

Forecasting long-run causal effects

David Rhys Bernard*
Paris School of Economics
david.rhys.bernard@gmail.com

August 25, 2023

Abstract

I investigate the accuracy of forecasts for short and long-run treatment effects of seven randomized experiments, collecting over 25,000 forecasts from 1,400 respondents, including academics, expert forecasters and nonexperts. I explore the accuracy of different types of forecasters, forecasts over different time horizons, the wisdom-of-crowds effect and gains from aggregation, and forecast calibration. I explore mechanisms driving forecast accuracy by looking randomising additional information on forecasting, and the contexts and interventions under study, as well as looking at how vertical and horizontal expertise, familiarity and effort are correlated with accuracy. I study the process by which forecasters update their long-run forecasts after receiving short-run information. I show that expert forecasters outperform academics and provide evidence that better forecast calibration is what drives this. I show that measures of horizontal expertise and experience with the context under study are more correlated with accuracy than vertical expertise. I show that forecasters strongly overestimate the strength of the relationship between short and long-run outcomes. Finally, I show that although forecasters act like Bayesian updaters when considering the uncertainty around their prior long-run forecasts, they do not appropriately consider the uncertainty around the short-run information signal they receive.

1 Introduction

Forecasting plays an important role in various aspects of decision-making processes, including across the design of experiments, the provision of policy advice, and the evaluation of research findings (DellaVigna and Pope, 2018; DellaVigna et al., 2019). These forecasts, whether implicit or explicit, can serve as valuable inputs for policymakers, researchers, and practitioners, helping them navigate uncertain environments and make informed decisions. When formulating policies or developing interventions, the ability to anticipate potential effects for which there is little empirical evidence

*I thank Nina Ruer for excellent research assistance on this project and the researchers who shared their long-term RCT results with me. I also acknowledge support from funding by a French government subsidy managed by the Agence Nationale de la Recherche under the framework of the Investissements d'avenir programme reference ANR-17-EURE-001. All mistakes remain my own.

either way, is essential for weighing the costs and benefits of various options as well as for prioritizing resources.

In contexts where information about the impacts of policies or interventions is limited, such as when considering long-run effects, forecasts become relatively more valuable. The long-run implications of policies often remain uncertain due to the difficulty of collecting long-run data and the increasing noise in results over extended periods. In these situations, decision-makers can rely on forecasts as a potential alternative means of estimating potential impacts. Accurate long-run forecasts can help identify promising interventions and effective policies.

I collect incentivized forecasts of the long-run results of 7 different randomized experiments before the long-run results were publicly available and assess their accuracy. I collect forecasts of 93 different causal effects from these 7 studies. I collect over 25,000 forecasts from around 1,400 different respondents, including academics (domain experts), expert forecasters, and nonexperts recruited from Prolific. I use this dataset to assess forecast accuracy across different groups and time horizons, the correlates of accuracy and how forecasters update.

For each experiment I run a separate survey, all of which follow the same structure. First, I give a basic description of the context and the study to all respondents. Next, I ask respondents to provide forecasts of the short and long-run results of the experiment. In particular, for each causal effect, I elicit three forecasts, a most likely case, a best case and a worst case. Finally, I tell participants what the short-run results were and ask them to update their long-run forecasts.

Additionally, I experimentally vary the information set provided to the respondents while describing the study. I cross randomize (1) tips on how to forecast better versus a control of no forecasting tips, and (2) additional information on the context of the experiment versus additional information on the results of studies of similar interventions in different contexts versus a control of no additional information.

This study design allows me to look at six questions. First, who is better at forecasting causal effects, domain experts with the advantage of contextual knowledge or forecasting experts with the advantage of forecasting experience? I find that while both groups outperform non-experts, expert forecasters exhibit a higher degree of accuracy than academics across all time horizons. I show evidence that this is likely due to expert forecasters being better calibrated. Their superior performance is clear on the log score accuracy measure that takes into account the full forecast distribution, but the two groups are statistically equivalent when it comes to identifying the most probable point estimate within this range.

Second, how well can we forecast treatment effects over different time horizons? How does accuracy change with time horizon length? I explore this by examining forecasts across short-run and prior and posterior long-run effects. Interestingly, prior long-run forecasts are as accurate, if not more so, than short-run forecasts. I also observe an increase in overoptimism regarding the benefits of treatment in long-run forecasts compared to short-run, a pattern that intensifies when

respondents are updated about the short-run effects. Nonetheless, the posterior long-run forecasts, made with knowledge of the short-run effects, are more accurate than the prior long-run forecasts, suggesting that short-run information provides valuable insight into long-term impacts.

Third, how much more accurate are groups of forecasters than individual forecasters? Is there a strong wisdom-of-crowds effect? I compare the accuracy of the aggregate median forecast of each group to the average accuracy of forecasters within that group. The data reveals a consistent pattern where the average score and negative absolute error of individual forecasters are worse than those of the aggregated median forecast across all participant groups. This indicates a clear wisdom-of-crowds effect. Nonetheless, despite this improvement, even the best aggregate forecasts still have significant room for improvement in terms of accuracy, often failing to outperform basic benchmarks.

Fourth, what information matters for making accurate long-term forecasts? For this, I exploit the randomised provision of information while describing the study to respondents. The results indicate no significant effect on the accuracy of forecasts from any type of randomised information. One possible explanation could be the bounded rationality of forecasters, suggesting that the complexity of forecasting causal effects and the volume of information provided might have exceeded the cognitive bandwidth of the forecasters, thereby failing to enhance accuracy. This aligns with the "less is more" paradigm, suggesting that providing a curated subset of crucial information might be more effective than overwhelming forecasters with extensive data.

Fifth, what correlates with forecast accuracy within and across different types of forecasters? I look at the relationship between accuracy and different measures of vertical and horizontal expertise, familiarity with the context and intervention under study, and effort. Among academics, horizontal expertise, notably experience with the intervention and field, significantly influenced accuracy. Among nonexperts, context familiarity significantly improved accuracy. Effort, measured by time spent on the task, had a substantial influence on accuracy, although its effect reduced when accounting for the type of forecaster, suggesting that effort plays a role in the accuracy disparity between experts and nonexperts.

Finally, how do forecasts update their long-run forecasts? There's a strong correlation between short-run and long-run treatment effects in practice. However, forecasters expect this correlation to be twice as strong as it actually is, overestimating the persistence of effects from short to long-run. After receiving short-run information, forecasters tend to significantly reduce their posterior forecasts, and give more weight to the short-run results relative to their initial expectations. However, short-run information doesn't completely replace prior beliefs. Forecasters are partially consistent with Bayesian updating as when updating their long-run forecasts, they place less weight on their long-run priors the more uncertain they initially are about them. However, they are also inconsistent with Bayesian updating as they neglect the precision of short-run effect estimates when updating.

This paper contributes to the recent literature on forecasts of causal effects in social science

contexts and the literature on long-run effects. There are several studies of the forecasts of the binary question of whether an experiment will replicate ([Camerer et al., 2016, 2018](#); [Forsell et al., 2019](#); [Fraser et al., 2023](#)). More relevant to policy decisions are forecasts of the magnitude of the treatment effect, not just whether there is a treatment effect that will replicate. The seminal paper in this literature is [DellaVigna and Pope \(2018\)](#) who study the accuracy of forecasts of 15 different treatments in an online effort experiment. [Otis \(2021\)](#) extends this research by studying the accuracy of forecasts of the results of three different short-run field experiments in Kenya. I replicate many of the key results in [DellaVigna and Pope \(2018\)](#) and [Otis \(2021\)](#) and extend them by (1) focusing on long-run effects and (2) collecting distributional forecasts rather than point estimates. I also provide evidence supporting the main results [Vivalt \(2020\)](#) on optimism bias and variance neglect.

There are also many other papers that, while their focus is on the results of a randomized experiment, also collect forecasts of the causal effects from the RCT. [Bernard and Vivalt \(nd\)](#) identify 29 studies of this sort and evaluate the relationship between forecast accuracy and time horizon across studies. Some prominent examples of papers in this vein include [Groh et al. \(2016\)](#), [Bloom et al. \(2020\)](#) and [Casey et al. \(2023\)](#).

Finally, I also contribute to the literature on long-run effects. [Bouguen et al. \(2019\)](#) review many long-term RCTs and highlight practical issues that occur when trying to econometrically estimate long-term effects. [Bernard \(2020\)](#) studies the performance of the surrogate index methodology for estimating long-term effects ([Athey et al., 2019](#)). I extend this literature by looking at the performance of a non-econometric method for predicting long-term results, judgmental forecasting.

This paper proceeds as follows: in section 2 I describe the studies for which I collect forecasts and the forecasters themselves. In section 3 I describe the design of the forecasting surveys and the accuracy measures used. In section 4 I present the main results of the study, and in section 5 I present the conclusion.

2 Data

2.1 Studies

I use long-term RCTs where the results were not publicly available at the time of forecast collection. To find the studies, I searched the AEA RCT Registry for RCTs where the planned endline was at least five years after the study start date. I wrote to researchers asking if they were interested in collecting forecasts of the long-run results of their studies, resulting in the seven studies below.

2.1.1 Boarding school France

[Behaghel et al. \(2017\)](#) investigate the impact of a French "boarding school of excellence" on students' cognitive and noncognitive outcomes. The school in question, established in 2009, caters to the

underserved suburban population. The policy initiative underlying this school was informed by the concern that detrimental factors such as poor academic quality, adverse peer influences, and unfavorable home study environments might hinder the progress of motivated students. It was anticipated that this boarding school could rectify these concerns by offering an enriched learning environment, higher academic expectations, and a peer group with stronger academic leanings.

The study's treatment group consists of students who were randomly selected to be offered a place at the boarding school, drawn from a larger pool of eligible applicants. The research focuses on the first two cohorts of students admitted to the school—in September 2009 and September 2010, respectively. During these two years, the demand for admission exceeded the school's capacity, leading to the conduct of a lottery. Out of the 395 eligible applicants, 258 were randomly selected for the treatment group and extended an offer to attend the school. The acceptance and retention rates were encouraging: 86 percent of lottery winners enrolled, with 76 percent remaining until the end of the first academic year. A small proportion of lottery non-winners also enrolled (6 percent) because one of their siblings was admitted to the school, and of these, five percent completed the first academic year. My analysis focuses on the intent-to-treat effects.

The study's authors analyze the impact on student cognitive and noncognitive outcomes from one to nine years post-experiment. In the short term, I use the outcomes of math test scores, French test scores, hours dedicated to homework, and intrinsic motivation, captured one and two years after the random admission offer. The long-term outcomes, assessed nine years after treatment assignment, encompass various measures of academic achievement and progression. These are completion of high school and attainment of any baccalaureate qualification (expected by eight years post-assignment or sooner), applications to higher education institutions, enrollment in higher education, and enrollment in elite academic institutions (Sciences Po, Dauphine, or preparatory class). These elite institutions are notably selective, setting them apart from the broader French university landscape.

2.1.2 Personal initiative Togo

[Campos et al. \(2017\)](#) evaluate the effectiveness of two business training programs introduced as part of a World Bank initiative in Lomé. Such training programs, common across developing countries, aim to elevate the income levels of small business owners through instruction in basic business operations like record-keeping, inventory control, and marketing.

A sample of 1,500 small businesses in Lomé, the capital of Togo, was selected from the applicants for the World Bank project. Eligibility criteria for applicants included having been in business for at least 12 months, having fewer than 50 employees, operating outside of the agricultural sector, and not being formally registered as a company. These firms were randomly distributed into three groups, each comprising 500 firms. The first treatment group, referred to as the 'traditional' group, received a conventional business training through the Business Edge program, encompassing four core topics: accounting and financial management, human resource management, marketing, and formalization.

The second treatment group, termed the 'initiative' group, underwent a novel program designed to inculcate a proactive, self-starter mindset. The control group did not receive any business training. Both training programs spanned three half-day sessions per week over a period of 4 weeks in April 2014, followed by monthly trainer check-ins for the following four months.

A majority of invited firms, 84 percent, participated in the trainings. My analysis again considers the intent-to-treat effect, contrasting the firms offered the training against the control group, irrespective of whether the firm actually partook in the training.

Survey data were collected at multiple points following the intervention, with four short-term surveys administered over the first 2.5 years (September 2014 - September 2016), and a long-term follow-up survey conducted from October 2021 to January 2022, seven years post-intervention. The key outcomes for this study are observed in both the short and the long-run, including business survival (whether the person is still operating any business), monthly sales, monthly profits, and personal initiative. Personal initiative was assessed via a seven-item agreement scale with statements such as "I actively attack problems" and "whenever something goes wrong, I search immediately for solutions". For the short-run effect, I use the average effect across the four short-term surveys in the first 2.5 years.

The study's core finding is that personal initiative training led to a more pronounced increase in firm profits compared to traditional training. Firms undergoing personal initiative training not only experienced a larger enhancement in personal initiative compared to traditional training recipients but also observed an increase in monthly sales relative to the control group.

2.1.3 Literacy Uganda

[Kerwin and Thornton \(2021\)](#); [Buhl-Wiggers et al. \(2022\)](#) conduct a Randomized Control Trial (RCT) to evaluate the Northern Uganda Literacy Project (NULP). With only 80% of grade 7 students in Uganda exiting primary school with the ability to read a short story, an educational firm devised a literacy program intended to improve these outcomes. The program was administered in two versions, full-cost and reduced-cost, spanning grades 1 to 3 (ages 6-8), thereby benefiting the children over a three-year period.

In the full-cost program, instruction is provided in the local language, Leblango, the lingua franca of the majority of the local population. This contrasts with the previous practice of using English, which the students were less familiar with. The program incorporates three residential (i.e., off-site) teacher training sessions per year, each roughly a week long, along with five classroom support visits per term to bolster teachers' capacity to instruct in Leblango. Teachers are trained to adopt a more engaged pedagogical approach, progressing at a slower pace to ensure students' reading ability keeps up. Detailed guides for daily and weekly lesson planning are provided, along with essential school supplies. Conversely, in the reduced-cost program, teacher training is conducted by government staff rather than the education firm's staff, fewer support visits are made, and no school supplies are

provided to the teachers.

The study includes 128 schools, randomized into three groups. The first treatment group consists of 42 schools (3,838 students) receiving the full-cost program. The second treatment group includes 44 schools (4,017 students) undergoing the reduced-cost program. The control group, comprising 42 schools (3,755 students), continues with their conventional teaching methods without the intervention.

The short-run outcomes measured student test scores at the conclusion of grades 2 and 3, including Leblango and English test scores. A long-term follow-up conducted five years post-program (eight years from the program's inception) included the same test scores along with mathematics test scores. Additionally, it evaluated students' school attendance in 2021 and their enrolment in secondary school in 2021 (assuming no dropouts and on-schedule grade progression, students should be in their first year of secondary school by this point). The study found significant short-run effects on literacy, as reflected in both Leblango and English test scores.

2.1.4 Financial education Spain

[Bover et al. \(2018\)](#) conduct a randomized controlled trial involving 3,000 9th grade students who participated in a financial education course named 'Finance for All' at various times throughout the year. This 10-hour course, delivered by high school teachers, aimed to equip 9th grade students with adequate financial literacy skills to enable sound financial decision-making. The program curriculum encompassed several areas, including saving and interest rates, budgeting, responsible consumption, types of bank accounts, and specific investment vehicles such as pension funds.

The study commenced in January 2015 with 78 participating schools, which were randomly assigned to either the treatment or control group. Surveys were administered to 9th and 10th graders, although only 9th graders received the course. The 9th grade students began the course at different times, meaning some received the course three months earlier than others. The control group of 9th graders undertook the course from April to June 2015, while the treatment group did it in January-March 2015.

