Title
Forecasting the Results of Experiments: Piloting an Elicitation Strategy

Authors
DellaVigna, Stefano
Otis, Nicholas
Vivalt, Eva

Publication Date
2020-03-05

DOI
10.26085/C3RC7J

Series Name: WPS
Paper No.: 128
Issue Date: 5 Mar 2020

Forecasting the Results of Experiments: Piloting an Elicitation Strategy

Stefano DellaVigna, Nicholas Otis and Eva Vivalt

Forecasting the Results of Experiments: Piloting an Elicitation Strategy
Stefano DellaVigna, Nicholas Otis, and Eva Vivalt
January 2020
JEL No. O1,O17

ABSTRACT

Forecasts of experimental results can clarify the interpretation of research results, mitigate publication bias, and improve experimental designs. We collect forecasts of the results of three Registered Reports preliminarily accepted to the *Journal of Development Economics*, randomly varying four features: (1) small versus large reference values; (2) whether predictions are in raw units or standard deviations; (3) text-entry versus slider responses; and (4) small versus large slider bounds. Forecasts are generally robust to elicitation features, though wider slider bounds are associated with higher forecasts throughout the forecast distribution. We make preliminary recommendations on how many forecasts should be gathered.

Stefano DellaVigna
University of California, Berkeley
Department of Economics
549 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
sdellavi@econ.berkeley.edu

Nicholas Otis
University of California, Berkeley
2121 Berkeley Way West
Berkeley, CA 94720
notis@berkeley.edu

Eva Vivalt
Research School of Economics
Australian National University
Acton ACT 2601
Australia
eva.vivalt@anu.edu.au

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, 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).

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. (In the table we translate predictions in raw units into standard deviations to allow for comparison.) Third, the average forecast is comparable when using slider bars or text entry.

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.

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

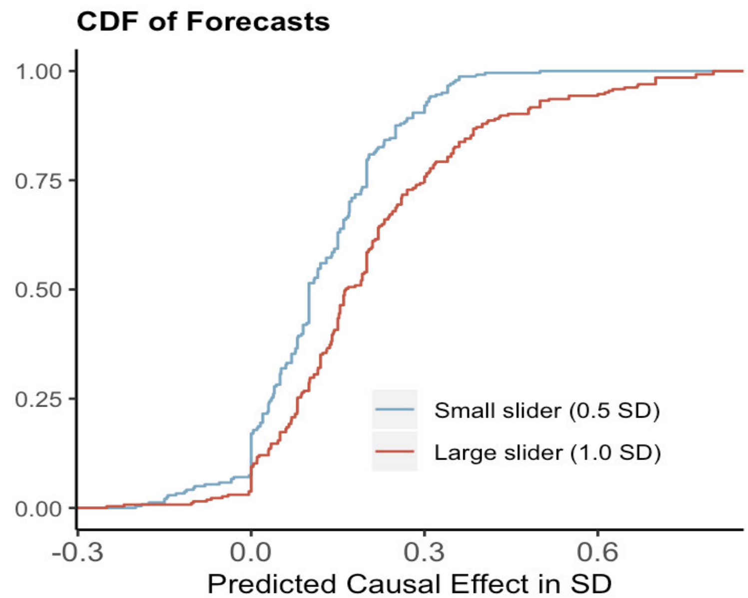


Figure 1. Forecasts by Small Versus Large Slider Bounds

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.

ONLINE APPENDIX

Forecasting the Results of Experiments:

Piloting an Elicitation Strategy

Stefano DellaVigna — UC Berkeley and NBER
Nicholas Otis — UC Berkeley
Eva Vivalt — Australian National University

Contents

1 Additional Tables

Table A1: Forecasts by Sample and Type

	Mean (1)	SD (2)	SE (3)	n_i (4)	n_f (5)	p (6)
Panel A: Sample						
Listserv	0.15	(0.16)	(0.01)	39	336	
Twitter	0.19	(0.22)	(0.02)	39	271	0.16
Other	0.15	(0.21)	(0.02)	28	165	0.89
Panel B: Type						
Faculty	0.15	(0.23)	(0.02)	33	193	
PhD student	0.17	(0.17)	(0.01)	40	331	0.48
Researcher	0.17	(0.18)	(0.02)	27	222	0.59
Practitioner	0.13	(0.22)	(0.02)	6	26	0.45

Notes: This table reports summary statistics for forecasts of results from three randomized controlled trials by study sample (Panel A) or self-reported type (Panel B). In Panel A, "Other" includes respondents from the Busara Center for Behavioral Economics, GiveWell, the Global Priorities Institute, IDinsight, and the World Bank. Columns 1, 2, and 3 present the forecast means (raw units are standardized relative to a reference mean), standard deviations, 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. This analysis was not pre-registered.

