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*

December 20, 2025

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 53 development RCTs with imperfect compliance (ICRCTs), we estimate the parameters required to construct our confidence intervals, and illustrate their use. Our confidence intervals allow observational estimates to be used even when there are doubts about identification, have close to nominal coverage, lead to power gains in meta-analysis, and enable researchers to choose between RCT and observational methods based on power. A key takeaway of our findings is that observational methods have significantly lower power than suggested by conventional confidence intervals.

*Bernard: Open Philanthropy (david.rhys.bernard@gmail.com). Bryan: London School of Economics (g.t.bryan@lse.ac.uk). Chabé-Ferret: Toulouse School of Economics (sylvain.chabe-ferret@tse-fr.eu). de Quidt: Queen Mary University of London and Institute for International Economic Studies (j.dequidt@qmul.ac.uk). Fliegner: University of Manchester (jasmin.fliegner@manchester.ac.uk). Rathelot: Institut Polytechnique de Paris (ENSAE) (roland.rathelot@ensae.fr). 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, Kieran Byrne, Louise Demoury, 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.

1 Introduction

The past decades have seen large advances in quasi-experimental program evaluation, and a rise in the use of Randomized Controlled Trials (RCTs) (Angrist and Pischke 2010; Duflo et al. 2007). Despite this, observational studies remain popular, likely reflecting the difficulty of finding naturally-occurring exogenous variation, and the high perceived logistical and monetary cost of RCTs.[1] How much should we trust these studies, which lack clear identifying variation, how can they be made more useful for those who worry about the biases they may contain, and how can a researcher decide when it is worth paying the cost of running an RCT?

We address these questions by estimating the *distribution* of observational bias, and using those estimates to provide bias corrected confidence intervals for observational estimates. We further establish consistency of the proposed confidence intervals. Our confidence intervals increase the usefulness of observational studies by: 1) providing a quantitative estimate of their uncertainty and allowing them to be used even when there are doubts about identification; and 2) allowing them to be combined with RCT estimates in meta-analyses, adding to power. Our approach also allows a researcher to decide between an RCT and observational study using power calculations. Our proposal moves beyond the influential approach of LaLonde (1986), which has shown that observational studies can be biased, and allows a researcher to *use* an observational estimate, while being clear about its limitations.

Our analysis is restricted to observational methods that leverage a cross section of data. To motivate our confidence intervals, 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 $\hat{\sigma}_\epsilon$.[2] 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 what we need to estimate and how our confidence intervals behave. First, in addition 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

[1] Appendix Figure D.1 shows the continued popularity of matching methods, a leading observational method, and the recent rapid growth of double debiased machine learning.

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

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 fundamental variation in observational bias (τ^2), that matter. This is the key area in which our work differs from the seminal paper of LaLonde (1986) and the literature that followed. That literature shows that observational estimates can sometimes be adjusted to recover a related experimental estimate, but that this is not always the case, and gives no guidance about when this will work or how to use observational estimates in general, given this uncertainty.[3]

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 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 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 dominates the effective standard error, meaning observational bias becomes especially 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 yields 53 different trials, with an average of about 48 potentially-biased parameter estimates 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.[4] This assumption is common in the literature, especially in the literature seeking to estimate treatment effects in a new study setting by extrapolating from existing estimates (e.g. Menzel (2024) and Gechter (2022)). We discuss this assumption and its validity at length in the robustness section of our paper.

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

[3] 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. (1998a); Smith and Todd (2005).

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

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,[5] 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 common support all hold.[6] In an encouragement design, again with standard assumptions,[7] an ICRCT allows an instrumental variables 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.[8] 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; we will also show that the findings are robust to restricting attention to eligibility designs.

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).[9] 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.

[5] *Independence*: assignment to treatment (eligibility) is independent of potential outcomes and potential take-up. *First stage*: assignment to treatment increases the probability of take-up. *SUTVA* (Stable Unit Treatment Value Assumption): i's potential outcomes are independent of j's take-up. *Exclusion*: assignment to treatment only affects outcomes through take-up. See Appendix A.1 for formal definitions.

[6] 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.1 for formal definitions.

[7] 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.1.

[8] We later show robustness to potential biases in the RCT estimates themselves, for example failure of Exclusion.

[9] 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 presence of some extreme outliers.

Our use of ICRCTs also means that experimental and observational estimates are created using 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 53 studies and recover our three key parameters. This requires that all biases are measured on a common scale, so we make two normalizations. We measure bias in standard deviations of the observational control group outcome.[10] And we align outcomes such that a positive treatment effect always indicates an increase in welfare (based on a manual coding of each outcome variable) and thus a positive bias indicates an overestimate of the welfare benefits of the program being evaluated. Our main results take a single primary bias estimate from each study (based on an index of the most important welfare measures), but we find extremely similar results under other aggregations including an omnibus of 2540 individual bias estimates.

The results 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 negative mean bias of -0.025 standard deviations. This implies that observational studies do not systematically over- or underestimate the welfare impact of the programs they evaluate. Second, variability is large. The standard error of the average bias is 0.038, while our estimate of τ is 0.204. Interpreting these numbers through the lens of the confidence interval in (1), the effective standard error of an infinite-N observational study is 0.207 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 two to three times larger than this.[11] 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 cases increases uncertainty relative to a specification without covariates.

We then turn to illustrating the usefulness of our confidence intervals. We first show that our adjusted intervals give far better coverage than unadjusted intervals, and hence allow a researcher using observational methods to give a more honest, quantitative sense of estimation uncertainty. We demonstrate this within our own sample, and in an entirely new setting. Within our own sample, we re-estimate the meta-analysis leaving out each study one by one, and use the new values of $\hat{\mu}, \hat{\sigma}^2_\mu$, and $\hat{\tau}$ to adjust the confidence interval of the left-out study. We measure coverage as the proportion of observational confidence intervals that include the experimental

[10] To be precise, we use the standard deviation of outcomes among the pooled untreated units within each study, regardless of the randomization arm. We choose this standardization (rather than one based on the RCT control group) because it is always available to our hypothetical policymaker that wants to bias-correct their observational study.

[11] 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. Minimal detectable effect size reflects the fact that it is difficult to distinguish true effects from statistical artefacts because estimates vary across samples taken from the same population, all the more so if samples are small and outcomes are highly variable. We argue that variation in selection bias across programs, outcomes, locations and sets of control variables also generates variability for observational estimates, and this variability persists even with infinitely large sample sizes. We use the rule of thumb of $MDE = 2.8 * \text{effective SE}$.

point estimate. The results are stark: for example, after using DDML to try and correct for bias, only 68% of uncorrected 95% confidence intervals contain the experimental estimate, rising to 90% after our bias correction (much closer to the nominal 95% level). For a tougher, more clearly out-of-sample test, we collected data from 11 LaLonde-type studies that compare observational and experimental estimates. This gives us 580 new point estimates in which the researcher has taken some effort to eliminate observational bias. Nevertheless, conventional 95% confidence intervals around these estimates contain the experimental estimate only 60% of the time, while our bias-corrected confidence intervals (based on $\hat{\mu}, \hat{\sigma}^2_{\mu}, \hat{\tau}$ in our sample) are almost dead on the nominal level, at 94%. Perhaps unsurprisingly this improvement in honesty comes at a substantial cost to power, which we discuss at length throughout the paper.

We next show how our approach can support meta-analyses that combine experimental and observational estimates. There is no current agreed upon approach to do this. A researcher can include potentially-biased observational studies, knowing that they may skew the findings and be overweighted due to exaggerated precision; or exclude them, throwing away valuable data. Our confidence intervals place observational and experimental estimates on the same footing, with appropriate measures of precision. We show the value of this approach by re-analyzing a meta-analysis of interventions to prevent sexual violence (Porat et al., 2024). In a meta-analysis of impacts on perpetration, RCT-based estimates are positive (i.e., a welfare-improving reduction in perpetration), but noisy, meaning there is no statistically detectable effect of the programs. Observational estimates are smaller in magnitude on average, but less noisy, nevertheless a researcher may be concerned that this is driven by selection bias rather than a true effect. Correcting the observational estimates using our method and combining observational and RCT estimates resolves the problem: there is a large and statistically significant reduction in perpetration caused by the programs on average.

Finally, we show how our approach allows for a better-informed choice between observational and experimental methods. We ask at what sample size an RCT has a smaller expected standard error than an infinite-N bias-corrected observational study. We find that a perfect-compliance RCT can have a smaller expected standard error with a sample of just 93 observations. Things look better for observational studies if there is imperfect compliance, but even with only 25% compliance an RCT would still only need 1487 observations to dominate.

These results demonstrate the validity and usefulness of our method. Next, we assess whether relaxing our main assumptions can increase the power of bias-corrected observational estimates.

Our key assumption is exchangeability: the policy maker has no good reason to estimate the distribution of bias for her study using only a subset of our studies, so uses the full sample. We have already shown that this assumption seems reasonable overall, as it leads to good coverage both within and outside of our sample. Many readers may still worry that we are lumping together studies that draw from *predictably different* bias distributions. For example, they might argue that an expert could divide studies into those likely to be positively biased and those likely to be negatively biased, and by so doing decrease uncertainty about residual bias. We show

empirically that this is not justified. We first use a frontier Large Language Model (GPT 4.1) to mimic an incredibly well-informed literature expert, and ask the model to predict bias direction for each study in our sample based on detailed information about the intervention and eligible population. While the LLM shows some ability to distinguish studies with more-negative or more-positive bias, and is able to provide convincing rationales behind each prediction, this has at best a modest effect on reducing effective standard errors.

A second concern might be that our sample includes poor-quality bias estimates that in turn bias $\hat{\mu}$ and $\hat{\tau}$. We show that our findings are robust to diverse attempts to remove potentially poor-quality estimates: we drop RCTs with encouragement designs; a weak first stage; few regression covariates; doubts about SUTVA; doubts about the exclusion restriction. Our estimate of the effective standard error decreases by at best 20%, far from enough to transformatively improve power of a bias-corrected observational study. We also show robustness to other modeling and estimation choices such as our parametric Normality assumption.

Our paper is inspired by the pioneering work of LaLonde (1986), which showed that RCT estimates were hard to recover with observational approaches. We move beyond testing *whether* observational methods can recover experimental estimates, and provide a tool that can make them more useful *even when they cannot*. We have additional advantages over much of that literature. Our use of ICRCTs and access to rich micro data means we can use the same datasets to estimate experimental and observational estimates; and we emphasize the use of hands-off estimators to reduce researcher degrees of freedom.

The recent paper by Gechter (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, Gechter is 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 he concentrates on studying average bias, while we place more emphasis on uncertainty. He also concentrates on two literatures—microcredit and cash transfers—while we consider a broader set of studies. His paper does however raise the possibility that our RCT database may not be representative of all observational settings, due to site-selection bias. This limits the application of our methods to places where it would be plausible to run an ICRCT. Empirically, he finds limited, although noisily estimated site-selection bias. Our paper is also related to Angrist et al. (2017), who use school admission lotteries to validate and improve observational value-added models in education.

The paper is structured as follows. Section 2 summarizes the methods we use to estimate and aggregate bias. Section 3 describes our data set of ICRCTs. Section 4 summarizes the results of our main meta-analysis. Section 5 presents our three applications: adjusting individual studies, adjusting meta-analyses, and deciding between conducting an RCT or an observational study. Section 6 evaluates whether we can improve the power of corrected observational studies using expert (LLM) predictions of bias, by dropping low-quality bias estimates, or relaxing other

identifying assumptions. Section 7 concludes.

2 Overview of Methods

In this section we give an overview of the methods we use to estimate $\{\mu, \sigma_\mu^2, \tau^2\}$, 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.

2.1 Study-Level Estimators of Bias

Our goal is to provide adjusted 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. Under the three assumptions of conditional independence, common support and SUTVA,[12] 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 observational methods as 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 estimand 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 TOT^{EXP} is close to TOT, then $\widehat{B}$ will be a good estimate of the bias B_0 that we are interested in. Our experimental estimand 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 in the experiment. 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. Following Bloom (1984), we define TOT^{EXP} as the ratio of the intent-to-treat estimand and the take-up rate

[12] Appendix A.1 provides formal definitions of all the identification assumptions discussed in this section.

in the treatment group.

It is well known that, under the standard IV assumptions for the validity of the RCT (independence, first stage, SUTVA, and exclusion), the Bloom estimand recovers TOT, the average treatment effect among those 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 can use observations from the eligible group to form an observational estimator that also identifies TOT by comparing those who take up to those who do not, conditional on observables (see below for details of the estimators we use). It then follows that, provided the assumptions for the validity of the RCT hold, $TOT^{OBS} - TOT^{EXP}$ identifies the observational bias.

One issue is worth noting at this point. The set of people who choose to select in will be differently selected in each study, and so the relevant population to which the TOT applies will vary from study to study. At face value that might suggest a failure of exchangeability. Our approach to this issue (and more generally) is to observe that a priori there is a complete lack of knowledge about differences in the distribution of bias between population groups, thus no reason to consider any study in particular to be exchangeable or not with the policy maker's study of interest. We revisit whether this is reasonable when we analyze Robustness in Section 6.

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, under the standard IV assumptions (independence, non-zero first stage, SUTVA, and exclusion) plus monotonicity, the Wald estimand (the intent-to-treat effect divided by the difference in compliance rates across treatment arms) identifies the average treatment effect for the compliers (those who are induced to take up by the incentive), which we denote TOC^{EXP} (Imbens and Angrist 1994). It is also well known that, with an encouragement design, it is not possible to identify TOT from the experiment alone, which appears to create a problem for us.

We address this problem as follows. Theorem 1 in Appendix A.1 states 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. The effect on the treated in the treatment group reflects the average effect for both always-takers and compliers, since both groups receive treatment when assigned

to treatment. The effect on the treated in the control group reflects only the effect for always-takers, as they are the only units who receive treatment when assigned to control. Therefore, by appropriately weighting out the contribution of always-takers, we can recover the effect for compliers.

With an experimental and an observational estimator for the treatment effect for compliers in hand, if the assumptions for experimental validity hold, then $TOC^{OBS} - TOC^{EXP}$ identifies observational bias for TOC.

Our final concern is that our goal is to estimate the bias in observational estimates of TOT, not TOC. We start from the observation 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. Note 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 the taker and complier populations. We also show in Section 6 that excluding encouragement designs altogether has little impact on our overall conclusions.

In summary then, for each eligibility-design study $s = 1, ..., S$, and each individual regression specification $o = 1, ..., N_s$ available within that study, our bias estimate is:

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

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 specifications per study later in this section, after we talk about our chosen estimators. We also calculate a standard error $\hat{\sigma}_{B,os}$ for each bias estimate. Appendix B explains how we do this.

Validating experimental identification assumptions. Our approach to ensuring the experimental identification assumptions hold is twofold. First, we have concentrated on gathering data from high-quality RCTs, as we discuss below. Second, it is possible to exclude potentially problematic bias estimates from our sample. We pursue this approach in section 6 below, and argue there that our results are robust to the exclusion of these studies.

2.2 Experimental Estimator

We produce our experimental estimates using a basic 2SLS regression including dummies for all strata on which the randomization was stratified.

2.3 Choice of Observational (Hands-off) Estimators

To create our bias estimates we need to decide on observational 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).[13] 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 that 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 (Table D.1).

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 test 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, 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.[14] Splitting helps reduce concerns about over-fitting. When implementing this approach we use all available covariates in the study dataset; the regularization in the ML method effectively chooses which controls to use. Chernozhukov et al. (2018) show that DDML yields 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.[15]

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

[14] 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.

[15] When implementing DDML we always use a random forest as the machine learning method because this means

2.4 Aggregating Estimates of Bias and Forming Confidence Intervals

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

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. Condition (2) may seem like a strong assumption, but it is a simple way to capture our key exchangeability assumption, and we show in section 6 that it approximates the data well.

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. 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 less beneficial than they believe. A negative bias has the opposite interpretation. τ^2 measures the fundamental variance in observational bias across studies, and is in some sense a measure of our ignorance.

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. Equation (3) describes a random-effect meta-analysis in which, as standard in this literature, the variance $\sigma^2_{\hat{B},s}$ is replaced by our estimated variance $\hat{\sigma}^2_{\hat{B},s}$. It can be efficiently and consistently estimated using Restricted Maximum Likelihood (Raudenbush, 2009; Chabé-Ferret, 2023).

Performing this analysis requires that biases 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

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

study. Second, in line with our interpretation of positive bias as exaggerating welfare benefits, we manually align the signs of all outcome variables such that more positive values imply higher social welfare, all else equal.[16] Our meta-analysis then returns $\{\hat{\mu}, \hat{\tau}^2, \hat{\sigma}_\mu^2\}$ 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 $\widehat{TOT}_p^{OBS}$ has been constructed, with standard error $\hat{\sigma}_{\epsilon,p}$. It follows from equation (3), and the normality of the error, that $\widehat{TOT}_p^{OBS} \sim \mathcal{N}(TOT_p + \mu, \sigma_{\epsilon,p}^2 + \tau^2)$, with the implication that

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

which leads to the confidence interval formula (1) discussed in the introduction when replacing the variances with their estimated values. Theorem 2 in Appendix A.2 proves the consistency of these corrected confidence intervals.

