How Much Should We Trust Observational Estimates? Accumulating Evidence Using Randomized Controlled Trials with Imperfect Compliance

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

January 12, 2024

Abstract

The use of observational methods remains common in program evaluation. How much should we trust these studies, which lack clear identifying variation? We propose adjusting confidence intervals to incorporate the uncertainty due to observational bias. Using data from 44 development RCTs with imperfect compliance (ICRCTs), we estimate the parameters required to construct our confidence intervals. The results show that, after accounting for potential bias, observational studies have low effective power. Using our adjusted confidence intervals, a hypothetical infinite sample size observational study has a minimum detectable effect size of over 0.3 standard deviations. We conclude that – given current evidence – observational studies are uninformative about many programs that in truth have important effects. There is a silver lining: collecting data from more ICRCTs may help to reduce uncertainty about bias, and increase the effective power of observational program evaluation in the future.

*Bernard: Paris School of Economics. Bryan: London School of Economics (g.t.bryan@lse.ac.uk). Chabé-Ferret: Toulouse School of Economics. de Quidt: Queen Mary University of London and Institute for International Economic Studies. Fliegner: University of Manchester. Rathelot: Institut Polytechnique de Paris (ENSAE). We gratefully acknowledge financial support from IPA and CEDIL. de Quidt acknowledges financial support from Handelsbanken's Research Foundations, grant no. P2017-0243:1. Fliegner thanks the International Association for Applied Econometrics (IAAE) for the IAAE travel grant for the 2018 IAAE Conference in Montreal. We thank Greg Fischer for early collaboration, and Steven Glazerman for wide-ranging support at multiple stages of the project. We thank Mitch Downey, Michael Gechter, Marc Gurgand, Pascal Lavergne, Rachael Meager, Christoph Rothe and Beth Tipton for comments and suggestions, as well as a host of great seminar and conference participants. We thank Sree Ayyar, Davi Bhering, Dominik Biesalski, Angie Ibrahim, Enora Messi, Ritu Muralidharan, Michael Rosenbaum, Daphne Schermer, Luis Schmidt, and Fabian Sinn for excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect those of any institution. All errors are our own.

The past decades have seen large advances in quasi-experimental program evaluation (Angrist and Pischke 2010). Despite this, naturally-occurring exogenous variation is hard to find, and there remains demand for methods that can be applied when there is no plausible natural experiment.[1] Two leading options are observational methods – such as matching and regression – that try to adjust for observable differences, and randomized controlled trials (RCTs), which explicitly generate their own exogenous variation. There is a strong trade-off between these methods. RCTs are often held up as the gold standard for identification, but they are costly to implement and non-trivial to manage.[2] Observational studies, in contrast, are logistically less challenging and probably cheaper, but have a remaining *observational bias* (OB) of unknown direction and magnitude.[3] Choosing between these two approaches requires weighing their costs and benefits. This is typically done through analytical argumentation, rather than empirical validation. Users of observational studies argue for an unconfoundedness assumption, while RCT advocates reply that these assumptions are rarely plausible, meaning that we learn little from observational studies.

We seek to move this debate onto an empirical footing, by treating observational bias as an object to be estimated. By doing so we can provide quantitative measures of the extent of uncertainty surrounding observational bias, which can be incorporated into standard statistics that summarize confidence in observational estimates.[4] We depart from much of the existing literature, inspired by LaLonde (1986), in emphasising that the primary problem with observational bias is *uncertainty*: we do not know its size nor its direction, so we cannot adjust for it. We have three goals for our approach. First, by incorporating measures of observational bias, researchers can be more honest about the uncertainty surrounding their estimates and can better understand whether observational approaches generate useful information about program impacts. Second, different observational methods can be compared in terms of how effective they are at reducing uncertainty

[1] Despite the lack of clear identifying variation, observational studies remain very popular, perhaps reflecting the difficulty of finding quasi-experimental variation or running an RCT. Appendix D Figure 10 shows the continued popularity of matching methods, a leading observational method, and the recent rapid growth of double debiased machine learning.

[2] RCTs may have their own sources of bias such as lack of blinding, implementation problems, demand effects etc. In addition, they cannot be applied to study all programs. We restrict ourselves to programs to which it would at least be plausible to implement an RCT.

[3] Observational methods, such as regression, attempt to control for observables in order to remove selection bias. We can decompose selection bias into two parts: selection on observables and selection on unobservables. The sum of these two is the bias in a standard comparison of means, while it is selection on unobservables that remains after an attempt to control. Beyond selection bias, breaches of SUTVA, or failure of common support may lead to bias. We group these together throughout, under the moniker *observational bias* or OB for short.

[4] We incorporate uncertainty regarding observational bias into a classic confidence interval. We believe this is the simplest addition to current practice, and hence the right place to start a research project in this domain.

about observational bias. Finally, RCTs and observational methods can be placed on an equal footing and compared using empirically-informed methods, such as power calculations.

Our analysis is restricted to observational methods that use a cross section of data. We consider a policy maker who has access to a large observational data set that includes variation in uptake of a program that they wish to evaluate. With the data they are able to generate an observational estimate, $\widehat{TOT}^{OBS}$, of the average treatment effect on the treated, with a standard error $\hat{\sigma}_\epsilon$.[5] We show that if this policy maker believes that the observational bias of their estimate is drawn from a Normal distribution with mean μ and standard deviation τ, then an appropriate two-sided confidence interval of size δ would be

$$\widehat{TOT}^{OBS} - \hat{\mu} \pm \Phi^{-1}\left(\frac{1+\delta}{2}\right)\sqrt{\hat{\sigma}_\epsilon^2 + \hat{\sigma}_\mu^2 + \hat{\tau}^2}, \tag{1}$$

where $\hat{\mu}$ and $\hat{\tau}$ are empirical counterparts for μ and τ, and $\hat{\sigma}_\mu$ is the standard error of $\hat{\mu}$.

This formula incorporates uncertainty about observational bias directly into a standard representation of parameter uncertainty, and helps clarify our goals. First, in additional to the usual estimates, our policy maker requires estimates $\{\hat{\mu}, \hat{\sigma}_\mu^2, \hat{\tau}^2\}$ of $\{\mu, \sigma_\mu^2, \tau^2\}$ (the mean observational bias, its standard error, and the true variability in observational bias). We can think of the square root term in (1) as an *effective standard error* that incorporates uncertainty about observational bias. Second, mean bias is not really a problem. If μ is known with precision (e.g., if $\widehat{TOT}^{OBS}$ is known to have a specific positive bias), it can easily be adjusted for. It is *uncertainty* in the estimate of μ ($\hat{\sigma}_\mu^2$), and the true variance of observational bias (τ^2) that matter. As noted, this is a key area in which our work differs from the seminal paper of LaLonde (1986) and the literature that followed.[6]

Third, efforts to increase the precision of observational estimates may be better focused on reducing uncertainty about bias than increasing sample size to reduce $\hat{\sigma}_\epsilon^2$. In this sense, studies like ours

[5] We assume throughout that TOT is the object of policy interest as it is the parameter most obviously identified in an observational study.

[6] LaLonde (1986), and other studies that focus on a single program, cannot estimate uncertainty about bias. However, even papers that report on multiple studies, so that there is some hope of estimating τ, focus on reporting bias for each study independently, or average bias across studies. For example, Glazerman et al. 2003; Chaplin et al. 2018; Forbes and Dahabreh 2020; Wong et al. 2017 all report estimates from multiple studies, but concentrate on average bias, rather than uncertainty. Without expecting to be exhaustive, additional papers in this literature also include Agodini and Dynarski (2004); Arceneaux et al. (2006); Dehejia and Wahba (2002, 1999); Eckles and Bakshy (2021); Ferraro and Miranda (2014); Fraker and Maynard (1987); Friedlander and Robins (1995); Gordon et al. (2019, 2023); Griffen and Todd (2017); Heckman and Hotz (1989); Heckman et al. (1998); Smith and Todd (2005).

that seek to increase understanding of observational bias can improve all future observational studies. Fourth, even with an infinite-sized observational study (so $\hat{\sigma}_\epsilon^2$ vanishes) uncertainty does not disappear: τ^2 will always remain and represents the uncertainty about identification that we tend to discuss in seminars and referee reports. Indeed, in large samples, uncertainty from observational bias will dominate the effective standard error, meaning observational bias becomes relatively more important for large studies that attempt to discover small effects, a fact that seems particularly important with the increased availability of very large observational data sets.

To estimate our three new objects ($\{\hat{\mu}, \hat{\sigma}_\mu^2, \hat{\tau}^2\}$) we proceed as follows. First, we build a new dataset containing micro data from a large number of randomized controlled trials with imperfect compliance (ICRCTs). The dataset was created using the Dataverses of the Abdul Latif Jameel Poverty Action Lab (J-PAL) and Innovations for Poverty Action (IPA), and we have 44 different trials, with an average of about 40 outcome variables per trial. These pioneering organizations have spearheaded the movement to evaluate development policy using RCTs, and their advocacy and hard work is what allows for our approach. The key assumption of our paper, and one that we discuss and defend throughout, is "exchangeability": given the information available to them, the policy maker would be willing to exchange estimates of bias from one of the studies with estimates from any of the others.[7]

Second, we show how to generate observational and experimental estimates of treatment effects that apply to the same population *within* each ICRCT. This ensures that any differences between estimates is driven by observational bias rather than differences in the population to which the estimates apply. We distinguish between two kinds of ICRCT. In *eligibility designs* the control group has no access to a program but the treatment group does. In *encouragement designs* both groups have access, but the treatment group receives some additional encouragement, for example a subsidy. In an eligibility design, under standard assumptions,[8] the RCT can be used to recover an experimental estimate of the TOT. It is also possible to form an observational estimate of the TOT using observations from the treatment group, if conditional independence, SUTVA, and

[7] Formally, we assume that the joint distribution of bias estimates is invariant to permutations of study IDs, see e.g. Higgins et al. (2008).

[8] Independence, First stage, SUTVA, and Exclusion. Independence says that assignment to treatment (eligibility) is independent of potential outcomes and potential take-up. First stage says that assignment to treatment increases the probability of take-up. SUTVA (Stable Unit Treatment Value Assumption) says that i's potential outcomes are independent of j's take-up. Exclusion says that assignment to treatment only affects outcomes through take-up. See Appendix A for formal definitions.

common support all hold.[9] In an encouragement design, again with standard assumptions,[10] an ICRCT allows an IV estimate of the causal effect of the program on those induced to take-up by the encouragement. We refer to this as the treatment effect on compliers (TOC). We show that the TOC can also be recovered as an observational estimate under the assumptions of conditional independence, common support, and SUTVA, using a scaled weighted average of observational estimates of the TOT in the treatment and control groups.

Third we compute, for each study, the difference between the experimental and observational estimates of either the TOT or TOC. Since we assume the RCT provides a consistent estimate of the true effect of interest, the difference yields an estimate of observational bias.[11] Naturally each bias estimate applies to a different sub-population, due to variation in study setting and design. Within the set of eligibility designs our bias estimates apply to takers within the treatment group, a group that will differ across studies. Within encouragement designs, our estimates apply to compliers who are a subset of the takers within the treatment group. Our primary results treat all these bias estimates as exchangeable: given current information, there is no clear reason to predict that the distribution of bias will differ systematically between sub-populations.

Our estimation methods are chosen to minimize any differences between observational and experimental estimates that are not caused by observational bias. We create observational estimates using "hands-off" procedures that do not require researcher input. This removes the possibility of deliberately or inadvertently tuning the observational estimate to match known experimental results, a potential weakness in the prior literature. We use three methods: naive comparison of means between those treated and not ("with-without", or WW); post double selection lasso (PDSL Belloni et al. 2014); and double-debiased machine learning (DDML Chernozhukov et al. 2018).[12] These methods were chosen as they can consistently estimate treatment effects in the presence of many nuisance parameters, while fulfilling our desire to remove researcher degrees of freedom. Our use of ICRCTs also means that experimental and observational estimates are created using

[9] Conditional independence says that potential outcomes are independent of take-up conditional on observables. Common support says that, there are comparable takers and non takers. See Appendix A for formal definitions.

[10] Independence, First stage, SUTVA, Exclusion, and Monotonicity. Monotonicity says that take-up is weakly increasing in assignment to the treatment (encouragement). See again Appendix A.

[11] We explore robustness of our results to different reasons that the RCT estimates themselves may be biased, for example failure of SUTVA. This does not qualitatively alter our results.

[12] We also experimented with a hands-off propensity score matching estimator that uses LASSO and cross validation for covariate and bandwidth selection. We did not pursue this further due to the difficulty of computing appropriate standard errors, and presence of some extreme outliers.

the same data set and surveying methods. This removes a concern with many studies following Lalonde where experimental and observational estimates were created with different data sets.

Finally, we use random effects meta-analysis to combine estimates from our 44 studies and recover our three key parameters.[13] This requires that all estimate use a common scale, so we make two normalizations. We measure bias in standard deviations of the control group outcome, and we align outcomes (based on a manual coding of "social desirability") such that a positive treatment effect always indicates an increase in welfare, i.e., a positive bias overestimates the welfare benefits of the program while a negative bias underestimates it.

The results of applying these methods to our 44 studies are surprising. First, we find that there is little bias on average. Using our best-performing observational method (DDML), there is a statistically insignificant and modest bias of -0.047 standard deviations. This implies that observational studies do not systematically over or under estimate the welfare impact of the programs they evaluate. Second, variability is large. The standard error of the average bias is about 0.035, while our estimate of τ is about 0.161. Interpreting these numbers through the lens of the confidence interval in (1), the effective standard error of an infinite-N observational study is 0.165 standard deviations. In many areas of study, for example health programs, a 0.2 standard-deviation impact is considered large. The minimal detectable effect size (MDE) for an infinite-N observational study using our confidence intervals would be more than 50% larger than this.[14] Third, we find substantial variation in the performance of observational methods. While DDML does reduce variance relative to a naive comparison of means, decreasing the effective standard error, PDSL performs less well and in some specifications increases uncertainty. Finally, we ask at what sample size an RCT has a smaller expected standard error than an infinite-N observational study. We find that a perfect-compliance RCT can have a smaller expected standard error with just 148 observations. Things look better for observational studies if there is imperfect compliance, but with only 25% compliance an RCT would still only need about 2400 observations to dominate.

Overall, we summarize the results as follows: while RCTs have their own weaknesses, given a

[13] Most studies have a large number of outcome variables. We take two approaches to deal with this. First, we combine all outcomes in a single index in the method of Anderson (2008). Second, we treat each outcome in each study as a separate estimate and then deal with intra-study correlation when we aggregate our results.

[14] Minimal detectable effect size is a notion often used in experimental design and records the smallest possible true effect that can reliably be estimated with statistical significance.

realistic assessment of current knowledge, observational studies that use a cross section of data produce very limited information about the effectiveness of important social programs with effect sizes that many would deem very large.

We take several steps to validate our methods. Perhaps the most striking is in terms of coverage rates. The coverage rate is the average number of times the experimental estimate falls into the confidence interval of the observational method. Using our preferred specification, we find that regular confidence intervals have a coverage of only 70%, while our corrected confidence intervals lead to a coverage of 94% – tantalizingly close to the nominal value of 95%. Our method achieves this by lowering the significance rate of observational estimates from 23% to just 4.2%, implying that about 20% of the observational estimates are incorrectly declared significant using conventional confidence intervals. Naturally this entails lower power: the implied power of observational methods falls from 41% to 14%.

Our key assumption is what we have called exchangeability – the policy maker has insufficient information to base her effective standard error on bias estimates from a subset of studies, so uses all available studies to estimate the distribution of bias. We argue empirically that, given our data, using the full set of studies is the policy maker's best option if her goal is to maximize power. We begin by arguing that it is appropriate to measure the gains from restricting the set of bias estimates by looking at the *expected* effective standard error across reasonable subsets. This is the effective standard error that an uninformed policy maker should anticipate. We then show empirically that the expected effective standard error for reasonable subsets is always higher than that from using all the data. This empirical finding reflects a theoretical trade-off. Using a subset of data may reduce the variance of bias (τ^2), but reduces sample size and foregoes shrinkage. Overall we believe that exchangeability across all studies is the right place to start, but we also argue there are hints of the value of continuing to run more ICRCTs, because in additional to answering its own research question, a new ICRCT can also contribute to more precise measures of observational bias in specific settings.

Our paper is inspired by the pioneering work of LaLonde (1986). Relative to that paper, and much of the literature that follows it, we concentrate on quantifying uncertainty about observational bias. Our use of ICRCTs and access to micro data means we can use the same data sets to estimate experimental and observational estimates, and we emphasise the use of hands-off estimators to

reduce researcher degrees of freedom. The contemporary paper Gechter and Meager (2022) is complementary to our work. That paper shows how to use an instrumental variable (arrival of J-PAL in a country, which lowers the cost of implementing an RCT), to estimate the extent of observational bias for a set of complier studies. By comparing the results of these complier studies to observational estimates, from which they subtract their estimate of average bias, they are also able to estimate the extent of site-selection bias under the assumption that complier and always-taker RCTs have the same average estimates. Relative to our work they concentrate on studying average bias, while we place more emphasis on uncertainty. They also concentrate on two literatures – microcredit and cash transfers – while we consider a broader set of studies. Their paper, however, raises the possibility that our RCT data base may not be representative of all observational settings, because of site-selection bias. This limits the application of our methods to places where it would be plausible to run an ICRCT. Empirically, they find very limited although noisily estimated site-selection bias.

The paper is structured as follows. Section 1 summarizes the methods we use to estimate and aggregate bias, section 2 describes our data set of ICRCTs and provides some model diagnostics and section 3 summarize the results of our main meta-analysis. Section 4 discusses robustness to relaxing the exchangeability and other assumptions, section 5 provides our analysis of the value of collecting more ICRCTs, and section 6 concludes.

1 Overview of Methods

In this section we give an overview of the methods we use to estimate $\{\mu, \sigma^2_\mu, \tau^2\}$, and the assumptions under which the confidence interval in equation (1) makes sense. We produce our estimates in two steps, we first estimate bias in each of our studies, then we combine these estimates using meta-analysis. We describe each step in turn.

