Centre of
Excellence
for Development
Impact and Learning

April 2023
Research project paper 9

How Biased are Observational Methods in Practice? Accumulating Evidence Using Randomised Controlled Trials with Imperfect Compliance

David Rhys Bernard, Gharad Bryan, Sylvain Chabé-Ferret,
Jon de Quidt, Jasmin Claire Fliegner, Roland Rathelot

London International Development Centre
UKaid from the British people
Oxford Policy Management

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL research project paper

CEDIL research project papers are the research outputs of the projects supported by CEDIL. They include impact evaluations, systematic reviews, evidence and gap maps and exploratory papers on research methods and approaches.

About this research project paper

This paper was submitted to CEDIL by the “How Biased are Observational Methods in Practice? Accumulating Evidence Using Randomised Controlled Trials with Imperfect Compliance” S.218 team.

Please direct any comments or queries to the corresponding author, Roland Rathelot at Roland.Rathelot@ensae.fr.

Suggested citation: Bernard, D. R., Bryan, G., Chabé-Ferret, S., de Quidt, J., Fliegner, J. C., Rathelot. R. 2023. How Biased are Observational Methods in Practice? Accumulating Evidence Using Randomised Controlled Trials with Imperfect Compliance, CEDIL Research Project Paper 9. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford. Available from: https://doi.org/10.51744/CRPP9

This project was funded by the Centre of Excellence for Development Impact and Learning (CEDIL), supported by UK aid from the UK Government. The views expressed in this research project paper do not necessarily reflect the UK Government's official policies or CEDIL.

All content is the sole responsibility of the authors. CEDIL has not commissioned a professional copy-edit of this paper. Any errors and omissions are the sole responsibility of the authors.

CEDIL Final Report: How Biased are Observational Methods in Practice? Accumulating Evidence Using Randomised Controlled Trials with Imperfect Compliance [*]

David Rhys Bernard Gharad Bryan Sylvain Chabé-Ferret
Jon de Quidt Jasmin Claire Fliegner Roland Rathelot

March 17, 2023

Abstract

Consider a policy maker choosing between programs of unknown impact. She can inform her decision using observational methods, or by running a randomised controlled trial (RCT). The proponents of RCTs would argue that observational approaches suffer from bias of an unknown size and direction, and so are uninformative. Our study treats this as an empirical claim that can be studied. By doing so we hope to increase the value of observational data and studies, as well as better inform the choice to undertake RCTs. We propose a large-scale, standardised, hands-off approach to assessing the performance of observational methods. First, we collect and categorise data from a large number of RCTs in the past 20 years. Second, we implement new methods to understand the size and direction of expected bias in observational studies, and how bias depends on measurable characteristics of programmes and settings. We find that the difference between observational estimators and the experimental benchmark is on average zero, but the resulting observational bias distribution has high variance.

Keywords: randomised experiments, matching, selection bias, public policy evaluation, conditional independence assumption.
JEL: C21, C26, D04.

[*] Bernard: Paris School of . Bryan: London School of Economics. Chabé-Ferret: Toulouse School of Economics. de Quidt: Institute for International Economic Studies, Stockholm University. Fliegner: University of Manchester. Rathelot: Institut Polytechnique de Paris. We gratefully acknowledge financial support from IPA and CEDIL. The views expressed herein are those of the authors and do not necessarily reflect those of any institution. All errors are our own.

Contents

1 Introduction

1.1 Description of the project

Consider a policymaker deciding whether to expand a social program, and who wishes to base her decision on whether the program has a positive causal impact on key measurable outcomes. Suppose she has access to observational data that records outcomes for users and non-users, allowing her to form an observational estimate of the program's impact. Her immediate concern is selection bias: differences in outcomes might reflect differences in who chooses to take up the program rather than the causal effect of program participation. Acknowledging this concern, she has two broad pathways forward. First, she can do her best with the data, using information about participants to form an estimate that controls for any observable differences. This will work if treatment assignment is unconfounded conditional on the observables she has, but many believe this is unlikely in most practical cases and argue that the resulting estimates have a remaining *observational bias* of unknown direction and magnitude. Alternatively, she can opt to run an RCT, gathering experimental data at an additional expenditure that allows unbiased estimation of the program's effects. What should she do?

If she is willing to assume that the RCT is always reliable, and estimates the impact that she is interested in, and she knows its cost, the size of the observational bias is what she needs to make her choice.[1] If she knew the observational bias was negligible, there would be no reason to run an RCT. Similarly, if there were a known positive bias (i.e., a bias that favours program adoption), or a known negative bias, she could simply adjust her observational estimate. The harder case, is when the sign and magnitude of bias are uncertain. If she knew its distribution, she could adjust her confidence in the observational estimate, and make a decision on that basis. But without further information on the distribution of bias, she must simply take a guess.

Our goal is to take the guesswork out of this decision. We treat selection bias as a parameter to be estimated, and the numerous imperfect compliance RCTs run in the last 20-30 years as a large untapped data source. An imperfect compliance RCT (ICRCT) is one in which some "treated" individuals do not take up the offered program or some "control" individuals do take it up. Each ICRCT can yield both an experimental and an observational estimate of program impact in the same setting, and hence allows estimation

[1] An RCT may be biased for many reasons, and in many cases with estimate at LATE. We discuss these issues in more detail below.

of the extent of the observational bias.

We differentiate between two types of ICRCTs: eligibility and encouragement designs. In eligibility designs, the treatment group is eligible for a program while the control group is prevented from taking it up. In this situation an experimental estimate will recover the Average Treatment Effect on the Treated (TT). In encouragement designs, individuals in both treatment and control groups can choose to take up the program. In that situation an experimental estimate will recover the Local Average Treatment Effect (LATE). With data from an ICRCT we can compare those that select into the program to those that do not to form an observational estimate, which can be compared to the experimental LATE or TT. An important interpretational point is that we assume throughout that the policy maker is interested in the estimate that can be formed from the RCT. The accuracy of this assumption will always depend on the context. If the policy makers is considering expanding a program by encouraging people to take it up, or lowering its cost, then the LATE will be the estimate she wants, but if she is hoping to make a program available to people then the TT is the relevant measure. Some argue that the TT is often more policy relevant, and the majority of our collected data estimate the TT and not a LATE (28 vs. 17 studies). To be precise, the here estimated observational bias, is not a bias in the statistical sense but the difference between an observational estimate and an experimental benchmark. Combining data from multiple ICRCTs using meta-analytic tools, we can provide evidence on the distribution of the observational bias.

1.2 Objectives of the project

We use this basic method to reach four objectives. First, we provide guidance on how to minimise bias in the analysis of an observational dataset. Intuitively, we can calculate bias estimates for a range of observational methods, and select those methods that tend to generate smaller biases. Second, we estimate the distribution of selection bias and how well different methods perform in reducing the bias. We do this using meta-analytic techniques. Thirdly, we explore how the bias distribution changes across different contexts to see whether we can successfully predict the bias in a new context. Finally, we provide guidance on how to evaluate the tradeoff between using the observational estimate, and conducting an RCT. This ultimately depends on how much additional precision the RCT would provide, and the social value of that improvement in precision.

1.3 Contribution to the literature

Our work builds on LaLonde (1986) who compares experimental and observational estimates of the impact of a job-training program, finding that observational methods failed to recover the experimental estimates. Since LaLonde many more RCTs have been run, and stronger observational methods have been developed, which allow us to overcome three challenges in prior work: it is small-scale, not standardised, and not hands-off.

LaLonde (1986) is small scale: it only uses data from one RCT, creating two limitations. First, the external validity of the project is limited as we only learn about the bias in the context of a job-training programme in the US. Second, the work cannot recover the distribution of bias across contexts meaning that while LaLonde is critical of observational methods, it is possible that observational methods perform better elsewhere. By taking our large-scale approach and looking at many contexts simultaneously, we will be able to draw more general conclusions.

LaLonde and most following work use an additional non-standardised, non-experimental comparison group to assess observational methods. This introduces two other potential biases: the new sample may not be perfectly comparable to the experimental sample, and the survey instruments or variable definitions may not match perfectly. Our standardised approach means we always apply observational and experimental methods within the same dataset. This avoids the problems and cost of searching for additional data in every context.

There has been a 20-year debate over LaLonde's results, in part because the approach is not hands-off. Dehejia and Wahba (1999, 2002) re-analyse the same data with different methods and find that observational estimates are closer to experimental ones. However, Smith and Todd (2005) find that Dehejia and Wahba's results are sensitive to technical details such as the trimming method used. Observational methods require many choices to be made by the researcher and these degrees of freedom can generate conflicting results even within one dataset. We tie our hands to the greatest extent possible by standardising these choices across all datasets and automating the observational methods.

Previous studies have addressed some of these issues, but ours is the first to solve all three at once. Attempts at solving the small-scale problem typically perform meta-analysis of prior studies of observational bias, for example, Glazerman et al. (2003) for job training, Wong et al. (2017) for education, Forbes and Dahabreh (2020) in medicine and Chaplin et al.

(2018) for regression discontinuity design (RDD). However, meta-analysis is constrained by the methodological choices of the original study authors. For example, some studies in Chaplin et al. (2018) use parametric RDD and some non-parametric, complicating the comparison. In single context studies of bias, these choices might have been tailored and so overstate the performance of observational methods.

Others have noted the standardisation problem, and focused on imperfect compliance RCTs as a way to solve this issue, e.g., Arceneaux et al. (2006) who assess matching methods in the context of a voter mobilisation experiment, and Gill et al. (2016) who focus on the intention-to-treat estimate and assess observational methods in the context of charter-school lotteries, but still suffer from the small-scale problem. More recently, Franklin et al. (2020) try to reproduce the results of RCTs in medicine by using large databases on patient-level from US commercial and Medicare payers. It is still unclear whether their outcome measurements are consistent between RCTs and observational methods.

Hands-off approaches are new. Duflo (2018) looks at the performance of double machine learning estimators in the context of scholarships in Ghana and explicitly calls for more LaLonde style research. We are able to incorporate a broad set of modern observational methods into our analysis including popular double machine learning methods.

The closest study to ours is Gordon et al. (2019), who evaluate two observational methods in 15 Facebook advertising experiments. Our broader focus is more relevant to policymakers and practitioners, and the development context, and will allow us to draw more general conclusions about the interpretation of existing evidence and generation of new evidence. In the development literature, Pritchett and Sandefur (2015) compare experimental and non-experimental estimates of the effects of microcredit on consumption and profits, using six studies. The estimates differ although the direction does not seem to be systematic. They also compare the magnitude of these biases to those incurred by extrapolating estimates across contexts, and find they tend to be smaller

1.4 Policy relevance

The number and policy influence of RCTs has exploded in the last two decades, confirmed by both proponents, like Banerjee et al. (2016b), and critics, like Ravallion (2020). More debated is the extent to which RCTs have crowded out observational studies and whether this is good or not. Both sides call for more evidence about the relative merits of the two families of methodologies. Highlighting the demand for this research, Ravallion (2020) writes that there is "need for more research on the distribution of estimates from O[bservational] S[tudie]s, such as by comparing OLS estimates with RCTs for the same setting".

Our project is relevant to policy makers who care about the impact of their policies. The project will help them better understand the strengths and weaknesses of different methods used in existing and new evaluations and therefore enable them to make better decisions. Given the large-scale and standardized approach of this project, it will be beneficial for policy makers and evaluators in all countries and thoese considering many different topics.

Policy makers are often presented with conflicting evidence from different methodologies on the effectiveness of a programme without a clear way to aggregate them. Furthermore, the hierarchy of evidence, with randomised experiments sitting above observational studies may not be clear in general. This project suggests an approach to inform policy makers how much they should trust the results from observational studies in the particular area they are working on. We find that the observational bias has large standard deviation. Taking this source of uncertainty into account it is very hard to learn about treatment effects from observational studies. For example, education researchers often see a 0.2 standard deviation increase in tests scores to be an extremely large improvement, but we estimate the standard deviation of the observational bias to be at least this large, implying that an observational study that sees a 0.2 standard deviation improvement cannot be sure that this is not the result of remaining observational bias. This suggests that in our setting, RCTs provide more reliable evidence of the programme's impact - potential threats to our identification are discussed in the text.

The project might also be useful for funding organisations, in order to assess which methodologies are more cost-effective to fund in each context. It could also benefit development organisations, such as Oxfam, who use routinely methods relying on an unconfoundedness assumption to measure the quantitative impact of the programmes they implement. This project supports these organisations to have more valid measures of the impact of their programmes. At a higher level, it also helps to inform how budgets for evaluation should be spent.

1.5 Innovation and relevance to CEDIL

There has been much debate in the development evidence community about the relative merits of RCTs contra other methods. The strongest proponents would argue that we should discard non-experimental evidence in cases where we have RCT evidence available. This is because the observational approach suffers from a bias of unknown size and direction and therefore is uninformative. The RCT estimate on the other hand is unbiased and will therefore give us the right answer on average.

In practice, policymakers need to use evidence from multiple sources when making policy decisions and cannot disregard all other sources of evidence. The decision to run an RCT might also be affected by other factors such as budget, time-constraints and feasibility. This is recognised in CEDIL's objective (Programme of Work 3, Evidence Synthesis) of helping policymakers effectively use evidence from multiple sources – both existing studies and new data collection – to make decisions. Our study is able to provide guidance on how two important different types of evidence, randomised controlled trials and observational studies, should be weighted against each other in the policymakers calculus. By observational studies, we mean studies based on observational data as opposed to experimental data. Specifically we consider estimators based on unconfoundedness in a cross-sectional setting (e.g. matching, regression, machine-learning techniques). We do not consider cases in which panel data may be available allowing for other ewstimation methods. Being able to quantify how right or wrong the observational evidence might be relative to an RCT is important in knowing how much weight to place on it, both in cases where policymakers have RCT-level evidence and where they do not.

2 Methodology

In this section we first explain the econometric theory of how we identify the bias. Then we describe our selection criteria, search process and data collection for the datasets we use to estimate the bias. Finally we describe how we synthesis the data to produce empirical bias estimates to match our theoretical constructs.

2.1 Identification of bias

The key to our approach is to use the same data to form an experimental and an observational estimator of the treatment effect. Randomised experiments with imperfect compliance offer that possibility under assumptions that would not go beyond classical assumptions for both estimands to be valid separately which we show in this section.

Imperfect compliance can typically be subsumed under two forms: *encouragement design* and *eligibility design*, also called two-sided imperfect compliance and one-sided imperfect compliance. We call an encouragement design a randomised experiment where individuals in both the treatment and control groups can decide to take up the treatment. An eligibility design can be considered a subset of the encouragement design and describes a randomised experiment where the control group can be prevented from taking up the treatment. In this section, we describe how to identify both, the experimental and non-experimental estimand, in each design and how to identify the resulting bias as the difference between the experimental and non experimental estimand. Note that the eligibility design can be considered a subset of the encouragement design which is why we are going to focus on the identification for the former.

Formally, we are in the typical setup of programme evaluation where we are interested in estimating the causal effect of a binary treatment on an outcome. In randomised experiments with imperfect compliance, individuals $i = 1, ..., N$ receive a randomised offer R_i to take up a programme. They can then choose to take-up the programme or not. We call the randomised offer the *manipulation variable* R_i where $R_i = 1$ if the individual is randomised into the treatment group. Let us denote the actual *programme participation* by D_i where $D_i = 1$ if the individual chooses to participate. If D_i was equal to R_i we would be in a situation of perfect compliance not suitable for our identification strategy. We will refer to the groups defined by R_i as the treatment or manipulated group whereas we call the individuals defined by D_i the participants. In the potential outcomes framework, one typically assumes that an individual can have an outcome under both states of the world: a potential outcome under programme participation and one without denoted by Y^d for $d \in \{0, 1\}$ respectively. Both are functional forms of the covariates X_i. Hereafter, we assume absence of interactions among units:

Assumption 1 (SUTVA) $(Y_i^1, Y_i^0) \perp D_j$ *for* $i \neq j$.

While this is not a trivial assumption, it is one of the most pervasive ones in the economic literature. Most programme evaluation papers rely on SUTVA, and only a subset of specific papers attempt to identify the presence and measure the magnitude of spillovers, or others consequences of the failure of SUTVA.[2]

The observed outcome is a function of the actual programme participation $Y_i = Y_i^1 D_i + Y_i^0(1 - D_i)$.

2.1.1 Encouragement design

In an encouragement design an individual can choose to participate in the programme regardless of their randomisation status: if she is randomised in, she might choose to participate or not and likewise if she is randomised out because it might be impossible to prevent the individuals in the control group from taking up the treatment. The participation probability conditional on the manipulation arm is in both cases non zero: $P(D_i = 1|R = r) > 0, r \in \{0,1\}$.

Let us denote the potential participation given the manipulation status by D_i^r. Realised programme participation is a function of the potential participation $D_i = D_i^1 R_i + D_i^0(1 - R_i)$ which leads us to split the individuals into four groups: (i) always takers (AT) are individuals who always choose to participate regardless their randomisation status ($D_i^1 = D_i^0 = 1$); (ii) never takers (NT) are individuals who never participate regardless their randomisation status ($D_i^1 = D_i^0 = 0$); (iii) compliers (C) are individuals who comply to the manipulation - they participate if they are randomised in and they don't otherwise ($D_i^1 - D_i^0 = 1$); and (iv) defiers (D) are individuals who do the opposite of what the encouragement in form of the manipulation induces them to do ($D_i^1 - D_i^0 = -1$).

On top of SUTVA, the standard assumptions to recover the Local Average Treatment effect ($LATE$) are summarised as follows.

Assumption 2 (Valid Encouragement Design) *Let Y_i^{dr} with $d, r \in \{0,1\}$ be the potential outcome given manipulation and participation status for individual $i = 1, ..., N$.*

1. *Independence: $(Y_i^{dr}, D_i^1, D_i^0) \perp R_i$*

2. *Exclusion restriction: $Y_i^{dr} = Y_i^d$*

3. *First Stage: $E(D_i^1 - D_i^0) \in]0,1[$*

[2] Maintaining SUTVA might be especially problematic in cluster RCTs where identification for the experimental estimator will come from between-cluster variation, while identification with the observational estimator will come from within-cluster variation.

4. *Monotonicity:* $D_i^1 - D_i^0 \geq 0$ *for all* i

Key to identification is a good randomisation of the manipulation to ensure independence. For the exclusion exclusion restriction, random assignment is not enough and would need to be discussed for each included study. We try to only include studies for which we consider it likely that the exclusion restriction holds. The monotonicity assumption leads us to assume away defiers. Even though this does not hold by design, the monotonicity assumption is often considered to be plausible in randomised experiments if they are designed in a way that treatment assignment increases the take-up incentives. The first stage ensures that there is a non zero share of compliers but that we still do have a degree of imperfect compliance. If the first stage were to be equal to 1, we would be in the case of perfect compliance. The fist stage can easily be verified by computing the participant shares on each manipulation arm. Under the classical assumptions 1 and 2, the experimental Wald estimand recovers the $LATE$ (see e.g. Angrist and Pischke (2009)):

$$E = \frac{E\left[Y_i|R_i = 1\right] - E\left[Y_i|R_i = 0\right]}{P(D_i = 1|R_i = 1) - P(D_i = 1|R_i = 0)} \tag{1}$$

The $LATE$ is formally defined as $LATE = E[Y_i^1 - Y_i^0|D_i^1 - D_i^0 = 1]$.[3]

In order to form the observational estimand, note that we can build two separate observational estimands, one in the treated group (NE^1) and one in the control group (NE^0). One of our contributions is to show that, for encouragement designs, a Wald-like combination of the observational estimand from each treatment arm recovers the $LATE$ under standard conditional independence conditions summarised below.

Assumption 3 (Valid Non-Experimental Wald-Ratio) *For each manipulation arm* $r \in \{0, 1\}$*:*

1. *Conditional Independence:* $(Y_i^1, Y_i^0) \perp D_i|X_i, R_i = r$

2. *Common Support:* $0 < P(D_i = 1|X_i, R_i = r) < 1$

The conditional independence assumption ensures that the participation is as good as random once we control for confounding covariates. The common-support assumption ensures that we have comparable individuals in the groups of participants and non

[3] To be precise, the Wald estimand is the empirical counterpart of the Wald estimator that we can estimate using the experimental data. The causal effect that want to identify is the LATE parameter.

participants.