Figure 1 gives a simple visual presentation of this confidence region, with solid lines representing the usual confidence intervals based only on sampling error, and dashed lines representing bias-adjusted confidence intervals. The diagram assumes $\hat{\mu} = 0$ (to centre both confidence intervals on zero), while the y-axis is $\hat{\sigma}_{\epsilon,p}$, which is specific to our policymaker's observational study. In both cases, study effects that lie outside of the funnel would be considered to have statistically significant effects, and studies within 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 *uncertainty* about the extent of the bias, captured by $\hat{\tau}$ and $\hat{\sigma}_\mu^2$, 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.[17] 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 $\hat{\sigma}_{\epsilon,p}$ 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 indeed 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 $\hat{\sigma}_\mu^2$, or allow

[16] A positively-coded outcome is one where a positive effect would increase social welfare, all else equal (e.g., income, health), a negatively-coded 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.

[17] 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.

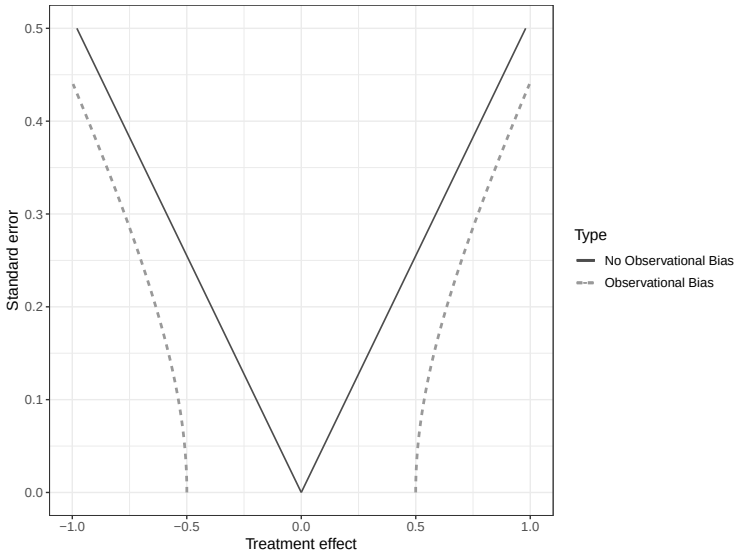


Figure 1: Funnel plot showing examples of bias-corrected and uncorrected rejection regions

Note: This figure presents the rejection regions under the null of no treatment effect for the classical approach which only accounts for sampling noise (*No Observational Bias*), and our proposed corrected approach which accounts for both sampling noise and uncertainty about observational bias (*Observational Bias*). Both sets are built using the formula in Equation (1) as a function of $\hat{\sigma}_\epsilon^2$ (represented as is usual in the y-axis), with $\delta = 0.95$. Uncorrected regions are built using $\hat{\mu} = \hat{\sigma}_\mu = \hat{\tau} = 0$. Corrected regions are built using $\hat{\mu} = 0$, $\hat{\sigma}_\mu = 0.05$ and $\hat{\tau} = 0.25$.

the policy maker to concentrate on a set of ICRCTs that are more similar to her own, which might reduce $\hat{\tau}$.

Finally, our concerns about observational bias are relatively less important for small-N observational 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, given the increasing availability of very large observational data sets.

2.5 Extension to Multiple Bias Estimates Per Study

Each of our studies includes multiple outcome variables, each of which capture some aspect of welfare (Table D.1); potentially multiple randomized manipulations (e.g. offer of different types of credit product); and potentially multiple different definitions of a "program" that is being taken up.[18] As a result, we can obtain multiple bias estimates per study. In principle this provides useful additional data, but introduces two challenges. First, we need to account for the correlation structure between bias estimates within a study. Second, it is a priori unclear which outcome variables are most representative of the welfare measure our hypothetical policy maker would be interested in.

Accounting for within-study correlation. We do this in two ways. First, we construct an index

[18]The definition of "program" can be subtle and in our analysis is synonymous with a measure of take-up. For example, a program could correspond to a taking specific microfinance loan, or to borrowing from any microfinance lender, both of which are influenced by the randomized entry of a microfinance lender into the community.

that aggregates all outcome measures within a single study, 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—when sampling variances differ—provides the efficient estimator of a latent index that places equal welfare weight on each outcome, an approach that we find appropriate given the lack of detailed information on policy makers' preferences.

Second, we allow for multiple bias estimates per study, but adjust for within-study correlation, so that we do not exaggerate the precision of our findings. To do this, we remain in the classical meta-analytical framework, but follow Pustejovsky and Tipton (2021) in allowing for within-study correlation in both effects and errors. Specifically, we generalize (3) to:

$$\hat{B}_{os} = \mu + \omega_s + \iota_{os} + \nu_{os} \tag{4}$$

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.[19] 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

$$\widehat{TOT}^{OBS} - \hat{\mu} \pm \Phi^{-1}\left(\frac{1+\delta}{2}\right)\sqrt{\hat{\sigma}_\epsilon^2 + \hat{\sigma}_\mu^2 + \hat{\xi}_\omega^2 + \hat{\xi}_\iota^2}. \tag{5}$$

From now on we denote $\hat{\tau}^2 = \hat{\xi}_\iota^2 + \hat{\xi}_\omega^2$, 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.

Distinguishing primary and secondary outcomes. Some outcomes available in a given study 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 effects on her primary outcome(s) of interest are unbiased. We distinguish between primary and secondary outcomes 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 aggregated using the indexing approach described above, or individually). We also produce estimates for all outcomes in the paper, either individually, or aggregated.

Taken together, our approach yields four different meta-analyses which we label Aggregated primary outcomes; Aggregated all outcomes; Individual primary outcomes; and Individual all outcomes.

[19] We follow the literature by assuming a value for ρ (see e.g. Viechtbauer (2021)) which we set to 0.6. Results are robust to different parameter values as discussed in Section 6. We also take into account additional variation by clustering our standard errors in the meta-analysis at the study level.

2.6 Quality Checks

To ensure the quality of our estimates we take the following steps. First, we automatically determine the experimental design (eligibility/encouragement) of each specification, where a specification is a combination of estimator, study, and outcome.[20] Second, we remove outliers, defined as any specification where the absolute value of the experimental estimate is larger than two standard deviations. Third, we remove specifications with a weak first stage by requiring a Kleibergen-Paap F-statistic larger than 10.

When forming aggregate outcomes, we group by experimental treatment, program, and unit of analysis (e.g. we separate individual vs. household outcomes) and aggregate within these groups. In practice that means that we often have several specifications per study remaining after aggregation (e.g., corresponding to two randomized treatment arms). For example, we might have an individual-level aggregate, and a group level aggregate. To come to a single outcome per study we multiply the share of compliers by the number of experimental units and select the estimate with the highest value.

3 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.

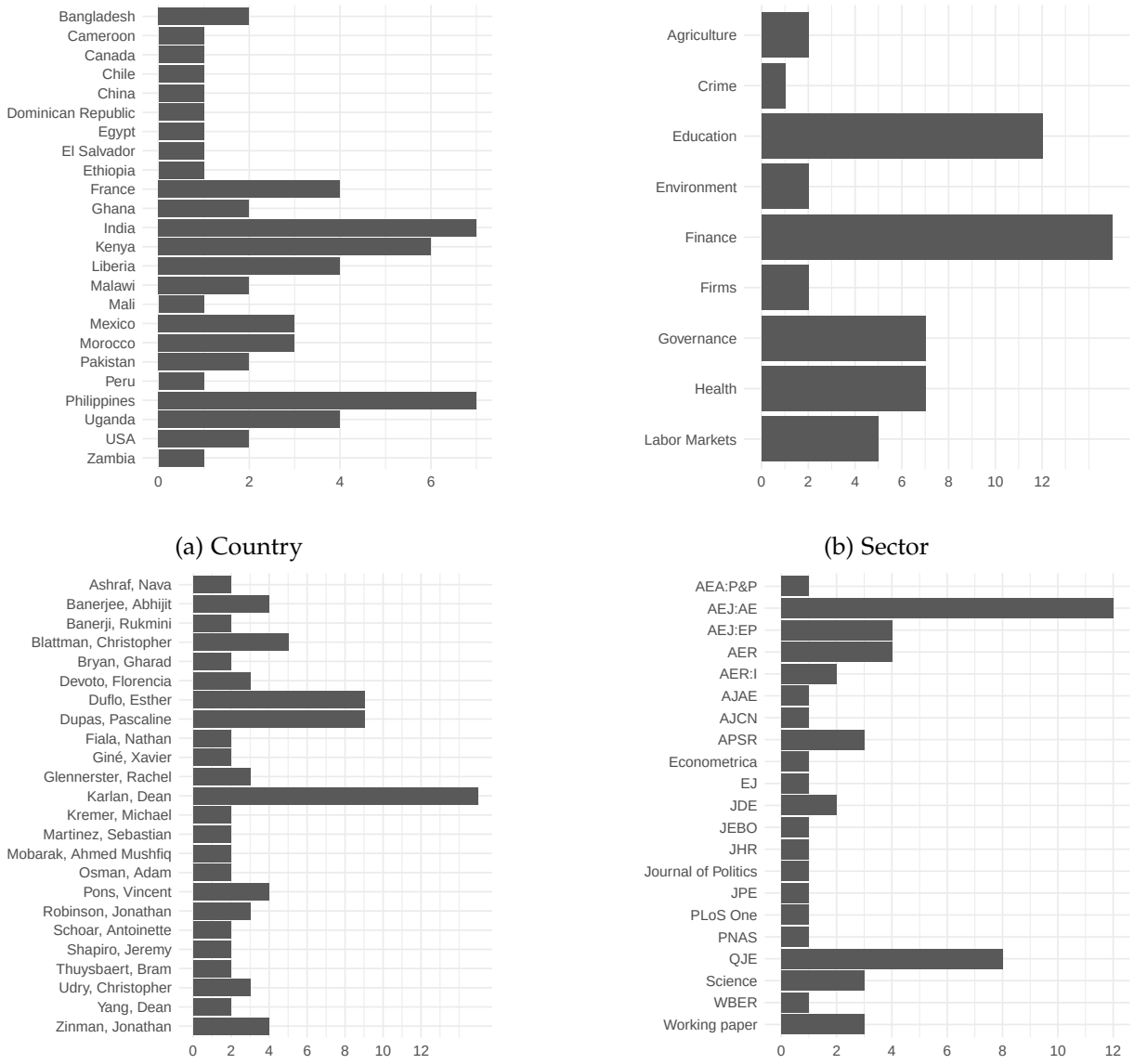
3.1 Study Selection

We start with 207 studies from the IPA and J-PAL dataverses. Within this set of studies, we select all ICRCTs where the data record treatment assignment and take-up, and at least one outcome variable (left-hand-side variable in a regression reported in the paper). This leaves us with a sample of 53 ICRCTs (see Appendix C for details about the screening process and a list of the

[20] 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, in which case we force take-up to be zero (if in the experimental control arm) or one (if in the treatment arm). The design is then determined as 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.

included studies). We have on average 48 distinct specifications (randomized manipulation×take-up measure×outcome combinations) per study, and 10 primary specifications per study. Our "Individual all outcomes" meta-analysis consists of 2540 bias estimates. For additional study-level summary statistics, see Table D.1.

3.2 Description of ICRCT Sample



(a) Country

(b) Sector

(c) Authors (with at least 2 RCTs)

(d) Journal

Figure 2: Study characteristics

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 Canada) and Europe (France). Studies from countries with IPA or J-PAL hubs are strongly represented. India and the Philippines appear the most in our analysis, with Kenya, Uganda, Liberia and France 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 and governance, all common areas of study within development economics. Panel 2c shows authors who appear at least twice in our dataset. Almost all of these authors are prominent development economists.

Sampling from repositories may give us broader coverage of types of study, and a weaker publication filter compared with e.g. sampling publications in top-5 economics journals (see Appendix E.3 for more discussion of publication bias). Panel 2d shows that we indeed have broad coverage of high-impact economics journals, as well as a number of journals from other disciplines and working papers.

It is difficult to assess whether our sample is representative of the kinds of studies a policy maker would be interested in, but we can at least descriptively compare the included studies to those that appear in the repositories but were excluded in our screening process. Appendix Figure D.2 shows the distributions of locations (by continent) and sector for the two samples. There are some differences but no clear patterns of exclusion, except that no Gender studies are included in our analysis while around 6% of excluded studies are assigned to this "sector."

4 Meta-analytic 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.[21] Our preferred specification is presented in panel A, which shows results for one aggregated primary outcomes measure per study. The second panel shows results aggregating all study-level outcomes, the third panel show the results for each individual primary outcome, and the fourth panel for every outcome variable.

Column "TE" presents a meta-analysis of the experimental treatment effects. Unsurprisingly, average effects on primary outcomes are larger than when looking at all outcomes within a paper (some of which correspond to robustness checks or low-priority analyses). On average, the programs we study improve our aggregate measure of welfare by 0.173 standard deviation units.

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 -0.025 standard deviations, around 14% as large as the treatment effect of the average program. In addition to the small size, average bias is never significant. We conclude that there is little evidence that observational studies systematically over

[21] While we have $N = 53$ studies in total, two of the aggregated primary outcome specifications do not survive our screening for outliers/first stage hence $N = 51$ in panel A.

Table 1: Meta-analysis of Bias

	TE	WW	PDSL	DDML
Panel A: Aggregated primary outcomes				
Mean ($\hat{\mu}$)	0.173	-0.030	-0.055	-0.025
SE ($\hat{\sigma}_{\mu}$)	(0.041)	(0.044)	(0.050)	(0.038)
Standard deviation ($\hat{\tau}$)		0.251	0.277	0.204
Effective SE		0.255	0.281	0.207
Num. obs.	51	51	50	51
Panel B: Aggregated all outcomes				
Mean ($\hat{\mu}$)	0.085	0.023	0.038	0.007
SE ($\hat{\sigma}_{\mu}$)	(0.032)	(0.038)	(0.041)	(0.035)
Standard deviation ($\hat{\tau}$)		0.213	0.218	0.182
Effective SE		0.216	0.222	0.185
Num. obs.	53	53	52	53
Panel C: Individual primary outcomes				
Mean ($\hat{\mu}$)	0.154	-0.051	-0.058	-0.053
SE ($\hat{\sigma}_{\mu}$)	(0.030)	(0.037)	(0.032)	(0.030)
Total standard deviation ($\hat{\tau}$)		0.280	0.231	0.226
Effective SE		0.282	0.233	0.228
Num. obs.	540	540	534	540
Panel D: Individual outcomes				
Mean ($\hat{\mu}$)	0.077	-0.010	-0.033	-0.029
SE ($\hat{\sigma}_{\mu}$)	(0.017)	(0.023)	(0.024)	(0.017)
Total standard deviation ($\hat{\tau}$)		0.257	0.324	0.213
Effective SE		0.258	0.325	0.213
Num. obs.	2540	2540	2368	2540

Notes: This table presents estimates of the meta-analytic regression introduced in Equation (3) (Panels A and B) and its generalized version introduced in Equation (4) (Panels C and D) for experimental treatment effects (column 1); the bias of the simple with-without estimator (*i.e.* selection bias) (column 2); the bias of the post double selection lasso estimator (column 3); and the bias of the DDML estimator (column 4). We also present estimates of the effective standard error: *Effective SE* $= \sqrt{\hat{\sigma}_{\mu}^2 + \hat{\tau}^2}$. *Num. obs.* reports the number of treatment effect or bias estimates used in each regression. 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. The total standard deviation $\hat{\tau} = \sqrt{\hat{\xi}_{\iota}^2 + \hat{\xi}_{\omega}^2}$ reported in panels C and D captures both the between-study and within-study components of variance.

or underestimate program impacts *on average*.

Small average bias does not mean that observational biases are small. We define the *effective standard error* as the standard error of a hypothetical infinite-N observational study (Effective SE := $\sqrt{\hat{\sigma}_{\mu}^2 + \hat{\tau}^2}$). These are always large, regardless of the observational method or sample. Looking across the table, the smallest effective standard error is 0.185. That in turn implies that an infinite-N observational study would still have a minimum detectable effect size of $2.8 \times 0.185 = 0.52$ standard deviations, which is large relative to typical treatment effect magnitudes (including those estimated in our "TE" column). 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 all panels in terms of having a smaller effective standard error. Further, PDSL often performs worse than the naive with-without comparison (and we had to trim some extreme outlier estimates, explaining the slight reduction in sample sizes). This poor performance potentially reflects over-fitting and finite-sample instability. Noting this, we focus the ensuing discussion on results from DDML and WW.

