \$0.80$ should lead to the same effort as the loss condition for $G = \$0.40$ for parameters $\eta = 1$ and $\lambda = 3$ (these are the standard parametrizations of loss aversion under Koszegi-Rabin gain-loss utility, parallel to the parametrization $\lambda = 2$ in Kahneman and Tversky's prospect

theory).

Social Comparison/Ranking/Task Significance Treatments. The basic model accommodates these treatments with a change in the intrinsic reward parameter s.

2 Appendix B: Planned Analysis

For the analysis of the data, we envision three main sections. In the first section, we document the forecasts themselves, including analyzing the treatments for which experts differ most in their beliefs, and network effects in belief formation. The second section relates the forecasts to the average actual effort by the mTurkers. The closer the forecasts are to the actual effort, the higher the accuracy and thus the expertise (in forecasting experimental results). We construct several measures of such accuracy first for the overall sample and then by group, examining horizontal and vertical expertise. In the third section, we relate these results to (over)confidence, and we collect information on beliefs of accuracy as elicited in page 2 of the survey.

Analysis of Forecasts. In this first step, we plot the forecasts for the 15 treatments (recall that the results for 3 treatments are provided). We intend to plot not just the mean, but also statistics to document the variance of forecasts, plotting for example the 90th and 10th percentile of forecasts. This captures in a snapshot the beliefs of the experts regarding the impact of the different behavioral levers, at least in our context. For example, does the discipline believe in crowd-out of motivation with low pay? Does it believe in gift exchange? Does it believe that a delayed payment will affect effort in a β, δ fashion?

The data provides us with some qualitative answers on the questions above by comparing the forecast for the behavioral treatment to the relevant control (comparison) treatment. For example, the test for whether people expect crowd-out of effort when pay is very low (1 cent every 1,000 point) is to compare the forecast $f_{i,CO}$ to the effort for the treatment without pay, $\bar{e}_{p=0}$. Notice that since the latter treatment is revealed to the subjects, we are comparing to the actual effort $\bar{e}_{p=0}$, not to another forecast. Thus, a measure of belief in crowd-out is the share of experts i for which $f_{i,CO} - \bar{e}_{p=0}$ is negative.

Similarly, the test for gift exchange (a 40 cent unconditional payment) is again comparing to the effort in the no-payment treatment: what share of experts believe that the 'gift exchange' payment will lead to a higher effort then with no gift exchange payment, that is $f_{i,GE} - \bar{e}_{p=0} > 0$? The comparison for the delayed-payment treatments is to the treatment with the same payment but no delay; thus, we will focus on $f_{i,2w} - \bar{e}_{p=.01}$ and on $f_{i,4w} - \bar{e}_{p=.01}$.

For most of the behavioral treatments examined, the relevant comparison point is to one of the three piece rate treatments which are revealed to the forecasters. Thus, the comparison is to a known effort. For the case or loss aversion, though, the comparison is instead to another

treatment to forecast, namely comparing the loss treatment to the two gain treatments. In these cases, the relevant measure of belief is $f_{i,.4Loss} - f_{i,.4Gain}$ and $f_{i,.4Loss} - f_{i,.8Gain}$.

The above comparisons use the reduced-form difference between forecasts in different treatments to evaluate the beliefs about the effectiveness of a particular behavioral factor. While this constituted the main envisioned focus of the project, the above model makes clear that, at least for some of the treatments it is possible to obtain structural estimates. That is, we can translate a forecast into a value for a relevant parameter (such as altruism α, time discounting β and δ, or loss aversion λ), provided that we can estimate the parameters of the cost of effort function. That is, the comparison $f_{i,2w} - \bar{e}_{p=.01}$ can be used to estimate the belief regarding $\beta\delta$ for expert i. In order to back out the cost of effort parameters, we can use, under some assumptions, the 3 piece rate treatments which we make public.

Having established the framework to examine the forecasts, we can compare the forecasts for the different groups – experts versus PhD students versus MBAs versus mTurkers. We also construct simple networks based on presence in the same institution and institution of PhD to examine whether closeness by this network measure reduces the distance in the forecasts.

Finally, as a separate point, we also consider the response rate to the survey within each group, since the response rate affects the representativeness of the results, and is of interest in itself.

Accuracy and Expertise. In this second part of the analysis, we relate forecasts to the mean effort in each treatment. The main set of graphs is the one in the main pre-analysis document, plotting a scatterplot with each treatment t as observation, on one axis having the actual effort $\bar{e}_t$ and on the other the (average) forecast $\bar{f}_t$. Closeness to the 45-degree line indicates higher accuracy of the forecasters. We aim to draw this plot overall using the average forecast $\bar{f}_t$, and then for each group g, that is, using $\bar{f}_{g,t}$. Indeed, such a graph can in principle be drawn individually for each forecaster i, using his/her forecasts $f_{i,t}$ to create the expert-specific graph. We indeed aim to send one such graph as feedback to the individual expert forecaster at the end of the experiment (these graph will remain confidential and only the forecaster will receive it).