Under these assumptions, we can estimate an observational or *non experimental estimand* on each treatment arm which assuming we use an estimand with the objective to recover the Average Treatment Effect on the Treated $TT^r = E(Y_i^1 - Y_i^0 | D = 1, R = r)$ (see e.g. Angrist and Pischke (2009)).

$$NE^r = E[Y_i^1 | D_i = 1, R_i = r] - E[E[Y_i^0 | X_i, D_i = 0, R_i = r] | D_i = 1, R_i = r] \tag{2}$$

By the balancing property one can show that covariates can be replaced by the propensity score $P(X_i) = P(D_i = 1 | X_i)$.

Under standard conditional independence conditions in both the treated and control groups as summarised in Assumption 1 and 3, we propose an observational estimand which recovers the $LATE$ as a weighted ratio of both NE^rs: [4]

$$NE = \frac{NE^1 \Pr(D_i = 1 | R_i = 1) - NE^0 \Pr(D_i = 1 | R_i = 0)}{\Pr(D_i = 1 | R_i = 1) - \Pr(D_i = 1 | R_i = 0)} \tag{3}$$

It might seem frustrating not to be able to recover the bias of the observational method for the TT parameter, but the experimentally-induced variation in an encouragement design only reveals the causal effect for compliers, not for the entire group of participants. As we noted above, the relative value of the LATE and TT estimates depends on the question the policy maker is interested in. For example, if she is interested in subsidising gym access to reduce obesity, she will want the LATE, but if she wants to know what will happen if she make gyms available to people in a new place she wants the TT. While we recognize that it is plausible that observational bias for LATE and TT estimates are very different, in the absence of evidence showing that we think it is reasonable to start investigating the issue by assuming that the policy maker does not have strong beleifs and is willing to treat evidence on observational bias in either as equally relevant.

This leads to the following result:

[4] To be precise, all these population quantities can be estimated using data from observational studies if we have an estimator for the TT's on each randomization arm. In that sense, we are actually proposing an estimator. We explain how to estimate the quantities in the following of this section. The theorem shows that this population counterpart of an estimator based on observational data identifies the LATE parameter.

Theorem 1 *Under Assumptions 1, 2 and 3,* $NE = E[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1] = E$*.*

We provide the proof in Appendix A.

The theorem shows that under standard assumptions, the here proposed observational estimand also recovers the LATE (as opposed to the LATE estimand traditionally recovered using experimental variation). This way, we can compare estimands based on observational data with estimands based experimental data. Of course, this requires that the standard assumptions for the experimental estimand and for the observational estimand are fulfilled.

2.1.2 Eligibility design

In the eligibility design, the control group individuals are prevented from participating.[5] We can form an experimental estimand E based on Equation 1 with $P(D_i = 1|R_i = 0) = 0$ and a single observational estimand on the treatment arm group $NE = NE^1$ according to Equation 2. Note that we can also obtain the same observational estimand using the observational LATE in equation 3 since this is equal to NE^1 when $Pr(D_i = 1|R_i = 0) = 0$.

Key differences in the identification are that the monotonicity assumption holds by design and that we obtain directly two comparable effects: both estimands, the experimental and observational one, recover under classical assumptions for a valid eligibility design 2 and for a valid observational method under conditional independence assumptions 3 on the manipulation arm $R_i = 1$ the same treatment effect, namely the average treatment effect on the treated TT.

Corollary 1 *Under Assumptions 1, 2 and 3, for* $P(D_i = 1|R_i = 0) = 0$*,*

$$NE = E[Y_i^1 - Y_i^0 | D_i = 1] = E.$$

The proof follows directly from the Bloom result (see e.g. Angrist and Pischke (2009)).

[5] There is also a reverse eligibility design case where $\Pr(D_i = 1|R_i = 1) = 1$ and $\Pr(D_i = 1|R_i = 0) > 0$ (i.e. there is perfect compliance in the treatment group but imperfect compliance in the control group) but none of the RCTs we use in this paper follow this design.

2.1.3 Bias of the observational method

We are able to identify an experimental and an observational estimand which recover both (i) the TT in case of an eligibility design and (ii) the $LATE$ in case of an encouragement design.

If we consider the experimental estimand as our benchmark and assuming that only the assumptions for the observational estimand are violated, the difference between both estimands yields the bias of the observational method: $B_{NE} = NE - E$. Due to the importance of the unconfoundedness assumption, in this paper, we mainly attribute the bias of observational methods to a failure of this assumption. However, it is possible to decompose the bias into two components: a bias due to the failure of the unconfoundedness assumption B_u and a bias due to the failure of the common support assumption B_{tts} (see Appendix B for further details).

Assuming that not only the assumptions for the experimental estimand but also the assumptions for the observational estimand hold, it follows from Theorem 1 that the bias vanishes.

Corollary 2 *Under Assumptions 1 and 2,*

$$B_{NE} = NE - E = B_u + B_{tts}.$$

If Assumption 3 holds as well, $B_{NE} = B_u = B_{tts} = 0$.

The Corollary follows directly from the Theorem 1.

This identification has a caveat: any difference between both estimands is entirely attributed to a failure of the assumptions required for the observational estimand. If any assumption to recover the experimental estimand is violated, then the difference between both estimands might be difficult to interpret. We will discuss the implications of a failure of this assumption in section 2.5.2.3.3.

2.2 Selection criteria

We use imperfect compliance RCTs for this project. An imperfect compliance RCT is an RCT where the randomised manipulation does not perfectly determine program take-up, for instance, if take-up depends on a choice by the participant(s). In other words, if there is a correlation of less than 1 between assignment to treatment and take-up of treatment then there is imperfect compliance. We make a distinction between three types of imperfect compliance RCT:

1. Eligibility designs: RCTs in which there is imperfect compliance in the manipulated group only. No-one takes up the program in the non-manipulated group and only some of the members of the manipulated group take up the program.

2. Reverse Eligibility designs (these are rare): RCTs in which there is imperfect compliance in the non-manipulated group only. Everyone takes up the program in the manipulated group, but some of the members of the non-manipulated group also take up the program.

3. Encouragement designs: RCTs in which there is imperfect compliance both in the manipulated and the non-manipulated groups. There is a positive but not 100% take up of the program in both groups and usually greater take-up in the manipulated group. Designs are only feasible encouragement designs if take-up of the program can be observed in both the manipulated and the non-manipulated group.

A study is included in our analysis if all of the following are present:

- Variable(s) measuring the experimental manipulation(s) (e.g. eligibility/encouragement for a program). Usually these will be binary, if not we have to transform them into a binary variable.

- Variable(s) measuring take-up of a program of interest. Usually these will be binary, if not we have to transform them into a binary variable.

- At least one outcome variable that we believe is influenced by the program

- Imperfect compliance with the experimental manipulation in program take-up.

We can use RCTs with any of the three types of imperfect compliance described above and we can handle imperfect compliance at the individual or cluster level.

2.3 Search

Our search domain was all of the datasets from the J-PAL and IPA Dataverses. Our final search of the two Dataverses was on 3rd August 2022, at which point there were 207 datasets available.

We used the J-PAL and IPA Dataverses for a number of reasons. Firstly, these are the two most prominent organisations that run randomised controlled trials in development economics. Secondly, these repositories had a large number of studies available on them so we expected to find many suitable datasets for our project.

Other repositories we considered including:

- International Initiative for Impact Evaluation Development Evidence Portal
- DIME data collection
- Impact Evaluation Surveys Collection
- David McKenzie's website
- MDRC
- Mathematica
- REES (within ICPSR)
- openICPSR
- NCES / IES
- Head Start Impact Study
- Journal Websites

These repositories were less well structured than the J-PAL and IPA repositories. They might also be less representative of the development economics literature.

2.4 Data collection

We scraped the meta-data from all 207 of the studies on both IPA and J-PAL Dataverses. This includes author names, paper title, year of publication, DOI where available, and so on. After we scrape the meta-data, each study goes through a three-step screening process from the initial scrape to being included in our study.

Pre-screening. At *Level 1*, for each repository, we pre-screen all projects to eliminate those datasets that are definitely not suitable for our analysis – often non RCT data or RCTs with full compliance.

Screening. At *Level 2*, we perform an in-depth screening of the projects that could proceed from *Level 1* to *Level 2*. The objective of this step is to get an understanding of the information potentially available in the dataset to a) once again eliminate papers that are not deemed suitable after further scrutinizing. This could for example happen if the authors do not collect a measure of imperfect compliance. b) To obtain a set of basic information about the paper such as the available outcome measures, the randomization and participation variables and other metadata relevant for *Level 3*.

Data preparation. The papers that pass *Level 2*, move on to *Level 3*. We now collect information from the dataset itself to prepare the econometric analysis. The goal of this stage is to prepare a clean dataset for each project where outcome, treatment, treatment uptake and control variables are stored. This step involves *data cleaning* (which we describe in more detail in section 2.5.1) Eventually, each project dataset stores the relevant variables in a harmonized way with one row for each specification ready to be read by our bias estimation code package.

Of the 207 datasets:

- 123 of them passed level 1 screening
- 85 of them passed level 2 screening
- 40 of them passed level 3 screening and are in our analysis

2.5 Data synthesis

The data synthesis follows two main steps. Firstly, we clean and merge the raw datafiles associated with each study to produce an analysis dataset for that file and collate the information on outcome, treatment, take-up and covariate variables in that dataset. Secondly, we run our bias estimation code on each of the analysis datasets to produce bias estimates for each outcome-treatment combination that are later used in the meta-analysis

2.5.1 Data cleaning

The process for cleaning each dataset is similar. First we download the data from the repository and identify the names of key variables and store them in a spreadsheet.

- Outcomes
- Treatment status
- Take-up measures
- Baseline covariates
- Strata
- Clusters
- Weights

For the outcomes, we use all of the variables that are included in outcome tables in the associated paper. For the baseline covariates, we use all possible variables available in the dataset that are either measured before treatment or are time-invariant.

We convert the raw data to a single wide dataset by merging and reshaping. We ensure variables are correctly classified as numeric or categorical. We create dummy variables to indicate whether baseline covariates have missing values and replace the missing values with the median for numeric variables or the mode for categorical variables.

Next we use the information we collected on the key variables to estimate many biases for each wide dataset. To do this, we first estimate an experimental version of each treatment effect. This is the ground truth which the observational estimates try to replicate. Then we use three different observational estimators to produce three different observational treatment effect estimations. The difference between the observational estimate and the experimental estimate is an estimate of the bias of the observational method. We first explain how we identify the bias, and we then explain the different estimators used.

2.5.2 Bias estimation

We are aiming at following a procedure that is as hands-off and normalised as possible.

2.5.2.1 Determining the design. First, we use an automatic procedure to detect the design (encouragement or eligibility or full compliance). Eligibility designs are defined so that there are never-takers but no always-takers. Encouragement designs are defined so that there are both never-takers and always-takers. Full compliance designs have no never-takers and no always-takers. Where never-takers or always-takers are present in the data but are few, there is a decision to make. If never-takers or always-takers are too few, the precision of the non-experimental estimator will be too low in the corresponding branch, and the analysis will be uninformative. In these cases, we decide to ignore these never-takers or always-takers, recoding their treatment variables. This leads us to accept a small amount of bias in the estimation of our parameter of interest, but we believe that this it is a better solution to discarding all datasets where this issue may happen. When should we consider that never-takers or always-takers are too few? The threshold of the share of never-takers/always-takers depends on the sample size. We use a power calculations to determine the threshold: if a relatively large minimum detectable effect (equal to 30% of a standard deviation) cannot be recovered using this dataset, we recode the never-takers/always-takers. We choose this value for the MDE, as it corresponds to a 90%/10% split between participants and non-participants in a treatment arm of 1000 individual observations. With our procedure, with a sample size of 1000, in treatment arms with more than 90% participants, all observations are considered to be participants. In control arms with less than 10% participants, all observations are considered to be non participants. The threshold increases with sample size, so as to reflect increasing precision on each arm.

2.5.2.2 Design matrix. Our default approach is to include all of the control variables that are available in the underlying RCT dataset. The estimation of treatment effects uses several machine learning-based econometric methods. For all these methods, we build the same matrix of covariates. We build dummy variables for each dichotomous variable. For the rest of the covariates, we include them linearly without higher order polynomials or interactions unless already prevalent in the original dataset. We then feed these covariates into several machine learning-based econometric methods in order to compute experimental and observational estimates of the treatment effects.

2.5.2.3 Estimators. We feed these covariates into several machine learning-based econometric methods in order to compute experimental and observational estimates of the treatment effects.

2.5.2.3.1 Experimental estimator $\hat{E}$.

Chernozhukov et al. (2018) propose a set of estimators that rely on orthogonalization and sample splitting / cross-fitting to overcome regularization bias and overfitting.

- *Partially Linear Instrumental Variable regression* This estimator follows the procedure outlined in Bach et al. (2021) and Chernozhukov et al. (2018). The model takes the form:

$$\begin{aligned} Y - E * D &= g_0(X) + \zeta, & \mathbb{E}(\zeta \mid R, X) = 0, \\ R &= m_0(X) + V, & \mathbb{E}(V \mid X) = 0. \end{aligned}$$

  Where U_i and V_i are error terms with $\mathbb{E}[U_i|R_i, X_i] = \mathbb{E}[V_i|X_i] = 0$. The estimation procedure works as follows:

  1. Split the sample randomly into k subsamples.
  2. Using $k-1$ subsamples, use a ranger learner to make the best predictions of Y and D using X: $\hat{g}_0(X)$ and $\hat{m}_0(X)$.
  3. Using the remaining subsample, compute $\tilde{Y}_i = Y_i - \hat{g}_0(X_i)$ and $\tilde{D}_i = D_i - \hat{m}_0(X_i)$.
  4. Using the remaining subsample, perform the partially linear IV regression of $\tilde{Y}_i$ on $\tilde{D}_i$ and R_i: obtain $\hat{E}_1$. The intuition is that the effect of the covariates on the take-up are partialled out but not the instrument.
  5. Repeat the last three steps using different splits of the k susamples to obtain k estimates of $\hat{E}_K$.
  6. Average the different estimators: get the DML2 estimator of $\hat{E} = \frac{1}{K}\sum_1^K E_k$.

2.5.2.3.2 Three non-experimental estimators $\widehat{NE}$.

We apply three different non-experimental estimators, the first two of which are based on machine-learning algorithms.

- *Post double selection lasso* Belloni et al. (2014):

  1. Lasso regression of D_i on X_i
  2. Lasso regression of Y_i on X_i

3. Run an OLS estimator of Y_i on D_i, controlling for the controls selected in both regressions.

- *Partially linear regression* following Bach et al. (2021) and Chernozhukov et al. (2018). The model takes the form:

$$\begin{aligned} Y &= NE * D + g_0(X) + \zeta, & \mathbb{E}(\zeta \mid D, X) &= 0, \\ D &= m_0(X) + V, & \mathbb{E}(V \mid X) &= 0. \end{aligned}$$

  The estimation procedure works as follows:

  1. Split the sample randomly into k subsamples.
  2. Using $k-1$ subsamples, use a ranger learner to make the best predictions of Y and D using X: $\hat{g}_0(X)$ and $\hat{m}_0(X)$.
  3. Using the remaining subsample, compute $\tilde{Y}_i = Y_i - \hat{g}_0(X)$ and $\tilde{D}_i = D_i - \hat{m}_0(X)$.
  4. Using the remaining subsample, perform the partially linear regression of $\tilde{Y}_i$ on $\tilde{D}_i$ and $\hat{g}_0(X)$: obtain $\widehat{NE}_1$.
  5. Repeat the last three steps using different splits of the K subsamples to obtain $\{\widehat{NE}_k\}_{k=1...K}$.
  6. Average the different estimators: get the DML estimator of $\widehat{NE} = \frac{1}{K}\sum_1^K \widehat{NE}_k$.

  The approach is similar to the experimental estimator but without an instrument: it relies on orthogonalization and sample splitting to overcome regularization bias and overfitting. Compared to Belloni et al. (2014) or Chernozhukov et al. (2017) the method relies on weaker assumptions through sample-splitting. Intuitively, the effect of the covariates on the take-up are partialled out. The nuisance function is estimated via random forest learner with a number of trees of 100. We use the DML2 algorithm described in the paper.

- *With-without comparison*. This is simply a naive comparison of the outcomes of those who took the treatment against those who did not take the treatment.

  1. Run a regression of Y_i on D_i without including any X_i variables

2. The coefficient on D_i is the estimated treatment effect

Note that based on this estimator, we can obtain a measure of selection bias (see Appendix B).

For the eligibility design, we estimate the observational estimator on $R_i = 1$, the manipulation arm. For the encouragement design, we estimate two observational estimators on $R_i = 1$ and $R_i = 0$ and construct the Wald-Type Ratio.

All these estimators aim at estimating either the LATE or the TT consistently if the assumptions are fulfilled. The methods, including the experimental estimators, are not unbiased in the statistical sense, which means that we require larger samples for consistency. We estimate the variance for the experimental and observational estimators using the variance estimators proposed by the respective authors. In case of the encouragement design, for the observational LATE, we provide weighted standard error estimates based on the standard errors estimated on each randomization branch.

2.5.2.3.3 Estimation of the Bias We compute our estimator of the bias of the observational methods by taking the difference between the observational estimator and the experimental estimator: $\hat{B} = \hat{E} - \widehat{NE}$. We estimate the variance of the bias term as the sum of the variance of the experimental and observational estimates adjusted by the correlation of both estimators across datasets. We normalise our bias terms by expressing them in units of effect sizes. In order to do so, we divide them by the standard deviation of the outcome in the control group.

In order to meaningfully interpret the results, we take into account the social desirability of the outcome. By a socially desirable outcome, we mean one which we want to see more of, or equivalently, one we would prefer to see a positive treatment effect on. This may be outcomes like years in school or income. Socially undesirable outcomes are those we want to see negative treatment effects, for example the presence of disease. For each paper, we collect information on whether or not a positive or a negative treatment effect is desirable (ceteris paribus). A positive bias on an desirable outcome means that the social desirability is overestimated. A negative bias on an undesirable outcome also means that the social desirability is overestimated. Therefore, we flip the sign of the bias of undesirable outcomes so that a positive bias always represents an overestimation of the program's social desirability.

All of these estimates are sent to a MySQL database for later use. The MySQL database is automatically updated with more recent estimates as they come.

It is important to emphasize that this estimation procedure relies heavily on the quality of the experimental estimator. As laid out in section 2.1.3. While we would assume that most assumptions for the LATE or TT estimand hold in the context of a controlled experiment, particularly the exclusion restriction needs to be argued for in social science RCTs as doubly blind RCTs are typically not feasible. [6] A failure of this assumption would lead to an inconsistent estimation of our experimental benchmark. Another problem would arise if the First Stage assumption only holds weakly. This would lead us to a weak instrument setup and one would expect that the variance of the IV estimator that we chose to recover this estimand is larger inducing more uncertainty about the estimated experimental effect. Furthermore, any violation of the exclusion restriction would be exacerbated by a weak instrument.

We try to mitigate this problem by collecting data from high-quality RCT repositories and collecting information on potential failures of these assumptions (see Appendix C and D). The objective is to eventually conduct robustness analysis using this information.

One could also argue that in case of the eligibility design, the observational estimator uses a smaller sample than the experimental estimator which might lead to a larger variance. One way to address this problem would be to take a random subsample for the experimental estimator of the same size as the experimental one. Yet this would affect the initial randomization procedure.

2.5.2.4 Inference on the bias As a results of the estimation procedure, we now have a dataset of bias terms and their corresponding standard errors: $(\hat{B}_s, \hat{\sigma}_{B_s})$. The amount of variation in the raw distribution of $\hat{B}_s$ reflects at the same time the effect of true selection bias B_s and the effect of sampling noise: both $\widehat{NE}_s$ and $\hat{E}_s$ are indeed estimated with error. In order to estimate the amount of variation in the data that is purely due to selection bias, we model the bias of observational methods as follows:

$$B(\hat{\theta}^o_s) = g(Z_s; \beta) + \eta_s + \epsilon_s, \tag{4}$$

[6] The monotonicity assumption required for the identification of the LATE might also fail in some settings, yet in most cases a well carried out RCT mitigates this risk and this is only required for encouragement designs.

with g a parametric function known up to parameter vector β, η_s a random effect term distributed with a density $f_\eta(.;\gamma)$, known up to a vector of parameters γ and ϵ_s being sampling noise, distributed as a $\mathcal{N}(0, \hat{\sigma}_{B_s})$, which is not restrictive in view of the Central Limit Theorem.