Table A2: Forecasts by Experiment and Survey Format

	Reference		Units		Entry		Slider bounds	
	Small	Large	Raw units	Std. dev.	Text	Slider	Small	Large
Panel A: Yang et al.								
HIV testing	0.23	0.22	0.22	0.24	0.25	0.22	0.20	0.23
	(0.18)	(0.18)	(0.13)	(0.22)	(0.26)	(0.13)	(0.11)	(0.15)
Panel B: Blimpo et al.								
Dropout (reversed)	-0.01	0.06	0.04	0.00	0.01	0.03	0.01	0.06
	(0.10)	(0.15)	(0.11)	(0.15)	(0.07)	(0.16)	(0.12)	(0.18)
Business participation	0.11	0.14	0.13	0.12	0.09	0.14	0.11	0.17
	(0.11)	(0.13)	(0.11)	(0.13)	(0.10)	(0.13)	(0.10)	(0.15)
Test scores	0.24	0.34	0.35	0.24	0.38	0.25	0.18	0.32
	(0.21)	(0.42)	(0.43)	(0.22)	(0.53)	(0.2)	(0.13)	(0.23)
Panel C: Bouguen et al.								
Food consumption								
T1 (Cash)	0.19	0.19	0.18	0.19	0.19	0.19	0.15	0.22
	(0.13)	(0.11)	(0.10)	(0.13)	(0.15)	(0.10)	(0.07)	(0.11)
T2 (T1+Asset)	0.21	0.20	0.16	0.24	0.23	0.19	0.12	0.26
	(0.2)	(0.16)	(0.16)	(0.18)	(0.2)	(0.16)	(0.12)	(0.17)
T3 (T2+Nutrition)	0.22	0.20	0.17	0.24	0.21	0.22	0.15	0.28
	(0.26)	(0.16)	(0.14)	(0.25)	(0.29)	(0.16)	(0.11)	(0.18)
Health consumption								
T1 (Cash)	0.11	0.10	0.11	0.11	0.09	0.12	0.08	0.15
	(0.11)	(0.07)	(0.13)	(0.07)	(0.07)	(0.10)	(0.08)	(0.12)
T2 (T1+Asset)	0.15	0.12	0.11	0.15	0.11	0.15	0.11	0.19
	(0.14)	(0.09)	(0.14)	(0.10)	(0.09)	(0.13)	(0.10)	(0.15)
T3 (T2+Nutrition)	0.15	0.14	0.12	0.16	0.10	0.17	0.10	0.23
	(0.17)	(0.16)	(0.19)	(0.14)	(0.10)	(0.19)	(0.12)	(0.21)
n_i	50	56	52	54	36	70	33	37
n_f	393	379	332	440	266	506	241	265

Notes: This table reports summary statistics for predictions of results of three randomized controlled trials by randomly assigned elicitation strategy. Predictions are of causal treatment effects standardized relative to a reference mean for raw-unit elicitations. Standard deviations are presented in parentheses. Panel A reports forecasts of the impact of a bundled health and education intervention on self-reported HIV testing. Panel B presents forecasts of the impact of a teacher training intervention 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 C reports forecasts of the impact of cash, cash and asset, and cash, asset, and nutrition transfers on food and health consumption. n_i and n_f are the number of individuals making forecasts and the total forecasts for each column. Columns 1 and 2 present forecasts by whether a small (0.1 SD) or large (0.3 SD) reference was used in an example. Columns 3 and 4 present forecasts made in raw units or standard deviations. Columns 5 and 6 present forecasts made using text or slider responses. Columns 7 and 8 present slider responses from small (0.5 SD) or large (1.0 SD) slider bounds.

2 Survey Instruments

This section contains the entire forecasting survey for one randomization. A direct comparison of the different randomizations (for the HIV testing outcome) can be found at the end of the survey.