1.1 Study-Level Estimators of Bias

Our goal is to provide corrected confidence intervals that account for uncertainty about observational bias. We envisage these being used by a policy maker who has access to an observational data set in which some subjects have adopted a program. It is well understood that under the

three assumptions of conditional independence, common support and SUTVA,[15] a data set of this kind can be used to form an estimate of the population treatment effect on the treated (TOT). Given this result we consider the TOT to be the policy maker's parameter of interest. We assume that the policy maker is able to form an observational estimate of TOT, which we denote $\widehat{TOT}^{OBS}$. To avoid confusion we refer to the population analog of this estimate, TOT^{OBS}, as an estimand.

We aim to estimate the bias $B_0 = TOT^{OBS} - TOT$. If the conditional independence, SUTVA, or common support assumptions fail, TOT^{OBS} may not be equal to TOT. We want to include all these sources of bias in our estimates, after making our best effort to minimize them using methods discussed below. Because we do not directly observe TOT, we will form our estimand and eventually estimator of bias as $B = TOT^{OBS} - TOT^{EXP}$, where TOT^{EXP} is the plim $\widehat{TOT}^{EXP}$ of an experimental estimator formed from an ICRCT. We denote $\widehat{B} = \widehat{TOT}^{OBS} - \widehat{TOT}^{EXP}$ our estimate of bias, formed by taking the differences between observational and experimental estimates.

If the estimator resulting from TOT^{EXP} is close to the true TOT, then $\widehat{B}$ will be a good estimate of the bias B_0 that we are interested in. Our experimental estimator may differ from TOT for two broad reasons. First, in the presence of heterogeneous treatment effects, the experimental estimator may apply to a different subset of the population than the population-level TOT that we are aiming to estimate. Second, we will need standard identification assumptions to hold. We discuss each of these issues in this section, first for eligibility designs, and then for encouragement designs.

Eligibility designs make a program available to a randomly chosen subset of the study population (the treatment, or eligible group). Imperfect compliance in this design occurs when not all eligible subjects take up the program. With an eligibility design it is relatively easy to ensure that both experimental and observational estimates apply to the same population. To obtain TOT^{EXP}, we use the Bloom estimand, which is the ratio of the intent to treat estimand and the compliance rate, to estimate an experimental treatment effect (Bloom 1984). It is well known that under standard assumptions for the validity of the RCT (Independence, first stage, SUTVA, and exclusion) the Bloom estimator recovers an experimental estimate of TOT, or the average treatment effect among the set of people who take up the program (e.g., Angrist and Pischke 2009). It is also well known that under two additional assumptions – conditional independence, and common support – we

[15] Appendix A provides formal definitions of all the identification assumptions discussed in this section.

can use observations from the eligible group to form an observational estimator that also estimates the *TOT* (essentially comparing those that take up to those that do not, conditional on observables; see below for details of the estimators we use). It then follows that $\widehat{B} = \widehat{TOT}^{OBS} - \widehat{TOT}^{EXP}$ is a good estimator of observational bias so long as the assumptions for the validity of the RCT hold.

Our approach to validating the experimental identification assumptions is threefold. First, we have concentrated on gathering data from high-quality RCTs, most of which have been published in top economics journals, as we discuss below. Second, Appendix F provides a description of each study, where the reader can evaluate the assumptions themselves. Finally, it is possible to exclude potentially problematic studies from our sample. We pursue this approach in section 4 below, and argue there that our results are qualitatively robust to the exclusion of these studies.

The next section discusses in detail how we aggregate these estimates across studies, but one issue is worth noting at this point. The set of people who choose to select in will be different in each study, and so the treated group to which the *TOT* applies will change. Our approach to this issue is similar to our approach throughout. We believe that there is a complete lack of knowledge about differences in the distribution of bias across population groups, and so we will treat the estimates as exchangeable with the policy maker's study of interest.

Encouragement designs, in contrast, randomly incentivize take-up of a program that is available to everyone. Imperfect compliance can occur in this design in the treatment *and* control groups when not all subjects take up the program. For studies of this type it is well known that under the same assumptions (independence, first stage, SUTVA, and exclusion) plus monotonicity, the Wald ratio, which is the intent-to-treat effect divided by the difference in compliance rates across treatment arms, results in an experimental estimand of the treatment effect for the compliers (those who are induced to take up by the incentive): TOC^{EXP} (Imbens and Angrist 1994). It is also well known that it is not possible to form an experimental estimand of the *TOT* with an encouragement design, which would appear to create a problem for us.

We address this problem as follows. We show in Appendix A that under the assumptions of conditional independence, SUTVA, and common support

$$TOC^{OBS} = \frac{TOT^{OBS}_{treat} Pr(D = 1|treat) - TOT^{OBS}_{cont} Pr(D = 1|cont)}{Pr(D = 1|treat) - Pr(D = 1|cont)}$$

is an estimand for the treatment effect on compliers. In this expression, TOT^{OBS}_{treat} is an observational estimand of the TOT based on the observations of the study's treatment group, TOT^{OBS}_{cont} is the same for the study's control group, and $Pr(D = 1|t)$ is the probability of take-up in group $t \in \{cont, treat\}$. If we have consistent estimators for TOT^{OBS}_{treat} and TOT^{OBS}_{cont}, the empirical counterpart of TOC^{OBS} results in a consistent estimator for the treatment effect on compliers. This expression makes intuitive sense. The term $TOT^{OBS}_{treat} Pr(D = 1|treat)$ tells us how much the average outcome in the treatment group is increased by the program, while $TOT^{OBS}_{cont} Pr(D = 1|cont)$ tells us the same for the control group. Any difference between these two averages must come from a combination of two effects: a difference in the share of takers; and the size of the treatment effect for the compliers (TOC). Dividing by $Pr(D = 1|treat) - Pr(D = 1|cont)$ removes the first effect, leaving only the TOC.

With an experimental and an observational estimator for the treatment effect for compliers in hand, if the assumptions for experimental validity hold, then $\widehat{B} = \widehat{TOC}^{OBS} - \widehat{TOC}^{EXP}$ is a good estimator of observational bias in the estimator of TOC. Our argument regarding validity of the experimental assumptions is the same as made above for eligibility designs.

Our final concern is that our goal is to estimate the bias in observational estimates of TOT, not TOC. Once again, our response is to note that we have very little information that could be used to rank the extent of bias in an estimate of TOC, relative to an estimate of TOT. Given this, we think it is reasonable to argue that our hypothetical policy maker would be willing to assume that an estimate of the bias in TOC is exchangeable with her desired estimate of the bias in TOT for her setting. It should be noted that this is essentially the same assumption that was required to aggregate estimates of the bias in TOT: the policy maker is willing to assume exchangeability across different complier populations. We also show in Appendix D that excluding encouragement designs altogether only serves to increase the effective standard error, and thus reduce power.

In summary then, for each eligibility-design study $s = 1, ..., S$, and each outcome variable $o = 1, ..., N_s$ available within that study, our bias estimator is:

$$\widehat{B}_{os} = \widehat{TOT}^{OBS}_{os} - \widehat{TOT}^{EXP}_{os}$$

whereas for encouragement-design studies we have:

$$\widehat{B}_{os} = \widehat{TOC}_{os}^{OBS} - \widehat{TOC}_{os}^{EXP}.$$

We discuss how we deal with having multiple outcomes per study later in this section. We also calculate a standard error $\hat{\sigma}_{B,os}$ for each outcome-study pair. Appendix B explains how we do this.

1.2 Choice of Observational (Hands-off) Estimators

To create our bias estimates we need to decide on estimators. The choice of estimator has been a concern in much of the literature that builds on LaLonde (1986). If a researcher has access to the experimental estimate prior to choosing an observational estimator, then the researcher has some latitude to choose an estimator that comes close to approximating the experimental estimate. This does not need to be intentional, the researcher may be influenced by results in the literature or contemporaneous theorizing (the garden of forking paths).[16] To overcome this problem we exclusively use "hands-off" estimators, which allow very limited researcher degrees of freedom. Here we are greatly helped by recent econometric advances which build on machine-learning methods to consistently estimate treatment effects in the presence of a high-dimensional set of nuisance parameters (e.g., Belloni et al. 2014 and Chernozhukov et al. 2018). In essence these methods use machine learning to select from a very large set of potential covariates, an approach that is helpful in our setting where we have an average of over 400 covariates per study.

We implement three hands-off estimators. First, a naive "with and without" estimator (WW), which simply compares outcomes for those who chose to take up the program ("with"), to those who did not ("without"). Second, the post double selection lasso (PDSL) of Belloni et al. (2014). Third, the double debiased machine learning (DDML) approach of Chernozhukov et al. (2018). The PDSL and DDML approaches are similar in spirit, so here we give only a brief discussion of DDML, see Appendix B for full details.

We apply the DDML method to a partial linear model, and proceed (roughly) as follows. First, the sample is split into a training and testing set. On the training set, we use a regularized machine-learning method to create a prediction, for each subject, of the outcome without take-up,

[16] The researcher might also face incentives to choose an observational estimator that poorly reproduces the experimental estimate, depending on their motivations.

and the probability of take-up. This "double" prediction, one for outcome and one for take-up, is what gives the approach its name. In the testing set we then regress the difference between the observed outcome and predicted outcome without take-up on the difference between observed take-up status and predicted take-up status. We repeat this process with multiple splits and report the average coefficient on take up.[17] Splitting helps reduce concerns about over-fitting. When implementing this approach we use all available covariates X and the regularization in the ML method implicitly chooses which controls to use.

Chernozhukov et al. (2018) show that this approach leads to consistent estimates of treatment effects when conditional independence holds given the set of covariates X, even if the set of covariates is large. Importantly for our application, it requires very little researcher input beyond choosing some tuning parameters for the learners.[18]

When implementing DDML we always use a random forest as the machine learning method, because this means we do not have to choose whether to include interactions or higher order terms in the control set. When we use PDSL we include only linear terms.

1.3 Experimental Estimator

We produce our experimental estimates using a basic 2SLS regression including all strata dummies, but no other controls.[19]

1.4 Aggregating Estimates of Bias and Forming Confidence Intervals

We first discuss how we aggregate outcomes assuming there is only one outcome per study. Then we show how we extend the analysis to the case of multiple outcomes per study.

[17] One way to get intuition for why this works is to note that it can be interpreted as using the deviation from predicted take-up as an instrument, in a regression with deviation from predicted outcome as the left hand-side variable. The deviation from predicted take-up is excluded in this setup because, by the conditional independence assumption, the deviation from prediction is purely random noise which determines why some individuals take-up despite having the same observables.

[18] We make use of default software parameters throughout to further minimize researcher degrees of freedom, see Appendix B.

[19] We could in principle include additional covariates when generating experimental estimates, e.g. again using PDSL or DDML, but since randomization implies that covariates are not needed for identification we focus on the simple experimental estimator.

Assume the policy maker believes her observational estimate is drawn from a normal distribution

$$\widehat{TOT}_p^{OBS} \sim \mathcal{N}(TOT_p + B_p, \sigma^2_{\epsilon,p}),$$

where p denotes the policy maker's study of interest. $\sigma^2_{\epsilon,p}$ is the standard error of her estimate based on sampling error, while B_p is the unknown observational bias of her study.

Next, we assume that the policy maker believes that B_p is drawn from the same distribution as the bias in each of our studies:

$$B_p \sim \mathcal{N}(\mu, \tau^2), \textit{ and } B_s \sim \mathcal{N}(\mu, \tau^2), \textit{ for } s \neq p \tag{2}$$

where μ is the true mean bias, and τ^2 the true variance of bias across studies. Introducing this notation immediately raises the question of how to interpret μ, in particular its sign. We will define a positive bias as one that exaggerates the welfare benefits of the program studied, and code our data accordingly. A finding of a positive mean bias would then suggest that the types of people that choose to select into programs are the types of people who would have done relatively well, even without the program. A positive mean bias would also imply that, all things being equal, policy makers relying on observational studies will tend to recommend programs that are in fact not beneficial. A negative bias has the opposite interpretation. τ^2 measures the variance in observational bias across programs, and is in some sense a measure of our ignorance. We discuss in detail below how one might go about reducing τ^2.

Condition (2) may seem like a strong assumption, but it is a simple way to capture our key exchangeability assumption, and we show later that it approximates the data well.

We wish to use our set of estimates $\{\hat{B}_s, \hat{\sigma}_{B,s}\}$ to form estimates of μ and τ^2. To do this, we assume that for each study s in our set of studies

$$\hat{B}_s = \mu + \eta_s + \nu_s \tag{3}$$

where, in line with (2), $\eta_s \sim \mathcal{N}(0, \tau^2)$, and ν_s is a sampling noise distributed $\mathcal{N}(0, \sigma^2_{\hat{B},s})$, which follows from the Central Limit Theorem. As standard in this literature, the variance $\sigma^2_{\hat{B},s}$ is replaced by our estimated variance $\hat{\sigma}^2_{\hat{B},s}$. Equation (3) describes a random-effect meta-analysis, which can

be efficiently and consistently estimated using Restricted Maximum Likelihood (Raudenbush, 2009; Chabé-Ferret, 2023).

Performing this analysis requires that our outcomes are measured in a common metric, so we make two normalizations. To make units of measurement comparable across studies, we express all bias estimates in units of standard deviations of the control-group outcome variable in that study. Second, in line with our interpretation of positive bias as exaggerating welfare benefits, we align the sign of all outcome variables by coding outcomes for "social desirability."[20] Our meta-analysis then returns $\{\hat{\mu}, \hat{\tau}^2, \hat{\sigma}^2_\mu\}$ as desired.

Finally, we can use these estimates to build an appropriate confidence interval for a hypothetical policy maker study p for which an observational estimate $\widehat{TOT}^{OBS}_p$ has been constructed, with standard error $\hat{\sigma}_{\epsilon,p}$. It follows from equation (3), and the normality of the error, that $\widehat{TOT}^{OBS}_p \sim \mathcal{N}(TOT_p + \mu, \sigma^2_{\epsilon,p} + \tau^2)$, with the implication that

$$\widehat{TOT}^{OBS}_p - \hat{\mu} \sim \mathcal{N}(TOT_p, \hat{\sigma}^2_{\epsilon,p} + \hat{\sigma}^2_\mu + \hat{\tau}^2),$$

which leads to the confidence interval formula (1) discussed in the introduction.[21]

Figure 1 gives a useful visual presentation of this confidence region, with solid lines representing the usual confidence intervals, and dashed lines representing bias adjusted confidence intervals. The x-axis (labelled "treatment effect") represents either $\widehat{TOT}^{OBS} - \hat{\mu}$ when considering a bias corrected confidence interval, or $\widehat{TOT}^{OBS}$ when considering a regular confidence interval, and the y-axis is $\hat{\sigma}_{\epsilon,p}$, which is specific to our policymaker's observational study. In both cases, studies outside of the funnel would be considered to have statistically significant effects, and studies with effects that lie inside the "tram lines" between the solid and dashed lines would be declared significant with standard confidence intervals, but not with our bias-adjusted intervals.

The diagram helps motivate several important observations. First, as we have already noted, it is

[20] A socially desirable outcome is one where a positive effect would increase social welfare, all else equal (e.g., income, health), a socially undesirable outcome has the opposite interpretation (e.g. child mortality, crop losses), and some outcomes are ambiguous (e.g. voting outcomes). We flip the sign of socially undesirable outcomes, and drop ambiguous cases.

[21] The result follows from the fact that $\hat{\mu} \perp \widehat{TOT}^{OBS}_p$ and that they are both normally distributed. As a consequence, $\text{Var}(\widehat{TOT}^{OBS}_p - \hat{\mu}) = \sigma^2_{\epsilon,p} + \tau^2 + \sigma^2_\mu$. Replacing the variance terms by their estimates gives the formula that we actually use.

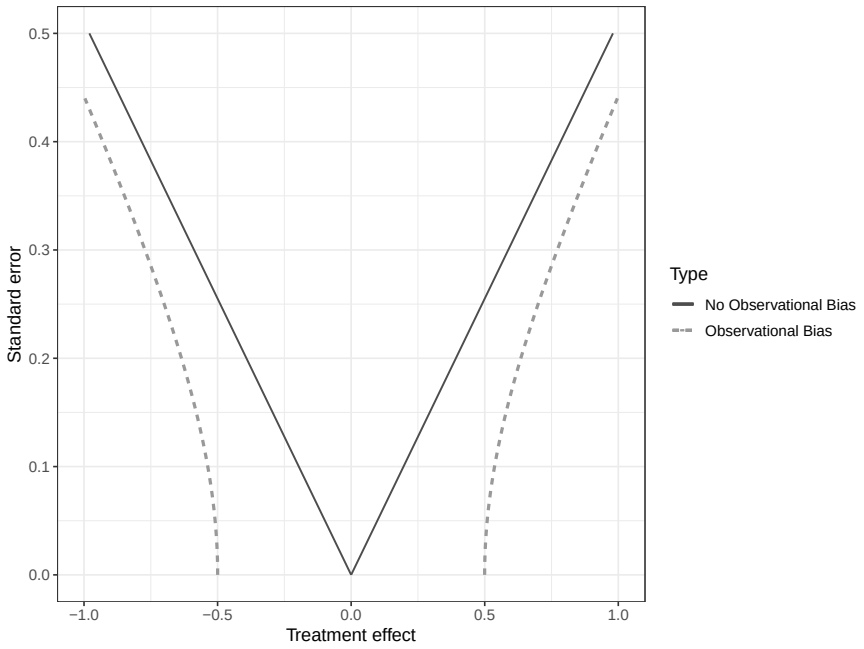


Figure 1: Funnel Plot Showing Examples of Adjusted and Unadjusted Confidence Intervals

uncertainty about the extent of the bias, captured by $\hat{\tau}$ and $\hat{\sigma}^2_{\mu}$, that poses a problem when using observational methods, rather than the mean bias itself. Our policy maker does not need her observational method to accurately estimate the treatment effects, as long as she knows the size and direction of the bias. This is a key area in which we depart from earlier work building on Lalonde (1986). The majority of this work, even where there are multiple studies so that there is some hope of estimating τ, focus on reporting bias for each study, or average bias across studies.[22] Second, we are used to thinking of large-N studies as having high power, but that need not be the case here. Even a very large observational study with $\hat{\sigma}_{\epsilon}$ approaching zero may have little power to detect policy-relevant effects if there is much uncertainty about the extent of observational bias. One interpretation of our empirical results below is that observational studies have significantly less power than is usually thought. A corollary of this observation is that the only way to increase power across a range of observational studies that already have large sample size is to increase precision in estimates of observational bias, which will tend to decrease $\hat{\sigma}^2_{\mu}$, or allowing the policy maker to concentrate on a set of ICRCTs that are more similar to her own, and allow an expected reduction in $\hat{\tau}$. This can potentially be achieved by running and aggregating evidence from more ICRCTs. Finally, our concerns about observational bias are less relevant for small-N observational

[22]For example, Glazerman et al. 2003; Chaplin et al. 2018; Forbes and Dahabreh 2020; Wong et al. 2017 all report estimate data from multiple studies, but concentrate on average bias, rather than uncertainty.

studies (where large conventional standard errors drive most of the uncertainty), but dominate for large-N studies whose conventional standard errors approach zero. This observation seems quite important to us, give the increasing availability of very large observational data sets.