Figure 3 provides another way to look at the results. Each point represents the effect size and standard error of an observational estimate in our sample. As in Figure 1, the solid lines show a standard confidence interval that accounts only for sampling error; estimates outside the funnel would be deemed statistically significant at the 5% level. The dashed lines show our adjusted confidence intervals, which take into account uncertainty about observational bias. Panels (a) and (b) show results for all outcomes, (c) and (d) for aggregated primary outcomes. We see that adjusted confidence intervals are much wider than the standard intervals, and there are few effect-size estimates large enough to remain significant after adjusting for uncertainty about bias (the table notes go into greater detail on how many lose significance in each case).

We can also use the same figure to compare across different observational methods. The left figure of each pair shows that the naive with-without estimator has a larger confidence region than the DDML method shown on the right. By adjusting for selection on observables and thereby reducing uncertainty about bias, DDML improves the power of the observational study (MDE falls from 0.71 to 0.58 standard deviations in the case of our preferred aggregate primary specification).

5 Applications

This section applies our findings to three domains: constructing more honest confidence intervals for observational studies; increasing power in meta-analyses by combining experimental and observational estimates; and choosing between experimental and observational approaches.

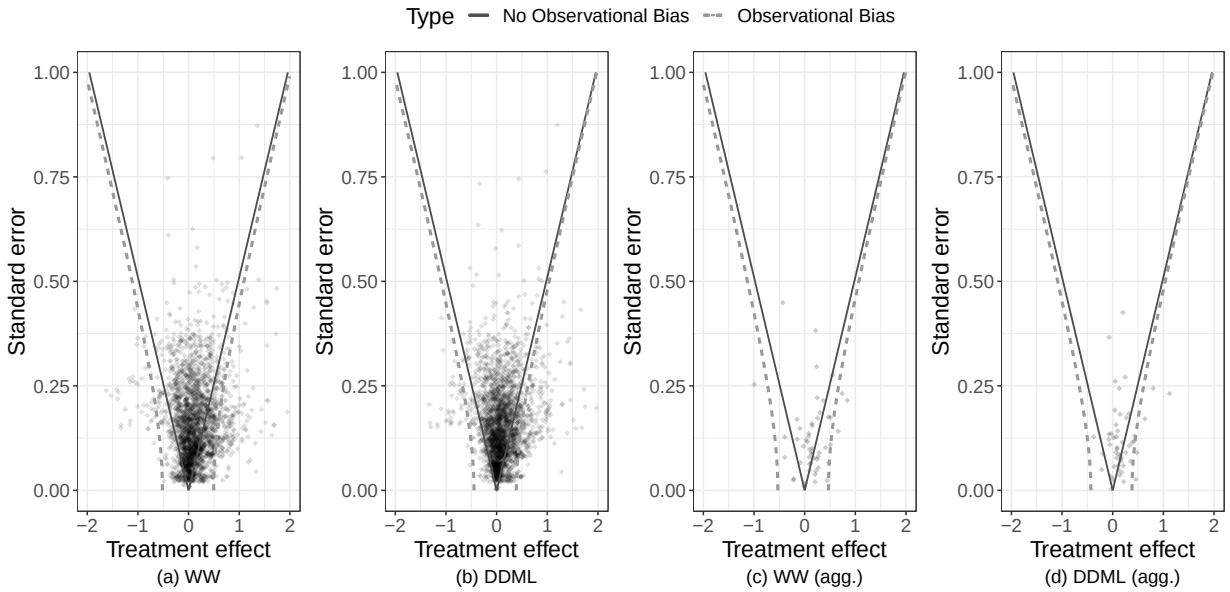


Figure 3: Funnel plot of observational treatment effect estimates with corrected and uncorrected confidence regions

Note: 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), 883 treatment effects are statistically significant whereas only 215 remain statistically significant after correction. For the DDML in panel (b), 732 uncorrected treatment effects are statistically significant whereas only 251 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), 23 uncorrected treatment effects are statistically significant and only 4 remain so. For the DDML in panel (d), 23 uncorrected treatment effects are statistically significant and only 5 remain so.

5.1 More Honest Confidence Intervals

A first goal of our approach is to provide confidence intervals for observational methods that are comparable to those from an RCT. That is, confidence intervals that are highly likely to include the true *causal* effect.

5.1.1 Leave-One-Out Confidence Interval Correction in Our Sample

For each study in the sample, we remove it from the dataset and re-estimate the bias distribution using only the remaining studies. We then construct a bias-adjusted confidence interval for the treatment effect in the omitted study. This simulates a practical scenario in which a researcher uses previously accumulated evidence to correct an observational estimate in a new setting.

Figure 4 plots the results. For each study in our sample we show the point estimate and confidence interval for three approaches: the experimental estimate, the observational estimate based on DDML, and the bias-corrected observational estimate. All results use only aggregated primary outcomes, and impacts are measured in standard deviations of the outcome variable. We can see numerous cases where the experimental and uncorrected DDML confidence intervals have

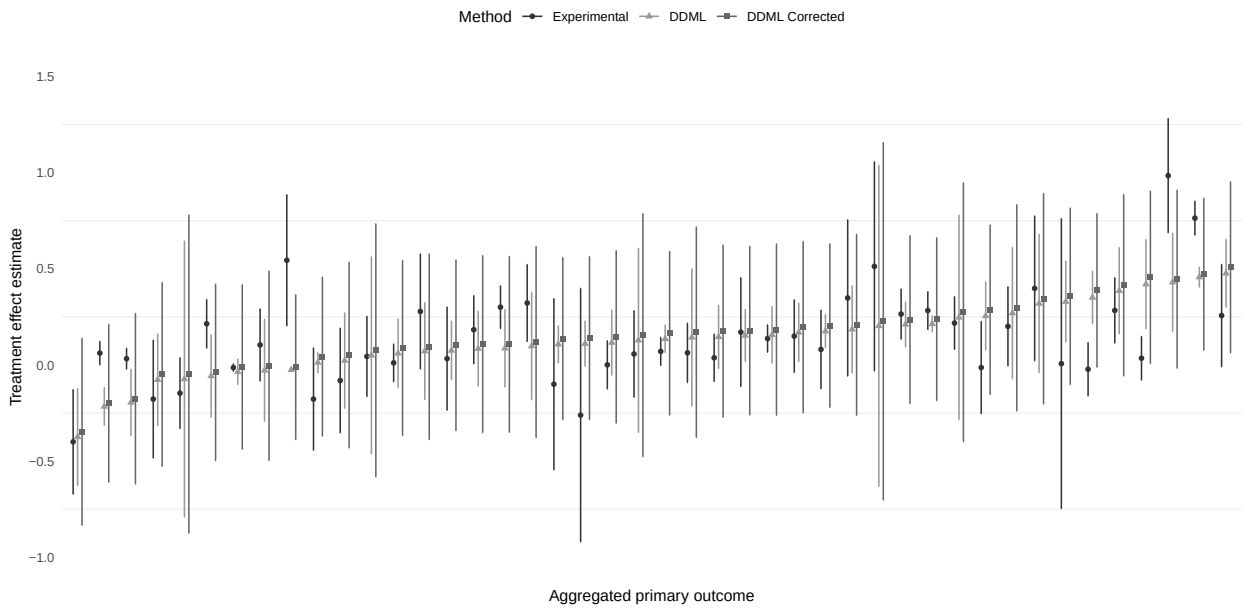


Figure 4: Corrected and uncorrected observational confidence intervals compared to RCT estimates

Note: This figure presents point estimates and confidence intervals of the effect of one treatment per study on aggregated primary outcomes. We report estimates for the treatment with the largest number of compliers in each study. Experimental estimates are from 2SLS regressions with strata fixed effects. Uncorrected observational estimates are DDML estimates. Corrected observational estimates apply the correction in Equation (1) using leave-one out estimates of $\hat{\mu}$, $\hat{\sigma}_\mu$ and $\hat{\tau}$. Leave-one out estimates are built by estimating Equation (3) by Restricted Maximum Likelihood on the sample of aggregated primary outcomes excluding the study for which we are building the correction.

zero overlap, whereas the corrected DDML confidence intervals do overlap with the experimental one, suggesting qualitatively that the correction is working. 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. Next, we quantify these observations, first within our sample of studies, and second by applying our correction to estimates from outside our sample.

Figure 5 provides a more focused summary of the results. The figure first shows that our proposed confidence intervals have much improved coverage rates. To calculate coverage we count the frequency with which the observational confidence interval contains the experimental point estimate. A valid 95% confidence interval should do so 95% of the time. Uncorrected coverage is just 67% under WW and 68% under DDML. Failing to account for observational bias leads to confidence intervals with very distorted size. In contrast, our corrected confidence intervals achieve 92% and 90% coverage respectively, much closer to the 95% target. If anything our proposed correction remains slightly conservative, and will still tend to over reject.[22]

The figure then shows that this improvement in coverage comes at a significant power cost. To

[22]Note that the coverage rate we report in Figure 5 is an estimate of the true coverage rate. Part of the distance between the coverage rate we report and the nominal rate is due to sampling noise in the experimental treatment effect, which has nothing to do with the actual performance of our corrected confidence interval.

calculate power we look at the sample of cases in which the experimental estimate is statistically significant, and report the proportion where the observational estimate is also significant and of the same sign. Our proposed correction decreases power from about 55% to about 25%, suggesting that observational studies have much lower power than usually reported. In section 6 we consider whether we can increase power by relaxing some of our assumptions.

Finally, Figure 5 reports a small drop in the reversal rate. The reversal rate is the proportion of studies where the experimental estimate is significant and the observational estimate is also significant but of the opposite sign. Within our sample reversals are rare – 1% uncorrected, 0% after correction, this is not the case for our second exercise below.

In sum, uncorrected observational confidence intervals have poor coverage, and overstate power. Our approach significantly improves coverage, and provides a more realistic reflection of power.

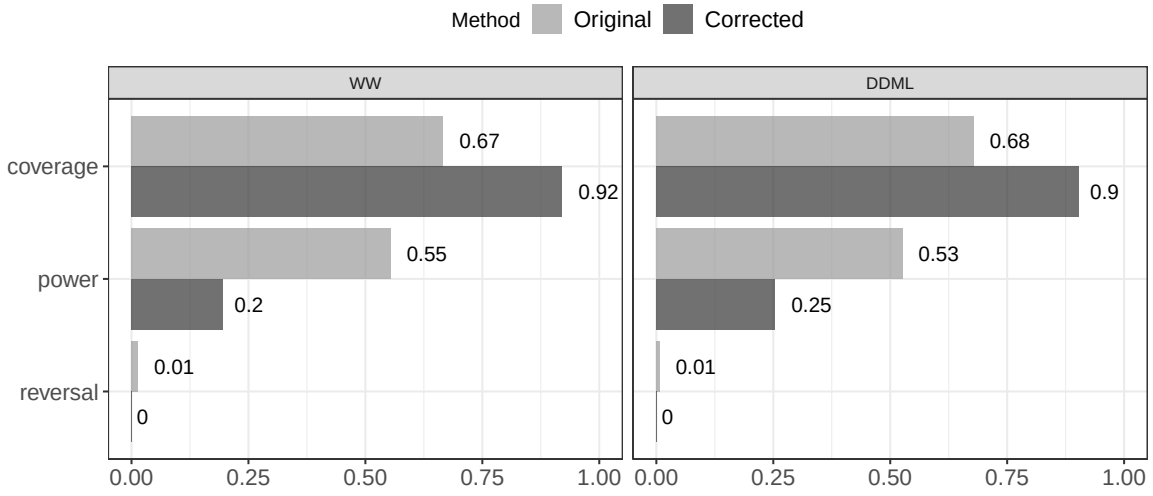


Figure 5: Leave-one-out estimates of the properties of corrected and uncorrected observational estimates

Note: This figure presents leave-one out estimates of the performance of corrected and uncorrected confidence intervals for observational estimators on the sample of aggregated primary outcomes. Corrected observational estimates apply the correction in Equation (1) using leave-one out estimates of $\hat{\mu}$, $\hat{\sigma}_\mu$ and $\hat{\tau}$. Leave-one out estimates are built by estimating Equation (3) by Restricted Maximum Likelihood on the sample of aggregated primary outcomes excluding the study for which we are building the correction. We measure *coverage* as the proportion of times the observational confidence intervals contain the experimentally estimated treatment effect. We measure *power* as the proportion of times the observational confidence intervals detect a statistically significant treatment effect of the same sign as the corresponding statistically significant experimental treatment effect. We measure *reversal* as the proportion of times the observational and experimental confidence intervals are fully located on opposite sides of zero.

5.1.2 Validation Using LaLonde-Style Studies

Next, we conduct an out of sample validation exercise. We compiled a set of 11 LaLonde-style studies that compare observational estimates to experimental benchmarks from similar settings.[23] For each study we correct the observational estimates and confidence intervals using the bias estimates from our meta-analysis of all outcomes. To determine which correction to use, we classify each observational method into two categories:

- **Unadjusted comparisons**, where no attempt is made in the source paper to adjust for covariates (e.g., raw difference in means). For these, we apply the correction derived from the WW bias distribution. ($P = 44$ comparisons).
- **Adjusted comparisons**, where the source paper explicitly attempts to control for confounding (e.g., matching, regression, difference-in-differences). For these, we apply the correction derived from the DDML bias distribution. ($P = 580$ comparisons).

This exercise differs from our earlier leave-one-out analysis in that it is fully out-of-sample: the LaLonde-style papers come from a different literature and use a variety of observational methods, whereas our bias distribution is estimated from JPAL and IPA RCTs, mostly on development themes, and we always use the same observational methods. This is a substantially more challenging test of our approach.

Figure 6 shows the results of this exercise. Once again our approach performs well with respect to coverage. When the authors make no attempt to adjust observational estimates, only 14% of naive observational confidence intervals include the experimental estimate. This proportion rises to 41% when we apply our correction. Better, but still not great. When the authors do attempt to adjust for potential confounding, naive observational estimates include the experimental estimate 57% of the time, suggesting that the adjustments are useful, but do not go far enough. Finally coverage jumps to 94% when we apply our correction, tantalizingly close to the 95% goal. Again, this improvement in coverage rates comes at the cost of power, which drops to about 5%. Finally, and in distinction to the within sample test, we see an important reduction in reversals. Concentrating on the adjusted comparisons, 5% of the cases in which the experimental estimate is statistically significant the observational approach reports a statistically significant result of the opposite sign when naive confidence intervals are used. Our approach removes all of these cases.

Our method performs well in practice, particularly when the observational estimate includes an effort to address selection bias. While the correction still improves inference for unadjusted methods, its real strength appears when applied to estimates that already account for covariates. This supports the idea that the distribution of bias is more stable across studies once observable confounding has been addressed.

[23] LaLonde (1986); Heckman and Hotz (1989); Dehejia and Wahba (1999); Smith and Todd (2005); Friedlander and Robins (1995); Heckman et al. (1998a); Arceneaux et al. (2006); Bléhaut and Rathelot (2014); Agodini and Dynarski (2004); Dehejia and Wahba (2002); Ferraro and Miranda (2014)

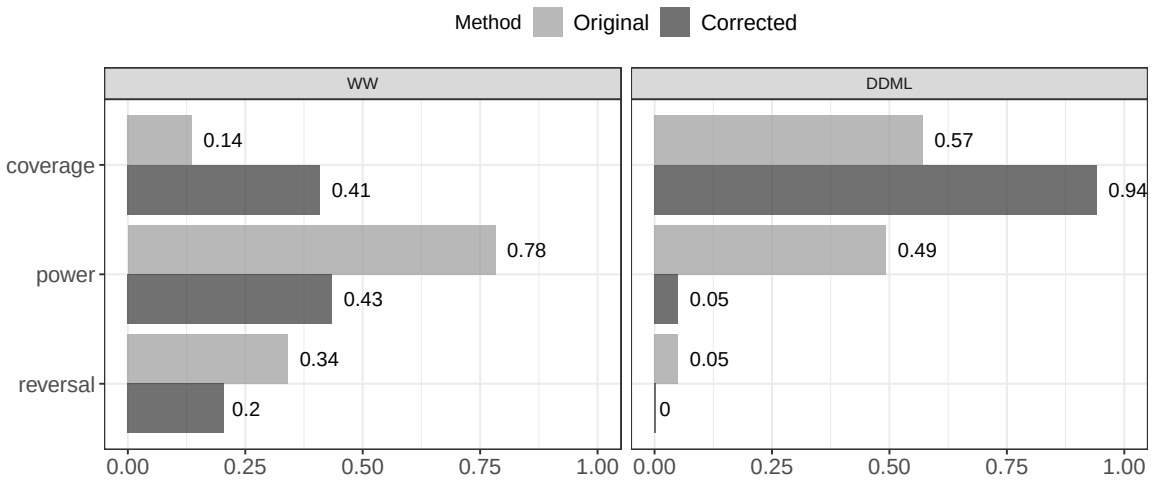


Figure 6: Errors and coverage rates for corrected and uncorrected observational LaLonde-type estimates

Note: This figure presents estimates of the performance of corrected and uncorrected confidence intervals for observational estimators on the sample of eleven LaLonde-style studies described in the text. Corrected observational estimates apply the correction in Equation (1) using parameter estimates $\hat{\mu}$, $\hat{\sigma}_\mu$ and $\hat{\tau}$ from Panel D in Table 1. We measure *coverage*, *power*, and *reversal* as in Figure 5.