The simplest version of such a model that we can think of is a random effects meta-regression with $g(Z_s;\beta) = \mu$ and $\eta_s \sim \mathcal{N}(0, \tau^2)$. With that model, we say that the bias of our observational estimator is on average equal to μ, but that, for each study, an additional term $\hat{B}_s$ is drawn with $\hat{B}_s \sim \mathcal{N}(0, \tau^2)$.

It is well known that a random effects meta-regression model can be estimated efficiently and consistently using Restricted Maximum Likelihood, which will give estimates of $\hat{\mu}$ and of $\hat{\tau}$ and of their respective standard errors. More complex models can of course be estimated, for example one with sector and/or country dummies, or with types of covariates dummies, etc.

In order to not only account for the imprecision of the bias estimates but also for the within paper correlation, we apply the correlated Hierarchical Model with Robust variance estimation following Pustejovsky and Tipton (2021) and Viechtbauer (2010) described in section 3.3.2.

3 Results

3.1 Description of the studies

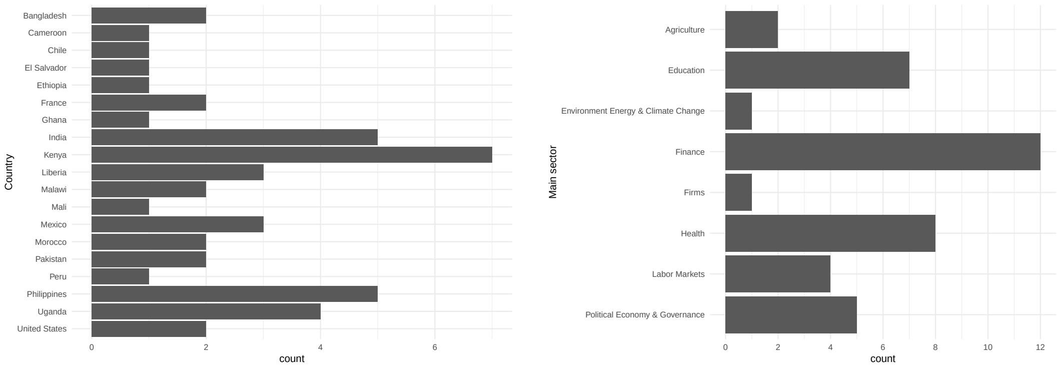
Currently, we include in our study the data from the following papers: Galiani et al. (2016), Ashraf et al. (2006), Blattman et al. (2014), Giné et al. (2010), Dupas and Robinson (2013a), Dupas and Robinson (2013b), Dupas (2011), Guiteras et al. (2015), Ashraf et al. (2009), Duflo et al. (2015), Dupas et al. (2016), Cohen et al. (2015), Ambler et al. (2015), Dupas et al. (2018b), Blattman and Dercon (2018), Karlan et al. (2017), Hicken et al. (2018), Karlan et al. (2016), Blattman et al. (2020), Romero et al. (2017), Chong et al. (2015), Karlan et al. (2019), Beaman et al. (2013), Banerjee et al. (2010), Devoto et al. (2012), Hanna et al. (2016), Khan et al. (2016), Mohammed et al. (2016), Banerjee et al. (2007), Banerjee et al. (2016a), Braconnier et al. (2017), Dupas et al. (2018a), Finkelstein et al. (2012), Angelucci et al. (2015), Crépon et al. (2015), Blattman and Annan (2016), Blattman et al. (2017), Bruhn et al. (2018), Pons and Liegey (2019), Bryan et al. (2014).

We provide information on each of these 40 studies in Appendix C. For the remainder of the analysis, we discard outliers by removing the top and bottom 1% estimates, across all studies. Furthermore, we remove outcomes with ambiguous social desirability.

3.1.1 Overview

In figures 1a to 2b we plot four characteristics of the studies used in our analysis: country, sector/topic, journal and authors.

Figure 1: Characteristics of the studies used in analysis



(a) Countries

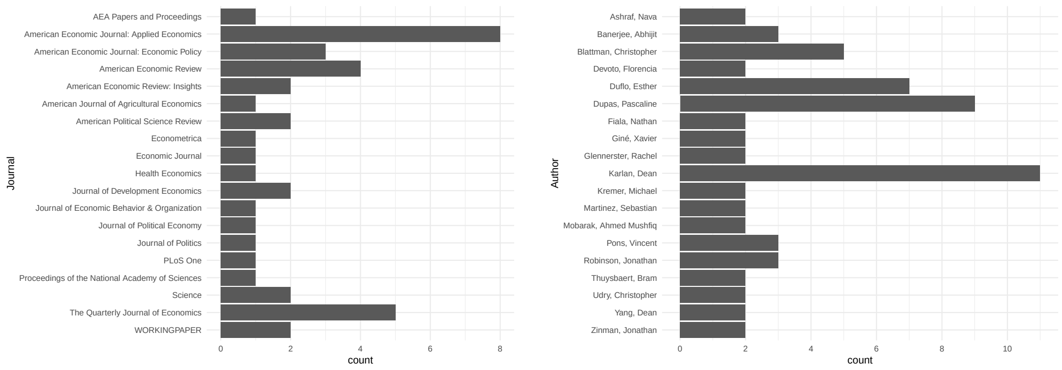
(b) Main sector

Figure 1a shows that our studies come almost entirely from developing countries, as this is where J-PAL and IPA mostly operate. We have studies from Africa, South America, and Asia, as well as North America (USA) and Europe (France). Studies from countries with IPA or J-PAL hubs are strongly represented in our dataset, similarly to the development economics literature more broadly. Kenya appears the most in our analysis, with India, the Philippines, Uganda and Mexico also being highly represented.

We use JPAL's eleven sectors to categorise our datasets into a topic of study, shown in figure 1b. The most represented sectors are finance, education and health, all common areas of study within development economics. Note that we only have one study in each of environment and firms, so when we look at heterogeneity according to these sectors later, we should not expect the results to be particularly generalisable, whereas we can be more confident in the sectors with more results.

The studies in our analysis also represent those published in top journals as shown in figure 2a. We have nine papers from top-five journals in economics: four papers from The Quarterly Journal of Economics, three from the American Economics Review, and one

Figure 2: Characteristics of the studies used in analysis (cont.)



(a) Journals

(b) Authors with > 1 studies

each from Econometrica and the Journal of Political Economy. Eight of our studies come from the American Economic Journal: Applied Economics. This journal publishes many randomised controlled trials and enforces its data availability policy which means it is the most strongly represented journal in our study. We also have a few studies published in non-economics journals, signifiying our breadth of coverage: American Political Science Review, the Journal of Politics, PLoS One, PNAS and Science. We do not cover many development field journals, only having two studies from the Journal of Development Economics.

Finally, figure 2b shows authors which appear at least twice in our dataset. Almost all of these authors are prominent development economists, with Dean Karlan, Pascaline Dupas and Esther Duflo all appearing most frequently, with at least seven papers each. This suggests that even though not many of our studies are published in development field journals, they are still likely to be representative of the development economics literature more broadly.

3.2 Quality of the studies

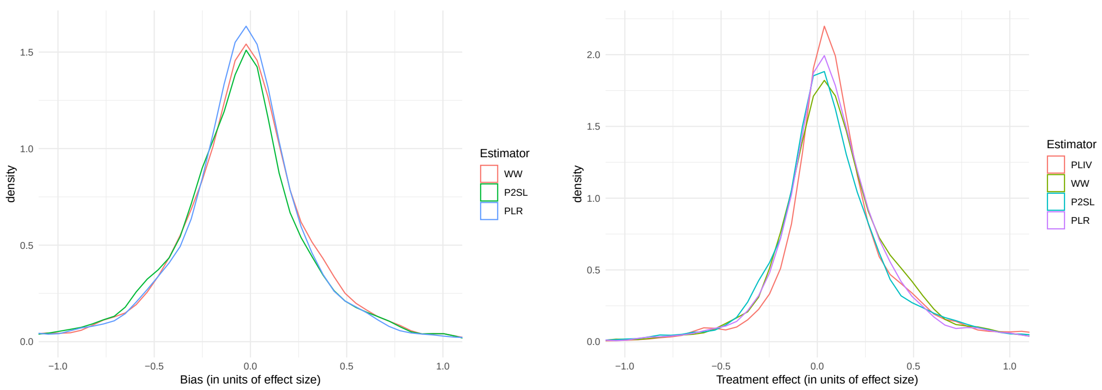
We rate the quality of each of our 40 studies in Appendix D. Overall, we think the quality is high. All of our datasets are RCTs, run by J-PAL or IPA and almost all are published in high quality journals. This suggests that the RCTs provide data that estimate causal effects consistently. We should note that not all studies originally estimate the $LATE$ / TT but only focus on the Intention-to-Treat Effect. In such cases, there is no need to argue for the exclusion restriction. Subject to the caveats mentioned in section 2.5.2.3.3, we believe that overall, our comparison between observational estimators applied on the subset of the data and the RCT causal effect should provide us with a good estimate of the bias of observational methods.

3.3 Effects of the intervention

In this section we describe the main results we find regarding the distribution of the observational bias of the different sectors, across different dimensions of heterogeneity. First we start by plotting and providing summary statistics of our raw observational bias estimates. However, these should be interpreted with caution as they are just estimates of the observational bias and contain noise. As such, next we run a meta-analysis of the observational bias. As described in section 2.5.2.4, the meta-analysis deals with two problems: (1) our bias estimates are just estimates with varying precision, and (2) we have multiple bias estimates per study which are likely to be correlated with each other. To deal with (1) the meta-analysis gives more weight to more precisely estimated biases and to deal with (2) the meta-analysis allows for intra-study correlation when estimating clustered standard errors.

3.3.1 Raw observational bias analysis

Figure 3: Distributions of observational bias and treatment effects resulting from the different estimators


(a) Raw observational bias

(b) Treatment effects

Figure 3a and Table 1 summarise the distribution of our raw observational bias estimates for our three observational estimators, with-without (WW), post double selection lasso (P2SL) and partially linear regression (PLR). As a reminder, the estimate of an observational bias for a given outcome is the estimated observational treatment effect minus the estimated

Table 1: Summary statistics for raw bias of estimators

	P2SL	PLR	WW
Mean bias	-0.055	-0.071	-0.061
Std dev bias	1.423	0.486	0.474
Mean absolute bias	0.501	0.294	0.299
% abs bias < 0.1	0.306	0.322	0.309
% abs bias 0.1 - 0.3	0.336	0.374	0.361
% abs bias 0.3 - 0.5	0.154	0.148	0.170
% abs bias > 0.5	0.205	0.156	0.160

experimental treatment effect, and the units of bias are normalised to a common effect size by dividing by the control group standard deviation.

Visually, the distribution of the estimators are all roughly similar from the -1 to +1 range. The distributions are centered at approximately -0.06. This means that on average, the estimated observational treatment effect is 0.06 standard deviations smaller than the experimental treatment effect. This small difference in effect sizes is not often likely to be economically significant.

However, this is merely an average. Furthermore, it is an average of some positive biases and some negative biases and it may be that the positive and negative biases average out. If we want to know what bias we should expect in any single draw from the distribution we also need to look at higher order moments such as the variance or standard deviation. Although the standard deviation of the three estimators look the same in from the -1 to +1 range, if we look at table 1 we can see that although WW and PLR have a standard deviation of approximately 0.48, the standard deviation of P2SL is 1.51, over three times the size. This is because P2SL has some extreme outliers which aren't shown which inflate the standard deviation. This suggests that the PLR estimate performs better than the P2SL estimate as PLR does not produce these extreme outliers. We can also look at the % abs bias > 0.5 row in table 1 and see that while 20.5% of the P2SL estimates have a large bias greater than 0.5 standard deviations, only 15.6% of the PLR estimates are of an equivalent size.

By looking at the peak of the distributions in figure 3a we can see some further evidence of this. The density of P2SL is lower than that of PLR at 0. This means that PLR has more estimates that are the same or very similar to the experimental estimate.

Table 2: Summary statistics for TE of different estimators

	PLIV	P2SL	PLR	WW
Mean TE	0.168	0.113	0.097	0.107
Std dev TE	0.454	1.370	0.332	0.331
% TE < 0	0.307	0.404	0.361	0.360
% TE 0 - 0.1	0.241	0.194	0.213	0.189
% TE 0.1 - 0.3	0.240	0.216	0.249	0.245
% TE > 0.3	0.212	0.185	0.178	0.207

Another way to look at the results is to look at the distribution of standardised treatment effects of the observational estimators and compare them to the distribution of treatment effects from the experimental estimator. This is what we do in figure 3b and Table 2. As a reminder, the experimental estimator is the Partially Linear Instrumental Variable estimator (PLIV).

By looking at the distribution of treatment effects, we can confirm that on the average treatment effect of the observational methods is 0.05 to 0.07 standard deviations less than the average experimental treatment effect. We can also confirm that the standard deviation of the P2SL estimator is significantly higher than the standard deviations of the other observational estimators as well as the experimental estimator.

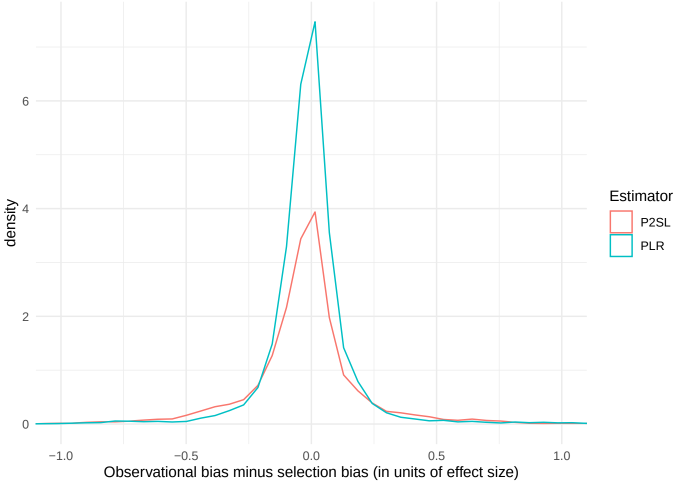


Figure 4: Distribution of the 'change' (observational bias - selection bias) of different estimators

A final metric we can look at is what we call 'change'. This is the difference between the observational bias of P2SL/PLR and the observational bias of WW. Recall that the bias of

Table 3: Summary statistics for ‘change’ (observational bias - selection bias) of estimators

	P2SL	PLR
Mean change	0.006	-0.010
Std dev change	1.347	0.176
% change > 0	0.423	0.438
% change 0 to -0.1	0.317	0.401
% change -0.1 to -0.3	0.162	0.129
% change < -0.3	0.098	0.032

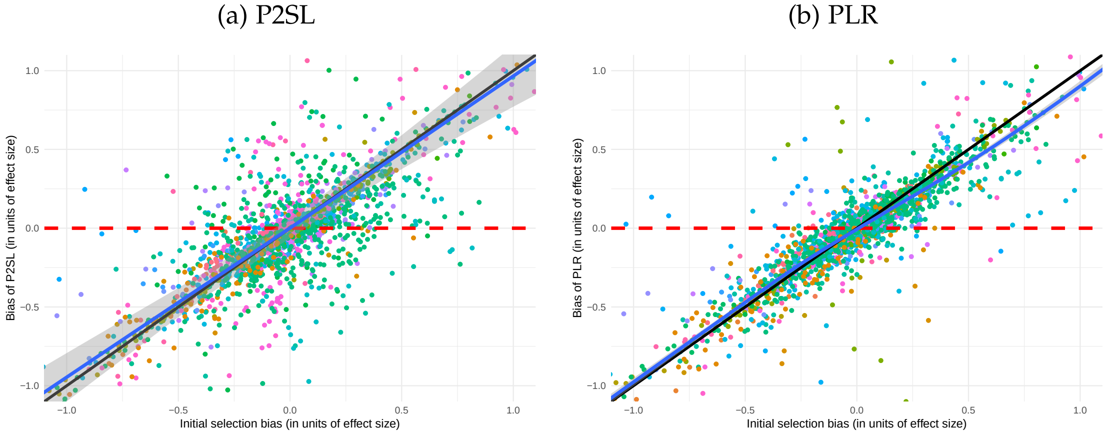
WW is equivalent to a measure of selection bias. As such, the difference between the bias after adding controls to adjust for selection bias (as P2SL and PLR do), and the selection bias, is a measure of how much selection bias is solved by the inclusion of controls. If the change measure is less than zero, then the bias is smaller after adding controls, or equivalently, the estimated observational treatment effect is closer to the experimental treatment effect. If the change measure is greater than zero, then adding controls as increased bias and resulted in a treatment effect estimate that is further away from the experimental effect, relative to simply comparing takers and non-takers.

We can see that both estimators have a substantial mass at 0 which means that in many cases they do not solve any of the selection bias problem. The mean change for P2SL is 0.01 while the mean change for PLR is -0.01. This means that on average P2SL makes the selection bias problem slightly worse, while PLR makes the selection bias problem slightly better. A 0.01 standard deviation improvement/worsening is not significant and is unlikely to be economically important.

Again however, note that, as with the other measures, the variance of the P2SL change measure is significantly higher than the PLR estimator. This means that not only does the P2SL estimator perform worse on average than the PLR estimator, but also that has significantly higher variance performance as well.

Another way of exploring how much the methods are able to reduce selection bias is by plotting the selection bias against the bias of the observational estimator as we do in figure 5a for P2SL and figure 5b for PLR. We show a scatter plot and a loess line of best fit for both estimators. For both plots we also show the 45 degree line in black and a horizontal red dashed line intercepting at 0. Ideally, our bias estimates would lie on the horizontal red line, as this would mean that whatever the initial selection bias, the observational estimator is able to reduce the bias to 0. On the other hand, if the bias estimates lie on the

Figure 5: Observational Bias against initial selection bias

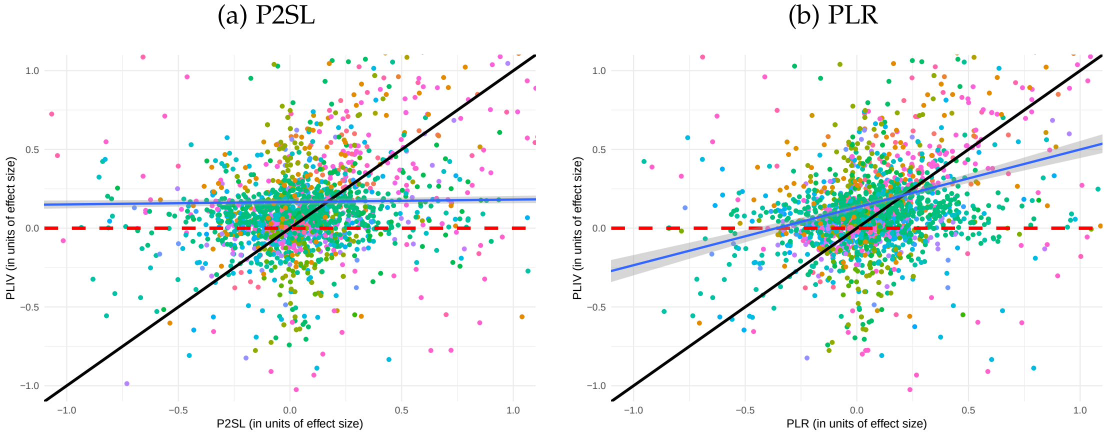

(a) P2SL

(b) PLR

45 degree line, this implies that the observational methods are unable to reduce selection bias if present. If there is no selection bias to begin with and observational methods do not increase the bias, then we would expect the estimates to lie on the 45 degree line.

For both figures 5a and 5b, we can see that the points are much closer to lying on the 45 degree line than the horizontal. We can see that the loess line of best fit for both plots is slightly shallower than the 45 degree line, meaning that the estimators are able to slightly reduce the bias, at least at the extremes, but they are far from solving the entire selection bias problem.

Figure 6: Experimental against observational treatment effect


(a) P2SL

(b) PLR

Figures 6a and 6b plot our estimates of the PLIV against the estimates of the P2SL and PLR. Ideally, the estimates should be on the 45 line for observational methods to recover the experimental ones. However, we can see that the slope is smaller than that especially for the P2SL. The PLR seems to be closer to the experimental estimates. However, these are merely correlations unadjusted for the estimated standard errors.

3.3.2 Meta-analysis

The previous analysis suffers from two problems. First, it doesn't account for the fact that some biases are more precisely estimated than others and thus should be more informative. Second, it doesn't account for the fact that we have multiple estimated biases per study, with intra-study correlation likely to be high. We address both these problems in this section by using meta-analysis.

