Forecasting the Results of Experiments: Piloting an Elicitation Strategy

BY STEFANO DELLAVIGNA, NICHOLAS OTIS, AND EVA VIVALT*

In the last decade, economics has increasingly focused on ways to encourage research transparency, such as through pre-registration and pre-analysis plans. These efforts are intended to improve the informativeness and interpretation of research results, but relatively little attention has been paid to another practice that could help to achieve this goal: relating research findings to the views of the scientific community, policy-makers, and the general public by eliciting forecasts of research results. The idea of this practice is to collect and store predictions of research results before the results are known. This makes it possible ex post to relate the findings to prior expectations. Such forecasts can improve the informativeness of research results in three main ways, as discussed in more detail in DellaVigna, Pope, and Vivalt (2019).

First, forecasts can improve the interpretation of research results since they put those results in context and are often of independent interest. For example, in research on the replication of experiments, Camerer et al. (2016) capture the expected probability that a study would replicate. In a behavioral context, DellaVigna and Pope (2018) compare the effects of different behavioral motivators to experts' predictions about which motivators would be most effective. In both cases, the predictions are highly correlated with the actual outcomes; this is important to know, since it implies that researchers' intuition about which studies would replicate, and about behavioral motivators, are on average mostly correct. In a third example, Vivalt and Coville (2017) document that policy-makers overestimate the effectiveness of RCT interventions. These three examples illustrate how predictions can add an extra layer of understanding to the study itself. Importantly,

* DellaVigna: University of California-Berkeley, 515 Evans Hall, Berkeley, CA, 94720 and NBER (e-mail: sdellavi@econ.berkeley.edu). Otis: University of California-Berkeley, 2121 Berkeley Way West, Berkeley, CA, 94720 (e-mail: notis@berkeley.edu). Vivalt: Research School of Economics, Australian National University, Canberra ACT 2601, Australia (e-mail: eva.vivalt@anu.edu.au). We are grateful for financial support from the Alfred P. Sloan Foundation (G-2019-12325), and an anonymous foundation. Vivalt is also supported by the John Mitchell Economics of Poverty Lab at the Australian National University. We thank seminar participants at the ASSA 2020, and especially our discussant, David McKenzie. We also thank the authors of the forecast studies for their support and contributions to the survey and the respondents for their time.

predictions must be collected in advance, to avoid hindsight bias ("We knew it already").

Second, forecasts can mitigate publication bias against null results. Null results are less likely to be published, even when authors have used rigorous methods to answer important questions (Franco et al. 2014). If priors are collected before a study is carried out, the results can be compared to the average expert prediction, rather than to the null hypothesis of no effect.

Third, forecasts may help with experimental design. For example, suppose that a research team could select one of ten different interventions to be evaluated in a randomized controlled trial. Forecasts could be used to gauge which potential treatment arm would have a higher value of information.

With these three motivations in mind, we are developing an online platform researchers can use to collect forecasts of social science research results (www.socialscienceprediction.org). The platform aims to make it easier to elicit forecasts by providing survey templates and making it possible to track forecasts for an individual across different studies. This in turn enables research on the determinants of forecast accuracy. A centralized platform can also help by coordinating requests for forecasts so as to reduce forecaster fatigue.

Before this platform can be a useful tool for the profession, however, important questions must be answered about how to elicit predictions. In particular, we focus on four survey design considerations.

First, prior to eliciting predictions, we may wish to give forecasters an example to ensure that they understand what their responses could mean. To what extent might this example anchor subsequent forecasts? Second, raw units may be more familiar or intuitive to forecasters, but in some contexts only forecasts of standard deviations (SDs) can be elicited, such as for indices. Thus, we would like to understand whether forecasts differ if predictions were gathered using raw units or standard deviations. Third, there is no consensus on whether it is preferable to use slider bars or a text entry response. Compared to slider bars, text entry may avoid anchoring effects, but could increase errors such as typos. Finally, if slider bars are used, does the width of the slider bars affect the predictions?

In this pre-registered pilot, we experimentally test whether these four features affect the predictions of researchers and practitioners (DellaVigna et al., 2020).