5.2 Combining Experimental and Observational Estimates in a Single Meta-Analysis

There is currently no agreed upon approach to combine observational and experimental estimates in a single meta-analysis. Standard meta-analytic approaches report a weighted sum of estimates, where the weight is inversely proportional to the square of the estimated standard error. This creates two problems for an analysis that includes both observational and experimental estimates. First, if observational studies are biased, so will be the meta-analysis. Second, if observational studies have artificially small standard errors, they will receive too much weight. At the same time, throwing out observational estimates is a waste of potentially valuable information. Our proposed approach is to bias-correct the observational estimates, inflate their variances by τ^2, and then combine all estimates in (almost) the standard way. We nevertheless have to carefully account for the estimation error of our bias-correction factor, and for the fact that it gives rise to correlation across the bias-corrected estimates. In Appendix A.3, we introduce a FGLS estimator which combines observational and experimental estimates optimally, and prove its consistency, as well as the consistency of our proposed estimator of its variance.

We illustrate our approach using the meta-analysis from Porat et al. (2024), which studies the effect of interventions designed to prevent sexual violence by targeting attitudes, beliefs, and knowledge. We focus on the most important and policy-relevant outcomes: perpetration (engaging in sexual violence), victimization (being the victim of sexual violence), and bystander intervention (helping a person who is being victimized). We also exclude quasi-experimental studies due to definitional ambiguity and lack of a clear correction strategy. After these restriction

we have 89 estimates of impact: 63 experimental and 26 observational.

Figure 7 presents the results of four meta-analyses on this data set. The most striking result appears for perpetration, arguably the most policy relevant outcome. Looking at experimental estimates alone does not reveal a statistically significant change in perpetration. A policy maker facing this set of results is likely to find them weak. Looking at observational estimates alone suggests a borderline significant impact, but a policy maker may well question whether these estimates are biased. What type of person selects into an anti-sexual violence program? We might reasonably worry that the worst offenders will not take up the program, upwardly biasing these estimates. Naively combining experimental and observational estimates suffers from the same concern, especially because it overweights the observational studies.

Combining results using our correction leads to a clear policy lesson. The resulting estimate is statistically significant: 0.12 ± 0.10 and positive, indicating that interventions can be designed to reduce perpetration. Our correction contributes by pushing the observational estimates upward, and downweighting them relative to the experiments. Notably, even when we set $\hat{\mu} = 0$, isolating the weighting effect, the result is nearly identical.

This example illustrates a key insight: while our method often results in wider intervals and lower power for individual estimates, it can increase power in a meta-analytic context by allowing more honest use of noisy observational data. It offers a way to synthesize diverse evidence without compromising a desire to establish causal effects.[24]

5.3 Choosing Between an RCT and Observational Study

Our methods also give an interesting way to compare between an RCT and observational study. Suppose that a policy maker has the option to assemble data for an observational study, or to conduct an RCT. For simplicity we assume the potential observational study has an infinite sample size and therefore its effective standard error will be 0.207, corresponding to our preferred "aggregate primary" estimate. We ask: how large would the potential RCT need to be to deliver more statistical power (precisely: a smaller expected standard error)?

Figure 8 plots a few scenarios, assuming an individually-randomized RCT with 50% assigned to treatment.[25] With 100% compliance, an experimental sample size of just 93 is sufficient to achieve the same expected standard error on the TOT estimate 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 1487 observations to dominate. Still, $1487 \ll \infty$; more seriously this is still a relatively modest trial when compared to recent published papers in economics. Overall, while conducting RCTs is of course costly, so is the low power that comes

[24] Gechter (2022) proposes a Bayesian method for combining experimental and observational evidence stemming from different studies, unlike in our case. His approach has to correct not only for selection bias in observational studies, but also for potential site-selection bias in RCTs. It requires an instrument for the context in which experiments are conducted. In contrast, our approach is simpler but rests on a stronger assumption of no site-selection bias.

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

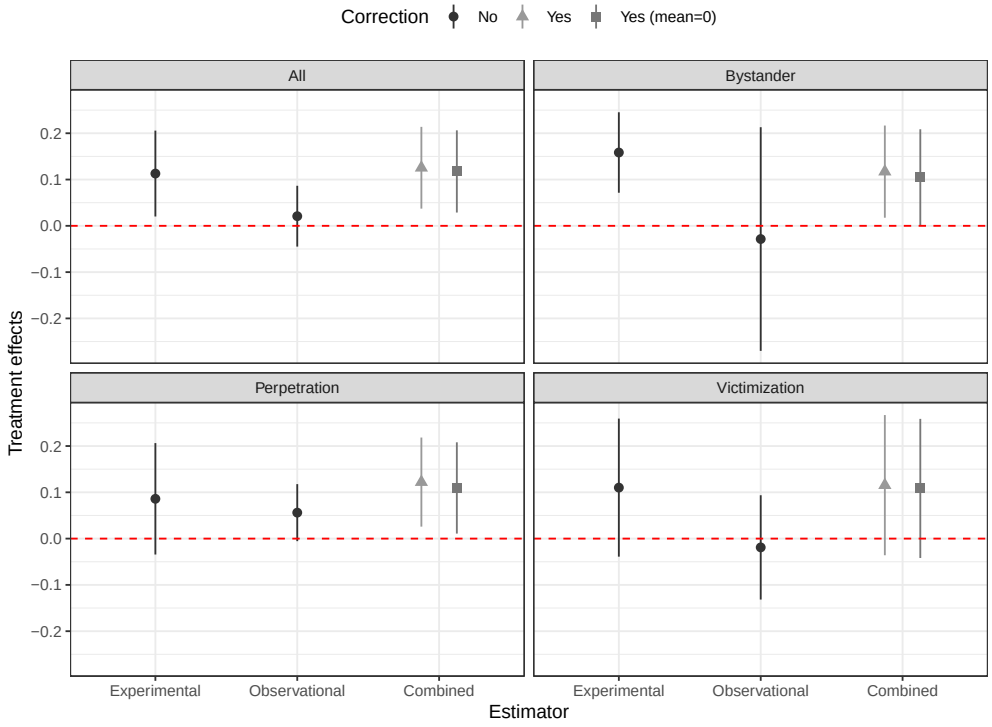


Figure 7: Meta-analysis on sexual violence: corrected meta-analytical estimates

Notes: Data is from Porat et al. (2024), focusing on behavioral outcomes with effects measured either using Experimental or Observational methods. *All* refers to all behavioral outcomes combined, *Bystander* measures whether individuals tend to act when they witness sexual violence perpetrated on others, *Perpetration* measures whether individuals report having perpetrated acts of sexual violence, and *Victimization* measures whether individuals report having been victims of sexual violence. Each point is a meta-analytical estimate (with associated 95% confidence interval based on standard errors clustered at the study level). Meta-analytical estimates combine either all the Experimental estimates (*Experimental*), all the Observational estimates (*Observational*) or both Observational and Experimental estimates (*Combined*). *Correction* indicates whether individual Observational estimates are corrected with our proposed approach before being included in the meta-analysis (*Yes*), or not (*No*). We show results for the *IND_PRI* correction, which corresponds to the sample of outcomes produced by the study. We also show results where we set the mean selection bias estimate $\hat{\mu}$ to zero (*Yes (mean=0)*). *Experimental* and *Observational* estimates are obtained using Random Effects meta-analysis estimated with Restricted Maximum Likelihood, using the same specification as the original authors. The combined estimates are obtained using the FGLS estimator introduced in Appendix A.3.

from uncertainty about bias in observational studies.

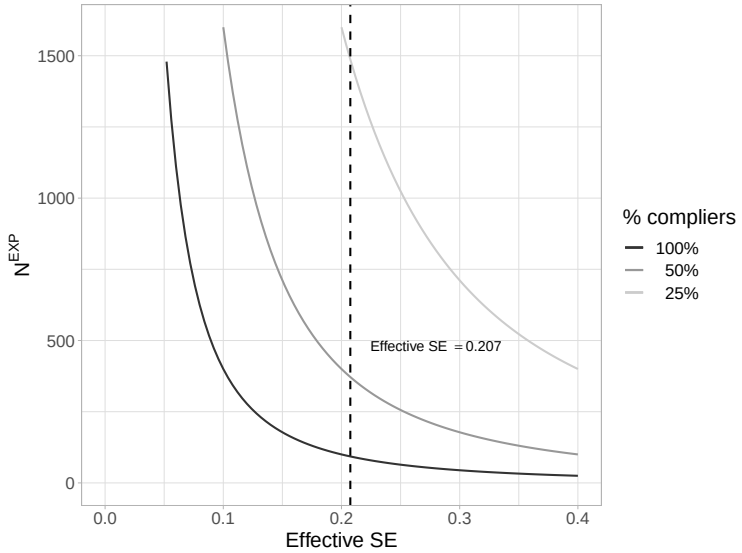


Figure 8: Required experimental sample size to match effective standard error of an infinite-N observational study

6 Robustness

The three use cases above show the value of our approach: confidence intervals are more accurate; meta-analytic precision can be improved; and power comparisons can be used to decide between an RCT and an observational study. These applications, however, all show that our confidence intervals come with a stark power reduction for observational approaches. In this section we test whether we can avoid this power loss by relaxing our two crucial assumptions: exchangeability and Normality.

6.1 Exchangeability Violations due to Predictable Variation in Bias

Our meta-analysis may be mixing together studies with *predictably* heterogeneous biases, artificially inflating estimates of bias variability. For the sake of illustration, suppose we knew that observational estimates were positively biased by 0.5 standard deviations for half our studies, and negatively biased by 0.5 s.d. for the other half. Mixing them together in our meta-analysis, we would erroneously conclude that each study draws its bias from a distribution with mean zero and large variance, and our bias correction would result in wide corrected confidence intervals. If we could predict, ex-ante, which study belonged in which set we could avoid this outcome.[26]

[26]We tried two approaches to this prediction problem, one based on regression trees using a set of observable study characteristics, and the other based on LLMs. We were unable to generate trees (indicating a lack of predictable heterogeneity), so we report here only our LLM based approach.

Economists tend to believe that they can *reason* about bias direction in a given setting, based on theory and contextual knowledge, and referee reports and seminar discussions often center on this type of discussion. If so, we could use expert knowledge to divide studies into predicted-positively and predicted-negatively biased and this should reduce the unexplained variation.

There are, however, reasons to be skeptical. It is often easy to come up with alternative stories that predict the exact opposite direction of selection: do business people take out credit because they are in trouble and unlikely to pay (Stiglitz and Weiss 1981), or because they have high performing projects (De Meza and Webb 1987)?[27] If bias direction is not predictable, then conditioning on expert predictions will neither reduce variability nor improve power.

We use our data to adjudicate between these two views. In principle we could elicit predictions via an expert survey of academics or practitioners à la DellaVigna and Pope (2018). However, our setting is different from the usual application, which asks experts to predict a small number of parameters from a given study. We have over fifty studies, some containing multiple programs, each of which is described in a long paper. The burden on fellow researchers' time of such a survey would be substantial.

Instead, we turn to a novel form of expert, namely a large language model (LLM), OpenAI's GPT 4.1. There are two good reasons to expect a frontier LLM to mimic expert human behavior in predicting bias in our setting. First, frontier LLMs are trained on a vast corpus of text data, including enormous amounts of scientific research (Bubeck et al., 2023), and, presumably, many instances of experts reasoning about observational bias.[28] Indeed, the training data likely include some or all of the papers that we analyze, though they do not have access to our bias estimates since those do not appear in the original papers nor have we reported them at the study level. Second, recent studies have shown empirically that LLMs can match or surpass human experts in predicting neuroscience and behavioral science results, as well as more general prediction tasks (Luo et al., 2025; Lippert et al., 2024; Schoenegger et al., 2024).

To get LLM predictions, we manually coded a precise description of each program in our sample, and a precise description of the population that was eligible to take up that program. We then asked for a binary prediction: would a member of the study population that selected into the program be expected, in general, to have *better* or *worse* outcomes, *in the absence of the intervention*, than those who did not select in. We focus on an "in general" prediction to align with our aggregated measures of bias.

We designed four prompts to elicit LLM bias predictions (see Appendix E.2 for the full text). We always provided detailed descriptions of program and population. We either asked the model to predict selection unconditionally (WW), or after conditioning using DDML on a rich set of observables, and we varied whether we supplied the text of the original research paper or not. Access to the paper might aid prediction, whereas prediction without access was intended to mimic an expert trying to predict bias before deciding whether to conduct a study (with the

[27] The argument is akin to concerns about the "garden of forking paths" that motivate pre-registration of studies.

[28] OpenAI does not disclose the training datasets used (OpenAI, 2023) but it is estimated that GPT-4 (precursor to 4.1) was trained on around 4.9 trillion words of text, see https://epoch.ai/data/ai-models (accessed 17 Oct. 2025)

obvious caveat that the paper may be in the training data!)[29]

We split our sample into predicted-positive and predicted-negative bias studies, then perform a separate 'aggregate primary' meta-analysis for each. Table 2 presents the findings. The results show that the LLM does have some ability to separate studies by bias direction, but it is far from perfect. Predicted-positive biases are indeed larger (more positive) on average than predicted-negative biases. But mean bias is still negative in all subsets, and is never statistically significantly different between pairs of subsets. The LLM seems to overestimate the prevalence of positive biases in general: 23 out of 51 point estimates of bias were negative, but the LLM predicts that the vast majority of studies will be positively biased (44–45 out of 51 when no paper is provided). Providing the research paper causes the LLM to be slightly less likely to predict a positive bias (38–39 out of 51).

Table 2: Meta-analysis Split by Bias Predictions from GPT 4.1

		Predicted negative bias		Predicted positive bias	
	Overall	No paper	With paper	No paper	With paper
Panel A: WW					
Mean ($\hat{\mu}$)	-0.030	-0.097	-0.109	-0.020	-0.003
SE ($\hat{\sigma}_{\mu}$)	(0.044)	(0.088)	(0.073)	(0.051)	(0.053)
Standard deviation ($\hat{\tau}$)	0.251	0.183	0.185	0.268	0.268
Effective SE	0.255	0.203	0.199	0.273	0.273
Num. obs.	51	7	12	44	39
Panel B: DDML					
Mean ($\hat{\mu}$)	-0.025	-0.069	-0.036	-0.020	-0.025
SE ($\hat{\sigma}_{\mu}$)	(0.038)	(0.091)	(0.063)	(0.043)	(0.049)
Standard deviation ($\hat{\tau}$)	0.204	0.174	0.155	0.216	0.235
Effective SE	0.207	0.196	0.167	0.221	0.240
Num. obs.	51	6	13	45	38

Notes: We re-estimate our meta-analysis for aggregate primary outcomes on subsets of the data where GPT 4.1 predicted negative bias ("adverse selection") or positive bias ("advantageous selection"), either unconditional (WW) or conditional on observables (DDML). For "No paper" estimates we provided the LLM with a detailed description of the program and eligible population. For "with paper" estimates we additionally provided the original research paper within the prompt.

[29]Our prompts included two hypothetical examples of reasoning about positive and negative bias, neither of which would match any individual study in our database (an example of "few-shot learning"). Although the mechanics of selection bias are a bit more subtle in encouragement designs, we used the same prompt for eligibility and encouragement design studies. LLMs exhibit inherent randomness in their responses, captured by the "temperature" parameter. We set temperature to 1 (the default) and repeated each prompt 11 times, determining the binary prediction by majority vote. We observed a high degree of concordance between predictions for a given prompt: 99% of predictions were unanimous. We also asked the model to supply a one-sentence rationale for each of its predictions. We found these rationales to be sensible, and similar to how a social scientist might reason about bias. For example: "Takers are more likely to be individuals with greater self-control, higher financial awareness, or stronger intrinsic motivation to save, which means they would have higher counterfactual welfare than non-takers even in the absence of the program." "High-risk men with fewer opportunities and lower counterfactual welfare are more likely to be motivated to take up the program, even after conditioning on observables."

Leveraging LLM predictions does not uniformly reduce the effective standard error. It goes down for the predicted-negative, but up for the predicted-positive set. In the best case scenario for power, the subset of studies that the LLM predicts will have a negative bias after DDML when provided with the paper, the effective standard error drops from 0.207 to 0.167. In other words, a policy maker whose study is also predicted to be negatively biased and chooses to focus on this subset will experience a modest but probably not transformative power gain.

6.2 Exchangeability Violations due to Inclusion of Poor-Quality Bias Estimates

We assume exchangeability across all our studies, but a subset of them may produce noisy or biased estimates of bias. For example, we have already seen that observational studies that make no attempt to control for selection (WW estimates) exhibit more variability in bias. Here we explore whether excluding potentially lower quality studies changes our estimates, and allows for an increase in power for observational methods.