In particular, we use a Correlated Hierarchical Model with Robust Variance Estimation following Pustejovsky and Tipton (2021) and Viechtbauer (2010). This allows us to cluster our standard errors at the study level as well as weight estimates by their precision. It however does not allow us to selectively weight studies with higher weights that have a small standard error for the underlying experimental estimate to address concerns about IV estimators being (statistically speaking) biased in small samples. First, we estimate the meta-analysis using all of the 2108 contrasts from the 40 studies in our dataset to understand the unconditional distribution of the observational bias. Then, we perform a series of heterogeneity and robustness checks, by re-estimating the meta-analysis using different moderator variables to study various conditional bias distributions.

Let B_i be the asymptotic bias of the observational approach. Of course, our estimates are produced with finite data so we only have noisy estimates of the underlying quantities of the bias and this is indicated by the hats.

To deal with the issues of noisy estimates and within cluster correlations, we run a study-level random-effects meta-analysis on the bias estimates for each estimator. The random-effects model takes the following structure.

$$\begin{aligned} \hat{B}_{ij} &= B_{ij} + \varepsilon_{ij} \\ B_{ij} &= \kappa_j + \nu_{ij} \\ \kappa_j &= \mu + \mu_j \end{aligned}$$

$\hat{B}_{ij}$ is the estimate of a true bias B_{ij} of outcome i in study j. κ_j is the average bias in study j and μ is the overall average bias. $\varsigma^2_{(2)} \equiv V(\nu_{ij})$ is the within-study variation or heterogeneity in true effect sizes and $\varsigma^2_{(3)} \equiv V(\mu_j)$ is the between-study variation or heterogeneity in study-average true effect sizes. By substituting we can reduce the model to:

$$\hat{B}_{ij} = \mu + \mu_j + \nu_{ij} + \varepsilon_{ij}.$$

For each estimator we estimate its mean observational bias, the standard error of that mean, and $\hat{\varsigma}$, the standard deviation estimation of the distribution. We do not separately report the within-study standard deviation, $\varsigma_{(2)}$, and the between-study standard deviation, $\varsigma_{(3)}$. Instead we combine them into a single standard deviation, $\varsigma = \sqrt{\varsigma^2_{(2)} + \varsigma^2_{(3)}}$. We estimate this model, accounting for within-study correlation, with robust standard errors clustered at the study level.

Table 4: Meta-analysis of bias estimates from different estimators

	WW	PLR	P2SL
Mean	-0.0231	0.0031	-0.0237
SE	(0.0405)	(0.0231)	(0.0257)
$\hat{\varsigma}$	0.3007	0.2108	0.433
No. studies	40	40	40
No. contrasts	2108	2108	2108

Table 4 shows the results of our main meta-analysis using all of the available data. We estimate a mean observational bias of -0.02, 0.00 and -0.02 for WW, PLR and P2SL respectively. These estimates are all smaller (closer to zero) than the estimates from the raw biases in table 1. We can also see that none of the estimates are statistically significantly different from zero. This means that once we account for the issue of varying precision and correlated outcomes, the average bias is even smaller and less economically significant than when we don't account for it. We still should keep in mind that these estimates are averaging out positive and negative biases, hence an over or underestimation of the program's social desirability.

We can also compare our estimate of the standard deviation, $\hat{\varsigma}$, in table 4 to the raw standard deviations from table 1. We can see that our meta-analysis estimates of the standard deviation are significantly smaller than the raw standard deviations. This suggests that the outliers that inflated the raw standard deviations are actually imprecisely estimated and thus are not given much weight in the meta-analysis estimate which accounts for varying levels of precision. We see that P2SL still has a higher standard deviation than WW or PLR, but the difference between them is much less now.

From the meta-analysis results, we can conclude that the average difference between experimental and observational estimators is zero with the distribution of the observational bias of each estimator being centered around zero. However, the estimators do differ in their standard deviation. The WW selection bias distribution has a standard deviation of 0.30. The PLR standard deviation is only 0.21. This suggests that PLR is able to somewhat deal with the selection bias and pull the tails of the distribution with large positive or negative bias towards zero.

P2SL on the other hand has a standard deviation of 0.43, greater than the standard deviation of the WW estimator. This suggests that, even when accounting for varying levels of precision and correlated biases, the P2SL estimator might perform less well than a simple comparison of takers to non-takers.

We now move to looking at heterogeneity in the bias distribution according to some key characteristics: whether a lagged outcome is available, the number of covariates available, whether the outcome is important and in the abstract, the sector of the study, whether the randomisation is clustered and the design of the RCT. We include indicators for each of these characteristics as moderators in separate meta-analysis. By doing this we can see if there are important predictors of the mean observational bias. This also functions as a form of robustness check.

In table 5 we report the results of separate meta-analyses with the different moderators. Overall we find that not much predicts the direction of the observational bias with very few moderaters being statistically significant.

In the first panel, we look separately at outcomes which do or do not have a lagged outcome available as a potential covariate. A lagged outcome is potentially an important variable for dealing with selection bias as it allows us to compare individuals who are similar on the exact outcome we care about at baseline, and thus potentially have similar unobservables as well. We can see that the mean bias for all estimators with or without a lagged outcome control variable available are not statistically significantly different from 0. The outcomes with lagged outcomes available as controls tend to have more negative biases. This is true for the WW estimator which does not use the lagged outcome and the

Table 5: Heterogeneity meta-analysis

	WW	WW SE	PLR	PLR SE	P2SL	P2SL SE
By whether lagged outcome is available						
Intercept (No lagged outcome)	-0.019	(0.037)	0.008	(0.022)	-0.003	(0.030)
Lagged outcome	-0.011	(0.026)	-0.013	(0.019)	-0.054	(0.037)
$\hat{\varsigma}$		0.299		0.210		0.433
By number of covariates						
Intercept (Below median no. X)	-0.037	(0.043)	0.003	(0.027)	-0.030	(0.023)
Above median no. X	0.054	(0.045)	-0.000	(0.038)	0.016	(0.047)
$\hat{\varsigma}$		0.302		0.212		0.433
By outcome importance						
Intercept (Not in abstract)	-0.014	(0.039)	0.011	(0.023)	-0.008	(0.027)
In abstract	-0.033	(0.017)	-0.031	(0.015)	-0.075	(0.033)
$\hat{\varsigma}$		0.297		0.210		0.432
By RCT clustering level						
Intercept (Individual)	0.070	(0.058)	0.039	(0.044)	-0.023	(0.033)
Clustered	-0.133	(0.077)	-0.052	(0.052)	-0.005	(0.047)
$\hat{\varsigma}$		0.298		0.211		0.434
By design of RCT						
Intercept (Eligibility)	0.008	(0.052)	0.031	(0.032)	-0.024	(0.024)
Encouragement	-0.079	(0.086)	-0.071	(0.067)	-0.001	(0.069)
$\hat{\varsigma}$		0.290		0.206		0.433
No. studies		40		40		40
No. contrasts		2108		2108		2108

PLR and P2SL estimators which may select the lagged outcome as a potential control.[7]

Next, we look at how the number of covariates available affects the performance of the observational estimators in the second panel. Our prior is that having more covariates available should improve the performance of the observational estimator as there are more chances for selection bias to be reduced. We split outcomes into those below and above the median number of covariates, where the median is 407.

We again find no statistically significant differences between outcomes with more or less potential covariates. One explanation for this could be that the initial covariates collected by researchers are those which do the most to alleviate selection bias, and further covariates do little to further resolve the problem.

Next, we look at whether important outcomes are more likely to suffer from selection bias than less important outcomes in the third panel. We define important outcomes as those that are mentioned in the abstract of their corresponding paper and we include a moderator for more important outcomes. 264 of the 2108 outcomes (12.5%) we study are mentioned in the abstract of their respective paper. Here we find negative coefficients on the in abstract moderator which are (marginally) statistically significant. This suggests that important outcomes are more likely to have negative biases than less important outcomes. However, due to our multiple hypothesis testing this finding may be spurious.

As a robustness check, we test whether the bias estimates are similar across RCTs where the randomisation is done individually or cluster randomised. Recall that our approach (relevant for both, the validity of the observational and the experimental method) relies on a SUTVA assumption to be valid. However, we can see that there are no statistically significant differences in the mean bias of clustered RCTs.

As a final robustness check, we test whether the bias distributions are similar across eligibility and encouragement design RCTs. Recall that we use eligibility design RCTs where there is non-compliance in just one treatment group, and encouragement design RCTs where there is non-compliance in both the treatment and control group. Broadly, we don't see statistically significant differences between the bias estimates of the eligibility and encouragement design RCTs.

[7] As the original authors ran RCTs they do not need to collect covariates to reduce selection bias, the randomisation eliminates selection bias in expectation. Instead authors are likely to collect variables that increase their power as much as possible, those that explain a large fraction of the variation in the outcome.

Next, we look at the results for the meta-analysis done separately for each sector in table 6

Table 6: Heterogeneity meta-analysis by sector

	WW	WW SE	PLR	PLR SE	P2SL	P2SL SE
Intercept (Agriculture)	0.168	(0.089)	0.099	(0.104)	0.005	(0.045)
Education	-0.160	(0.147)	-0.062	(0.120)	0.018	(0.061)
Environment	-0.250	(0.319)	-0.228	(0.129)	0.048	(0.111)
Finance	-0.144	(0.096)	-0.041	(0.108)	-0.015	(0.053)
Firms	0.077	(0.323)	-0.012	(0.152)	-0.127	(0.067)
Health	-0.178	(0.155)	-0.160	(0.134)	-0.108	(0.079)
Labor Markets	-0.309	(0.155)	-0.182	(0.141)	-0.059	(0.114)
Political Economy	-0.352	(0.191)	-0.103	(0.107)	0.003	(0.050)
$\hat{\varsigma}$		0.3164		0.2158		0.4353
No. studies		40		40		40
No. contrasts		2108		2108		2108

We can see that the coefficients of the bias distributions for the sectors is not statistically significantly different from 0. However, we have to keep in mind that we have sometimes few studies in each sector so we can interpret this result only with caution.

4 Conclusions

4.1 Limitations of the project

We cover two limitations of our study (1) that observational studies are done differently in practice, and (2) the external validity of our results.

Firstly, we use an idealised form of an observational study to estimate the bias of observational methods. Our observational study is nested inside experimental data so it is in the exact same context and measures variables in the exact same way. This means that any difference between the observational estimate and the experimental estimate is due to selection bias and not due to other changes if our assumptions on the validity of the experimental method hold true.

However, if we were to compare an observational and experimental study of the same intervention, these would likely differ on many different dimensions. The construction of variables such as income may be different across the two studies, or the studies make take place in different populations. Our approach focuses only on the difference caused by different assignment protocols, either self-selection or randomisation. As such, our study may underestimate the differences between observational and experimental treatment effect estimates, as we are not able to account for these other differences with our design. One might also argue that the observational LATE is a constructed estimator that cannot be estimated outside the framework of an ICRCT.

Furthermore, we are only able to study a subset of non-experimental program evaluation methods. We use PLR and P2SL as representatives of a broader class of estimators that rely on an unconfoundedness assumption as we described in assumption 3. However, there are other types of non-experimental program evaluation methods which rely on other assumptions and it is not clear how much our approach tells us about the performance of such methods. Difference-in-differences is a canonical, widely used method that relies on a different, parallel trends assumption to make causal inferences. With a simple difference-in-difference where you compare takers and non-takers, before and after treatment, there are many parallels to our approach, but with more complex difference-in-difference methods where you compare regions who gradually phase-in to a policy, our approach is less informative.

Secondly, one may worry about the external validity of our results. To be clear, we do not explicitly address the question of whether results generalise from one context to another with our approach, we instead focus on the question of whether non-experimental program evaluation approaches are able to replicate the results of experimental ones, within the same context.

Our approach requires us to use imperfect compliance randomised controlled trials and we focus on those trials run by J-PAL and IPA with publicly available data. Many policies and corresponding RCTs are implemented with perfect compliance but as we cannot include these in our approach, we are not able to say how well hypothetical observational studies of these sorts of policies and programs would perform. Furthermore, to the extent that you think studies run by J-PAL and IPA are different to the broader population of studies, you might further worry about external validity.

To be clear though, the external validity issue is one that arises with all LaLonde (1986)-style, within study comparison (WSC) studies. Indeed, an initial motivation for this study is that the existing literature looks at one context at a time, checks whether observational estimates can replicate experimental ones in that context and then generalises. As we take the results from 40 different studies, we hope that our results are more generalisable than previous results in the literature.

4.2 Main results

Our main results are the mean and standard deviation of our observational bias estimates for the WW, PLR and P2SL estimators, as shown in table 4. We find that the mean observational bias of each of these estimators is zero, that means that positive and negative biases seem to average out. However, they differ in terms of their standard deviation, with PLR having a smaller standard deviation than the WW and and P2SL having a larger standard deviation than the WW. While the mean estimates are marginally the same, we can observe that the PLR seems closest to zero. The bias distribution of WW is a measure of the bias distribution of selection bias. This suggests that although PLR is able to reduce selection bias, P2SL might make it worse in our setting.

Overall, even though observational and experimental methods seem to estimate on average similar causal effects, the experimental estimates do not suffer from the large variation that we find for the observational bias.

4.3 Quality and applicability of the evidence

Overall, we think the evidence we use to study the question of the bias of observational methods is high quality and highly applicable. All of the evidence we use is randomised controlled trial evidence. The RCTs have been implemented by J-PAL and IPA, two organisations that are well known for running RCTs. Most of the RCTs are published in good journals as well, providing another signal of quality.

Furthermore, as the RCTs are mostly development economics RCTs, they are highly applicable to the question of what methods researchers should be using in development economics. The fact that we are able to use the results from 40 different RCTs also increases the reliability and at least somewhat generalisability of our results.

However, if the underlying experimental estimate is not a good estimate of the LATE or TT, the resulting observational bias is difficult to interpret. We discuss the quality of each of the individual studies used in the analysis, including the potential threats for identification, in section 3.2.

4.4 Implications for practice

A surprising result is that our meta-analysis shows that observational methods have an average bias of approximately zero, however, the resulting bias term seems to have a large variance. As such, this implies that when a practitioner takes the result from an observational study to inform a decision, the observational study could still be over or underestimate of the treatment effect that an RCT would estimate. This might suggest that running an RCT yield more reliable estimates of the causal effect in our setting.

However, practitioners and grantmakers may have different objective functions. The decision to run an RCT might depend on various factors. Social science RCTs are typically time consuming, costly, not always implementable and can be disrupting.

Especially given the surprising result of a mean zero observational bias, a policy maker or grantmaker might be ready to take higher risk and accept a potentially less reliable observational method. As such we recommend a careful weighting of the advantages and disadvantages for running an RCT compared to using observational methods taking into account the large distribution of the bias of observational methods found in this project.

4.5 Implications for research

Firstly, one main implication for research is that it is important to keep in mind that there might be a wide distribution of observational bias when using observational methods.

Secondly, we can consider what our results imply for researchers who do chose to run observational studies. Our results suggest that using the PLR estimator over the P2SL estimator will give you estimates that are closer to the experimental benchmark. Indeed, in our sample, using a P2SL seems to be further away from the experimental benchmark with larger standard deviation than a naive comparison of takers and non-takers without even attempting to control for selection bias. This suggests that more research is needed to understand why different observational methods seem to be more reliable than others and what are the driving factors.

Our results can be extended in multiple directions by other researchers. Firstly, researchers could use our approach on other imperfect compliance RCTs and increase the evidence base on the performance of observational methods. Secondly, the set of methods tested could also be expanded. This could include other methods that rely on an unconfoundedness assumption, but with more innovative designs, other researchers may be able to expand to other methods which rely on different identification assumptions such as difference-in-differences or regression discontinuity designs.

References

Ambler, K., D. Aycinena, and D. Yang (2015): "Channeling remittances to education: A field experiment among migrants from El Salvador," *American Economic Journal: Applied Economics*, 7, 207–32.

Angelucci, M., D. Karlan, and J. Zinman (2015): "Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco," *American Economic Journal: Applied Economics*, 7, 151–82.

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Arceneaux, K., A. S. Gerber, and D. P. Green (2006): "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis*, 14, 37 – 62.

Ashraf, N., X. Giné, and D. Karlan (2009): "Finding missing markets (and a disturbing epilogue): Evidence from an export crop adoption and marketing intervention in Kenya," *American Journal of Agricultural Economics*, 91, 973–990.

Ashraf, N., D. Karlan, and W. Yin (2006): "Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines," *The Quarterly Journal of Economics*, 121, 635–672.

Bach, P., V. Chernozhukov, M. S. Kurz, and M. Spindler (2021): "DoubleML – An Object-Oriented Implementation of Double Machine Learning in R," ArXiv:2103.09603 [stat.ML].

Banerjee, A., E. Duflo, R. Chattopadhyay, and J. Shapiro (2016a): "The long term impacts of a "Graduation" program: Evidence from West Bengal," *Unpublished paper, Massachusetts Institute of Technology, Cambridge, MA*.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010): "Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India," *American Economic Journal: Economic Policy*, 2, 1–30.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007): "Remedying education: Evidence from two randomized experiments in India," *The Quarterly Journal of Economics*, 122, 1235–1264.

Banerjee, A. V., E. Duflo, and M. Kremer (2016b): "The influence of randomized controlled trials on development economics research and on development policy," *The state of Economics, the state of the world*, 439–498.

Beaman, L., D. Karlan, B. Thuysbaert, and C. Udry (2013): “Profitability of fertilizer: Experimental evidence from female rice farmers in Mali,” *American Economic Review*, 103, 381–86.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): “Inference on Treatment Effects after Selection among High-Dimensional Controls,” *The Review of Economic Studies*, 81, 608.

Blattman, C. and J. Annan (2016): “Can employment reduce lawlessness and rebellion? A field experiment with high-risk men in a fragile state,” *American Political Science Review*, 110, 1–17.

Blattman, C. and S. Dercon (2018): “The impacts of industrial and entrepreneurial work on income and health: Experimental evidence from Ethiopia,” *American Economic Journal: Applied Economics*, 10, 1–38.

Blattman, C., N. Fiala, and S. Martinez (2014): “Generating skilled self-employment in developing countries: Experimental evidence from Uganda,” *The Quarterly Journal of Economics*, 129, 697–752.

——— (2020): “The long-term impacts of grants on poverty: Nine-year evidence from Uganda’s youth opportunities program,” *American Economic Review: Insights*, 2, 287–304.

Blattman, C., J. C. Jamison, and M. Sheridan (2017): “Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia,” *American Economic Review*, 107, 1165–1206.

Braconnier, C., J.-Y. Dormagen, and V. Pons (2017): “Voter registration costs and disenfranchisement: experimental evidence from France,” *American Political Science Review*, 111, 584–604.

Bruhn, M., D. Karlan, and A. Schoar (2018): “The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in Mexico,” *Journal of Political Economy*, 126, 635–687.

Bryan, G., S. Chowdhury, and A. M. Mobarak (2014): “Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh,” *Econometrica*, 82, 1671–1748.

Chaplin, D. D., T. D. Cook, J. Zurovac, J. S. Coopersmith, M. M. Finucane, L. N. Vollmer, and R. E. Morris (2018): “The Internal and External Validity of the Regression Discontinuity Design: A Meta-Analysis of 15 Within-Study Comparisons,” *Journal of Policy Analysis and Management*, 37, 403–429.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017): “Double/Debiased Machine Learning for Treatment and Structural Parameters,” *The Econometrics Journal*, C1–C68.

——— (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Chong, A., A. L. De La O, D. Karlan, and L. Wantchekon (2015): "Does corruption information inspire the fight or quash the hope? A field experiment in Mexico on voter turnout, choice, and party identification," *The Journal of Politics*, 77, 55–71.

Cohen, J., P. Dupas, and S. Schaner (2015): "Price subsidies, diagnostic tests, and targeting of malaria treatment: evidence from a randomized controlled trial," *American Economic Review*, 105, 609–45.

Crépon, B., F. Devoto, E. Duflo, and W. Parienté (2015): "Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco," *American Economic Journal: Applied Economics*, 7, 123–50.

Dehejia, R. H. and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

——— (2002): "Propensity Score-Matching Methods For Nonexperimental Causal Studies," *The Review of Economics and Statistics*, 84, 151–161.

Devoto, F., E. Duflo, P. Dupas, W. Parienté, and V. Pons (2012): "Happiness on tap: Piped water adoption in urban Morocco," *American Economic Journal: Economic Policy*, 4, 68–99.