By comparing the treatment group of 9th graders who completed the course in March 2015 with the control group who hadn't started the course yet, the authors estimated the immediate effect of the program. A fade-out effect was assessed by comparing the treatment group with the control group in June 2015, evaluating how much knowledge the treatment group lost over three months. Finally, long-term impacts of the course are estimated by comparing all former 9th graders to all former 10th graders in summer 2020. The immediate and fade-out effects are identified with experimental variation, but the long-run effect is not experimentally identified. Instead to identify the long-run effect we must assume that 9th and 10th grade students are comparable, except for the treatment that 9th graders received.

The immediate effects analyzed in my study include whether students have a bank account or

money card, whether students have a source of labor income, a financial knowledge index, and a hypothetical savings choice. I measure the fade-out effect on financial knowledge. For long-term impacts, I collect forecasts of the effect on a financial knowledge index and a financial awareness index. The authors conclude that while financial education enhances students' awareness about the value of resources and future implications of current choices, it does not necessarily bolster their intertemporal decision-making, such as helping them create a budget or improve their ability to stick to it.

2.1.5 Targeting the Ultra Poor Afghanistan

[Bedoya et al. \(2019\)](#) examine the impact of a big-push graduation program, also known as the 'Targeting the Ultra Poor' program (TUP), in Afghanistan. Eligible households had to meet specific criteria and were individually randomized with stratification at the Participatory Rural Appraisal (PRA) level.

As part of the TUP program, women received a one-off package that included: a transfer of livestock (typically cows, occasionally sheep and goats), worth approximately 1,312PPP(357 nominal); a consumption stipend of 54PPP(15 nominal) delivered in 12 monthly installments; training on livestock rearing and entrepreneurship; access to savings accounts and savings encouragement; facilitation of access to healthcare services; and coaching through biweekly visits for one year.

Eighty villages from four districts in the Balkh province were selected, and the ultra-poor households within each village were identified. Public lotteries were held in each village, resulting in 491 households assigned to the treatment group and 728 households to the control group. In May 2016, the treatment group began receiving the full TUP program for one year as outlined above, while the control group received none of the components.

A short-run follow-up survey was conducted two years after the initial livestock transfer and one year after the program ended. A long-run follow-up survey, intended to assess the five-year impacts of the program, was carried out from January-June 2021. The follow-up survey was successfully completed among 458 treatment households (93%) and 689 control households (95%).

The primary outcomes we ask people to predict in both the short and long run include monthly per capita consumption, total household income and revenues, total household savings, the number of cows owned (as this is the primary asset in the TUP program), and school enrollment (for children aged 6 to 18 years).

One year after the program's end, the labor choices of ultra-poor women have expanded, and the overall wellbeing of recipient households has improved. The share of households below the national poverty line decreased by 20 percentage points from 82% in the control group. Household savings increased by 2,195% (%106 USD PPP), and indebtedness decreased by 53%. These impacts were driven by an increase in income from livestock, due to the asset transfer, and a concurrent increase in women's labor participation by 22 percentage points.

2.1.6 Social signaling for vaccination Sierra Leone

[Karing \(2018\)](#) explores the concept of social signaling in the realm of childhood immunization in Sierra Leone. While 99% of children receive the first vaccine in Sierra Leone, the full immunization course of five vaccines by the age of one is completed by only 69% of children. The researcher introduced color-coded bracelets as a means for parents to publicly signal that their child had been vaccinated.

Sixty clinics, along with their corresponding communities, were randomized into treatment and control groups. In the treatment communities, all children were given a yellow "1st visit" bracelet upon receiving the first vaccine. If a child received all vaccines up to the fifth one in a timely manner, health workers replaced the yellow bracelet with a green "5th visit" bracelet. If a child was late for any vaccination, they were given an identical yellow "1st visit" bracelet. This process served as a public indication of whether a parent had ensured that their child received all five vaccines on time (green bracelet) or had failed to do so (yellow bracelet). The control communities followed the regular procedure without the use of bracelets, hence, there were no public indicators of vaccination status.

The experiment began in June 2016 and ran until August 2018. Children born within this period in treatment communities received the bracelets. An endline survey was conducted to assess the program's effects. A follow-up survey was implemented in 2020/2021 to compare immunization behaviors between the treatment and control groups for children born after the experiment ended. The aim was to estimate the long-term effects after the experiment ended and potentially the bracelets ran out, with a focus on whether there were lasting changes in vaccination rates due to social learning or habit formation. Separate forecasts were elicited for effects on children who had an older sibling participating in the experiment, and those who did not.

The study found that the signaling treatment at the 5th vaccine led to a significant and substantial increase of 14.4 percentage points in the rate of receiving all five vaccinations in a timely manner..

2.1.7 General equilibrium effects of cash Kenya

[Egger et al. \(2022\)](#) conducted a randomized controlled trial to study the effects of providing one-time cash transfers to over 10,500 impoverished households across 653 randomized villages in rural Kenya. Households living in homes with thatched roofs were eligible for the transfer. Eligible households received three cash transfers over a six-month period, totaling \$1871 USD (in 2014 PPP dollars, or 87,000 Kenyan Shillings).

In this study, the researchers manipulated the intensity of village-level saturation of unconditional cash transfers within sublocation clusters of two to 15 villages in Siaya County, Kenya. This allowed them to identify general equilibrium effects. In high-intensity sublocations, two-thirds of eligible households in the randomly assigned villages received cash. Meanwhile, in low-intensity sublocations,

cash was received by eligible households in only one-third of the villages. The study focused on two primary household outcomes: annual household consumption and total household assets. Due to the experimental design, households were classified into eight different types: eligible and ineligible households, in high or low-intensity sublocations, within treated or untreated villages. Forecasts were collected for six of these groups, with average outcomes for eligible and ineligible households living in low-intensity sublocations in cash-free villages serving as the control group baseline.

The first household follow-up survey was conducted approximately 1.5 years after the study commenced, and 11 months after the final cash installment was transferred. A long-term follow-up survey was carried out seven years after the study began. While all eligible households received the cash transfer, only a subset was surveyed. In each village, eight eligible and four ineligible households were randomly selected for the survey, yielding a final sample size of 8,242.

In the short term, recipient households reported significantly higher total expenditure—an increase of 11.6% over the control village in low-saturation areas. Coinciding with increased expenditure on durables, asset stocks also increased. Specifically, asset measures increased by 26% of the mean for eligible households in control villages located in low-saturation sublocations.

2.2 Forecasters

I recruited three groups of respondents to provide forecasts: academics, expert forecasters, and nonexperts. For the first two groups of respondents, I hosted the survey on the Social Science Prediction Platform (SSPP). I recruited respondents both passively and actively. The passive recruitment was from the SSPP’s regular pool of forecasters and their mailing list. The active recruitment was done by emailing and advertising on Twitter. Respondents on the SSPP were offered two incentives: (1) a \$40 payment if they completed four or more of the surveys, and (2) for each study 10% of respondents were randomly selected and paid an accuracy bonus of up to \$100 that depended on the accuracy of their final long-run forecasts.

2.2.1 Academics

Some academics were recruited passively to take part in the study from the SSPP’s regular pool of respondents. Each survey was available on the SSPP for approximately three months. Additionally, for each study, I collected the names and emails of 50-100 researchers who had either been cited in or cited the paper studying the short-run results of the relevant RCT. Then I (or the study authors) personally invited the researchers to provide forecasts for that study. I also invited people via various departmental mailing lists.

Lastly, I sent emails to staff at a number of development organizations inviting them to take the surveys. This was often after I gave a presentation on the results from (Bernard, 2020) and Bernard et al. (nd) and invited people to participate in the study at the end of the presentation.

The organizations invited were GiveWell, Open Philanthropy, IDinsight, Charity Entrepreneurship, Rethink Priorities, the Center for Global Development, and the Development Innovation Lab. There were few respondents from these organisations, but I include them in the academic sample.

2.2.2 Expert forecasters

I invited Pro Forecasters from Metaculus and superforecasters from Good Judgement (two popular online forecasting platforms) to participate in the study. These groups of forecasters have a strong track record in forecasting, with Pro Forecasters being selected among the highest ranked Metaculus forecasters and superforecasters being among the top 2% of forecasters on Good Judgment Open. However, they do not necessarily have much exposure to the social sciences and the contexts under study, with their forecasting mostly relating to geopolitical events. Due to concerns about low uptake from these groups (since they do not work/research in the social sciences), I offered these forecasters an additional \$100 for completing all 7 surveys. There are also other forecasters in my sample who have experience on these platforms, but are not necessarily Pro Forecasters or superforecasters.

2.2.3 Nonexperts

I recruited nonexperts using Prolific, a popular survey research recruitment platform, similar to Amazon Mechanical Turk. For each study, I posted a survey on Prolific entitled “Forecasting long-run effects” and recruited approximately 150 people for each survey. For each survey, I recruited 50 respondents from each of three countries: the USA, Spain and France. I included Spanish and French respondents to increase the variation in how familiar respondents were with the context under study in the Boarding school France and Financial education Spain papers.¹ I paid respondents £3 per completed survey and offered a bonus of £10 if they completed all 7 surveys. Since I had to pay nonexpert forecasters for completing the survey, their maximum accuracy reward was limited to £50, rather than \$100. To improve comprehension and survey completion rates, I restricted participants to fluent English speakers, with an undergraduate degree or higher, who had completed at least 50 studies on Prolific before and had a Prolific approval rate of 99 or 100.

3 Methodology

3.1 Survey design

There is a separate survey for each RCT and they all follow the same format:

1. **Consent:** Participants confirmed their agreement to take part in the experiment.

¹Ideally I would have recruited respondents from Afghanistan, Kenya, Uganda, Togo and Sierra Leone as well, but there are very few potential participants from these countries on Prolific.

2. **Introduction:** Participants were provided with an example of how forecasts will be elicited and informed about the available participation and accuracy incentives.
3. **RCT Description:** Respondents read a brief description of the RCT, which included details on the context, the intervention(s) under study, the experimental design, data collection, and key outcomes for which forecasts were required. The descriptions of the RCTs are included in Appendix A.
4. **Comprehension Questions:** Respondents then answered three questions to test their understanding of the study. Those who answered incorrectly were asked to read the study description again. Respondents who failed the comprehension questions twice were still allowed to complete the survey, but their responses are excluded from the main analysis.
5. **Randomised Information:** Respondents were then randomized to receive either (1) tips on forecasting well, based on Tetlock’s 10 commandments for aspiring superforecasters, or (2) no additional forecasting tips. They were also cross-randomized to receive either (1) additional contextual information about the study, (2) information about the effects of similar interventions in different contexts, or (3) no additional information. The contextual information includes additional baseline characteristics about the participants in the RCT and broader context for the setting of the study. The intervention information includes the results of a meta-analysis where possible or the short and long-run results of 2-3 similar randomized controlled trials. The randomised information is included in Appendix A.
6. **Prior Forecasts:** Respondents then provided their initial prior forecasts of the causal effect on short and long-run outcomes.² For each causal effect, participants provided three forecasts: a most likely case, a worst case, and a best case. They enter these forecasts in textboxes using the natural units of each outcome as in figure 1. Respondents’ worst case forecasts are forced to be lower than their most likely case and their best case is forced to be higher than their most likely case and the survey will not proceed to the next page unless these constraints are satisfied. Bounds are also placed on the minimum and maximum forecast. For naturally bounded outcomes such as a school enrollment rate, the natural bounds are used e.g. -100 percentage points to +100 percentage points. For outcomes measured in standard deviations (e.g. test scores) bounds of -5 to +5 standard deviations are used. For other unbounded outcomes, I attempted to set bounds at roughly the equivalent of -5 to +5 standard

²For some studies, the short and the long-run outcomes are the same but measured at different points in time, whereas for other studies the short and long-run outcomes are entirely different. Some studies have short-run outcomes from multiple time points (e.g. one and two years after treatment) and some studies have long-run outcomes from multiple time points. Some studies have multiple treatments so for these participants provide forecasts of the causal effect of each treatment on each outcome. Some studies estimate treatment effects on multiple subgroups so respondents provide forecasts of the effect on each outcome for each subgroup.

The average effect on mathematics test scores in standard deviations.

	Worst case	Most likely	Best case
1 year after offer	<input type="text"/>	<input type="text"/>	<input type="text"/>
2 years after offer	<input type="text"/>	<input type="text"/>	<input type="text"/>

Figure 1: Example of how forecasts are elicited from respondents

deviations.³ Respondents also provide their qualitative, Likert confidence for their short-run forecasts and their long-run forecast separately.

- Short-run Effect Information:** Participants were then informed about the short-run effects. They were shown a table (like Figure 2) detailing their most likely forecast, the actual treatment effect, the p-value of the actual treatment effect, and the levels of the treatment and control groups.
- Updated Posterior Forecasts:** Participants were asked to update their long-run forecasts in light of the short-run effects. Their prior long-run forecasts were pre-filled in the forecast elicitation boxes, which they could edit. The same constraints and bounds on forecasts are imposed here as before.
- Questionnaire:** Finally, participants were asked a series of questions about their familiarity with the context, the intervention, and forecasting.

By following this structure, each survey was consistent, aiding in the comparison of results across RCTs.

3.2 Accuracy measures

I assess the quality of forecasts using three distinct measures. The primary measure is the log score, which is used for determining the accuracy reward that 10% of randomly chosen respondents to each survey receive. The secondary measures are the error and negative absolute error of the most likely case forecast. Before calculating these accuracy measures, I normalize all forecasts and treatment

³As I did not know the long-run causal effects or the standard deviations of the outcomes when designing the surveys, sometimes the bounds are larger than -5 to +5 standard deviations. I erred on the side of providing wider bounds because I wanted to minimise the probability of the actual treatment effect falling outside the bounds. If respondents end up giving forecasts less than -5 standard deviations or greater than 5 standard deviations I winsorise these to ± 5 standard deviations.

Outcome	Year	Your forecast	Actual treatment effect	Treatment group	Control group	p-value
Math score	1	0.2	-0.037	-0.014	0.023	0.70
Math score	2	0.13	0.280	0.303	0.023	0.01
French score	1	0.29	-0.065	0.043	0.022	0.54
French score	2	0.18	-0.115	-0.093	0.022	0.35
Intrinsic motivation	1	0.45	0.047	0.037	-0.01	0.71
Intrinsic motivation	2	0.65	0.367	0.357	-0.01	0.004
Homework hours	1	3	0.100	6.198	6.098	0.83
Homework hours	2	2	1.601	7.699	6.098	0.003

Figure 2: Example of how short-run results were shown to respondents

effects by dividing by the control group standard deviation. I then winsorize the forecasts to be within $[-5, 5]$ standard deviations to screen for misinterpretations and carelessness.

3.2.1 Log score

In the realm of probability forecasting, proper scoring rules play a crucial role. These rules have the desirable property that they incentivize forecasters to report their true beliefs, rather than hedging their bets or manipulating their forecasts. In other words, the expected score under a proper scoring rule is maximized when a forecaster reports their true probability distribution (Gneiting and Raftery, 2007).

Among the various types of proper scoring rules, log scoring rules are particularly useful and commonly used. For this study, I employ a variant of the standard log scoring rule proposed by Greenberg (2018). This approach calculates the score of a forecast for causal effect τ as follows:

$$score_{\tau} = 0.1 \times \log \left(\frac{P(x)}{P(u)} \right) \quad (1)$$

where $P(x)$ is the probability density the respondent assigns to the correct answer and $P(u)$ is the probability density a uniform distribution between -5 and $+5$ standard deviations assigns to the correct outcome (i.e. 0.1). When calculating the score, if $P(x) < P(u)$, I set $P(x) = P(u)$. This correction ensures that the score for each forecast has a lower bound of 0 (as $\log(1) = 0$). This avoids heavily penalising respondents with an infinite negative score if they assign a 0 probability density to the correct answer, an issue with the regular log score.