We use several proxies for quality, and the results are presented in Table 3. We discuss each in turn, but the bottom line is that effective standard errors remain large and power remains low even when we concentrate on the best quality studies.

Table 3: Removing Potentially Poor-Quality Bias Estimates

	Full sample	Eligibility	1st stage	# covariates	SUTVA	Exclusion
Panel A: WW						
Mean ($\hat{\mu}$)	-0.030	-0.021	-0.032	-0.038	0.025	-0.036
SE ($\hat{\sigma}_{\mu}$)	(0.044)	(0.051)	(0.054)	(0.057)	(0.078)	(0.051)
Standard deviation ($\hat{\tau}$)	0.251	0.235	0.231	0.231	0.275	0.197
Effective SE	0.255	0.240	0.237	0.238	0.286	0.204
Num. obs.	51	35	25	25	18	25
Panel B: DDML						
Mean ($\hat{\mu}$)	-0.025	-0.035	-0.036	-0.027	0.025	-0.034
SE ($\hat{\sigma}_{\mu}$)	(0.038)	(0.042)	(0.044)	(0.045)	(0.059)	(0.043)
Standard deviation ($\hat{\tau}$)	0.204	0.176	0.178	0.159	0.181	0.159
Effective SE	0.207	0.181	0.184	0.165	0.190	0.164
Num. obs.	51	35	25	25	18	25

Notes: We re-estimate our meta-analysis for aggregate primary outcomes on subsets of the data that might in principle increase power by removing poor-quality estimates of bias. "Full sample" reproduces our primary estimates for comparison. "Eligibility" includes only eligibility-design studies (which identify *TOT*, the policy-maker's parameter of interest). "1st stage" includes only studies with an above-median F-statistic in the experimental first stage (reducing concerns about biased experimental estimates and standard errors). "# covariates" includes only studies where the number of covariates available to the DDML algorithm is above the median in our sample (where we hope to better capture selection on observables). "SUTVA" includes only studies where the experiment was individually randomized (from which we infer the researchers are not concerned about SUTVA violations). "Exclusion" includes only studies where the original authors report a specification that instruments for takeup with assignment (from which we infer they believe the exclusion restriction is satisfied).

6.2.1 Encouragement Versus Eligibility Designs

Our hypothetical policy maker cares about the bias in her estimate of the treatment-on-the-treated (TOT). As we showed in section 2, encouragement-design studies instead recover the bias in the treatment effect on compliers (TOC). If the bias distribution for TOC differs from that for TOT, mixing them could increase variance. Table 3 shows that if we exclude encouragement designs, the bias distribution for eligibility designs exhibits slightly less variation, but not enough to change our qualitative conclusion: the effective SE falls from 0.207 to 0.181.

6.2.2 First Stage

Our experimental estimates use instrumental variables and some studies may have a weak IV problem leading to bias estimates that are biased toward the OLS estimate, with spuriously small standard errors (Staiger and Stock, 1997). That variation then may get misattributed to fundamental variation in bias. Note that we already pruned all studies where the first stage F-stat was less than 10 (a standard cutoff for weak IV), but we further explore the issue by re-estimating our meta-analysis on the sample of studies with above-median first stage F-statistics. This does reduce our estimate of $\hat{\tau}$, but the change is small with the effective standard error falling to 0.184.

6.2.3 Limited Regression Covariates

If a study in our sample contains few or poor-quality covariates, we would not expect DDML to be able to remove selection bias. That might lead us to overestimate the bias faced by a policy maker who could collect rich covariates for their observational study. We therefore re-estimate the meta-analysis for studies with an above-median number of covariates available (see Appendix Table D.1 for descriptives). Unsurprisingly this has minimal effect on the WW bias variance, but does reduce $\hat{\tau}$ after applying DDML, lowering the effective SE by around 20%, to 0.165.

We performed a second exercise, where we restrict attention to regression specifications in which the pre-treatment ("lagged") value of the outcome variable is available as a regression control, i.e. an ANCOVA specification (McKenzie, 2012) (individual-outcome specifications only). Surprisingly, we find that this has effectively no impact on the effective SE.

6.2.4 Failure of Identification

Estimation of TOT or TOC requires identification assumptions that may be violated (Section 2). We concentrate on two. Our experimental estimates will be poorly identified if the exclusion restriction fails. Both experimental and observational estimates will be biased if SUTVA fails.

We can construct indicators for whether the researchers have either of these concerns. First, for a subset of studies the RCT design is clustered, indicating that the researcher was concerned about spillovers and breaches of SUTVA. Excluding these studies has a very small impact on estimated effective standard error (it drops to 0.190). Second, we check whether the original study

authors reported estimates in which they instrumented for takeup using treatment assignment (i.e. estimating TOT/TOC). If not, we infer they were not confident the exclusion restriction held in their experiment. Keeping only the studies where an IV estimate was reported (suggesting more confidence in the exclusion restriction), we find a 20% reduction in the effective standard error, to 0.164.

Overall, we see no evidence that power can be significantly improved by removing lower quality estimates of bias studies.

6.3 Violation of the Normality Assumption

Beyond exchangeability, the main assumption in our meta-analysis is Normality of the observational bias distribution (condition 2). If violated we may draw incorrect conclusions about the mean and variance of bias.

To test this assumption, we estimate the distribution of observational bias nonparametrically. In our data, we observe a combination of true bias and sampling noise: $\hat{B} = B_s + \nu_s$. We can assume that the sampling noise ν_s is normally distributed $\mathcal{N}(0, \sigma^2_{B,s})$ as $B_s = \mu + \eta_s$, and hence is the sum of underlying normally distributed estimators. Extracting the distribution of B leads us to a deconvolution problem with a nonstandard component: while Normal, our error term is heteroskedastic. We use the estimator from Delaigle and Meister (2008) to account for this heteroskedasticity. The estimator is coded up by Wang and Wang (2011) and we use optimal bandwidth selection via bootstrap and a rule-of-thumb starting value to find the optimal bandwidth (specifically our starting value is $S^{-1/5} * \sqrt{\hat{\sigma}^2_\mu + \hat{\tau}^2}$).

Figure 9 plots the nonparametrically estimated density together with the Normal distribution with average bias and variance from our meta-analytic results.

Using all individual outcomes, we obtain relatively overlapping distributions and, if anything, the variance of the parametrically estimated Normal distribution appears to be smaller. For the aggregated outcomes, we notice that the Normal distribution is much tighter than the nonparametrically estimated distribution, with the caveat that this is based on just 51 datapoints. In both cases, our non-parametric distribution appears reasonably symmetric and bell-shaped.

We test more formally the shape of the deconvolved estimators by estimating their skewness and kurtosis. Table 4 shows estimates of these parameters, obtained using the composite Simpson's rule for irregularly spaced data, after a change of variable to project the distribution from $\mathbb{R}$ to $[0, 1]$. We show the results for the whole support of the distribution as estimated by the deconvolution procedure and with the support restricted to $[-2, 2]$ to avoid giving too much influence to estimation errors far away in the tails. After filtering, most of the deconvolved distributions have skewness and kurtosis quite close to 0 and 3, the values corresponding to a Normal distribution. In sum, we conclude that our parametric Normality assumption appears reasonable, and that relaxing it would only lead us to conclude that uncertainty about bias is even greater.

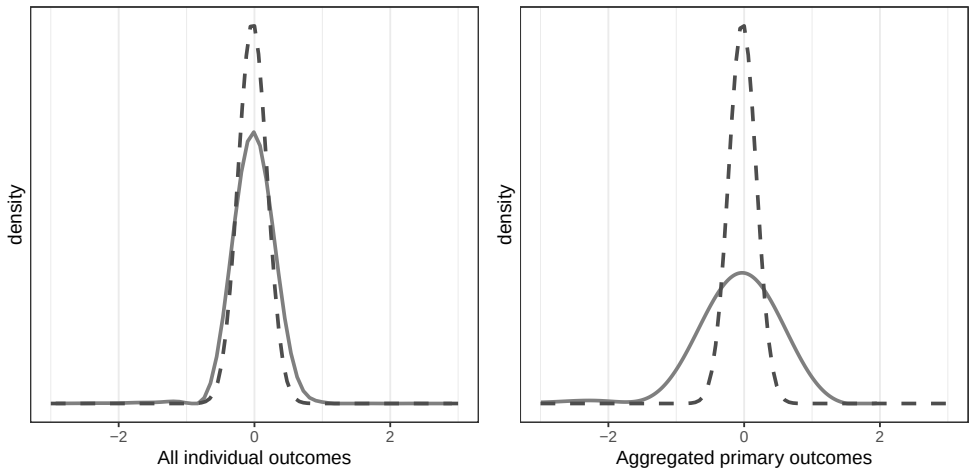


Figure 9: Nonparametric estimation of the bias density for the DDML estimator.

Note: This Figure shows nonparametric estimates of the distribution of the bias of observational estimators. The left panel uses all individual outcomes. The right panel uses the aggregated primary outcomes. The solid line represents the nonparametrically estimated density via Delaigle and Meister (2008)'s heteroskedastic deconvolution estimator. The dashed line represents the Normal density with average observational bias $\hat{\mu}$ and standard error $\hat{\tau}$ from the meta analytic results.

6.4 Other Assumptions

6.4.1 Incorrect adjustment of multi-outcome studies.

Our "aggregate" specifications include just one bias estimate per study. However, our "individual" specifications estimates from multiple outcomes and/or programs per study. To adjust for intra-study correlation in outcomes, we calibrate an intra-cluster correlation parameter which we set to 0.6, following existing literature (see Section 2.5). If our estimates are sensitive to this parameter, we might overestimate (or underestimate) τ. We experiment with adjusting the parameter to 0.5 or 0.7 in Appendix table D.3; this does not meaningfully change the results.

6.4.2 Incorrect standard error calculations.

When computing the standard errors of our bias estimates we treat the experimental and observational estimates as independent. In reality they are likely to be correlated, in which case our standard errors will be biased. We discuss this point further in Appendix B.2, and demonstrate using a bootstrap approach that this does not matter in practice.

6.4.3 Site Selection Bias and Publication Bias

Our exchangeability assumption requires that the policy maker's observational study of interest is drawn from the same population as the set of studies we use for estimation; this would fail if there is selection into conducting an RCT (site selection bias) or publishing the data (publication bias). Appendix E.3 discusses how these might influence our estimates.

Table 4: Estimated moments of the deconvolved distribution of selection bias

Filtered	Mean	Sd	Skewness	Kurtosis
Panel A: Aggregated primary outcomes				
N	-0.10	0.68	-1.85	14.38
Y	-0.05	0.55	-0.12	2.83
Panel B: Aggregated all outcomes				
N	-0.03	0.57	-0.96	5.76
Y	0.00	0.50	-0.27	3.21
Panel C: Individual primary outcomes				
N	0.00	0.43	-0.45	5.36
Y	0.01	0.42	-0.12	3.40
Panel D: Individual outcomes				
N	-0.01	0.44	24.49	1637.19
Y	-0.01	0.31	-0.37	5.06

Notes: This table presents estimates of features of the deconvolved density distributions of selection bias. All estimates are computed using the composite Simpson's rule for irregularly spaced data, after a change of variable to project the distribution from $\mathbb{R}$ to $[0, 1]$. We show the results for the whole support of the distribution as estimated by the deconvolution procedure (*Filtered*=*N*) and with the support restricted to $[-2, 2]$ to avoid excessive influences from estimation errors in the tails (*Filtered*=*Y*). *Mean* reports the value of the first moment. *Sd* reports the value of the standard deviation, computed as the square root of the second centered moment. *Skewness* reports the value of the skewness, computed as the third standardized moment. *Kurtosis* reports the value of the kurtosis, computed as the fourth standardized moment.

7 Conclusion

Observational studies are likely to remain a mainstay of program evaluation. We study the bias in these studies, with an emphasis on quantifying uncertainty, which is often treated as having unknown size and magnitude. We show that we can construct corrected observational confidence intervals with good nominal coverage, which appropriately account for uncertainty about bias. We find substantial uncertainty, which implies that observational studies have low power to detect program impacts of a policy-relevant size: even an infinite-N observational study has an effective standard error equal to around 0.2 standard deviations of the outcome variable.

It is reasonable to question whether the heterogeneous sample of studies that underlie our estimates can provide generalizable evidence about the distribution of bias. But, we show that our correction has close to nominal coverage both within our sample (using a leave-one-out approach), and out of sample, applied to a large number of estimates from LaLonde-style studies in very different settings. The correction performs best when applied to point estimates that have already been adjusted for selection on observables, supporting a notion of "conditional exchangeability" whereby the predictable component of bias can differ a lot between settings

while the unpredictable component follows a common distribution.

Our correction is useful for three primary exercises: it can be used to bias correct individual studies, generating honest confidence intervals that include appropriate uncertainty about bias. It can be used to bias correct and increase the precision of collections of studies, combining observational and experimental estimates in a single meta-analysis. And it can be used in ex-ante power calculations to inform the decision to conduct an RCT in a new setting.

A couple of auxiliary findings stand out to us. We provide useful evidence that Double-Debiased Machine Learning performs better than a leading alternative (Post Double Selection LASSO), which sometimes decreased precision relative to a regression without covariates. And we provide a novel test case for using machine experts (LLMs) as stand-ins for human experts in a bias-prediction exercise (contrasting the more standard case of predicting headline findings). The LLM behaved like a human expert: it provided coherent predictions, alongside persuasive rationales, and to a modest degree separated positive and negative biases (albeit substantially overestimating the prevalence of positive biases). Nevertheless this did not lead to any meaningful improvement in the statistical power of bias-corrected estimates.

Future research can build on our work in a couple of ways. In principle there might emerge better ways to predict bias in our sample, for instance perhaps an even more-powerful AI will do better at predicting bias, though there is a risk of overfitting (with enough exploration we could certainly find a subset of point estimates that yield smaller $\hat{\sigma}_\mu$ and $\hat{\tau}$s). Our sample of imperfect compliance RCT datasets can also grow in future, and a sufficiently large sample might open the door to estimating richer bias-prediction models. More generally, we have demonstrated a disciplined, hands-off approach to assessing bias of different methodologies that could be applied in new settings. For instance, to quasi-experimental studies in which identification of the non-RCT results hinges on a different set of identification assumptions than in observational studies.

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.

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.

Andrews, I. and M. Kasy (2019): "Identification of and Correction for Publication Bias," *American Economic Review*, 109, 2766–2794.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *The Quarterly Journal of Economics*, 132, 871–919.

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.

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].

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.

Bléhaut, M. and R. Rathelot (2014): "Expérimentation contrôlée contre appariement : le cas d'un dispositif d'accompagnement de jeunes diplômés demandeurs d'emploi," *Economie & Prévision*, 204-205, 163–181.

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

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang (2023): "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," *arXiv preprint arXiv:2303.12712*.

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.

De Meza, D. and D. C. Webb (1987): "Too much investment: A problem of asymmetric information," *The Quarterly Journal of Economics*, 102, 281–292.

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.

Delaigle, A. and A. Meister (2008): "Density estimation with heteroscedastic error," *Bernoulli*, 14, 562 – 579.

DellaVigna, S. and D. Pope (2018): "Predicting experimental results: who knows what?" *Journal of Political Economy*, 126, 2410–2456.

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

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.

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.

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

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.

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.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

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 (1998a): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1099.

Heckman, J. J., H. Ichimura, and P. E. Todd (1998b): "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies*, 65, 261–294.

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.

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

Lippert, S., A. Dreber, M. Johannesson, W. Tierney, W. Cyrus-Lai, E. L. Uhlmann, Emotion Expression Collaboration, and T. Pfeiffer (2024): "Can large language models help predict results from a complex behavioural science study?" *Royal Society Open Science*, 11, 240682.

Luo, X., A. Rechardt, G. Sun, K. K. Nejad, F. Yáñez, B. Yilmaz, K. Lee, A. O. Cohen, V. Borghesani, A. Pashkov, et al. (2025): "Large language models surpass human experts in predicting neuroscience results," *Nature Human Behaviour*, 9, 305–315.

Mammen, E., C. Rothe, and M. Schienle (2016): "Semiparametric Estimation with Generated Covariates," *Econometric Theory*, 32, 1140–1177.

McKenzie, D. (2012): "Beyond baseline and follow-up: The case for more T in experiments," *Journal of Development Economics*, 99, 210–221.

Menzel, K. (2024): "Transfer Estimates for Causal Effects across Heterogeneous Sites," *arXiv:2305.01435*.

Newey, W. K. and D. McFadden (1994): "Chapter 36 Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Elsevier, vol. 4, 2111–2245.

OpenAI (2023): "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*.

Porat, R., A. Gantman, S. A. Green, J.-H. Pezzuto, and E. L. Paluck (2024): "Preventing Sexual Violence: A Behavioral Problem Without a Behaviorally Informed Solution," *Psychological Science in the Public Interest*, 25, 4–29.

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

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.