Duflo, E. (2018): "Machinistas meet randomistas: Useful ML tools for empirical researchers," NBER Summer Institute Master Lectures.

Duflo, E., P. Dupas, and M. Kremer (2015): "Education, HIV, and early fertility: Experimental evidence from Kenya," *American Economic Review*, 105, 2757–97.

Dupas, P. (2011): "Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya," *American Economic Journal: Applied Economics*, 3, 1–34.

Dupas, P., V. Hoffmann, M. Kremer, and A. P. Zwane (2016): "Targeting health subsidies through a nonprice mechanism: A randomized controlled trial in Kenya," *Science*, 353, 889–895.

Dupas, P., E. Huillery, and J. Seban (2018a): "Risk information, risk salience, and adolescent sexual behavior: Experimental evidence from Cameroon," *Journal of Economic Behavior & Organization*, 145, 151–175.

Dupas, P., D. Karlan, J. Robinson, and D. Ubfal (2018b): "Banking the unbanked? Evidence from three countries," *American Economic Journal: Applied Economics*, 10, 257–97.

Dupas, P. and J. Robinson (2013a): "Savings constraints and microenterprise development: Evidence from a field experiment in Kenya," *American Economic Journal: Applied Economics*, 5, 163–92.

——— (2013b): "Why don't the poor save more? Evidence from health savings experiments," *American Economic Review*, 103, 1138–71.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group (2012): "The Oregon health insurance experiment: evidence from the first year," *The Quarterly journal of economics*, 127, 1057–1106.

Forbes, S. P. and I. J. Dahabreh (2020): "Benchmarking Observational Analyses Against Randomized Trials: a Review of Studies Assessing Propensity Score Methods," *Journal of General Internal Medicine*, 35, 1396–1404.

Franklin, J. M., E. Patorno, R. J. Desai, R. J. Glynn, D. Martin, K. Quinto, A. Pawar, L. G. Bessette, H. Lee, E. M. Garry, N. Gautam, and S. Schneeweiss (2020): "Emulating Randomized Clinical Trials with Nonrandomized Real-World Evidence Studies: First Results from the RCT DUPLICATE Initiative," *Circulation*.

Galiani, S., P. Gertler, N. Ajzenman, and A. Orsola-Vidal (2016): "Promoting handwashing behavior: The effects of large-scale community and school-level interventions," *Health economics*, 25, 1545–1559.

Gill, B., J. Furgeson, H. Chiang, B. ru Teh, J. Haimson, and N. V. Savitz (2016): "Replicating Experimental Impact Estimates with Nonexperimental Methods in the Context of Control-Group Noncompliance," *Statistics and Public Policy*, 3, 1–11.

Giné, X., D. Karlan, and J. Zinman (2010): "Put your money where your butt is: a commitment contract for smoking cessation," *American Economic Journal: Applied Economics*, 2, 213–35.

Glazerman, S., D. M. Levy, and D. Myers (2003): "Nonexperimental versus Experimental Estimates of Earnings Impacts," *The Annals of the American Academy of Political and Social Science*, 589, 63–93.

Gordon, B. R., F. Zettelmeyer, N. Bhargava, and D. Chapsky (2019): "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," *Marketing Science*, 38, 193–225.

Guiteras, R., J. Levinsohn, and A. M. Mobarak (2015): "Encouraging sanitation investment in the developing world: A cluster-randomized trial," *Science*, 348, 903–906.

Hanna, R., E. Duflo, and M. Greenstone (2016): "Up in smoke: the influence of household behavior on the long-run impact of improved cooking stoves," *American Economic Journal: Economic Policy*, 8, 80–114.

Hicken, A., S. Leider, N. Ravanilla, and D. Yang (2018): "Temptation in vote-selling: Evidence from a field experiment in the Philippines," *Journal of Development Economics*, 131, 1–14.

Karlan, D., S. Mullainathan, and B. N. Roth (2019): "Debt traps? Market vendors and moneylender debt in India and the Philippines," *American Economic Review: Insights*, 1, 27–42.

Karlan, D., A. Osman, and J. Zinman (2016): "Follow the money not the cash: Comparing methods for identifying consumption and investment responses to a liquidity shock," *Journal of Development Economics*, 121, 11–23.

Karlan, D., B. Savonitto, B. Thuysbaert, and C. Udry (2017): "Impact of savings groups on the lives of the poor," *Proceedings of the National Academy of Sciences*, 114, 3079–3084.

Khan, A. Q., A. I. Khwaja, and B. A. Olken (2016): "Tax farming redux: Experimental evidence on performance pay for tax collectors," *The Quarterly Journal of Economics*, 131, 219–271.

LaLonde, R. J. (1986): "Evaluating the Econometric Evaluation of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

Mohammed, S., R. Glennerster, and A. J. Khan (2016): "Impact of a daily SMS medication reminder system on tuberculosis treatment outcomes: a randomized controlled trial," *PloS one*, 11, e0162944.

Pons, V. and G. Liegey (2019): "Increasing the electoral participation of immigrants: Experimental evidence from France," *The Economic Journal*, 129, 481–508.

Pritchett, L. and J. Sandefur (2015): "Learning from Experiments when Context Matters," *American Economic Review*, 105, 471–475.

Pustejovsky, J. and E. Tipton (2021): "Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models." *Prev Sci.*

Ravallion, M. (2020): "47C1Should the Randomistas (Continue to) Rule?" in *Randomized Control Trials in the Field of Development: A Critical Perspective*, Oxford University Press.

Romero, M., J. Sandefur, and W. A. Sandholtz (2017): "Can Outsourcing Improve Liberia's Schools? Preliminary Results from Year One of a Three-Year Randomized Evaluation of Partnership Schools for Liberia," *Center for Global Development Working Paper*.

Smith, J. A. and P. E. Todd (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125, 305–353.

Viechtbauer, W. (2010): "Conducting meta-analyses in R with the metafor package," *Journal of Statistical Software*, 36, 1–48.

Wong, V. C., J. C. Valentine, and K. Miller-Bains (2017): "Empirical Performance of Covariates in Education Observational Studies," *Journal of Research on Educational Effectiveness*, 10, 207–236.

A Appendix - Proofs

Proof: Theorem 1

Note that the proof can be written down with or without conditioning on the common support $S_i = 1$.

Classical results imply that the Wald estimands E and E_s recover $\mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1]$ (*eligibility*) and $\mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1]$ (*encouragement*) respectively under Assumption 2. Adding common support, we have $E = E_s$.

Under Assumption 3, classical results imply that $NE_s^r = \mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, S_i = 1, R_i = r] = \mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, R_i = r]$ since $S_i = 1$ is implied by the common support assumption.

What remains to be shown is that the non experimental Wald ratio recovers the LATE, $NE_s = \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1]$ under Assumptions 2 and 3: First note that the observational estimate on the manipulation arm is the sum of the treatment effects for the always-takers and the compliers weighted by the respective occurrence probabilities:

$$\begin{aligned}
NE_s^1 &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, S_i = 1, R_i = 1] \\
&= \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = D_i^0 = 1, S_i = 1] \Pr(D_i^1 = D_i^0 = 1 | D_i = 1, S_i = 1, R_i = 1) \\
&\quad + \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1] \Pr(D_i^1 - D_i^0 = 1 | D_i = 1, S_i = 1, R_i = 1),
\end{aligned}$$

where the second equality comes from Independence and Monotonicity.

Now let us consider the conditional complier probabilities:

$$\begin{aligned}
\Pr(D_i^1 = D_i^0 = 1 | D_i = 1, S_i = 1, R_i = 1) &= \frac{\Pr(D_i^1 = D_i^0 = 1 \wedge D_i = 1 | S_i = 1, R_i = 1)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)} \\
&= \frac{\Pr(D_i^1 = D_i^0 = 1 | S_i = 1, R_i = 1)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)} \\
&= \frac{\Pr(D_i = 1 | S_i = 1, R_i = 0)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)},
\end{aligned}$$

where the first equality comes from Bayes rule, the second equality from the fact that $D_i^1 = D_i^0 = 1$ imply $D_i = 1$ and the third equality from Monotonicity and Independence.

Using the same approach, we have:

$$\begin{aligned}\Pr(D_i^1 - D_i^0 = 1 | D_i = 1, S_i = 1, R_i = 1) &= \frac{\Pr(D_i^1 - D_i^0 = 1 \wedge D_i = 1 | S_i = 1, R_i = 1)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)} \\ &= \frac{\Pr(D_i^1 - D_i^0 = 1 | S_i = 1, R_i = 1)}{\Pr(D_i = 1 | S_i = 1, R_i = 1)},\end{aligned}$$

where the first equality uses Bayes rule and the second equality uses the fact that $D_i^1 - D_i^0 = 1$ implies $D_i = 1$ when $R_i = 1$. Under Monotonicity and Conditional Independence, we also have:

$$\begin{aligned}NE_s^0 &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, S_i = 1, R_i = 0] \\ &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = D_i^0 = 1, S_i = 1].\end{aligned}$$

Combining the formulas for NE_s^1 and NE_s^0, the numerator of the NE_s estimand in equation 1 is:

$$\begin{aligned}&NE_s^1 \Pr(D_i = 1 | R_i = 1, S_i = 1) - NE_s^0 \Pr(D_i = 1 | R_i = 0, S_i = 1) \\ &\quad = \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1] \Pr(D_i^1 - D_i^0 = 1 | S_i = 1, R_i = 1) \\ &\quad\quad + \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = D_i^0 = 1, S_i = 1] \Pr(D_i = 1 | S_i = 1, R_i = 0) \\ &\quad\quad - \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = D_i^0 = 1, S_i = 1] \Pr(D_i = 1 | S_i = 1, R_i = 0) \\ &\quad = \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1, S_i = 1] \Pr(D_i^1 - D_i^0 = 1 | S_i = 1, R_i = 1).\end{aligned}$$

Finally, Monotonicity and Independence imply that:

$$\Pr(D_i^1 - D_i^0 = 1 | S_i = 1, R_i = 1) = \Pr(D_i = 1 | S_i = 1, R_i = 1) - \Pr(D_i = 1 | S_i = 1, R_i = 0),$$

which proves the result.

Note that if $S_i = 1$ for everyone, hence every individual has covariates on the common support, then $NE = NE_s = \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 = 1]$. ■

B Appendix - Bias decomposition

The validity of the observational methods relies on two key assumptions: conditional independence and common support. We are interested in how much both components contribute to the bias the observational method and therefore propose to decompose the measure of the bias into two separate components which reflect both assumptions. Recall that the observational method is only defined over the set of X values common for participants and non participants: $S_X = S_{1X} \cap S_{0X}$ with $S_{1X} = Support\,\{X|D = 1\}$ and $S_{0X} = Support\,\{X|D = 0\}$. Let us define the variable S_i which is $S_i = 1$ if X_i is on the common support S_X. Then we can define the the different estimands on and off the common support.

Define E_s conditional on being on the common support $S_i = 1$. Note that through the assumptions for the non experimental estimator, we already implicitly conditioned on the common support, hence $NE_s = NE$.

Define for the encouragement design:

B.1 Decomposition of the bias

$$
\begin{aligned}
NE_s^d = & \,\mathbb{E}[Y_i|D_i = 1, R_i = d, S_i = 1] \\
& - \mathbb{E}[\mathbb{E}[Y_i|X_i, D_i = 0, R_i = d, S_i = 1]|D_i = 1, R_i = d, S_i = 1]. \\
& \textit{And the Wald-Ratio:} \\
NE_s = & \,\frac{NE_s^1 \Pr(D_i = 1|R_i = 1, S_i = 1) - NE_s^0 \Pr(D_i = 1|R_i = 0, S_i = 1)}{\Pr(D_i = 1|R_i = 1, S_i = 1) - \Pr(D_i = 1|R_i = 0, S_i = 1)} \\
E_s = & \,\frac{\mathbb{E}[Y_i|R_i = 1, S_i = 1] - \mathbb{E}[Y_i|R_i = 0, S_i = 1]}{\Pr(D_i = 1|R_i = 1, S_i = 1) - \Pr(D_i = 1|R_i = 0, S_i = 1)}.
\end{aligned}
$$

For the eligbility design, we have : $NE_s = NE_s^1$ and E_s with $\Pr(D_i = 1|R_i = 0, S_i = 1) = 0$.

Consequently, the bias of the observational method can be defined as:

$$B_{NE} = NE_s - E$$

The bias of the observational method can therefore be decomposed into two components:

$$B_{NE} = B_u + B_{tts}$$

Where $B_u = NE_s - E_s$ and $B_{tts} = E_s - E$. B_u measures the failure of the observational estimand to account for all the relevant confounders on the common support. B_{tts} measures the bias due to the fact that the treatment effect on the common support is not representative of the treatment effect outside of the common support. B_u might decrease as we condition on more covariates but B_{tts} cannot. Whether the bias of the observational method seems mainly from B_u or B_{tts} determines whether there is some scope for improving the observational methods by observing more covariates.

Naturally, we can see that if Assumptions 1, 2 and 3 hold, then, $NE_s = E_s = E$, and as a consequence, $B_{ne} = B_u = B_{tts} = 0$.

C Appendix - Description of studies

	Study	Context	Treatment	Non-compliance	Examples of outcome variables
1	**Title**: Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Phillippines. **Authors**: Ashraf, Nava; Karlan, Dean; Yin, Wesley. **Journal**: Quarterly Journal of Economics. **Year published in repository:**2014	Although much has been written, little has been resolved concerning the representation of preferences for consumption over time. From models in economics, individuals who voluntarily engage in commitment devices ex ante may improve their welfare. If individuals with time-inconsistent preferences are sophisticated enough to realize it, one should observe them engaging in various forms of commitment. The authors designed a commitment savings product for a Philippine bank and implemented it using a randomized control methodology.	The authors designed a commitment savings product for a Philippine bank. The savings product was intended for individuals who want to commit now to restrict access to their savings, and who were sophisticated enough to engage in such a mechanism. The authors randomly assigned these individuals to three groups: commitment-treatment (T), marketing-treatment (M), and control (C) groups. The treatment group received access to “SEED” (Save, Earn, Enjoy Deposits) account. This account was a pure commitment savings product that restricted access to deposits as per the client’s instructions upon opening the account, but did not compensate the client for this restriction.	The authors offered the commitment product to a randomly chosen subset of 710 clients; 202 (28.4 percent) accepted the offer and opened the account.	Change in total balance (6 months, 12 months). Change in non-seed balances (12 months).
2	**Title:** Northern Uganda Social Action Fund - Youth Opportunities Program (YOP) (published as Generating skilled self-employment in developing countries: Experimental evidence from Uganda. **Authors**: Blattman, Christopher; Fiala, Nathan; Martinez, Sebastian. **Journal**: Quarterly Journal of Economics. **Year published in repository**: 2014	The authors study a government program in Uganda designed to help the poor and unemployed become self-employed artisans, increase incomes, and thus promote social stability. Young adults in Uganda’s conflict-affected north were invited to form groups and submit grant proposals for vocational training and business start-up.	Funding was randomly assigned among screened and eligible groups. A list of 535 groups eligible for randomisation was given to the research team, and they randomly assigned 265 groups to the treatment and 270 groups to the control, stratified by district. Treatment groups received unsupervised grants of $382 per member.	11% of groups assigned to treatment did not receive a grant.	Enrolled in vocational training (2-year) Business assets (2 and 4-year) Average employment hours per week (2 and 4-year) Engaged in any skilled trade (4-year) Enterprise is formally registered (2 and 4-year) No. of paid and unpaid laborers hired in past month, family and nonfamily (4-year)

3	**Title**: Put Your Money Where your Butt Is: A Commitment Contract for Smoking Cessation. **Authors**: Giné, Xavier;Karlan, Dean;Zinman, Jonathan. Journal: American Economic **Journal**: Applied Economics. **Year published in repository**: 2014	The authors designed and tested a voluntary commitment product to help smokers quit smoking. Their study sample consists of 2,000 smokers aged 18 or older who reside on the island of Mindanao in the southern Philippines.	The product (CARES) offered smokers a savings account in which they deposit funds for six months, after which they take a urine test for nicotine and cotinine. If they pass, their money is returned; otherwise, their money is forfeited to charity.	Of smokers offered CARES, 11 percent took it up.	Passing urine test 6 months and 1 year later.
4	**Title**: Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh. **Authors**: Bryan, Gharad; Chowdhury, Shyamal; Mobarak, Ahmed Mushfiq. **Journal**: Econometrica. **Year published in repository**: 2014	This paper studies the causes and consequences of internal seasonal migration in northwestern Bangladesh, a region where over 5 million people live below the poverty line, and must cope with a regular pre-harvest seasonal famine. This seasonal famine—known locally as monga—is emblematic of the widespread lean or “hungry” seasons experienced throughout South Asia and Sub-Saharan Africa, in which households are forced into extreme poverty for part of the year.	The authors randomly assign an $8.50 incentive to households in rural Bangladesh to temporarily out-migrate during the lean season. 100 villages are split into four groups: Cash, Credit, Information, and Control.	The informational manipulation has perfect take-up. However, in the pooled encouragement design manipulation, where migration is the program, these do not have perfect take-up.	Total consumption, total calories, total savings, total earnings.
5	**Title**: Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya. **Authors**: Dupas, Pascaline; Robinson, Jonathan **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2015	Many microentrepreneurs do not have access to basic financial services such as savings account, which may impede business success. The authors test this directly by expanding access to bank accounts for a randomly selected sample of small informal business owners in one town of rural Western Kenya.	The authors randomised access to noninterest-bearing bank accounts among two types of self-employed individuals in rural Kenya: market vendors (who are mostly women) and men working as bicycle taxi drivers.	A total of 156 respondents had the opportunity to open a savings account through this program. Twenty-one of them (13%) refused to open the account, while another 40% opened an account but never made a single deposit.	Bank savings, business investment and daily private expenditure.