1.5 Extension to Multiple Outcomes Per Study

As we will discuss in detail below, each of our studies includes multiple outcome variables. This has become the norm in project evaluations run by development economists, with each study reporting a range of different outcomes that might be considered to be positive or negative from a welfare perspective. In principle this can provide useful additional data — multiple bias observations per study — that could be informative about our parameters of interest. But biases within a study may be correlated and we need to deal with that correlation structure. Moreover, it is a priori unclear which of those outcomes best represents the welfare measure our policy maker would be interested in.

We pursue two different approaches to this problem. First, we aggregate all outcomes in a single study into one indexed outcome following Anderson (2008). We think of this as being a (hands-off) approximation of the welfare function that a policy maker might have in mind. Consistent with the arguments in Viviano et al. (2021) we use a precision-weighted average, which corresponds to a welfare function that puts equal weight on each outcome, an approach that we find appropriate given the lack of detailed information on policy makers' preferences.

Second, we persist with the multiple outcomes, but adjust for the possibility that within-study outcomes are correlated, so that we do not exaggerate the precision of our own estimates. To do this, we remain in the classical meta-analytical framework, but follow Pustejovsky and Tipton (2021) in allowing for some within-study correlation in both effects and errors. Specifically, we generalize (3) to:

$$\hat{B}_{os} = \mu + \omega_s + \iota_{os} + \nu_{os}$$

where $\iota_{os} \sim N(0, \xi_\iota^2)$, $\omega_s \sim N(0, \xi_\omega^2)$ and ν_{os} is again a normal error term, but with $Cov(\nu_{os}, \nu_{o's}) = \rho\hat{\sigma}_s^2$ and ρ is a "known" parameter that we set to 0.6. Let N_s be the number of outcomes per study s. Each bias estimate has a standard error $\hat{\sigma}_{B,os}$ and $\hat{\sigma}_s^2 = \frac{1}{N_s}\sum_{o=1}^{N_s}\hat{\sigma}_{B,os}^2$ is the average sampling variance for study s. The interpretation is that each study draws a bias $\mu + \omega_s$, there is an additional draw ι_{os} for each outcome within s and that the sampling errors are potentially

correlated within study. We then report confidence intervals

$$\widehat{TOT}^{OBS} - \hat{\mu} \pm \Phi^{-1}\left(\frac{1+\delta}{2}\right)\sqrt{\hat{\sigma}^2_{\epsilon} + \hat{\sigma}^2_{\mu} + \hat{\xi}^2_{\omega} + \hat{\xi}^2_{\iota}}. \tag{4}$$

From now on we denote $\hat{\tau}^2 = \hat{\xi}^2_{\iota} + \hat{\xi}^2_{\omega}$, so that the total variance is the sum of the within and between variances. This approach amounts to assuming that the policy maker has one outcome in mind, and believes that it is exchangeable with any outcome in our data set.

1.6 Distinguishing primary and secondary outcomes

An obvious critique of an approach that uses all outcomes available from a given study is that many of them may have been collected for robustness checks or secondary analysis. The policy maker may not be too concerned if those estimates suffer from observational bias, provided her primary outcome(s) of interest are unbiased.

We therefore attempt to distinguish between primary and secondary outcomes, once again using a hands-off approach. Namely, we code as primary any outcome that is mentioned in the abstract of a paper, and produce analysis for only these outcomes (either individually or aggregated using the indexing approach described above). We also produce estimates for all outcomes in the paper, either aggregated or individually (dropping those that are neutral with respect to social welfare). Together this gives us four different meta-analyses.

1.7 Quality Checks

To ensure the quality of our estimates we take the following steps. First, we automatically determine the experimental design of a specification, where a specification is a combination of estimator study and outcome. To do this we calculate the normalized minimum detectable effect (NMDE) on each treatment arm. If the NMDE is greater than 1 we conclude that there is insufficient take-up in that arm to form a reliable observational estimate. If the NMDE is greater than one in the control, we force take-up to be zero for all observations in that arm. For the treatment we force take-up to be 1. The design is then determined accordingly: perfect compliance if take-up is always zero in control and one in treatment; eligibility if take-up is always zero in control and a mix of zeros and ones in treatment; and encouragement if there are a mix of zeros and ones in both treatment and control. Second, we remove outliers from both the

aggregated and individual outcomes. We remove all outcomes where the absolute value of the normalized experimental estimator is larger than two standard deviations.[23] Third, we remove weak instruments by only keeping specifications with a Kleibergen-Paap F-statistic larger than 10.

When aggregating outcomes, we group the outcomes by study, treatment, take-up and unit of analysis (e.g. individual vs. group level outcomes) and aggregate the outcomes within these groups. That means that we still have several specifications per study left after aggregation. For example, we might have an individual level aggregate, and a group level aggregate. To come to a single outcome per study we select the most powerful specification in each study: we multiply the share of compliers by the number of experimental units and select the estimate with the highest value.

2 Data Description

Two important advances make our approach feasible, one methodological and one practical. On the methodological side, modern approaches such as DDML allow us to create hands-off observational estimates, even in the presence of very large sets of covariates. On the practical side, our approach requires a large set of ICRCTs. Here we are in debt to the pioneering work of two organizations, the Abdul Latif Jameel Poverty Action Lab (J-PAL) and Innovations for Poverty Action (IPA). Since their founding in 2002 and 2003 respectively, these two organizations have worked to encourage the use of randomized policy evaluations across the developing world. Because our approach requires access to micro-data, we access data from many of these RCTs hosted on their respective Dataverses. In this section, we describe how we select studies, and describe the studies that are in our sample.

2.1 Study Selection

We start with 207 studies from the IPA and J-PAL dataverse. Within this set of studies, we select those studies that have imperfect compliance, a variable recording random treatment assignment, a variable recording program take-up, and at least one outcome variable. This leaves us with a sample of 44 ICRCTs (see Appendix C for details about the screening process). We have on average

[23] To be conservative we do not only remove outliers based on he experimental estimate resulting from the 2SLS regression without covariates but also outliers based on an experimental estimate resulting from an estimation of a partially linear IV regression model (Chernozhukov et al. (2018)) including the same controls as for the estimation of the observational estimate.

41 specifications (outcome–treatment–take-up–level-of-analysis combinations) per study, and six primary specifications (those mentioned in the abstract) per study. Our largest meta-analysis has 1797 outcome-study pairs. For additional details on study-level summary statistics, see Appendix G.

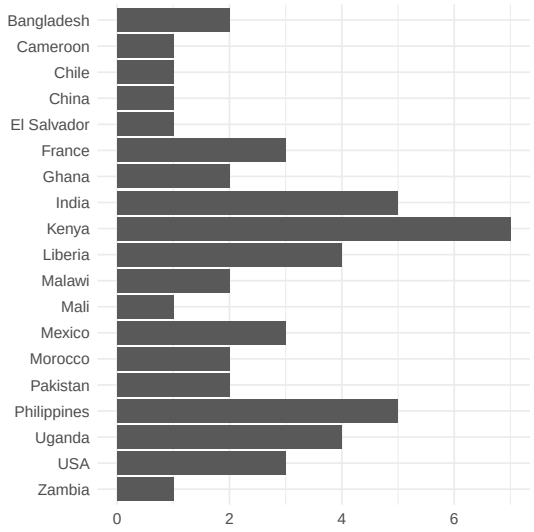
2.2 Description of ICRCT Sample

Here we provide a high-level overview of the 44 studies that we use in our analysis. Summaries of each individual study are provided in Appendix F.

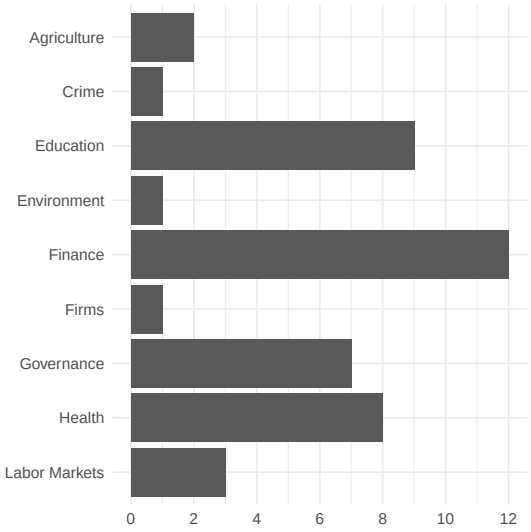
Figure 2 shows counts of four characteristics of our studies: country, sector/topic, journal and author. Panel 2a shows that our studies come almost entirely from developing countries, reflecting the goals of J-PAL and IPA. We have studies from Africa, South America, and Asia, as well as North America (USA) and Europe (France). Studies from countries with IPA or J-PAL hubs are strongly represented, similar to the development economics literature more broadly. Kenya appears the most in our analysis, with India, the Philippines, Uganda and Liberia also being highly represented.

We use J-PAL's eleven sectors to categorise our studies by topic in panel 2b. The most represented sectors are finance, education and health, all common areas of study within development economics. Our studies are published in a range of journals as shown in panel 2c. We have twelve papers from top-five journals in economics: six papers from The Quarterly Journal of Economics, four from the American Economic Review, and one each from Econometrica and the Journal of Political Economy. Ten 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. We also have a few studies published in non-economics journals, signifying 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.

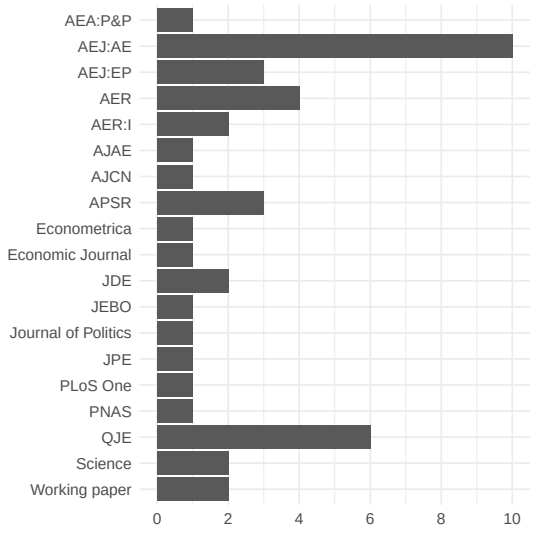
Finally, panel 2d shows authors who appear at least twice in our dataset. Almost all of these authors are prominent development economists, with Dean Karlan, Pascaline Dupas and Esther Duflo appearing frequently.



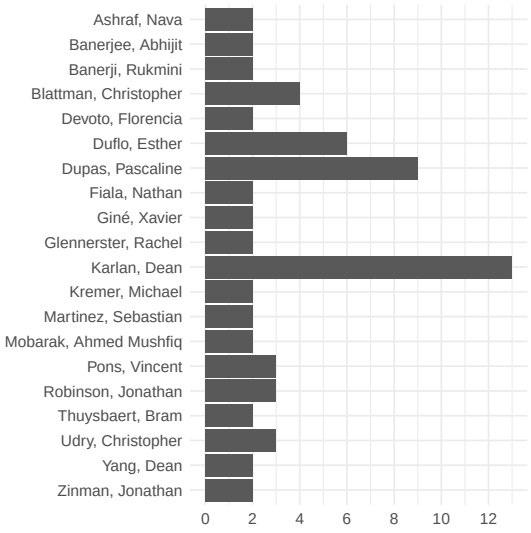
(a) Country



(b) Sector



(c) Journal



(d) Authors (with at least 2 RCTs)

Figure 2: Study Characteristics

We provide detailed information on each of the 44 included studies in Appendix F.[24] We rate the quality of each in Appendix G. In particular, we provide information relevant to determining whether the key assumptions for RCT validity, including SUTVA and exclusion, hold. Overall, we think the quality is high. All of our studies are RCTs, run by J-PAL or IPA and almost all are published in high quality journals. This reassures us that the RCTs identify consistent causal effects and as such, our comparison between observational estimators and RCT estimators should provide a good estimate of observational bias.

2.3 Model Diagnostics

In this section we provide some evidence on the appropriateness of our model. Because of its importance we provide a detailed analysis of exchangeability in Section 4. As noted above, the key parametric assumption we used to model exchangeability is that bias is drawn from a normal distribution. Figure 3 shows the raw distributions of estimated biases in our sample of studies, the top two panels show the distributions for the aggregated outcomes while the bottom two panels show the distribution for all outcomes. Interpreted through the lens of our meta-analysis model, these raw biases are a combination of the true bias in the study and a normally distributed sampling error. If the underlying true bias distribution is normal, then the raw bias distribution will also be normal. Visual inspection suggests that the distributions are sufficiently close to normal that there is no obvious alternative distribution to use.

3 Main Results

Table 1 summarizes the results of our meta-analysis, and gives our estimates of $\{\hat{\mu}, \hat{\sigma}^2_\mu, \hat{\tau}^2\}$ broken down by observational method.[25] The first panel shows results for the aggregated primary

[24] The papers we have in our study are: Ashraf et al. (2006), Blattman et al. (2014a), Giné et al. (2010), Bryan et al. (2014), Dupas and Robinson (2013a), Dupas and Robinson (2013b), Dupas (2011), Guiteras et al. (2015), Angelucci et al. (2015), Ashraf et al. (2009), Duflo et al. (2015), Crépon et al. (2015), Dupas et al. (2016), Cohen et al. (2015), Baldwin et al. (2016), Blattman and Annan (2016), Ambler et al. (2015), Blattman et al. (2017), Dupas et al. (2018b), Karlan et al. (2017), Bruhn et al. (2018), Fink 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), Banerji et al. (2017), Banerjee et al. (2007), Braconnier et al. (2017), Dupas et al. (2018a), Pons and Liegey (2019), Blattman et al. (2014b), Bloom et al. (2015), Behaghel et al. (2017), Gerber et al. (2009) and Finkelstein et al. (2012).

[25] All individual outcomes are based on 44 different studies after applying robustness checks for outliers in the experimental effects and removing weak instruments. When aggregating all (primary) outcomes, Dupas et al. (2018a) is removed because of outliers in the experimental effects. Karlan et al. (2016) does not have any primary outcomes after applying our robustness checks.

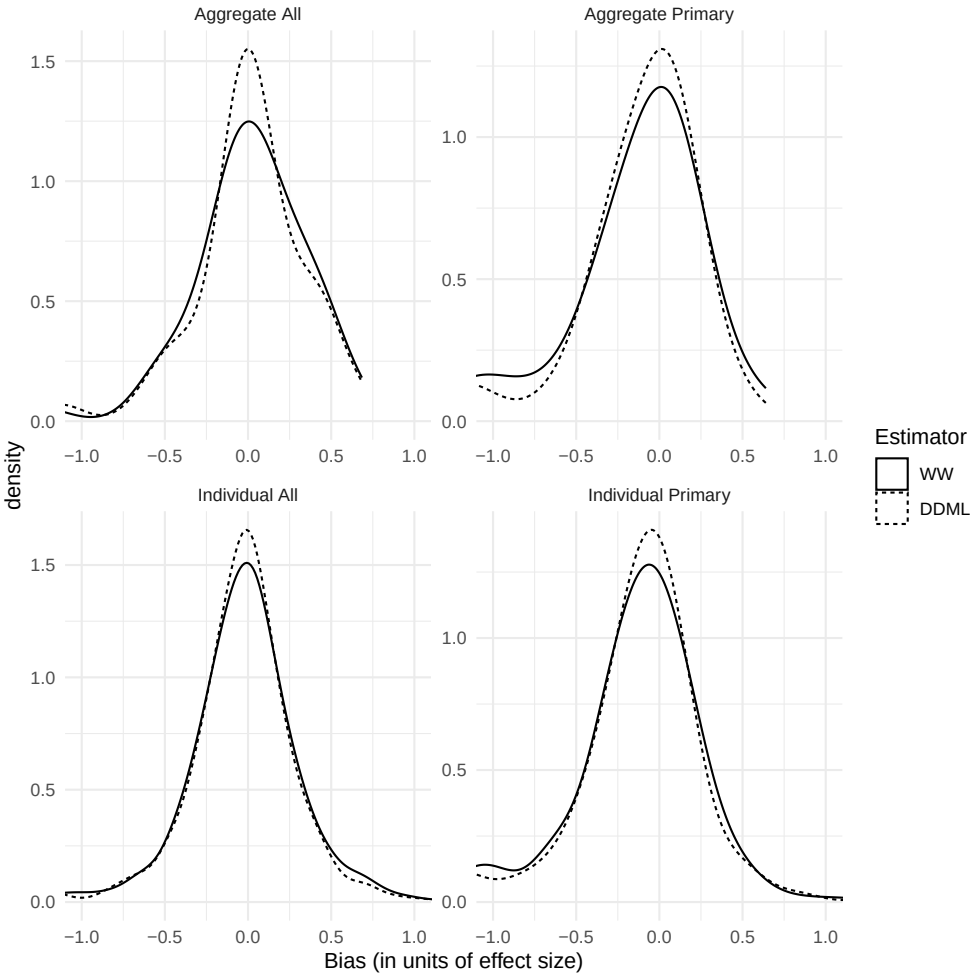


Figure 3: Kernel Density Plots of Raw Bias

outcome variables (our preferred specification), while the second panel shows results aggregating all outcomes, the third panel show the results for individual primary outcomes, and the fourth panel for all outcomes individually. The first column shows a meta-analysis of the experimental treatment effects, while columns (2)-(4) show meta-analyses of bias for our three observational methods.