I construct each respondent's $P(x)$ for each outcome by using their most likely, best and worst case forecasts to construct a triangular distribution of their belief. The probability density function for the constructed triangular distribution is:

$$\begin{aligned} f(x) &= 2(x - W)(M - W)(B - M), & W < x < M \\ &= 2(B - x)(B - M)(B - W), & M < x < B \\ &= 0, & otherwise \end{aligned}$$

where W is the respondent's worst case forecast, M is their most likely forecast, and B is their best case forecast.

The primary advantage of using a triangular distribution is its simplicity, requiring respondents to specify only three easily understood parameters. Unlike a normal distribution, it allows people to specify asymmetric distributions, which is advantageous for treatment effects that are rarely below 0 . However, the assumption of linearity imposes limitations on the complexity of the distribution that can be specified and may lead respondents to assign lower probabilities to tail events.

Respondents are incentivized based on their score to encourage accurate forecasting. A randomly selected 10% of respondents are paid an amount of \$100 (or £50 for Prolific respondents) multiplied by the average score of their final long-run forecasts. This reward incentivises respondents to report their true beliefs.

3.2.2 Error

Apart from the log score, I also use the error of the forecast as a measure of accuracy. The error of a forecast for causal effect τ is defined as:

$$Error_{\tau} = M_{\tau} - \beta_{\tau} \tag{2}$$

where β_{τ} is the observed experimental effect and M_{τ} is the most likely case forecast provided by the respondent (or the average of M for a group of respondents). This measure of error is intuitive and straightforward as it calculates the raw difference between the forecasted effect and the actual effect. However, it should be noted that this method of calculating error leads to overestimation and underestimation of the effect cancelling out. However, it is informative about whether respondents tend to be overoptimistic or overpessimistic about the effects of treatments.

3.2.3 Negative absolute error

To overcome the limitation of the error calculation mentioned above, I also use the negative absolute error (NAE) of the forecast for causal effect τ :

$$NAE_{\tau} = - | M_{\tau} - \beta_{\tau} | \tag{3}$$

where β_{τ} is the observed experimental effect and M_{τ} is the most likely case forecast provided by the respondent (or the average of M of a group of respondents). Unlike the previous error measure, the absolute error takes into account the magnitude of the error irrespective of whether the effect was overestimated or underestimated.

While this measure is not a proper scoring rule like the log score, it is useful for quantifying the size of errors in a more intuitive manner. By taking the negative of the absolute error, the measure maintains the convention that higher scores are better, facilitating comparison across different measures. It helps to understand the average deviation of forecasts from the actual outcomes, making it a useful tool in the analysis of forecast accuracy.

4 Results

4.1 Forecast accuracy

First, we look at how accurate forecasters are. The main results are summarised in table 1 which shows the accuracy of the three over different time horizons and using different types of forecasts; individual and aggregate. I compute the aggregate forecast for each group by taking the median of the most likely, best and worst case forecasts (separately for each outcome) and using these average values to construct the accuracy measures. The accuracy measures used are the log score, where higher scores indicate better accuracy, the negative absolute error, where less negative values indicate better accuracy, and the error, with values closer to 0 indicating better accuracy. In Panel D I provide benchmarks by showing the score or error you would have received if you used the given uniform or triangular distribution for all forecasts.

There are a number of results evident from table 1. We first compare the accuracy of different types of forecasters and discuss a key difference between academics and forecasters, differences in calibration. Next we look at how the accuracy results change between the short and the long-run forecasts, and the difference between prior long-run accuracy and posterior long-run accuracy after participants are told the short-run impacts. Finally, we look at the gains from aggregating forecasts.

Table 1: Accuracy of individual and averaged forecasts

Time \times Attempt	Score			Negative Absolute Error			Error	
	Average (SD) (1)	Aggregate forecast (2)	% better than aggregate (3)	Average (SD) (4)	Aggregate forecast (5)	% better than aggregate (6)	Average (SD) (7)	Aggregate forecast (8)
<i>Panel A: Academics</i>								
Short-run	0.108 (0.169)	0.156	23.2	-0.342 (0.472)	-0.195	34.5	0.022 (0.583)	-0.101
Long-run 1 Prior	0.107 (0.170)	0.143	23.4	-0.338 (0.497)	-0.158	35.0	0.089 (0.595)	-0.029
Long-run 2 Posterior	0.133 (0.186)	0.198	23.2	-0.332 (0.529)	-0.149	33.9	0.150 (0.606)	0.044
<i>Panel B: Expert forecasters</i>								
Short-run	0.154 (0.178)	0.183	26.9	-0.280 (0.427)	-0.172	33.7	0.049 (0.509)	-0.057
Long-run 1 Prior	0.167 (0.187)	0.188	27.8	-0.232 (0.356)	-0.171	33.3	0.097 (0.414)	0.044
Long-run 2 Posterior	0.182 (0.191)	0.214	30.2	-0.223 (0.396)	-0.165	33.7	0.107 (0.442)	0.070
<i>Panel C: Prolific nonexperts</i>								
Short-run	0.091 (0.155)	0.127	19.7	-0.487 (0.636)	-0.310	35.0	0.213 (0.772)	0.083
Long-run 1 Prior	0.093 (0.159)	0.121	23.3	-0.450 (0.628)	-0.278	38.3	0.314 (0.706)	0.158
Long-run 2 Posterior	0.105 (0.17)	0.185	21.5	-0.454 (0.648)	-0.190	36.9	0.347 (0.711)	0.115
<i>Panel D: Uniform and Triangular benchmark</i>								
$U[-0.5, 0.5]$	0.206							
$U[-1, 1]$	0.158							
$T[-0.5, 0, 0.5]$	0.245			-0.162			-0.144	
$T[-1, 0, 1]$	0.205			-0.162			-0.144	

Notes: Column 1 shows the average score of individual forecasters from the three groups (academics, forecasters and Prolific participants) with the standard deviation of the score in parentheses. Column 2 shows the score of the aggregate forecast from that group. The aggregate forecast is constructed by taking the median of each group's worst case, most likely case and best case forecast and constructing a triangular distribution with these parameters, separately for each causal effect being forecast. Column 3 shows the percentage of forecasters of that group who had a higher score than the median. Columns 4 to 6 give the same information for the negative absolute error. Columns 7 and 8 give the same information for the error (the % who a better error than median is not included). Panel D represents the scores, negative absolute errors and errors that would have been achieved if forecasters provided for all forecasts (a) a uniform distribution between -0.5, (b) a uniform distribution between 0.5 or -1 and 1, (c) a triangular distribution with minimum -0.5, mode 0, maximum 0.5, or (d) a triangular distribution with minimum -1, mode 0, maximum 1.

4.1.1 Forecaster type

We can see that for each of the three accuracy measures and in each time horizon of forecast, the non-expert Prolific group performs worst. We can see that this is true for both the individual forecasts in columns (1), (4) and (7), and the aggregated forecasts in columns (2), (5), (8). Of particular note is the average error of the Prolific forecasts being between around 0.3. This means that the Prolific group tends to overestimate the benefit of treatment by 0.3 standard deviations on average, a substantial overestimation. Academics and expert forecasters also tend to overestimate the benefits of treatment on average, but by a substantially smaller margin. This consistent overoptimism of the Prolific forecasters is likely to drive a large share of the differences in accuracy between them and the experts.

Now we move to comparing the academics with the expert forecasters. We can see that in the descriptive statistics, for every time frame, the expert forecasters have better average scores and negative absolute errors than academics. However, we investigate this further in table 2. We run regressions of the different accuracy measures on forecaster type, with the Prolific nonexpert group as the omitted group. We test whether the coefficients on the academic and forecaster indicators are statistically significantly different. Importantly, in table 2 we include study fixed effects so our comparisons are between the accuracy of academics and expert forecasters in a given study. This is important as there may be selection into studies of varying difficulty by either group and this selection is not captured in table 1.

Firstly, table 2 confirms that the academics and expert forecasters are significantly more accurate than the Prolific group at every time frame and on every accuracy measure. According to the score measure, across all time frames, academics are 23% more accurate and expert forecasters are 61% more accurate on average than nonexperts. This is a large, significant difference in accuracy between the experts and the non-experts. However, the comparison between academics and expert forecasters is more complex.

The expert forecasters have better log scores than the academics for every type of forecast. This difference is statistically significant at the one percent level for all forecasts in Panel A, at the 5% level for prior short and long-run forecasts in Panel B and C, and is insignificant for the posterior long-run forecasts in panel D. However, the differences between the negative absolute error of academics and forecasters is not significant in any of the panels although the academics have higher coefficients. This result is different to the result in the descriptive statistics because here we include study fixed effects, which makes this result a more fair comparison between the two groups.

Recall that the negative absolute error only depends on the most likely case forecast, whereas the score depends on the full distribution as calculated from the most likely, best and worst case forecasts given by participants. Given this, the results suggest that academics and expert forecasters are equally good at giving most likely case forecasts, but forecasters are better at giving best and worst case forecasts. In other words, the expert forecasters are better at forecasting a range of

Table 2: Comparison of forecasters type

	Score	NAE	Error
<i>Panel A: All forecasts, n = 19585</i>			
Academic (A)	0.022*** (0.006)	0.147*** (0.027)	-0.196*** (0.031)
Forecaster (F)	0.059*** (0.012)	0.127*** (0.035)	-0.094** (0.043)
p-value (A = F)	0.008	0.689	0.089
Prolific mean	0.096	-0.463	0.294
<i>Panel B: Short-run, n = 6292</i>			
Academic (A)	0.013* (0.007)	0.153*** (0.028)	-0.207*** (0.033)
Forecaster (F)	0.055*** (0.018)	0.120*** (0.044)	-0.065 (0.055)
p-value (A = F)	0.042	0.543	0.042
Prolific mean	0.091	-0.487	0.213
<i>Panel C: Long-run 1, n = 6622</i>			
Academic (A)	0.021*** (0.008)	0.136*** (0.034)	-0.200*** (0.038)
Forecaster (F)	0.064*** (0.016)	0.128*** (0.042)	-0.094* (0.050)
p-value (A = F)	0.021	0.893	0.137
Prolific mean	0.093	-0.450	0.314
<i>Panel D: Long-run 2, n = 6671</i>			
Academic (A)	0.034*** (0.008)	0.153*** (0.036)	-0.181*** (0.039)
Forecaster (F)	0.057*** (0.016)	0.133*** (0.041)	-0.130*** (0.045)
p-value (A = F)	0.251	0.755	0.453
Prolific mean	0.105	-0.454	0.347

Notes: Standard errors clustered at the study \times person level, with study fixed effects included (and time \times attempt fixed effects included in panel A). P-values for a test of the equality of the academic coefficient and forecaster coefficient are given. Omitted group is nonexperts from Prolific.

plausible treatment effects but no better at specifying the most likely effect within that range.

4.1.2 Calibration

Another way to explore this result is by looking at the calibration of the different types of forecasters. In the context of forecasting, calibration refers to the degree of correspondence between predicted probabilities and actual outcomes. A forecaster is considered perfectly calibrated if, over time, the proportion of their predictions that end up being correct matches the predicted probability. For example, if a forecaster predicts an event with a 70% probability, then ideally, out of 100 such forecasts, the event should happen 70 times.

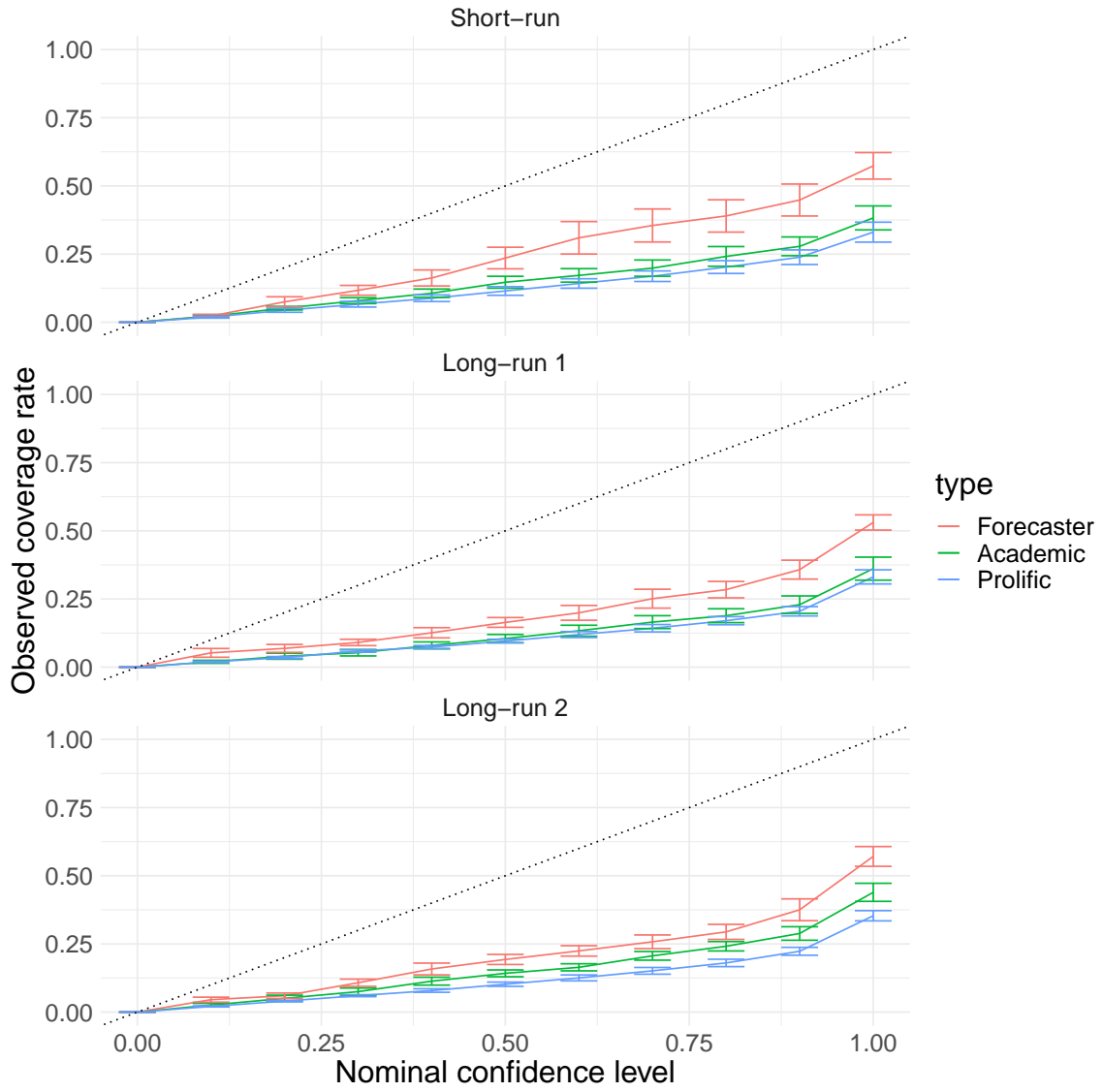
Calibration is not about individual predictions but is a property that emerges over a set of forecasts. It provides insight into the reliability of a forecaster's (or a group of forecasters') probability estimates and therefore speaks to the quality of the forecaster's judgment. Poor calibration could indicate that a forecaster is overconfident (if actual outcomes occur less frequently than predicted) or underconfident (if actual outcomes occur more frequently than predicted).