I. Forecast Studies

We collected forecasts of the results of three large field experiments preliminarily accepted by the Journal of Development Economics, using their “pre-results review” track, which evaluates research on the basis of rigor, feasibility, and importance (Journal of Development Economics, 2018). The three studies have undergone peer review and their results are unknown, making them excellent targets for prediction.

Study 1. Yang et al. (2019) are running an experiment in Mozambique examining the effects of health and education interventions targeting households with orphaned and vulnerable children on a variety of HIV outcomes. We collected forecasts of the impact of being assigned to receive home visits from a local community worker; these visits were supposed to include referrals for HIV testing, to provide information related to HIV/AIDS, and to involve discussions to reduce concerns about stigma. Our forecast outcome was whether households reported having any member receive HIV testing in the last year.

Study 2. In 2016, Rwanda reformed an entrepreneurship course required for all students in grades 10–12. Blimpo and Pugatch (2019) are examining the effects of a teacher-training program implemented in the same year, which included multiday training sessions, exchange visits across participating schools, and support from trained “Youth Leaders.” We collected forecasts of the impact of this intervention on (1) the percentage of students who dropped out (reverse coded); (2) the percentage of students who reported earning money from a business in the prior month; and (3) standardized entrepreneurship test scores.

Study 3. Bouguen and Dillon (2019) are running a randomized controlled trial evaluating the impact of an unconditional cash, asset, and nutrition transfer program. Randomization took place at the village level, with poor households in treated villages receiving (1) a cash transfer, (2) a combined cash and asset transfer, or (3) a combined cash, asset, and nutrition transfer. We collected forecasts of the impact of these interventions on (1) food consumption and (2) health consumption.

II. Forecast Elicitation

We worked with each of the three project teams to develop a short description of the study, including information on setting, experimental design, and outcomes of interest. Each team reviewed and approved our surveys before we began data collection.

Consenting respondents were randomized to provide predictions for one of the three studies described above. They first read the study description, which included a link to the registered report. We then asked them to forecast the experimental impacts of the treatments. Participants were able to revisit the study description in a new window while providing responses. They were also given the mean and SD of the predicted outcomes from a reference condition to contextualize responses. When a study had more than one outcome, we randomly varied the order in which participants provided their forecasts. After participants completed predictions for one study, they were given the choice to continue and provide predictions for one of the other two studies (of their choosing), or to end the survey. Those predicting the results of a second study were given a similar choice for the third study.

A. Randomized Survey Features

We randomized four features of our forecast elicitation at the individual level. (1) We randomized the reference value (±0.1 or ±0.3 SDs) used in an example just before forecasts were provided. (2) We varied whether responses were given in SDs or in raw units. (3) We randomized whether respondents gave their predictions via a slider scale or simple text entry. For text entries, we bounded responses at 2.0 SDs. (4) Among the sample providing responses on a slider scale, we varied whether the slider was bounded at ±0.5 or ±1.0 SDs.

B. Sample of Forecasters

We sent our forecasting survey to individuals in several research organizations (the Busara Center for Behavioral Economics, GiveWell, the Global Priorities Institute, IDinsight, and the World Bank). We also sent it to the Berkeley development economics Listserv and posted a link to the survey on Twitter. Finally, the authors of the three studies provided a list of 35 total respondents they wanted to send their survey to (for these, the first predicted study was not randomized).

We offered incentives to Listserv and Twitter respondents who completed all three studies. Listserv respondents received a $10 Amazon Gift Card, and Twitter respondents with an academic email address had a 10% chance of receiving a $50 Visa Cash Card. Overall, 106 people responded to our survey, for a total of 772 predictions.

III. Results

We compare forecasts of experimental treatment effects for the three predicted studies across our four experimental elicitation conditions. To compare results across studies and outcomes, we standardize predictions made in real units using the SD of a reference condition.

Table 1 summarizes predictions across the three forecast studies. The average predicted effect size is 0.16 SD, which is comparable to the average treatment effect of 0.12 SD estimated from 635 results from development interventions (Vivalt, forthcoming). Even within a study, forecasters are differentiating across outcomes. For example, the average forecast effect of teacher training on student dropout (reverse coded) is 0.02 SD, compared to a predicted 0.29 effect on entrepreneurship test scores (Panel C).