The results are striking. Regardless of the method used, or the approach we take with respect to the outcome variables, we find very small average bias. For example, for aggregated primary outcomes, the average bias using the DDML estimator is -0.047 standard deviations, which compares to an average treatment effect of 0.171 standard deviations across all studies in our data. In addition to the small size, average bias is uniformly insignificant, with the exception of one coefficient when we use PDSL and the individual primary outcomes. We conclude that there is no evidence that observational studies systematically over or underestimate program impacts.

We also see that the minimum effective standard error – defined as the effective standard error of a hypothetical infinite N observational study ($\sqrt{\hat{\sigma}_{\mu}^2 + \hat{\tau}^2}$) – is large, regardless of the method used. Looking across the table, the smallest effective standard error is 0.141, leading to a smallest minimum detectable effect size of 0.28 standard deviations for an observational study. Many development economists use a rule of thumb that suggests a 0.2 standard deviation impact is a large impact when considering power. This in turn implies that there are large and policy important impacts that simply cannot be detected with an observational approach, given our current knowledge about observational bias.

The table also shows that the choice of observational method matters. DDML outperforms both a naive with-without comparison and PDSL in almost all panels in terms of having a smaller effective standard error. Further, PDSL is occasionally worse than the naive with-without comparison. Noting this, we focus much of the ensuing discussion on results from DDML and the simple with-without.

Figure 4 provides another way to look at the results. Each point in the figures represents an observational estimate from our data set, with the x-axis recording the effect size in standard deviations, and the y-axis recording the standard error. The figures also show two potential confidence intervals. The straight lines show a standard confidence interval, with those observational estimates outside the funnel being deemed statistically significant at the 5% level. The

Table 1: Meta-analysis of Bias

	TE (1)	WW (2)	PDSL (3)	DDML (4)
Panel A: Aggregated primary outcomes				
Mean ($\hat{\mu}$)	0.171	−0.046	−0.053	−0.047
SE ($\hat{\sigma}_{\mu}$)	(0.042)	(0.042)	(0.037)	(0.035)
Standard deviation ($\hat{\tau}$)		0.201	0.165	0.161
Effective.SE		0.206	0.169	0.165
Num.obs.	42	42	42	42
Panel B: Aggregated all outcomes				
Mean ($\hat{\mu}$)	0.061	0.057	0.036	0.036
SE ($\hat{\sigma}_{\mu}$)	(0.033)	(0.040)	(0.046)	(0.033)
Standard deviation ($\hat{\tau}$)		0.189	0.228	0.137
Effective.SE		0.193	0.232	0.141
Num.obs.	43	43	43	43
Panel C: Individual primary outcomes				
Mean ($\hat{\mu}$)	0.126	−0.052	−0.074	−0.052
SE ($\hat{\sigma}_{\mu}$)	(0.031)	(0.036)	(0.031)	(0.031)
Standard deviation ($\hat{\tau}$)		0.231	0.199	0.199
Effective.SE		0.234	0.202	0.202
Num.obs.	264	264	264	264
Panel D: Individual outcomes				
Mean ($\hat{\mu}$)	0.043	0.039	0.004	0.036
SE ($\hat{\sigma}_{\mu}$)	(0.018)	(0.055)	(0.041)	(0.039)
Standard deviation ($\hat{\tau}$)		0.394	0.359	0.286
Effective.SE		0.398	0.362	0.289
Num.obs.	1797	1797	1795	1797

Notes: Column 1 presents the results of the meta-analysis on experimental treatment effects, column 2 is the bias of the simple with-without estimator (selection bias), column 3 is the bias of the post double selection lasso estimator, and column 4 is the bias of the DDML estimator. Effective SE $= \left(\sqrt{\hat{\sigma}_{\mu}^2 + \hat{\tau}^2}\right)$. Panel A includes one aggregated outcome generated from the primary outcomes for each study, panel B includes one aggregated outcome generated from the all outcomes for each study, panel C shows the results from using all primary outcomes in each study, panel D shows the results from using all individual outcomes in each study.

dashed lines show our adjusted confidence intervals, which take into account uncertainty about observational bias. The two figures to the left display the results for all outcomes. The figures to the right focus only on the aggregated primary outcomes. The figures show the key points that we have made before: the adjusted (dashed) confidence intervals are much wider than the standard intervals, and even with an infinite observational sample, which gives a zero standard error, it is never possible to reject a positive treatment effect of less than about 0.3 standard deviations, a remarkably large treatment effect.

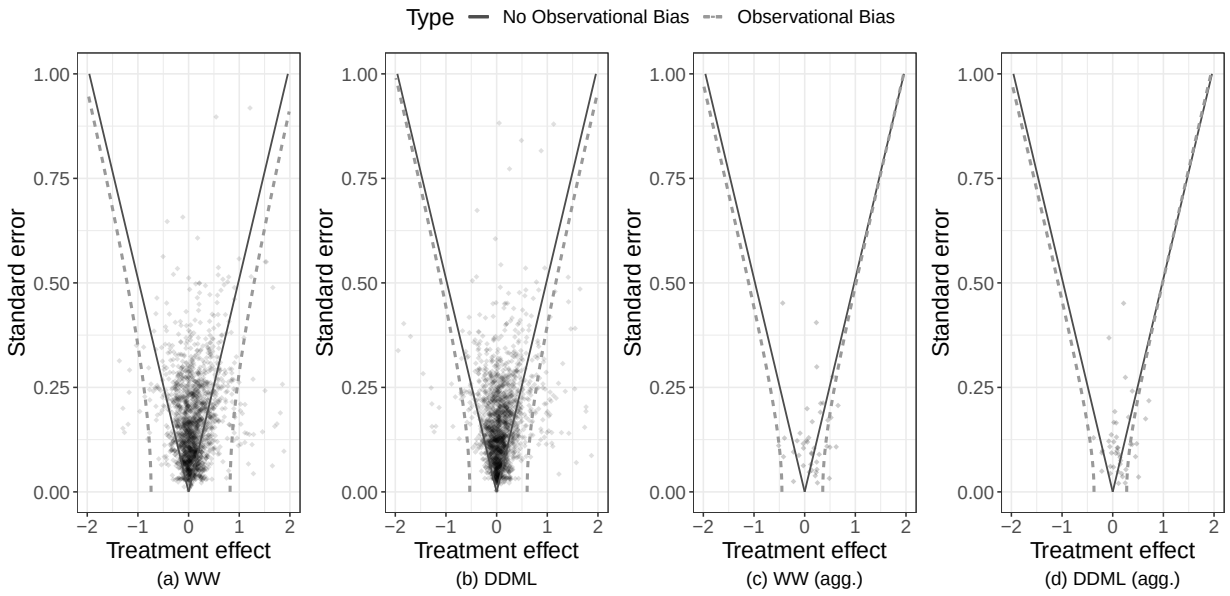


Figure 4: Funnel Plot of Treatment Effect Estimates

Notes: The solid lines represent the uncorrected confidence regions and the dashed lines represent the corrected confidence regions. The two figures on the left plot the treatment effects associated with all outcomes: for the with-without in panel (a), 512 treatment effects are statistically significant whereas only 48 remain statistically significant after correction. For the DDML in panel (b), 413 uncorrected treatment effects are statistically significant whereas only 76 remain statistically significant after applying the correction. The two figures to the right plot the treatment effects associated with the aggregated primary outcomes. For the with-without in panel (c), 20 uncorrected treatment effects are statistically significant and only 7 remain so. For the DDML in panel (d), 19 uncorrected treatment effects are statistically significant and only 7 remain so.

We can also use the same figure to compare across different observational methods. The left figure of each panel shows that the naive with-without estimator has a much larger confidence region than the DDML method shown on the right of each panel. One interpretation of this is in terms of effective power for an infinitely sized observational study. If using with-without and our preferred specification, this hypothetical study would have a minimal detectable effect size of 0.40 standard deviations, while if it made use of DDML it would have an MDE of 0.32 standard deviations.

Similar calculations can be used to illuminate the trade-off between observational approaches and an RCT. Suppose that a policy maker has access to a infinitely sized observational study, with effective standard error equal to 0.165. We can then ask what sample size she would need in an eligibility-design RCT to obtain a smaller expected standard error? Figure 5 plots a few scenarios, assuming individual randomization with 50% assigned to treatment.[26] With 100% compliance, an experimental sample size of just 148 is sufficient to achieve the same expected standard error as an infinite-N observational study. The required sample sizes increase if there is imperfect compliance in the RCT. For example, with 25% compliance the RCT would need 2364 observations to dominate, which is still a relatively modest trial when compared to some of the more recent studies run by development economists.

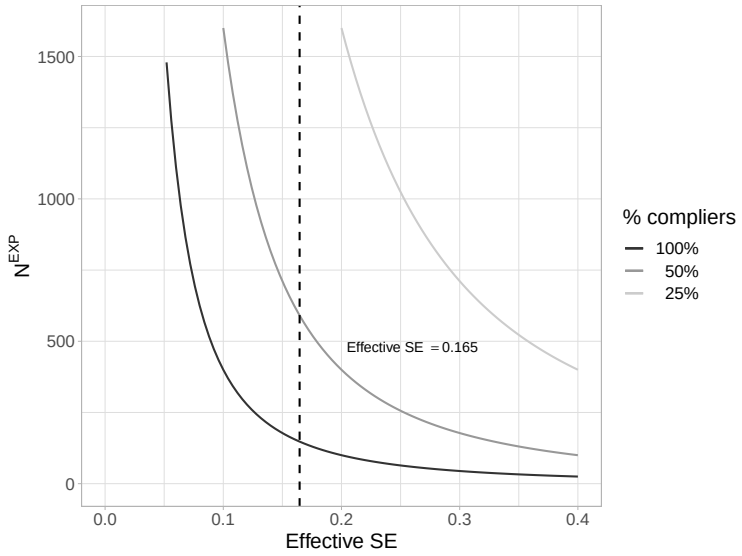


Figure 5: Required Experimental Sample Size to Match Effective Standard Error of an Infinite-N Observational Study

Figure 4 also shows that a large proportion of our observational estimates lose their significance when confidence intervals are adjusted for observational bias and the notes summarizes the information by looking at the significance of corrected and uncorrected estimates. When using our preferred observational method, DDML, around 19% of all observational estimates would be declared incorrectly statistically significant using uncorrected confidence intervals.

Figure 6 gives more detail for the aggregated primary outcome from each study. Each circle shows

[26] For sample size N, fraction P treated, and compliance rate C, we calculate the expected standard error on the experimental TOT estimate as $\frac{1}{C}\sqrt{\frac{1}{P(1-P)N}}$SD (Duflo et al., 2007).

the experimental treatment effect estimate and its 95% confidence interval (in terms of standard deviation effect sizes). The triangle and line shows the uncorrected observational estimate and confidence interval of the DDML estimator, while the square shows the observational estimate and confidence interval after we apply our correction. In many cases (e.g. the second line) we can see that the experimental and uncorrected DDML confidence intervals do not overlap, whereas the corrected DDML confidence intervals do overlap with the experimental estimate. Overall, uncorrected confidence intervals for observational estimators appear to be too tight, and our correction allows a researcher to be honest about the uncertainty generated by observational bias.

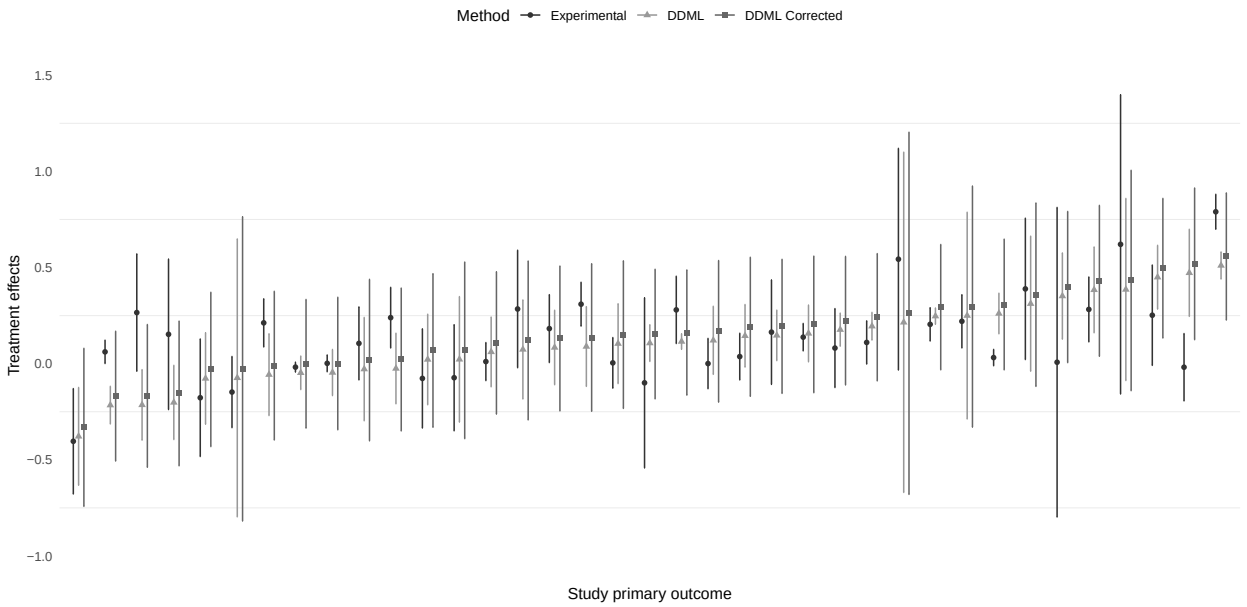


Figure 6: Corrected and Uncorrected Observational Confidence Intervals Compared to RCT Estimates

Figure 7 provides a summary of how our correction affects inference, and shows that our correction performs significantly better than the original intervals based on all individual outcomes. Uncorrected 95% confidence intervals from DDML only contain the experimental treatment effect 70% of the time, in contrast our corrected intervals manage this 94% of the time, close to the ideal 95%.

Figure 7 also provides information about the power of observational methods in general. Type II errors (false negatives where the observational estimator fails to reject a zero effect when the experimental estimator rejects zero effect) increase when our correction is applied, from 59% to 86%. Although this seems like our correction performs worse, this really shows the limited power

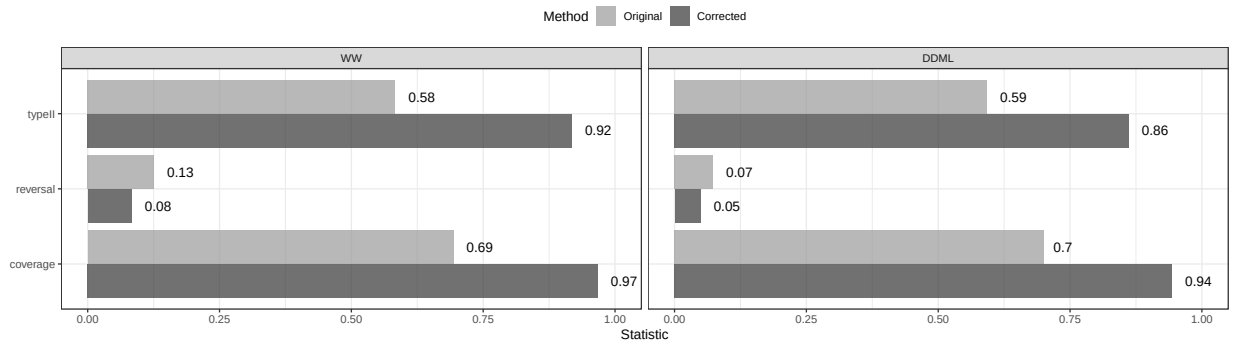


Figure 7: Errors and Coverage Ratios for Corrected and Uncorrected Observational Estimates

of observational methods when we are honest about the uncertainty surrounding observational bias. The original DDML estimates claim to have a power of 0.41 ($1 - 0.59$), but once observational bias is accounted for power drops to 0.14 ($1 - 0.86$). We also find strong gains in terms of power when conditioning on covariates. The power of the corrected confidence intervals for the with-without estimator is only 8% (1-0.92) whereas we gain 6 percentage points of power using the corrected confidence intervals for the DDML estimator.

Finally, the reversal row of figure 7 shows that without correcting the confidence intervals 7% of the outcomes for which the experimental estimator indicates a significant treatment effect in one direction, DDML declares statistical significance *in the opposite direction*. This drops to 5% when using our corrected confidence intervals.

4 Robustness to Relaxing our Assumptions

4.1 Exchangeability and Precision of the Bias Correction

Our headline results rely strongly on the assumption of exchangeability, which essentially says that our policy maker believes that all the studies in our data, and her own, draw their biases from the same distribution. Under this assumption our evidence implies that observational studies have very low effective power. A reasonable response might be to argue that domain experts and policy makers do in fact have sufficient information to exclude studies from our data that are not exchangeable with their programs of interest, and that this may reduce the effective standard error. For example, given an observational estimate of the impact of a microfinance program, a policy maker might be happy to focus on the subset of studies evaluating finance interventions.

In this section we argue that, given the current number of ICRCTs available, there are no power gains to be had from taking this approach.

The argument for using all of the data, rather than a subset, is similar to the argument for the use of average treatment effects in general, and the analogy can be used to highlight the trade-offs. Take a setting in which we have an RCT that creates within-community variation in treatment assignment across a large set of communities (and in which we are sure SUTVA holds). Combining within-community estimates across communities gives us an ATE that applies to no specific community, but might be a good estimate for what would occur if the treatment were applied to a randomly-selected community from the study area. But a policy maker never actually wants to know what will happen to a randomly selected community, rather she wants to know what would happen if treatment were scaled up in one of the communities. The use of ATE is not motivated by the randomly-selected community argument, but rather two reasons for *not* using the data from just one community. First, if only data from one community is used sample size and power are reduced. Second, it is well known that shrinking estimates from each community back toward the mean of the across-community estimates reduces the expected mean squared error. For example, a single community with a very high estimated ATE is likely to be an overestimate, and the fact that it is higher than the average of the ATEs reveals information to this effect. Thus, it is sensible to shrink estimates even if it is done by using data from communities that are less relevant.