6	**Title**: Why Don’t the Poor Save More? Evidence from Health Savings Experiments. **Authors**: Dupas, Pascaline; Robinson, Jonathan **Journal**: American Economic Review. **Year published in repository**: 2015	In developing countries, the returns to many types of investments in human or physical capital appear to be high, yet investment levels remain quite low. Credit constraint's arise as an obvious culprit, but cost of these investments are not massive. As a result, household should be able to save up to these investments. Using data from a field experiment in Kenya, the authors document that providing individuals with simple informal savings technologies can substantially increase investment in preventative health and reduce vulnerability to health shocks.	They worked with 113 ROSCAs in one district of Kenya, and randomly assigned these ROSCAs to one of five study arms. Treatments are a safebox, lockbox, health pot and health savings account, HSA.	Imperfect compliance in each of the five study arms, varying from 65% to 93%	Amount spent on preventative health products since baseline, whether participant could not afford medical treatment in last 3 months, participant reached health goal and finally ROSCA exists at 33 months
7	**Title**: Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya. **Authors**: Dupas, Pascaline **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2015	Nearly 2 million people become infected with HIV/AIDS every year in sub- Saharan Africa, the great majority of them through sex, and a quarter of them before the age of 25. The author uses a randomized experiment to test whether and what information changes teenagers’ sexual behavior in Kenya.	The study provides participants information on the relative risk of HIV infection by partner’s age. There were 4 treatment groups: (1) Schools with the teachers who received the training program on the national HIV/AIDS curriculum that focuses on abstinence (TT); (2) School with 8th grade classrooms that received the relative risk of partners' age, implemented by an NGO on the prevalence of HIV disaggregated by age and gender group (RR); Schools that received both of these treatments (TT & RR); and schools that received neither program.	The 164 schools selected for the HIV Education program were asked to send three upper primary teachers to participate in a five-day training program. Since schools have 14 teachers on average, the training program covered around 21 percent of teachers in program schools. Compliance with the training was high, with 93 percent of training slots filled.	Age difference between teenage girl and her partner, whether girls have ever had sex but never used a condom, and whether boys have ever had sex but never used a condom.
8	Title: Encouraging Sanitation Investment in the Developing World: A Cluster-Randomized Trial. Authors: Guiteras, Raymond; Levinsohn, James; Mobarak, Ahmed Mushfiq Journal: Science. **Year published in repository**: 2015	Poor sanitation contributes to morbidity and mortality in the developing world, but there is disagreement on what policies can increase sanitation coverages.	The authors assigned 380 communities in rural Bangladesh to different marketing treatments – community motivation and information; subsidies; a supply-side market access intervention; and a control – in a cluster-randomised trial.	Up-take of higienic latrine ownership did not increase in the community motivation and information, but it did increase in the subsidy group by 22 percentage points, as well as to their unsubsidied neighbors within that group.	Open Defecation or Hanging Toilet usage

9	**Title**: Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco. **Authors**: Angelucci Manuela, Karlan Dean, and Zinman Jonathan **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2015	Expanded access to credit may improve the welfare of its recipients by lowering transaction costs and mitigating information asymmetries. Compartamos Banco is the largest microlender in Mexico and targets women who operate a business or are interested in starting one.	The authors use a clustered randomized trial to estimate impacts at the community level from a group lending expansion at 110 percent APR. Specifically, they randomized credit access and loan promotion across 238 geographic clusters. Both baseline and endline surveys were administered to potential borrowers.	Treatment assignment strongly predicts the depth of Compartamos penetration: according to Compartamos administrative data, 18.9 percent (1,563) of those surveyed in the treatment areas had taken out Compartamos loans during the study period, compared to only 5.8 percent (485) of those surveyed in the control areas.	The authors measure effect in 37 outcomes across 6 domains: microentrepreneurship, income, labor supply, expenditures, social status, and subjective well-being. Examples of these are revenues, value of assets and expenses in food and health.
10	**Title**: Finding Missing Markets (and a disturbing epilogue): Evidence from an Export Crop Adoption and Marketing Intervention in Kenya. **Authors**: Ashraf, Nava; Giné, Xavier; Karlan, Dean. **Journal**: American Journal of Agricultural Economics. **Year published in repository:** 2014	In much of the developing world, many farmers grow crops for local or personal consumption despite export options which appear to be more profitable. The authors report here on a randomized controlled trial conducted by DrumNet in Kenya that attempts to help farmers adopt and market export crops. DrumNet provides smallholder farmers with information about how to switch to export crops, makes in-kind loans for the purchase of the agricultural inputs, and provides marketing services by facilitating the transaction with exporters.	The experimental evaluation design randomly assigns pre-existing farmer self-help groups to one of three groups: (1) a treatment group that receives all DrumNet services, (2) a treatment group that receives all DrumNet services except credit, or (3) a control group.	41% of the members from credit groups joined DrumNet, only 27% did so when credit was not included as a DrumNet service.	Whether farmer produced a crop for export, total spent in marketing, household income.
11	**Title**: Education, HIV and Early Fertility: Experimental Evidence from Kenya. **Authors**: Duflo, Esther; Dupas, Pascaline; Kremer, Michael. **Journal**: American Economic Review. **Year published in repository:** 2015	Early fertility and sexually transmitted infections (STIs), chief among them HIV, are arguably the two biggest health risks facing teenage girls in sub-Saharan Africa. A seven-year randomised evaluation suggests education subsidies reduce adolescent girls' dropout, pregnancy, and marriage but not sexually transmitted infection (STI).	The study took place in all 328 public primary schools in 7 divisions of 2 districts in Western Kenya: Butere-Mumias and Bungoma. Schools were stratified and assigned one of four arms using a random number generator: (i) Control (82 schools); (ii) Stand-alone education subsidy program i.e., providing free school uniforms (83 schools); (iii) Stand-alone HIV education program (83 schools); (iv) Joint program (80 schools).	The 164 schools selected for the HIV Education program were asked to send three upper primary teachers to participate in a five-day training program. Since schools have 14 teachers on average, the training program covered around 21 percent of teachers in program schools. Compliance with the training was high, with 93 percent of training slots filled.	Dropped out of primary school, ever married, ever pregnant, HIV positive blood test.

12	**Title**: Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco. **Authors**: Crépon, Bruno; Devoto, Florencia; Duflo, Esther; Parienté, William, **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2016	The authors present results from a randomized evaluation of microcredit in rural areas of Morocco.The design of our study tracked the expansion of Al Amana, their partner microcredit institution (MFI) into non-densely populated areas between 2006 and 2007.	Selected villages were matched in pairs based on observable characteristics. In each pair, one village was randomly assigned to treatment, and the other to control. In total, 81 pairs belonging to 47 branches were included in the evaluation. In treatment villages, credit agents started to promote microcredit and to provide loans immediately after the baseline survey. They visited villages once a week and performed various promotional activities: door-to-door campaigns, meetings with current and potential clients, contact with village associations, cooperatives, and women's centers, etc.	13% of the households in treatment villages took a loan, and none in control villages did.	Assets, income from labor and salaried labor, expenses and investments.
13	**Title**: Targeting health subsidies through a nonprice mechanism: A randomized controlled trial in Kenya. **Authors**: Dupas, Pascaline; Hoffman, Vivian; Kremer, Michael; Zwane, Alix Peterson. **Journal**: Science. **Year published in repository:** 2016	Free provision of preventive health products can markedly increase access in low-income countries. A cost concern about free provision is that some recipients may not use the product, wasting resources. Yet, charging a price to screen out nonusers may screen out poor people who need and would use the product. The authors report on a randomized controlled trial of a screening mechanism that combines the free provision of chlorine solution for water treatment with a small nonmonetary cost.	This study compares three mechanisms for allocating dilute-chlorine water treatment solution: (1) Cost sharing program (50% discount off the retail prices); (2) Voucher program where 12 vouchers were provided, each redeemable for one 150-mL bottle of water treatment solution at either a local shop or at the clinic, and (3) Free delivery program. The free delivery program functions as a control group because there was perfect compliance with this treatment group.	Take-up of the cost-sharing treatment starts in 52% with the voucher of one bottle, and take-up of the vouchers starts with 85% of participants that redeemed at leas one voucher. Cotrol group reports perfect compliance.	Positive chlorine test at follow-up

14	**Title**: Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial. **Authors**: Cohen, Jessica; Dupas, Pascaline; Schaner, Simone. **Journal**: American Economic Review. **Year published in repository:** 2017	Both under- and over-treatment of communicable diseases are pub- lic bads. But efforts to decrease one run the risk of increasing the other. Using rich experimental data on household treatment-seeking behavior in Kenya, the authors study the implications of this trade-off for subsidizing life-saving antimalarials sold over-the-counter at retail drug outlets.	The study selected four drug shops, in four rural market centers and sampled all households in the catchment area (within a 4-kilometer radius) of each of these shops. Then they visited each household to administer a baseline survey. At the end of the survey two vouchers for artemisinin combination therapies (ACTs) and, when applicable, two vouchers for rapid diagnostic tests (RDTs) were distributed. Surveyors explained that ACTs are the most effective type of antimalarial and, if the household received an RDT voucher, what the RDT was for and how it worked. Households were randomly assigned to one of three core groups, corresponding to the three policy regimes of interest: ACT voucher (no subsidy), subsidised ACT voucher, and subsidised ACT voucher + subsidised RDT voucher. Both the ACT and RDT subsidies had three levels of subsidisation.	Only 19% of illnesses in the control group were treateed with ACT. Any ACT subsidy over 80% increased take -up by 16 to 23 percentage points.	Actual malaria status, whether theey reported any illness episode, number of episodes and patient age.
15	**Title**: Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State. **Author**: Blattman, Christopher; Annan, Jeannie. **Journal**: American Political Science Review. **Year published in repository:** 2015	States and aid agencies use employment programs to rehabilitate high-risk men in the belief that peaceful work opportunities will deter them from crime and violence. Rigorous evidence is rare.	The authors experimentally evaluate a program of agricultural training, capital inputs, and counseling for Liberian ex-fighters who were illegally mining or occupying rubber plantations. Action on Armed Violence (AoAV) rebuilt and operated two training centers and designed a job training program with a large productive asset and conditional cash transfer.	Men were randomly assigned to an offer to enter the program in this order within blocks until a target number per block was reached. 75% of those assigned to treatment complied.	Whether respondent does any farming, or farming and animal raising, and cash earnings over the past month.

16	Title: Channeling Remittances to Education: A Field Experiment among Migrants from El Salvador. Author: Ambler, Kate; Aycinena, Diego; Yang, Dean Journal: American Economic Journal: Applied Economics. **Year published in repository**: 2017	Migrant remittances are one of the largest types of inter- national financial flows to developing countries, amounting in 2012 to over US$400 billion.	The authors implement a randomized experiment offering Salvadoran migrants matching funds for educational remittances, which are channeled directly to a beneficiary student in El Salvador chosen by the migrant. There are 3 treatment groups and 1 control group: a) 3:1 where each dollar was matched with $3 in project funds, b) 1:1 match, c) No match where migrants were simply offered the EduRemesa product without matching funds and d) control group.	18.5% of migrants in the 3:1 match, treatment executed at least one EduRemesa transaction, compared to 6.9% in the 1:1 match group and exactly zero in the no match group. A total of 15.1% and 6.0% of migrants with the 3:1 and 1:1 matches, respectively, sent an EduRemesa to their target student.	Total annualized target student expenditure (migrant) and average hours per week any work (student)
17	**Title**: Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia. **Authors**: Blattman, Christopher; Jamison, Julian; Koroknay-Palicz, Tricia; Rodrigues, Katherine; Sheridan, Margaret. **Journal**: American Economic Review. **Year published in repository:** 2017	In many countries, poor young men exhibit high rates of violence, crime, and other antisocial behaviors. In addition to their direct costs, crime and instability hinder economic growth by reducing investment or diverting productive resources to security. In fragile states, such men are also targets for mobilization into election intimidation, rioting, and rebellion.	The authors recruited criminally engaged men and randomized one-half to eight weeks of cognitive behavioral therapy designed to foster self-regulation, patience, and a noncriminal identity and lifestyle. They also randomized $200 grants. They show that a number of noncognitive skills and preferences, including patience and identity, are malleable in adults, and that investments in them reduce crime and violence.	Of men assigned to the grant, 98% received it. Of men assigned to therapy, 5% attended none, another 5% dropped out within the first three weeks, and two-thirds attended at least 80% of all sessions	Antisocial behaviors, drug trade and economic performance at different points in time.

18	**Title**: Banking the Unbanked? Evidence from Three Countries. **Authors**: Dupas, Pascaline; Karlan, Dean; Robinson, Jonathon; Ubfal, Diego. **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2017	Bank accounts are essential to daily economic life in developed countries but are still far from universal in developing countries: only 54 percent of adults in developing countries report having a bank account, compared to 94 percent in OECD countries.	The authors experimentally test the impact of expanding access to basic bank accounts in Uganda, Malawi, and Chile. The experiment contained a control group and a treatment group within each country for the given subject population. In Malawi and Uganda, treatment respondents were given a voucher that could be redeemed for the free account at the bank branch; paperwork assistance was also extended to respondents. While in Chile, treatment respondents were informed of the existence of the main account features [which entailed no fees] and were invited to open an account with BancoEstado.	Account take varies on average from 17% in Chile, 54% in Uganda and 69% in Malawi.	Savings stocks in various categories, labor income and total expenditures.
19	**Title**: The Impacts of Industrial and Entrepreneurial Work on Income and Health: Experimental Evidence from Ethiopia. **Authors**: Blattman, Christopher; Dercon, Stefan. **Journal**: Review of Economic Studies. **Year published in repository:** 2017	A large body of evidence suggests that formal firms pay premium wages, especially large, foreign-owned or exporting firms. However, industrial jobs may be attractive only compared to poor people’s largely informal alternatives. It is not clear that wage premiums exist in general, or that steady employment makes up for the disamenities of industrial work. This paper uses an experiment in Ethiopia to investigate a) the relative qualities of informal and industrial work, b) the benefits or risks to the choice of one occupation over the other, and c) how does the quality of self-employment options affect this occupational choice.	Working with five Ethiopian firms, the authors randomized applicants to an industrial job offer, an "entrepreneurship" program of $300 plus business training, or control status. Industrial jobs offered more and steadier hours but low wages and risky conditions.	Of those offered the job, 10% did not show up the first day. Meawhile, 13% of the control group was hired by a study firm during the year. Of those in the Entrepreneur treatment, 97% were offered the grant and training, and 94% received the grant and training.	Weekly cash earnings, hours worked and physical health.

20	Title: Impact of savings groups on the lives of the poor. Authors: Karlan, Dean; Savonitto, Beniamino; Thuysbaert, Bram; Udry, Christopher. Journal: Proceedings of the National Academy of Sciences (PNAS). **Year published in repository:**2017	The poor make complex financial decisions and use the limited range of financial instruments available to them to address their varying needs. The available formal and informal tools, however, are often risky and expensive or lack necessary flexibilities. Savings-led microfinance programs operate in poor rural communities in developing countries to establish groups that save and then lend out the accumulated savings to each other. Nonprofit organizations train villagers to create and lead these groups.	In a clustered randomized evaluation spanning three African countries (Ghana, Malawi, and Uganda), the authors present the results of the Village Savings and Loan Association (VSLA) program across a total of 561 clusters, 282 of which were randomly assigned to treatment and the remaining of which were randomly assigned to control.	Program up-take at the end of the study in the treatment groups are 36% in Ghana and Uganda, and 22% in Malawi. In the control group are 8%, 6% and 3% respectively.	Income and revenue, assets, consumption, women's empowerment.
21	**Title**: The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico. **Authors**: Karlan, Dean; Bruhn, Miriam; Schoar, Antoinette. **Journal**: Journal of Political Economy. **Year published in repository:** 2017	A large literature in development economics and entrepreneurship aims to understand the impediments to firm growth. Capital alone cannot explain the entirety of firm growth and therefore “managerial capital” is needed to know how to employ the capital best. The authors argue that managerial capital can directly affect the firm by improving strategic and operational decisions, and by increasing the productivity of other factors.	The intervention aims to expand the managerial skills of the managers by giving them access to subsidized consulting and mentoring services. Treated enterprises met with their consultants for 4 hours per week over a 1-year period. The randomized controlled trial took place in Puebla, Mexico, in which 432 micro, small, and medium-sized enterprises applied to receive subsidized consulting services, and 150 out of the 432 were randomly chosen to receive the treatment.	Out of the 150 enterprises in the treatment group, 80 then took up the consulting services. The remaining 70 treatment group enterprises declined to participate in the program although they had initially signed a letter of interest saying that they would participate if offered a spot.	Number of employees, daily wage bill, entreprenurial spirit and full-time employees.

22	**Title**: Temptation in vote-selling: Evidence from a field experiment in the Philippines. **Authors**: Hicken, Allen; Leider, Stephen; Ravanilla, Nico; Yang, Dean **Journal**: Journal of Development Economics. **Year published in repository:** 2019	Vote-buying and vote-selling are pervasive phenomena in many developing democracies. Vote-buying and other forms of clientelism can undermine the standard accountability relationship that is central to democracy, as well as hampering the development of and trust in the political institutions and is associated with larger public deficits and public sector inefficiencies.Because of these potential inimical effects, NGOs, and international donors have directed significant attention and resources towards combating vote-buying and vote-selling.	The authors report the results of a randomized field experiment in the Philippines on the effects of two common anti-vote-selling strategies involving eliciting promises from voters. There were two treatment groups and one control group, where a third of participants were assigned to each. The treatment group participants were invited to make a promise in terms of their voting behavior in the upcoming mayoral, vice-mayoral, and city council elections. For treatment 1 promess reads "to not accept money from any candidate", and for treatment 2 "to vote their conscience, even if money was accepted"	In each treatment group, slightly more than half of respondents make the promise—51% for Promise 1 (“Don’t take the money”) and 56% for Promise 2 (“Take money, vote conscience”)—and these proportions are not different from one another at conventional levels of statistical significance.	Whether respondent switched vote for mayor, vice-mayor, city council or any race.
23	**Title**: Follow the money not the cash: Comparing methods for identifying consumption and investment responses to a liquidity shock. **Authors**: Karlan, Dean; Osman, Adam; Zinman, Jonathan. **Journal**: Journal of Development Economics. **Year published in repository:** 2019	Measuring the impacts of liquidity shocks on spending is difficult but important for theory, practice and policy. They shed light on perceived returns to investment, and on the extent to which constraints bind more for some types of household spending than others. Estimating impacts of liquidity shocks matters in many do- mains, for example in understanding household leveraging and deleveraging decisions in the wake of credit supply shocks, as well as evaluating interventions such as business grants, unconditional cash transfers, and microcredit expansions.	In the counterfactual analysis of this paper, the authors take advantage of a randomized trial in which marginal applications were randomly assigned to either treatment or control (i.e., compare cash outflows of those who borrowed to a counterfactual group that did not borrow). Then, at both two weeks and two months post-randomization, independent surveyors asked about all cash outflows from the individual's household or business that exceeded a certain amount, and compare treatment to control to estimate the impact of the liquidity shock on specific outcomes.	67% of the treated group reports having a loan from an experimenting lender, compared to 34% in the control group.	Business expenditures, assets for business, utilities for business, merchandise for business, business renovations, salaries for employees.

24	**Title**: The long-term impacts of grants on poverty: 9-year evidence from Uganda's Youth Opportunities Program. **Authors**: Blattman, Christopher; Fiala, Nathan; Martinez, Sebastian. **Journal**: AER: Insights. **Year published in repository:** 2019	In 2008, Uganda gave $400 per person to thousands of young people to help them start skilled trades, work more, and raise incomes (The Youth Opportunities Program (YOP)). Four years on, an experimental evaluation found grants raised work by 17 percent and earnings by 38 percent. After nine years, the authors find these gains have dissipated. Grantees' investment leveled off; controls eventually increased their incomes and so both groups converged in employment, earnings, and consumption levels.	Funding was randomly assigned among screened and eligible groups. A list of 535 groups eligible for randomisation was given to the research team, and they randomly assigned 265 groups to the treatment and 270 groups to the control, stratified by district. Treatment groups received unsupervised grants of $382 per member.	11% of groups assigned to treatment did not receive a grant.	Income after 4 and 9 years, monthly earning, nondurable consumption, average employment hours, whether the respondent engaged in any skilled trade.
25	**Title**: Can Outsourcing Improve Liberia's Schools? Preliminary Results from Year One of a Three-Year Randomized Evaluation of Partnership Schools for Liberia. **Authors**: Romero, Mauricio; Sandefur, Justin; Sandholtz, Wayne. **Journal**: American Economic Review. **Year published in repository:** 2018	Governments often enter into public-private partnerships as a means to raise capital or to leverage the efficiency of the private sector. This paper studies the Partnership Schools for Liberia (PSL) program, which delegated management of 93 public schools (3.4 percent of all public primary schools, serving 8.6 percent of students enrolled in public primary or preschool) to 8 different private organizations.	93 randomly selected public schools are delegated to private providers. Providers received US$50 per pupil, on top of US$50 per pupil annual expenditure in control schools.	The percentage of students originally assigned to treatment schools who are actually in treatment schools at the end of the school year is 81%.	English and math test scores, composite test scores, pupil/teacher ratio, instruction time.

26	**Title**: Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification. **Authors**: Chang, Alberto; De La O, Ana L.; Karlan, Dean; Wantchekon, Leonard. **Journal**: The Journal of Politics. **Year published in repository:** 2020	Retrospective voting models assume that offering more information to voters about their incumbents' performance strengthens electoral accountability. However, it is unclear whether incumbent corruption information translates into higher political participation and increased support for challengers. The authors provide experimental evidence that of the effects of such information in local elections in Mexico.	Households within the boundaries of an experimental voting precinct were assigned to receive a flyer. There are 3 treatment groups (1) "Corruption Information": flyer included information about the percentage of resources the mayor spent in a corrupt [public spending w/ some form of irregularity] manner, (2) Placebo – "Budget expenditure": only information about the percent of resources mayors spent by the end of the fiscal year, (3) Placebo – "Poverty expenditure": information about the percent of resources mayors directed toward improving services for the poor and 1 control – received no information.	Compliance with treatment assignment was overall high. Among voting precincts in the state of Jalisco, 97% received full treatment; among voting precincts in Morelos, 89% received full treatment; and among voting precincts in Tabasco, 60% of precincts were fully treated, 20% were partially treated, and 20% failed to receive any treatment.	Turnout, incumbent party votes over registered voters, challenger party votes over registered voters, whther the respondent identifies with the incumbent party or the challenger party.
27	**Title**: Debt Traps? Market Vendors and Moneylender Debt in India and the Philippines. **Authors**: Karlan, Dean; Mullainathan, Sendhil; Roth, Benjamin N. **Journal**: AER: Insights. **Year published in repository:** 2020	A debt trap occurs when someone takes on a high-interest-rate loan and is barely able to pay back the interest, and thus perpetually finds themselves in debt (often by refinancing). Studying such practices is important for understanding financial decision-making of households in dire circumstances, and also for setting appropriate consumer protection policies. This paper reports three experiments: Chennai, India in 2007 (1000 market vendors), Cagayan de Oro, Philippines in 2007 (250 market vendors), and Cagayan de Oro, Philippines in 2010 (701 market vendors, from different markets than in 2007).	Both the experiments in Chennai (India 07) and in Cagayan de Oro (Phillipines 07) included the same four equal-sized treatment arms: 1) debt payoff; 2) financial education; 3) debt payoff and financial education; and 4) control. In the 2010 Philippines experiment, participants were randomised into one of four groups: 1) debt payoff; 2) savings account; 3) debt payoff and savings account; and 4) control. All three treatment groups in this study also received a 5-10 minute financial education lesson.	In the Philippines 07 experiment, 105 out of the 125 vendors invited to the training attended and only nominal compensation was given for attendance. In India 07, 434 out of 500 individuals attended the financial training. Because of problems with insufficient compliance with account opening requirements in the Phillipines 10 experiment, only 10 savings accounts were opened, and thus there is nothing to analyze with respect to the savings account treatment arms. Financial training was not tested separatelly in this last experiment.	Household expenditures, take-hope profit, total working capital, whether they hold any moneylender debt.