Schoenegger, P., I. Tuminauskaite, P. S. Park, and P. E. Tetlock (2024): "Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy," *arXiv preprint arXiv:2402.19379*.

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

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stiglitz, J. E. and A. Weiss (1981): "Credit rationing in markets with imperfect information," *The American Economic Review*, 71, 393–410.

Viechtbauer, W. (2021): "A general workflow for complex meta-analyses with dependent effect sizes," URL: https://www.wvbauer.com/doku.php/presentations (02/12/2025).

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

Wang, X.-F. and B. Wang (2011): "Deconvolution Estimation in Measurement Error Models: The R Package decon," *Journal of Statistical Software*, 39, 1–24.

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.

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

Online Appendix (Not for Publication)

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

A Theoretical results

Here we formally prove the main theoretical results in the text. We show how to estimate the bias of observational estimators using ICRCTs, prove asymptotic validity of our bias-adjusted confidence intervals, and prove consistency of the Feasible Generalized Least Squares (FGLS) estimator we propose to combine experimental and observational estimates in a single meta-analysis.

A.1 Identifying Observational Bias using ICRCTs

In this appendix we show how to identify 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*[30]

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:* $(Y_i^1, Y_i^0) \perp D_i | X_i, R_i = r, \forall r \in \{0, 1\}$.

[30] 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.

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

A.1.1 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.

Under the classical assumptions SUTVA, Independence, Exclusion, First Stage and Monotonicity, the experimental Wald estimand

$$TOC^{EXP} = \frac{E\left[Y_i|R_i = 1\right] - E\left[Y_i|R_i = 0\right]}{P(D_i = 1|R_i = 1) - P(D_i = 1|R_i = 0)} \tag{6}$$

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 estimand, 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 TOT^{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

$$TOC_X^{OBS} = \frac{TOT_X^{OBS,1} \Pr(D_i = 1|R_i = 1) - TOT_X^{OBS,0} \Pr(D_i = 1|R_i = 0)}{\Pr(D_i = 1|R_i = 1) - \Pr(D_i = 1|R_i = 0)}. \tag{7}$$

Theorem 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 from the fact that $D_i^1 = D_i^0 = 1$ imply $D_i = 1$ and the third 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 6 is:

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

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. ■

Theorem 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 TOC:

$$\begin{aligned} TOC^{OBS} - TOC^{EXP} = SBC \\ TOC_X^{OBS} - TOC^{EXP} = BC_X. \end{aligned}$$

Where SBC stands for selection bias on compliers and BC_X stands for observational bias on compliers after covariate adjustment. In section 2 we refer to these term simply as B.

A.1.2 Eligibility Designs

In an eligibility design, the control group are prevented from participating.[31] We can form an experimental estimand TOT^{EXP} based on Equation 6 with $P(D_i = 1|R_i = 0) = 0$ and a single observational estimand on the treatment arm $TOT^{OBS} = TOT^{OBS,1}$. 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:

$$\begin{aligned} TOT^{OBS,1} - TOT^{EXP} = SBT \\ TOT_X^{OBS,1} - TOT^{EXP} = BT_X. \end{aligned}$$

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 2 we refer to these terms simply as B.

[31] 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.

A.2 Asymptotic Validity of Bias-Adjusted Confidence Intervals

We consider a policy-maker interested in the treatment effect on the treated, TOT_p, for a new study p. The policy-maker observes an observational estimator $\widehat{TOT}_p^{OBS}$ derived from a sample of size n_p, with estimated sampling noise $\hat{\sigma}^2_{e,p}$. To construct a confidence interval, we rely on historical data from S independent studies to estimate the distribution of the observational bias.

Assumption 3 (Asymptotic Normality of Observational Estimator) *For the policy-maker's study p, there exists a finite constant $V_{e,p} > 0$ such that, as $n_p \to \infty$,*

$$\sqrt{n_p}\left(\widehat{TOT}_p^{OBS} - TOT_p^{OBS}\right) \xrightarrow{d} \mathcal{N}(0, V_{e,p}),$$

where TOT_p^{OBS} is the probability limit of the estimator (the observational estimand). We also assume that there is a consistent estimator $\hat{\sigma}^2_{e,p}$ such that $n_p\hat{\sigma}^2_{e,p} \xrightarrow{p} V_{e,p}$.

Assumption 3 follows from the $\sqrt{N}$-consistency and asymptotic normality of observational estimators (Hahn, 1998; Heckman et al., 1998b; Mammen et al., 2016; Chernozhukov et al., 2018).

Assumption 4 (Consistency of Bias Distribution Parameters) *Let $(\hat{\mu}, \hat{\tau}^2)$ be estimators of the mean and variance of the observational bias derived from S historical studies. There exists a finite constant $V_\mu > 0$ such that, as $S \to \infty$,*

$$\sqrt{S}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, V_\mu).$$

We also assume that we have estimators $\hat{\sigma}^2_\mu$ and $\hat{\tau}^2$ such that $S\hat{\sigma}^2_\mu \xrightarrow{p} V_\mu$ and $\hat{\tau}^2 \xrightarrow{p} \tau^2$. Finally, we assume that $\tau^2 > 0$.

Assumption 4 posits the $\sqrt{N}$-consistency and asymptotic normality of random-effects meta-analytic estimators. This follows from recasting these estimators as M-estimators and using classical results (Newey and McFadden, 1994). The last part of Assumption 4 eschews the cases where τ^2 is at a boundary. It is fully satisfied in our applications.

Assumption 5 (Exchangeability) *The unobserved bias of the policy-maker's study, $B_p = TOT_p^{OBS} - TOT_p$, is exchangeable with the biases of the historical studies. Specifically, B_p is drawn from the same distribution:*

$$B_p = \mu + \eta_p, \quad \eta_p \sim \mathcal{N}(0, \tau^2).$$

Assumption 6 (Independence) *The policy-maker's study p is independent of the set of historical studies used to estimate $\hat{\mu}$ and $\hat{\tau}^2$. Consequently, the sampling error of the study ($\widehat{TOT}_p^{OBS} - TOT_p^{OBS}$), the specific bias realization (η_p), and the estimation error of the mean bias ($\hat{\mu} - \mu$) are mutually independent.*

Theorem 2 (Asymptotic Validity) *Under Assumptions 3 through 6, the bias-adjusted estimator $\widehat{TOT}_p^{OBS} - \hat{\mu}$ satisfies:*

$$\frac{\widehat{TOT}_p^{OBS} - \hat{\mu} - TOT_p}{\sqrt{\hat{\sigma}^2_{\epsilon,p} + \hat{\tau}^2 + \hat{\sigma}^2_\mu}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n_p \to \infty$ and $S \to \infty$. Consequently, the confidence interval

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

has asymptotic coverage probability δ.

Proof: Decompose the total estimation error as

$$\widehat{TOT}_p^{OBS} - \hat{\mu} - TOT_p = \left(\widehat{TOT}_p^{OBS} - TOT_p^{OBS}\right) + (B_p - \mu) - (\hat{\mu} - \mu).$$

By Assumption 5, $B_p - \mu = \eta_p \sim N(0, \tau^2)$. By Assumption 3, $\widehat{TOT}_p^{OBS} - TOT_p^{OBS}$ is asymptotically normal with variance $\sigma^2_{e,p} = V_{e,p}/n_p$. By Assumption 4, $\hat{\mu} - \mu$ is asymptotically normal with variance $\sigma^2_{\mu} = V_{\mu}/S$. By Assumption 6, these three components are mutually independent. Therefore their sum is asymptotically normal with mean 0 and variance

$$V_{tot} = \sigma^2_{e,p} + \tau^2 + \sigma^2_{\mu} = \frac{V_{e,p}}{n_p} + \tau^2 + \frac{V_{\mu}}{S}.$$

Let $\hat{V}_{tot} := \hat{\sigma}^2_{e,p} + \hat{\tau}^2 + \hat{\sigma}^2_{\mu}$. By the consistency assumptions in Assumptions 3 and 4, $\hat{V}_{tot} \xrightarrow{p} V_{tot}$. Slutsky's theorem then implies that as $n_p \to \infty$ and $S \to \infty$

$$\frac{\widehat{TOT}_p^{OBS} - \hat{\mu} - TOT_p}{\sqrt{\hat{V}_{tot}}} \xrightarrow{d} \mathcal{N}(0,1).$$

The stated asymptotic coverage of CI_δ follows immediately. ■

A.3 A meta-analytic estimator combining experimental and observational estimates

We now consider a meta-analysis combining K_E experimental studies and K_O observational studies, with $K = K_E + K_O$ the total number of studies. Let $\widehat{TOT}_k$, $k \in \{1, \ldots, K\} = \mathcal{K}$ denote a set of estimates of the effect of an intervention from K independent studies, with $\mathcal{K}_{EXP} \subset \mathcal{K}$ and $\mathcal{K}_{OBS} = \overline{\mathcal{K}_{EXP}} \subset \mathcal{K}$ the corresponding sets of indices of experimental and observational estimates.

We first derive general conditions for a FGLS estimator to be consistent, along with a consistent estimator of its asymptotic variance. We then check that these conditions are met with our bias corrected estimator, derive closed forms for the weights, and provide consistent estimators for the variance components.

A.3.1 A consistent FGLS meta-analytical estimator

Let $\hat{\mathbf{y}}$ be the full vector of estimates and $\mathbf{V}(\psi_0)$ be their covariance matrix, which depends on variance parameters $\psi_0 = (t^2, \tau^2, \sigma^2_{\mu})$.

Assumption 7 (Linear Model) *The data generating process is* $\hat{\mathbf{y}} = TOT\mathbf{1} + \mathbf{u}$, *where TOT is the true average treatment effect,* $\mathbf{1}$ *is a vector of ones, and* $\mathbf{u}$ *is a zero-mean random vector with covariance* $\mathbf{V}(\psi_0)$.

Assumption 8 (Asymptotics) *The number of studies grows such that* $K \to \infty$, *with the proportion of observational studies* $K_O/K \to \rho \in [0, 1[$.

Assumption 9 (Consistent Variance Estimation) *There exists an estimator* $\hat{\psi}$ *such that* $\|\hat{\psi} - \psi_0\| \xrightarrow{p} 0$. *Furthermore, the estimated covariance matrix* $\widehat{\mathbf{V}} = \mathbf{V}(\hat{\psi})$ *satisfies* $\|\widehat{\mathbf{V}}^{-1} - \mathbf{V}(\psi_0)^{-1}\| \xrightarrow{p} 0$ *in operator norm.*

Assumption 10 (Non-degeneracy) *The information quantity* $A_K = \mathbf{1}'\mathbf{V}(\psi_0)^{-1}\mathbf{1}$ *satisfies* $A_K \to \infty$ *as* $K \to \infty$. *This ensures the variance of the optimal GLS estimator vanishes asymptotically.*

Theorem 3 (Consistency of FGLS) *Under Assumptions 7 through 10, the GLS estimator* $\widehat{TOT}_{GLS} = \left(\mathbf{1}'V(\psi_0)^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'V(\psi_0)^{-1}\hat{\mathbf{y}}$ *is the best linear unbiased estimator (BLUE). The Feasible GLS estimator defined as:*

$$\widehat{TOT}_{FGLS} = (\mathbf{1}'\widehat{\mathbf{V}}^{-1}\mathbf{1})^{-1}\mathbf{1}'\widehat{\mathbf{V}}^{-1}\hat{\mathbf{y}}$$

is consistent for TOT ($\widehat{TOT}_{FGLS} \xrightarrow{p} TOT$*). Moreover, the estimated variance* $(\mathbf{1}'\widehat{\mathbf{V}}^{-1}\mathbf{1})^{-1}$ *is consistent for the true asymptotic variance.*

Proof: By the Gauss-Markov theorem for the general linear model, the GLS estimator is BLUE. Combining with Assumption 10, we have that $\text{Var}(\widehat{TOT}_{GLS}) = A_K^{-1} \to 0$, implying $\widehat{TOT}_{GLS} \xrightarrow{p} TOT$. The difference between FGLS and GLS depends on the convergence of the weights. Let $\widehat{W} = \mathbf{1}'\widehat{\mathbf{V}}^{-1}\mathbf{1}$ and $W = \mathbf{1}'\mathbf{V}(\psi_0)^{-1}\mathbf{1}$. By Assumption 9, $\left|\frac{\widehat{W}}{W} - 1\right| = o_p(1)$, which also proves the consistency of the asymptotic variance. Similarly, the weighted sum of outcomes converges. Using the decomposition $\widehat{TOT}_{FGLS} = TOT + (\mathbf{1}'\widehat{\mathbf{V}}^{-1}\mathbf{1})^{-1}\mathbf{1}'\widehat{\mathbf{V}}^{-1}\mathbf{u}$, and applying the convergence rates of the quadratic forms, we obtain that the estimation error vanishes. ■

A.3.2 Specification of the combined meta-analysis

We specify the model and covariance structure in our proposed meta-analysis combining experimental estimates with bias-corrected observational estimates. For each experimental study $k \in \mathcal{K}_{EXP}$, the estimator $\widehat{TOT}_k^{EXP}$ is modeled as:

$$\widehat{TOT}_k^{EXP} = TOT + \kappa_k + e_k,$$

where $\kappa_k \sim \mathcal{N}(0, t^2)$ represents between-study heterogeneity in treatment effects, and $e_k \sim \mathcal{N}(0, \sigma_{e,k}^2)$ represents sampling error. For each observational study $k \in \mathcal{K}_{OBS}$, we apply the bias correction: $\widehat{TOT}_k^{OBS,c} = \widehat{TOT}_k^{OBS} - \hat{\mu}$. The model for the corrected observational estimator is therefore:

$$\widehat{TOT}_k^{OBS,c} = TOT + \kappa_k + (\eta_k - \mu) - (\hat{\mu} - \mu) + e_k,$$

where $\kappa_k \sim \mathcal{N}(0, t^2)$ is the treatment effect heterogeneity (shared with experimental studies); $\eta_k \sim \mathcal{N}(0, \tau^2)$ is the random component of the observational bias; $(\hat{\mu} - \mu) \sim \mathcal{N}(0, \sigma_\mu^2)$ is the estimation error of the mean bias correction; $e_k \sim \mathcal{N}(0, \sigma_{e,k}^2)$ is the sampling error. This implies that Assumption 7 is verified. Note that Assumption 7 requires that the set of observational studies in the meta-analysis $\mathcal{K}_{OBS}$ is exchangeable with the set of studies in the bias-correction sample, a generalization of Assumption 5. If there was a systematic bias-component specific to $\mathcal{K}_{OBS}$ and different from μ, Assumption 7 would fail and our bias-corrected meta-analytic estimator would be both biased and inconsistent.

The covariance matrix $\mathbf{V}$ of the vector of estimates $\hat{\mathbf{y}}$ is block-diagonal:

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{EE} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{OO} \end{pmatrix}.$$

The experimental block is diagonal, $\mathbf{V}_{EE} = \text{diag}(t^2 + \sigma_{e,k}^2)$. The observational block combines a diagonal block with a common term due to the shared variance σ_μ^2: $\mathbf{V}_{OO} = \mathbf{D} + \sigma_\mu^2 \mathbf{1}\mathbf{1}'$, where $\mathbf{D} = \text{diag}(t^2 + \tau^2 + \sigma_{e,k}^2)_{k \in \mathcal{K}_{OBS}}$. Using the fact that the inverse of a block diagonal matrix is the block diagonal matrix of the inverses of each block, and leveraging the Sherman-Morrison formula to invert $\mathbf{V}_{OO}$, we derive the normalized GLS weights w_k:

$$w_k \propto \begin{cases} \frac{1}{t^2 + \sigma_k^2} & \text{for } k \in \mathcal{K}_{EXP} \\ \frac{1}{d_k(1 + \sigma_\mu^2 S_O)} & \text{for } k \in \mathcal{K}_{OBS}, \end{cases}$$

where $d_k = t^2 + \tau^2 + \sigma_{e,k}^2$ and $S_O = \sum_{j \in \mathcal{K}_{OBS}} \frac{1}{d_j}$. The variance of the pooled estimator is therefore:

$$\text{Var}(\widehat{TOT}_{GLS}) = \left(\sum_{k \in \mathcal{K}_{EXP}} \frac{1}{t^2 + \sigma_{e,k}^2} + \frac{S_O}{1 + \sigma_\mu^2 S_O} \right)^{-1}.$$

This variance term verifies Assumption 10: $\sum_{k \in \mathcal{K}_{EXP}} \frac{1}{t^2 + \sigma_{e,k}^2}$ tends to infinity as $K \to \infty$ (since $\rho < 1$, $K_{EXP} \to \infty$ as well) while $\frac{S_O}{1 + \sigma_\mu^2 S_O} \to \frac{1}{\sigma_\mu^2}$. Note that a meta-analysis including only observational estimates would be inconsistent (even if unbiased): its asymptotic variance would always be bounded below by σ_μ^2. The confidence intervals based on our FGLS estimator would still have correct coverage in that case, though.