The *Journal of Development Economics* recently became the first economics journal to review and approve projects for publication before the results are known. These articles are evaluated based on the importance of the research questions and the quality of the research design (such as having sufficient statistical power). Building on this work--which emphasizes open and transparent science--we are collecting predictions of empirical findings from the three Registered Reports that were preliminarily accepted before October 2019 that publicly posted their proposals and have not yet publicly released results:

- Direct and Spillover Impacts of a Community-Level HIV/AIDS Program: Evidence from a Randomized Controlled Trial in Mozambique
  Authors: Dean Yang, Arlete Mahumane, James Riddell, Hang Yu
- The Impact of a Multidimensional Program on Nutrition and Poverty in Burkina Faso
  Authors: Adrien Bouguen, Andrew Dillon
- Entrepreneurship Education and Teacher Training in Rwanda
  Authors: Todd Pugatch, Moussa Blimpo

We ask you to predict the results of one of these three studies, after which you can either end the survey or provide predictions for the additional projects. Predictions for each project should take approximately 10 minutes to complete. The survey is approved by UC Berkeley IRB 2019-10-12690, and Australian National University IRB 2019/836. The principal investigators on this project are Stefano DellaVigna (sdellavi@econ.berkeley.edu), Eva Vivalt (eva.vivalt@anu.edu.au), and Nicholas Otis (notis@berkeley.edu).

Survey details can be found in the box below:

Researchers: The primary investigators in this study are Stefano DellaVigna, an academic staff member in the Department of Economics at the University of California, Berkeley, Eva Vivalt, an academic staff member at the Research School of Economics within the College of Business and Economics at the Australian National University, and Nicholas Otis, a graduate student at the University of California, Berkeley.

Project Title: Social Science Prediction Survey.

You may advance to the survey by selecting "I consent" below.

- (●) I consent
- ( ) I do not consent

→

Which of the following best describes you?

- ( ) Faculty
- ( ) PhD student
- ( ) Researcher
- ( ) Practitioner

← →

Which of the following best describes your degree?

- ◯ Economics
- ◯ Public Policy
- ◯ Political Science
- ◯ Psychology
- ◯ Other

How familiar are you with the following types of interventions in developing countries?

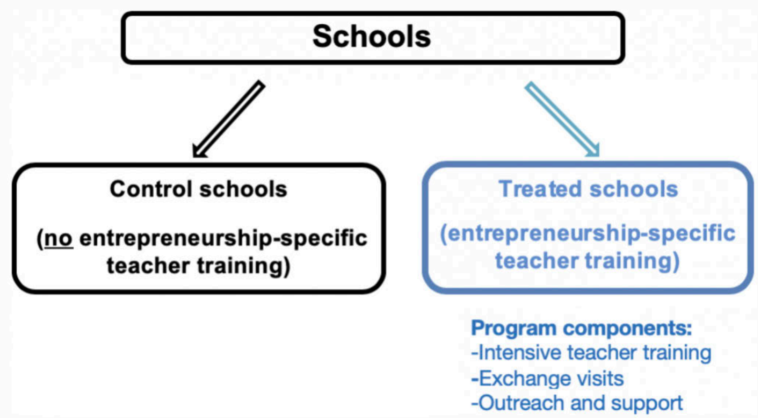
	Not at all familiar	Slightly familiar (Haven't heard of any specific studies)	Familiar (Have heard of some studies)	Very familiar (Know the details of several studies)
Unconditional cash transfers	◯	◯	◯	◯
Education / teacher training interventions	◯	◯	◯	◯
HIV / public health interventions	◯	◯	◯	◯

← →

Study: Entrepreneurship education and teacher training in Rwanda
Authors: Todd Pugatch, Moussa P. Blimpo

Background on education in Rwanda: Primary school (grades 1-6) in Rwanda is compulsory. All Rwandan secondary students are required to enroll in entrepreneurship courses through six years of secondary school (S1-S6, equivalent to grades 7-12). In 2016, Rwanda reformed its required upper secondary (S4-S6, equivalent to grades 10-12) entrepreneurship course by introducing interactive pedagogy and a focus on business skills, which covers the full cycle of business creation and development, including product development, registration and legal issues, marketing, accounting, and customer relations.