The same two basic trade-offs apply in our setting. A policy maker who does not wish to use our effective standard errors cannot condition on what she observes in our results to decide whether to restrict to a subset (i.e., she cannot throw out studies *because* they increase the effective standard error). She must make the decision without knowledge. Given this, the effective standard error that can be achieved is the expectation of the effective standard error across reasonable subsets. This expectation takes into account both the trade-offs: reduced sample size and shrinkage. Table 2 shows the empirical trade-off as it exists in our data set. Panel A shows results of our meta analysis restricted to the subset of finance studies, Panel B only has health studies, while Panel C is only education studies. These are the largest sectors in our data, and are the three subsets for which there are enough studies to consider doing a meta-analysis. Effective standard errors using DDML are 0.175, 0.35 and 0.05 respectively.

If there were systematic precision gains to be had from restricting the set of studies, we would predict that doing so would decrease the *expected* effective standard error, i.e. the average of these three should be smaller than our main estimate. But that is not what we observe: the average is 0.19 standard deviations, which is greater than the effective standard error from using all the data (0.165 sd). This is consistent with a view that while ex-post sample restrictions sometimes increase precision (entirely unsurprisingly), we see no evidence that ex-ante restrictions would improve expected precision. We couldn't have known ex-ante that restricting to education would improve precision, in expectation it would worsen precision.

In short, we see no power gains from relaxing exchangeability, unless the policy maker is willing to commit, and risk having the very large standard errors found in the health subset. We intend, in future work, to understand whether policy makers and experts are able to predict which subsets reduce variability, and so whether there could be gains in the presence of commitment.

The case of education is also interesting. Here we seem to have enough similar studies to gain power from restricting attention to this subset. Whether that represents a systematic pattern or is just sampling variation (in the sampling of studies) is unknown, but provides some hope that adding ICRCTs to our dataset, especially in sectors that are presently small, may enable future researchers or policy makers to more precisely bias-correct their observational estimates. We wish to emphasise that these gains are only available if the set of studies across domains is sufficiently large, or external evidence sufficiently strong, that the policy maker is willing to commit to a subset ex-ante.

4.2 The Quality of RCTs and the Availability of Covariates

This section considers robustness to two additional assumptions. First, we have assumed that the RCTs we use are not themselves biased, and so give good estimates of the true treatment effect. This may not be the case if, for example, there is a breach of SUTVA. Second, we are implicitly assuming that we have all the covariates that a policy maker would usually have available to make use of observational estimates. We argue that our results are robust to relaxing these assumptions.

Our approach is similar to that taken above. Appendix D contains meta-analytic estimates for a large number of subsets of our data. The subsets relevant to RCT quality are whether the paper reports an experimental estimate of LATE or just ITT (which we think of as a proxy for the

Table 2: Bias distributions of different subsets

	TE (1)	WW (2)	DDML (3)
Panel A: Finance			
Mean ($\hat{\mu}$)	0.212	−0.067	−0.102
SE ($\hat{\sigma}_{\mu}$)	(0.034)	(0.072)	(0.073)
Standard deviation ($\hat{\tau}$)		0.150	0.159
Effective.SE		0.166	0.175
Num.obs.	11	11	11
Panel B: Health			
Mean ($\hat{\mu}$)	0.301	−0.109	−0.078
SE ($\hat{\sigma}_{\mu}$)	(0.194)	(0.180)	(0.131)
Standard deviation ($\hat{\tau}$)		0.475	0.326
Effective.SE		0.508	0.351
Num.obs.	8	8	8
Panel C: Education			
Mean ($\hat{\mu}$)	0.105	−0.015	−0.023
SE ($\hat{\sigma}_{\mu}$)	(0.052)	(0.040)	(0.042)
Standard deviation ($\hat{\tau}$)		0.000	0.032
Effective.SE		0.040	0.053
Num.obs.	8	8	8
Panel D: All			
Mean ($\hat{\mu}$)	0.171	−0.046	−0.047
SE ($\hat{\sigma}_{\mu}$)	(0.042)	(0.042)	(0.035)
Standard deviation ($\hat{\tau}$)		0.201	0.161
Effective.SE		0.206	0.165
Num.obs.	42	42	42

Notes: Column 1 presents the results of the meta-analysis on experimental treatment effects, column 2 is the bias of the simple with-without estimator (selection bias), and column 3 is the bias of the DDML estimator. Both studies show the meta-analyses of aggregated primary outcomes and panel A is for finance studies, while panel B is for health studies and panel C is for education studies. Effective SE $= \sqrt{\hat{\sigma}_{\mu}^2 + \hat{\tau}^2}$.

researchers' belief in the plausibility of the exclusion restriction) and whether the experimental estimate is produced using across-cluster variation (a possible indicator of the plausibility of the SUTVA assumption). Subsets relevant to covariate availability are the number of covariates (we would expect more covariates to improve the precision of the bias correction), and whether a pre-treatment ("baseline") measure of the outcome variable is available or not (controlling for the outcome variable at baseline is a common way to try to alleviate selection bias). We also subset according to whether the study has an eligibility or encouragement design, since as discussed in Section 1 these imply different estimands and we might doubt whether exchangeability holds across them. We see no qualitative changes in the effective standard errors, which remain large throughout.

5 The Value of Additional ICRCTs

We are used to thinking about the power of a study as driven mostly by the sample size in that study. Figure 1 shows that N is not always the dominant determinant of power when using our corrected confidence intervals – uncertainty about bias potentially matters more. This opens the possibility that the best way to increase power in observational studies may be to increase the number of ICRCTs that are run. There are two senses in which this is true. First, continuing to assume exchangeability across all studies, an additional study is expected to leave $\hat{\tau}^2$ and $\hat{\mu}^2$ unchanged, but to decrease $\hat{\sigma}^2_\mu$, increasing power in observational studies. Second, increasing the number of ICRCTs within a particular domain, for example education, can allow the policy maker to more readily commit to focus on a subset of studies that she believes are more likely to be exchangeable with her own setting, without having to face the sample size and shrinkage trade-offs discussed above. To return to our analogy with average treatment effects, if the set of observations from a particular community becomes large enough, then it makes sense to look only at results from that community when deciding whether to increase treatment rollout.

Figure 8 illustrates the empirical value of more studies in our data set. It plots confidence interval lengths as a function of the number of included studies, S. Each dot represents the average length of a corrected confidence interval taken across all possible combinations of $p = 2, ..., S$ studies for the DDML estimator. The Figure uses our aggregated primary outcomes and assumes that $\sigma^2_\epsilon = 0$.[27] The figure shows confidence interval length for different subsets, as well as the average

[27] Including σ^2_ϵ moves results up by roughly a constant.

of subsets for the reasons discussed above.

Concentrating first on the curve showing all studies, we see a sharp gain from increasing from 2 to 5 ICRCTs which stabilizes at around 12 studies. We do not display more than 20 included studies because the line asymptotes. It seems striking that the convergence materializes much earlier than at $S = 42$. The marginal gain of an additional ICRCT seems indeed to converge to zero with as little as 12 included studies. This tends to suggest that we already have a relatively large data set for our purposes, so long as we are committed to assuming exchangeability across all studies.

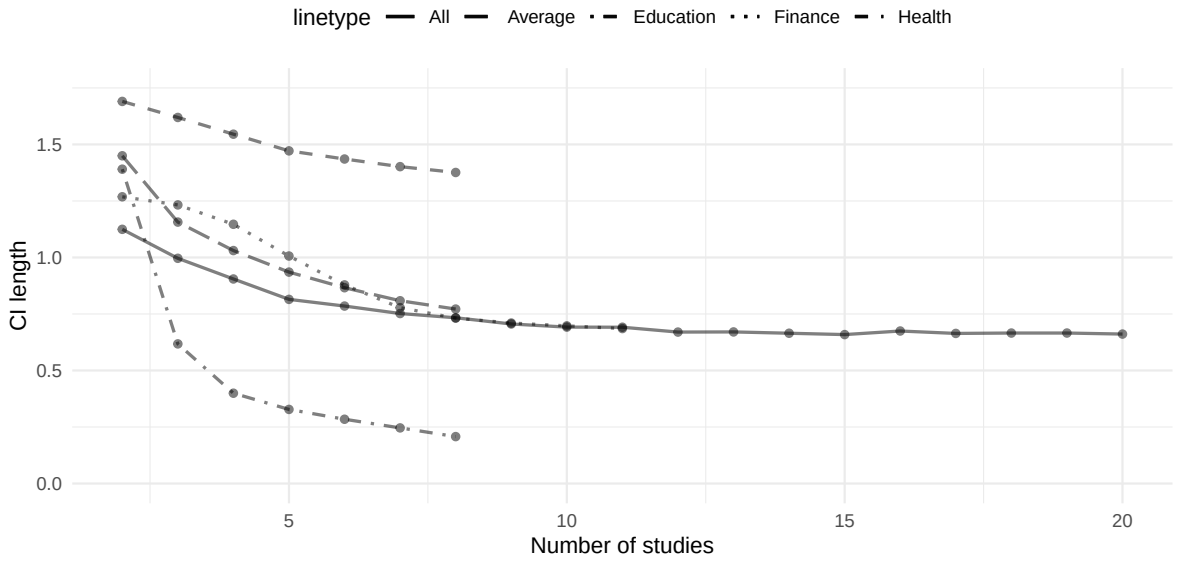


Figure 8: Theoretical and Empirical Confidence Interval Length for the DDML Estimator

Notes: The dotted lines represent the empirical results for the aggregated primary outcomes. Each dot represents the average corrected confidence interval lengths from including $p = 1, ..., S$ studies in the meta-analysis using the effective standard error only $\sqrt{\hat{\tau}^2 + \hat{\sigma}_{\hat{\mu}}^2}$. For the Finance, Health and Education subsets, the average is computed by estimating a meta-analysis for each possible combination of p included studies in our sample and averaging over the resulting confidence interval lengths. The "Average" represents the average confidence interval lengths over these three subsets. Without sub-setting ("All"), a random subset of 1000 combinations is chosen to compute the average length if the number of combinations exceeds 1000.

Next, consider the three smaller subsets that we considered above: finance, education and health, as well as their average. Four points are worth noting. First, all three curves show the strong reductions in confidence interval length to be had within subset from increasing the number of ICRCTs, suggesting gains for collecting more data in sectors where we have less data. Second, the average taken across the three subsets starts far from the full data set, but converges quickly. This suggests that gathering more data from ICRCTs is likely to lead to a reduction in the expected cost of concentrating on subsets of the data. Third, the very sharp decline in confidence interval

length for education studies, and the low effective standard error in general shows the potential gains from subsetting if a policy maker is willing to commit, and that these gains become larger with more data. Finally, the large difference between the standard errors for health and education studies highlights the risks of committing to a subset of the data. It may well be that this risk can be removed, but we would need more data to allow concentrating on a further subset of education studies, or to feel confident that the high effective standard error for those studies reflects true variability in bias, rather than sampling error.

Overall, we are very bullish about the value of continuing to run ICRCTs. Perhaps in the long run we will have enough data to be able to forego running RCTs, and simply use adjusted confidence intervals for observational studies that draw on highly specific estimates for a given setting.

6 Conclusion

Observational studies are likely to remain a mainstay of program evaluation for some time. We study the bias in these studies, with an emphasis on quantifying uncertainty, which is often treated as having unknown size and magnitude. Our main results suggest that observational studies have very little power to detect program effects that are of a policy relevant size. We find that some observational approaches, notably DDML, can improve power, but only by a small amount. This may be seen as quite a negative outcome, but we see it as suggesting strong value in collecting more data from ICRCTs to help reduce uncertainty and improve the power of observational studies. More practically, our proposed correction to standard errors and confidence interval enables to adequately reflect the uncertainty around observational estimates. Our correction enables the inclusion of observational estimates in meta-analysis, with weights reflecting their actual precision.

References

Agodini, R. and M. Dynarski (2004): "Are Experiments the Only Option? A Look at Dropout Prevention Programs," *The Review of Economics and Statistics*, 86, 180–194.

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.

Anderson, M. L. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481–1495.

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.

——— (2010): "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *Journal of economic perspectives*, 24, 3–30.

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].

Baldwin, K., D. Karlan, C. Udry, and E. Appiah (2016): "Does community-based development empower citizens? Evidence from a randomized evaluation in Ghana," J-PAL (Working Paper), available at URL: https://www.povertyactionlab.org/evaluation/does-community-based-development-empower-citizens-evidence-randomized-evaluation-ghana (01/08/2024).

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.

Banerji, R., J. Berry, and M. Shotland (2017): "The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India," *American Economic Journal: Applied Economics*, 9, 303–37.

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.

Behaghel, L., C. De Chaisemartin, and M. Gurgand (2017): "Ready for boarding? The effects of a boarding school for disadvantaged students," *American Economic Journal: Applied Economics*, 9, 140–164.

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., N. Fiala, and S. Martinez (2014a): "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., A. C. Hartman, and R. A. Blair (2014b): "How to Promote Order and Property Rights under Weak Rule of Law? An Experiment in Changing Dispute Resolution Behavior through Community Education," *American Political Science Review*, 108, 100–120.

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.

Bloom, H. S. (1984): "Accounting for no-shows in experimental evaluation designs," *Evaluation review*, 8, 225–246.

Bloom, N., J. Liang, J. Roberts, and Z. J. Ying (2015): "Does working from home work? Evidence from a Chinese experiment," *The Quarterly journal of economics*, 130, 165–218.

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.

Chabé-Ferret, S. (2023): *Statistical Tools for Causal Inference*.

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

Duflo, E., R. Glennerster, and M. Kremer (2007): *Chapter 61 Using Randomization in Development Economics Research: A Toolkit*, Elsevier, 3895–3962.

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.

Eckles, D. and E. Bakshy (2021): "Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects," *Journal of the American Statistical Association*, 116, 507–517, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1796393.

Ferraro, P. J. and J. J. Miranda (2014): "The performance of non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark," *Journal of Economic Behavior & Organization*, 107, 344 – 365.

Fink, G., R. Levenson, S. Tembo, and P. C. Rockers (2017): "Home-and community-based growth monitoring to reduce early life growth faltering: an open-label, cluster-randomized controlled trial," *The American journal of clinical nutrition*, 106, 1070–1077.

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.

Fraker, T. and R. Maynard (1987): "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *The Journal of Human Resources*, 22, 194–227.

Friedlander, D. and P. K. Robins (1995): "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *The American Economic Review*, 85, 923–937.

Gechter, M. and R. Meager (2022): "Combining Experimental and Observational Studies in Meta-Analysis: A Debiasing Approach," Working Paper, available at URL: https://michaelgechter.com/research/ (01/08/2024).

Gerber, A. S., D. Karlan, and D. Bergan (2009): "Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions," *American Economic Journal: Applied Economics*, 1, 35–52.

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., R. Moakler, and F. Zettelmeyer (2023): "Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement," *Marketing Science*, 42, 768–793.

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, publisher: INFORMS.

Griffen, A. S. and P. E. Todd (2017): "Assessing the Performance of Nonexperimental Estimators for Evaluating Head Start," *Journal of Labor Economics*, 35, S7–S63.

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.

Heckman, J. J. and V. J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: the Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–874.

Heckman, J. J., H. Ichimura, J. A. Smith, and P. E. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1099.

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.

Higgins, J. P. T., S. G. Thompson, and D. J. Spiegelhalter (2008): "A Re-Evaluation of Random-Effects Meta-Analysis," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 172, 137–159.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

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.

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

Raudenbush, S. W. (2009): "Analyzing Effect Sizes: Random-Effects Models," in *The Handbook of Research Synthesis and Meta-Analysis*, ed. by H. Cooper, L. V. Hedges, and J. C. Valentine, Russell Sage Foundation, 295–316.

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.

Viviano, D., K. Wuthrich, and P. Niehaus (2021): "(When) should you adjust inferences for multiple hypothesis testing?" Tech. rep., UC San Diego.

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 Estimating Observational Bias in Randomised Controlled Trials with Imperfect Compliance

In this appendix we show how to produce estimates of average observational bias for a well defined population for both of our study types: eligibility designs and encouragement designs.

First some notation. In randomized experiments with imperfect compliance, individuals $i = 1, ..., N$ receive a randomized offer $R_i \in \{0, 1\}$. They can then choose to take-up a program or not. The randomized offer divides the sample into two groups with $R_i = 1$ if the individual is randomized into the treatment group and $R_i = 0$ for the control. We denote program take-up $D_i \in \{0, 1\}$ where $D_i = 1$ if the individual chooses to participate and $D_i = 0$ otherwise. If D_i were equal to R_i we would have perfect compliance. We denote the potential participation given treatment group by D_i^r and we let Y_i^{dr} be the potential outcome given treatment and take-up.

Below we use subsets of the following classical assumptions:

Assumption 1 *Assumptions for Valid RCTs*[28]

1. *SUTVA:* $(Y_i^1, Y_i^0) \perp D_j$ *for* $i \neq j$.
2. *Independence:* $(Y_i^{dr}, D_i^r) \perp R_i, \forall (d, r) \in \{0, 1\}^2$.
3. *Exclusion restriction:* $Y_i^{dr} = Y_i^d, \forall (d, r) \in \{0, 1\}^2$.
4. *First Stage:* $E(D_i^1 - D_i^0) \in (0, 1]$.
5. *Monotonicity:* $D_i^1 - D_i^0 \geq 0$ *for all* i.

Assumption 2 *Additional Assumptions for Observational Estimators*

1. *Conditional Independence:* $(Y_i^1, Y_i^0) \perp D_i | X_i, R_i = r, \forall r \in \{0, 1\}$.
2. *Common Support:* $0 < P(D_i = 1 | X_i, R_i = r) < 1, \forall r \in \{0, 1\}$.