Figure 3 shows calibration curves for the individual forecasts from academic, expert forecaster and Prolific respondents for all time frames. A calibration curve is used to evaluate the quality of probabilistic predictions by representing the degree to which forecasters' confidence intervals align with the actual outcomes. On the x-axis, we have the nominal confidence level, which represents the expected likelihood that the outcome will fall within the given interval. This nominal confidence level is derived from the forecasters' triangular distributions - for instance, the 10%, 20%, 30% intervals, and so on, up to 100% which spans from their worst case to their best case forecast. On the y-axis, we have the observed coverage rate, which is the proportion of times that the actual treatment effect fell within the corresponding confidence interval.

In an ideal scenario, the calibration curve should coincide with the diagonal line, which signifies perfect calibration - that is, the forecasters' stated confidence levels match the observed coverage rates. If the curve lies above the diagonal, it suggests that forecasters are underconfident, with outcomes falling within their confidence intervals more often than expected. Conversely, if the curve lies below the diagonal, it suggests overconfidence, with outcomes falling within their confidence intervals less frequently than predicted.

In figure 3 we see that all respondents are poorly calibrated. As the calibration curves lie below the 45 degree line of perfect calibration, this implies that respondents are overconfident and their stated confidence intervals are too narrow. This pattern of overconfidence replicates what is shown in [Otis \(2021\)](#).

However, we see important differences between the calibration of the different groups. Expert forecasters are clearly more calibrated than academics and nonexperts, with their calibration curve lying above the curves of the other groups and the differences often being statistically significant. Academics, on the other hand, are only slightly more calibrated than the nonexperts with the



Notes: Bootstrap standard errors clustered at the study level shown

Figure 3: Calibration curves

difference between the two groups often being insignificant, especially in the short-run and first long-run forecasts. The estimated treatment effect only lies inside the worst and best case forecasts a third of the time for the Prolific group, 40% of the time for academics and 55% of the time for expert forecasters.

This difference in calibration across groups is to be expected. The expert forecasters have a history of making, refining and evaluating forecasts in the past and have had the opportunity to improve their calibration. Academics on the other hand have typically not engaged in the practice of forecasting, evaluating their forecasts, updating their confidence levels and improving their calibration. Notably however, in the second round of long-run forecasting academics get closer in calibration to the forecasters and are consistently statistically significantly better than the nonexperts. This suggests that seeing the short-run effects and how their forecasts compare already helps to improve the calibration of academics.

Figure 4 also studies the calibration of forecasters, but this time by looking at the relationship between their average score stated confidence (Not at all confident, not very confident, neither confident nor unconfident, fairly confident and very confident). We can see that for the academic group, there is actually negative relationship between their stated confidence and their accuracy, with more confident academics tending to have lower scores. The forecasters on the other hand have a positive relationship between their stated confidence and their score, at least for the two long-run forecasts. Notably, forecasters never say they are very confident in their forecasts, displaying more epistemic humility than the academic and Prolific groups.

Better calibration of forecasters seems to be a big part of what gives them a higher accuracy than academics as measured by log score. Although there is no difference in the accuracy of the most likely forecast provided by both groups, forecasters are better at knowing what they know and display less overconfidence than academics.

4.1.3 Time horizon

We turn our attention back to table 1 to examine how forecast accuracy varies across three time horizons: short-run, prior long-run, and posterior long-run. The short-run outcomes are typically observed one to two years after treatment commenced whereas the long-run effects are observed five to nine years after treatment. In particular, we focus on (1) comparing short-run forecasts with prior long-run forecasts and (2) comparing prior long-run forecasts with posterior long-run forecasts after informing respondents about the short-run effects. It's important to remember that only the long-run outcomes are true forecasts, as the short-run outcomes were publicly available in published articles or working papers at the time of the survey.

Despite this, our results reveal an intriguing pattern where prior long-run forecasts tend to be as accurate, if not more so, than short-run forecasts. Upon examining the average scores and negative absolute errors (columns 1 and 4 respectively), we find that in 5 out of 6 cases, the average

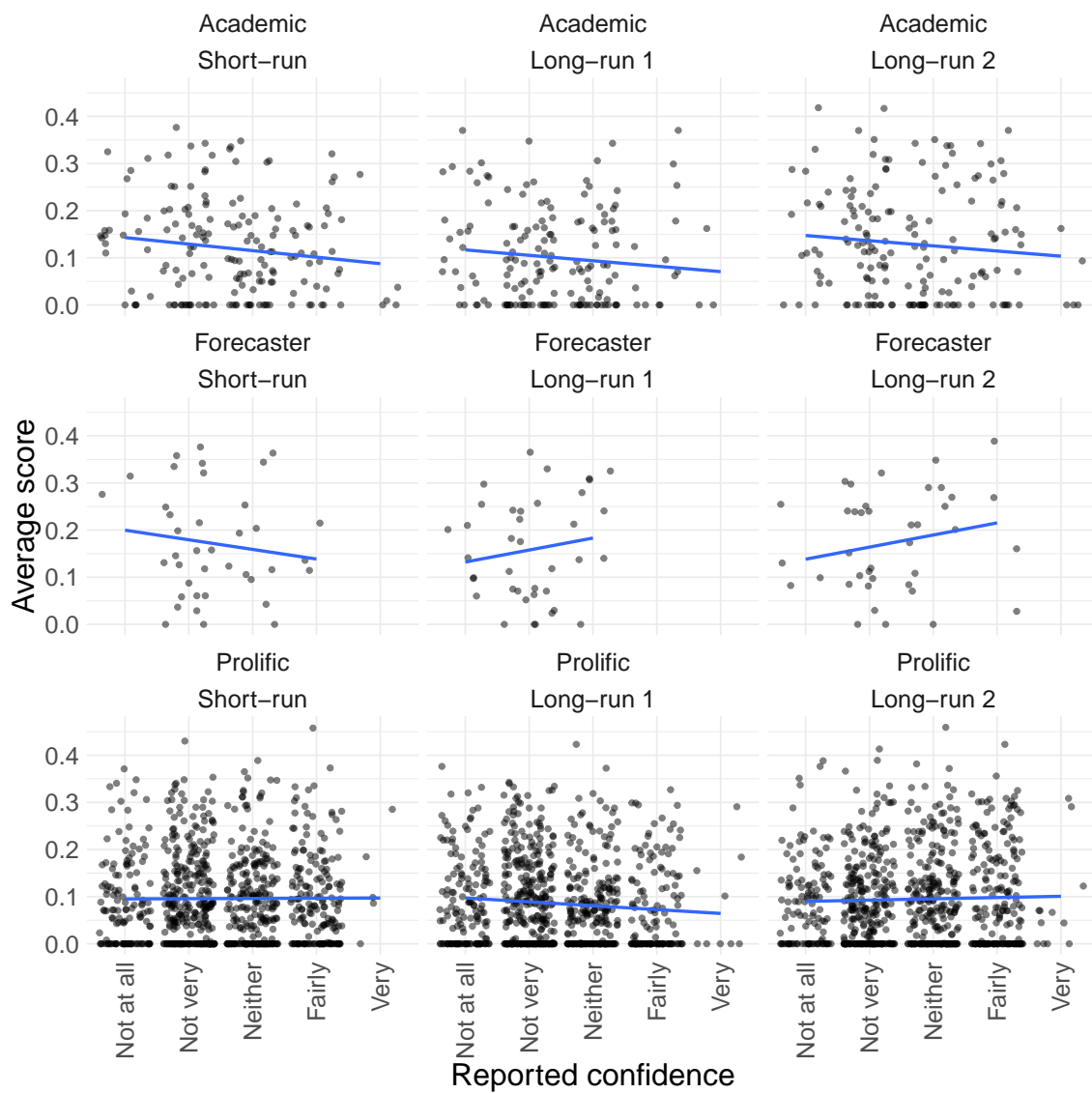


Figure 4: Score and qualitative confidence

prior long-run forecast is more accurate than the average short-run forecast. The exception to this is in panel A, column 1 where academics' average short-run forecast score (0.108) marginally outperforms their prior long-run forecast score (0.107). This exception is in line with the academics being more likely to know the short-run results already, although the difference in accuracy between the supposedly known short-run 'forecasts' and the unknown long-run forecasts is smaller than would be expected.

One explanation for this could be that short-run effects have a higher variance than long-run effects. One model of the world could be that interventions have varying effects in short-run, but in the long-run the effect of everything is 0 (to quote Keynes, "in the long-run we are all dead"). If this were the case, and forecasters knew that this was the case, we might expect to see this pattern of long-run forecasts being more accurate than short-run forecasts. Unfortunately, as the long-run treatment effects for each of my studies are currently embargoed, I cannot present analysis on the relative accuracy of small versus large long-run effects.

However, table 1 also shows that people are relatively more overoptimistic about the benefits of treatment in the long-run than in the short-run. In column 7 we can see that the average error tends to be larger and more positive in the long-run than the short-run for all groups. Furthermore, the posterior long-run forecasts are even more overoptimistic than the prior long-run forecasts. Being told the short-run impacts does not seem to curb overoptimism about long-run effects. A possible explanation for this is [Vivalt and Coville \(nd\)](#)'s asymmetric optimism where "good news" is favoured over "bad news" when people update their beliefs. As respondents receive information about the effects on multiple short-run outcomes, it's possible the short-run outcomes with the largest positive effects are most salient when they are updating their long-run forecasts.

More in line with expectations, we can see that the posterior long-run forecast is more accurate than the prior long-run forecast (with the one exception being in panel C, column 4 where nonexperts' negative absolute error of -0.450 for the prior long-run and -0.454 for the posterior long-run). This is to be expected, although there are at least two possible explanations for why this could be the case.

Firstly, it could be that the short-run information is informative about the long-run impacts and forecasters are able to improve their long-run forecasts once they know it. Secondly, it could be that forecasters update on their own ability and calibration after receiving feedback on how accurate their short-run forecasts are. However, in table 8 we show that short-run effects are positively correlated with long-run effects, supporting the first explanation. Additionally, if the only effect were through forecasters updating on their ability, it's not clear why we would see an improvement in the negative absolute error as well as the score. The score may improve as forecasters realise they were overconfident and widen their range, increasing the probability that their range contains the true effect (although lowering their probability density on any given point in that range). On the other hand, it's not obvious how learning about their ability would change their forecast of the most likely case to be systematically closer to the true treatment effect and therefore improve their

negative absolute error.

4.1.4 Aggregation

Table 1 also shows that there are accuracy gains from aggregating forecasts, a wisdom of the crowds effect. Columns 2, 5 and 8 show the score, negative absolute error and error respectively, of the aggregated median forecast. This aggregation is done by taking the median of each group’s worst case, most likely case and best case forecast and constructing a triangular distribution with these parameters, separately for each causal effect being forecast.

We can see that the average score of individual forecasters is worse than the aggregated median forecast. This is true across all forecast types and participant groups. For instance, within the academic group, the average short-run forecast score is 0.108 (with a standard deviation of 0.169), which is notably lower than the score for the median forecast, 0.156. This trend also holds in the case of expert forecasters and Prolific nonexperts. A small number of forecasters, however, did outperform the median. For example, 23% of academics achieved higher scores than the aggregate of their forecasts.

Looking at the negative absolute error, in columns 4 to 6, a similar pattern emerges. The average negative absolute error for individual forecasters is consistently larger (indicating less accuracy) than that of the median forecast across all forecast types and participant groups. Yet again, a significant percentage of forecasters, often over a third, achieved better negative absolute errors than the median forecast.

The fact that more forecasters exceed the median in terms of negative absolute error (NAE) than in terms of score can be interpreted as suggesting that the benefits of aggregating distributions are more significant than those of aggregating point estimates. The NAE is reliant on the most likely case forecast - a single point estimate - whereas the score incorporates the worst and best case forecasts, thus reflecting not only the forecasters’ expected outcomes, but also their uncertainty around those expectations. It may be the case that individual forecasts of the worst and best cases are noisier than the forecasts of the most likely case, and thus the gains from aggregating them are larger.

However, it must be noted that even the aggregate forecasts are not particularly accurate. In panel D, I show what scores would have been received if your forecast for each causal effect was either (a) a uniform distribution⁴ between -0.5 and 0.5, (b) a uniform distribution between -1 and 1, (c) a triangular distribution between -0.5 and 0.5 with a mode of 0, and (d) a triangular distribution between -1 and 1 with a mode of 0.

The aggregate academic and Prolific forecasts tend to outperform the lowest benchmark score of 0.158 from the uniform distribution from -1 to 1. The expert forecasters aggregate consistently

⁴Note that respondents could not actually provide uniform distributions, they were restricted to providing triangular distributions. I use a uniform distribution just for illustration.

outperforms this benchmark, and their second long-run aggregate forecasts are only outperformed by the triangular distribution from -0.5 to 0.5. However, these benchmarks are not particularly demanding as they simply forecast the same distribution for every outcome without using any inside information. The benchmarks should not be viewed as rigorous targets but rather as baseline values and achieving them should be a minimal requirement for forecasting to be informative.

In terms of negative absolute error, the aggregate forecasts of academics and expert forecasters are comparable to the benchmark of forecasting a 0 effect for everything. The nonexperts on the other hand are substantially worse than the benchmark, suggesting substantial room for improvement.

In light of these observations, we must acknowledge that even though aggregation does improve forecast quality, even the best aggregate forecasts leave much to be desired in terms of accuracy. This should lead us to being more epistemically humble in our forecasts of both short and long-run effects of policies and more accepting of the uncertainty that surrounds policy evaluation and forecasting. The median aggregation method used is relatively simplistic and it may be the case that more complex methods which give different weights to different forecasters perform better.

4.2 Mechanisms

I now turn to studying what drives forecast accuracy within groups. First, I use the randomised provision of information to assess the causal impact of different information sets on forecast accuracy. Then I turn to a non-causal, more exploratory assessment of the correlates of accuracy within the academic and nonexpert groups.⁵ Finally, I look at another potential determinant of accuracy across groups, the effort made by forecasters, to complement the previous analysis on calibration.

4.2.1 Randomised information

As a reminder, respondents are randomised to receive different types of information in two stages. In the first stage, they are randomized to either (1) receive tips on how to forecast well (Forecast), or (2) a control group with no forecasting tips. In the second stage, respondents are cross-randomized to either (1) additional contextual information about the study (Context), (2) additional information about the effects of similar interventions in different contexts (Intervention), or (3) a control group with no additional information.⁶ As such, in table 3 I show the effect on accuracy of the interactions of forecast information with context and intervention information, as well as the individual indicators. Additionally, I run the regressions separately for the Prolific group and the combined group academics and nonexperts (SSPP since they did the surveys on the Social Science Prediction Platform). I combine the academic and expert forecaster group due to the relatively small sample size of expert forecasters.

⁵I do not look at correlates of accuracy within the expert forecaster group due to limited sample size.

⁶You can see the information respondents were provided in each survey in appendix A

Table 3: Effect of randomised information on accuracy

	SSPP		Prolific	
	Score (1)	NAE (2)	Score (3)	NAE (4)
Forecast	0.008 (0.020)	-0.042 (0.076)	0.005 (0.010)	0.008 (0.046)
Context	0.004 (0.018)	-0.079 (0.076)	-0.001 (0.010)	-0.056 (0.043)
Intervention	0.009 (0.018)	0.012 (0.054)	0.006 (0.010)	-0.002 (0.045)
Forecast \times Context	0.000 (0.026)	0.139 (0.106)	0.000 (0.014)	0.015 (0.067)
Forecast \times Intervention	-0.015 (0.027)	-0.030 (0.092)	-0.011 (0.014)	0.029 (0.061)
Control Mean	0.127	-0.276	0.092	-0.481
Num. Obs.	3516	3516	16 069	16 069

Notes: Standard errors clustered at the study \times person level, with time \times attempt and study fixed effects included. SSPP is the combined academic and expert forecaster types. Omitted group is control who received no additional information. Short-run, long-run 1 and long-run 2 forecasts are used jointly in this analysis.