Intervention: In that year, a subset of schools was randomly selected for two years of intensive teacher training and support (treated schools). The program covered more than 100 schools, 260 teachers, and 6,800 students, and was implemented by the government and a large international NGO. A control group of equal size received the curriculum and standard government training only. The training received by treated teachers was subject-specific (entrepreneurship), incorporated peer feedback meetings, and included follow-up support.



The training had three main components:
-*Intensive teacher training*: Entrepreneurship teachers received multi-day training sessions each academic term beginning April 2016 through January 2018. Each of the six sessions was held during holidays between terms and lasted four days. Training emphasized lesson planning, engaging students in classroom discussions, encouraging students to create entrepreneurship "portfolios" of their work, and assisting student business clubs to form and grow. Trainings culminated in a "mock day" in which teachers rehearsed upcoming lessons.
-*Exchange visits*: Teachers participating in the intervention visited each other's schools to learn from and provide feedback to their peers.
-*Outreach and support*: Teachers received ongoing outreach to support their implementation of the curriculum, including visits from trained "Youth Leaders" which contained product-making demonstrations (e.g., for household goods such as soap or candles) co-taught with the teacher, advising of student business clubs, classroom observation, participating in teacher exchange visits, and addressing any other concerns. Student business clubs were encouraged to submit their ideas to regular business competitions held for treated schools.

Target population: The study focused on the cohort entering S4 (10th grade) in 2016, with training provided to this cohort's entrepreneurship teacher as they progressed to S6 (12th grade). The control group and the treated group received the new entrepreneurship curriculum. Teachers in control schools did not receive the intensive training, exchange visits, or outreach provided to treatment schools.

Outcomes overview: Outcomes were measured at the student level. Approximately 15 students were sampled from each school. We ask you to predict the experimental results for three outcomes: scores on a standardized entrepreneurship test, whether students dropped out of school, and business participation.

← →

We are interested in what you think the impact of these treatments will be. For each outcome, please provide your prediction. Even if you do not have strong beliefs about the effects of the interventions, we are still interested in your best guess. As a reference, we will provide the mean value of the control group at endline.

← →

One outcome we are interested in is the percent of respondents who dropped out of school. This outcome measures dropout at any time after baseline (April 2016) through when the endline surveys were completed June-October 2018. The final training was in January 2018, with final exchange visits and outreach in April.

Please predict the difference in the percent of respondents who dropped out of school between the group in which teachers received entrepreneurship-specific training and the control group (the average treatment effect).

Notes:

- Click here for a reminder of the intervention and study background, which will open in a new window.
- **Reference**: In the control group, an average of 9% of respondents dropped out over the duration of the study (with a standard deviation of 29 percentage points).
- As an example, if you enter **8.7** it means you think student dropout will be **8.7** percentage points **higher** in the treatment group. If you enter **-8.7** it means you think student dropout will be **8.7** percentage points **lower** in the treatment group. If you enter **0** it means you think the program had **no impact**.

Difference in student dropout between treatment and control condition (percentage points)

→

One outcome we are interested in is scores on an entrepreneurship test. Students took the test in November 2018. The final training was in January 2018, with final exchange visits and outreach in April.

Please predict the difference in entrepreneurship test scores between the group in which teachers received entrepreneurship-specific training and the control group (the average treatment effect).

Notes:

- Click here for a reminder of the intervention and study background, which will open in a new window.
- **Reference**: The mean entrepreneurship test score at endline for the control group is 2.20 (with a standard deviation of 1.47), on 1-6 scale, where 6 is the highest number of points that can be scored.
- As an example, if you enter **0.45** it means you think test scores will be **0.45** test points **higher** in the treatment group. If you enter **-0.45** it means you think test scores will be **0.45** test points **lower** in the treatment group. If you enter **0** it means you think the program had **no impact**.

Difference in exam score between treatment and control condition (test points)

← →

One outcome we are interested in is business participation. The final training was in January 2018, with final exchange visits and outreach in April, and at endline (June-October 2018) respondents were asked if they earned money from running a business (in the last month).

Please predict the difference in the percent of students who reported earning money from running a business in the last month between the group in which teachers received entrepreneurship-specific training and the control group (the average treatment effect).