28	Title: Profitability of Fertilizer: Experimental Evidence from Female Rice Farmers in Mali. Authors: Beaman, Lori; Karlan, Dean; Thuysbaert, Bram; and Udry, Christopher Journal: AEA: Papers and proceedings. **Year published in repository:** 2020	Intensified use of agricultural inputs, par- ticularly fertilizer, is a possible route to im- proved agricultural productivity. The authors use a field experiment to provide free fertilizer to women rice farmers in southern Mali to measure how farmers choose to use the fertilizer, what changes they make to their agricultural practices, and the profitability of this set of changes.	The experiment was conducted in 23 villages in the district of Bougouni of southern Mali. 383 women were randomly assigned to one of 2 treatment cells or a control group: (1) 135 received the total recommended quantity per hectare, (2) 123 received half of the recommended quantity per acre, and (3) 125 were in the control group and received no fertilizer.	In control, 32% of women used fertilizer, whereas the two treatments had almost perfect compliance, generating treatment effects of 64 percentage points (se=0.04) for both the half and full treatments (96%)	Family labor, fertilizer expenses, total inputs, value of output and profits.
29	**Title**: Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. **Authors**: Duflo, Esther;Banerjee, Abhijit;Banerji, Rukmini;Glennerster, Rachel;Khemani, Stuti. **Journal**: American Economic Journal: Economic Policy. **Year published in repository:** 2009	The deplorable state of publicly provided social services in many developing countries has attracted considerable attention in recent years. Participation of beneficiaries in the monitoring of public services is increasingly seen as a key to improving their quality. The authors conducted a randomized evaluation of three interventions to encourage beneficiaries' participation to India. The evaluation took place in 280 villages in the Jaunpur district in the state of Uttar Pradesh, India.	In the first treatment, mobilization, teams facilitated a meeting, got discussions going, and encouraged village administrators to share information about the structure and organization of local service delivery. The second treatment also provided that information, but administered a reading test for children, and invited them to create "report cards" on the status of enrollment and learning in their village. The third intervention had the features of the first two, but added a "reading course" that lasted two to three months, with classes held every day outside of school. This intervention offered the opportunity to improve learning among children.	On average, only 8 percent of children (including 13 percent of those who could not recognize letters) in our sample attended the reading class in intervention 3 villages.	Whether children could read letters, words or paragraphs and stories.

30	**Title**: Happiness on Tap: Piped Water Adoption in Urban Morocco. **Authors**: Devoto, Florencia; Duflo, Esther; Dupas, Pascaline; Parienté, William; Pons, Vicent. **Journal**: American Economic Journal: Economic Policy. **Year published in repository:** 2012	Worldwide, 1.1 billion people have no access to any type of improved drinking source of water within 1 kilometre. Furthermore, only about 42 percent of the people with access to water have a household connection. Connecting private dwellings to the water main is expensive and typically cannot be publicly financed. The authors worked in collaboration with Amendis, a private utility company, which operates the drinking water distribution in Tangiers, Morocco. In 2007, Amendis launched a social program to increase household direct access to piped water.	The Amendis program (BSI) provided an interest-free loan to cover the cost of the water connection. The loan was to be repaid in regular instalments with the water bill over three to seven years. The authors conducted a door-to-door awareness and facilitation campaign in early 2008 among 434 households, randomly chosen from the 845 that were eligible for a connection on credit. Those households received information about the credit offer as well as help with the administrative procedures needed to apply for the credit and the water connection. The remaining households (the comparison group) were eligible to apply for a connection on credit if they wanted to, but they received neither individualized information nor procedural assistance.	69% of treatment households purchased a home connection by August 2008, while 10% in of control households did.	Income generated by female head, household wellbeing, respondent wellbeing.
31	**Title**: Up in Smoke: The Influence of Household Behavior on the Long-Run Impact of Improved Cooking Stoves. **Authors**: Duflo, Esther;Greenstone, Michael;Hanna, Rema. **Journal**: American Economic Journal: Economic Policy. **Year published in repository:** 2015	A third of the world's population, and up to 95 percent in poor countries, rely on solid fuels, including biomass and coal, to meet their energy needs. Laboratory studies suggest that improved cooking stoves can reduce indoor air pollution, improve health, and decrease greenhouse gas emissions in developing countries. The authors provide evidence, from a large-scale randomized trial in India, on the benefits of a common, laboratory-validated stove.	A public lottery determined the order in which stoves were constructed within each village for 2,600 households. The first third of households within each village received the stoves at the start of the project, the second third received the stoves about two years after the first wave, and the remaining households received them at the end.	Over 70 percent of households that won Lottery 1 built a GV stove during the first six months of the program. Lottery 2 winners did not look very different than Lottery 1 winners.	Carbon monoxide exposure, any illnes, health expenditures, BMI of children aged 13 and under, infant mortality.

32	**Title**: Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors. **Authors**: Khan, Adnan Q;Khwaja, Asim I;Olken, Benjamin **Journal**: Quarterly Journal of Economics. **Year published in repository:** 2015	Although much has been written, little has been resolved concerning the representation of preferences for consumption over time. From models in economics, individuals who voluntarily engage in commitment devices ex ante may improve their welfare. If individuals with time-inconsistent preferences are sophisticated enough to realize it, one should observe them engaging in various forms of commitment. The authors designed a commitment savings product for a Philippine bank and implemented it using a randomized control methodology.	The authors designed a commitment savings product for a Philippine bank. The savings product was intended for individuals who want to commit now to restrict access to their savings, and who were sophisticated enough to engage in such a mechanism. The authors randomly assigned these individuals to three groups: commitment-treatment (T), marketing-treatment (M), and control (C) groups. The tratment group received access to “SEED” (Save, Earn, Enjoy Deposits) account. This account was a pure commitment savings product that restricted access to deposits as per the client’s instructions upon opening the account, but did not compensate the client for this restriction.	The authors offered the commitment product to a randomly chosen subset of 710 clients; 202 (28.4 percent) accepted the offer and opened the account.	Change in total balance (6 months, 12 months). Change in non-seed balances (12 months).
33	**Title**: Impact of a Daily SMS Medication Reminder System on Tuberculosis Treatment Outcomes: A Randomized Controlled Trial. **Authors**: Mohammed, Shama;Glennerster, Rachel;Khan, Aamir J. **Journal**: PlosOne. **Year published in repository:** 2016	Tuberculosis is the second-leading cause of death from infectious diseases globally, with nine million people infected and 1.5 million deaths in 2013. The rapid uptake of mobile phones in low and middle-income countries over the past decade has provided public health programs unprecedented access to patients. For that reason the authors measure the impact of Zindagi SMS, a two-way SMS reminder system, on treatment success of people with drug-sensitive tuberculosis.	The authors conducted a two-arm, parallel design, effectiveness randomized controlled trial in Karachi, Pakistan. Individual participants were randomized to either Zindagi SMS or the control group. Zindagi SMS sent daily SMS reminders to participants and asked them to respond through SMS or missed (unbilled) calls after taking their medication. Non-respondents were sent up to three reminders a day. They enroll 2,207 participants, with 1,110 randomized to Zindagi SMS and 1,097 to the control group.	Of the 1,069 participants who were sent messages, 912 (85%) responded at least once. Over the course of treatment, average response rates fell from 48% in the first two weeks to 24% (eight-month regimen) and 20% (six-month regimen) in the last two weeks.	Clinically recorded treatment success, whether the participant took medication in the last 24 hours, self reported treatment completion.

34	**Title**: Remedying Education: Evidence from two randomized experiments in India. **Authors**: Banerjee, Abhijit;Cole, Shawn;Duflo, Esther;Linden, Leigh. **Journal**: Quarterly Jourrnal of Economics. **Year published in repository:** 2017	There is a tension in the public conversation about primary education in developing countries. On the one hand, primary education should be universal. On the other hand, there is dismal quality of the educational services that developing countries offer to the poor. This paper presents the results of two randomized experiments conducted in schools in urban India (Vadodara and Mumbai)	The first is remedial education program hired young women ("Balsakhi") to teach students lagging behind in basic literacy and numeracy skills. An instructor typically meets with a group of approximately 15–20 children in a class for two hours a day during school hours. The second is a computer-assisted learning program where children in grade 4 are offered two hours of shared computer time per week during which they play games that involve solving math problems.	There is perfect compliance in year 1 of the intervention in Mumbai, and year 1 and 2 in Vadodara. However, the implementation in year 2 in Mumbai experienced some administrative difficulties. For various reasons, only two-thirds of the schools assigned balsakhis actually re- ceived them. Nevertheless, all children were tested, regardless of whether or not they participated in the program.	Test score in math, language and total.
35	**Title**: The Long term Impacts of a "Graduation" Program: Evidence from West Bengal. **Authors**: Banerjee, Abhijit; Chattopadhyay, Raghabendra; Duflo, Esther; Shapiro, Jeremy **Journal**: AER: Insights. **Year published in repository:** 2017	Development economics has long posited that the poor may be poor for no good reason other than the fact that they started poor. This is the idea of a poverty trap, which has the implication that a one-time capital grant that makes very poor households significantly less poor. This paper studies the long-run effects of a "big-push" program providing a large asset transfer to the poorest Indian households.	Households in the treatment group who chose to participate chose a productive asset from a menu of options (two cows, four goats, one cow and two goats, non- farm microenterprise inventory, etc). About 82 percent chose livestock. In addition to the asset, they received weekly consumption support for 30–40 weeks, 3 access to savings, and weekly visits from Bandhan staff over a span of 18 months.	56 percent of selected households received the transfer.	Assets, per-capita consumption, food security, income and revenues, financial inclusion.

36	Title: Voter Registration Costs and Disenfranchisement: Experimental Evidence from France. Authors: Braconnier, Céline;Dormage, Jean-Yves;Pons, Vincent Journal: American Political Science Review. **Year published in repository:**2017	Elections in established democracies regularly attract less than half of the voting-age population, raising concerns not only for the equal representation of all citizens, but also for the overall legitimacy and stability of the democratic regimes. A large-scale randomized experiment conducted during the 2012 French presidential and parliamentary elections shows that voter registration requirements have significant effects on turnout, resulting in unequal participation.	20,500 apartments, located at 4,118 addresses, were assigned to one control group or six treatment groups: 1) early canvassing and 2) late canvassing: canvassers encouraged people to register and provided information about the proces. In 3) early home registration and 4) late home registration: the canvassers offered to register people at home so that they would not have to register at the town hall. In 5) early canvassing and late home registration, and 6) early home registration and late home registration.	Number of new registrations in the treatment groups vary between 0.18 and 0.26, and for the control group are 0.17.	Electoral participation, interest in politics.
37	**Title**: Risk information, risk salience, and adolescent sexual behavior: Experimental evidence from Cameroon. **Authors**: Dupas, Pascaline; Huillery, Elise; Seban, Juliette **Journal**: Journal of Economic Behavior & Organization. **Year published in repository:**2017	Every day young people engage in risky behaviors, including teen drinking and driving, smoking, drug use, criminal activity, and unprotected sex. Future costs of these behaviors are often immense. For example, unprotected sex presents the dual risk of unwanted pregnancy and HIV infection. These risks are disproportionately borne by young women. This paper tests the hypothesis that the behavior of adolescents responds to risk information and risk salience. The authors consider one type of risky behavior: risky sex, in one context: Cameroon.	318 schools in 3 regions participated in the program, with a sample totaling 2907 girls. There are four interventions. The first (In-Class Quiz) students were simply asked to fill in an anonymous questionnaire with questions on HIV as well as on their own sexual behavior and that of their peers. Two of the others consisted of general information on HIV prevention methods and the average HIV prevalence at the national level. These two could be delivered by a teacher that received special training (Teacher Training) or by an external consultant. A third one mimicked the “sugar daddy risk information”.	3 schools out of 80 in the Teacher Training (TT) group had nobody from the school staff attending the training.	Knowledge about HIV, ways of prevention, whether they are pregnant and whether has started childbearing.

38	**Title**: Promoting Handwashing Behaviour: The Effects of Large-scale Community and School-level Interventions. **Authors**: Galiani, Sebastian; Gertler, Paul; Ajzenman, Nicolas; Orsola-Vidal, Alexandra **Journal**: Health Economics. **Year published in repository:** 2017	Handwashing with soap is generally easy and relatively cheap, and the medical evidence of its benefits is conclusive (significant reduction in diarrhea and respiratory illnesses in children). However, only 3 to 34% of the population in developing countries routinely washes their hands with soap at critical junctures during the day. This paper analyzes a randomized experiment that uses novel strategies to promote handwashing with soap at critical points in time in Peru	The intervention comprised two main components: 1) a mass media campaign at the province level and 2) a more comprehensive district-level community treatment that included, in addition to media campaigns, training of community agents; capacity-building for mothers, caregivers, and children; and the inclusion of handwashing promotion as part of primary school curricula. 85 districts (with between 1500 and 100,000 inhabitants) were randomly selected, with 44 randomly assigned to receive the district-level community treatment and the other 41 randomly assigned to serve as the control group.	Exposure to treatments varies from 14.7 percent to 71 percent, depending on intensity of the exposure.	Knowledge on events that require handwashing, best method to wash hands, handwashing behaviour.
39	Title: Increasing the Electoral Participation of Immigrants: Experimental Evidence from France. Authors: Pons, Vincent;Liegey, Guillaume Journal: Economic Journal. **Year published in repository:** 2018	As the number of first- and later-generation immigrants continues to increase among the population of the United States and Europe, the question of their integration gains ever more importance. Policies implemented to foster immigrants' integration fall into three groups, broadly speak- ing. Laws regulating the access to citizenship, citizenship tests, and related civic integration policies directly affect immigrant's efforts and attitudes to integrate. In this study, 23,800 citizens were randomly assigned to receive visits from political activists during the lead-up to the 2010 French regional elections.	678 addresses were randomly allocated to the manipulated group, which received the visits of the canvassers, and the remaining 669 addresses to the non-manipulated group, which did not receive any visit. All citizens living in the same building thus belonged to the same group by design	92 percent of buildings in the teament group were visited by canvassers.	Participation in regional and catonal elections.

40	**Title**: The Oregon Health Insurance Experiment: Evidence from the First Year. **Authors**: Finkelstein, Amy;Baicker, Katherine;Taubman, Sarah;Wright, Bill;Bernstein, Mira;Gruber, Jonathan;Allen, Heidi;Newhouse, Joseph P;Schneider, Eric;Zaslavsky, Alan. **Journal**: Quarterly Journal of Economics. **Year published in repository:** 2018	In early 2008, Oregon opened a waiting list for a limited number of spots in its Medicaid program for low-income adults, which had previously been closed to new enrollment. The state drew names by lottery from the 90,000 people who signed up. This lottery presented an opportunity to study the effects of access to public insurance using the framework of a randomized controlled design. In this article the authors examine the effects of the Oregon Medicaid lottery after approximately one year of insurance coverage.	In January 2008, Oregon determined it had the budget to enroll an additional 10,000 adults in the Oregon Health Plan (OHP) Standard program. New members would be added through random lottery draws from a new reservation list. Anyone could be added to the lottery list and a total of 89,824 individuals were placed on the list during the five-week window it was open. The state conducted eight lottery drawings from the list with roughly equal numbers selected from each drawing. Selected individuals won the opportunity to apply for OHP Standard coverage. In total, 35,169 individuals were selected by lottery.	About 30 percent of selected individuals successfully enrolled in OHP.	Out of pocket medical expenses, whether respondent owes money for medical expenses, utilization, self-reported health and access.

D Appendix - Quality of studies

	Study	Exclusion restriction	Attrition	Spill-overs	Sample size
1	**Title**: Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Phillippines. **Authors**: Ashraf, Nava; Karlan, Dean; Yin, Wesley. **Journal**: Quarterly Journal of Economics. **Year published in repository:**2014	The offer to open a "SEED" bank account does not affect outcomes in ways other than the program.	Not reported.	Not discussed.	Up to 1777 observations.
2	**Title:** Northern Uganda Social Action Fund - Youth Opportunities Program (YOP) (published as Generating skilled self-employment in developing countries: Experimental evidence from Uganda. **Authors**: Blattman, Christopher; Fiala, Nathan; Martinez, Sebastian. **Journal**: Quarterly Journal of Economics. **Year published in repository**: 2014	Being offered the grant does not affect training, business assets and employment in ways other than the program.	Nearly 40% of the YOP applicants had moved or were temporarily away at each endline survey. To minimise attrition, the authors used a two-phase tracking approach. Their response rate was 97% at baseline, and effective response rates at endline (weighted for selection into endline tracking) were 85% after two years and 82% after four.	Spillovers between study villages were unlikely as the 535 groups were spread across 454 communities in a population of more than five million, and control groups are typically very distant from treatment villages.	Up to 2029 observations.
3	**Title**: Put Your Money Where your Butt Is: A Commitment Contract for Smoking Cessation. **Authors**: Giné, Xavier;Karlan, Dean;Zinman, Jonathan. Journal: American Economic **Journal**: Applied Economics. **Year published in repository**: 2014	The offer of CARES does not affect smoking behaviors in ways other than the program. However, the authors highlight that the instrument may not satisfy the exclusion restriction as there is the possibility that the CARES offer itself may influence quit behavior among those who are offered, but do not take the product.	Practical reasons required that subject compensation for taking the six-month test vary across treatment arms (CARES users did not receive compensation, while all other subjects did). In principle, this could generate sample selection bias. The 12-month test does not suffer from this problem, since all subjects were offered equal compensation for taking the test. 64% of people were found in each manipulation group, conditional on being found 95% take urine test.	Not discussed.	Up to 2000 observations.

4	**Title**: Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh. **Authors**: Bryan, Gharad; Chowdhury, Shyamal; Mobarak, Ahmed Mushfiq. **Journal**: Econometrica. **Year published in repository**: 2014	The offer of cash or loan does not affect consumption, calorie intake, earnings and savings in ways other than the program.	Not discussed	There are four sources of possible spillovers: 1) migration will affect village labor supply for non-agricultural tasks, and non-migratory household may receive different compensation as a result. 2) potential general equilibrium effects on local goods production due to migration Information may affect financial and labor behavior during upcoming draught 3) remittances may affect migrants' household member's labor supply, 4) migration may affect household dynamics and bargaining that could result in expenditure changes	Up to 2147 observations.
5	**Title**: Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya. **Authors**: Dupas, Pascaline; Robinson, Jonathan **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2015	The offer of the noninterest-bearing bank accounts does not affect savings, business investment and daily private expenditure in ways other than the program.	Two main sources of attrition: (1) some respondents could not be found and asked to keep logbooks and (2) some people refused to fill the logbooks (17 percent of the sample) The post-attrition treatment and control groups that make it into the final analysis do not differ along most observable characteristics	Spouses (and other family members) of bank account owners benefit from increased capability to save.	Up to 250 observations.
6	**Title**: Why Don't the Poor Save More? Evidence from Health Savings Experiments. **Authors**: Dupas, Pascaline; Robinson, Jonathan **Journal**: American Economic Review. **Year published in repository**: 2015	The offering of the safebox, lockbox, health pot and healt savings account does not affect spending on preventative health products, affordability of medical treatment and reaching a health goal in ways other than the program.	5% of individuals recontacted after 6 months and 8% after 12, not differential across experimental arms. ROSCAs may or may not have survived Loss of 21 percent of ROSCAs after random assignment, however the groups seemed relatively balanced, suggesting that ROSCA attrition was orthogonal to the experimental treatment assignment.	Control groups were also ROSCA participants in the same administrative area in Western Kenya, so they could have heard about any of the four treatments and individually implemented them	Up to 771 observations.