Table 3 regresses our three accuracy measures on indicators for the randomised information received, for all types of forecast (short-run, long-run 1 and long-run 2), separately for the SSPP sample and the Prolific sample. In short, none of the randomised information has a significant effect on accuracy according to any measure.

In table 4, I further split the sample by running separate regressions on the short-run (panel A), long-run 1 (panel B) and long-run 2 (panel C) forecasts. We might hypothesise that we would see larger effects of the randomised information on short and prior long-run forecasts, as when forecasters make their posterior long-run forecasts, they have the information on the short-run effects which is presumably more valuable. However, we again see very little in the way of statistically significant effects on performance, above what would be expected by chance.

So why might there be consistently null effects of information on forecast accuracy? One possibility is a bounded rationality hypothesis. Bounded rationality suggests that when individuals make decisions, their rationality is limited by the information they have, the cognitive limitations of their minds, and the finite amount of time they have to make a decision. In this case, the task of forecasting causal effects is already complex and cognitively demanding, requiring the individual to hold, process, and synthesize a substantial amount of information in real-time. Given these considerations, it is plausible that the addition of further information, may have exceeded the cognitive bandwidth of the forecasters and thereby not resulted in improved accuracy.

This hypothesis aligns with the "less is more" paradigm, suggesting that when it comes to information provision for forecasters, it might be more effective to focus on delivering a curated subset of crucial information rather than overwhelming them with copious amounts of data. This is in contrast to a "more is more" approach, which the results suggest does not necessarily yield better outcomes.

For the SSPP group, another possible explanation is that academics possibly know the extra information already due to their domain expertise. As such, the randomly provided information does not result in a difference in the information sets between those who receive it and those who do not. However, if this were the case and bounded rationality were not an issue, we would expect to see an effect in the nonexpert Prolific group. The nonexperts do not have domain expertise to begin with so the randomised information provision should result in meaningful differences between their information sets. Given that we still see null effects in the Prolific group, the evidence suggests bounded rationality is a larger concern.⁷

For the intervention information in particular, there may be an external validity problem. [Vivalt \(2020\)](#) finds that there is a large degree of heterogeneity in treatment effects of similar programs. As such, receiving information about the effect of a similar intervention in a different context would not inform a forecaster much about what they should expect the effect to be in their context.

⁷One other explanation is that the randomised information provided is just not informative. I attempted to make the information as informative as possible within the constraint of a survey, but I leave the interested reader to peruse appendix A and judge for themselves.

Table 4: Heterogeneity of information randomisations by forecast type

	SSPP		Prolific	
	Score	NAE	Score	NAE
<i>Panel A: Short-run</i>				
Forecast	-0.015 (0.029)	-0.002 (0.072)	0.002 (0.010)	-0.095* (0.051)
Context	-0.020 (0.027)	0.039 (0.056)	0.002 (0.010)	-0.089** (0.043)
Intervention	-0.021 (0.026)	0.033 (0.055)	0.005 (0.009)	-0.001 (0.044)
Forecast x Context	0.020 (0.036)	0.008 (0.089)	0.003 (0.015)	0.136* (0.073)
Forecast x Intervention	0.011 (0.034)	-0.091 (0.110)	0.001 (0.013)	0.106 (0.066)
Num. Obs.	1240	1240	5052	5052
<i>Panel B: Long-run 1 (Prior)</i>				
Forecast	0.013 (0.025)	-0.035 (0.099)	0.003 (0.013)	0.065 (0.058)
Context	0.015 (0.022)	-0.094 (0.110)	-0.005 (0.012)	-0.048 (0.057)
Intervention	0.022 (0.022)	0.035 (0.071)	0.004 (0.013)	0.021 (0.056)
Forecast x Context	-0.002 (0.035)	0.144 (0.144)	0.002 (0.018)	-0.058 (0.084)
Forecast x Intervention	-0.015 (0.034)	-0.047 (0.118)	-0.018 (0.018)	-0.050 (0.075)
Num. Obs.	1127	1127	5495	5495
<i>Panel C: Long-run 2 (Posterior)</i>				
Forecast	0.025 (0.026)	-0.093 (0.083)	0.010 (0.013)	0.047 (0.058)
Context	0.017 (0.024)	-0.202* (0.114)	0.002 (0.012)	-0.033 (0.054)
Intervention	0.027 (0.026)	-0.031 (0.058)	0.008 (0.013)	-0.029 (0.059)
Forecast x Context	-0.013 (0.034)	0.289* (0.149)	-0.005 (0.018)	-0.024 (0.082)
Forecast x Intervention	-0.041 (0.037)	0.051 (0.107)	-0.014 (0.018)	0.037 (0.078)
Num. Obs.	1149	1149	5522	5522

Notes: Standard errors clustered at the study \times person level, with study fixed effects included. SSPP is the combined academic and expert forecaster types. Omitted group is control who received no additional information. Short-run, long-run 1 and long-run 2 forecasts are used separately in this analysis.

4.2.2 Academic expertise

Now, we focus on just the academic group to see what correlates with accuracy most amongst academics. I consider various measures of self-reported horizontal and vertical expertise. For vertical expertise I consider whether the respondent is a professor against a PhD student or post doctoral researcher. For horizontal expertise, I consider whether the respondent has done research in the same country as the study being forecast, research on the same or similar intervention, research in the broader field (e.g. education or health), and having seen the short-run results before. Additionally, I consider whether the respondent has prior forecasting experience.

In table 5 I focus on the prior and posterior long-run forecasts as academics are more likely to already know the short-run results and these would therefore not be true forecasts. I run a regression including all measures of expertise and a set of separate bivariate regressions. Panel A looks at the effect on score, whereas panel B looks at the effect on negative absolute error.

We see that the coefficients on the different measures of expertise are always positive in the bivariate regressions and often significant. However, in the multivariate regression the coefficients are almost all insignificant (the one exception being research in the same field’s effect on negative absolute error). This is likely due to correlation between the different measures and the more limited sample size in this analysis.

The measures of horizontal expertise seem to be more important for accuracy than vertical expertise. Having done research on the intervention itself and research in the same field as the intervention is associated with statistically significantly better scores and negative absolute errors, between a 30 to 40% increase in accuracy across both measures. The effects of having done research in the context and having seen short-run results before both have a significant effect on the negative absolute error and positive but insignificant effects on the score.

On the other hand, the effect of being a professor is only significant at the 10% level on the score and insignificant for the negative absolute error. The effect of having forecasting experience is insignificant for both measures, and the coefficient size is roughly an order of magnitude smaller than the others, suggesting that forecasting experience is not particularly valuable in the academic sample. This suggests that there might be other factors that drive the difference in accuracy between academics and expert forecasters, although it is worth noting that the expert forecasters likely have significantly more experience forecasting than the even the most experienced academics.

These results contrast with the results of [DellaVigna and Pope \(2018\)](#) who find that “academic rank, field, and contextual experience do not correlate with accuracy” and [Otis \(2021\)](#) who finds that “rank, citations, and conducting research in East Africa [where the forecasted studies took place] ... do not correlate with accuracy.”. One possible explanation for this is that I collect distributional forecasts of the treatment effects whereas the other two studies collect point estimate forecasts. It could be possible that horizontal expertise helps with understanding the range of possible effects more than the mean or mode of the effect distribution. However, this does not fully explain the

Table 5: Effects of expertise for academics on long-run accuracy

<i>Panel A: Score</i>							
Research in context	-0.002 (0.016)	0.021 (0.016)					
Research on intervention	0.024 (0.026)		0.046** (0.022)				
Research in field	0.024 (0.016)			0.030** (0.014)			
Seen short-run results before	0.022 (0.025)				0.033 (0.021)		
Professor	0.020 (0.014)					0.025* (0.015)	
Forecasting experience	0.021 (0.014)						0.008 (0.013)
Control Mean	0.083	0.121	0.113	0.108	0.110	0.115	0.111
<i>Panel B: Negative absolute error</i>							
Research in context	0.031 (0.051)	0.108** (0.048)					
Research on intervention	0.004 (0.064)		0.122*** (0.041)				
Research in field	0.136* (0.082)			0.138** (0.062)			
Seen short-run results before	0.074 (0.058)				0.102* (0.053)		
Professor	0.033 (0.058)					0.063 (0.062)	
Forecasting experience	0.062 (0.077)						0.005 (0.065)
Control Mean	-0.531	-0.337	-0.365	-0.389	-0.363	-0.354	-0.353

Notes: Standard errors clustered at the study \times person level, with time \times attempt and study fixed effects included. Only long-run 1 and long-run 2 forecasts are included, meaning $n = 1870$.

Table 6: Effects of familiarity for non-experts

	Score			Negative absolute error		
	Familiar with context	-0.003 (0.010)	-0.003 (0.010)		0.132*** (0.048)	0.133*** (0.049)
Familiar with intervention	-0.002 (0.008)		-0.002 (0.008)	0.004 (0.038)		0.005 (0.038)
Control Mean	0.104	0.101	0.102	-0.467	-0.449	-0.465

Notes: Standard errors clustered at the study \times person level, with time \times attempt and study fixed effects included. Only non-expert sample and long-run 1 and long-run 2 forecasts used, meaning $n = 11017$.

difference between our results as I also find a positive relationship when looking at the negative absolute error metric, which only depends on the respondents' central point estimate.

4.2.3 Non-expert familiarity

Next, we turn to assessing determinants of accuracy within the nonexpert group. As respondents did not necessarily do research or work in academia, I instead asked them whether they had personal experience with the intervention being studied or in the country in which the study took place. In table 6, I regress these measures of familiarity on the score and negative absolute error, again both in a multivariate and set of bivariate regressions. Even though the nonexperts are exceedingly unlikely to know the short-run effects, I again focus on only the long-run forecasts to maintain comparability with the analysis done in the previous section for academics.

We can see that neither familiarity with the context or the intervention has an effect on the score. Similarly, familiarity with the intervention has no effect on the score. On the other hand, familiarity with the context does have a substantial and significant effect on negative absolute error, improving it by approximately 30%. This effect remains robust in both bivariate and multivariate regressions.

This positive effect of context familiarity but not intervention might again be explained by the limited external validity of the effect of interventions across different contexts (Vivalt, 2020). For example, consider an educational intervention involving the distribution of textbooks in schools. If forecasters are familiar with the intervention but not the context, they might overlook how factors specific to the country, such as the local language, literacy rates, existing curriculum, or teaching methods, might affect the intervention's success. On the other hand, forecasters familiar with the context (i.e., the country and its education system) may be better equipped to make accurate forecasts about the intervention's outcome, even if they aren't as familiar with the intervention itself. This is because they would have a better understanding of how these local factors might interact with the intervention. This suggests that the familiarity with context might allow for a more nuanced understanding of the potential implications and effects of the intervention, leading to

Table 7: Effects of effort on posterior long-run accuracy

	Score		Negative absolute error	
	(1)	(2)	(3)	(4)
Minutes (10)	0.009*** (0.003)	0.005 (0.003)	0.021* (0.012)	0.007 (0.012)
FE: study	X	X	X	X
FE: type		X		X
Num.Obs.	6671	6671	6671	6671

Notes: Standard errors clustered at the study \times person level. Study fixed effects are included in all columns and forecaster type (Academic, Expert Forecaster, or Prolific) fixed effects are included in columns 2 and 4. Only posterior long-run 2 forecasts are included meaning $n = 6671$

more accurate predictions.

It is important to note that the variance in the familiarity with context mostly comes from the boarding school France and financial education Spain studies. Recall that for each survey, on Prolific I recruited 50 people from the USA, 50 from France and 50 from Spain. As such, very few of the Prolific respondents are familiar with the five other countries in which studies took place (Togo, Uganda, Afghanistan, Sierra Leone and Kenya). Since I include study fixed effects in this regression, there will be little residual variation in the familiar with context indicator for these other studies, whereas there will be a mix of people familiar and unfamiliar with France and Spain.

4.2.4 Effort

Effort is likely to be important for forecast accuracy because it implies a deeper engagement with the information and context of the forecasting task. Greater effort, in terms of time spent on the task, may allow for a more thorough analysis of the available information, a more careful consideration of various factors and potential outcomes, and a more deliberate and well-informed decision-making process, all of which can enhance the accuracy of forecasts. However, it's also possible that effort could be a proxy for other attributes such as conscientiousness, diligence, or a serious attitude towards the task, which could themselves be related to forecasting skill.

In order to test this hypothesis, we regressed the accuracy of posterior long-run forecasts on the amount of time spent on the survey, in increments of 10 minutes. To prevent skewness and limit the influence of outliers, we winsorized the duration variable at a maximum of 60 minutes to account for instances where participants might have left the survey open while attending to other tasks.

The results show a positive and significant effect of effort on both accuracy measures when fixed effects for type of forecaster are not included (columns 1 and 3). However, when we include fixed

effects for the type of forecaster (columns 2 and 4), the effect of effort diminishes to about one third or half of its previous size and loses its statistical significance. This suggests that effort is an important mediator in explaining the differences in accuracy between experts and nonexperts. Experts might provide better forecasts simply by trying harder.

4.3 Updating behaviour

In this final section, I delve into the process by which forecasters update their long-run forecasts in response to short-run information. The analysis is narrowed down to the four studies where the outcomes are the same in both the short-run and the long-run - Togo, Afghanistan, Sierra Leone, and Kenya. Moreover, the analysis is restricted to forecasters who altered their posterior long-run forecast upon receiving the short-run information. Table 8 presents the relationship between actual short and long-run effects (column 1) compared to forecasters' priors about those relationships (column 2). Columns 3 to 5 cover how forecasters' long-run posteriors depend on their long-run priors and the observed short-run effects and the uncertainty surrounding both these estimates.

Beginning with Column 1, we observe that there is a robust correlation between short-run and long-run treatment effects for the same outcome in our sample. Specifically, for each unit increase in the short-run effect, we see a corresponding 0.44 unit increase in the long-run effect. This reveals a positive relationship between short-term and long-term outcomes in practice.

Column 2 however shows that when giving their priors, forecasters expect this relationship to be twice as strong as it actually is. Forecasters anticipate an almost one-to-one correspondence in the magnitude of short-run and long-run effects when the actual correspondence as shown in column 1 is half the size. Forecasters essentially expect the effect to persist as-is from the short to the long-run, whereas in actual fact, effect sizes tend to decay, at least in our sample of studies.

Turning our attention to Columns 3 to 5, which look at how the posterior long-run effect is formed, a series of results emerge. From Column 3, we infer that a one unit increase in the prior long-run forecast tends to be associated with a 0.5 increase in the posterior long-run forecast, made after receiving the short-run effect information. This suggests forecasters realise they are overoptimistic about the size of the long-run effect and substantially downscale their posterior forecasts.

Column 4 shows that forecasters do indeed respond to the short-run information. The weight forecasters give to the observed short-run effect is roughly double that which they give to their long-run priors. The difference between the 0.902 coefficient on the observed short-run effect and the 0.451 coefficient on the long-run prior is significant at the 10% level but not the 5% level. This suggests that forecasters are responsive to short-run information but it does not fully replace the information contained in their priors.