Notes:

- Click here for a reminder of the intervention and study background, which will open in a new window.
- **Reference**: In the control group, an average of 30% of respondents reported earning money from running a business in the last month (with a standard deviation of 46 percentage points).
- As an example, if you enter **13.8** it means you think the number of respondents who report earning money from running a business will be **13.8** percentage points **higher** in the treatment group. If you enter **-13.8** it means you think the number of students respondents who report earning money from running a business will be **13.8** percentage points **lower** in the treatment group. If you enter **0** it means you think the program had **no impact**.

Difference in business participation between treatment and control condition (percentage points)

← →

How confident are you in your predictions for this study? If you are confident it means that you believe your predictions are very accurate.

- ◯ Not at all confident
- ◯ Not very confident
- ◯ Somewhat confident
- ◯ Very confident

If you have any comments, please enter them below. We would love to hear your feedback.

→

Thank you for your predictions! You can now choose to end the survey, or you can continue and provide predictions for an additional study.

- ◯ Continue to the project *Direct and Spillover Impacts of a Community-Level HIV/AIDS Program: Evidence from a Randomized Controlled Trial in Mozambique.*
- ◯ Continue to the project *The Impact of a Multidimensional Program on Nutrition and Poverty in Burkina Faso.*
- ◯ End the survey

← →

Study: The Impact of a Multidimensional Program on Nutrition and Poverty in Burkina Faso
Authors: Adrien Bouguen and Andrew Dillon

Study introduction: A large number of people in Burkina Faso live in poverty. This problem is compounded by the Sahel environment, which leaves households vulnerable to food insecurity. This impact evaluation examines the effects of a multi-faceted anti-poverty program in rural Burkina Faso implemented by two local partner nonprofits.

Target population: The program targets ultra-poor and poor households with a child under the age of five, and/or a pregnant/breastfeeding woman. Ultra-poor and poor households were identified before randomization using quantitative and qualitative targeting methods. On average, 21 households per village were selected to be eligible for the program.

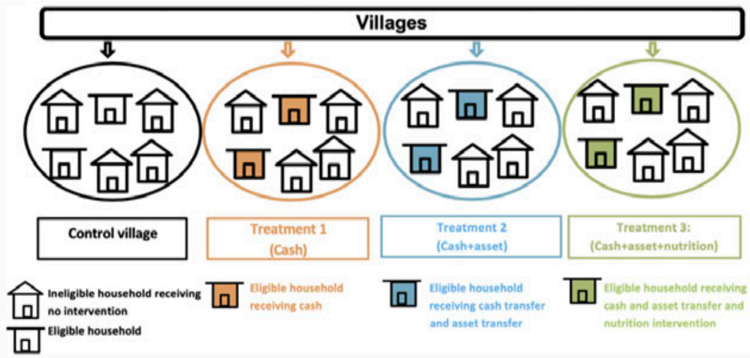
Randomization: This study is a cluster randomized controlled trial. Randomization took place at the village level: 168 villages were randomized into one of three treatment conditions or a control condition. Note that treatments varied slightly based on region. The main components of these three treatments are described below:

Treatment 1: Unconditional cash transfer. The cash transfer lasted a total of two years. In the first year, households received 20,000 FCFA per month for four months (~40% of monthly household consumption; ~100 USD, PPP adjusted), for a total annual transfer of 80,000 FCFA (~400 USD, PPP adjusted). In the second year they received 15,000 FCFA per month for four months (~30% of monthly household consumption; ~75 USD, PPP adjusted), for a total annual transfer of 60,000 FCFA (~300 USD, PPP adjusted). Cash was distributed during the lean season between planting and harvesting (June-September).

Treatment 2: Unconditional cash and asset transfer. The cash transfer is as above. These households also received a coupon for animals which can be exchanged at markets. The coupons varied by area in two ways. First, whether a household receives a coupon for sheep or poultry varied based on the suitability of the village for raising each animal, as determined by a local implementing partner. Second, the amount of the voucher varied by region: poultry coupons were worth either 25,000 (~125 USD, PPP adjusted) or 44,000 FCFA (~220 USD, PPP adjusted), and sheep coupons were worth either 90,000 (~450 USD, PPP adjusted) or 114,000 FCFA (~570 USD, PPP adjusted).