To proceed to a statistical analysis, we envision the following regression framework. Denote by $\bar{e}_t$ the average observed effort in treatment t and with $f_{i,t}$ the forecast of expert i regarding treatment t. We can then define as measures of accuracy (expertise) the distance between observed effort and the forecast, $d\left(\bar{e}_t, f_{i,t}\right)$. We envision two main measures, absolute distance ($d_A\left(\bar{e}_t, f_{i,t}\right) = -\left|\bar{e}_t - f_{i,t}\right|$) and quadratic distance ($d_Q\left(\bar{e}_t, f_{i,t}\right) = -\left(\bar{e}_t - f_{i,t}\right)^2$). For ease of interpretation, we introduce a negative sign, so expertise is less distance. We can also build overall measures of distance, which are averages across the 15 treatments of the distance measure. Indeed, the average of the quadratic distance (with negative sign) factors into our compensation scheme. We chose it because an individual attempting to maximize

$-E_i \sum_t (\bar{e}_t - f_{i,t})^2$ will provide $f_{i,t} = E_i e_t$. We can then use either measure of distance to examine vertical expertise, expertise by group, and horizontal expertise.

Consider first *vertical expertise*. Denote by v_i the relevant measure of vertical expertise for expert i, such as the academic rank or citation indicators (more on this below). Then the basic regression-based specification would be

$$d(\bar{e}_t, f_{i,t}) = \alpha + \beta_V v_i + \varepsilon_{i,t}, \tag{5}$$

where each observation is a forecast by person i for treatment t. (We envision clustering the standard error at the person level for this analysis) Specification (5) could contain treatment fixed effects η_t and can be done for each treatment separately. That is, we are interested in asking in which treatments vertical expertise is associated with higher accuracy. The benchmark measure would be the overall test of vertical expertise.

As for the actual measures of vertical expertise, we envision using (i) indicators for Professor versus Associate Professor versus Assistant Professor versus PhD student; (ii) Google Scholar counts stored in groups, both unconditional and conditional on years since PhD[1]; (iii) Years since PhD.

On vertical expertise, we envision three hypotheses. The first hypothesis ($\beta_V > 0$) is that individuals with more expertise in the field (by any measure above) are more accurate in their forecasts, presumably due to their superior knowledge. A second hypothesis ($\beta_V = 0$) is that experience in a discipline does not translate into ability to forecast experimental results. The third hypothesis ($\beta_V < 0$) is that in fact experience in a field leads to reduced ability to forecast future experimental results, presumably through higher time costs. We provide evidence on the latter channel using the time taken to answer the survey, and the share of experts who click through the detailed instructions and the trial test.

The above analysis is envisioned at the individual level. However, we also intend to compute the accuracy of the *average* forecast for a group to examine phenomena such as the *wisdom of crowds*. It is possible, for example, that experts individually have higher accuracy compared to MBAs, but that the accuracy for the group on average for the two groups may be the same. To examine this, we can similarly examine

$$d(\bar{e}_t, \bar{f}_{g,t}) = \alpha + \beta_V v_g + \varepsilon_{g,t}, \tag{6}$$

where in this case an observation is a group g in treatment t, and the forecast has been averaged across experts i in group g. (Here, we cannot cluster by group if comparing just two groups)

Next, we turn to expertise *by group*. In comparing MBAs to PhD students, there is no obvious vertical difference in expertise, but the two groups differ in the type of expertise:

[1]For the researchers who did not create a Google Scholar page, we stored the Google Scholar citation for their ten most cited papers, and use it to impute overall citations.

MBAs have more practical experience, PhD students have more theoretical background. We are interested in examining the role of (i) work experience (MBAs), especially as some of these levers are used in business settings; (ii) knowledge of the setting, captured by the forecasts made by MTurk workers themselves; (iii) importance of direct experience, captured by the forecasts made by mTurkers who actually do a full task for 10 minutes. Regarding this latter case, it is possible both that attempting the task makes people better forecasters, or to the opposite it makes them worse forecasters because it leads them to focus too much on a particular treatment; (iv) impact of having a PhD in psychology versus in economics. The manipulations have both psychological and economic components, we can compare the two groups in this particular setting.

We also compare the accuracy of experts in different areas within the (broadly construed) group of behavioral economists. We code the behavioral theorists, lab experimenters, field behavioral economists, and applied economists with interests in behavioral economics as, to the extent possible, separate groups. We also use the conference participation to separation some of the above, i.e., authors of papers in SITE Experimental Economics versus authors of papers in SITE Psychology and Economics.