Following a Bayesian updating model, we should expect forecasters who are less certain about the long-run effects to update more in response to the short-run information. Similarly, we should

Table 8: How individuals update

	Observed LR	Prior LR	Posterior LR		
	(1)	(2)	(3)	(4)	(5)
(Intercept)	0.021 (0.032)	0.062 (0.029)	0.099 (0.045)	-0.020 (0.041)	0.140 (0.062)
Observed SR	0.441** (0.134)			0.902** (0.271)	-0.614 (0.655)
SR Prior		0.968*** (0.158)			
LR Prior			0.505*** (0.060)	0.451*** (0.044)	0.500*** (0.032)
LR Prior Range					0.075** (0.022)
Observed SR SE					-2.897 (1.312)
LR prior * LR Prior Range					-0.062* (0.026)
Observed SR * SR SE					22.566 (10.339)
Num. Obs.	23	1373	1373	1373	1373

Notes: SR stands for short-run and LR stands for long-run. Standard errors clustered at the study level. Only outcomes which are observed once in short-run and once in long-run are included. Only forecasts where the forecasters updated their long-run forecasts are included. Column 1 regresses the actual long-run effect on the actual short-run effect. Column 2 regresses forecasters most likely first long-run forecast on their most likely short-run forecast. Columns 3 to 5 regress forecasters' most likely second forecast after they receive information on the actual short-run effect, their first long-run forecast, and two measures of uncertainty, the standard error of the short-run effect, and the range between the forecaster's best case and worst case forecast.

expect forecasters who receive more precise short-run information signals to update more from them. Column 5 includes interactions with measures of long-run prior uncertainty and short-run effect uncertainty. The long-run prior uncertainty is measured by the distance between the forecaster’s worst and best case forecast for that outcome, while the short-run uncertainty is measured by the size of the standard error on the short-run effect.⁸

The interaction term between long-run (LR) prior and long-run uncertainty is negative and significant at the 10% level. This indicates that forecasters whose prior long-run forecasts are more uncertain (i.e. have a wider range) place less weight on these long-run priors when providing their posterior long-run forecasts as a result. This is in line with what we would expect from Bayesian updaters.

On the other hand, the interaction term between the short-run effect and its standard error has an insignificant positive coefficient. This suggests that the precision of the short-run forecast does not affect the how much forecasters update their long-run forecasts. A Bayesian updater would place more weight on more precise short-run estimates and we would thus expect to see a negative coefficient here, but we do not.⁹

This replicates and extends the ‘variance neglect’ result of [Vival and Coville \(nd\)](#). Quoting them, “variance neglect is a novel bias closely related to extension neglect, in which individuals place less weight on the precision of results than a Bayesian would”. They show that participants are relatively insensitive to the confidence intervals associated with the results of policymakers. I replicate this result with forecasters being insensitive to the precision of short-run effects. However, this cannot just be due to forecasters neglecting uncertainty altogether, as forecasters are still sensitive to the uncertainty surrounding their prior forecasts. Future work should investigate the relative impacts of different types of uncertainty on forecasters updating procedure.

5 Conclusion

This paper makes contributions to our understanding of long-term forecasts, including their accuracy, the factors influencing them, and how they are updated over time. Utilizing incentivized forecasts from diverse groups, including academics, expert forecasters, and nonexperts, the study analyzes forecasts from seven different randomized experiments, evaluating them against the actual long-run results once these became available.

I find that while domain experts and expert forecasters outperform nonexperts, expert forecasters exhibit higher accuracy across all time horizons. This is likely due to superior calibration skills. I

⁸Note that forecasters were not explicitly given the short-run effect standard error but were instead given the short-run effect size and associated p-value, which provides the same information.

⁹This is under the assumption that the short-run estimates are informative signals of the long-run effects. Fortunately, columns 1 and 2 of table 8 show us that short-run effects are informative signals of long-run effects and forecasters *believe* that short-run effects are correlated with long-run effects respectively

also demonstrate a wisdom-of-crowds effect with aggregated forecasts being more accurate than the average forecaster. However, it's important to note that even these improved aggregate forecasts show considerable room for accuracy enhancement.

A unique element of this study is the experimental variation of the information set provided to the forecasters. It revealed no discernible improvement in accuracy from additional information, including forecasting tips, context-specific and intervention specific information. This supports the "less is more" paradigm, suggesting that effective forecasting may be less about quantity of information and more about the quality and relevance of the information provided within the cognitive bounds of forecasters. I also finds that expertise and familiarity with the context or intervention improve forecast accuracy, as does effort.

I show that forecasters can somewhat accurately predict long-term treatment effects, and these forecasts improve upon the receipt of short-term results. However, they tend to overestimate the correlation between short and long-term effects, suggesting an overoptimism bias. This bias, though partially mitigated upon receiving short-term results, remains a significant factor influencing the forecasts.

The paper's findings have important implications for policymakers and practitioners. Understanding the strengths and limitations of long-term forecasts can inform better decision-making processes, and the evidence that short-term results can provide valuable insights for long-term impacts should be taken into account in policy design.

Future research should aim to explore these areas further, investigate how to improve forecast accuracy and calibration, and better understand the role of various types of expertise and information in the forecasting process. Also, given the finding that the precision of short-run effect estimates is neglected when updating long-run forecasts, more research is needed to understand why this is the case and how it can be mitigated. The exploration of these areas could improve the science of forecasting and increase our ability to plan for the future, making more accurate and informed decisions in a range of fields, from policy to business to scientific research.

References

- Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research.
- Bedoya, G., Coville, A., Haushofer, J., Isaqzadeh, M., and Shapiro, J. P. (2019). No household left behind: Afghanistan targeting the ultra poor impact evaluation. Technical report, National Bureau of Economic Research.
- Behaghel, L., De Chaisemartin, C., and Gurgand, M. (2017). Ready for boarding? the effects of a boarding school for disadvantaged students. *American Economic Journal: Applied Economics*, 9(1):140–164.
- Bernard, D. R. (2020). Estimating long-term treatment effects without longterm outcome data. Technical report, Global Priorities Institute Working Papers.
- Bernard, D. R., Chabe-Ferret, S., de Quidt, J., Fliegner, J. C., and Rathelot, R. (n.d.). How biased are observational methods in practice? accumulating evidence using randomized controlled trials with imperfect compliance. Working paper.
- Bernard, D. R. and Vivalt, E. (n.d.). *Essays on Longtermism*, chapter What are the Prospects of Forecasting the Far Future? Oxford University Press.
- Bloom, N., Mahajan, A., McKenzie, D., and Roberts, J. (2020). Do management interventions last? evidence from india. *American Economic Journal: Applied Economics*, 12(2):198–219.
- Bouguen, A., Huang, Y., Kremer, M., and Miguel, E. (2019). Using randomized controlled trials to estimate long-run impacts in development economics. *Annual Review of Economics*, 11:523–561.
- Bover, O., Hospido, L., and Villanueva, E. (2018). The impact of high school financial education on financial knowledge and choices: Evidence from a randomized trial in spain. *Banco de España Working Paper*.
- Buhl-Wiggers, J., Kerwin, J. T., Muñoz-Morales, J., Smith, J., and Thornton, R. (2022). Some children left behind: Variation in the effects of an educational intervention. *Journal of Econometrics*.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature human behaviour*, 2(9):637–644.

- Campos, F., Frese, M., Goldstein, M., Iacovone, L., Johnson, H. C., McKenzie, D., and Mensmann, M. (2017). Teaching personal initiative beats traditional training in boosting small business in west africa. *Science*, 357(6357):1287–1290.
- Casey, K., Glennerster, R., Miguel, E., and Voors, M. (2023). Long-run effects of aid: Forecasts and evidence from sierra leone. *The Economic Journal*, 133(652):1348–1370.
- DellaVigna, S. and Pope, D. (2018). Predicting experimental results: who knows what? *Journal of Political Economy*, 126(6):2410–2456.
- DellaVigna, S., Pope, D., and Vivalt, E. (2019). Predict science to improve science. *Science*, 366(6464):428–429.
- Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. (2022). General equilibrium effects of cash transfers: experimental evidence from kenya. *Econometrica*, 90(6):2603–2643.
- Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., and Dreber, A. (2019). Predicting replication outcomes in the many labs 2 study. *Journal of Economic Psychology*, 75:102117.
- Fraser, H., Bush, M., Wintle, B. C., Mody, F., Smith, E. T., Hanea, A. M., Gould, E., Hemming, V., Hamilton, D. G., Rumpff, L., et al. (2023). Predicting reliability through structured expert elicitation with the replicats (collaborative assessments for trustworthy science) process. *Plos one*, 18(1):e0274429.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.
- Greenberg, S. (2018). Calibration scoring rules for practical prediction training. *arXiv preprint arXiv:1808.07501*.
- Groh, M., Krishnan, N., McKenzie, D., and Vishwanath, T. (2016). The impact of soft skills training on female youth employment: evidence from a randomized experiment in jordan. *IZA Journal of Labor & Development*, 5(1):1–23.
- Karing, A. (2018). Social signaling and childhood immunization: A field experiment in sierra leone. *University of California, Berkeley*, 2.
- Kerwin, J. T. and Thornton, R. L. (2021). Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures. *The Review of Economics and Statistics*, 103(2):251–264.
- Otis, N. G. (2021). Forecasting in the field. Technical report, Working paper.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association*, 18(6):3045–3089.

Vivalt, E. and Coville, A. (n.d.). How do policymakers update their beliefs? *Journal of Development Economics*. Forthcoming.

A Appendix A - Survey information

A.1 Forecasting tips

Here are some additional forecasting tips to help you make better forecasts.

1. It can sometimes be very difficult to know where to start with forecasting. One general approach is to break down difficult problems into more tractable sub-problems.
2. People are often overoptimistic with their forecasts. Make sure you don't focus on how you would like the world to be, instead focus on how the world actually is. Very large impacts are rare, we should typically expect to see small to moderate impacts from any intervention.
3. People are generally overconfident with their forecasts. Maintain humility and recognise that the world is a noisy, random place and it is difficult to know things with certainty.
4. People naturally take an inside view, focusing on the specifics of the situation or process. Try to complement this by also taking an outside view, ignore the details and make an estimate based on a group of roughly similar previous cases.
5. People often justify or excuse their forecasting mistakes. After you make your first set of forecasts, we will tell you what the true short-term impacts were. Think about how your forecast differs from the truth and if you were wrong, try to work out why that happened and how you can do better in your second set of long-term forecasts.

A.2 Boarding school France

A.2.1 Study description

Introduction:

Boarding schools are a form of education where students live and learn at the school. Although traditionally boarding schools were mostly for upper-class English and American families, there has been recent interest in using boarding schools to improve the progress of disadvantaged students. This study took place in France, with a boarding school which serves relatively high ability students from poor families.

The school:

The school we study was created in 2009 and is to the southeast of Paris and serves those who live in the relatively poor suburbs. Policymakers were concerned that poor school quality, negative influences from peers and bad studying conditions at home could impair the success of motivated students, and hoped that the boarding school could solve these issues by providing a new environment, higher academic demands and more academically able peers. In the first year, the

school was open to those in grades 8 to 10 (ages 14 to 16), but from the second year onwards it was open to those in grades 6 to 12 (ages 12 to 18).

Experimental design:

Motivated students who might benefit from boarding school were identified by school principals and encouraged to apply to the school. In 2009 and 2010, the school was oversubscribed, so a lottery was run, and of the 395 eligible applicants, 258 of them were randomly selected to be in the treatment group and offered a place at the school. Since they were randomly selected, the treatment group is similar to the 137 students who weren't offered a place, the control group, in terms of demographic and educational characteristics. 86 percent of lottery winners enrolled in the boarding school, and 76 percent of winners stayed until the end of the first academic year. Six percent of lottery losers enrolled because one of their siblings had been admitted to the school and five percent stayed until the end of the first year. We ask you to compare the outcomes of lottery winners (treatment group) with lottery losers (control group) regardless of whether winners or losers enrolled in the school or not. This is known as the intent to treat effect.

Applicants characteristics:

The students who applied to the school did slightly better academically than their classmates with an average mathematics mark of 12.6/20 compared to 10.5, and an average French mark of 12.3 compared to 10.5. 46% of them received a means-tested grant in middle school, as opposed to 28% of the student population. 25% of them have a parent who has completed high school and 60% of them speak a language other than French at home. 57% of the applicants were girls. The average age was 14.

Follow-up:

For the short-run outcomes, students were tested on cognitive and non-cognitive outcomes one and two years after they were randomly either offered or not offered a place at the boarding school. We will collect your forecasts for their math test scores, French test scores, hours spent on homework and intrinsic motivation. Our long-run outcomes are measured nine years after students were assigned to the boarding school and they include: whether students had completed high school and received any baccalaureate qualification (they were expected to pass by eight years after assignment or earlier), whether they had applied for higher education, whether they had enrolled in higher education, and whether they had enrolled in an elite higher education institute (Sciences Po, Dauphine or preparatory class), which are selective, unlike most French universities. We have data on all 395 students nine years after the treatment group started at the school.

A.2.2 Context

The school studied is to the southeast of Paris and it serves students from the deprived eastern Parisian suburbs. It was the first boarding school of its type to open. It is now the largest boarding school operating in France.

French secondary education is usually divided into middle school (collège) from age 11 to 15 and high school (lycée) from age 15 to 18. Students prepare for the baccalaureate (baccalauréat / bac) during high school. There are three types of baccalaureate; the general bac, the technological bac, and the professional bac. The general bac is the stepping stone to university degrees, while the technological and professional bac are vocational qualifications.

Normal schools have around 26 hours of schooling per week, with 4-5 hours of French and 4 hours of mathematics, depending on the type of baccalaureate. There are different teachers for each subject. Classes are usually not streamed by ability, there are students of different abilities in each class. Repeating grades is common in France, with 28% of students repeating a school grade at least once by age 15.

Tertiary education in France is divided into public universities and grandes écoles. Grandes écoles require students to attend two years of preparatory classes before admission. Preparatory classes are highly selective as only 5% of each cohort is admitted to preparatory classes. They prepare students for competitive national exams and the ranking in these exams determine which grandes écoles the students attend. Admission to public universities is less selective but high school grades still matter.

A.2.3 Intervention

Short-run

A study of the SEED Boarding school in Washington, DC, found that it raised mathematics test scores by 0.23 standard deviations and raised English test scores by 0.20 standard deviations per year spent in the school. For reference, a rule of thumb people typically use for interpreting standard deviation treatment effects is that a difference of 0 between the treatment group and control group is no effect, 0.1 is a small positive effect, 0.3 is a moderate positive effect and greater than 0.5 is a large positive effect.

Long-run

There is no long-run evidence on the impacts of other boarding schools yet. However, there are studies of charter schools which similarly provide more resources for education, but do not include the boarding component.

A study of The Harlem Promise Academy, a charter school, found that it increased high school graduation rates by 3.7 percentage points and increased college enrollment by 5.5 percentage points but these effects were statistically insignificant. A study of Boston charter schools found that they decreased the high school graduation rate by 0.3 percentage points and increased college enrollment by 1 percentage point, but these effects were also statistically insignificant.

A.3 Personal initiative Togo

A.3.1 Study description

The context:

Business training programs are common in developing countries. They aim to increase the incomes of small business owners by teaching them basic business practices such as record-keeping, stock control, and marketing. A World Bank project in Lomé, the capital city of Togo, aimed to help businesses improve and grow their business by offering two different types of training, a traditional business training and a psychology-based personal initiative program.

