Combining Experimental and Observational Studies in Meta-Analysis: A Mutual Debiasing Approach[*]

Michael Gechter[†] Rachael Meager[‡]

September 30, 2021

Abstract

We propose a method for aggregating evidence from observational studies, which may be subject to internal selection bias, and randomized controlled trials (RCTs) which may be subject to site selection bias. We show that it is possible to nonparametrically debias both types of studies using an Instrumental Variables (IV) strategy, uncovering the true distribution of treatment effects for complier studies. As we often have a small number of studies, imperfect instruments, and observe study results with error, parametric hierarchical Bayesian models work well in practice. Our specific implementation uses the presence of a facilitating organisation such as the Jameel Poverty Action Lab (JPAL) or Innovations for Poverty Action (IPA) as a novel Differences in Differences "Plausibly Exogenous" IV. This model point-identifies the internal selection bias for studies switching to the experimental design due to JPAL-IPA entry which is the negative local average treatment effect (LATE) in our setting. Applying this strategy to Conditional Cash Transfers (CCTs) and Microcredit shows substantial internal selection bias in observational studies of CCTs and much less for Microcredit; neither application shows evidence of RCT site selection bias, though credible intervals are wide with minimal updating.

JEL codes: C21, C93, O1

Keywords: meta-analysis, experiments, observational studies, microfinance, conditional cash transfers

[*] We thank the faculty affiliates and staff of the Abdul Latif Jameel Poverty Action Lab, and Innovations for Poverty Action, for generously sharing their time, feedback and data with us. In particular we thank David Sears, Dean Karlan, Anne Thibault, Laura Posner, Loic Watine and Ilf Bencheik. We are grateful to audiences at the WEAI Annual Conference special session on Combining Observational and Experimental Data, the Second Annual Y-RISE Conference, and the EGAP and The Policy Lab at Brown Multi-Site Trials Meeting for helpful comments. We thank Charlotte Wachter, David Cai, Ed Jee and Yootha Yong for excellent research assistance.

[†] The Pennsylvania State University, mdg5396@psu.edu

[‡] The London School of Economics, r.meager@lse.ac.uk

1 Introduction

The popularity of randomized controlled trials (RCTs) in economics is partially driven by the desire to avoid classical selection bias, which can arise in observational studies when individuals or firms can select into behaviours or interventions in a manner plausibly correlated to their outcomes. Yet RCTs are not immune to other kinds of bias, particularly "site selection bias", which may be of concern if the locations, study contexts or implementing partners who are willing or able to do RCTs differ systematically from the broader population of locations or implementers (Allcott (2015), Reid, Fagan, and Zahawi (2018), Coe, Njoloma, and Sinclair (2019), Laajaj, Macours, Masso, Thuita, and Vanlauwe (2020), Andor, Peters, Gerster, and Schmidt (2020)). In this paper, we propose a new approach to combine and thereby mutually debias these two types of studies. We show that the true distribution of treatment effects is nonparametrically identified in theory using an instrumental variables (IV) approach, and parametrically estimable in practice using a Bayesian Hierarchical implementation.

At present, meta-analyses in economics have generally pursued one of two strategies when confronted with different types of evidence: either omit the observational evidence due to the internal selection bias problem (e.g. Hsiang, Burke, and Miguel (2013), Bandiera, Fischer, Prat, and Ytsma (2021), Meager (2019)) or combine both types of studies together[1] and treat them equivalently in the analysis (Chliova, Brinckmann, and Rosenbusch (2015), Miller, Reichelstein, Salas, and Zia (2015), Vivalt (2020)). Simply combining all studies on equal footing may seem undesirable, but many other differences across studies could be more problematic, including the RCT site selection bias itself. Throwing observational evidence away is costly; RCTs are few and expensive to conduct, making it difficult to estimate average treatment effects across literatures in the presence of heterogeneous effects, let alone to conduct meta-regression. It is unclear which strategy is preferable in any given literature partially because we do not know how severe either type of bias is. In this paper we build models that quantify both types of biases and use that information to improve our estimate of the average treatment effect in the literature as a whole.

We first present theory showing conditions under which it is possible to nonparametrically debias the observational and experimental studies. If we observe study results directly, even if only at the aggregate level, then we can use an IV strategy to characterise the site-selection bias in the RCT results, and thus uncover the average internal selection bias in the observational studies. We do this by mapping the RCT site selection problem onto a switching regression or Roy (1951) model such as the Heckman (1976) two-sided selection model, whose nonparametric form Vytlacil (2002)

[1] Sometimes after screening the observational studies for quality as in Vivalt (2020).

showed to be equivalent to the Imbens and Angrist (1994) local average treatment effect (LATE) assumptions. We assume that each study can be implemented using one of two research designs, experimental or observational, where observational research designs cannot guarantee their estimand identifies a treatment effect free from internal selection bias. Through this assumption, we exclude from our analysis studies which could not have been implemented experimentally even with almost infinite resources such as work in economic history or global general equilibrium trade models.

Choosing the experimental research design imposes an unobserved and potentially large net cost, which may be correlated with the main result of the study under either or both designs. When this correlation is non-zero, there is site selection bias.[2] With an instrument that changes the net cost of the experimental design without affecting the potential results of the study authors would obtain under either or both designs, we show that the average internal selection bias for complier studies switching research design due to the instrument is identified. Given this, we can also identify a measure of site selection bias by comparing the experimental results obtained by complier studies to those obtained by studies which would choose the experimental design regardless of the cost change offered by the instrument.

We further show that when RCTs have partial compliance and analysts have access to the full microdata from all studies it is possible to uncover the full joint distribution of treatment effects and internal selection biases in the population of studies using the same IV. Intuitively, RCTs with partial compliance nest hypothetical observational studies, and one can extrapolate the correlation between the treatment effects and would-be selection biases within the RCTs across to the observational studies using the above instrument following the approach of Arellano and Bonhomme (2017)'s distributional sample selection model. This identifies the joint distribution of treatment effects and selection biases in the population of studies. Importantly, identifying the distribution of treatment effects and selection biases allows the meta-analyst to generate a predictive distribution for the true (experiment-identified) treatment effect of an observational study after conditioning on its main result.

The strategy above requires an instrument for selection into study type that is independent of treatment effects and internal selection bias; we propose and construct one such instrument based on the presence of a facilitating organisation such as the Jameel Poverty Action Lab (JPAL) or Innovations for Poverty Action (IPA). Although the establishment of JPAL-IPA presence is likely to be endogenous to the fundamental contextual factors in a given location, once established, presence is highly persistent and makes RCTs more likely to occur in a location even conditional on the given

[2] Note that the observational studies are also selected since they have relatively high unobserved costs of experimental implementation.

fundamentals. This suggests a Differences in Differences IV strategy that uses the switching on of JPAL-IPA presence as the exogenous variable. To implement this idea, we develop a new approach to Instrumented Difference in Differences or DD-IV (Hudson, Hull, and Liebersohn (2017), also known as Fuzzy DID, de Chaisemartin and D'Haultfoeuille (2018)) which allows us to point-identify the LATE for studies whose research design is changed due to JPAL-IPA's entry into their country, which is the negative internal selection bias.

Our method uses three parallel trend assumptions across locations JPAL-IPA entered and those where the organizations never established a presence. The first parallel trend is on the probability that a given study is implemented as an experiment and the latter two are on the time trends in experimental and observational results. The first assumption yields the DD-IV first stage and is common in applied work, as documented by de Chaisemartin and D'Haultfoeuille (2018). The latter two replace the typical assumption of a parallel trend in *average* results, which, as we discuss, is hard to square with a parallel trend in the probabilty of choosing an experimental design. Together, the three parallel trends identify the counterfactual average post-entry period result had JPAL-IPA never entered. The difference between the actual and counterfactual post-entry average result is the DD-IV reduced form. When divided by the DD-IV first stage, it forms a Wald ratio much like that of Imbens and Angrist (1994).

Due to practical concerns about the parallel trends assumption and thus instrument validity, we pursue a "Plausibly Exogenous" version of this strategy which replaces the strict exogeneity assumption with a smooth penalty on violations of the exclusion restriction (Conley, Hansen, and Rossi (2012)). The small number of studies typically found in applied microeconomic research areas motivates a parametric implementation of the above procedure. Moreover, even in RCTs, treatment effects are always observed with error due to sampling variation – but the arguments above rest on the direct observation of study results for the same reason that the typical LATE argument requires one to accurately observe outcomes for individuals in the sample. This sampling variation affects the debiasing procedure: in threshold-crossing models for site selection, the covariation between the latent cost of choosing the experimental design and the observed outcome data is a key parameter, but the variation in the "outcome data" itself is challenging to estimate when these data are composed of "estimated treatment effects." We provide a parametric Bayesian hierarchical implementation to perform the necessary deconvolution of the two sources of variation jointly with the estimation of the two types of biases, and we show via simulations that it works reasonably well in small samples even with incorrect priors.

We then apply this model with the JPAL-IPA instrument in a DD-IV strategy to two literatures in development economics, and find no conclusive evidence of site selection bias in the experimental studies, though the uncertainty intervals are wide. We first examine the literature on Conditional Cash Transfers, and find little evidence of site selection bias (though there is little updating in general on this parameter). We find strong evidence of internal selection bias in the observational studies, and the direction of this bias is positive within government studies and negative within non-government studies. We next examine the literature on expanding access to microcredit services and find a somewhat more precise zero on both site selection bias and internal selection bias once time trends are accounted for; this result arises because the hierarchical structure of the model effectively discards a large and imprecise outlying observational study.

Our approach complements various alternative approaches to combining experimental and observational data in economics and more broadly, such as Hartman, Grieve, Ramsahai, and Sekhon (2015), Hull (2018), Peysakhovich and Lada (2016), Gui (2020), and Wang and Yang (2021). We are most closely aligned with Athey, Chetty, and Imbens (2020) who combine a single experiment with a single observational study to correct long-run treatment effect estimates. Their approach uses control functions for the latent factors driving the observational study bias and can therefore be used with only two studies even with different outcome variables, but it assumes there is no site-selection bias among the RCTs; our approach uses the multi-study IV strategy to measure and correct for site-selection bias, but requires more than 2 studies with the same outcome variable. Most of the remaining econometric literature combining observational and experimental data in economics does so in the context of comparing the LATE to the average treatment effect within a single site's population, with examples including Brinch, Mogstad, and Wiswall (2017) and Mogstad, Santos, and Torgovitsky (2018).

With respect to DD-IV, our approach is distinct from de Chaisemartin and D'Haultfoeuille (2018). Applied to our context, de Chaisemartin and D'Haultfoeuille (2018) aim to estimate the average internal selection bias for all studies in areas which would have been observational *had they been implemented prior to JPAL-IPA entry.* Rather than using the trend in the probability of choosing an experimental design in areas JPAL-IPA never entered to identify studies which would have been conducted experimentally regardless of JPAL-IPA entry, de Chaisemartin and D'Haultfoeuille (2018) would focus on comparison groups where the experimental design probability does not change over time since these are the comparison groups where they secure point identification. To emphasize the difference in estimands and approaches we use the DD-IV terminology from Hudson et al. (2017) rather than de Chaisemartin and D'Haultfoeuille (2018)'s Fuzzy DID term. Our parametric implementation takes as its starting point the control function specifications shown by Kline and Walters (2019) to produce LATE estimates equivalent to directly applying the Imbens and Angrist (1994) Wald ratio-based estimator.

The remainder of this paper is organised as follows. Section 2 describes the nonparametric identification results, for a completely exogenous instrument Section 3 discusses one potential instrument, the establishment of JPAL or IPA presence in a country, and our suggested strategy for making use of this variation. Section 4 lays out our DD-IV approach and makes a more detailed comparison to de Chaisemartin and D'Haultfoeuille (2018). Section 5 provides a parametric Bayesian hierarchical implementation of this strategy for plausibly exogenous IV (Conley et al. (2012)) building on the insights from the nonparametric theory. For this section we focus on meta-analyses with few studies from which there are only summary data available, as this is the most common case in practice. Section 6 presents applications to the literatures on conditional cash transfers and to microcredit interventions, and Section 7 concludes.

2 Nonparametric Identification

This section demonstrates that the site-selection bias of the randomized experiments and internal selection bias of the observational studies are nonparametrically identified. We show in Proposition 1 that under conditions analogous to the conventional Imbens and Angrist (1994) local average treatment effect (LATE) assumptions, it is possible to debias the average observational treatment effect estimate for a sub-population of potential studies if one has an instrument for selection of studies into experimental versus observational research designs. Proposition 2 shows that experiments with partial compliance nest hypothetical observational studies, and so in the presence of the instrument above the RCTs further identify the distribution of experimental treatment effects conditional on a particular value of the observational treatment effect in the same study. After adjusting this conditional distribution for selection into experimental research designs, we can debias each of these observational studies.

Even if researchers must use parametric methods for meta-analysis due to the constraints of a particular application, the nonparametric identification results in this section offer some hope that empirical results may generalize beyond any particular functional form.

2.1 Set-up of the Problem

Consider a literature on a particular economic policy or intervention comprised of both experimental and observational studies indexed by $j = 1, 2, 3, \dots J$. Each study j is performed in a certain time period t and in a particular context c. Context encompasses not only geographic location of the study but also the implementation protocol, implementing partner if any, and research team conducting the evaluation. We differentiate between experimental and observational studies using a variable that captures this study design feature, denoted by D_{jct}. This design indicator takes two possible values:

$$D_{jct} = \begin{cases} e & \text{if study } j \text{ is experimental} \\ o & \text{if study } j \text{ is observational} \end{cases}$$

In each study, at a minimum, we observe a main result R_{jct} on some outcome of interest; this result is intended to identify a treatment effect.[3] We generally have in mind studies concerned with policy evaluation and assume throughout this paper that the assumptions made by the authors of these studies which permit them to conceive of identifying such effects, most notably the Stable Unit Treatment Value Assumption, are valid in their contexts. From the point of view of identification, R_{jct} is a population object free from sampling error. In the case of the standard difference-in-means estimate of the experimental ATE,

$$R_{jct} = E[Y_{ijct}|T_{ijct} = 1] - E[Y_{ijct}|T_{ijct} = 0],$$

where we index study subjects by i, denote their measured outcome of interest as Y_{ijct}, and their treatment status as $T_{ijct} \in \{0, 1\}$ with 1 denoting treated and 0 denoting untreated. We may observe information about the studies other than D_{jct}, and we generically denote such information by covariates X_{jct}. The following analysis should be thought of as implicitly be carried out conditional on X_{jct} should heterogeneity be of interest or this make assumptions more credible.