Treatment 3: Unconditional cash and asset transfer and nutrition intervention. The cash and asset transfers are as above. These households also received nutritionally fortified flour for children aged 6 to 23 months. Pregnant or lactating women received bread flour on a monthly basis.

The intervention contents are summarized in the figure and table below:



Type	Group/Period	Region 1	Region 2
Unconditional cash	First year	20,000 FCFA (~100 USD) per month for 4 months (~400 USD **total**)	20,000 FCFA (~100 USD) per month for 4 months (~400 USD **total**)
	Second year	15,000 FCFA (~75 USD) per month for 4 months (~300 USD **total**)	15,000 FCFA (~75 USD) per month for 4 months (~300 USD **total**)
Asset	Poultry	44,000 FCFA (~220 USD **total**)	25,000 FCFA (~125 USD **total**)
	Sheep	114,000 FCFA (~570 USD **total**)	90,000 FCFA (~450 USD **total**)
Nutrition	Children (6-23 months)	2.5 kg per child per month for 4 months (10 kg **total**)	2.5 kg per child per month for 3 months (7.5 kg **total**)
	Pregnant / breastfeeding women	0.7 kg per month for 4 months (2.8 kg **total**)	2.1 kg per month for 3 months (6.3 kg **total**)

Note: All USD are PPP adjusted.

Timing: The second wave of the intervention was delivered in November, 2019. Endline is estimated to take place in April-May, 2020.

Treatment delivery: The table below depicts the proportion of households in each group that had received an intervention at midline (households should have received treatment by this point).

- For the **bold values**, treatment should be 100% if all intended recipients were treated. For example, only **75.3%** of households in T1 (cash) had received some cash at midline.
- Red text indicates households that should not have received a treatment. For example, 1.8% of households in the control group had received some cash at midline.

	Experimental group			
Component	Control	T1 (cash)	T2 (cash+asset)	T3 (cash+asset+nutrition)
Cash	1.8%	**75.3%**	**81.7%**	**83.5%**
Asset	<1%	<1%	**62.0%**	**76.3%**
Nutrition	<1%	<1%	<1%	**53.5%**

→

We are interested in what you think the impact of these treatments will be. For each outcome, please provide your prediction. Even if you do not have strong beliefs about the effects of the interventions, we are still interested in your best guess. As a reference, we will provide the mean value of the control group at midline.

← →

One outcome we are interested in is the average household health expenditure. This includes expenses related to outpatient visits, hospitalization, medical transportation costs, insurance fees, and all other medical expenses.

Please predict the difference in monthly health expenditure between households assigned to each of the three treatment groups and the control group.

Notes:

- Click here for a reminder of the interventions and study background, which will open in a new window.
- **Reference**: At midline, the average monthly household health consumption expenditure in the control group was about 13 USD (PPP adjusted; 2,600 FCFA) with a standard deviation of about 25 USD (PPP adjusted; 5,000 FCFA).
- As an example, if you enter **7.5** it means you think average monthly household health consumption expenditure will be **7.5 USD higher** in the treatment group. If you enter **-7.5** it means you think average monthly household health consumption expenditure will be **7.5 USD lower** in the treatment group. If you enter **0** it means you think the program had **no impact**.

	Difference: Health consumption expenditure (USD)
Difference between (T1) unconditional cash transfer and control condition	
Difference between (T2) unconditional cash + asset transfer and control condition	
Difference between (T3) unconditional cash + asset transfer + nutrition intervention and control condition	

→

One outcome we are interested in is monthly household food consumption expenditure. This includes purchased food, home produced food, food received from other household members, friends and in the form of in-kind payments.

Please predict the difference in monthly food expenditure between households assigned to each of the three treatment groups and the control group.

Notes:

- Click here for a reminder of the interventions and study background, which will open in a new window.
- **Reference**: At midline, the average monthly household food consumption expenditure in the control group was about 200 USD (PPP adjusted; 40,000 FCFA) with a standard deviation of about 130 USD (PPP adjusted; 26,000 FCFA).
- As an example, if you enter **39** it means you think average monthly household food consumption expenditure will be **39 USD higher** in the treatment group. If you enter **-39** it means you think average monthly household food consumption expenditure will be **39 USD lower** in the treatment group. If you enter **0** it means you think the program had **no impact**.