Given the exclusion restriction, observed take-up is a function of treatment assignment $D_i = D_i^1 R_i + D_i^0 (1 - R_i)$, and the observed outcome is a function of the actual program participation

[28] In addition, because we restrict to ICRCTs, it must be that $E(D_i^1 - D_i^0) < 1$, but this is not an identification condition so we leave it out of the below statements.

$Y_i = Y_i^1 D_i + Y_i^0(1 - D_i)$.

A.1 Encouragement Designs

We show how to generate observational and experimental estimates of average treatment effects for the same sub-population (the compliers).

In an encouragement design everyone in treatment and control can choose to participate, but the treatment receives an encouragement to do so. To use this design, we require imperfect compliance in both treatment arms: $P(D_i = 1|R = r) > 0$, $r \in \{0, 1\}$. As is well known, there are four potential groups of subjects: (i) always takers (AT) are individuals who always choose to participate regardless of randomization status ($D_i^1 = D_i^0 = 1$); (ii) never takers (NT) are individuals who never participate regardless of randomization status ($D_i^1 = D_i^0 = 0$); (iii) compliers (C) comply with the manipulation - they participate if they are randomized 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 suggests ($D_i^1 - D_i^0 = -1$). We use the notation T_i to refer to these groups, where, for example $T_i = C$ refers to the complier group.

It is well known that under the classical assumptions SUTVA, Independence, Exclusion, First Stage and Monotononicity, the experimental Wald estimand

$$TOC^{EXP} = \frac{E\left[Y_i|R_i = 1\right] - E\left[Y_i|R_i = 0\right]}{P(D_i = 1|R_i = 1) - P(D_i = 1|R_i = 0)} \tag{5}$$

recovers a local average treatment effect $LATE = E[Y_i^1 - Y_i^0|D_i^1 - D_i^0 = 1]$. We refer to this as the treatment on compliers, or TOC in the text to differentiate it from a different late, the treatment on the treated. The notation TOC^{EXP} refers to an experimental estimand and we will denote non-experimental, or observational, estimands that conditions on X by TOT_X^{OBS}. We denote by TOC^{OBS} the naive observational estimand that does not condition on any covariate.

In order to form an observational estimand, note that we can build two separate observational estimands, one in the treated group ($TOT_X^{OBS,1}$) and one in the control group ($TOT_X^{OBS,0}$). 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 the additional assumptions of conditional independence and common support. As is well known, under these

assumptions, it is possible to recover an estimate of the treatment on the treated in each treatment arm $TOT_X^{OBS,r} = E[E[Y_i|X_i, D_i = 1, R_i = r] - E[Y_i|X_i, D_i = 0, R_i = r]|D_i = 1, R_i = r] = TOT^r = E[Y_i^1 - Y_i^0|D = 1, R = r]$. We propose to combine these estimates in a Wald-type estimand

$$TOC_X^{OBS} = \frac{TOT_X^{OBS,1} \Pr(D_i = 1|R_i = 1) - TOT_X^{OBS,0} \Pr(D_i = 1|R_i = 0)}{\Pr(D_i = 1|R_i = 1) - \Pr(D_i = 1|R_i = 0)}. \tag{6}$$

Proposition 1 (Observational Estimand of LATE) *Under Assumptions 1 and 2:*

$$TOC_X^{OBS} = E[Y_i^1 - Y_i^0|T_i = C] = LATE$$

Proof: First note that the observational estimand on the treatment arm is the sum of the treatment effects for the always-takers and the compliers weighted by their respective proportions:

$$\begin{aligned} TOT_X^{OBS,1} &= \mathbb{E}[Y_i^1 - Y_i^0|D_i = 1, R_i = 1] \\ &= \mathbb{E}[Y_i^1 - Y_i^0|T_i = AT] \Pr(T_i = AT|D_i = 1, R_i = 1) \\ &\quad + \mathbb{E}[Y_i^1 - Y_i^0|T_i = C] \Pr(T_i = C|D_i = 1, R_i = 1), \end{aligned}$$

where the second equality comes from Independence and Monotonicity. Now let us consider the proportions of each type conditional on treatment arm and participation status:

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

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:

$$\begin{aligned}\Pr(T_i = C|D_i = 1, R_i = 1) &= \frac{\Pr(T_i = C \wedge D_i = 1|R_i = 1)}{\Pr(D_i = 1|R_i = 1)} \\ &= \frac{\Pr(T_i = C|R_i = 1)}{\Pr(D_i = 1|R_i = 1)},\end{aligned}$$

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:

$$\begin{aligned}TOT_X^{OBS,0} &= E[Y_i^1 - Y_i^0|D_i = 1, R_i = 0] \\ &= E[Y_i^1 - Y_i^0|T_i = AT].\end{aligned}$$

Combining the formulas for $TOT_X^{OBS,1}$ and $TOT_X^{OBS,0}$, the numerator of the TOC_X^{OBS} estimand in equation 5 is:

$$\begin{aligned}&TOT_X^{OBS,1} \Pr(D_i = 1|R_i = 1) - TOT_X^{OBS,0} \Pr(D_i = 1|R_i = 0) \\ &= E[Y_i^1 - Y_i^0|T_i = C] \Pr(T_i = C|R_i = 1) \\ &\quad + E[Y_i^1 - Y_i^0|T_i = AT] \Pr(D_i = 1|R_i = 0) \\ &\quad - E[Y_i^1 - Y_i^0|T_i = AT] \Pr(D_i = 1|R_i = 0) \\ &= E[Y_i^1 - Y_i^0|T_i = C] \Pr(T_i = C|R_i = 1).\end{aligned}$$

Finally, Monotonicity and Independence imply that:

$$\Pr(T_i = C|R_i = 1) = \Pr(D_i = 1|R_i = 1) - \Pr(D_i = 1|R_i = 0),$$

which proves the result. ■

Proposition 1 implies that we can generate observational and experimental estimands which, under Assumptions 1 and 2 should be equal to each other. We use as estimands of observational bias on compliers the difference between the observational and experimental estimands of the

treatment effect on compliers:

$$TOC^{OBS} - TOC^{EXP} = SBC$$
$$TOC_X^{OBS} - TOC^{EXP} = BC_X.$$

Where SBC stands for selection bias on compliers and BC_X stands for observational bias on compliers after covariate adjustment. In section 1 we refer to these term simply as B.

A.2 Eligibility Designs

Eligibility designs are much more straightforward to analyse. In an eligibility design, the control are prevented from participating.[29] We can form an experimental estimand TOT^{EXP} based on Equation 5 with $P(D_i = 1|R_i = 0) = 0$ and a single observational estimand on the treatment arm $TOT^{OBS} = TOT^{OBS,1}$ according to Equation 6. It is well known that $TOC^{EXP} = TOT^{EXP} = TOT$, the Treatment on the Treated ($TOT = E[Y_i^1 - Y_i^0|D_i = 1]$) under Assumption 1 and that $TOT_X^{OBS,1} = TOT$ under SUTVA, Assumption 2 and the fact that $D_i = 1$ implies $R_i = 1$ in this setup. We use as estimands of observational bias on the treated the difference between the observational and experimental estimands of TOT:

$$TOT^{OBS,1} - TOT^{EXP} = SBT$$
$$TOT_X^{OBS,1} - TOT^{EXP} = BT_X.$$

Where SBT stands for selection bias on the treated and BT_X stands for observational bias on the treated after covariate adjustment. Again, in section 1 we refer to these term simply as B.

[29] 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.

B Appendix - Estimators

We first present our observational estimators before explaining how we estimate observational bias and its precision. For simplicity, since estimation for the encouragement and eligibility design on each treatment arm follows the same procedure, we denote the experimental estimates that identify depending on the design either a TOT^{EXP} or TOC^{EXP} as EXP or $\widehat{EXP}$ for the resulting estimator. For the observational estimate on each treatment arm, we denote the estimands and resulting estimators on each treatment arm as OBS^r and $\widehat{OBS^r}$ respectively (with a subscript X if we condition on covariates). The resulting observational estimator is denoted as $\widehat{OBS}$ estimating either a treatment effect on the compliers or on the treated depending on the design. We name all estimates of observational bias $\hat{B}$ regardless of the design and underlying estimator.

B.1 Observational estimators

We apply three different observational estimators, the first two of which are based on machine-learning algorithms:

- *Post double selection lasso PDSL* (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 covariates selected in both regressions.

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

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

  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 $\widehat{OBS^r}_{X,1}$.

5. Repeat the last three steps using different splits of the k subsamples to obtain k estimates of $\widehat{OBS^r}_{X,k}$.

6. Average the different estimators: get the DML estimator of $\widehat{OBS^r}_X = \frac{1}{K}\sum_1^K \widehat{OBS^r}_{X,k}$.

Compared to Belloni et al. (2014), Chernozhukov et al. (2018) the method relies on weaker assumptions through sample-splitting. Intuitively, the effect of the covariates on take-up are partialled out. The nuisance function is estimated via random forest learner with 100 trees. We use the DML2 algorithm.

- *With-without comparison WW*. 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 $\widehat{OBS^r}$.

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

B.2 Estimates of the bias of observational estimators and their standard errors

With eligibility designs, we obtain, for each study s and outcome o, one observational estimate $\widehat{OBS}_{os} = \widehat{OBS^1}_{os}$ for each of the three observational methods (DDML, PDSL and WW) along with their respective standard errors $\hat{\sigma}_{OBS,os}$.[30] We also obtain an experimental estimate $\widehat{EXP}_{os}$ and its respective standard error $\hat{\sigma}_{EXP,os}$ using an IV regression of Y on D using R as an instrument, with strata fixed effects. For standard errors on both the observational and experimental estimates, we assume the same covariance structure as the authors of the original papers, i.e. if they cluster their

[30] Note that in the main text, we have denoted the standard error of the observational estimate as $\hat{\sigma}_{\epsilon,os}$. We change the notation in this section to improve readability.

standard errors, we cluster at the same level, otherwise we use heteroskedasticity robust standard errors.

With encouragement designs, we obtain two observational estimates $\widehat{OBS^1}_{os}$ and $\widehat{OBS^0}_{os}$ for each of the three observational methods (DDML, PDSL and WW) along with their respective standard errors $\hat{\sigma}_{OBS^1,os}$ and $\hat{\sigma}_{OBS^0,os}$, one for each treatment arm. We combine the estimates obtained on each treatment arm using Equation (6), replacing the population values by the sample values to obtain $\widehat{OBS}_{os}$. We estimate the standard error of the resulting estimate $\hat{\sigma}_{OBS,os}$ by using the delta method and the fact that, because of randomization, $\widehat{OBS^1}_{os} \perp \widehat{OBS^0}_{os}$, for a given outcome and study pair.

Finally, we combine our observational and experimental estimates to build an estimate of observational bias $\hat{B}_{os} = \widehat{OBS}_{os} - \widehat{EXP}_{os}$. We estimate the standard error of the resulting parameter as $\hat{\sigma}_{B,os} = \sqrt{\hat{\sigma}^2_{OBS,os} + \hat{\sigma}^2_{EXP,os}}$. This assumes independence of the observational and experimental estimator. We argue in Appendix E that assuming independence gives a lower bound on $\hat{\tau}^2$.

We also provide nonparametric bootstrap with replacement standard errors for the WW and DDML bias estimators and they are very close to our standard errors. We also considered estimating the standard errors as $\hat{\sigma}_{B,os} = \sqrt{\hat{\sigma}^2_{OBS,os} + \hat{\sigma}^2_{EXP,os} - 2\hat{\sigma}_{OBS,EXP}}$, where $\hat{\sigma}_{OBS,EXP}$ is the estimated covariance between observational and experimental estimators across outcome$\times$study pairs. Instead of another robustness table, we provide the $\hat{\tau}^2$ that we would obtain using that approach which is indeed much higher than when assuming independence.

C Appendix - Selecting and screening studies and cleaning data

In this section we describe our selection criteria, search process and data collection for the datasets we use to estimate the bias. We also describe how we clean data.

C.1 Selection and Screening

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: 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 transform them into a binary variable.

- Variable(s) measuring take-up of a program of interest. Usually these will be binary, if not we 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.

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 amongst the 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.[31]

We scraped the meta-data from all 207 of the studies on both 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 C.2).

[31] Other repositories we considered included: International Initiative for Impact Evaluation Development Evidence Portal, DIME data collection (The World Bank), Impact Evaluation Surveys Collection (The World Bank), David McKenzie's website, MDRC, Mathematica, REES (within ICPSR), openICPSR, NCES / IES, Head Start Impact Study, journal websites. These repositories were less well structured and typically less representative of the development economics literature than the J-PAL and IPA repositories. We plan to use them in future work.

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. During this stage, we notice that, for some projects, not all inclusion criteria hold. These projects are said to be excluded at Level 3.

Figure 9 shows how many studies pass each stage of screening.

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.

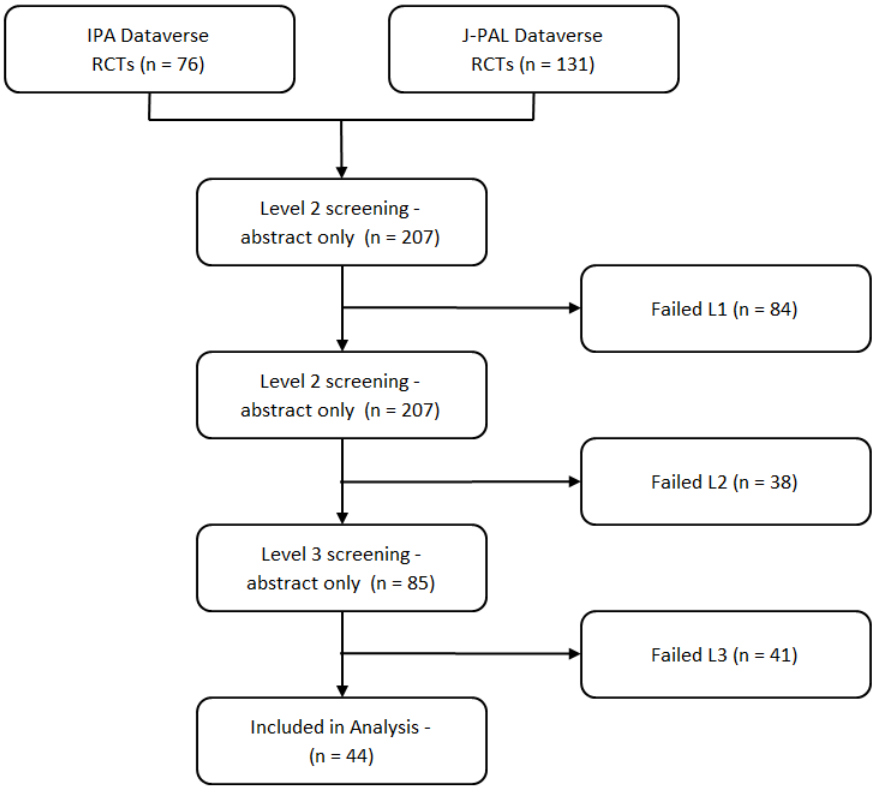


Figure 9: Flow diagram of studies passing through our selection process

C.2 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 measured before treatment and/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. We use the missingness indicators as potential controls as well.

D Appendix - Additional Results

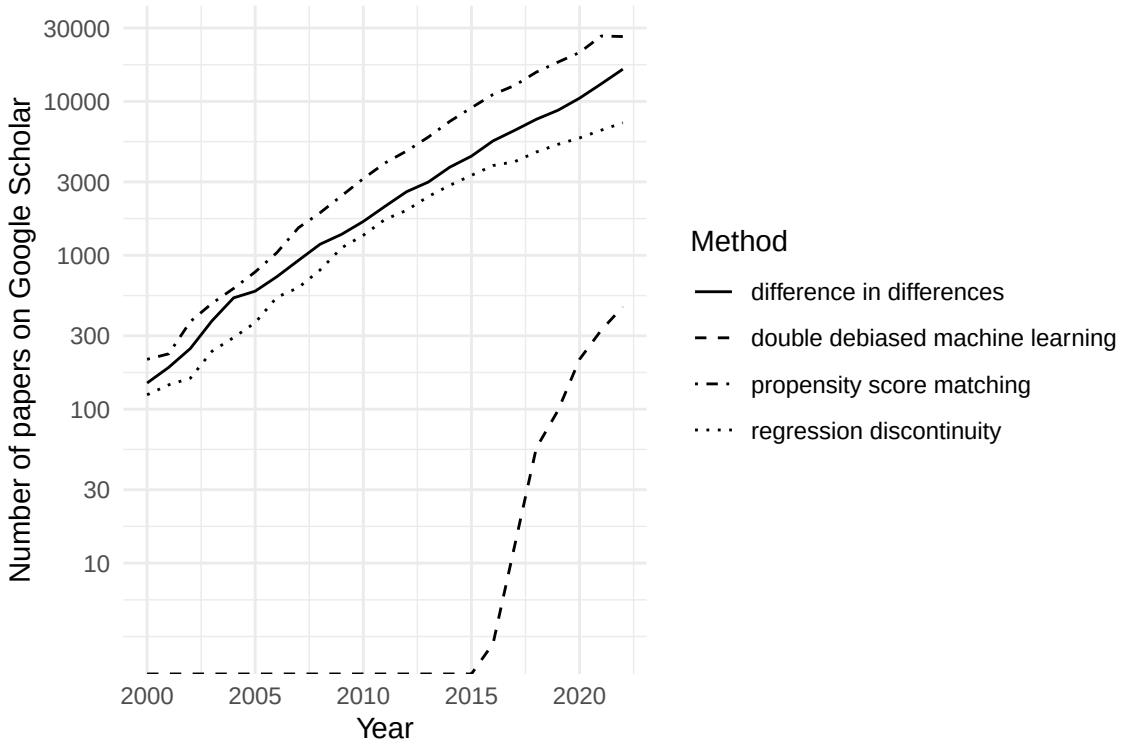


Figure 10: Number of papers mentioning method on Google Scholar

Table 3: Finance studies meta-analysis - alternate specifications

	TE (1)	WW (2)	DDML (3)
Panel A: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.179	0.037	0.004
SE ($\hat{\sigma}_{\mu}$)	(0.059)	(0.051)	(0.039)
Standard deviation ($\hat{\tau}$)		0.087	0.000
Effective.SE		0.101	0.039
Num.obs.	12	12	12
Panel B: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.201	-0.080	-0.123
SE ($\hat{\sigma}_{\mu}$)	(0.018)	(0.052)	(0.052)
Standard deviation ($\hat{\tau}$)		0.233	0.227
Effective.SE		0.239	0.233
Num.obs.	80	80	80
Panel C: Individual outcomes			
Mean ($\hat{\mu}$)	0.060	0.043	0.024
SE ($\hat{\sigma}_{\mu}$)	(0.047)	(0.038)	(0.035)
Standard deviation ($\hat{\tau}$)		0.201	0.180
Effective.SE		0.204	0.183
Num.obs.	764	764	764

Notes: Column 1 presents the results of the meta-analysis on experimental treatment effects, column 2 is the bias of the simple with-without estimator (selection bias), and column 3 is the bias of the DDML estimator. Effective SE $= \left(\sqrt{\hat{\sigma}_{\hat{\mu}}^2 + \hat{\tau}^2}\right)$. Panel A shows the results from including one aggregated outcome generated from all outcomes in each study, panel B shows the results from using all primary outcomes, and panel C uses all outcomes. Results based on aggregated primary outcomes, where one aggregated outcome generated from all primary outcomes in each study is included, can be found in the main text.