The experimental design gives researchers control over assignment to treatment so that they can assure $R_{jct} = TE_{jct}$. However, these studies may not be drawn randomly from the distribution of potential intervention locations; selection into the sample of experimental studies may be correlated to the treatment effect itself, which in this case creates the potential for "site selection bias." The site-selection bias is a form of classical selection bias: if the probability that we observe a treatment effect is correlated to the value of the effect itself, then the average of the observed set may be a poor indicator of the average of the full set of studies. As such, in Section 2.3 we analyze it in a manner which extends Heckman (1979).

We define observational studies as research designs which do not provide researchers with direct control over the treatment assignment process and where the researchers cannot therefore guarantee that $R_{jct} = TE_{jct}$. We refer to departures of R_{jct} from TE_{jct} as internal selection bias, denoted SB_{jct}. For observational studies $R_{jct} = TE_{jct} +$

[3] This could be a LATE, an average treatment effect on the treated, etc.

SB_{jct}, and we do not know SB_{jct}. Summarizing:

$$R_{jct} = TE_{jct} + \mathbb{1}\{D_{jct} = o\}SB_{jct}.$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. We can equivalently write this in the potential outcomes notation we will use throughout the paper:

$$R_{jct} = \mathbb{1}\{D_{jct} = e\}R^e_{jct} + (1 - \mathbb{1}\{D_{jct} = e\})R^o_{jct} \tag{1}$$

where $R^e_{jct} = TE_{jct}$ and $R^o_{jct} = TE_{jct} + SB_{jct}$. The potential outcome pair for each study, (R^e_{jct}, R^o_{jct}), encodes the idea that any evaluation can be executed via an observational or experimental research design, though potentially at very high cost for the latter, with the research team choosing which is executed.

In this set-up, when we observe the average results in each set of studies $(\overline{R^e_{jct}}, \overline{R^o_{jct}})$, neither makes for an attractive estimate of the average treatment effect for the set of all studies. $\overline{R^o_{jct}}$ is affected by internal selection bias, while $\overline{R^e_{jct}}$ is affected by site selection bias. The presence of the internal selection bias in $\overline{R^o_{jct}}$ also makes the simple average of all the results biased. When we observe the gap between the two averages, $\overline{R^e_{jct}} - \overline{R^o_{jct}}$, we know the difference is comprised of both types of bias. If we can quantify one type of bias, and thus debias the average of one type of study, then we will be able to debias the average of the other type of study using the results of the first debiasing. Proposition 1, in the next section, shows that it is possible to do this average debiasing nonparametrically with an instrument for research design choice. This result allows us to identify the marginal distributions of the true treatment effects $R^e_{jct} = TE_{jct}$ identified using the experimental design and the selection-bias corrupted treatment effects $R^o_{jct} = TE_{jct} + SB_{jct}$ as well as their average difference for complying studies of the instrument. Complying studies are those whose research design choice responds to the instrument, which in our empirical implementation will reduce the cost of choosing the experimental design.

For a given observational study, we would ideally like to do better than debias using the average difference between true and selection-bias-corrupted reported results. We would like to condition on the information encoded in the study's observational treatment effect to arrive at a tailored and potentially lower-variance measure of its internal bias. This kind of conditional debiasing requires identifying dependence between internal selection bias and true treatment effects. We do this in Proposition 2 by leveraging the nested observational studies available in RCTs with imperfect compliance (Section 2.3).

2.2 Proposition 1: Debiasing On Average

We first provide a brief intuitive sketch of the argument that underpins Proposition 1. Begin by noting that Equation (1) invites a switching regression or Roy (1951) model representation as in Heckman (1976) and following papers. In our model researchers choose the experimental design if the unobserved cost of doing so is sufficiently low relative to the observable net benefits. This latent cost may be correlated to treatment effects; this produces a situation in which experimental studies' locations are selected in ways correlated to their treatment effects. But if there is an instrumental variable that predicts research design while being independent of treatment effects, internal selection bias under an observational design, and latent costs this situation is equivalent to the conditions under which the conventional LATE theorem applies (Vytlacil (2002)). The LATE allows us to debias the complying observational studies on average. We additionally invoke the results of Imbens and Rubin (1997) to show that the entire marginal distributions of both true and selection-bias-corrupted treatment effects are nonparametrically identified for complying studies.

Equivalent to the monotonicity assumption in the LATE setup, we assume researchers choose the experimental research design when a function ψ mapping a binary instrument Z_{Djct} to the real line exceeds the 1-dimensional unobserved cost of running an experiment.

Assumption 1. ***Threshold crossing.*** *Let V_{jct} denote the one-dimensional unobserved cost of choosing the RCT design for study jct. The researchers will choose the experimental design if the net benefits due to observed factors $Z_{Djct} \in \{0, 1\}$ exceed the unobserved cost:*

$$D_{jct} = e \iff \psi(Z_{Djct}) \geq V_{jct}.$$

Following Heckman and Vytlacil (2005), we can apply $F_V(\cdot)$ to both sides of the inequality to obtain

$$\begin{aligned} D_{jct} = e &\iff F_V(\psi(Z_{Djct})) \geq F_V(V_{jct}) \\ &\iff P(Z_{Djct}) \geq U_{jct} \end{aligned} \tag{2}$$

where $U_{jct} \sim \text{Uniform}(0, 1)$ and $P(Z_{Djct}) = P(D = e|Z_{Djct})$ is the propensity score. We make use of this normalization for the rest of Section 2.

Analogous to the independence and exclusion LATE assumptions, we assume that Z_{Djct} is independent of the unobserved cost V_{jct} of running study jct as an experiment and also independent of the potential study results under both research designs.

Assumption 2. ***Independence.*** $V_{jct}, R^e_{jct}, R^o_{jct} \perp\!\!\!\perp Z_{Djct}$.

The following proposition shows that under Assumptions 1 and 2, we can identify the marginal distributions of complying studies' true and bias-corrupted treatment effects.

Proposition 1. *Suppose Assumptions 1 and 2 are satisfied. Then the marginal distributions of* R^e_{jct} *and* R^o_{jct} *are identified for studies with* $U \in (P(Z_{Djct} = 0), P(Z_{D_{jct}} = 1)]$. *We will refer to such studies as complying studies since their choice of research design switches according to the value of* Z_{Djct}.

Proof. Vytlacil (2002) shows that Assumptions 1 and 2 are equivalent to the Imbens and Angrist (1994) LATE assumptions with $U \in (P(Z_{Djct} = 0), P(Z_{Djct} = 1)] \iff D(Z_{Djct} = 0) \neq D(Z_{Djct} = 1)$ defining the complying subpopulation. Imbens and Rubin (1997) show that under the LATE assumptions, the marginal distributions of potential outcomes (here R^e_{jct} and R^o_{jct}) are identified for compliers. □

2.3 Proposition 2: The Joint Distribution of Effects and Biases

As mentioned at the beginning of this section, a more tailored debiasing of individual observational studies is theoretically possible using the distribution of the treatment effect observational study jct would have obtained had it been implemented experimentally, conditional on its observational treatment effect. That is:

$$R^e_{jct} | R^o_{jct} = r^o.$$

This can be acheived, once again, with an instrument for research design choice.

The intuition is as follows. If some individuals assigned or encouraged to take the treatment within an experimental study do not comply, and overall subject behavior can be characterized by the LATE/Vytlacil (2002) assumptions then subject-level data allow us to identify both the underlying average treatment effect on the treated and the selection bias in a hypothetical observational study comparing treated compliers to never-takers.[4] In the hypothetical study, the LATE is the treatment effect and the internal selection bias is the difference between the untreated outcome mean for compliers and the (untreated) outcome mean for never-takers. We can identify the internal selection bias because we observe the never-takers' untreated outcome (in the group assigned to treatment, they are the only ones who do not take it up) and we know the untreated in the control group are a mix of never-takers and compliers, so we can also identify the average untreated outcome for compliers.

[4] To make our hypothetical observational studies as comparable to the observational studies we wish to debias, we effectively assume we have access to microdata from the observational studies. We can compute the analogous quantity, the simple treated vs. untreated outcome contrast, as opposed to the estimand of the specific observational research design employed.

Thus, we can identify (TE_{jct}, SB_{jct}) pairs within RCTs, where TE_{jct} is the within-study LATE and SB_{jct} is the complier vs. never-taker contrast, giving us the full joint distribution of these objects. However, to make use of this distribution, we need a way to extrapolate from the behaviour of internal selection bias in RCTs to the behaviour of internal selection bias in observational studies. This is not automatic because of the site selection problem: since the distribution of effects differs across study design type, so too might the internal selection bias. As the distribution of (TE_{jct}, SB_{jct}) cannot be assumed to be the same across the study types, we use the instrument for research design choice to point- or partially-identify a conditional distributional sample selection model following Arellano and Bonhomme (2017). To emphasize the distinction between across-study instruments and identification which were the focus of the previous subsection, and their within-study counterparts, as well as to underscore the role of cost in determining research design choice, we use the Vytlacil (2002) formulation across studies and the LATE formulation within-study.

2.3.1 Within-experiment behavior

We begin by laying out the aforementioned assumptions on within-experiment behaviour. Recall that within jct:

$$T_{ijct} = \begin{cases} 1 & \text{if subject } i \text{ takes the treatment} \\ 0 & \text{if subject } i \text{ does not take the treatment.} \end{cases}$$

$Y_{ijct} \subseteq \mathbb{R}$ denotes i's outcome of interest, related to the potential outcomes Y_{0ijct} and Y_{1ijct} through the equation:

$$Y_{ijct} = T_{ijct}Y_{1ijct} + (1 - T_{ijct})Y_{0ijct}.$$

And for studies with $D_{jct} = e$, Z_{Tijct} denotes experimental treatment assignment so that

$$Z_{Tijct} = \begin{cases} 1 & \text{if subject } i \text{ is assigned to the treatment group} \\ 0 & \text{if subject } i \text{ is assigned to the control group.} \end{cases}$$

Note that the study-specific instruments $\{Z_{Ti1}, \ldots, Z_{TiJ}\}$ (ct subscripts omitted for clarity) form a vector where all elements are zero for all subjects in observational studies and only Z_{Tij} can be equal to one for a subject in study j. Finally, T_{ijct} is itself the result of a potential outcomes equation:

$$T_{ijct} = Z_{Tijct}T_{1ijct} + (1 - Z_{Tijct})T_{0ijct}.$$

Given this setup we make the following assumptions. First, we formally state our definition of imperfect compliance, which is that the probability of taking up the treatment is bounded away from one when assigned to the treatment group.[5]

Assumption 3. ***Imperfect compliance.***

$$E[T_{ijct}|Z_{Tijct}=1]<1\ \forall jct : D_{jct}=e.$$

Second, we make the standard LATE assumptions within experiment, assuming treatment assignment Z_{Tijct} is independent of both sets of within-experiment potential outcomes, (Y_{0ijct}, Y_{1ijct}) and (T_{0ijct}, T_{1ijct}), that treatment assignment has an impact on takeup in expectation, and finally that assignment to the treatment group affects takeup in the same direction for all individuals.

Assumption 4. ***Within-experiment LATE.*** *$\forall jct : D_{jct}=e$ the following conditions hold.*

1. *Independence and Exclusion.* $(Y_{0ijct}, Y_{1ijct}, T_{0ijct}, T_{1ijct}) \perp\!\!\!\perp Z_{Tijct}$.
2. *First stage.* $E[T_{1ijct}-T_{0ijct}]\neq 0$.
3. *Monotonicity.* $T_{1ijct}-T_{0ijct}\geq 0\ \forall i$ *or* $T_{1ijct}-T_{0ijct}\leq 0\ \forall i$.

We define the untreated complier vs. never-taker contrast which represents SB_{jct} in our hypothetical observational studies according to the following equation:

$$SB_{jct}=E[Y_{0ijct}|T_{0ijct}\neq T_{1ijct}]-E[Y_{0ijct}|T_{1ijct}=T_{0ijct}=0].$$

This is the expected difference between the untreated outcome for subjects whose takeup decision changes due to treatment assignment (the compliers) and the untreated outcome for individuals who do not take up the treatment regardless of their experimental]assignment (the never-takers). As shown in Katz, Kling, and Liebman (2001), Abadie (2003), and Kowalski (2016), this object can be identified from experimental data satisfying Assumptions 3 and 4. To see this, first let the proportions of compliers and never-takers in the population of experiment jct be given by p_c and p_n, respectively. These are defined and identified as follows

$$\begin{aligned}p_n&=P(T_{1ijct}=T_{0ijct}=0)=P(T_{ijct}=0|Z_{ijct}=1)\\p_c&=P(T_{0ijct}\neq T_{1ijct})=P(T_{ijct}=1|Z_{ijct}=1)-P(T_{ijct}=1|Z_{ijct}=0)\end{aligned}$$

where the second equality in each row follows from Assumptions 4.1 and 4.3. $p_n>0$ by Assumption 3.

[5] We do not need to bound the treatment probability away from zero for subjects assigned to the control group because always takers do not play a role in our hypothetical observational studies.

SB_{jct} can then be identified according to the following argument. First, by the law of iterated expectations:

$$
\begin{aligned}
&E[Y_{ijct}|T_{ijct}=0, Z_{ijct}=0] \\
&= \frac{p_c}{p_c+p_n}E[Y_{ijct}|T_{1ijct}\neq T_{0ijct}, Z_{ijct}=0] + \frac{p_n}{p_c+p_n}E[Y_{0ic}|T_{1ijct}=T_{0ijct}=0, Z_{ijct}=0] \\
&= \frac{p_c}{p_c+p_n}E[Y_{0ijct}|T_{1ijct}\neq T_{0ijct}] + \frac{p_n}{p_c+p_n}E[Y_{0ic}|T_{1ijct}=T_{0ijct}=0].
\end{aligned}
$$

We can then recover SB_{jct} as

$$
\frac{p_c+p_n}{p_c}(E[Y_{ijct}|T_{ijct}=0, Z_{ijct}=0] - E[Y_{ijct}|T_{ijct}=0, Z_{ijct}=1]) = SB_{jct}.
$$

Now for our experiment-derived observational study with

$$
\begin{aligned}
R^o_{jct} &= E[Y_{1ijct}|T_{1jct}\neq T_{0jct}] - E[Y_{0ijct}|T_{1jct}=T_{0jct}=0] \\
&= TE_{jct} + SB_{jct} \qquad (3)
\end{aligned}
$$

we observe both potential results since $R^e_{jct} = LATE_{jct} \equiv E[Y_{1ijct} - Y_{0ijct}|T_{1jct}\neq T_{0jct}]$. So we can characterize

$$
R^e_{jct}|R^o_{jct}=r^o, D_{jct}=e. \qquad (4)
$$

However, this object is not immediately useful for our goal of characterizing the distribution of the true treatment effect conditional on the bias-corrupted treatment effect[6] in observational studies:

$$
R^e_{jct}|R^o_{jct}=r^o, D_{jct}=o. \qquad (5)
$$

In contrast to the "two-sided" Roy sample selection problem explored in Section 2.2, this is now a classical sample selection problem as in Heckman (1979). $R^e_{jct}|R^o_{jct}$ is identified for experimental studies, but R^e_{jct} is missing for observational studies. To identify (5), we use an instrument for research design choice, Z_{Djct}, and apply the distributional sample selection model of Arellano and Bonhomme (2017).