	Difference: Food expenditure (USD)
Difference between (T1) unconditional cash transfer and control condition	
Difference between (T2) unconditional cash + asset transfer and control condition	
Difference between (T3) unconditional cash + asset transfer + nutrition intervention and control condition	

← →

How confident are you in your predictions for this study? If you are confident it means that you believe your predictions are very accurate.

- ◯ Not at all confident
- ◯ Not very confident
- ◯ Somewhat confident
- ◯ Very confident

If you have any comments, please enter them below. We would love to hear your feedback.

→

Thank you for your predictions! You can now choose to end the survey, or you can continue and provide predictions for an additional study.

- ◯ Continue to the project *Direct and Spillover Impacts of a Community-Level HIV/AIDS Program: Evidence from a Randomized Controlled Trial in Mozambique.*
- ◯ End the survey

← →

Study: Direct and Spillover Impacts of a Community-Level HIV/AIDS Program: Evidence from a Randomized Controlled Trial in Mozambique

Authors: Dean Yang, Arlete Mahumane, James Riddell, Hang Yu

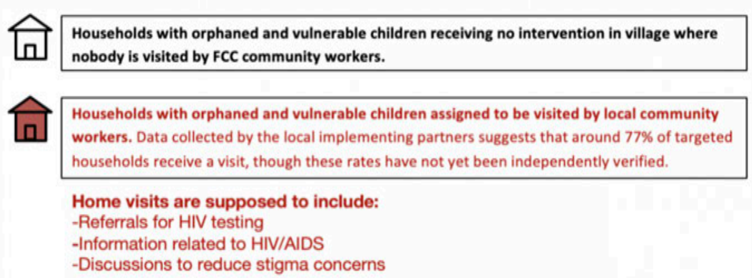
Introduction: Mozambique has high levels of HIV (7.1% of the population). This study examines the impacts of Força à Comunidade e Crianças (FCC, "Strengthening Communities and Children"), a U.S. government-funded program targeting households with orphaned and vulnerable children which is designed to combat HIV/AIDS. The focus of this study is to understand how home visits by local community workers (hired by a local implementing partner of the FCC program) impact HIV-related outcomes in these households.

Home visits: Local community workers conduct home visits to identify households with orphans and vulnerable children, which are then linked to appropriate programs and services in communities, schools, and health facilities. Local community workers are 80% female and usually between 25 and 40 years old.

Home visits and HIV testing: A key component of the home visits is referrals for HIV testing at the nearest affiliated health clinics. All FCC beneficiaries (both adults and children of all ages) who do not know their HIV status (or were negative and have not been tested in the last 12 months) are supposed to be referred by the community workers to HIV testing services. Those testing positive for HIV are referred to receive antiretroviral therapy (ART) through a nearby affiliated clinic. Local community workers follow up with individuals initiating ART to promote ART adherence on an ongoing basis. During these home visits, the local community workers try to increase HIV testing rates through:

- **Information related to HIV/AIDS**: FCC beneficiaries receive information on HIV/AIDS, such as on methods of disease transmission, progression of the disease, treatment, HIV testing, and locations of health clinics providing testing.
- **Discussions to reduce stigma concerns**: FCC beneficiaries also engage in discussions to reduce stigmatizing attitudes among program beneficiaries. Community workers are expected to provide psychosocial support, gradually gaining program beneficiaries' trust over repeated interactions.
- **Education**: In home visits, community workers are also expected to give caregivers advice and encouragement regarding children's education.
- **Other components**: Households are connected to other relevant services after the home visits, based on needs assessments conducted by the local community workers. These other services are expected to reach only a relatively small fraction of those reached by home visits. More information on these subcomponents can be found in this document (pages 5-6).

Experimental design
Randomization took place at multiple levels. In this survey, we focus on the two types of households depicted in the figure below:



At baseline, the average household contained 5.9 members. The experiment contains several other levels of randomization, including an individual level of randomization that took place in all villages. For simplicity, this survey focuses on those households that were not assigned to receive any intervention beyond the FCC program. For more information, see this document (pages 7-9)

Study sites: Communities were selected on the basis of being close to health clinics offering HIV testing and treatment, having sufficient populations of orphans and vulnerable children, and having no other active donor-funded HIV/AIDS programs.