A.3.3 Estimation of each variance component

To compute the FGLS estimator, we simply need consistent estimates of t^2, τ^2 and σ_μ^2. We take $\hat{\tau}^2$ and $\hat{\sigma}_\mu^2$ from the REML estimator in our sample of bias estimates. We use a method-of-moments estimator to recover treatment effect heterogeneity t^2. Let $\hat{\sigma}_{tot}^2$ be the sample variance of all the

uncorrected $\widehat{TOT}_k$ estimates. We define the average within-study variance $\hat{\bar{\sigma}}^2$ as:

$$\hat{\bar{\sigma}}_e^2 = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_{e,k}^2.$$

Our estimator is $\hat{t}^2 = \max(0, \hat{\sigma}_{tot}^2 - \hat{\bar{\sigma}}_e^2 - \frac{K^{OBS}}{K}\hat{\tau}^2)$. We have $\mathbb{E}[\hat{\sigma}_{tot}^2] = \frac{1}{K}\sum_{k=1}^{K}\text{Var}(\widehat{TOT}_k) = t^2 + \frac{K^{OBS}}{K}\tau^2 + \bar{\sigma}^2$. The consistency of our estimator follows from the consistency of each individual estimator of the variance components.

B 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:

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

  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 $\tilde{D}_i$: 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.1).

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 $\hat{\sigma}_{OBS,os}$.[32] 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 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 (7), 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

[32] 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.

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.1 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×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 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.[33]

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). 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 C.1 shows how many studies pass each stage of screening.

[33]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.

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. During this analysis we further clean the data as described in the next section which leads us to lose some datasets prepared for analysis.

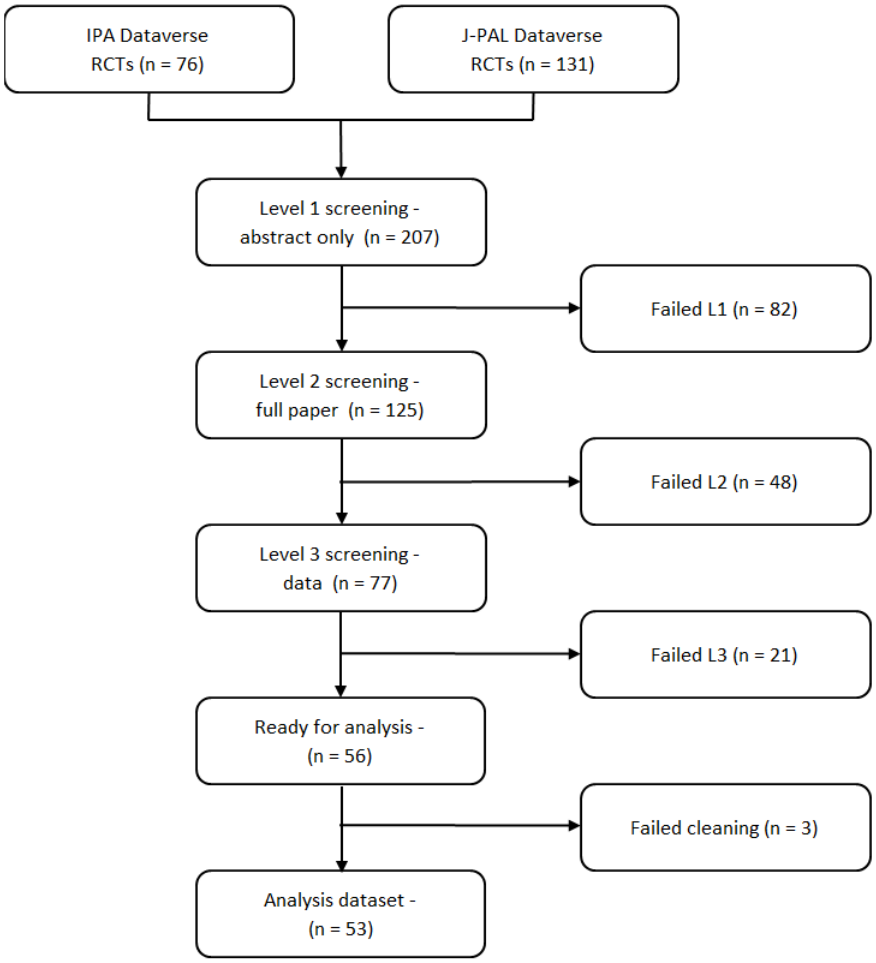


Figure C.1: Flow diagram of studies passing through our selection process

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.

C.3 Studies included in the meta-analysis (in alphabetical order by author)

- Ambler, K., Aycinena, D., and Yang, D. (2015): "Channeling remittances to education: A field experiment among migrants from El Salvador," American Economic Journal: Applied Economics, 7(2), 207–32.
- Angelucci, M., Karlan, D., and Zinman, J. (2015): "Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco," American Economic Journal: Applied Economics, 7(1), 151–82.
- Ashraf, N., Karlan, D., and Yin, W. (2006): "Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines," The Quarterly Journal of Economics, 121(2), 635–672.
- Ashraf, N., Giné, X., and Karlan, D. (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(4), 973–990.
- Atkin, D., Khandelwal, A. K., and Osman, A. (2017): "Exporting and Firm Performance: Evidence from a Randomized Experiment," The Quarterly Journal of Economics, 132(2), 551–615.
- Baldwin, K., Karlan, D., Udry, C., and Appiah, E. (2016): "Does community-based development empower citizens? Evidence from a randomized evaluation in Ghana," Working paper.
- Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., and Khemani, S. (2010): "Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India," American Economic Journal: Economic Policy, 2(1), 1–30.
- Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007): "Remedying education: Evidence from two randomized experiments in India," The Quarterly Journal of Economics, 122(3), 1235–1264.
- Banerjee, A., Duflo, E., Glennerster, R., and Kinnan, C. (2015): "The Miracle of Microfinance? Evidence from a Randomized Evaluation," American Economic Journal: Applied Economics, 7(1), 22–53.
- Banerjee, A., Duflo, E., Chattopadhyay, R., and Shapiro, J. (2016): "The long term impacts of a "Graduation" program: Evidence from West Bengal," Unpublished paper, MIT.
- Banerji, R., Berry, J., and Shotland, M. (2017): "The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India," American Economic Journal: Applied Economics, 9(4), 303–37.
- Beaman, L., Karlan, D., Thuysbaert, B., and Udry, C. (2013): "Profitability of fertilizer: Experimental evidence from female rice farmers in Mali," American Economic Review, 103(3), 381–86.
- Behaghel, L. and De Chaisemartin, C. (2017): "Ready for boarding? The effects of a boarding school for disadvantaged students," American Economic Journal: Applied Economics, 9(1), 140–164.
- Benhassine, N., Devoto, F., Duflo, E., Dupas, P., and Pouliquen, V. (2015): "Turning a Shove into a Nudge? A "Labeled Cash Transfer" for Education," American Economic Journal: Economic Policy, 7(3), 86–125.
- Blattman, C., Fiala, N., and Martinez, S. (2014): "Generating skilled self-employment in developing countries: Experimental evidence from Uganda," The Quarterly Journal of Economics, 129(2), 697–752.
- Blattman, C. and Annan, J. (2016): "Can employment reduce lawlessness and rebellion? A field experiment with high-risk men in a fragile state," American Political Science Review, 110(1), 1–17.
- Blattman, C., Jamison, J. C., and Sheridan, M. (2017): "Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia," American Economic Review, 107(4), 1165–1206.

- Blattman, C., Fiala, N., and Martinez, S. (2020): "The long-term impacts of grants on poverty: Nine-year evidence from Uganda's youth opportunities program," American Economic Review: Insights, 2(3), 287–304.
- Blattman, C., Hartman, A. C., and Blair, R. A. (2014): "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(1), 100–120.
- Blattman, C. and Dercon, S. (2018): "The impacts of industrial and entrepreneurial work on income and health: Experimental evidence from Ethiopia," American Economic Journal: Applied Economics, 10(3), 1–38.
- Bloom, N., Liang, J., Roberts, J., and Ying, Z. J. (2015): "Does working from home work? Evidence from a Chinese experiment," The Quarterly Journal of Economics, 130(1), 165–218.
- Braconnier, C. (2017): "Voter registration costs and disenfranchisement: Experimental evidence from France," American Political Science Review, 111(3), 584–604.
- Bruhn, M., Karlan, D., and Schoar, A. (2018): "The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in Mexico," Journal of Political Economy, 126(2), 635–687.
- Bryan, G., Chowdhury, S., and Mobarak, A. M. (2014): "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh," Econometrica, 82(5), 1671–1748.
- Bryan, G., Choi, J. J., and Karlan, D. (2021): "Randomizing Religion: The Impact of Protestant Evangelism on Economic Outcomes," The Quarterly Journal of Economics, 136(1), 293–380.
- Chong, A., De La O, A. L., Karlan, D., and Wantchekon, L. (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(1), 55–71.
- Chong, A., Karlan, D., Shapiro, J., and Zinman, J. (2015): "(Ineffective) messages to encourage recycling: evidence from a randomized evaluation in Peru," The World Bank Economic Review, 29(1), 180–206.
- Crépon, B., Devoto, F., Duflo, E., and Parienté, W. (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(1), 123–150.
- Devoto, F., Duflo, E., Dupas, P., Parienté, W., and Pons, V. (2012): "Happiness on tap: Piped water adoption in urban Morocco," American Economic Journal: Economic Policy, 4(4), 68–99.
- Drexler, A., Fischer, G., and Schoar, A. (2014): "Keeping It Simple: Financial Literacy and Rules of Thumb," American Economic Journal: Applied Economics, 6(2), 1–31.
- Duflo, E., Dupas, P., and Kremer, M. (2015): "Education, HIV, and early fertility: Experimental evidence from Kenya," American Economic Review, 105(9), 2757–97.
- Dupas, P. and Robinson, J. (2013): "Savings constraints and microenterprise development: Evidence from a field experiment in Kenya," American Economic Journal: Applied Economics, 5(1), 163–92.
- Dupas, P. and Robinson, J. (2013): "Why don't the poor save more? Evidence from health savings experiments," American Economic Review, 103(4), 1138–71.
- Dupas, P. (2011): "Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya," American Economic Journal: Applied Economics, 3(1), 1–34.
- Dupas, P., Hoffmann, V., Kremer, M., and Zwane, A. P. (2016): "Targeting health subsidies through a nonprice mechanism: A randomized controlled trial in Kenya," Science, 353(6302), 889–895.
- Dupas, P., Karlan, D., Robinson, J., and Ubfal, D. (2018): "Banking the unbanked? Evidence from three countries," American Economic Journal: Applied Economics, 10(2), 257–97.
- Dupas, P., Huillery, E., and Seban, J. (2018): "Risk information, risk salience, and adolescent sexual behavior: Experimental evidence from Cameroon," Journal of Economic Behavior & Organization, 145, 151–175.
- Fink, G. (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(4), 1070–1077.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Oregon Health Study Group (2012): "The Oregon health insurance experiment: evidence from the first year," The Quarterly Journal of Economics, 127(3), 1057–1106.

- Giné, X., Karlan, D., and Zinman, J. (2010): "Put your money where your butt is: a commitment contract for smoking cessation," American Economic Journal: Applied Economics, 2(4), 213–235.
- Guiteras, R., Levinsohn, J., and Mobarak, A. M. (2015): "Encouraging sanitation investment in the developing world: A cluster-randomized trial," Science, 348(6237), 903–906.
- Hanna, R., Duflo, E., and Greenstone, M. (2016): "Up in smoke: the influence of household behavior on the long-run impact of improved cooking stoves," American Economic Journal: Economic Policy, 8(1), 80–114.
- Hicken, A., Leider, S., Ravanilla, N., and Yang, D. (2018): "Temptation in vote-selling: Evidence from a field experiment in the Philippines," Journal of Development Economics, 131, 1–14.
- Karlan, D., Savonitto, B., Thuysbaert, B., and Udry, C. (2017): "Impact of savings groups on the lives of the poor," Proceedings of the National Academy of Sciences, 114(12), 3079–3084.
- Karlan, D., Osman, A., and Zinman, J. (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., Mullainathan, S., and Roth, B. N. (2019): "Debt traps? Market vendors and moneylender debt in India and the Philippines," American Economic Review: Insights, 1(1), 27–42.
- Karlan, D. and Zinman, J. (2011): "Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation," Science, 332(6035), 1278–1284.
- Khan, A. Q., Khwaja, A. I., and Olken, B. A. (2016): "Tax farming redux: Experimental evidence on performance pay for tax collectors," The Quarterly Journal of Economics, 131(1), 219–271.
- Mohammed, S., Glennerster, R., and Khan, A. J. (2016): "Impact of a daily SMS medication reminder system on tuberculosis treatment outcomes: a randomized controlled trial," PloS One, 11(11), e0162944.
- Oreopoulos, P., Angrist, J., and Williams, T. (2014): "When Opportunity Knocks, Who Answers? New Evidence on College Achievement Awards," Journal of Human Resources, 49(3), 572–610.
- Pons, V. and Liegey, G. (2019): "Increasing the electoral participation of immigrants: Experimental evidence from France," The Economic Journal, 129(617), 481–508.
- Pons, V. (2018): "Will a Five-Minute Discussion Change Your Mind? A Countrywide Experiment on Voter Choice in France," American Economic Review, 108(6), 1322–1363.
- Romero, M., Sandefur, J., and Sandholtz, W. A. (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.

D Additional Figures and Tables

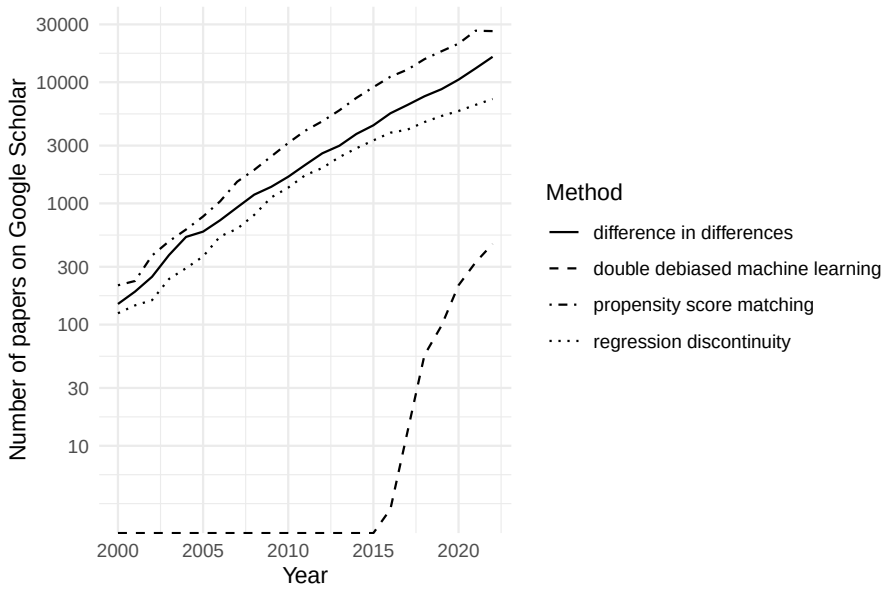
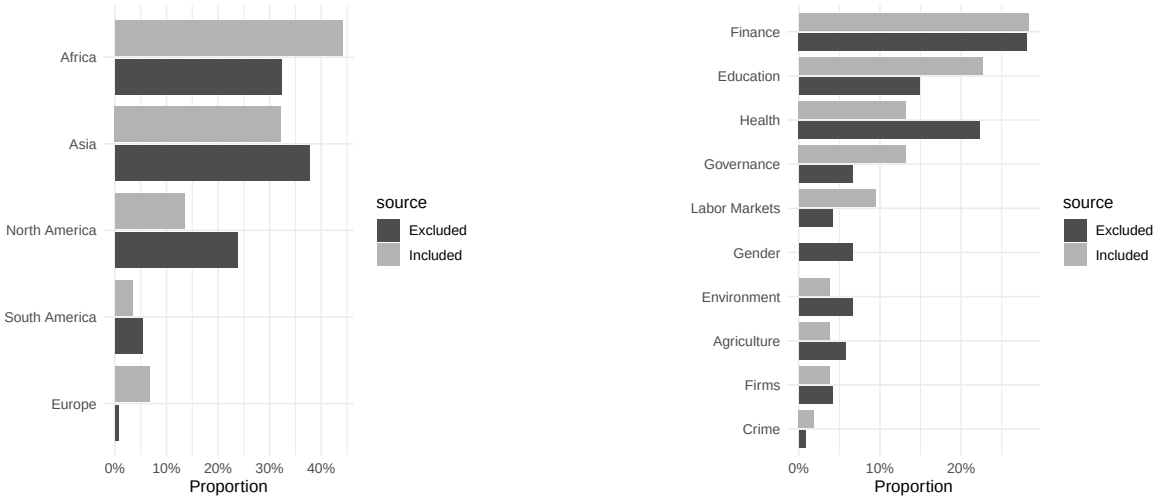


Figure D.1: Number of papers on Google Scholar mentioning various methods



(a) Continent

(b) Sector

Figure D.2: Characteristics of excluded and included papers

Table D.1: Summary statistics by study

Study	# Specifications	Average # covariates	Average # observations	Average take-up ($R = 1$)
1	5	34	1311	0.24
2	32	49	1935	0.88
3	18	18	965	0.28
4	62	61	1139	0.46
5	27	175	1775	0.67
6	11	15	244	0.43
7	12	37	247	0.74
8	5	9	2054	0.91
9	6	30	7405	0.47
10	5	210	875	0.42
11	72	22	14954	0.18
12	21	39	6585	0.99
13	34	115	4927	0.17
14	10	3	1070	0.73
15	2	112	652	0.31
16	101	125	1921	0.56
17	55	658	1024	0.74
18	53	16	717	0.12
19	376	1613	644	0.94
20	25	413	5879	0.63
21	170	1467	826	0.83
22	15	927	11524	0.44
23	60	392	333	0.53
24	10	23	343	0.85
25	8	72	511	0.53
26	6	23	1661	0.66
27	91	49	1585	0.87
28	36	8	2381	0.87
29	3	16	2039	0.90
30	58	541	471	0.87
31	20	16	249	0.68
32	18	12	2590	0.65
33	8	116	14393	0.08
34	33	885	597	0.76
35	38	6	4486	0.24
36	107	87	2528	0.96
37	42	650	2151	0.82
38	47	26	3720	0.91
39	24	114	4982	0.73
40	75	39	5149	0.83
41	12	7	11982	0.37
42	137	2	903	0.55
43	70	16	5877	0.46
44	8	14	13905	0.08
45	13	64	679	0.95
46	6	105	19598	0.91
47	96	13	1183	0.58
48	50	29	3987	0.90
49	19	1	3224	0.93
50	168	45	326	0.80
51	105	117	21585	0.35
52	36	63	540	0.49
53	19	124	683	0.52
Mean	48	185	3836	0.62

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 5 displays the average take-up in the treatment group.