[6] Note that other conditional distributions are identified according to our argument. For instance we also also identify the distribution of LATE conditional on the expected treated outcome in our experiment-derived observational study as well as the expected untreated outcome

$$
R^e_{jct}|\bar{Y}_{T=1} = E[Y_{1ijct}|T_{1jct}\neq T_{0jct}], \bar{Y}_{T=0} = E[Y_{0ijct}|T_{1jct}=T_{0jct}=0], D_{jct}=e.
$$

We can also augment the conditioning information with average characteristics of the "treated group" using Abadie (2003) and of the untreated group. To keep exposition simple, and to keep the emphasis on the potential outcomes R^e_{jct} and R^o_{jct} we focus on the conditional distribution $R^e_{jct}|R^o_{jct}$ for the rest of the section.

2.3.2 Across experiments

The instrument for research design helps identify the conditional distribution of experimental treatment effects in the non-site-selected contexts because it impacts the probability of choosing the experimental design without changing the distribution of experimental costs and treatment effects (conditional on a value for the observational treatment effect R^o_{jct}). In the extreme, the instrument pushes the probability of choosing the experimental design to 1 so that site selection is no longer a problem and the full conditional distribution of experimental treatment effects is identified, including for studies where experimental implementation is so costly that they would have been observational at other values of the instrument. When the instrument cannot move the conditional experimental design probability all the way to 1, we obtain bounds on the non-site-selected distribution of experimental effects.

Our exposition in this subsection closely follows that of Arellano and Bonhomme (2017) with additional details, and adapted to our context. The model works as follows. Let $q(\epsilon, r^o)$ be the quantile function of $R^e|R^o = r^o$, representing the ϵth quantile of results free from internal selection bias, unconditional on research design choice. Let U_{jct} be a latent cost of performing an experiment distributed uniformly on the interval [0, 1] following the argument in Equation (2). Let R^e_{jct} refer to the LATE from study jct, observed only if $D = e$. Then, we have

$$\begin{aligned}
&R^e_{jct} = q(\epsilon_{jct}, R^o_{jct}) \\
&D_{jct} = e \iff \mathbb{1}\{U_{jict} \leq P(D_{jict} = e|Z_{Djct} = z, R^o = r^o)\}.
\end{aligned}$$

The distributional sample selection model requires the following conditions.

Assumption 5. ***Arellano and Bonhomme (2017)***.

1. *Exclusion. The joint distribution of $(\epsilon_{jct}, U_{jct}) \perp\!\!\!\perp Z_{Djct}|R^o_{jct}$.*
2. *Copula function. The bivariate distribution $\epsilon_{jct}, U_{jct}|R^o_{jct} = r^o$ is absolutely continuous with respect to the Lebesgue measure. Its CDF is the copula function for a given value r^o of R^o_{jct}, $C(\epsilon, u|R^o_{jct} = r^o)$.*
3. *Continuous outcomes. For all r^o: the conditional CDF $F_{R^e|R^o}(r^e|R_{jct} = r^o)$ and its inverse $q(\epsilon, r^o)$ are strictly increasing and $C(\epsilon, u|R^o = r)$ is strictly increasing in ϵ.*
4. *Positive probability of choosing the experimental design on the support of (Z_{Djct}, R^o_{jct}):*

$$P(D_{jct} = e|Z_{Djct}, R^o_{jct}) > 0 \text{ with probability } 1.$$

The exclusion restriction in Assumption 5.1 is satisfied under Assumption 2 but is weaker since it does not restrict the relationship between R^o_{jct} and Z_{Djct}.[7] Assumption 5.2 gives the conditional bivariate distribution of $(\epsilon_{jct}, U_{jct})$ a well-defined density. Note that this rules out cases where there is no internal selection bias since the degenerate distribution $R^e_{jct}|R^o_{jct} = R^o_{jct}$ is not absolutely continuous with respect to the Lebesgue measure. The theory should extend to this limiting case, though we do not undertake the extension here in order to maintain focus on the central argument.[8] Assumption 5.3 restricts our attention to cases where R^e_{jct} is continuously distributed given R^o and allows us to invert the copula function in ϵ. Finally, correcting $R^e_{jct}|R^o_{jct}, D_{jct} = e$ for experimental site selection using Z_{Djct} will only be possible for values r^o and z_d where the probability of choosing an experimental design is positive. Assumption 5.4 ensures this is true across the support of R^o_{jct} and Z_{Djct}. Partial identification results could be derived when $P(D_{jct} = e|Z_{Djct}, R^o_{jct}) > 0$ for only part of the support of R^o_{jct}, Z_{Djct}.

Let $p(r^o, z) = P(D = e|R^o_{jct} = r^o, Z_{Djct} = z)$. Following Arellano and Bonhomme (2017), under Assumptions 3, 4, 5, and 1 where we augment $\psi(\cdot)$ to take R^o_{jct} as a second argument:

$$\begin{aligned}
&P(R^e_{jct} \leq q(\epsilon, r^o)|D = e, R^o_{jct} = r^o, Z_{Djct} = z) && (6)\\
&= P(\epsilon_{jct} \leq \epsilon|D = e, R^o_{jct} = r^o, Z_{Djct} = z)\\
&= P(\epsilon_{jct} \leq \epsilon|U_{jct} \leq p(r^o, z), R^o_{jct} = r^o)\\
&= \frac{C(\epsilon_{jct}, p(r^o, z)|R^o_{jct} = r^o)}{p(r^o, z)}\\
&\equiv G_{r^o}(\epsilon, p(r^o, z)) && (7)
\end{aligned}$$

where the first equality follows from applying $F_{R^e|R^o}(\cdot|R^o_{jct} = r^o)$ to both sides of the inequality, the second follows from the augmented Assumption 1 and the exclusion restrictio,n Assumption 5.1, and the third from the definition of the copula in Assumption 5.2. The function $G_{r^o}(\epsilon, p(r^o, z))$ takes a rank ϵ in the nonselected distribution $R^e_{jct}|R^o_{jct} = r^o$ and a probability of choosing the experimental design given r^o and z, and outputs the corresponding rank in the selected distribution $R^e|D = e, R^o_{jct} = r^o, Z_{Djct} = z$. If identified, it allows us to back out the nonselected distribution $R^e_{jct}|R^o_{jct}$ from observing the selected distribution $R^e_{jct}|R^o_{jct}, D_{jct} = e, Z_{Djct} = z$. As an example, if implementing partners only allowed experimental evaluations of particularly successful programs such that the distribution of $R^e_{jct}|R^o_{jct}, D_{jct} = e, Z_{Djct} = z$ first order stochastically dominated that of $R^e_{jct}|R^o_{jct}, D_{jct} = o, Z_{Djct} = z$ then $G_{r^o}(\epsilon, p(r^o, z)) = \alpha < \epsilon$ for at least one ϵ.

Consider some output of $G_{r^o}(\epsilon, p(r^o, z)) = \alpha$ and two values z and z' on the support

[7] Recall that Assumption 2 has Z_{Djct} independent of the joint distribution of R^e_{jct}, U_{jct} *and* R^o_{jct}.

[8] No internal selection bias is a limiting case in the sense that, under Assumption 5, the variance of $R^e_{jct}|R^o_{jct}$ can be made arbitrarily small as long as it is not exactly zero.

of $Z_{Djct}|R^o_{jct} = r^o$ Then, using Assumption 5.3 for invertibility, Equations (6) and (7) show that:

$$G_{r^o}(G^{-1}_{r^o}(\alpha, p(r^o, z)), p(r^o, z'))$$
$$= F_{R^e|D=e,R^o,Z}(F^{-1}_{R^e|D=e,R^o,Z}(\alpha|D = e, R^o_{jct} = r^o, Z_{Djct} = z)|D = e, R^o_{jct} = r^o, Z_{Djct} = z') \tag{8}$$

Equation (8) provides an opportunity to identify features of $G_{r^o}(\cdot, \cdot)$, as in the example below.

Example 1. *Suppose the joint distribution of ϵ_{jct} and U_{jct} is independent of R^o_{jct} so that $G_{r^o}(\cdot, \cdot) = G(\cdot, \cdot)$. Additionally, let the underlying copula function linking ϵ_{jct} and U_{jct} be the Ali, Mikhail, and Haq (1978) copula with parameter θ,*

$$C(\epsilon, u) = \frac{\epsilon u}{1 - \theta(1 - \epsilon)(1 - u)},$$

so that

$$G(\epsilon, p(r^o, z)) = \frac{\epsilon}{1 - \theta(1 - \epsilon)(1 - p(r^o, z))}$$

and

$$G^{-1}(\alpha, p(r^o, z)) = \frac{\alpha - \theta\alpha(1 - p(r^o, z))}{1 - \theta\alpha(1 - p(r^o, z))}.$$

Applying (8),

$$\frac{\frac{\alpha - \theta\alpha(1-p(r^o,z))}{1-\theta\alpha(1-p(r^o,z))}}{1 - \theta\left(1 - \frac{\alpha-\theta\alpha(1-p(r^o,z))}{1-\theta\alpha(1-p(r^o,z))}\right)(1 - p(r^o, z))} \tag{9}$$
$$= F_{R^e|D=e,R^o,Z}(F^{-1}_{R^e|D=e,R^o,Z}(\alpha|D = e, R^o_{jct} = r^o, Z_{Djct} = z)|D = e, R^o_{jct} = r^o, Z_{Djct} = z') \tag{10}$$

where θ is the only remaining unidentified object and α is arbitrary.

We are now ready to state Proposition 2 regarding non-parametric identification when experimental subjects follow the LATE assumptions and compliance is imperfect.

Proposition 2. *Suppose Assumptions 1, 3, 4, and 5 hold. For r^o on the support of*

$$R^o_{jct} = E[Y_{1ijct}|T_{1jct} \neq T_{0jct}] - E[Y_{0ijct}|T_{1jct} = T_{0jct} = 0]$$

across jct such that $D_{jct} = e$, the nonselected quantile function of R^e_{jct} conditional on the value for the observational treatment effect $R^o_{jct} = r^o$, $q(\epsilon, r^o)$, is:

- *point identified if*
    1. *There exists a value of z such that $P(D = e|R^o_{jct} = r^o, Z_{Djct} = z) = 1$, which Arellano and Bonhomme (2017) refer to as identification at infinity,*
    2. *or, Z_{Djct} is continuous, the support of $P(D = e|R^o_{jct} = r^o, Z_{Djct} = z)$ contains an open interval,s and $G_{r^o}(\epsilon, p(r^o, z))$ is real analytic on the unit interval.*
- *and partially identified otherwise with bounds given by*

$$q(\epsilon, r^o) \leq \inf_{z \in \mathcal{Z}_{r^o}} F^{-1}_{R^e|D,R^o,Z}\left(\min\left\{\frac{\epsilon}{p(z, r^o)}, 1\right\}\middle| D_{jct} = e, R^o_{jct} = r^o, Z_{Djct} = z\right) \tag{11}$$

$$q(\epsilon, r^o) \geq \sup_{z \in \mathcal{Z}_{r^o}} F^{-1}_{R^e|D,R^o,Z}\left(\max\left\{\frac{\epsilon + p(z, r^p) - 1}{p(z, r^o)}, 0\right\}\middle| D_{jct} = e, R^o_{jct} = r^o, Z_{Djct} = z\right) \tag{12}$$

where $\mathcal{Z}_{r^o}$ denotes the support of Z_{Djct} conditional on $R^o_{jct} = r^o$.

Proof. Under Assumptions 3 and 4 both R^e_{jct} and R^o_{jct} are identified for $jct : D = e$ as described in the text so the conditional distribution $R^e_{jct}|D_{jct} = e, R^o_{jct}, Z_{Djct}$ is also identified. Assumptions 1 and 5 satisfy the assumptions of Proposition 1 in Arellano and Bonhomme (2017), which provides the point identification results. Bounds (11) and (12) are due to Manski (1994), as adapted by Arellano and Bonhomme (2017). □

We close this section with several remarks on the implications of Proposition 2. First, the width of the bounds (11) and (12) will tend to be large for values of R^o_{jct} where the share of experiments is low for all values of Z_{jct}. The bounds shrink to a point in the limiting case where the share of experiments is actually equal to 1.

Second, Condition 2 listed under the subheading "point identified" in Proposition 2 provides non-parametric identification if we restrict the conditional dependence structure between the unobserved cost of conducting evaluation jct via RCT and the true treatment effect to be generated by a real analytic copula. That unique copula is identified from within the class of real analytic copulae because Arellano and Bonhomme (2017) show it is unique on the $P(D = e|R^o_{jct} = r^o, Z_{Djct} = z)$ open interval specified in the statement of the result and real analytic functions which coincide on an open neighborhood coincide on their whole domain (the bivariate unit interval in this case). Hence they term this form of point identification as "analytic extrapolation." Note also that we need the instrument to be continuously supported to achieve this result, which will not be the case in our application.

In practice meta-analyses are typically done with relatively small numbers of studies and instruments could have limited support such that a parametric copula would have to be specified.[9] This approach will yield point identification and parametric extrapolation (see Example 1) but researchers and policymakers should seriously consider the possibility of misspecification when using the resulting predictive distributions of true treatment effects for studies with a specific, potentially bias-corrupted treatment effect estimate. Depending on instrument support and data availability, more flexibility can be introduced in the dependence of the copula on R^o_{jct}[10] as well as in the parametric form of the copula itself.

Third, comparing Propositions 1 and 2, note that more is identified in the environment of Proposition 2, where hypothetical observational studies can be generated due to imperfect compliance so long as the LATE assumptions hold within-experiment. For these hypothetical studies both the average treatment effect on the treated R^e_{jct} (the LATE for within-experiment compliers) and the observational treatment effect R^o_{jct} are identified so that the full distribution of selection bias is identified for all experimental studies rather than only those whose research design shifts due to the instrument for research design. However, as we have laid out, it is the *non-hypothetical* set of observational studies for which we would like to know the distribution of the unknown experimental treatment effect, and this requires extrapolation using the instrument. In addition, for Proposition 1 we did not assume anything about the way in which experimental and observational studies are conducted. The observational studies could have been based on unconfoundedness, difference-in-differences (DID), or some other quasi-experimental approach.[11] In contrast, for Proposition 2 we assumed the actual observational studies were based on a simple comparison of treated vs. untreated outcome means when constructing the hypothetical observational studies. Thus Proposition 1 may be best suited to environments where the meta-analyst does not have access to study microdata, while Proposition 2 is more appropriate when all microdata are available and its conditions are met.