Timing: The FCC program began activities in early 2017. Over the calendar year they gradually enrolled beneficiaries and scaled up program activities. The follow-up survey began in May 2019, and was scheduled to be completed near the end of 2019. Households in treated communities can therefore have had up to two years of exposure to the FCC program at the time of the follow-up survey, but some households may have had a few months' less program exposure, if they happened to have been enrolled in the program towards the end of 2017.

→

We are interested to hear your predictions about the effects of this intervention on one outcome: self-reported HIV testing. Even if you do not have strong beliefs about the effects of the intervention on this outcome, we are still interested in your best guess.

Self-reported HIV testing was measured in the endline survey. Respondents were asked if anyone in the household had been tested for HIV in the last 12 months. The outcome is a household-level variable equal to 1 if at least one household member is reported to have had an HIV test in the last 12 months, and 0 otherwise.

Please predict the difference in self-reported HIV testing (in percentage points) between households assigned to be visited by local community workers and the control group.

Notes:

- Click here for a reminder of the interventions and study background, which will open in a new window.
- **Timing**: Households in treated communities can have had up to two years of exposure to the FCC program at the time of the follow-up survey, but some households may have had a few months' less program exposure, depending on enrollment date.
- **Reference**: As a reference, at baseline about 41.9% (with a standard deviation of 49.4 percentage points) of households self-reported having any household member receive HIV testing in the last 12 months.
- As an example, if you enter **14.7** it means that you think self-reported HIV testing will be **14.7** percentage points **higher** in the group assigned to be visited by local community workers. If you enter **-14.7** it means that you think self-reported HIV testing will be **-14.7** percentage points **lower** in the group assigned to be visited by local community workers. If you enter **0** it means you think the treatment had **no impact**.

Difference between treatment and control condition in percentage points

→

How confident are you in your predictions for this study? If you are confident it means that you believe your predictions are very accurate.

- ◯ Not at all confident
- ◯ Not very confident
- ◯ Somewhat confident
- ◯ Very confident

If you have any comments, please enter them below. We would love to hear your feedback.

→

We thank you for your time spent taking this survey.
Your response has been recorded.

3 Treatment Comparison

0.1 standard deviation reference

As an example, if you enter **0.10** it means that you think self-reported HIV testing will be **0.10** standard deviations **higher** in the group assigned to be visited by local community workers. If you enter **-0.10** it means that you think self-reported HIV testing will be **-0.10** standard deviations **lower** in the group assigned to be visited by local community workers. If you enter **0** it means you think the treatment had **no impact**.

0.3 standard deviation reference

As an example, if you enter **0.30** it means that you think self-reported HIV testing will be **0.30** standard deviations **higher** in the group assigned to be visited by local community workers. If you enter **-0.30** it means that you think self-reported HIV testing will be **-0.30** standard deviations **lower** in the group assigned to be visited by local community workers. If you enter **0** it means you think the treatment had **no impact**.

Standard deviations (text entry)

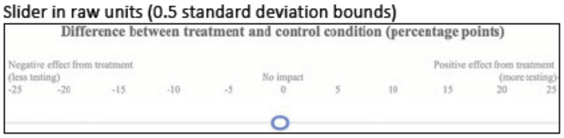
Difference between treatment and control condition in standard deviations

Raw units (text entry)

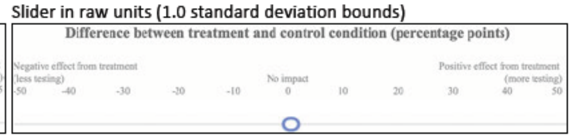
Difference between treatment and control condition in percentage points

Respondents providing forecasts in standard deviations are also provided with the following statement: ***As a reference, a recent survey of many impact evaluations in development economics suggests that the average effect size is around 0.10 standard deviations (Vivalt, 2019).***

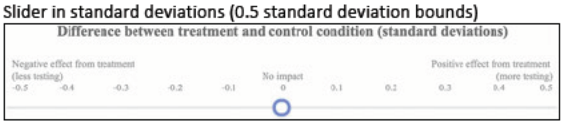
Slider in raw units (0.5 standard deviation bounds)



Slider in raw units (1.0 standard deviation bounds)



Slider in standard deviations (0.5 standard deviation bounds)



Slider in standard deviations (1 0 standard deviation bounds)