Table 4: Health studies meta-analysis - alternate specifications

	TE	WW	DDML
	(1)	(2)	(3)
Panel A: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.143	0.045	0.018
SE ($\hat{\sigma}_{\mu}$)	(0.178)	(0.147)	(0.118)
Standard deviation ($\hat{\tau}$)		0.372	0.282
Effective.SE		0.400	0.306
Num.obs.	8	8	8
Panel B: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.183	−0.030	−0.025
SE ($\hat{\sigma}_{\mu}$)	(0.194)	(0.182)	(0.158)
Standard deviation ($\hat{\tau}$)		0.523	0.458
Effective.SE		0.553	0.484
Num.obs.	48	48	48
Panel C: Individual outcomes			
Mean ($\hat{\mu}$)	0.177	0.004	−0.007
SE ($\hat{\sigma}_{\mu}$)	(0.185)	(0.158)	(0.144)
Standard deviation ($\hat{\tau}$)		0.500	0.451
Effective.SE		0.524	0.474
Num.obs.	150	150	150

Notes: See notes in previous table.

Table 5: Education studies meta-analysis - alternate specifications

	TE	WW	DDML
	(1)	(2)	(3)
Panel A: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.039	0.053	0.057
SE ($\hat{\sigma}_{\mu}$)	(0.043)	(0.094)	(0.066)
Standard deviation ($\hat{\tau}$)		0.201	0.116
Effective.SE		0.222	0.133
Num.obs.	8	8	8
Panel B: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.080	-0.028	-0.033
SE ($\hat{\sigma}_{\mu}$)	(0.028)	(0.055)	(0.065)
Standard deviation ($\hat{\tau}$)		0.126	0.141
Effective.SE		0.138	0.155
Num.obs.	53	53	53
Panel C: Individual outcomes			
Mean ($\hat{\mu}$)	0.033	0.102	0.141
SE ($\hat{\sigma}_{\mu}$)	(0.018)	(0.167)	(0.151)
Standard deviation ($\hat{\tau}$)		0.539	0.471
Effective.SE		0.564	0.495
Num.obs.	374	374	374

Notes: See notes in previous table.

Table 6: Meta-analysis on studies where authors estimate LATE/ATT

	TE	WW	DDML
	(1)	(2)	(3)
Panel A: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.201	−0.052	−0.041
SE ($\hat{\sigma}_{\mu}$)	(0.063)	(0.047)	(0.045)
Standard deviation ($\hat{\tau}$)		0.162	0.154
Effective.SE		0.168	0.161
Num.obs.	21	21	21
Panel B: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.061	0.100	0.058
SE ($\hat{\sigma}_{\mu}$)	(0.048)	(0.045)	(0.039)
Standard deviation ($\hat{\tau}$)		0.155	0.121
Effective.SE		0.161	0.127
Num.obs.	21	21	21
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.139	−0.057	−0.051
SE ($\hat{\sigma}_{\mu}$)	(0.043)	(0.035)	(0.035)
Standard deviation ($\hat{\tau}$)		0.171	0.168
Effective.SE		0.175	0.171
Num.obs.	117	117	117
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.061	−0.018	−0.022
SE ($\hat{\sigma}_{\mu}$)	(0.020)	(0.030)	(0.029)
Standard deviation ($\hat{\tau}$)		0.212	0.190
Effective.SE		0.215	0.192
Num.obs.	866	866	866

Notes: Column 1 presents the results of the meta-analysis on experimental treatment effects, column 2 is the bias of the simple with-without estimator (selection bias), and column 3 is the bias of the DDML estimator. Effective SE $= \left(\sqrt{\hat{\sigma}_{\hat{\mu}}^2 + \hat{\tau}^2}\right)$. Panel A includes one aggregated outcome generated from all primary outcomes in each study, panel B includes one aggregated outcome generated from all outcomes in each study, panel C shows the results from using all primary outcomes in each study, and panel D shows the results from using all individual outcomes in each study.

Table 7: Meta-analysis on clustered RCTs

	TE (1)	WW (2)	DDML (3)
Panel A: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.183	−0.080	−0.070
SE ($\hat{\sigma}_{\mu}$)	(0.060)	(0.044)	(0.036)
Standard deviation ($\hat{\tau}$)		0.165	0.122
Effective.SE		0.171	0.127
Num.obs.	28	28	28
Panel B: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.078	0.040	0.004
SE ($\hat{\sigma}_{\mu}$)	(0.040)	(0.038)	(0.019)
Standard deviation ($\hat{\tau}$)		0.133	0.000
Effective.SE		0.138	0.019
Num.obs.	28	28	28
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.127	−0.062	−0.056
SE ($\hat{\sigma}_{\mu}$)	(0.045)	(0.029)	(0.028)
Standard deviation ($\hat{\tau}$)		0.178	0.168
Effective.SE		0.180	0.170
Num.obs.	170	170	170
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.050	−0.022	−0.035
SE ($\hat{\sigma}_{\mu}$)	(0.024)	(0.025)	(0.023)
Standard deviation ($\hat{\tau}$)		0.229	0.194
Effective.SE		0.23	0.196
Num.obs.	982	982	982

Notes: See notes in previous table.

Table 8: Meta-analysis on indivdually randomised RCTs

	TE (1)	WW (2)	DDML (3)
Panel A: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.140	0.022	−0.017
SE ($\hat{\sigma}_{\mu}$)	(0.045)	(0.081)	(0.082)
Standard deviation ($\hat{\tau}$)		0.237	0.242
Effective.SE		0.251	0.255
Num.obs.	14	14	14
Panel B: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.025	0.090	0.094
SE ($\hat{\sigma}_{\mu}$)	(0.059)	(0.091)	(0.087)
Standard deviation ($\hat{\tau}$)		0.278	0.260
Effective.SE		0.292	0.274
Num.obs.	15	15	15
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.144	−0.021	−0.057
SE ($\hat{\sigma}_{\mu}$)	(0.040)	(0.086)	(0.085)
Standard deviation ($\hat{\tau}$)		0.312	0.300
Effective.SE		0.324	0.312
Num.obs.	94	94	94
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.033	0.152	0.124
SE ($\hat{\sigma}_{\mu}$)	(0.024)	(0.121)	(0.088)
Standard deviation ($\hat{\tau}$)		0.498	0.362
Effective.SE		0.513	0.373
Num.obs.	815	815	815

Notes: See notes in previous table.

Table 9: Meta-analysis on eligibility design studies

	TE (1)	WW (2)	DDML (3)
Panel A: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.129	−0.020	−0.032
SE ($\hat{\sigma}_{\mu}$)	(0.026)	(0.050)	(0.043)
Standard deviation ($\hat{\tau}$)		0.207	0.171
Effective.SE		0.212	0.176
Num.obs.	31	31	31
Panel B: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.064	0.099	0.058
SE ($\hat{\sigma}_{\mu}$)	(0.040)	(0.051)	(0.047)
Standard deviation ($\hat{\tau}$)		0.206	0.176
Effective.SE		0.212	0.182
Num.obs.	30	30	30
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.116	−0.034	−0.037
SE ($\hat{\sigma}_{\mu}$)	(0.021)	(0.046)	(0.040)
Standard deviation ($\hat{\tau}$)		0.238	0.207
Effective.SE		0.243	0.211
Num.obs.	183	183	183
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.033	0.118	0.094
SE ($\hat{\sigma}_{\mu}$)	(0.014)	(0.064)	(0.047)
Standard deviation ($\hat{\tau}$)		0.397	0.290
Effective.SE		0.402	0.294
Num.obs.	1335	1335	1335

Notes: See notes in previous table.

Table 10: Meta-analysis on encouragement design studies

	TE (1)	WW (2)	DDML (3)
Panel A: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.328	−0.112	−0.085
SE ($\hat{\sigma}_{\mu}$)	(0.179)	(0.077)	(0.059)
Standard deviation ($\hat{\tau}$)		0.194	0.135
Effective.SE		0.208	0.148
Num.obs.	11	11	11
Panel B: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.071	−0.013	−0.009
SE ($\hat{\sigma}_{\mu}$)	(0.074)	(0.026)	(0.024)
Standard deviation ($\hat{\tau}$)		0.000	0.000
Effective.SE		0.026	0.024
Num.obs.	13	13	13
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.207	−0.113	−0.089
SE ($\hat{\sigma}_{\mu}$)	(0.122)	(0.043)	(0.038)
Standard deviation ($\hat{\tau}$)		0.239	0.204
Effective.SE		0.243	0.207
Num.obs.	81	81	81
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.098	−0.079	−0.066
SE ($\hat{\sigma}_{\mu}$)	(0.059)	(0.053)	(0.051)
Standard deviation ($\hat{\tau}$)		0.275	0.250
Effective.SE		0.280	0.255
Num.obs.	462	462	462

Notes: See notes in previous table.

Table 11: Meta-analysis on studies where number of covariates is greater than median

	TE (1)	WW (2)	DDML (3)
Panel A: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.162	0.014	0.002
SE ($\hat{\sigma}_{\mu}$)	(0.070)	(0.061)	(0.050)
Standard deviation ($\hat{\tau}$)		0.198	0.147
Effective.SE		0.208	0.156
Num.obs.	18	18	18
Panel B: Aggregated all outcomes			
Mean ($\hat{\mu}$)	-0.008	0.094	0.057
SE ($\hat{\sigma}_{\mu}$)	(0.010)	(0.053)	(0.040)
Standard deviation ($\hat{\tau}$)		0.154	0.099
Effective.SE		0.163	0.107
Num.obs.	18	18	18
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.089	-0.007	0.001
SE ($\hat{\sigma}_{\mu}$)	(0.051)	(0.058)	(0.054)
Standard deviation ($\hat{\tau}$)		0.268	0.232
Effective.SE		0.274	0.239
Num.obs.	100	100	100
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.046	-0.002	-0.020
SE ($\hat{\sigma}_{\mu}$)	(0.030)	(0.040)	(0.031)
Standard deviation ($\hat{\tau}$)		0.235	0.187
Effective.SE		0.239	0.190
Num.obs.	981	981	981

Notes: See notes in previous table.

Table 12: Meta-analysis on studies where number of covariates less than median

	TE (1)	WW (2)	DDML (3)
Panel A: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.157	−0.097	−0.093
SE ($\hat{\sigma}_{\mu}$)	(0.043)	(0.057)	(0.051)
Standard deviation ($\hat{\tau}$)		0.208	0.183
Effective.SE		0.215	0.190
Num.obs.	24	24	24
Panel B: Aggregated all outcomes			
Mean ($\hat{\mu}$)	0.095	0.019	−0.001
SE ($\hat{\sigma}_{\mu}$)	(0.054)	(0.059)	(0.052)
Standard deviation ($\hat{\tau}$)		0.224	0.180
Effective.SE		0.231	0.187
Num.obs.	25	25	25
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.132	−0.069	−0.080
SE ($\hat{\sigma}_{\mu}$)	(0.028)	(0.035)	(0.034)
Standard deviation ($\hat{\tau}$)		0.172	0.171
Effective.SE		0.175	0.174
Num.obs.	164	164	164
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.042	0.055	0.060
SE ($\hat{\sigma}_{\mu}$)	(0.016)	(0.092)	(0.064)
Standard deviation ($\hat{\tau}$)		0.496	0.355
Effective.SE		0.504	0.360
Num.obs.	816	816	816

Notes: See notes in previous table.

Table 13: Meta-analysis on studies where lagged outcomes are present

	TE	WW	DDML
	(1)	(2)	(3)
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.203	-0.153	-0.126
SE ($\hat{\sigma}_{\mu}$)	(0.056)	(0.069)	(0.047)
Standard deviation ($\hat{\tau}$)		0.291	0.189
Effective.SE		0.299	0.195
Num.obs.	94	94	94
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.093	-0.003	-0.026
SE ($\hat{\sigma}_{\mu}$)	(0.042)	(0.044)	(0.031)
Standard deviation ($\hat{\tau}$)		0.271	0.200
Effective.SE		0.274	0.202
Num.obs.	497	497	497

Notes: See notes in previous table. Since the aggregated outcomes are based on several outcomes that may each have an individual lagged outcome, we do not provide Panel A (aggregated primary outcomes) and B (aggregated all outcomes).

E Appendix - Standard Error Robustness

As explained in Appendix B.2, and focusing on a single outcome per study, our main analysis computes the variance of each individual bias estimate assuming that $\widehat{EXP}_s$ and $\widehat{OBS}_s$ are independent, i.e., it does not take into account the covariance between our experimental and observational estimator. We use as our estimand of the variance of selection bias $\sigma^2_{B,s} = \sigma^2_{OBS,s} + \sigma^2_{EXP,s}$ instead of $\sigma^2_{B,s,true} = \sigma^2_{OBS,s} + \sigma^2_{EXP,s} - 2Cov(\widehat{EXP}_s, \widehat{OBS}_s)$. It is likely that $\widehat{EXP}_s$ and $\widehat{OBS}_s$ are positively correlated since the treated units are the same in both analyses. As a consequence, our approach gives an upper bound on the true variance of selection bias as $\sigma^2_{B,s} = \sigma^2_{B,s,true} + 2Cov(\widehat{EXP}_s, \widehat{OBS}_s)$.

This section explores robustness of our main result to relaxing the independence assumption, both theoretically, and using the bootstrap.

E.1 Bootstrap

Bootstrapping our estimates is computationally costly because it involves repeatedly re-running the machine-learning observational estimators. Table 14 does this just for the "aggregate primary" outcomes which reduces the number of specifications we must re-estimate. We find very similar albeit slightly smaller estimates to our primary analysis, with a mean bias of -0.032 and an effective SE of 0.154. Thus, our overall conclusions do not appear to be materially affected by the independence assumption.

E.2 Theoretical analysis

We estimate $\hat{\tau}^2$ using the restricted maximum likelihood estimator. To give intuition to how sensitive this estimator might be to our assumption that the experimental and observational estimates are independent, consider the closely-related Hedges' Estimator, which has a simpler

Table 14: Bias estimates using bootstrap standard errors

	TE (1)	WW (2)	DDML (3)
Panel A: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.139	−0.039	−0.032
SE ($\hat{\sigma}_{\mu}$)	(0.036)	(0.041)	(0.034)
Standard deviation ($\hat{\tau}$)		0.200	0.150
Effective.SE		0.204	0.154
Num.obs.	42	42	42

Notes: Column 1 presents the results of the meta-analysis on experimental treatment effects, column 2 is the bias of the simple with-without estimator (selection bias), and column 3 is the bias of the DDML estimator. All results are based on the aggregated primary outcomes using bootstrap standard errors. Effective SE $= \sqrt{\hat{\sigma}_{\hat{\mu}}^2 + \hat{\tau}^2}$. We provide results based on bootstrap standard errors solely for our main specification, the aggregated primary outcomes, due to computational constraints.

formula (see Chabé-Ferret (2023) for details):[32]

$$\begin{aligned} \hat{\tau}^2 &= \hat{\sigma}_{tot}^2 - \bar{\sigma}^2 \\ \text{where} \quad \hat{\sigma}_{tot}^2 &= \frac{1}{S}\sum_{s=1}^{S}(\widehat{B}_s - \overline{B})^2 \\ \overline{B} &= \frac{1}{S}\sum_{s=1}^{S}\widehat{B}_s \\ \bar{\sigma}^2 &= \frac{1}{S}\sum_{s=1}^{S}\hat{\sigma}_{B,s}^2. \end{aligned}$$

We have:

$$\begin{aligned} \bar{\sigma}^2 &= \frac{1}{S}\sum_{s=1}^{S}\hat{\sigma}_{B,s,true}^2 + 2\frac{1}{S}\sum_{s=1}^{S}Cov(\widehat{EXP}_s, \widehat{OBS}_s) \\ &= \bar{\sigma}_{true}^2 + 2\overline{Cov} \\ \hat{\tau}^2 &= \hat{\sigma}_{tot}^2 - \bar{\sigma}_{true}^2 - 2\overline{Cov} = \hat{\tau}_{true}^2 - 2\overline{Cov}. \end{aligned}$$

[32]The actual estimator we are using is

$$\hat{\tau}_{REML}^2 = \frac{\sum_{s=1}^{S}\left(\frac{1}{\hat{\sigma}_{B,s}^2+\hat{\tau}^2}\right)^2\left[(\widehat{B}_s - \hat{\mu})^2 - \hat{\sigma}_{B,s}^2\right]}{\sum_{s=1}^{S}\left(\frac{1}{\hat{\sigma}_{B,s}^2+\hat{\tau}^2}\right)^2} + \frac{1}{\sum_{s=1}^{S}\frac{1}{\hat{\sigma}_{B,s}^2+\hat{\tau}^2}}.$$

The solution is recursive estimation until convergence. This also involves re-estimating $\hat{\mu}$.