3 The JPAL and IPA IV Strategy

Our nonparametric identification results require an instrument for selection into study design: we now propose and construct one possible instrument for literatures in development economics. One need not accept the specific instrumental strategy proposed here make use of the nonparametric results above, and we hope that future work may generate or use alternative instruments, a point we return to in the conclusion.

[9] In Section 5.1 we discuss what can be done with a Gaussian copula and Gaussian marginal distributions for R^e_{jct}, R^o_{jct}, and V_{jct}.

[10] For example by allowing the parameters to depend linearly on R^o_{jct} as suggested in Arellano and Bonhomme (2017).

[11] These features could be elements of X_{jct}.

Our strategy is based on the observation that the establishment of a regional presence by an organisation intended to promote and facilitate the running of field experiments, such as the Jameel Poverty Action Lab (JPAL) or Innovations for Poverty Action (IPA), substantially lowers the costs of performing randomized trials in that region. We interviewed field staff and faculty affiliates from both organisations to understand how and why a presence is established in a given country, and construct a differences in differences based instrumental variable (DD-IV) approach from this information. We note that the exclusion restriction with this strategy is still somewhat tenuous, and describe a Bayesian implementation of the strategy that allows for minor violations of the exclusion restriction based on the approach in Conley et al. (2012).

The intuitive rationale for using JPAL-IPA presence to construct an instrument is based on understanding how randomized trials get done. A major practical barrier to conducting field experiments is their cost in terms of money, time and the institutional knowledge required to navigate the administrative burden of the bureaucratic or political environment. All of these costs are substantially reduced by the presence of JPAL, IPA or similar facilitating organisations in a particular country or region. According to JPAL's website, a major goal of the research lab since its founding in 2003 by Esther Duflo, Abhijit Banerjee and Sendhil Mullainathan has been "to support the use of randomized evaluations" and it has typically worked with IPA, a nonprofit partner organisation founded by JPAL affiliate Dean Karlan (JPAL (2021)). In addition to offering grants to run experiments, JPAL and IPA jointly support the running of experiments in the field via the establishment of a network of affiliated researchers and local in-country staff who have relevant knowledge of the institutional environment, as well as direct expertise in performing these types of studies. For the remainder of this paper we refer to the pair of organisations jointly as JPAL-IPA.

Despite the obvious impact of JPAL-IPA on the cost of doing RCTs, we cannot use mere presence of these organisations in a location as a simple instrument for RCT design, because this presence is not likely to be random with regard to potential results R^e_{jct} and R^o_{jct}. RCTs often require implementing partners - governmental or NGO - willing to subject themselves to scrutiny as well as extensive surveying if not direct monitoring. Moreover, even within this set of amenable contexts, JPAL-IPA do not decide where to establish presence randomly.

To understand how the decision to establish JPAL-IPA presence in a country is made, we conducted qualitative interviews with faculty and staff from across JPAL-IPA. We supplemented this with internal databases of projects, including location and implementation information generously shared with us by these two organisations. This research revealed that a major determinant of *establishing* presence is latent demand from researchers to run RCTs in c. We may imagine this is potentially correlated with

R^e_{jct} in some manner. However, data on JPAL-IPA office openings shows that once established, presence is strongly persistent with very few cases of mass in-country staff redundancies or office closures.

The information above leads to develop a particular DD-IV strategy using the JPAL-IPA presence instrument. Consider a panel data set of contexts indexed by c observed over multiple time periods t. As in our main set-up, we observe a set of J study results R_{jct} indexed by j, with design either experimental ($D_{jct} = e$) or observational ($D_{jct} = o$). JPAL-IPA presence in a location c signals different underlying fundamentals, and thus different results (R^e_{jct}, R^o_{jct}), to places JPAL-IPA never enters. Yet conditional on these, RCTs become less costly in a given location after JPAL-IPA enters relative to before, purely because JPAL-IPA is there facilitating RCTs regardless of fundamentals or trends in (R^e_{jct}, R^o_{jct}). This direct effect on study design in a location c persists even if fundamentals change, due to the strong temporal persistence in regional staff presence.

4 Our Approach to DD-IV

To implement the JPAL-IPA identification strategy we lay out our novel approach to DD-IV, first nonparametrically, then in a parametric specification like the one we take to the data. To simplify exposition, in this section we consider the classic 2-by-2 DID case with two time periods, $t \in \{0, 1\}$, and two groups indexed by $everJPAL_c \in \{0, 1\}$. $everJPAL_c = 1$ denotes c's ever experiencing substantial JPAL-IPA presence in the form of in-country staff or office residency. $JPALswitch_{ct} \in \{0, 1\}$ will indicate whether the JPAL-IPA office is actually operating in c at time t. For the purposes of our 2-by-2 analysis, $JPALswitch_{ct}$ will be 1 only when $everJPAL_c = 1$ and we are in the post-JPAL-IPA-entry period, $t = 1$.

4.1 Nonparametric Analysis

To illustrate our nonparametric identification approach, consider a threshold crossing model for the choice of research design in the style of Equation (2) during period 1 in the context JPAL-IPA enters:

$$D_{jc1} = e \iff P(JPALswitch_{c1}) \geq U_{jc1}.$$

In contrast to the analysis in Section 2.2, there is no variation in $JPALswitch_{ct}$ when $t = 1$ and $everJPAL_c = 1$ which we could assess or restrict in relation to U_{jct} as in Assumption 2. Our strategy will be different. We will use a parallel trend assumption for the probability of choosing the experimental research design to generate a ceteris paribus (causal) change of $JPALswitch_{ct}$ from one to zero. By the definition of ceteris paribus, the distribution of U_{jct} is unaffacted. In potential outcomes notation, we write

$$D_{jct} = JPALswitch_{ct} D_{jct}^{JPALswitch=1} + (1 - JPALswitch_{ct}) D_{jct}^{JPALswitch=0}.$$

In another parallel trend assumption, we assume the effect of time on both types of results is the same regardless of whether a context ever receives a JPAL-IPA office. This means that the effect of time on internal selection bias (the treatment effect in this case) is identified directly through a DID approach alone (without needing an instrument), but the initial level of internal selection bias is not. Finally, we make two assumptions analogous to the Imbens and Angrist (1994) assumptions. First, we assume JPAL-IPA entry weakly pushes researchers towards the experimental design. No research teams switch to the observational design because of JPAL-IPA entry. Second, JPAL-IPA entry has no direct impact on potential results under either research design: it changes average results only through its effect on researchers' choice of design.

We make these assumptions formal in Assumption 6 below.

Assumption 6. *DD-IV for research design choice.*

1. *Common trends in the treatment (experimental design) probability without the policy (JPAL-IPA entry):*

$$P(D_{jc1}^{JPALswitch=0} = e | everJPAL_c) - P(D_{jc0}^{JPALswitch=0} = e | everJPAL_c).$$

   does not depend on $everJPAL_c$.

2. *Common trends for potential outcomes (results):*

$$E[R_{jc1}^d | everJPAL_c] - E[R_{jc0}^d | everJPAL_c]$$

   does not depend on $everJPAL_c$.

3. *Monotonicity. The effect of the policy (an operational JPAL-IPA office) on choosing the treatment (experimental research design) is described by:*

$$\mathbb{1}\{D_{jc1}^{JPALswitch=1} = e\} \geq \mathbb{1}\{D_{jc1}^{JPALswitch=0} = e\} \; \forall c : everJPAL_c = 1.$$

4. *Exclusion. The policy only affects the potential outcomes (results) through its effect on choice of experimental design. That is,*

$$R_{jct}^{d,JPALswitch} = R_{jct}^d.$$

Assumption 6.1 is commonly made by applied researchers employing DD-IV designs, like Duflo (2001). Assumptions 6.4 and 6.3 are natural extensions of Imbens and Angrist

(1994) to this setting. Assumption 6.2 merits some additional discussion. In Proposition 3 below, we will show that Assumption 6 identifies the average treatment effect for studies which change from the observational design to experimental as a result of JPAL-IPA entry. The key step in the proof is to reconstruct the expected observed study result in the counterfactual case where JPAL-IPA did not enter any context in period 1. We refer to this quantity as

$$E[R_{jc1}|fix(JPALswitch_{c1}) = 0, everJPAL_c = 1], \tag{13}$$

where we have used the $fix(\cdot)$ notation from Heckman and Pinto (2014) to emphasize that (13) is a hypothetical object to be identified, like the expectation of an unobserved potential outcome. By the law of iterated expectations, we can decompose (13) into a component involving the counterfactual expected observational result and the counterfactual expected experimental result:

$$\begin{aligned}
&E[R_{jc1}|fix(JPALswitch_{c1}) = 0, everJPAL_c = 1] \\
&= E[R^o_{jc1}|D^{JPALswitch=0}_{jc1} = o, everJPAL_c = 1]P(D^{JPALswitch=0}_{jc1} = o|everJPAL_c = 1) \\
&\quad + E[R^e_{jc1}|D^{JPALswitch=0}_{jc1} = e, everJPAL_c = 1]P(D^{JPALswitch=0}_{jc1} = e|everJPAL_c = 1) \\
&= (\bar{R}^o_{01} + \bar{R}^o_{10} - \bar{R}^o_{00})(1 - (P_{01} + P_{10} - P_{00})) + (\bar{R}^e_{01} + \bar{R}^e_{10} - \bar{R}^e_{00})(P_{01} + P_{10} - P_{00}),
\end{aligned}$$

where $\bar{R}^d_{t,everJPAL} = E[R_{jct}|D_{jct} = d, everJPAL_c, t]$ and $P_{t,everJPAL} = P(D = e|everJPAL_c, t)$. The parallel trends in observational and experimental results Assumption, 6.2, allows us to construct $E[R^o_{jc1}|D^{JPALswitch=0}_{jc1} = o, everJPAL_c = 1]$ and $E[R^o_{jc1}|D^{JPALswitch=0}_{jc1} = o, everJPAL_c = 1]$ by adding the design-specific trend in contexts which never got a JPAL-IPA office to the initial value in places that did. One could in principle weaken this assumption to allow for differential selection of studies into the experimental design across contexts to affect result trends, likely at the cost of point-identification, or run placebo checks with more pre-treatment periods.[12] We now state our main nonparametric identification result for DD-IV.

Proposition 3. *Suppose Assumption 6 holds. Then the negative internal selection bias (the average treatment effect of choosing the experimental design) for complying studies*

[12] As documented in de Chaisemartin and D'Haultfoeuille (2018), it is common in the applied literature to assume parallel trends on $E[R_{jct}|fix(JPALswitch_{c1}) = 0, everJPAL_c]$ directly. However, this is hard to justify when $P_{01} \neq P_{00}$ and one simultaneously invokes Assumption 6.1 because it would require non-parallel trends in the potential results to offset the assumed difference in $P(D^{JPALswitch=0}_{jc1} = o|everJPAL_c)$ by $everJPAL_c$.

is given by the following Wald ratio.

$$\begin{aligned}
&E[R^e_{jc1} - R^o_{jc1} | everJPAL_c, D^{JPALswitch=1}_{jc1} \neq D^{JPALswitch=0}_{jc1}] \\
&= \frac{[P_{11}\bar{R}^e_{11} + (1 - P_{11})\bar{R}^o_{11}] - [P_{CF}\bar{R}^e_{CF} + (1 - P_{CF})\bar{R}^o_{CF}]}{(P_{11} - P_{01}) - (P_{10} - P_{00})}, \qquad (14)
\end{aligned}$$

where the CF subscript refers to "counterfactual" and

$$\begin{aligned}
\bar{R}^d_{CF} &= \bar{R}^d_{01} + \bar{R}^d_{10} - \bar{R}^d_{00}, \\
P_{CF} &= P_{01} + P_{01} - P_{00}.
\end{aligned}$$

Proof. See Appendix A.1. □

The numerator of Equation (14) looks like that of a standard Wald ratio, with the second term in square brackets equal to $E[R_{jc1}|fix(JPALswitch_{c1}) = 0, everJPAL_c = 1]$. The denominator is the difference-in-differences estimand applied to the probability of choosing the experimental design, which is typical in applied work, as documented in de Chaisemartin and D'Haultfoeuille (2018). We believe this formulation, in identifying the average effect of the treatment (experimental research design) for units (studies) induced by the policy (JPAL-IPA entry) to adopt the experimental research design in period 1, captures the spirit of what many applied researchers intend when using DD-IV.

Comparison to de Chaisemartin and D'Haultfoeuille (2018)

de Chaisemartin and D'Haultfoeuille (2018) take a fundamentally different approach from ours, and have a different estimand. Starting with the estimand, when applied to our setting de Chaisemartin and D'Haultfoeuille (2018) target the average treatment effect (negative average internal selection bias) for *all* units (studies) in contexts where JPAL-IPA entered in period 1 who would not have taken the treatment (experimental research design) in period 0. In contrast, our estimand excludes any unit in areas where JPAL-IPA entered who takes the treatment (chooses the experimental design) for reasons other than the establishment of the JPAL-IPA office. Their approach involves a weaker version of our Assumption 6.2, where trends in outcomes (results) are common only for the potential outcome corresponding to units' treatment (research design) choice in period 0, but at the cost of partially-identifying their estimand when the comparison ($everJPAL_c = 0$) group's fraction treated changes over time. Their preferred specification makes use of comparison groups where the fraction treated does not change over time, while for us trends in the treatment rate in the comparison group are useful for identifying the fraction of complying studies.

4.2 Parametric Implementation

Moving towards estimation, we now fully specify a parametric model which reproduces the Wald ratio from Proposition 3. We start with a first stage which embeds Assumption 6.1:

$$\begin{aligned}
&D_{jct} = e \iff \pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 t * everJPAL_c \geq U_{jct}, \\
&U_{jct} \sim \text{ Uniform(0,1)}, \\
&U_{jct} \perp\!\!\!\perp t, everJPAL_c. \qquad (15)
\end{aligned}$$

We specify the potential result equation for designs of type d as:

$$R^d_{jct} = \alpha_d + \beta_{1d} t + \beta_{2d} everJPAL_c + \epsilon^d_{jct}. \qquad (16)$$

We now introduce endogeneity through the two-sided version of the Olsen (1980) control function estimator analyzed in Kline and Walters (2019), specifying the residual in the study result equation as:

$$\epsilon^d_{jct} = \gamma_d (U_{jct} - 1/2) + \xi^d_{jct} \qquad (17)$$

where $\gamma_d(U_{jct} - 1/2)$ represents the conditional expectation of ϵ^d_{jct} given study jct's unobserved cost of the RCT research design. We assume that ξ_{jct} is mean-independent of time and the indicator for ever having a JPAL-IPA office in c.