7	**Title**: Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya. **Authors**: Dupas, Pascaline **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2015	The training does not affect the the age difference between girls and their partners in ways other than the program.	There is no evidence of differential attrition for any outcome, except for dropout information after five years	The RR program might have had negative spillovers onto nontreated students in the RR treatment schools. Indeed, the control cohort available is a younger cohort (the seventh graders of 2004). This cohort could have been indirectly and negatively affected by the RR information program if the “sugar daddies” newly turned down by informed eighth graders decided to try their luck with seventh graders instead. Alternatively, the seventh graders could have benefitted from positive information spillovers if the eighth graders shared the information with their younger schoolmates.	Up to 6074 observations.
8	Title: Encouraging Sanitation Investment in the Developing World: A Cluster-Randomized Trial. Authors: Guiteras, Raymond; Levinsohn, James; Mobarak, Ahmed Mushfiq Journal: Science. **Year published in repository**: 2015	The offer og higienic latrines does not affect open defecatiuon and hanging toilet usage in ways other than the program.	Not-discussed	The authors study the the extent of demand spillovers across neighbours by randomizing the share of lottery winners at the neighbourhood level into low, medium and high intensity (25, 50 and 75% of households receiving the subsidy). The researcher investigated whether there is a social multiplier in sanitation investments by analysing the effects of the share of other households in the neighbourhood offered subsidies on latrine investment.	Up to 13127 observations
9	**Title**: Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco. **Authors**: Angelucci Manuela, Karlan Dean, and Zinman Jonathan **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2015	Credit access and loan promotion do not affect microentrepreneurship, income, labor supply, expenditures and others in ways other than the program.	The authors attempted to track 2912 household from the baseline to test whether attrition correlates with observed characteristics or differs by treatment assignment. Although attrition is not random—the probability of being in the endline is correlated with some demographics, income and account ownership- neither the rate of attrition nor the correlates of attrition systematically differ in control and treatment areas.	These are possible but considering they find no effect it is not obvious how spillovers will arise.	Up to 16560 observations.

10	**Title**: Finding Missing Markets (and a disturbing epilogue): Evidence from an Export Crop Adoption and Marketing Intervention in Kenya. **Authors**: Ashraf, Nava; Giné, Xavier; Karlan, Dean. **Journal**: American Journal of Agricultural Economics. **Year published in repository:** 2014	The offer of DrumNet services does not affect the crops planted, marketing expenditures and household income in ways other than the program.	86% of the baseline individuals were surveyed in the follow-up survey	Not discussed	Up to 1983 observations.
11	**Title**: Education, HIV and Early Fertility: Experimental Evidence from Kenya. **Authors**: Duflo, Esther; Dupas, Pascaline; Kremer, Michael. **Journal**: American Economic Review. **Year published in repository:** 2015	The training does not affect human capital of girls, their partners and health outcomes in ways other than the program.	There is no evidence of differential attrition for any outcome, except for dropout information after five years.	Teachers getting the training then moving to schools who were not part of the treatment group, but still teaching the trained curriculum Could have positive spillover effects where sexual partners of students educated on condom use will benefit from their safe sex practices (and are therefore less likely to infect other sexual partners)	Up to 9461 observations.
12	**Title**: Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco. **Authors**: Crépon, Bruno; Devoto, Florencia; Duflo, Esther; Parienté, William, **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2016	Microcredit promotion does not affect assets, income, expenditure and investment in ways other than the program. However, the authors highlight that there are good reasons to believe that microcredit availability impacts not only on clients, but also on nonclients through a variety of channels. Thus, the exclusion restriction is likely to be violated.	8% attrition, with some differential attrition concerns.	There are good reasons to believe that microcre- dit availability impacts not only on clients, but also on nonclients through a variety of channels: equilibrium effects via changes in wages or in competition, impacts on behavior of the mere possibility to borrow in the future	Up to 4934 observations.
13	**Title**: Targeting health subsidies through a nonprice mechanism: A randomized controlled trial in Kenya. **Authors**: Dupas, Pascaline; Hoffman, Vivian; Kremer, Michael; Zwane, Alix Peterson. **Journal**: Science. **Year published in repository:** 2016	Discounts in dilute-chlorine water treatment solution do not affect chlorine tests in ways other than the program.	Attrition was 12.8% in the cost sharing group, 11.8% in the vouchers group, and 13.4% in the free delivery group, not statistically different accross groups.	Not discussed	Up to 385 observations.

14	**Title**: Price Subsidies, Diagnostic Tests, and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial. **Authors**: Cohen, Jessica; Dupas, Pascaline; Schaner, Simone. **Journal**: American Economic Review. **Year published in repository:** 2017	ACT subsidies do not affect malaria status and other health outcomes in ways other than the program.	Only 5 percent of households surveyed at baseline were not reached at endline, and attrition was balanced across treatment arms	Limiting the spread of infectious diseases has positive spillovers, and these can exist in members of the treated group that are not treated.	Up to 631 observations.
15	**Title**: Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State. **Author**: Blattman, Christopher; Annan, Jeannie. **Journal**: American Political Science Review. **Year published in repository:** 2015	The offer of training, capital inputs and counseling does not affect occupational choice and earnings in ways other than the program.	8.7% attrition of the sample in two categories: death, unable to be found	The authors expect within-community spillovers to the control group to be minor, given the low percentage of treated men over the adult work force of those communities, and high migration accross villages.	Up to 1025 observations
16	Title: Channeling Remittances to Education: A Field Experiment among Migrants from El Salvador. Author: Ambler, Kate; Aycinena, Diego; Yang, Dean Journal: American Economic Journal: Applied Economics. **Year published in repository**: 2017	The offer to participate in EduRemesa does not affect student expenditure and employment in ways other than the program.	27% of target households didn’t complete the follow-up survey; 26% of migrants didn’t complete the follow-up survey	Spillovers between participant migrants were avoided by a first-stage randomization that was conducted at the day-by-location level that assigned migrants to either the control group or to a group that would receive an offer of the EduRemesa. Spillover in targeted households are not discussed.	Up to 728 observations.
17	**Title**: Reducing Crime and Violence: Experimental	The offer of CBT and grant do not affect noncognitive skills and	7.6% attrition, not differentially in observables accross groups.	The authors work in large neighborhoods, recruiting less than 1%	Up to 947 observations

	Evidence from Cognitive Behavioral Therapy in Liberia. **Authors**: Blattman, Christopher; Jamison, Julian; Koroknay-Palicz, Tricia; Rodrigues, Katherine; Sheridan, Margaret. **Journal**: American Economic Review. **Year published in repository:** 2017	preferences in ways other than the program.		of adult men in those areas, and less than 15% of high-risk men we could identify on the street. They argue this was designed to reduce equilibrium effects such as a change in the returns to illicit work. Another potential spillover involves interactions within and between treatment arms, especially therapy. There could be positive spillovers from treating groups of friends or, alternatively, to the extent that control subjects interact with and learn from treat ment subjects, they may acquire some of the lessons. Without systematic data on networks we cannot estimate spillovers.	
18	**Title**: Banking the Unbanked? Evidence from Three Countries. **Authors**: Dupas, Pascaline; Karlan, Dean; Robinson, Jonathon; Ubfal, Diego. **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2017	Bank access does not affect savings, income and expenditures in ways other than the program.	Attrition in the follow-up surveys is low (~3%) and uncorrelated with treatment status.	Not discussed.	Up to 2159 households in Uganda, 2107 households in Malawi, and 1967 households in Chile.
19	**Title**: The Impacts of Industrial and Entrepreneurial Work on Income and Health: Experimental Evidence from Ethiopia. **Authors**: Blattman, Christopher; Dercon, Stefan. **Journal**: Review of Economic Studies. **Year published in repository:** 2017	The industrial job offer does not affect earnings, hours worked and outher outcomes in ways other than the program.	The authors tracked 88% of individuals after 11 months, 85% after 13 months, and also interviewed 90% of their households.	Not discussed.	Up to 1587 observations
20	Title: Impact of savings groups on the lives of the poor. Authors: Karlan, Dean; Savonitto, Beniamino; Thuysbaert, Bram; Udry,	The offer of VSLA does not affect business and household outcomes in ways other than the program.	8.5% of the sample cannot be found at endline.	Not discussed	Up to 15221 observations.

	Christopher. Journal: Proceedings of the National Academy of Sciences (PNAS). **Year published in repository:**2017				
21	**Title**: The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico. **Authors**: Karlan, Dean; Bruhn, Miriam; Schoar, Antoinette. **Journal**: Journal of Political Economy. **Year published in repository:** 2017	The offer of management consulting services does not affect firm size and managerial capital in ways other than the program.	88% of the 432 enterprises interviewed at baseline were reinterviewed at endline.	Not discussed.	Up to 378 observations.
22	**Title**: Temptation in vote-selling: Evidence from a field experiment in the Philippines. **Authors**: Hicken, Allen; Leider, Stephen; Ravanilla, Nico; Yang, Dean **Journal**: Journal of Development Economics. **Year published in repository:** 2019	The offer to make promisses 1 or 2 does not affect voting behavior in ways other than the program.	The share of the 883 baseline respondents who completed the endline survey, voted, and reported their mayoral vote was 86.0%. The corresponding shares for vice-mayor and city council are 85.0% and 90.0%.	Not discussed.	Up to 806 observations.
23	**Title**: Follow the money not the cash: Comparing methods for identifying consumption and investment responses to a liquidity shock. **Authors**: Karlan, Dean; Osman, Adam; Zinman, Jonathan. **Journal**: Journal of Development Economics. **Year published in repository:** 2019	The offer of a loan does not affect expenditures, assets, and other outcomes in ways other than the program.	Yes, after 2-3 Weeks is 18% and after two Months is 38%.	Not discussed.	Up to 1388 observations.
24	**Title**: The long-term impacts of grants on poverty: 9-year evidence from Uganda's Youth Opportunities Program. **Authors**: Blattman, Christopher; Fiala, Nathan; Martinez, Sebastian. **Journal**:	The offer of grant does not affect income, consumption and employment in ways other than the program.	Nearly 40% of the YOP applicants had moved or were temporarily away at each endline survey. To minimise attrition, we used a two-phase tracking approach…our response rate was 97% at baseline, and effective response rates at endline (where individuals found in phase 2 tracking	Spillovers between study villages were unlikely as the 535 groups were spread across 454 communities in a population of more than five million, and control groups are typically very distant from treatment villages.	Up to 2005 observations.

	AER: Insights. **Year published in repository:** 2019		were given higher weights) were 90.7% after two years (2010), 84% after four (2012) and 87% after nine (2017).”		
25	**Title**: Can Outsourcing Improve Liberia’s Schools? Preliminary Results from Year One of a Three-Year Randomized Evaluation of Partnership Schools for Liberia. **Authors**: Romero, Mauricio; Sandefur, Justin; Sandholtz, Wayne. **Journal**: American Economic Review. **Year published in repository:** 2018	The offer to delegate administration to a private provider does not affect students english and math scores in ways other than the program.	Attrition in the second wave of data collection from ther original sample is balanced between treatment and control and is below 4%	In this setting, while outsourcing management improves most indices of school quality on average, the effect varies across providers. In addition, some providers’ actions had negative unintended consequences and may have generated negative spillovers for the broader education system, underscoring the importance of robust contracting and monitoring for this type of program.	Up to 3508 observations.
26	**Title**: Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification. **Authors**: Chang, Alberto; De La O, Ana L.; Karlan, Dean; Wantchekon, Leonard. **Journal**: The Journal of Politics. **Year published in repository:** 2020	The flyers do not affect incumbent and challenger votes in ways other than the program.	Not discussed.	The corruption-information treatment could have spilled to the placebo and control groups. People who re ceived information about incumbent corruption could have talked to people in other treatment groups and these would dilute the magnitude of the effects. To deal with possible spillover effects, they estimated models without the three municipalities that are state capitals.	Up to 749 observations.
27	**Title**: Debt Traps? Market Vendors and Moneylender Debt in India and the Philippines. **Authors**: Karlan, Dean; Mullainathan, Sendhil; Roth, Benjamin N. **Journal**: AER: Insights. **Year published in repository:** 2020	The offer of training does not affect expenditures and other outcomes in ways other than the program.	In the India 07 experiment, 881 of 1000 completed all 4 follow-up surveys. In Phillipines 07 experiment, 206 of 250 completed all 4 follow-up surveys. In Phillipines 10 experiment, 569 of 701 completed all 4 follow-up surveys	Not discussed.	Up to 2643 observations in India 07, 824 in the Philippines 07, and 2272 in Philippines 10.
28	Title: Profitability of Fertilizer: Experimental Evidence from Female Rice Farmers in Mali. Authors: Beaman, Lori; Karlan, Dean; Thuysbaert, Bram; and Udry, Christopher Journal: AEA: Papers and proceedings. **Year published in repository:** 2020	The delivery of bags of fertilizer does not affect inputs, value of output and profitability in ways other than the program.	The authors were able to collect follow-up data for 378 primary respondents (out of 383).	Not discussed	Up to 378 observations.

29	**Title**: Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. **Authors**: Duflo, Esther;Banerjee, Abhijit;Banerji, Rukmini;Glennerster, Rachel;Khemani, Stuti. **Journal**: American Economic Journal: Economic Policy. **Year published in repository:** 2009	The offering of reading clases does not affect childrens' reading skills in ways other than the program.	In the endline survey, 17,419 children were tested, a sample that includes all but 716 of the children in the baseline.	Not discussed.	Up to 17500 observations.
30	**Title**: Happiness on Tap: Piped Water Adoption in Urban Morocco. **Authors**: Devoto, Florencia; Duflo, Esther; Dupas, Pascaline; Parienté, William; Pons, Vicent. **Journal**: American Economic Journal: Economic Policy. **Year published in repository:** 2012	Information does not affect household wellbeing and income in other way than the program.	Among the 845 households who participated in the baseline survey, 793 households (94 percent) could be resurveyed.	By August 2009 27% of control households had applied for a connection, up from 10% in 2008. Control households could have learned from neighbors the benefits of the connections. and this can be attributed to social learning effects. The results suggest important diffusion effects.	Up to 793 observations.
31	**Title**: Up in Smoke: The Influence of Household Behavior on the Long-Run Impact of Improved Cooking Stoves. **Authors**: Duflo, Esther;Greenstone, Michael;Hanna, Rema. **Journal**: American Economic Journal: Economic Policy. **Year published in repository:** 2015	Providing a stove does not affect outcomes in other way than the program (using the stove to cook)	94 percent of the households participate in the first main two surveys and about 81 percent in the last survey.	Treatment households could conduct all the cooking for the control group since they own the improved stove. The data are inconsistent with this possibility. Second, the experiment may cause control households to learn about the dangers of indoor air pollution, which leads them to change their cooking habits to protect themselves from smoke. Using data from their midline survey, we find no difference in the min utes spent cooking at arm's length from one's cooking stove.	Up to 2511 households.
32	**Title**: Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors. **Authors**: Khan, Adnan Q;Khwaja, Asim I;Olken, Benjamin **Journal**: Quarterly Journal of Economics. **Year published in repository:** 2015	Offering incentives to tax collectors does not affect service quality and tax revenue in ways other than the program.	Not discussed.	Revenue plus areas show higher satisfaction and quality of service appears generalized to other departments beyond just tax, suggesting that there may be positive spillovers, which is consistent with citizens attributing a positive interaction in one government service to other related services.	Up to 9870 observations.

33	**Title**: Impact of a Daily SMS Medication Reminder System on Tuberculosis Treatment Outcomes: A Randomized Controlled Trial. **Authors**: Mohammed, Shama;Glennerster, Rachel;Khan, Aamir J. **Journal**: PlosOne. **Year published in repository:** 2016	The SMS mesages to participant did not affect the outcomes in ways other than the program.	Attrition rate of less than 1 percent, similar across arms for treatment outcomes.	Spillovers were minimized as patients with another household member in the study were ineligible.	Up to 2207 observations.
34	**Title**: Remedying Education: Evidence from two randomized experiments in India. **Authors**: Banerjee, Abhijit;Cole, Shawn;Duflo, Esther;Linden, Leigh. **Journal**: Quarterly Jourrnal of Economics. **Year published in repository:** 2017	The offer of the Balsakhi remedial program, and the computarized program do not affect the test scores in ways other than the program.	For the Balsakhi Program, attrition was 17 and 18 percent, respectively, in the comparison and treatment groups in Vadodara in year 1, 4 percent in both the treatment and the comparison group in Vadodara in year 2. In Mumbai it was 7 and 7.5 percent, respectively, in the treatment and comparison groups in year 1, and 7.7 and 7.3 percent, respectively, in year 2.	Spillover effects of the computerized program on language skills could have occurred due to, for example, increased attendance.	Up to 21936 observations.
35	**Title**: The Long-term Impacts of a “Graduation” Program: Evidence from West Bengal. **Authors**: Banerjee, Abhijit; Chattopadhyay, Raghabendra; Duflo, Esther; Shapiro, Jeremy **Journal**: AER: Insights. **Year published in repository:** 2017	The offer to participate in the program does not affect the outcomes in ways other than the program.	Attrition varies between 11 percent and 17 percent and does not differ significantly accross groups.	The authors note that there are potential spill-overs. For example, gifts of food given may increase, and these probably go to ultrapoor households that are not participating in the program. There are also possible equilibrium effects: previous papers show that wages of casual workers increase in treated villages, because the trated households work more in their houehold businesses and sell less labor.	Up to 1730 observations.
36	Title: Voter Registration Costs and Disenfranchisement: Experimental Evidence from France. Authors: Braconnier, Céline;Dormage, Jean-Yves;Pons, Vincent Journal: American Political Science Review. **Year published in repository:**2017	The canvassing and home visits does not affect voting behaviour in ways other than the program.	Not discussed.	The assignment of all apartments of a particular building to the same treatment condition reduces the scope for spillovers between the control and treatment groups.	Up to 20458 observations.

37	**Title**: Risk information, risk salience, and adolescent sexual behavior: Experimental evidence from Cameroon. **Authors**: Dupas, Pascaline; Huillery, Elise; Seban, Juliette **Journal**: Journal of Economic Behavior & Organization. **Year published in repository:**2017	The offer to participate in the training does not affect girls behavior in ways other than the program.	Out of 3154 girls in our sample, they obtained information (in-person interview or relative interview) for 2907 of them. This constitutes an overall 7.8% attrition rate (247 girls lost) for objective outcomes (pregnancy history and school enrolment)	Consultant sessions may be more attractive thanks to the use of videos and the expertise of the messenger, however, they provide only one session while teachers are encouraged to provide several sessions. In case of positive inter-class spillovers, it gives an advantage to the teacher training treatment over the consultant treatment.	Up to 2732 observations
38	**Title**: Promoting Handwashing Behaviour: The Effects of Large-scale Community and School-level Interventions. **Authors**: Galiani, Sebastian; Gertler, Paul; Ajzenman, Nicolas; Orsola-Vidal, Alexandra **Journal**: Health Economics. **Year published in repository:** 2017	The promotion of handwashing does not affect outcomes in ways other than the program.	The survey records show an overall attrition rate of 20% after 3 years.	Not discussed.	Up to 2847 observations.
39	Title: Increasing the Electoral Participation of Immigrants: Experimental Evidence from France. Authors: Pons, Vincent;Liegey, Guillaume Journal: Economic Journal. **Year published in repository:** 2018	Being assigned to a canvasser visit does not affect outcomes in ways other than the program.	Not discussed.	The assignment of all apartments of a particular building to the same treatment condition reduces the scope for spillovers between the control and treatment groups.	Up to 23760 observations.
40	**Title**: The Oregon Health Insurance Experiment: Evidence from the First Year. **Authors**: Finkelstein, Amy;Baicker, Katherine;Taubman, Sarah;Wright, Bill;Bernstein, Mira;Gruber, Jonathan;Allen, Heidi;Newhouse, Joseph P;Schneider, Eric;Zaslavsky, Alan. **Journal**: Quarterly Journal of Economics. **Year published in repository:** 2018	The offer to enroll in the OHP does not affect outcomes in ways other than the program.	50% nonresponse rate in the subsample of survey respondents; 97% match rate i.e. 3% "attrition rate" in credit report data.	Not discussed.	Up to 74922 observations

Centre of
Excellence
for Development
Impact and Learning

www.CEDILprogramme.org