The program:

We study the impact of both the traditional business training and the personal initiative training. Both training programs were implemented in three half-day sessions per week over 4 weeks in April 2014, with monthly follow-ups by a trainer for the next 4 months.

The traditional business training program is the Business Edge training program, which is an internationally accredited program developed by the International Finance Corporation. It focused on four core topics: accounting and financial management, human resource management, marketing, and formalization.

The personal initiative training is very different from that of traditional business training programs, focusing on teaching a mindset of self-starting behavior, innovation, identifying and exploiting new opportunities, goal-setting, planning and feedback cycles, and overcoming obstacles.

Experimental design:

1,500 small businesses in Lomé, the capital of Togo, were selected from applicants to the World Bank project. Applicants had to be in business for at least 12 months, have fewer than 50 employees, operate outside of agriculture, and not be a formally registered company. Firms were randomly assigned into three groups, each of 500 firms.

1. A control group, which did not receive any business training.
2. A traditional business training group.
3. The personal initiative training group.

84% of the firms invited to take part in the trainings participated. We focus on the intent-to-treat effect, comparing all of the firms offered the training to all of the control group, even if the firm did not actually participate in the training.

Follow-up surveys:

Four short-run surveys took place over the first 2.5 years from September 2014 to September 2016. For the short-run outcomes, we ask you to forecast the average treatment effect over these four surveys.

The long-term follow-up survey took place from October 2021 to January 2022, 7 years after the intervention.

For both the short and long-run surveys, we ask you to forecast the effects on four main outcomes.

1. Business Survival: whether the person is operating any business
2. Value of last month's sales for that person's main business (coded as 0 if that person no longer has a business)
3. Value of last month's profits for that person's main business (coded as 0 if that person no longer has a business)
4. Personal initiative measured by a seven-item scale of agreement from 1-5 (Likert) with statements like "I actively attack problems" and "whenever something goes wrong, I search immediately for solutions"

A.3.2 Context

Here is an example to help highlight how the traditional training and the personal initiative training are taught differently, even on the same general topic of business finance. Traditional approaches explicitly teach owners to keep business records, explain what the different types of lending products banks offer are, and discuss how to apply for a loan. In contrast, personal initiative training teaches business owners to identify and approach unusual sources of money and not just banks (self-starting behavior), that they should do bootstrapping in order to not need to rely on external funds in the long-term (future-oriented behavior), and that they should not give up if they face financial problems, but develop plan B and Cs (persistent behavior).

The business owners taking part in the study were almost equally split by gender (53% female), had an average age of 41 years, and had an average of 9 years of education. The sample contained a broad mix of industries (27% manufacturing, 48% commerce, and 25% services), with the businesses earning a mean of \$185 in monthly profits before the training, with a mean \$1313 in sales. Firms had a mean of three employees and a median of two.

There was considerable scope for firms to improve their business practices. Only 37% of firms kept accounts books, and only 4.7% had a written budget. Only one-third of firms used advertising or publicity, 71% compared sales performance with objectives, and 66% visited competitors to compare prices or product offerings.

A.3.3 Intervention

A study in Peru randomised whether female microentrepreneurs received 30-60 minute entrepreneurship training sessions during their normal weekly or monthly microfinance sessions, or if they just

made loan and savings payments at these sessions. They found no evidence of changes in key outcomes such as business revenue, profits, or employment, but business knowledge was improved.

A study in the Dominican Republic randomly assigned a three-month standard accounting training versus a simplified, rule-of-thumb training that taught basic financial heuristics. The rule-of-thumb training significantly improved firms' financial practices, objective reporting quality, and revenues. For micro-entrepreneurs with lower skills or poor initial financial practices, the impact of the rule-of-thumb training was significantly larger than that of the standard accounting training.

A study in Tanzania included a randomised business training intervention consisting of 21 sessions. They found that the business training by itself had little effect on business outcomes like sales and profits, but when it was combined with a monetary grant, it had a positive effect for males but no significant effect for females.

A.4 Literacy Uganda

A.4.1 Study description

The intervention:

In Uganda children in lower primary are often taught in English instead of their local language. An Ugandan education firm developed a program to improve literacy by teaching in the local language. There are two versions of the program: full-cost and reduced-cost. Both versions of the program run from grade 1 to 3 (ages 6-8) so children benefit from them for 3 years.

The main feature of the full-cost program is that teachers teach in the local language, Leblango, which the vast majority of the local population speak, instead of English, which students are less competent in. The education firm also provides three residential (i.e. off-site) teacher trainings per year, each about a week long, and five classroom support visits per term to help teachers instruct in Leblango. The teachers are trained to be more engaged with students and move through materials at a slower pace to ensure students learn to read. Teachers are given detailed guides for daily and weekly lesson plans, as well as school supplies such as textbooks designed for the lesson plans, slates (for writing on), chalk, and wall clocks.

The reduced-cost program is mostly the same as the full-cost but there are three main differences. Firstly, the education firm no longer trains the teachers directly, instead they train government staff who are then responsible for training and supporting the teachers. Secondly, the schools only receive two support visits per term. Thirdly, the teachers are no longer given slates or wall clocks. This reduces the cost of the program by 64

Experimental design:

128 schools are taking part in this study. Schools were randomized into three groups:

- Full-cost treatment (42 schools, 3,838 students): these schools got the original program described above.

- Reduced-cost treatment (44 schools, 4,017 students): these schools got the reduced-cost program where government staff trained teachers instead of the education firm’s staff, fewer support visits were made, and slates and wall clocks were not given to the teacher.
- Control (42 schools, 3,755 students): these schools got nothing and continued with their normal teaching methods.

As the schools were randomly divided into groups, we can measure the impact of the full-cost treatment by comparing the average outcome of students in the full-cost treatment schools to the average outcome of students in the control schools (and the same for the reduced-cost treatment). This is known as the intent-to-treat effect and this is what we will be asking you to forecast.

Follow-up surveys:

Students did tests at the end of grades 1, 2 and 3. We will ask you to forecast the short-term effects on Leblango and English test scores at the end of grade 2 and 3. The outcomes we are interested in are Leblango test scores and English test scores. We also ask you to forecast long-term effects 8 years after the program began, or in other words 5 years after the program finished at the end of grade 3. You will again forecast impacts on Leblango and English test scores, and additionally, math test scores, whether the students are still enrolled in school, and whether they have made it to secondary school.

A.4.2 Context

Uganda is a low-income country in East Africa. Primary education in Uganda is free of charge and has relatively high net enrollment rates of more than 90%, but only 60% of students attend secondary school. In 2007, the government introduced a new curriculum which changed the language of instruction in lower primary (grade 1 to 3) from English to the local language (11 different languages throughout the country). This change was only partially implemented with many schools still teaching lower primary in English. Uganda still faces major learning challenges in primary schools. 94% of children in grade 4 in government primary schools are not able to read and understand a simple paragraph in English. Only 80% of grade 7 students leave primary school being able to read a short story, but this is an overestimate of reading skills amongst all students since weaker students are more likely to drop-out before grade 7.

Primary education in Uganda consists of seven years of schooling starting at age six. The net enrollment rate is above 90% but grade repetition and early drop out are still significant problems, with only 60% of students making it to secondary school. The 1997 reform that made school free of charge was successful in getting children into school, but also raised concern about lack of resources and falling school quality.

As well as 94% of children in grade 4 in government schools being unable to read and understand a simple paragraph in English, 54% cannot order numbers correctly, 47% cannot add double digit

numbers and 76% cannot subtract double digit numbers. Some of the problems are due to the poor quality of teaching.

A recent survey found that only 16% of teachers in Uganda have the minimum required knowledge in language, 70% have the minimum knowledge in math and only 4% have the minimum pedagogical knowledge. Few teachers plan their lessons in advance, or introduce and summarize their lessons. Effective teaching time is only 3 hours a day, despite scheduled teaching being 7 hours a day, since teachers are frequently absent from the classroom or from the school entirely, leading to almost half of the classrooms being without a teacher at any given time.

A.4.3 Intervention

Treatment effects on student learning are usually measured in terms of standard deviations (SD), and for reference, a 0.1 SD effect is typically called a small effect, 0.3 SD is a moderate effect, and 0.5 SD is a large effect.

Two of the key components of the interventions just described to you are mother tongue instruction, and teacher training.

A recent study in Kenya compared the effects of:

- Program 1: a teacher training program enabling teaching literacy in English and Kiswahili (the two national languages of Kenya),
- Program 2: a teacher training program enabling teaching literacy in English, Kiswahili, and the mother tongue,
- Control group with no teacher training or change in the language of instruction.

Program 1 (with teacher training but without mother tongue instruction) increased language learning by 0.10-0.35 SD relative to the control group. Program 2 (with teacher training and with mother tongue instruction) increased language learning by 0.37-0.56 SD, relative to the control group.

An older systematic review of 72 mother tongue instruction programs in the USA for non-native English speakers (mostly native Spanish speakers) with rigorous evaluations, found that only 22% of studies found that reading outcomes were better under bilingual education than English only education, 7% for language outcomes, and 9% for mathematics outcomes.

A review of six systematic reviews/meta-analyses, found that teacher training was recommended by four of the six reviews. One meta-analysis found that teacher training produced a 0.12 SD improvement in learning. Interventions involving long-term teacher mentoring or coaching increased learning by 0.25 SD.

A.5 Financial education Spain

A.5.1 Study description

The context:

Financial education is an important part of the curriculum in secondary education in many countries. However, the short and long-run effects of this education are subject to debate. Researchers in Spain launched a program “Finance for All” to improve financial knowledge among high schoolers.

The program:

We study the impact of the Finance for All program. It is a ten hour course given by high school teachers with the objective to help 9th grade students become sufficiently financially literate to make sound financial decisions.

The program covers several areas including: saving and interest rates, budgeting, responsible consumption, types of bank accounts, and specific investment vehicles such as pension funds. Teachers had discretion over the emphasis and order of the topics covered.

Experimental timeline

78 schools participated in the study starting in 2015. They were randomly assigned to a treatment group and a control group. 9th and 10th graders were surveyed, but only 9th graders received the course. The timeline of the study is as follows:

- January to March 2015: 9th graders in treatment schools took the program
- March 2015: All 9th and 10th graders did survey #1
- April to June 2015: 9th graders in control schools took the program
- June 2015: All 9th and 10th graders did survey #2 and an incentivised saving task
- Summer 2020: all (former) 9th and 10th graders did survey #3

We measure the impact of the program in three ways.

- Immediate: compare treatment 9th graders to control 9th graders in March 2015. The treatment 9th graders have just finished the course while the control 9th graders haven’t started it yet. This gives us the immediate impact of the course.
- Fade out: compare treatment 9th graders to control 9th graders in June 2015. The treatment 9th graders received the course 3 months previously whereas the control 9th graders just finished. This tells us how much the treatment 9th graders forgot over 3 months.
- Long-run: compare all former 9th graders to all former 10th graders in Summer 2020. The 9th graders all received the course 5 years before whereas the 10th graders did not. This allows us

to assess the long-run impacts of the course (assuming there are no other differences between the 9th and 10th graders).

Outcomes to be forecast

Survey #1 is the immediate survey in March 2015. The outcomes from this survey we ask you to make forecasts for are:

- Financial knowledge index, standardised score of answers to a financial literacy quiz
- Hypothetical savings choice. Students were asked to choose between receiving 100 euros today and more euros in three weeks' time that ranged from 120 to 180 euros. Please forecast the effect on the percentage of students that choose the later choice.
- Whether students have a bank account or money card.
- Whether students have a source of labor income.

Survey #2 is the fade-out survey in June 2015. The outcome from this survey we ask you to make forecasts for are:

- Financial knowledge index, standardised score of answers to a financial literacy quiz

Survey #3 in summer 2020 is the long-run survey. The outcomes from this survey we ask you to make forecasts for are:

- Financial knowledge index standardised score of answers to a financial literacy quiz
- Financial awareness index made up of familiarity with financial products, number of financial products owned and non-cash forms of saving

A.5.2 Context

Since 2012, every year about 400 high schools in Spain have voluntarily delivered a 10 hour financial education. Although the implementation varies across schools, participant students are typically 9th graders (that is, third grade in compulsory high school). Students complete 9th grade between their ages of 14 and 15. That particular grade was chosen to maximize the potential number of students who receive the material, as 9th grade is the last grade of compulsory schooling with few, if any, electives. Compulsory education finishes at age 16 in Spain in 10th grade.

The sample in this study are high schools applying to participate in the program for the first time during the 2014-2015 academic year. Neither the teaching body nor students in the school had had any previous experience on the contents of the specific program. Out of 169 schools contacted, 78 schools agreed to participate in the study.

The 78 schools cover a large geographic area of Spain but are not representative of all Spanish schools. Seventy percent of schools are located in three regions: Madrid, Aragon and Valencia. 58% of the schools are public and 8% are private schools. The remaining third of schools were “concerted” ones i.e., publicly funded but privately owned and managed.

A.5.3 Intervention

A meta-analysis of 76 randomized experiments with a total sample size of around 160,000 people finds that financial education has positive impacts on financial knowledge and financial behaviors. The average effect on financial knowledge is 0.2 standard deviations while the average effect on financial behaviors is 0.1 standard deviations. The meta-analysis also does not find evidence of a rapid decay in treatment effects, but it also does not find evidence of the sustainability of long-run effects.

A.6 Targeting the Ultra Poor Afghanistan

A.6.1 Study description

Introduction:

Afghanistan is one of the most fragile and conflict-affected countries with 55% of people living below the national poverty line of 112 USD Purchasing Power Parity (PPP) per person per month in 2016. Ultra-poor people in Afghanistan face many constraints. In the target group for this study, 80% of households live below the national poverty line. Five in six households have an illiterate household head. Only 1.5% of households save anything, while two-thirds of households are in debt. These many constraints may give rise to poverty traps, situations from which it is difficult to escape unless multiple constraints are relieved simultaneously.

The Targeting the Ultra Poor program:

“Big-push” interventions are often seen as key for reducing persistent poverty. By combining a package of different assets and services, the Targeting the Ultra Poor program (TUP) aims to lift households out of poverty.

As part of the Targeting the Ultra Poor program, women received the one-off package including: a transfer of livestock – typically cows, and occasionally sheep and goats worth approximately 1,312 USD PPP (357 USD nominal), a consumption stipend of 54 USD PPP (15 USD nominal) delivered in 12 monthly installments, training on livestock rearing and entrepreneurship, access to savings accounts and savings encouragement, facilitation of access to health care services, and coaching through biweekly visits for one year.

Experimental design:

80 villages from four districts in the Balkh province were selected and the ultra-poor households were identified in each village. Public lotteries took place in each village and 491 households were

assigned to the treatment group and 728 households assigned to the control group. In May 2016, the treatment group started receiving the full TUP program for one year as described above, while the control group did not receive any of the components. In practice, 99.5% of the treatment group received at least part of the project's benefits while only 3% of the control group did.

Follow-up:

A first short-run follow-up survey was done from July to October 2018, two years after the initial transfer of livestock and one year after the program ended. A second long-run follow-up survey was done from January-June 2021 to assess the long-run five-year impacts of the program. In both the short and long-run survey, we focus on the same five main outcomes:

1. Consumption per capita in the last month
2. Total household income and revenues in the last month
3. Household total savings
4. Number of cows owned (since this is the primary asset in the TUP program)
5. School enrollment (whether any child in the household aged 6 to 18 years is enrolled in school)

The follow-up survey was successfully completed among 458 treatment households (93%) and 689 control households (95%). The difference in attrition rates across treatment and control groups is not statistically significant.