Assumption 7. ***Transient shocks to potential results.***

$$E[\xi^d_{jct} | everJPAL_c, t] = 0 \textit{ for } d \in \{e, o\}.$$

Combining (16) and (17) the observed conditional expectation function for results of study type d is

$$E[R_{jct}|D_{jct} = d, everJPAL_c, t] = \alpha_d + \gamma_d E[U_{jct} - 1/2 | D_{jct} = d, everJPAL_c, t] + \beta_{1d} t + \beta_{2d} everJPAL_c \qquad (18)$$

Examining the conditional expectation function for experimental studies we make use of the Uniform(0,1) assumption for U_{jct} to write $E[R_{jct}|D_{jct} = e, everJPAL_c, t]$ as a function of observed quantities:

$$\begin{aligned}
&E[R_{jct}|D_{jct} = e, everJPAL_c, t] \\
&= \alpha_e + \gamma_e E[U_{jct} - 1/2 | U_{jct} \leq P(D = e | everJPAL_c, t)] + \beta_{1e} t + \beta_{2d} everJPAL_c \\
&= \alpha_e + \gamma_e \frac{P(D = e | everJPAL_c, t) - 1}{2} + \beta_{1e} t + \beta_{2e} everJPAL_c. \qquad (19)
\end{aligned}$$

To see how each parameter in (19) is identified, we now examine the structure of the conditional expectations for the four groups indexed by $(everJPAL_c, t)$. We can write

$$\bar{R}^e_{00} = \alpha_e + \gamma_e \frac{P_{00} - 1}{2}, \tag{20}$$

$$\bar{R}^e_{10} = \alpha_e + \gamma_e \frac{P_{10} - 1}{2} + \beta_{1e}, \tag{21}$$

$$\bar{R}^e_{01} = \alpha_e + \gamma_e \frac{P_{01} - 1}{2} + \beta_{2e}, \tag{22}$$

$$\bar{R}^e_{11} = \alpha_e + \gamma_e \frac{P_{11} - 1}{2} + \beta_{1e} + \beta_{2e}. \tag{23}$$

Taking differences over time for each of the $everJPAL_c$ groups:

$$\bar{R}^e_{10} - \bar{R}^e_{00} = \gamma_e \left(\frac{P_{10} - P_{00}}{2} \right) + \beta_{1e}$$

and

$$\bar{R}^e_{11} - \bar{R}^e_{01} = \gamma_e \left(\frac{P_{11} - P_{01}}{2} \right) + \beta_{1e},$$

so that the difference-in-differences identifies γ_e according to:

$$(\bar{R}^e_{11} - \bar{R}^e_{01}) - (\bar{R}^e_{10} - \bar{R}^e_{00}) = \frac{\gamma_e}{2}((P_{11} - P_{01}) - (P_{10} - P_{00})).$$

With γ_e identified, equation (20) identifies α_e:

$$\alpha_e = \frac{\bar{R}^e_{00}(P_{11} - P_{01} - (P_{10} - P_{00})) - (\bar{R}^e_{11} - \bar{R}^e_{01} - (\bar{R}^e_{10} - \bar{R}^e_{00}))(P_{00} - 1)}{P_{11} - P_{01} - (P_{10} - P_{00})}. \tag{24}$$

We can then go back and identify β_{1e} and β_{2e} from equations (21) and (22) respectively. The conditional distributions of ξ_{jct} can be identified from the distributions of R_{jct} conditional on t and $everJPAL_c$ with $D = d$ after subtracting the relevant objects defining their conditional mean. α_o and γ_o can be identified analogously, noting that:

$$\begin{aligned} &E[R_{jct}|D_{jct} = o, everJPAL_c, t] \\ &= \alpha_o + \gamma_o E[U_{jct} - 1/2 | U_{jct} > P(D = e|everJPAL_c, t)] + \beta_{1o} t + \beta_{2o} everJPAL_c \\ &= \alpha_o + \gamma_o \frac{P(D = e|everJPAL_c, t)}{2} + \beta_{1o} t + \beta_{2o} everJPAL_c. \end{aligned} \tag{25}$$

So

$$\gamma_o = -2 \frac{(\bar{R}^o_{11} - \bar{R}^o_{01}) - (\bar{R}^o_{10} - \bar{R}^o_{00})}{(P_{11} - P_{01}) - (P_{10} - P_{00})}$$

and

$$\alpha_o = \frac{\bar{R}^o_{00}(P_{11} - P_{01} - (P_{10} - P_{00})) + (1 - P_{00})(\bar{R}^o_{11} - \bar{R}^o_{01} - (\bar{R}^o_{10} - \bar{R}^o_{00}))}{P_{11} - P_{01} - (P_{10} - P_{00}))}.$$

We can now analyze the LATE in this model. The LATE applies to studies conducted in period 1 that would not have chosen the experimental research design had a JPAL-IPA office not been available in their context. Returning to the first stage equation, (15), these studies must have

$$\pi_0 + \pi_1 + \pi_2 < U_{jct}$$

but

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 \geq U_{jct}.$$

Since this is a linear probability model,

$$\begin{aligned}\pi_0 &= P_{00},\\ \pi_1 &= P_{10} - P_{00},\\ \pi_2 &= P_{01} - P_{00}\end{aligned}$$

so that $\pi_0 + \pi_1 + \pi_2 = P_{01} + P_{01} - P_{00}$, which is the counterfactual treatment probability under parallel trends, and $\pi_0 + \pi_1 + \pi_2 + \pi_3 = P_{11}$.

Bringing together the results above, the LATE is:

$$\begin{aligned}&E[R^e_{jct} - R^o_{jct} | P_{CF} < U_{jct} \leq P_{11}, everJPAL_c = 1, t = 1]\\ &= (\alpha_e - \alpha_o) + (\beta_{1e} - \beta_{1o}) + (\beta_{2e} - \beta_{2o}) + (\gamma_e - \gamma_o)E[U_{jct} - 1/2 | P_{CF} < U_{jct} \leq P_{11}].\end{aligned}$$

Combining the identification results based on Equations (19) and (25) yields a LATE equivalent to the one in Proposition 3:[13]

$$LATE = \frac{[P_{11}\bar{R}^e_{11} + (1 - P_{11})\bar{R}^o_{11}] - [P_{CF}\bar{R}^e_{CF} + (1 - P_{CF})\bar{R}^o_{CF}]}{(P_{11} - P_{01}) - (P_{10} - P_{00})}. \tag{26}$$

Our Bayesian likelihood-based implementation, however, will not support the linear probability model for the first stage implied by the Uniform(0,1) assumption for U_{jct} because it can produce predicted probabilities outside the unit interval. In Kline and Walters (2019), the authors showed that the control function $\gamma_d(U_{jct} - 1/2)$ in (17)

[13] See Appendix A.3 for the proof.

could be replaced by

$$\gamma_d(J(U_{jct}) - E[J(U_{jct})])$$

where $J(\cdot)$ is any strictly increasing continuous function, and numerical equivalence between control function methods and the standard Wald Estimator in binary-treatment, binary-instrument LATE estimands is maintained. Their framework also accomodates any first-stage structure which can perfectly fit the observed treatment probabilities. We show that in the DD-IV case, however, forms for $J(\cdot)$ which allow for Bayesian likelihood-based estimation produce slightly different LATEs. As in discussions over the correct functional form for parallel trend assumptions in other DID contexts (see e.g. Kahn-Lang and Lang (2020)), we will argue in the remainder of the paper that the alternative formulations of the LATE implied by these models are just as reasonable and equally relevant for research and policy.

Consider the following limited dependent variable difference-in-difference model discussed in Blundell and Costa Dias (2009):

$$\begin{aligned}&D_{jct} = e \iff \mathbb{1}\{\pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 everJPAL_c * t \geq \Phi^{-1}(U_{jct})\},\\ &U_{jct} \perp\!\!\!\perp t, everJPAL_c\end{aligned}$$

where we are using the quantile function of a standard Normal variable, $\Phi^{-1}(\cdot)$, as $J(\cdot)$. Under this setup, the counterfactual experimental design probability is

$$P_{CF} = \Phi(\pi_0 + \pi_1 t + \pi_2 everJPAL_c).$$

π_0, π_1, π_2 and π_3 are identifed from the following system, noting that $\Phi(\cdot)$ is invertible:

$$\begin{aligned}P_{00} &= \Phi(\pi_0)\\ P_{10} &= \Phi(\pi_0 + \pi_1)\\ P_{01} &= \Phi(\pi_0 + \pi_2)\\ P_{00} &= \Phi(\pi_0 + \pi_1 + \pi_2 + \pi_3).\end{aligned}$$

So the model does perfectly fit the observed choice probabilities.

With $\Phi^{-1}(\cdot)$ in place of $J(\cdot)$ the model is exactly the Bjorklund and Moffitt (1987) two-sided version of Heckman (1979), modified with a difference-in-difference-based first stage. To make notation more compact, let $V_{jct} = \Phi^{-1}(U_{jct})$. We can then examine the observed conditional expectation function for the experimental studies:

$$
\begin{aligned}
&E[R_{jct}|D_{jct} = e, everJPAL_c, t] \\
&= \alpha_e + \gamma_e E[V_{jct}|V_{\leq}\pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 JPALswitch_{ct}] + \beta_{1d} t + \beta_{2d} everJPAL_c \\
&= \alpha_e - \gamma_e \frac{\phi(\pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 JPALswitch_{ct})}{\Phi(\pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 JPALswitch_{ct})} + \beta_{1d} t + \beta_{2d} everJPAL_c
\end{aligned}
$$

where $\phi(\cdot)$ indicates the standard normal density function and $\Phi(\cdot)$ the standard normal CDF.

Since the terms multiplying γ_e have already been identifed, γ_e is identified using the same difference-in-difference strategy as when we worked with U_{jct}. We then proceed to identify α_e, β_{1e}, and β_{2e}. α_o and γ_o are identified analogously, noting that:

$$
\begin{aligned}
&E[R_{jct}|D_{jct} = o, everJPAL_c, t] \\
&= \alpha_o + \gamma_o \frac{\phi(\pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 JPALswitch_{ct})}{1 - \Phi(\pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 JPALswitch_{ct})} + \beta_{1o} t + \beta_{2o} everJPAL_c.
\end{aligned}
$$

However, this model does not yield a Wald ratio representation, as stated in the following proposition.

Proposition 4. *Suppose a model of the form*

$$
\begin{aligned}
&D_{jct} = e \iff \mathbb{1}\{\pi_0 + \pi_1 t + \pi_2 JPAL_c + \pi_3 everJPAL_c * t \geq J(U_{jct})\} \\
&E[R^d_{jct}|everJPAL, t, U] = \alpha_d + \beta_{1d} t + \beta_{2d} everJPAL + \gamma_d (J(U_{jct}) - \mu_J) \\
&U_{jct} \perp\!\!\!\perp t, everJPAL_c
\end{aligned}
$$

yields functions

$$
\begin{aligned}
\lambda_e(p) = E[J(U_{jct}) - E[J(U_{jct})]|U_{jct} \leq p] \\
\lambda_o(p) = E[J(U_{jct}) - E[J(U_{jct})]|U_{jct} > p]
\end{aligned}
$$

which are non-linear in p. Then the LATE does not take the Wald ratio form from Equation (26)*. Instead, the LATE is given by:*

$$
\begin{aligned}
LATE = &(\bar{R}^e_{01} + \bar{R}^e_{10} - \bar{R}^e_{00}) - (\bar{R}^o_{01} + \bar{R}^o_{10} + \bar{R}^o_{00}) \\
&- \left(\frac{DID^e_R}{DID^e_\lambda} (\lambda_e(P_{01}) + \lambda_e(P_{10}) - \lambda_e(P_{00})) - \frac{DID^o_R}{DID^o_\lambda} (\lambda_o(P_{01}) + \lambda_o(P_{10}) - \lambda_o(P_{00})) \right) \\
&+ \left(\frac{DID^e_R}{DID^e_\lambda} - \frac{DID^o_R}{DID^o_\lambda} \right) \left(\frac{P_{11}\lambda_e(P_{11})) - (P_{01} + P_{10} - P_{00})\lambda_e(P_{01} + P_{10} + P_{00})}{DID_P} \right).
\end{aligned} \tag{27}
$$

where

$$
\begin{aligned}
DID_\lambda^d &= \lambda_d(P_{11}) - \lambda_d(P_{01}) - \lambda_d(P_{10}) + \lambda_d(P_{00}) \\
DID_P &= P_{11} - J^{-1}(\pi_0 + \pi_1 + \pi_2) \\
DID_R^d &= \bar{R}_{11}^d - \bar{R}_{01}^d - \bar{R}_{10}^d + \bar{R}_{00}^d.
\end{aligned}
$$

Proof. See Appendix A.4. □

In the model with $J(\cdot) = \Phi^{-1}(\cdot)$ we can write

$$
\begin{aligned}
\lambda_e(p) = E[V_{jct}|V_{jct} \leq \Phi^{-1}(p)] &= -\frac{\phi(\Phi^{-1}(p))}{p} \\
\lambda_o(p) = E[V_{jct}|V_{jct} > \Phi^{-1}(p)] &= \frac{\phi(\Phi^{-1}(p))}{1-p}
\end{aligned}
$$

where the second equality on each line follows from the properties of the standard normal distribution. So the $\lambda_d(p)$ functions are nonlinear and the conditions for Proposition 4 hold. Despite the fact that the Wald ratio form does not hold here, all functions in Equation (27) are known; should it be of interest, one could quantify by how much the LATE from Equation 26 differs from that in Equation (27). The more fundamental question however is what the correct specification is for the parallel trends assumption in DID analysis, as raised by, for example, Kahn-Lang and Lang (2020). The conventional LATE is underpinned by a linear probability model in the selection equation; the alternative LATE here is underpinned by a Probit selection equation. The two may diverge in cases where the relevant changes in probability occur outside of the range in which the Probit is approximately linear, yet this is exactly the region in which one requires the nonlinearity of the link function (Probit, Logit, or any other tractable option) to ensure the probability bounds are respected. We discuss this point and additional rationale for our choice to focus on the Probit-LATE in more depth in the following sections outlining our specific Bayesian implementation.