[Insert Table 1 Here]

We obtain precise estimates of predicted treatment effects. For example, for Yang et al. (2019) (Panel B), with 73 responses the average predicted treatment effect is 0.23 SD, with a confidence interval of [0.19, 0.27]. When the experiment is complete and treatment effects are known, the authors could compare their estimates with these forecasted effects.

We can then consider whether forecasts differ across our four survey elicitation features. As Table 2 shows, three features of elicitation have no impact. First, the reference value used in an example (e.g., 0.1 vs. 0.3 SD) does not affect the results. Second, there is no difference in forecasts elicited in raw units (e.g., percentage of household members tested for HIV) or standard deviations.[1] Third, the average forecast is comparable when using slider bars or text entry.

[Insert Table 2 Here]

This last comparison, however, masks an important dimension of heterogeneity. When the slider has a wider range (±1.0 SD), the elicited forecasts are larger than when the slider has a narrower range (±0.5 SD).

Figure 1 shows that this is not due to censoring at the top in the narrow slider bar condition; only one respondent in this condition provided a prediction of 0.5 SD. In fact, the entire distribution is shifted to the right when wider slider bounds were presented. This may reflect that forecasters are making an inference from the bounds, or that the bounds are anchoring their responses. To the extent that the researcher is interested in comparing forecasts across studies, it is important to use consistent slider ranges.

[Insert Figure 1 Here]

Finally, one may wonder if the forecasts differ by type (faculty, PhD students, or

[1] In the table we translate predictions in raw units into standard deviations to allow for comparison.

researchers) or by recruitment channel (Twitter, the development Listserv, or direct emailing). In Appendix Table A1, we show that forecasts do not vary across these categories.

IV. Conclusion

In this paper we pilot approaches that researchers can use to collect predictions of research results for their own projects. We obtain estimates for the average forecast treatment effect for three development experiments. The average forecast is highly precise with a sample of 106 forecasters, suggesting that for similar projects a sample of 15-30 forecasters should be sufficient. Predictions are robust to most survey elicitation features, with the exception of slider bounds, where wider bounds shift the distribution of predicted treatment effects.

REFERENCES

Bouguen, Adrien, and Andrew Dillon. 2019. “The Impact of a Multidimensional Program on Nutrition and Poverty in Burkina Faso.” Accepted based on pre-results review at the *Journal of Development Economics*, June 20th, 2019.

Blimpo, Moussa and Todd Pugatch. 2018. "Teacher Training and Entrepreneurship Education: Evidence from a Curriculum Reform in Rwanda." Accepted based on pre-results review at the *Journal of Development Economics*, May 5th, 2019.

Camerer, Colin et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351, no. 6280: 1433-1436.

DellaVigna, Stefano, Nicholas Otis and Eva Vivalt. 2020. "Forecasting the Results of Experiments: Piloting an Elicitation Strategy." *AEA RCT Registry.* January 06. https://doi.org/10.1257/rct.5211-1.1.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. “Predict science to improve science.” *Science*, 366(6464), pp.428-429.

DellaVigna, Stefano, and Devin Pope. 2018. "Predicting experimental results: who knows what?" *Journal of Political Economy*, 126(6), pp.2410-2456.

Journal of Development Economics. 2018. “Pre-Results Review (Registered Reports) Guidelines for Authors.” https://www.elsevier.com/__data/promis_misc/JDE_RR_Author_Guidelines.pdf Accessed on 2020-1-5.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science*. 345, no. 6203: 1502-1505.

Yang, Dean et al. “Direct and Spillover Impacts of a Community-Level HIV/AIDS

Program: Evidence from a Randomized Controlled Trial in Mozambique." Accepted based on pre-results review at the *Journal of Development Economics*, July 22, 2019.

Vivalt Eva, and Aidan Coville. 2017. "How Do Policymakers Update?" Unpublished.

Vivalt, Eva. (forthcoming). "How Much Can We Generalize From Impact Evaluations?" *Journal of the European Economics Association*.