A.6.2 Context

Ultra-poor households are defined in this study as those that (1) the local village community ranked as ultra-poor in a participatory wealth ranking exercise and (2) met at least three of the six following criteria:

1. Household is financially dependent on women's domestic work or begging
2. Household owns less than 20 decimals (800 square meters) of land or is living in a cave
3. Targeted woman is younger than 50 years of age
4. There are no active adult men income earners in the household
5. Children of school age are working for pay
6. Household does not own any productive assets, based on a pre-defined list

44% of households were ranked as ultra-poor by the village communities, and 15% of these households met at least three of the six criteria.

80% of ultra-poor households lived below the national poverty line. 96% of women were illiterate and 84% of men. 53% of girls and 59% of boys aged 6-19 were enrolled in school. Only 2% of households have any savings and 68% of households are in debt. 69% of women report suffering from major depression. These levels of consumption, human capital, asset ownership and well-being are among the lowest in the world.

The TUP intervention was implemented by the “Microfinance Investment Support Facility for Afghanistan” (MISFA), an organization owned by the government, and Coordination for Humanitarian Assistance (CHA), a local NGO.

A.6.3 Intervention

Existing studies of TUP-like programs mostly come from contexts with less conflict and fragility than Afghanistan.

A multi-site randomised controlled trial study of TUP programs in Ethiopia, Ghana, Honduras, India, Pakistan and Peru found a 0.12 standard deviation increase (an 18-21% increase) in consumption two to three years after the asset transfer. An asset index increased by 0.25 standard deviations, while an income and revenues index increased by 0.27 to 0.38 standard deviations.

Another randomised study of TUP in Bangladesh surveyed households two and four years after the asset transfer. They found that after two years consumption had increased by 6% and 11% after four years. Similarly, earnings went up by 15% after two years and 21% after four years. The total value of cows owned went up by 110% after two years and 122% after four years.

A.7 Social signaling for vaccination Sierra Leone

A.7.1 Study description

The context:

Child immunization is one of the best ways to reduce child mortality. Globally, there is a standard course of 5 vaccines over the first year of life. In Sierra Leone, 99% of children get the first vaccine, but only 69% of children complete all 5 vaccinations by the age of one. This study asks if we can increase timely and complete vaccination by allowing parents to show to others that they have vaccinated their child. Researchers introduced a signal in the form of coloured bracelets that children receive upon vaccination in Sierra Leone.

Experimental design:

60 clinics and their catchment communities were randomized into a treatment group or control group.

In treatment communities, all children received a yellow “1st visit” bracelet when coming for the first vaccine. If a child comes on time for all vaccines up to vaccine five, health workers exchange the yellow bracelet for a green “5th visit” bracelet. If a child comes late, the bracelet is exchanged for an identical yellow “1st visit” bracelet. This provides a signal to the rest of the community about whether a parent vaccinated their child on time for all five vaccines (green bracelet) or failed to do so (yellow bracelet).

The control communities followed a business-as-usual procedure with no bracelets being given out at clinics, so there were no signals about whether a parent vaccinated their child or not.

Experimental timeline and groups:

The experiment was launched in June 2016 and ran until August 2018. Children born during this time in treatment communities received the bracelets. An endline survey captured the treatment effects of the program implementation. A follow-up survey that was implemented in 2020/2021 captured the differences in immunization behavior between the treatment and control groups for children who were born after the experiment ended.

We care about the impact of the treatment on children in four time periods.

1. We measure the main treatment effects (short-term) by looking at children who were fully exposed to the treatment. During this period, 70% of children who came timely for all five vaccinations received the green bracelet. These children were due for their final vaccine before the end of the experiment, in August 2018.
2. We measure the medium-term effects by looking at children who were due for the final vaccine after the end of the experiment and before May 2019. Clinics had a remaining stock of bracelets and therefore could hand out bracelets for longer if they wished to do so. However, nurses were not instructed to do so and monitoring of clinics by the research team had stopped. The remaining stock would on average have lasted for 6 to 9 months, if a clinic had continued the implementation in the same way as they did during the experiment.
3. We measure the long-term effects (before the COVID pandemic) by looking at children who were due for their last vaccine before March 2020.
4. We measure the long-term effects (during the COVID pandemic) by looking at children who were due for their last vaccine before October 2020, and were due for at least one vaccine during the COVID-19 pandemic.

We also want to know whether you think parents who had a child during the experiment were more or less likely to timely vaccinate their next child. We will ask you to separately forecast long-term treatment effects for children who were born after the experiment ended and:

- Group 1: had an older sibling who was part of the experiment, or

- Group 2: did not have a sibling that was part of the experiment.

On the one hand, parents in the first group may have formed a habit of vaccinating their children further and more timely, or improved their views about the benefits of vaccination. Similarly, parents in the second group observed the bracelets on other children and may have updated their beliefs about how socially desirable it is to timely and fully vaccinate their child in order to be viewed as a good parent. On the other hand, parents may feel demotivated to vaccinate their child if they are no longer rewarded with bracelets.

Outcomes

We will first ask you to forecast an outcome related to the implementation of bracelets: Bracelet handout by clinics: the proportion of children in treatment communities that received a green bracelet, when coming timely for all five vaccinations. This outcome is in percentage so your forecast must be between 0 and 100. No children in control communities received green bracelets so the proportion in treatment communities is the same as the treatment effect. We will then ask you to forecast the treatment effect on two different outcomes:

- Timely vaccinations: the proportion of children that completed all five vaccinations on time by the age of 3, 4, 5, 6 and 11.5 months respectively.
- Completed vaccinations: the proportion of children that completed all five vaccinations by 12 months of age, regardless of whether a vaccine was taken on time or not.

These outcomes are both in percentage points so your forecasts must be between -100 and +100.

A.7.2 Context

Sierra Leone is one of the poorest countries in the world, ranked 181 out of 188 in the Human Development Index and has one of the highest infant and under five mortality rates. One in every 11 Sierra Leonean children dies before reaching the age of one and one in every 7 does not survive to her fifth birthday.

A child under the age of one needs to receive five routine vaccinations. The first vaccination is due at birth, followed by a three-dose series of vaccines protecting against diphtheria, tetanus, and pertussis (DTP). The three doses must be given one month apart, with the first given at 1.5 months and the last dose given at 3.5 months of age. The fifth vaccine is the first of a two-dose series protecting against Measles and is due at nine months of age.

Vaccines are free of charge and readily available in clinics and readily available throughout Sierra Leone. At baseline, fewer than 14% of study clinics reported having a stock-out of one or more vaccines. Moreover, 94% of communities were aware of the health benefits of vaccines and 79% believed that negligence of parents is the most common reason for missed or delayed vaccination.

A.7.3 Intervention

This was the first field experiment designed to test whether social signaling could increase childhood vaccination. However, there are other studies looking at other ways to increase the uptake of vaccinations in low-income settings.

One randomized experiment in India found that offering 1 kg of raw lentils for each vaccination visit and a metal plate upon completion of the full series, increases complete vaccination rates from 18 to 39 percent, an increase of 21 percentage points.

A randomized experiment in Nigeria found that giving conditional cash transfers when parents brought their infants to get vaccinated for their first-year vaccines, increased the proportion of children who were fully immunized by 27 percentage points. It also improved the proportion of infants who received the measles vaccine within one month of the recommended age by 33 percentage points.

Another experiment in Nigeria found that in the control group where mothers were paid \$0.05 to get a single tetanus vaccine, vaccine take-up was 55%. If mothers were instead offered \$2, take-up was 74%, 19 percentage points higher. Furthermore, those who were offered \$5.30 had a vaccine take-up of 83%, a 28 percentage points increase.

A.8 General equilibrium effects of cash Kenya

A.8.1 Study description

The intervention


This study was done in collaboration with a large NGO which gives cash transfers directly to very poor households. To be eligible for a transfer, households had to live in homes with thatched roofs. Eligible households were given three cash transfers over six months, totalling \$1871 USD (in 2014 PPP purchasing power parity adjusted dollars), or 87,000 Kenyan Shillings. This is equivalent to 75% of the average household expenditure in the eligible households. Less poor households are ineligible and receive no cash.

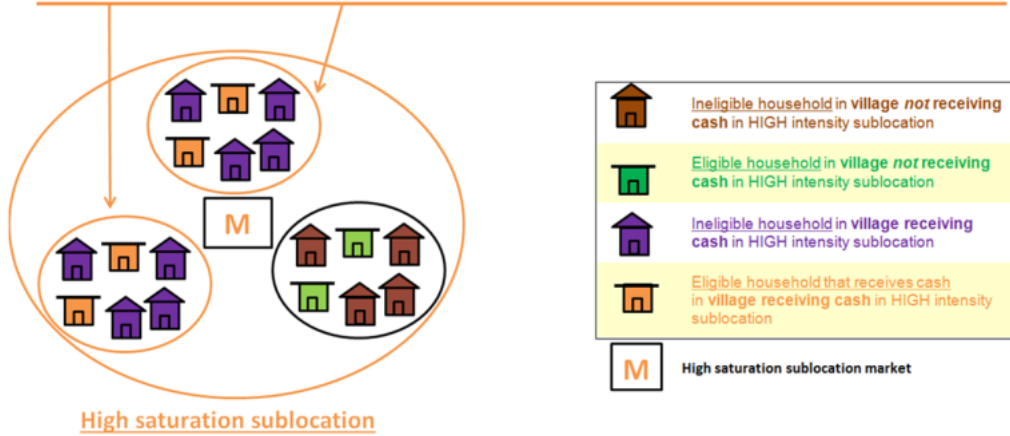
Experimental design


Villages close to each other are grouped in 68 sublocations of between two and fifteen villages. 33 sublocations are assigned to high saturation and 35 to low saturation. In the high saturation sublocations, two thirds of villages are assigned to be treated and in low saturation sublocations, one third of villages are assigned to be treated. This means that villages in high saturation sublocations have more nearby villages that receive cash transfers.

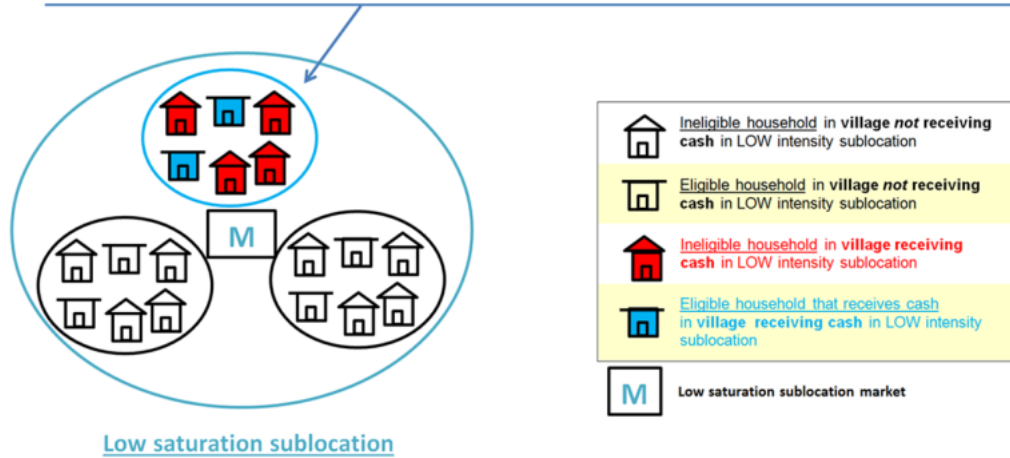
Within treated villages, eligible households receive the cash transfer of \$1871 USD PPP while ineligible households receive no cash transfer. However, the ineligible households may still be affected by the fact that their eligible neighbors received cash transfers.

Within control villages, neither eligible nor ineligible households received cash transfers. However, they may still be affected by whether people in nearby villages received cash transfers.

In **high intensity sublocations**, 2/3 of the villages in the sublocation were randomly assigned to have very poor (eligible) households  receive cash transfers.



In **low intensity sublocations**, 1/3 of the villages in the sublocation were randomly assigned to have very poor (eligible) households  receive cash transfers.



We study the effects on both eligible and ineligible households separately. We use households in the low saturation, untreated villages as the benchmark. This means there are six types of household we want to consider the effects on.

Eligible households

1. Eligible households in treated villages in high saturation sublocations (orange flat roof: received cash, eligible neighbors received cash and many nearby villages received cash).

2. Eligible households in treated villages in low saturation sublocations (blue flat roof: received cash, eligible neighbors received cash and few nearby villages received cash).
3. Eligible households in untreated villages in high saturation sublocations (green flat roof: did not receive cash, eligible neighbors did not receive cash and many nearby villages received cash).

These are the eligible households. We compare their outcomes to the eligible households in untreated villages in low saturation sublocations (white flat roof); in other words households who did not receive cash, their eligible neighbors did not receive cash and few nearby villages received cash.

Ineligible households

1. Ineligible households in treated villages in high saturation sublocations (purple sloped roof: did not receive cash, eligible neighbors received cash and many nearby villages received cash).
2. Ineligible households in treated villages in low saturation sublocations (red sloped roof: did not receive cash, eligible neighbors received cash and few nearby villages received cash).
3. Ineligible households in untreated villages in high saturation sublocations (brown sloped roof: did not receive cash, eligible neighbors did not receive cash and many nearby villages received cash).

These are the ineligible households. We compare their outcomes to the ineligible households in untreated villages in low saturation sublocations (white sloped roof); in other words, compared to households who did not receive cash, their eligible neighbors did not receive cash and few nearby villages received cash.

Study characteristics

The average village had approximately 100 households in it. There were 65,385 households across the 653 villages. Before the cash transfers were made, average annual consumption per household in the study area was \$2727 USD PPP.

Follow-up surveys

The first household follow-up survey took place on average 1.5 years after the study began and 11 months after the last cash installment was transferred. The long-run follow-up survey took place 6.5 years after the study began. All eligible households received the cash transfer, but not all of them were surveyed. In each village, 8 eligible households and 4 ineligible households were randomly chosen to be surveyed, resulting in a final sample size of 8,242.

A.8.2 Context

The study took place in Siaya County in western Kenya, which is a rural area that borders Lake Victoria. The area was selected for the study based on its high poverty levels. The area is relatively densely populated with 395 people per km² compared to an average of 91 in Kenya. The main road of Kenya that connects the port of Mombasa to the capital of Nairobi runs through Siaya County, helping the area be economically connected. The population is majority Luo, the second largest ethnic group in Kenya.

The average household in the study has 4.3 members, of which 2.3 are children. The average survey respondent was 48 years old at the start of the study and had about 6 years of schooling. 97% of households work in agriculture, 49% of them also work for a wage, and 48% are also self-employed.

A.8.3 Intervention

Short-run

Cash transfer programs have been studied across the world. A meta-analysis that combines the results of studies from 11 different programs found that cash transfers increased consumption by between 2.5% and 18%.

The study most similar to this one was a randomised controlled trial that also took place in Kenya in a nearby county. This study found that similarly large cash transfers increased monthly consumption by \$36, from \$157 to \$193, nine months after the transfer. There was also a significant positive effect on average asset holdings, moving from \$495 to \$797, a \$302 increase.

Although the design of the experiment was not identical, the researchers were also able to estimate a spillover effect on ineligible households by comparing ineligible households in treated villages to similar ineligible households in untreated villages. They found that there was no statistically significant effect on ineligible households' consumption or asset holdings.

Long-run

The same study described above also followed up with participants three years after the transfer. They find similarly large positive effects on eligible households, with consumption increases of \$47 and asset holding increases of \$416. However, they now find negative spillover effects on ineligible households, with ineligible households in treated villages having monthly consumption decreases between \$29 and \$38, and asset holding decreases between \$0 and \$140.