5 Bayesian Implementation

Our qualitative research combined with the project databases shared from JPAL and IPA suggest "established presence" occurs after 3 RCTs are performed by JPAL or IPA in context c; we denote this year y_c for context c. Denote the time period tracker covariate t with some abuse of notation. Under a conventional parallel trends assumption, t and $everJPAL_c$ may be directly correlated with (R_{jct}^e, R_{jct}^o), but the indicator for post-establishment of JPAL-IPA ($\mathbb{1}\{t > y_c\} * everJPAL_c$) is exogenous. Note that the IV switches on *after* the first three studies are conducted; these studies are excluded because they are part of the endogenous motivation for JPAL-IPA entry. We have access to all of this information from JPAL and IPA, and this set-up allows us to invoke the results from Sections 2 and 4.

One potential complication with the structure above is that the parallel trends assumption may not hold; in our applications, as in much of meta-analysis, we typically do not have enough studies to check pre-trends. An alternative assumption we will make use of in Section 5.2 is that JPAL-IPA entry itself may be correlated to results (R^e_{jct}, R^o_{jct}) to some degree even controlling for t and $everJPAL_c$. In this situation, is still possible to use it as an instrument within an estimation approach that does not rely on strict exogeneity: we can use the plausibly exogenous instrument framework from Conley et al. (2012). This approach uses Bayesian priors to suggest that the direct correlation between JPAL entry and changes in time trends in R_{jct} is *small* relative to the direct effect of JPAL entry on D_{jct}. The resulting model nests the usual strict exogeneity IV model assumption as a special case. The intuition and technical details of this approach in the context of our parametric evidence aggregation exercise are described in the following section.

Parametric estimation is practical for meta-analysis for several reasons, chief among them that the number of studies J is typically too small for nonparametric estimation. Another reason to use parametric models is that the argument in Section 2 relied on direct observation of study results R_{jct}, but in practice we observe $\hat{R}_{jct}$ with sampling error se_{jct}. We must deconvolve se_{jct} from true variance in effects to accurately estimate the site selection bias which is captured by the correlation between TE_{jct} and the latent cost V_{jct}. The need to perform this deconvolution points to the use of a hierarchical likelihood model, layering the sampling uncertainty on top of the conventional selection model structure (Rubin (1981), Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2014)). As these models can be challenging to estimate using frequentist methods, and our instrumental variable from Section 3 is only plausibly exogenous anyway, we pursue a Bayesian implementation building on work from Meager (2019) and Conley et al. (2012) and estimating a LATE of the general form given by Equation (27).

5.1 Hierarchical Selection Models

To build our parametric model, we start with the theoretical set-up laid out in Sections 2 and 4, which implies a particular structure on the conditional means of the study results for each study design type. Recall that for experimental studies ($D_{jct} = e$) the results are the treatment effects, so $R^e_{jct} = TE_{jct}$, while for observational studies ($D_{jct} = o$) the results are affected by internal selection bias, so $R^e_{jct} = TE_{jct} + SB_{jct}$. The design selection equation consists of an observed net cost of performing an RCT, which depends linearly on covariates X_{Djct}, and the latent net cost V_{jct}. As in Section

2, the selection equation is:

$$D_{jct} = \mathbb{1}\{X_{Djct}\pi + V_{jct} < 0\}e + \mathbb{1}\{X_{Djct}\pi + V_{jct} \geq 0\}o. \tag{28}$$

Here, as per usual, the threshold in the latent space is a normalization chosen without loss of generality. If one could observe simultaneous experiments and observational studies, one would see results on average depending on contextual factors X_{jct} which include t and $everJPAL_c$ though it seems possible that the dependence might differ across study types due to the presence of internal selection bias in the observational studies. This means that both the parameters governing the relationship to observed factors differ across study types, as do their structural errors:

$$\begin{aligned} R^e_{jct} &= TE_{jct} = X_{jct}\beta_e + \epsilon^e_{jct} \\ R^o_{jct} &= TE_{jct} + SB_{jct} = X_{jct}\beta_o + \epsilon^o_{jct} \end{aligned} \tag{29}$$

As we only observe each type of result according to the partially-latent selection procedure, then for each type of study, the observed mean conditional on relevant latent experimental net cost V_{jct} is:

$$\begin{aligned} E[TE_{jct}|V_{jct}] &= X_{jct}\beta_e + E[\epsilon^e_{jct}|V_{jct} < -X_{Djct}\pi] \\ E[TE_{jct} + SB_{jct}|V_{jct}] &= X_{jct}\beta_o + E[\epsilon^o_{jct}|V_{jct} \geq -X_{Djct}\pi]. \end{aligned} \tag{30}$$

The fact that these expected selected regression errors may not be mean zero captures the notion that there is selection on unobservables in the errors $(\epsilon^e_{jct}, \epsilon^o_{jct})$ correlated to the latent cost of performing an RCT $(V_{jct},)$ which is why even the average of the RCTs we do see in the data is not an unbiased estimate of the average treatment effect in the population. In this set-up, having an instrument Z_{Djct} in X_{Djct} excluded from X_{jct} allows us to invoke our Proposition 1 and nonparametrically debias these means, for the exact same reason that having an instrument for treatment assignment allows one to estimate the LATE in an RCT with partial compliance. Note that our construction does not assume that the observational studies are randomly drawn from the population of potential evaluations; they, too, are a subset of the total set of interventions, observed only when the latent cost variable realises above a threshhold.

Due to the impracticality of the nonparametric approach, as noted above, we now make a parametric assumption – but we do so in a way that allows us to retain an estimand that corresponds to the general LATE given in Proposition 4 for the compliers of the study design selection instrument Z_{Djct}. Specifically, we assume $(\epsilon^e_{jct}, \epsilon^o_{jct}, V_{jct})$ are trivariate Gaussian (with the usual normalization on the latent error V_{jct} without loss of generality) giving us:

$$\begin{bmatrix} V_{jct} \\ \epsilon^e_{jct} \\ \epsilon^o_{jct} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_e\sigma_e & \rho_o\sigma_o \\ \rho_e\sigma_e & \sigma_e^2 & \rho_{eo}\sigma_e\sigma_o \\ \rho_o\sigma_o & \rho_{eo}\sigma_e\sigma_o & \sigma_o^2 \end{bmatrix} \right). \tag{31}$$

This parametric structure aligns with the Gaussian random-effects meta-analysis models of Meager (2019), Vivalt (2020) and Andrews and Kasy (2019), and allows us perform inference on the general LATE from Equation (27) for compliers of the study design instrument. This will not be exactly the Wald ratio from Proposition 3, as discussed in Section 4, because the conventional LATE is underpinned by a linear probability model on the selection equation. We instead follow Bjorklund and Moffitt (1987) and use a Probit for two reasons: first, the linear probability model (LPM) violates support constraints on the probabilities which can lead to impossible inferences, and second, the region in which the LPM behaves well is also the region in which the Probit is relatively linear, so the wedge between the Probit LATE and the conventional Wald ratio from Proposition 3 should be small if the latter is valid. Therefore, we can aspire to robustness beyond the parametric functional form used to gain traction on the estimation in small samples.

The nonparametric identification set-up combined with the additional trivariate Gaussian error structure leads to a generalized Heckman log likelihood kernel as in Toomet and Henningsen (2008):

$$\sum_{j:D_{jct}=o} \left(-log\sigma_o - \frac{1}{2}\left(\frac{R_{jct} - X_{jct}\beta_o}{\sigma_o}\right)^2 + log\Phi\left(-\frac{X_{Djct}\pi_s + \frac{\rho_o}{\sigma_o}(R_{jct} - X_{jct}\beta_o)}{\sqrt{1-\rho_o^2}} \right) \right)$$
$$+ \sum_{j:D_{jct}=e} \left(-log\sigma_e - \frac{1}{2}\left(\frac{R_{jct} - X_{jct}\beta_e}{\sigma_e}\right)^2 + log\Phi\left(\frac{X_{Djct}\pi_s + \frac{\rho_e}{\sigma_e}(R_{jct} - X_{jct}\beta_e)}{\sqrt{1-\rho_e^2}} \right) \right) \tag{32}$$

Note the disappearance of the ρ_{eo} parameter: we do not need to estimate this parameter to debias the means of each type of study, which is fortunate as we cannot estimate it (this occurs because we never observe both an experimental and observational study with the same exact contextual features at the same moment in time). The log likelihood kernel above contains parameters which are all identified (Toomet and Henningsen (2008)). As mentioned in Section 2.3, ρ_{eo} is exactly the parameter which would be identified if imperfect compliance in the experiments meant we could see both an experimental and an observational result for the RCTs. In that case the model in Equation (31) has a Gaussian copula with Gaussian marginals for R^e_{jct}, R^o_{jct}, and V_{jct}.

Parallels to the classical Heckman (1979) sample selection model provide further intuition in our setting. Heckman considers an outcome distributed Normally around a linear conditional mean function, which is only observed when some latent Normal error - which is correlated with that outcome - crosses a threshold in the latent space, according to a selection equation. The true conditional mean can be identified either from the functional form of the model or nonparametrically if there is an instrument in the selection equation omitted from the outcome equation. In our set-up, the experimental study results R^e_{jct} are this outcome variable, and experimental design $D_{jct} = e$ is determined by this latent selection process. We in fact do observe the nonselected (observational) study results as in two-sided Heckman models beginning with Heckman (1976), albeit observing effects with some internal bias. But now we can estimate that internal bias on average, because we use the Heckman-style likelihood to first estimate the site selection bias separating the two types of studies in the literature.

To complete our likelihood model, we now need to address the fact that we do not observe study results R_{jct} directly, but rather we observe estimates $\hat{R}_{jct}$ with some sampling error se_{jct}, and this is true whether one observes only the reported results or the full microdata, since one never observes the true effects. Failing to observe the true effects means failing to observe the true $(\epsilon^o_{jct}, \epsilon^e_{jct})$ draws even conditional on V_{jct}; this means we cannot identify $(\sigma_e, \sigma_o, \rho_e, \rho_o)$, and that is a problem because these parameters characterize the site selection bias. On average, we expect convolving the results with random sampling variance to increase the observed variation in effects and thus attenuate the correlation between effects and study type, though in practice the reverse can occur depending on patterns in sampling variability. In any case, we must deconvolve the sampling error from the genuine variation in order to estimate $(\sigma_e, \sigma_o, \rho_e, \rho_o)$. This is the job for which hierarchical models are designed (Rubin (1981), Gelman et al. (2014), Meager (2019)).

We therefore modify the conventional likelihood above by introducing a hierarchy of variation: first, there is genuine variation in results across settings described by (σ_e, σ_o) captured in Equation (31). Then, there is additional sampling variation within each study, se_{jct}. Studies in economics typically use estimators that they feel comfortable assuming are approximately Normally distributed around the true parameters, and calculate confidence intervals and t-statistics or p-values based on that assumption. Hence, it is no additional structure beyond that already used in the underlying papers to specify the following additional layer on top of the likelihood in Equation (32):

$$\hat{R}_{jct} \sim \mathcal{N}(R_{jct}, se_{jct}). \tag{33}$$

Estimation of the entire hierarchical likelihood can be challenging, however, due to correlations in the uncertainty surrounding $(\sigma_e, \sigma_o, \rho_e, \rho_o)$ and $\{R_{jct}\}_{j=1}^{J}$, as the latter is now a set of J parameters. We discuss the performance of a Bayesian estimation procedure for this model with specific prior choice in Section 5.3.

5.2 Selection with Plausibly Exogenous IV

We now present the specific regression equations we use in order to implement our plausibly exogenous DD-IV strategy, as needed to use our chosen instrument described in Section 3. Recall that in our discussion, the time tracker (t) and binary indicator of JPAL or IPA ever establishing a presence in a country ($everJPAL_c$) were endogenous to the study results, but the variable indicating the "switching year" y_c for JPAL presence ($\mathbb{1}\{t > y_c\} * everJPAL_c$) was plausibly exogenous. To see how we implement this, first consider the DD-IV strategy laid out in Section 4. Here, the "JPALswitch" variable ($\mathbb{1}\{t > y_c\} * everJPAL_c$) is the excluded instrument for study design selection, denoted Z_{Djct} in all previous sections. This would lead to the following regression equations:

$$\begin{aligned} D_{jct} &= \mathbb{1}\{\pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 \mathbb{1}\{t > y_c\} * everJPAL + \pi_4 X_{jct} > V_{jct}\} \\ TE_{jct} &= \alpha_o + \beta_{1o} t + \beta_{2o} everJPAL_c + \beta_{3o} X_{jct} + \epsilon^e_{jct} \\ TE_{jct} + SB_{jct} &= \alpha_e + \beta_{1e} t + \beta_{2e} everJPAL_c + \beta_{3e} X_{jct} + \epsilon^o_{jct} \end{aligned} \tag{34}$$

We then make the trivariate Gaussian assumption described above in Section 5.1. Using the results in Section 4, the LATE associated with this model set-up for can be computed using the properties of Gaussians:

$$\begin{aligned} LATE =& (\alpha_e - \alpha_o) + (\beta_{1e} - \beta_{1o}) + (\beta_{2e} - \beta_{2o}) + (\beta_{3e} - \beta_{3o}) \\ &+ (\rho_e \sigma_e - \rho_o \sigma_o) \frac{\phi(\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4) - \phi(\pi_0 + \pi_1 + \pi_2 + \pi_4)}{\Phi(\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4) - \Phi(\pi_0 + \pi_1 + \pi_2 + \pi_4)} \end{aligned}$$

The fraction at the end of the expression gives the expected value of the latent cost V_{jct} conditional on it lying between the bounds defined by the difference in the Probit surface from switching the instrument on and off (thus, omitting π_3 or not). Note that in the Gaussian case $\rho_d \sigma_d = \gamma_d$ from Section 4. Note also that this formulation above implicitly takes the covariates X_{jct} to be meaningful to include in the LATE; if some of these are categorical, one might want to compute multiple LATEs for the different categories, by omitting some or all of the β_3 parameters (and correspondingly π_4) in the formula above.