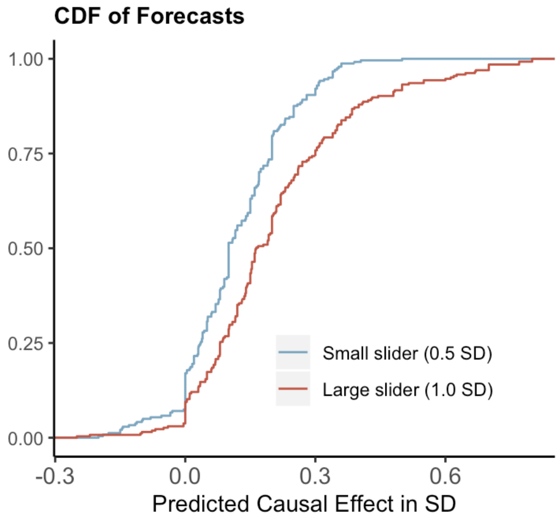

FIGURE 1. FORECASTS BY SMALL VERSUS LARGE SLIDER BOUNDS

Notes: This figure presents CDFs of forecasts from participants assigned to small (0.5 SD) versus large (1.0 SD) slider conditions. Forecasts elicited in raw units are standardized relative to a reference mean.

TABLE 1— FORECASTS BY EXPERIMENT

	Mean (1)	SD (2)	SE (3)	n_i (4)	n_f (5)
Panel A: All pred.	0.16	(0.20)	(0.01)	106	772
Panel B: Yang et al					
HIV testing	0.23	(0.18)	(0.02)	73	73
Panel C: Blimpo et al.					
Dropout (reversed)	0.02	(0.13)	(0.01)	85	85
Business participation	0.12	(0.12)	(0.01)	85	85
Test scores	0.29	(0.34)	(0.04)	85	85
Panel D: Bouguen et al.					
Food consumption					
T1 (Cash)	0.19	(0.12)	(0.01)	74	74
T2 (T1+Asset)	0.20	(0.18)	(0.02)	74	74
T3 (T2+Nutrition)	0.21	(0.21)	(0.02)	74	74
Health consumption					
T1 (Cash)	0.11	(0.09)	(0.01)	74	74
T2 (T1+Asset)	0.14	(0.12)	(0.01)	74	74
T3 (T2+Nutrition)	0.14	(0.16)	(0.02)	74	74

Notes: This table reports summary statistics for forecasts of causal effects from three randomized controlled trials. Columns 1, 2, and 3 present the forecast mean (raw units are standardized), standard deviation, and standard error. In Panel A, standard errors are clustered at the individual level. n_i (col. 4) and n_f (col. 5) are the number of respondents and forecasts per row. Panel A pools forecasts across all studies. Panel B reports forecasts of the impact of a bundled health and education program on self-reported HIV testing. Panel C presents forecasts of the impact of a teacher training program on student dropout (reverse coded), self-reports of earning money from a business in the last month (dichotomous), and scores on an entrepreneurship test. Panel D reports forecasts of the impact of cash, cash and asset, and cash, asset, and nutrition transfers on food and health consumption.

TABLE 2— FORECASTS BY SURVEY FORMAT

	Mean (1)	SD (2)	SE (3)	n_i (4)	n_f (5)	p (6)
Panel A: Reference						
Small (0.1 SD)	0.16	(0.18)	(0.01)	50	393	
Large (0.3 SD)	0.17	(0.21)	(0.02)	56	379	0.53
Panel B: Units						
Raw units	0.16	(0.21)	(0.01)	52	332	
Standard deviations	0.17	(0.18)	(0.02)	54	440	0.75
Panel C: Entry						
Text	0.16	(0.25)	(0.02)	36	266	
Slider	0.17	(0.16)	(0.01)	70	506	0.93
Panel D: Slider bounds						
Small (0.5 SD)	0.12	(0.12)	(0.01)	33	241	
Large (1.0 SD)	0.21	(0.18)	(0.02)	37	265	0.00

Notes: This table reports summary statistics for forecasts of results from three randomized controlled trials by randomly assigned survey format. Columns 1, 2, and 3 present the forecast mean (raw units are standardized), standard deviation, and standard errors (clustered at the individual level). n_i (col. 4) and n_f (col. 5) are the number of respondents and forecasts per row. Column 6 presents clustered p values comparing groups within each panel. Panel A presents forecasts by whether a small (0.1 SD) or large (0.3 SD) reference was used in an example. Panel B presents forecasts made in raw units or standard deviations. Panel C presents forecasts made using text or slider responses. Panel D presents slider responses from small (0.5 SD) or large (1.0 SD) slider bounds.