For all of these comparisons by group, we can capture them both graphically as well as with a variant of specification (5):

$$d\left(\bar{e}_t, f_{i,t}\right) = \alpha + \gamma G_i + \varepsilon_{i,t}, \tag{7}$$

where G_i are indicators for the different groups. As outlined above, we can also compare the average forecast at the group level, not just the expert-specific forecast.

We then turn to *horizontal expertise*. We define horizontal expertise as expertise of expert i that applies to a particular setting or treatment t. The first comparison is to compare the impact of expertise in behavioral economics. The main comparison here, by group as in specification (7), is comparing the forecasts of PhD students in economics to PhD students in economics in the same institution specializing in behavioral economics. (We envision doing this at UC Berkeley and at the University of Chicago, our home institutions).[2]

In addition, we are interested in examining the impact of expertise of expert i in treatment t. For example, does a researcher with a paper on gift exchange make more precise forecasts on the gift exchange treatment? Hence, we define indicators of match $m_{i,t}$ between expert i and treatment t by: (i) doing searches of highly-cited papers in a particular keyword, such as 'hyperbolic discounting', associated to the treatment; (ii) coding papers in an area using the CV of experts responding to the survey; (iii) using our coding of authors of papers motivating a particular treatment, such as, say, the authors of the original papers on Paying too much or

[2]Ideally, one would also compare non-behavioral economists to behavioral economists, but we did not deem it plausible to get a sizeable enough sample of standard economists respoding to the survey.

not at all; (iv) finally, we consider the treatments on social comparison, task significance, and ranking to be the more psychologically motivated, and thus examine whether experts with a PhD in psychology have higher accuracy on those treatments.

The specification here is

$$d\left(\bar{e}_t, f_{i,t}\right) = \alpha + \beta_H m_{i,t} + \eta_i + \zeta_t + \varepsilon_{i,t}. \tag{8}$$

In specification (8) we estimate with β_H the effect of horizontal expertise defined as match between the expertise of the expert i and the question t. Notice that in this case we can control for a question fixed effect ζ_t and an expert fixed effect η_i, thus holding constant the overall difficulty of forecasting one particular treatment and the overall expertise of expert i. We will estimate specification (8) without such fixed effects.

The hypothesis we consider are the following. The first ($\beta_H > 0$) is that horizontal expertise indeed helps for a forecast, while an alternative ($\beta_H = 0$) is that it does not. It is also possible that there may be a perverse effect of thinking of knowing too much, which would lead to lower accuracy ($\beta_H < 0$). We also examine whether expertise reduces the variance in the forecasts of experts, independent of their accuracy. Finally, we consider this separately for each treatment, since expertise in a particular topic may have differential effects on different behavioral levers (although we do not have a ex ante prediction of how).

We should also return to the point made above about the forecasting of accuracy of predictions for the gain-loss treatment. Assume that there is a sizably larger effect of incentives of the loss side, that is, $\bar{e}_{.4Loss} = 2,000$ while $\bar{e}_{.4Gain} = 1,800$. Thus, the framing of incentives as losses leads to an extra 200 units of effort on average. Consider now an expert i who makes the forecasts $f_{i,.4Loss} = 1,600$ and $f_{i,.4Gain} = 1,400$. By the measure above, this individual has a quite high error, since he/she is 400 units off on each treatment. This individual, however, captured perfectly the difference between the loss and the gain treatment, since $f_{i,.4Loss} - f_{i,.4Gain} = \bar{e}_{.4Loss} - \bar{e}_{.4Gain} = 200$. Notice that this arises because the relevant comparison to understand the effect of loss aversion is not to one of the three known treatments but to another treatment. When analyzing the case of loss aversion, thus, we will use as an alternative measure of distance a diff-in-diff distance. That is, in this case, we will use as distance $d\left(f_{i,.4Loss} - f_{i,.4Gain}, \bar{e}_{.4Loss} - \bar{e}_{.4Gain}\right)$. This issue also applies to the examination of probability weighting and of the response to an increase in the payment for the charity.

Confidence and Beliefs about Experts. Using the results on page 2, we analyze a final set of topics. First, we examine overconfidence by comparing the actual number of forecasts that are within 100 points of the actual effort versus the corresponding forecast of the expert. We then relate this measure of overconfidence to vertical and horizontal expertise. It is conceivable for example that expertise leads to no higher accuracy, but it leads to higher confidence. The opposite is also possible. We also intend to examine whether researchers who worked on overconfidence display less overconfidence themselves.

The second part of page 2 in the survey elicits what the different groups *themselves* believe about horizontal and vertical expertise. Do experts believe that more cited experts are more accurate (vertical expertise)? Do they believe that knowing the context matters (and hence they think that MTurkers will do well)? Do they believe that PhD students with expertise in behavioral economics will do better, compared to PhD students with no such expertise (horizontal expertise)? These questions thus allow us to move from directly examining the impact of expertise to examining the beliefs about expertise. We then relate the actual expertise results to the beliefs to see to what extent the beliefs are accurate.