Table D.2: Meta-analysis on specifications where lagged outcomes are available

	TE	WW	DDML
Panel C: Individual primary outcomes			
Mean ($\hat{\mu}$)	0.222	-0.115	-0.109
SE ($\hat{\sigma}_{\mu}$)	(0.048)	(0.046)	(0.040)
Total standard deviation ($\hat{\tau}$)		0.240	0.215
Effective SE		0.244	0.219
Num. obs.	204	204	204
Panel D: Individual outcomes			
Mean ($\hat{\mu}$)	0.094	-0.038	-0.047
SE ($\hat{\sigma}_{\mu}$)	(0.040)	(0.035)	(0.029)
Total standard deviation ($\hat{\tau}$)		0.257	0.217
Effective SE		0.259	0.219
Num. obs.	726	726	726

Notes: We re-estimate the meta-analysis restricting attention to specifications in which the a pre-treatment measure of the outcome variable available. Since aggregated outcomes are based on several outcomes that each may or may not have an individual lagged outcome, we estimate these specifications only for individual outcomes.

Table D.3: Meta-analysis on individual outcomes under alternative within-study correlation assumptions

	TE	Within-study corr. $\rho = 0.5$			Within-study corr. $\rho = 0.7$		
		WW	PDSL	DDML	WW	PDSL	DDML
Panel C: Individual primary outcomes							
Mean ($\hat{\mu}$)	0.157	-0.050	-0.057	-0.050	-0.053	-0.060	-0.056
SE ($\hat{\sigma}_{\mu}$)	(0.031)	(0.037)	(0.031)	(0.030)	(0.037)	(0.032)	(0.030)
Total standard deviation ($\hat{\tau}$)		0.278	0.224	0.221	0.283	0.241	0.234
Effective SE		0.280	0.226	0.223	0.286	0.243	0.236
Num. obs.	540	540	534	540	540	534	540
Panel D: Individual outcomes							
Mean ($\hat{\mu}$)	0.078	-0.010	-0.031	-0.027	-0.011	-0.035	-0.030
SE ($\hat{\sigma}_{\mu}$)	(0.017)	(0.022)	(0.024)	(0.017)	(0.024)	(0.025)	(0.017)
Total standard deviation ($\hat{\tau}$)		0.248	0.308	0.201	0.270	0.345	0.227
Effective SE		0.249	0.309	0.202	0.271	0.345	0.228
Num. obs.	2540	2540	2368	2540	2540	2368	2540

Notes: Our primary estimates assume a within-study correlation of 0.6 when there are multiple bias estimates per study. To assess robustness, this table reports the meta-analysis for individual outcomes assuming within-study correlations of 0.5 and 0.7.

E Robustness

E.1 Standard Error Robustness

As explained in Appendix B.2, and focusing on a single outcome per study, our main analysis computes the sampling variance (or within-study 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 sampling variance of observational 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 assigns, for each specification fed into the meta-analysis, an upper bound on the true variance of observational 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.1.1 Bootstrap

Bootstrapping our estimates is computationally costly because it involves repeatedly re-running the machine-learning observational estimators. Table E.1 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 -0.008 and an effective SE of 0.206. Thus, our overall conclusions do not appear to be materially affected by the independence assumption.

E.1.2 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 as explained in Appendix E.1 , consider the closely-related Hedges' Estimator, which has a simpler formula (see Chabé-Ferret (2023) for details):[34]

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

[34]The actual estimator we are using is

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

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

Table E.1: Bias estimates using bootstrap standard errors

	TE	WW	DDML
Panel A: Aggregated primary outcomes (bootstrap)			
Mean ($\hat{\mu}$)	0.145	-0.026	-0.008
SE ($\hat{\sigma}_{\mu}$)	(0.040)	(0.044)	(0.038)
Standard deviation ($\hat{\tau}$)		0.254	0.203
Effective SE		0.258	0.206
Num. obs.	51	51	51
Panel B: Aggregated primary outcomes			
Mean ($\hat{\mu}$)	0.173	-0.030	-0.025
SE ($\hat{\sigma}_{\mu}$)	(0.041)	(0.044)	(0.038)
Standard deviation ($\hat{\tau}$)		0.251	0.204
Effective SE		0.255	0.207
Num. obs.	51	51	51

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 $= \left(\sqrt{\hat{\sigma}_{\mu}^2 + \hat{\tau}^2}\right)$. We provide results based on bootstrap standard errors solely for our main specification, the aggregated primary outcomes, due to computational constraints.

We have:

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

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 ($\overline{Cov} > 0$).

Thus this back-of-the-envelope calculation is consistent with the claim that our main results do not materially overestimate the effective standard error. If anything, we are probably slightly underestimating the actual size of the effective standard error.

E.2 Prompts used for LLM bias prediction

We used four prompt variants in the API calls:

- WW, With paper == 0
- WW, With paper == 1
- DDML, With paper == 0
- DDML, With paper == 1

All four share the following common template. WW is the base prompt. DDML-only additions are shown in red.

You are an expert in social science, particularly economics. You are familiar with the research literature (both published and unpublished work) on program evaluation in high-and low-income countries. You also have deep expertise about human nature and human behavior informed by your research expertise.

Based on your knowledge of human behavior and the social science literature, please consider the following task. I will start by providing some background information and context before describing the task itself.

I am first going to describe a social program and an eligible population to whom that program was made available. Program Description: {ProgramDescription} Eligible Population: {EligiblePopulation}

Context

- "Program description" is a text description of a social program made available to a set of potential beneficiaries.
- "Eligible Population" is a text description of the population that is eligible to receive the program.
- However, not everyone in the eligible population chose to take up the program. We will call those who did take up the program "takers" and those who did not take up the program "non-takers".
- Multiple welfare-relevant outcomes are tracked; each is coded so that "more" means higher welfare.
- All comparisons below refer to the welfare level a person would reach if the program had never existed (the counterfactual world).
- We have access to a rich set of observable characteristics about the eligible population that we can condition on when performing our analysis. Imagine that we have the typical observable characteristics that are collected in a household survey such as the Demographic and Health Surveys (DHS) or the Current Population Survey (CPS). The precise observable characteristics will be those that are particularly useful or relevant given the program and eligible population.
- When conditioning on observable characteristics, we use a state-of-the-art method: double machine learning (DML). This approach flexibly selects and adjusts for observables while mitigating regularization bias and overfitting through cross-fitting and orthogonalization.

Definitions

- Positive selection -- *after conditioning on observable characteristics*, takers would still have **higher counterfactual welfare** than non-takers.
- Negative selection -- *after conditioning on observable characteristics* takers would instead have **lower counterfactual welfare** than non-takers.

(These labels describe welfare comparisons only; they do not refer to participation probabilities.)

Illustrative examples (for guidance only)

- "Unemployment insurance offered to the general population" -- we might expect people who are more likely to lose their jobs to enroll, so likely **negative** selection.

- "Free prenatal vitamins for pregnant women" -- we might expect more health-conscious mothers to enroll, so likely **positive** selection.

Task

Classify the direction of selection for the program above.

Response format -- exactly two lines

1. ‘label: positive‘ or ‘label: negative‘
2. ‘rationale: <one concise sentence explaining the key mechanism underlying the label>‘

For the two variants with paper context (With paper == 1), the following block was appended at the end of the base prompt (identical for WW and DDML):

To assist you in your judgment, here is the full text of the research paper that originally studied the program of interest. This paper provides additional background on the context, program, eligible population, and the data available.

Important Guidance:

- The answer to your task is likely not directly stated in the research paper. While the authors may discuss selection patterns, often they do not.
- The research paper describes a randomized experiment that evaluated the program of interest.
- Your task is to use your expertise to predict likely selection patterns if beneficiaries were free to choose whether or not to take up the program.
- Use the research paper to inform your contextual understanding, but do not simply extract or summarize information from it. Focus on making an informed prediction based on your knowledge, using the paper only as supporting context.

Now, here is the text of the research paper: {Text here}

E.3 Threats to Exchangeability

Our approach relies heavily on the assumption of exchangeability which requires that any observational study of interest is drawn from the same population as the set of studies we use for estimation. In this section, we discuss the several ways that this assumption might fail (site selection bias and publication bias) and how they might impact our estimates.

E.3.1 Site selection bias

One way exchangeability might fail is if the set of programs for which an experiment has been run differs substantially from the set of programs for which an experiment has not been run. One mechanism that can give rise to such a difference is site selection bias: researchers and their operational partners might be more likely to implement an experimental evaluation in locations where they expect a project to perform especially well, be it because the characteristics

of the agents served by the program are especially favorable to its effectiveness, or because the caseworkers in charge of implementing the program are especially knowledgeable and efficient.

To be slightly more formal, let us define a variable M_s which takes value 1 when study s is conducted using an RCT and 0 when it is conducted with an observational method. Other possible margins of selection would be whether to run a program or not, whether to conduct an evaluation or not and whether to conduct an ICRCT or not. The first ones are not relevant to our application since we consider the population of studies for which an observational evaluation exists. The last margin of selection does not affect the main thread in the argument.

Let us focus on recovering μ, the mean selection bias. We can only observe either an experimental estimate or an observational estimate for each study, the only feasible comparison that might give us μ is the difference between observational and experimental estimates. We have:

$$\mathbb{E}[OBS_s|M_s=0] - \mathbb{E}[EXP_s|M_s=1] = \underbrace{\mathbb{E}[OBS_s - TOT_s|M_s=0]}_{\mu_0} + \underbrace{\mathbb{E}[TOT_s|M_s=0] - \mathbb{E}[TOT_s|M_s=1]}_{\text{site selection bias}}$$

where μ_0 is our parameter of interest: average selection bias among programs that have been evaluated with an observational method. Extracting μ_0 from the observed data requires an instrument that moves M_s, enabling identification of site selection bias. This is the approach followed by Gechter (2022).

In contrast, our approach leverages variations within ICRCTs with $M_s = 1$ to identify selection bias directly:

$$\mathbb{E}[OBS_s - EXP_s|M_s=1] = \underbrace{\mathbb{E}[OBS_s - TOT_s|M_s=1]}_{\mu_1}.$$

Our approach is therefore not directly affected by site selection bias. We face a different problem though: whether selection bias varies with M_s. Exchangeability imposes that $\mu_1 = \mu_0$. One way that this assumption might fail is if the process of selection into the program responds to either M_s or to the expected program impacts which drives M_s. Programs that are evaluated with an RCT might yield caseworkers to select individuals based on their expected response to treatment, in order to prove effectiveness. If individuals with lower outcomes in the absence of the treatment respond more to the treatment, we expect μ_1 to be negative. Programs that are evaluated with observational methods might yield caseworkers to cream-skim, yielding a positive μ_0. As a consequence, we expect the bias to our estimate of μ_0 to be negative. This is compatible with the small and statistically insignificant negative estimates that we find.

One way to test whether $\mu_1 = \mu_0$ would be to use an instrument Z_s for M_s and to compare $\mathbb{E}[OBS_s - EXP_s|M_s = 1, Z_s = z]$ at different values of z. At higher values of z (for which $\Pr(M_s = 1|Z_s = z)$ is high), experimental evaluations should contain locations where programs

have lower effects, while at lower values of z (for which $\Pr(M_s = 1|Z_s = z)$ is low), only the best programs are experimentally evaluated, yielding a larger treatment effect. If our estimates of selection bias do not vary with z, exchangeability would be supported. If it were the case that our estimates of selection bias vary with z, we could estimate a selection model to be able to infer the distribution of selection bias when $M_S = 0$.[35]

E.3.2 Publication bias

Another way exchangeability might fail is because of publication bias. There is publication bias when the probability of publication of the result of an ICRCT is higher when its result is above the threshold of statistical significance. For simplicity, we are going to assume that publication bias is such that we only observe experimental results that are statistically significant and positive. We denote $P_s = 1$ when the experimental estimate for program s has been published and 0 otherwise. Without loss of generality, we also assume that sampling noise around our experimental estimate, v_s^{EXP} is distributed normally with mean zero and variance $\sigma^2_{EXP,s}$. We make the same assumptions for the observational estimator, replacing the EXP index by the OBS index. Finally, we assume that the correlation between the estimates due to sampling noise is ρ.

With this simple model, we have that $P_s = \mathbb{1}[\frac{EXP_s + v_s^{EXP}}{\sigma_{EXP,s}} \geq 1.96]$, and therefore:

$$\begin{aligned}\mathbb{E}[\widehat{EXP}_s|OBS_S, EXP_s, P_s = 1] &= \mathbb{E}[EXP_s + v_s^{EXP}|EXP_s, \widehat{EXP}_s \geq 1.96\sigma_{EXP,s}] \\ &= EXP_s + \sigma_{EXP,s}\frac{\phi\left(\frac{1.96\sigma_{EXP,s} - EXP_s}{\sigma_{EXP,s}}\right)}{1 - \Phi\left(\frac{1.96\sigma_{EXP,s} - EXP_s}{\sigma_{EXP,s}}\right)},\end{aligned}$$

with ϕ and Φ respectively the density and cumulative distribution function of the standard normal. We also have:

$$\begin{aligned}\mathbb{E}[\widehat{OBS}_s|OBS_S, EXP_s, P_s = 1] &= \mathbb{E}[OBS_s + v_s^{OBS}|EXP_s, \widehat{EXP}_s \geq 1.96\sigma_{EXP,s}] \\ &= OBS_s + \rho\sigma_{OBS,s}\frac{\phi\left(\frac{1.96\sigma_{EXP,s} - EXP_s}{\sigma_{EXP,s}}\right)}{1 - \Phi\left(\frac{1.96\sigma_{EXP,s} - EXP_s}{\sigma_{EXP,s}}\right)},\end{aligned}$$

As a consequence:

$$\mathbb{E}[\widehat{OBS}_s - \widehat{EXP}_s|OBS_S, EXP_s, P_s = 1] = \hat{B}_s + (\rho\sigma_{OBS,s} - \sigma_{EXP,s})\frac{\phi\left(\frac{1.96\sigma_{EXP,s} - EXP_s}{\sigma_{EXP,s}}\right)}{1 - \Phi\left(\frac{1.96\sigma_{EXP,s} - EXP_s}{\sigma_{EXP,s}}\right)},$$

which implies that, as long as $\sigma_{OBS,s} \approx \sigma_{EXP,s}$, publication bias tends to bias our estimates of selection bias downwards, since the experimental estimate of the treatment effect is more inflated

[35] One candidate instrument Z_s is the one used by Gechter (2022): the date at which an IPA or JPAL office has been introduced in a country. We leave the implementation of this estimation for further research.

by publication bias than the observational one. Note that this is also compatible with the slightly negative average selection bias that we estimate.[36]

[36]We could systematically correct for publication bias by estimating the joint distribution of experimental and observational estimates taking into account differential probabilities of publication based on the experimental estimate, extending Andrews and Kasy (2019). Note that we could simultaneously account for selection into the experimental evaluation as suggested by Gechter (2022). We leave the estimation of this full selection model for future work.