The "plausibly exogenous" IV strategy further relaxes the structure above by allowing the instrument to enter the outcome equations, albeit in a restricted fashion that still permits identification (Conley et al. (2012)). In our case, these restrictions form additional components of the model in Section 5.1: we replace the strict exogeneity assumption with a milder "expected exogeneity" structure. We allow the instrument $\mathbb{1}\{t > y_c\} * everJPAL_c$ to be present in the results equations but insist that its coefficient, δ_d, be in expectation small in magnitude and centered at zero. In our setting, with variables roughly on the standard Gaussian scale, this leads to the following structure:

$$
\begin{aligned}
D_{jct} &= \mathbb{1}\{\pi_0 + \pi_1 t + \pi_2 everJPAL_c + \pi_3 \mathbb{1}\{t > y_c\} * everJPAL + \pi_4 X_{jct} > V_{jct}\} \\
TE_{jct} &= \alpha_o + \beta_{1o} t + \beta_{2o} everJPAL_c + \delta_o \mathbb{1}\{t > y_c\} * everJPAL + \beta_{3o} X_{jct} + \epsilon^e_{jct} \\
TE_{jct} + SB_{jct} &= \alpha_e + \beta_{1e} t + \beta_{2e} everJPAL_c + \delta_e \mathbb{1}\{t > y\} * everJPAL + \beta_{3e} X_{jct} + \epsilon^o_{jct} \\
\gamma_e, \gamma_o &\sim \mathcal{N}(0, 0.1)
\end{aligned}
\tag{35}
$$

Here, our assumption that each $\delta_d \sim N(0, 0.1)$ replaces the stronger, strict exogeneity ($\delta_e = 0, \delta_o = 0$) in the conventional model in Equation (34); ours is an example of a model with more parametric equations but strictly weaker assumptions. The choice of variation in δs permitted around zero is made to the scale of the outcome data in the particular application; if a standard deviation of 0.1 units is not sufficiently small in a given application, the restriction can be tightened further.

5.3 Bayesian Model Performance

We now propose and evaluate a Bayesian estimation procedure for the likelihood described in Sections 5.1 and 5.2. We propose a set of priors for a problem of a given scaling, and we test the performance of the model on simulated literatures of experimental and observational studies. We examine performance of both the general Bayesian Hierarchical Selection model and the version with the plausibly exogenous instrument, and we focus on literatures with fewer than 50 studies, as is typical in applied microeconomics.

Bayesian estimation is particularly useful for hierarchical models on small numbers of studies J, as there may be tractability and overfitting issues with common frequentist approaches (see Meager (2019) for a deeper discussion). Moreover, the plausibly exogenous structure described in Section 5.2 is already inherently Bayesian, as the restriction that ex-ante the second-stage instrument parameters should be small in magnitude is expressed in what is essentially a prior: $\delta_e, \delta_o \sim \mathcal{N}(0, 0.1)$. We select the remaining priors based on a combination of statistical and economic concerns. Particularly notable are the priors placed on the (ρ_e, ρ_o) parameters, which capture the site selection bias and which can be challenging to estimate in small data sets, especially since the variable V_{jct} is itself inferred rather than observed, even with error. This motivates the use of Gaussian priors centered at 0 with standard deviation 0.3 truncated on $[-1, 1]$, which regularizes the estimates while still allowing extreme correlations if they are warranted by the data. Working with outcome data roughly on the standard scale (i.e. mean zero, unit variance), all other parameters are given standard Gaussian priors which are relatively weak for this scale, with the exception of the hyper-standard-deviations (σ_o, σ_e), where Gamma(2,2) offers more mass slightly away from 0. Combining these priors with the likelihoods in Sections 5.1 and 5.2 generates the full Bayesian Posterior.

We test these models on simulated data and find that the coverage of the central 50% and 95% posterior credible intervals is close to nominal even in small data sets, and even when priors are wrong. We consider $J = (50, 30, 15)$ for each model. Our simulations take the trivariate Gaussian error process as given, but all cases we test the models with incorrect priors. Specifically, the priors are as described above, but the true generating process for the trivariate error covariance matrix is governed by a Wishart distribution with a diagonal scale matrix of 3 degrees of freedom, truncated to avoid correlations larger than 0.8 in magnitude.[14] We therefore do not expect perfect coverage, but seek to understand the extent to which coverage is degraded by our incorrect priors in small samples. We fit the models using Hamiltonian Monte Carlo in Stan (for further details of this computational approach, see Appendix A of Meager (2019)). The results are in Table 1 and show that even in these small samples, with incorrect priors, the posterior coverage is close to nominal: the posterior 95% credible interval has approximately 90% coverage even for $J = 15$. The 50% credible interval is somewhat more degraded in the case of the plausibly exogenous model, where the priors have to do more heavy lifting, producing around 43% coverage on average relative to the fully exogenous Bayesian Heckman which has around 46% coverage on average for its nominal 50% interval. Given the incorrect priors and challenges of model fitting in this environment, we regard this as reasonable performance for small samples.[15]

6 Applications

We apply our Bayesian hierarchical selection model to two different literatures in development economics and find that it yields insights even on small literatures. We consider Conditional Cash Transfers (CCTs), a literature of 28 studies, and expanding Microcredit access, a literature of 12 studies. In the CCTs application, we find little evidence of site selection bias but substantial evidence of two different internal selection biases: observational studies of government-implemented programs have positive-selected treated groups while observational studies of NGO-implemented programs have negatively-selected treated groups. In the Microcredit application, we find stronger evidence of zero site selection bias and zero internal selection bias in the observational studies once a time trend is accounted for; this is because the hierarchical modelling structure substantially down-weights an imprecise outlying observational study. We also estimate the implied LATEs for the complier studies of our instrument using Equation (35). The

[14] This truncation avoids the local-to-unity problem in the correlation space.

[15] We interpret the rough or approximate values of the coverage because at 100 simulations for each row of the Table there is likely to be Monte Carlo error driving the exact figures.

LATE is large and positive for NGO-implemented CCTs, indicating negative internal selection bias, and zero for the Microcredit case, confirming the finding of substantial internal selection within observational studies in the CCTs case but not for Microcredit.

6.1 Conditional Cash Transfers

In this subsection, we apply the Bayesian hierarchical plausibly exogenous selection model from Section 5 to the literature on conditional cash transfers (CCTs). This intervention offers households in developing countries cash payments conditional on meeting certain requirements or targets, often related to children's school enrolment or healthcare attendance records. We use the data from Vivalt (2020), available on the AidGrade online database, and focus on studies which record the effect of conditional cash transfers on school enrolment. This leads to a total of 28 studies, 14 of which are RCTs and 14 of which are observational, shown in Table 2. AidGrade's database contains information on the study citations, effect sizes, confidence intervals, and research design.

In order to apply our identification approach, we then collected additional data on the country in which the study was done, and the year in which the program was implemented (*not* the publication year of the study), as this year is the relevant year by which it matters whether JPAL or IPA has an established presence in a region for our DD-IV strategy since this is also when researchers picked their study designs. From Vivalt (2020) we know that government involvement in studies predicts effects in this literature, so we also collect that variable. Finally, using the JPAL-IPA project databases shared with us, for each country c we then recover $everJPAL_c$ and year y_c at which JPAL-IPA switches on in that country; from this we can construct $JPALswitch_{jct} = \mathbb{1}\{t > y_c\} * everJPAL_c$.

We fit the model to this data set and find that there is a strong conditional correlation between study design and JPAL presence, as shown by the large positive coefficient on $JPALswitch_{jct}$ in the selection equation in columns 1 and 2 of Table 3. Although the coefficients are not precisely estimated, in part due to the small number of studies, both $JPALswitch_{jct}$ and $NonGov_{jct}$ both substantially predict selection into study type.[16] The model uses this information to estimate both the site selection bias captured by the ρ_d parameters and the internal selection bias captured by any difference in the conditional means of the two types of studies' outcome equations, in columns 4-6 of Table 3. Note in these columns that $JPALswitch_{jct}$ enters the outcome regressions, and it is slightly positive and covers zero, reflecting a broad alignment between the plausible exogeneity assumption and the data in this sample.

[16] We use $NonGov_{jct}$ as a covariate rather than Gov because a large majority of studies have $Gov = 1$, hence, the intercept would capture only a small minority of studies if Gov is entered directly as a covariate.

We find that while there is little evidence of site selection bias, there is substantial evidence of internal selection bias within observational studies. These results are visually displayed in Figure 1, and show that both ρ parameters are centered at zero, suggesting little to no site selection bias in this literature, although their posterior intervals are wide. In terms of internal selection bias, the intercept for observational studies is larger than for RCTs, suggesting positive selection. However, we also note that the coefficient on both $everJPAL_c$ and $NonGov_{jct}$ are substantially larger for RCTs, especially the latter, which is by far the largest coefficient in the entire results set. This provides evidence of two different internal selection bias dynamics: for studies with government involvement, the observational studies positively select households, but for studies without government involvement, the observational studies negatively select households. The coefficient on the non-government studies is much larger, leading to a LATE for that is very large and positive (see the last two rows of Table 3) for non-government studies only.

Finding no evidence of site selection bias in this literature is somewhat surprising, and the posterior intervals on ρ_e and ρ_o are wide with little updating away from the prior, so we seek further corroboration of this result in the data. The first suggestive evidence that there is unlikely to be major site selection bias in these studies is that the observed variation in RCT results is actually larger than the observed variation in observational results: that is, $\sigma_e > \sigma_o$ (Table 3). This is not the pattern one would expect if RCTs were subsampling a region of the distribution of true effects in a manner correlated with the value of the treatment effect.[17] The histogram of treatment effects split out by study type in Figure 2 confirms this variance differential in the raw studies, not just in the hierarchical model.

To understand how the inferences above are dependent on our sample and chosen procedure, we now investigate robustness of these results to changing both the data and the priors. We note the presence of a visually outlying RCT result in Figure 2 and seek to confirm that our results are not solely determined by this single study: the results of fitting the model with this study omitted are shown in Table 4 and Figure 3, and the main results are broadly preserved, though less precise; the non-governmental studies LATE in particular is both somewhat smaller and less precise, but still much larger than the LATE for the governmental studies.

We also note that for the site selection bias, the priors are particularly informative in the posterior – this was always a potential scenario in the small sample case, given the somewhat ambitious separation of variation implied by the DD-IV strategy. We note that the priors on the internal selection bias within the observational studies were

[17] It might arise if RCTs are selected according to variation, but pure selection on variation in TEs would not bias the ATE estimated from RCTs, and is therefore not the main concern that has been raised with regards to site selection bias in the previous literature.

also centered at zero and reasonably tight, so the fact that we update away to find bias along that aspect is informative. The relative evidence for the presence of internal selection bias within the observational studies is thus much stronger than any evidence about site selection bias. One might, however, be concerned that the regularization of the site selection bias to zero may be driving part of the finding on the internal selection bias (since this is a decomposition of differences, in some respects). To confirm this is not the case we re-fit the model using a prior to encourage a finding of positive RCT site selection bias, and find again little updating (i.e. the data are contributing little relative to the prior), but that this does not eliminate the conclusion of large negative internal selection bias in nongovernmental observational studies (Figure 4). Thus, while the prior has a large influence on the posterior interval for ρ_e, congruent with the wide credible intervals in the main results, this prior does not affect the finding of large internal selection bias in the observational studies.[18]

6.2 Microcredit

We now apply the Bayesian hierarchical plausibly exogenous selection model to the literature on the impact of access to Microcredit. This intervention offers households in developing countries the opportunity to take out small loans, usually uncollateralized, with the hope that they will be able to invest in their businesses and lift themselves out of poverty. This intervention, popularised by the Grameen Bank, was studied observationally for several years before it attracted the attention of economists running randomized trials, and as such we bring together a combination of studies from Aidgrade's database and from 2 previous meta-analyses: Meager (2019) and Chliova et al. (2015). We focus on household profit as the key outcome, as this is collected in the majority of studies, and this leads to a total of 12 studies, 7 RCTs and 5 observational, shown in Table 5. As with the CCTs application, in order to apply our identification approach, we recorded the country in which the study was done and the year in which the program was implemented. As above, using the JPAL-IPA project databases shared with us, for each country c we then recover $everJPAL_c$ and year y_c at which JPAL-IPA switches on in that country; from this we construct $JPALswitch_{jct} = \mathbb{1}\{t > y_c\} * everJPAL_c$.

We fit the model in Section 5 to this data set and report the results in Table 6. We see little evidence of either site selection bias or internal selection bias in this set of studies. Here, the LATE for both types of studies is also essentially zero, though with very wide uncertainty intervals due to the lack of information in the instrument in this case. The coefficients for experimental and observational studies are shown graphically alongside

[18] A useful alternative LATE-like measure of site selection bias could be to compare the average experimental results obtained in "always-taking" studies which would have been experiments regardless of the JPAL-IPA instrument to the average experimental results for complying studies.

each other in Figure 5. Finding no bias of either type is somewhat unexpected in this literature, especially given the raw patterns in the estimated effects, which seem to suggest a positive bias in the observational studies (see Figure 6). Here, the hierarchical structure of the model is playing a key role: finding that the conditional means of the two study types are essentially identical is due to the model essentially discarding the outlying observational study in Figure 6, both because it is so outlying and because its standard error is over half the size of the estimated effect, and double that of the next-largest standard error in the literature.

The lack of a strong residualized correlation between $JPALswitch_{jct}$ and study type, due to the much stronger correlation between the linear time trend and study type, gives reason for caution here. We take the results to be suggestive rather than conclusive. The model is still identified due to the different functional forms of the selection equation and the outcome equations, but it is interesting to understand how results would change if time - the stronger predictor - was selected as an additional excluded variable in the outcome regressions. Pursuing such an adjustment is no longer a DD-IV strategy, but rather an attempt to find ways to predict selection that are not correlated to the outcomes, and proceed by separating out the predictive surface of the two using the nonlinearity. The results of a model in which both $JPALswitch_{jct}$ and $time$ are treated as plausibly exogenous are shown in Table 7 and we still find no evidence of either type of bias. Again, this is because the hierarchical model discards the largest observational treatment effect as implausible given its distance from the rest of the studies and its large standard error.

7 Conclusion

We conclude with a few bigger picture thoughts on where our methodology and results sit with respect to current and future work. We hope that our approach will be picked up for other meta-analytical studies, both examining different empirical literatures and using different sources of exogenous variation in research design. For organizations and researchers able to take a larger view of program evaluation, our non-parametric identification results may be of particular interest. Funding organizations could generate instrumental variables by creating randomized incentives to choose experimental research designs while tracking the results of both experimental and non-experimental evaluations. In Section 2.3, we mentioned that nonparametric identification and extrapolation of distributions of experimental effects conditional on a particular value of a study's observational treatment effect is only possible with a continuously distributed instrument. There is no reason incentives for experimental research design should be binary: having them drawn from a continuous distribution would help facilitate the objective of generating predictive distributions for the effects of programs where only observational studies are cost-effective, the end goal of many methods for securing external validity.