Therefore, assuming $\widehat{EXP_s}$ and $\widehat{OBS_s}$ are independent will tend to lead us to underestimate the effective SE if they are in reality positively correlated ($\overline{Cov} > 0$).

Given these formulas, by calculating the mean covariance between $\widehat{EXP_s}$ and $\widehat{OBS_s}$ across our studies we can get a ballpark estimate of by how much we underestimate $\hat{\tau}^2$. Using the meta-analytic correlation between all included experimental and observational estimates, we compute (for the aggregated primary outcomes):

$$\hat{\tau}_{true} = \sqrt{\hat{\tau}^2 + \frac{2}{S}\sum_{s=1}^{S} \widehat{corr}(\widehat{EXP_s}, \widehat{OBS_s}) * \hat{\sigma}_{EXP,s}\hat{\sigma}_{OBS,s}} = 0.325.$$

Where $\overline{corr}(\widehat{EXP_s}, \widehat{OBS_s})$ is the estimated correlation. The calculation is based on an uncorrected estimated Hedges' estimator of $\tau = 0.277$.[33] Thus this back-of-the-envelope calculation is consistent with the claim that our main results do not materially overestimate the effective SE.

[33] Using the REML estimator, we find a corrected $\tau_{REML} = 0.196$.

F Appendix - Description of studies

In this appendix we provide a detailed description of each study included in our analysis.

nr	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 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%) 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% 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. 21 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% of teachers in program schools. Compliance with the training was high, with 93% 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.	Take-up of hygienic 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% 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% (1,563) of those surveyed in the treatment areas had taken out Compartamos loans during the study period, compared to only 5.8% (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% of teachers in program schools. Compliance with the training was high, with 93% 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 public 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 they reported any illness episode, number of episodes and patient age.
15	**Title:** Does Community-Based Development Empower Citizens? Evidence from a Randomized Evaluation in Ghana. **Authors:** Baldwin, Kate; Karlan, Dean; Udry, Christopher; Appiah, Ernest. **Journal:** Working Paper. **Year published in repository:** 2017.	The “community-based development” approach may empower citizens and improve outcomes through different mechanisms. Using a randomized evaluation of a nongovernmental-organization-led CBD program in Ghana, the authors examine whether community-based development results in citizens’ empowerment to improve their socioeconomic well-being through these mechanisms.	Randomized communities were invited to participate in The Hunger Project’s (THP) Vision, Commitment and Action (VCA) workshops and invited to build an epicenter.	28 of the 51 village groupings invited to take part actually began the THP process. All but three of these groupings successfully completed construction of the epicenter building, and four groupings built two epicenter buildings.	Quality of Village Leadership Index, contributions to public goods in non-THP sectors, number of candidates in district assembly election, proportion of Non-THP Sectors with local government-funded projects (education, road, power, agricultural processing).

16	**Title**: Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State. **Authors**: 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.
17	**Title:** Channeling Remittances to Education: A Field Experiment among Migrants from El Salvador. **Authors:** 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).
18	**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.

19	**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% of adults in developing countries report having a bank account, compared to 94% 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 expeditures.
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 take-up 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% repectively.	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**: Bruhn, Miriam; Karlan, Dean; 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**: Home- and community-based growth monitoring to reduce early life growth faltering: an open-label, cluster-randomized controlled trial. **Authors**: Fink, Günther; Levenson, Rachel; Tembo, Sarah; Rockers, Peter C. **Journal**: The American Journal of Clinical Nutrition. **Year published in repository:** 2018.	Despite the continued high prevalence of faltering growth, height monitoring remains limited in many low- and middle-income countries. The objective of this study was to test whether providing parents with information on their child's height can improve children's height and developmental outcomes.	Villages in Chiapata district, Zambia, were randomly assigned to 1 of 3 intervention groups to increase parents' awareness of their children's growth trajectories: (1) Home-based growth monitoring (HBGM) (2) Community-based growth monitoring including nutritional supplementation for children with stunted growth (CBGM+NS) and (3) Control.	More than 75% did attend the meeting. Caregivers reported actively using the poster at a measurement frequency similar to that. 97.5% of posters were still hanging at caregivers' homes at the study's end.	Individual height-for-age z score (HAZ), food diversity, and overall child development.
23	**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.

24	**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 conterfactual 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.
25	**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% and earnings by 38%. 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	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.

26	**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% of all public primary schools, serving 8.6% 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.
27	**Title**: Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification. **Authors**: Chong, 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.

28	**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.
29	**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 improved 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.

30	**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% of children (including 13% 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.
31	**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 kilometer. Furthermore, only about 42% 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 installments 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.

32	**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% 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% 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.
33	**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%) accepted the offer and opened the account.	Change in total balance (6 months, 12 months). Change in non-seed balances (12 months).

34	**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.
35	**Title**: The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India. **Authors**: Banerji, Rukmini; Berry, James; Shotland, Marc. **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2017.	Using a randomized field experiment in India, the authors evaluate the effectiveness of adult literacy and parental involvement interventions in improving children's learning.	In the states of Bihar and Rajasthan, 240 hamlets (village subdivisions) were randomly assigned in equal proportions to the control group or to one of the three treatment groups. Households were assigned to receive either adult literacy (language and math) classes for mothers, training for mothers on how to enhance their children's learning at home, or a combination of the two programs.	Self-reported attendance: 40% of mothers in ML and 45% of mothers in ML-CHAMP reported having attended ML Classes.19% of selected children in ML villages and 25% of selected children in ML-CHAMP villages were reported to have attended with the mother. ML attendance collected by Pratham volunteers: take-up of 76% in ML and 84% in ML-CHAMP.	Children's test scores (math) and mothers' test scores (language, math, total), and mother's participation.
36	**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 received them. Nevertheless, all children were tested, regardless of whether or not they participated in the program.	Test score in math, language and total.

37	**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.
38	**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.

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% of buildings in the teament group were visited by canvassers.	Participation in regional and catonal elections.
40	**Title**: How to Promote Order and Property Rights under Weak Rule of Law? An Experiment in Changing Dispute Resolution Behavior through Community Education. **Authors**: Blattman, Chris; Hartman, Alexandra; Blair, Robert. **Journal**: American Political Science Review. **Year published in repository:** 2018.	Dispute resolution institutions facilitate agreements and preserve the peace whenever property rights are imperfect. In weak states, strengthening formal institutions can take decades, and so state and aid interventions also try to shape informal practices and norms governing disputes.The authors study the short-term impact of a alternative dispute resolution campaign in Liberia using a randomized experiment.	Alternative dispute resolution (ADR) campaign in rural Liberian communities. Out of 246 communities, 116 were initially randomly assigned to treatment. 16 out of those were assigned to an intense treatment. Treatment was sequential Treated communities were randomly assigned to 1 phase over the 5 of the program (each phase represented a time range).	resource constraints meant UNHCR stopped in Phase 4, with 85 communities treated out of the 86 assigned to Phases 1 to 4. The 30 randomly assigned to Phase 5 were assigned to the control group.	Survey replies: any unresolved/ resolved land dispute, dispute resulted in property desctruction, and satisfied with outcome.
41	**Title**: Does working from home work? Evidence from a Chinese experiment. **Authors**: Bloom, Nicholas; Liang, James; Roberts, John; Ying, Zhichun Jenny. **Journal**: Quarterly Journal of Economics. **Year published in repository:** 2018.	A rising share of employees now regularly engage in working from home (WFH), but there are concerns this can lead to ‘‘shirking from home.’’ The authors conduct a WFH experiment to measure its impact.	Call center employees were randomly assigned to WFH or in the office. The WFH treatment was four shifts (days) a week at home and the fifth shift in the office on a fixed day of the week determined by the firm.	Individuals who are interested in WFH get selected to work from home but some may return to work after special circumstance. 80 - 90% of the treatment group was actually working at home.	Employee performance, Log phone calls per minute, employee satisfaction.

42	**Title**: Ready for Boarding? The Effects of a Boarding School for Disadvantaged Students. **Authors**: Behaghel, Luc; de Chaisemartin, Clément; Gurgand, Marc. **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2018.	The authors analyze the effects of a French “boarding school of excellence” on students’ cognitive and non-cognitive outcomes using a randomized experiment. The authors followed the treatment and the control groups over two years after the lottery.	The school was oversubscribed, and students offered a seat were randomly selected out of the pool of applicants.	86% of lottery winners enrolled in the school, and 76% of them stayed until the end of the academic year. By contrast, 6% of lottery losers managed to enroll because one of their siblings had been admitted to the school. 5% stayed until the end of the year.	Student's test scores in Mathematics and well-being related survey replies.
43	**Title**: Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions. **Authors**: Gerber, Alan S.; Karlan, Dean; Bergan, Daniel. **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2018.	The authors conduct a field experiment to measure the effect of exposure to newspapers (the Washington post or the Washington times) on political behaviour and opinion.	The authors sampled households in the Prince William County and selected individuals who did not already subscribe to either the Washington post and the Washington times. These households were randomly assigned to either one of two treatment groups or the control group. Treatment was a free subscription for ten weeks to the Times or the Post.	There were three noncompliance issues to note regarding treatment administration. (1) 6% of households in the treatment groups opted out of the free subscription. (2) Some addresses (76 for the Times, 1 for the Post) were deemed “undeliverable”. (3) 75 (out of 965) were already on the Post and 5 were already in the Times subscription.	Self-reported and administrative voting data, voted for Democrat, did not vote, but preferred Democrat.
44	**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% 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.

G Appendix - Quality of studies

nr	Study	Exclusion restriction	Attrition	Spillovers	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% 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% 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 of hygienic latrines does not affect open defectation 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% 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:** Does Community-Based Development Empower Citizens? Evidence from a Randomized Evaluation in Ghana. **Authors:** Baldwin, Kate; Karlan, Dean; Udry, Christopher; Appiah, Ernest. **Journal:** Working Paper. **Year published in repository:** 2017.	Broadly, a violation seems unlikely as the offer to participate in the workshop should not affect the outcomes other than through the programme. However if people attend the workshops and villagers do not mobilize as a whole, then a violation might be possible.	The research team was able to resurvey 74% of baseline households. They examined whether the treatment affects the likelihood of attrition, and have found no empirical evidence that suggests concerns of bias due to attrition from the survey sample frame.	There may be spillovers for individual level take-up of attending any VCA session.We decided not to record these as take-up measures.	Up to 3786 households and 122 electoral areas.
16	**Title**: Can Employment Reduce Lawlessness and Rebellion? A Field Experiment with High-Risk Men in a Fragile State. **Authors**: 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.

17	**Title:** Channeling Remittances to Education: A Field Experiment among Migrants from El Salvador. **Authors:** 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.
18	**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.	The offer of CBT and grant do not affect noncognitive skills and preferences in ways other than the program.	7.6% attrition, not differential in observables across groups.	The authors work in large neighborhoods, recruiting less than 1% 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.	Up to 947 observations.
19	**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.

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 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.
21	**Title**: The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico. **Authors**: Bruhn, Miriam; Karlan, Dean; 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**: Home- and community-based growth monitoring to reduce early life growth faltering: an open-label, cluster-randomized controlled trial. **Authors**: Fink, Günther; Levenson, Rachel; Tembo, Sarah; Rockers, Peter C. **Journal**: The American Journal of Clinical Nutrition. **Year published in repository:** 2018.	The offer to get any treatment should not affect the individual height or overall child development other than the program.	About 5% Attrition. No statistically significant differences were found in follow-up rates across groups.	Parents who attended the meeting could share information with others in the village who did not attend or who were not invited to attend.	Up to 497 Children.
23	**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.
24	**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.

25	**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.	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, the authors used a two-phase tracking approach. The response rate was 97% at baseline, and effective response rates at endline (where individuals found in phase 2 tracking were given higher weights) were 90.7% after two years (2010), 84% after four (2012) and 87% after nine (2017).	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.
26	**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.
27	**Title**: Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification. **Authors**: Chong, 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 received 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.

28	**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.
29	**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.
30	**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.
31	**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%) could be resurveyed.	By August 2009 27% of control households had appliad 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.

32	**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% of the households participate in the first main two surveys and about 81% 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.
33	**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.
34	**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%, similar across arms for treatment outcomes.	Spillovers were minimized as patients with another household member in the study were ineligible.	Up to 2207 observations.

35	**Title**: The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India. **Authors**: Banerji, Rukmini; Berry, James; Shotland, Marc. **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2017.	The authors explain that there exist the possibility that the programs affected children directly. They find suggestive evidence that in the case of ML the impact is limited but in case of CHAMP the impacts may play a greater role.	Approximately 3.5% of households reached for surveys and testing at baseline were not reached at endline. Endline child tests are available for 94% of children tested at the baseline. There does not seems to be evidence of differential attrition across treatment groups at the household level, but there is some imbalance of attrition levels among child test-takers between the CHAMP and ML-CHAMP groups and the control group.	No evidence of spillovers across program hamlets but 7% of mothers in the CHAMP and control groups reported attending ML classes.	Up to 18283 observations.
36	**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%, respectively, in the comparison and treatment groups in Vadodara in year 1, 4% in both the treatment and the comparison group in Vadodara in year 2. In Mumbai it was 7% and 7.5%, respectively, in the treatment and comparison groups in year 1, and 7.7% and 7.3%, 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.
37	**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.

38	**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 the 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.
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**: How to Promote Order and Property Rights under Weak Rule of Law? An Experiment in Changing Dispute Resolution Behavior through Community Education. **Authors**: Blattman, Chris; Hartman, Alexandra; Blair, Robert. **Journal**: American Political Science Review. **Year published in repository:** 2018.	It is unlikely that the invitation to participate in the workshop affects the outcomes directly. The authors discuss the potential of the impact of facilitators instead of the workshop but argue against it.	Endline data on 243 of the 246 communities. Nonresponse within village was typically less than 5-10% per community. Attrition of targeted residents was 13%.	Communities were located far from each other, with little risk of spillovers between them. However there might be spillovers effects on untrained individuals within communities.	Up to 5435 residents and 940 Leaders.
41	**Title**: Does working from home work? Evidence from a Chinese experiment. **Authors**: Bloom, Nicholas; Liang, James; Roberts, John; Ying, Zhichun Jenny. **Journal**: Quarterly Journal of Economics. **Year published in repository:** 2018.	The authors discuss the possibility of a violation of the exclusion restriction but provide additional robustness results to argue against such violation.	The authors acknowledge that the results may be biased by attrition, but biased downward, so the true impact of WFH is probably substantially larger.	Given that the employees work in the call center, there appear to be no obvious spillovers from the WFH employees to the rest of the team.	249 of 957 employeees took part in the experiment for 85 time periods.

42	**Title**: Ready for Boarding? The Effects of a Boarding School for Disadvantaged Students. **Authors**: Behaghel, Luc; de Chaisemartin, Clément; Gurgand, Marc. **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2018.	Not discussed. It is unlikely that the offer of a place changes the outcomes other than through the boarding school itself.	10% of the students didn't take the follow-up tests. Attrition was balanced in treatment and control groups.	Not directly discussed but if the applicants come from similar neighborhoods, the existence of spillovers might be possible. However, Students not enrolled in the boarding school were scattered among 169 schools. Most of them were in the local school district of Creteil, but some of them were in other areas of France. This may have limited spillovers.	Up to 381 students over 2 years.
43	**Title**: Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions. **Authors**: Gerber, Alan S.; Karlan, Dean; Bergan, Daniel. **Journal**: American Economic Journal: Applied Economics. **Year published in repository:** 2018.	Not discussed but there may be a small possibility for the outcomes being affected by the offered subscription and not by the take-up if the randomization is a reminder to stay well-informed.	32.3% of individuals interviewed at the baseline were re-interviewed at the follow up survey but for the main outcomes, the authors have administrative data. Attrition appears to be balanced across treatment and control group.	May be possible if households live nearby. Given the random selection of households within a county, they do however appear to be unlikely.	Up to 1081 respondents.
44	**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.

Table 15: Summary statistics by study

Study	# Specifications	Average # covariates	Average # observations	Average take-up ($R = 1$)
1	5	34.00	1777.00	0.24
2	32	49.00	1935.31	0.88
3	18	18.00	965.00	0.28
4	62	61.00	1138.90	0.46
5	11	15.00	243.64	0.43
6	12	37.00	246.75	0.74
7	10	7.30	2138.50	0.89
8	6	30.00	7405.17	0.47
9	5	210.00	875.20	0.42
10	72	22.00	14954.39	0.18
11	21	39.33	13103.52	0.99
12	34	115.00	4927.24	0.17
13	2	112.00	652.00	0.69
14	25	49.00	704.12	0.40
15	101	125.00	1920.72	0.56
16	55	658.00	1024.20	0.74
17	51	16.00	716.55	0.12
18	376	1613.00	643.52	0.94
19	50	412.76	5879.36	0.30
20	16	941.38	11474.00	0.45
21	60	392.00	332.87	0.53
22	10	23.00	322.10	0.84
23	8	72.00	511.38	0.53
24	6	23.00	1661.00	0.66
25	91	49.00	1584.98	0.87
26	36	7.69	2381.47	0.87
27	3	16.00	2039.33	0.90
28	59	541.00	780.24	0.91
29	19	16.00	248.42	0.68
30	8	116.00	6647.38	0.08
31	33	885.45	596.76	0.76
32	42	649.67	2151.17	0.82
33	39	24.92	10396.31	0.94
34	12	114.00	4981.50	0.86
35	123	38.52	5361.28	0.73
36	3	6.00	9986.00	0.64
37	49	16.43	5355.78	0.56
38	3	64.00	2688.67	0.94
39	6	105.00	19597.50	0.91
40	36	29.00	3616.50	0.86
41	19	1.00	5723.42	0.93
42	119	45.00	289.50	0.79
43	14	36.00	609.57	0.55
44	35	117.00	21584.54	0.43

Notes: Column 2 represents the number of different outcome-treatment-take-up combinations for each study. Column 3 provides the average number of covariates available to the DDML and PDSL estimator. The number of covariates can differ e.g. due to different units of analysis. Column 4 represents the average number of observations used in the estimation of the experimental estimator. Column 4 displays the average take-up in the treatment group.