Empirically, it would be incautious to extrapolate from the two applications here to any broad lessons about the typical amount of each type of bias in empirical economics literatures. The fact that neither of our applications found evidence of site selection bias in the RCT literature may be a lucky (or unlucky) draw, or it may be a typical result, reflecting either a general lack of bias or the presence of multiple competing selection biases working in opposite directions. We hope that the methods provided in this paper will be used to assess the site selection bias in experimental literatures beyond the applications provided here. Yet our applied results at least demonstrate that there do exist literatures in which RCTs, according to our point estimates if not strongly by our credible intervals, are not affected by site selection bias and for which aggregate effects estimated from experimental studies might plausibly be used to infer effects for a broader class of contexts.

Where possible, however, we recommend combining both experimental and observational studies for the purposes of meta-analysis using the approaches we have laid out in this paper. This allows us to study a broader set of contexts and benefit from greater estimation precision by using a larger set of studies, and permits an assessment of the amount of bias present in each type of study, which could be of interest for its own sake. Even if both types of bias are present in a given literature, quantifying the sign and magnitude of each allows us to learn more about how the literature itself is constructing and processing information about the policy or intervention of interest (consider, for example, the differential types of bias present in government and non-government CCTs studies). This substantive knowledge may turn out to be just as valuable as the ability to eliminate the bias from our inference on average treatment effects.

Table 1: Finite Sample Credible Interval Coverage in Simulation with Incorrect Priors

Model	N	50% CI	95% CI
Bayesian Hierarchical Heckman	50	0.45	0.92
Bayesian Hierarchical Heckman	30	0.45	0.91
Bayesian Hierarchical Heckman	15	0.47	0.91
Bayesian Hierarchical Plausibly Exogenous Heckman	50	0.43	0.89
Bayesian Hierarchical Plausibly Exogenous Heckman	30	0.43	0.88
Bayesian Hierarchical Plausibly Exogenous Heckman	15	0.42	0.89

Notes: Each row is based on 100 simulations of 5000 Hamiltonian Monte Carlo iterations across 4 chains, and takes the central posterior marginal interval of each parameter for the nominal coverage specified. There is likely to be some Monte Carlo noise present in these results, and we take them to be indicative of approximate magnitudes of coverage in practice.

Table 2: Conditional Cash Transfers Studies

Paper	Country	TE (SE)	RCT?
Akresh, de Walque and Kazianga (2013)	Burkina Faso	0.18 (0.05)	1
Angelucci et al. (2010)	Mexico	0.1 (0.02)	1
Arraiz and Rozo (2011)	Panama	0.08 (0.02)	0
Attanasio et al. (2010)	Colombia	0.03 (0.01)	0
Baird et al. (2011)	Malawi	0.38 (0.09)	1
Barrera-Osorio et al. (2008)	Colombia	-0.02 (0.01)	1
Behrman, Parker and Todd (2004)	Mexico	0.12 (0.03)	1
Chaudhury, Friedman and Onishi (2013)	Philippines	0.04 (0.01)	1
Davis et al. (2002)	Mexico	0.05 (0.01)	0
Dubois, de Janvry and Sadoulet (2012)	Mexico	0.03 (0)	1
Edmonds and Schady (2012)	Ecuador	0.19 (0.07)	1
Ferreira, Filmer and Schady (2009)	Cambodia	0.21 (0.02)	0
Ferro, Kassouf and Levison (2007)	Brazil	0.03 (0.01)	0
Fuwa (2001)	Bangladesh	-0.01 (0)	0
Galasso (2006)	Chile	0.09 (0.03)	0
Galiani and McEwan (2013)	Honduras	0.09 (0.03)	1
Garcia and Hill (2010)	Colombia	0.03 (0.05)	0
Gitter and Barham (2009)	Nicaragua	0.02 (0.03)	1
Glewwe and Kassouf (2012)	Brazil	0 (0)	0
Maluccio, Murphy and Regalia (2010)	Nicaragua	0.05 (0.02)	1
Mo et al. (2013)	China	-0.08 (0.03)	1
Olinto and Souza (2005)	Honduras	0.03 (0.02)	1
Perova (2010)	Peru	0.04 (0.01)	0
Rubio-Codina (2010)	Mexico	0.01 (0.02)	1
Ward et al. (2010)	Kenya	0.03 (0.02)	0
De Brauw and Gilligan (2011)	El Salvador	0.07 (0.02)	0
De Janvry, Finan and Sadoulet (2006)	Brazil	0.08 (0)	0
De Janvry, Finan and Sadoulet (2012)	Brazil	-0.06 (0)	0

Notes: Standard Errors computed from AidGrade Enrolment Effect Confidence Intervals assuming use of the log scale, rounded to 2 decimal places; numbers recorded as zero are not literally zero but very small.

Table 3: Conditional Cash Transfers Bayesian Selection Model Results

	Selection.Coef	95% CI	RCT.Coefs	95% CI	Obs.Coefs	95% CI
Intercept	-0.01	(-0.96,0.95)	0.05	(-0.07,0.17)	0.14	(0.01,0.26)
time	-0.01	(-0.14,0.11)	-0.01	(-0.02,0.01)	0	(-0.01,0.01)
ever JPAL	-0.08	(-1.13,0.97)	-0.02	(-0.14,0.11)	-0.1	(-0.23,0.03)
JPAL-switch	0.3	(-0.94,1.53)	0.04	(-0.1,0.17)	0.04	(-0.1,0.17)
NonGov	0.32	(-0.91,1.55)	0.27	(0.09,0.45)	-0.03	(-0.19,0.13)
rho			0	(-0.62,0.63)	0	(-0.62,0.63)
Population SD			0.07	(0.03,0.15)	0.06	(0.04,0.12)
LATE (Gov)	-0.02	(-0.12,0.09)				
LATE (Non-Gov)	0.28	(0.05,0.53)				

Notes: N = 28, Hamiltonian Monte Carlo Iterations = 5,000, Number of Chains = 4, R-hat criterion indicating good chain mixing lies below 1.001 in all cases.

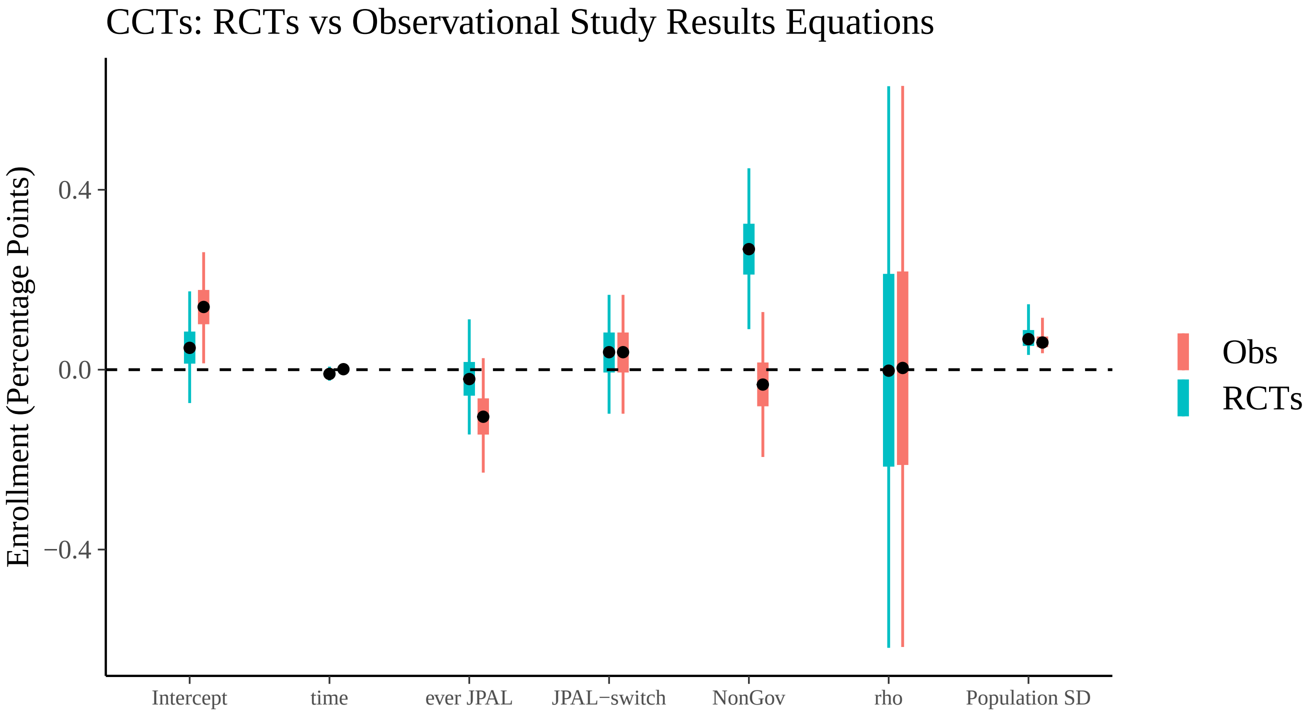

Figure 1: Conditional Cash Transfers Bayesian Selection Model Results

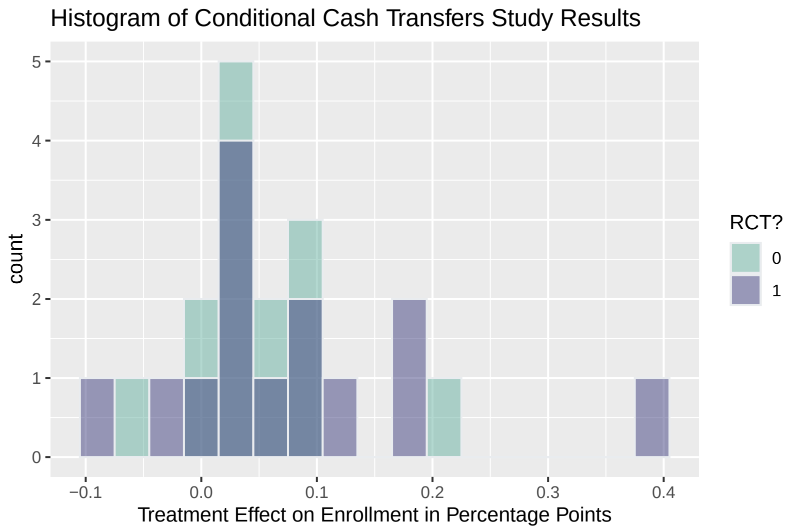

Figure 2: Histogram of Conditional Cast Transfer Study Results (Split by Design)

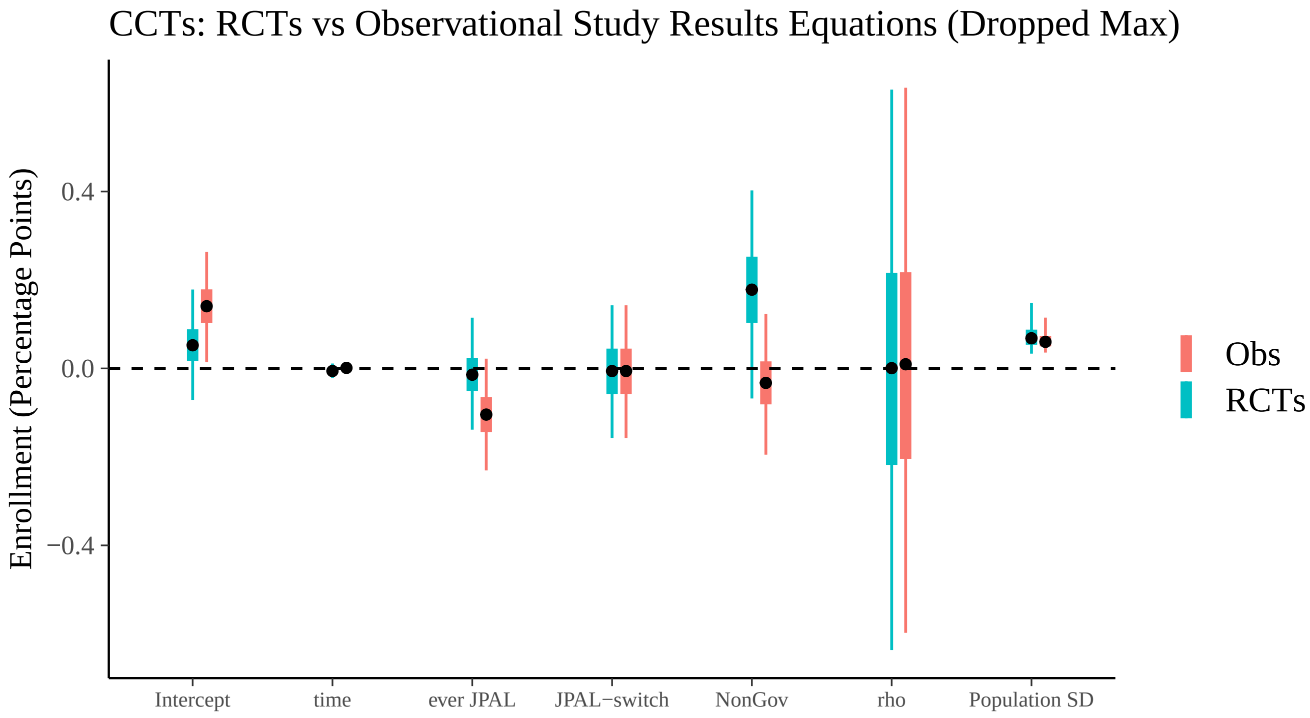

Figure 3: Conditional Cash Transfers Results with the Largest Recorded Result Removed

Table 4: Conditional Cash Transfers Bayesian Selection Model Results (Dropping Max Study)

	Selection.Coef	95% CI	RCT.Coefs	95% CI	Obs.Coefs	95% CI
Intercept	-0.01	(-0.97,0.92)	0.05	(-0.07,0.18)	0.14	(0.01,0.26)
time	-0.02	(-0.15,0.11)	-0.01	(-0.02,0.01)	0	(-0.01,0.01)
ever JPAL	-0.07	(-1.1,0.98)	-0.01	(-0.14,0.12)	-0.1	(-0.23,0.02)
JPAL-switch	0.14	(-1.12,1.41)	-0.01	(-0.16,0.14)	-0.01	(-0.16,0.14)
NonGov	0.04	(-1.32,1.4)	0.18	(-0.07,0.4)	-0.03	(-0.2,0.13)
rho			0	(-0.6,0.62)	0	(-0.61,0.62)
Population SD			0.07	(0.03,0.14)	0.06	(0.04,0.12)
LATE (Gov)	0	(-0.1,0.1)				
LATE (Non-Gov)	0.21	(-0.06,0.48)				

Notes: N = 28, Hamiltonian Monte Carlo Iterations = 5,000, Number of Chains = 4, R-hat criterion indicating good chain mixing lies below 1.001 in all cases.

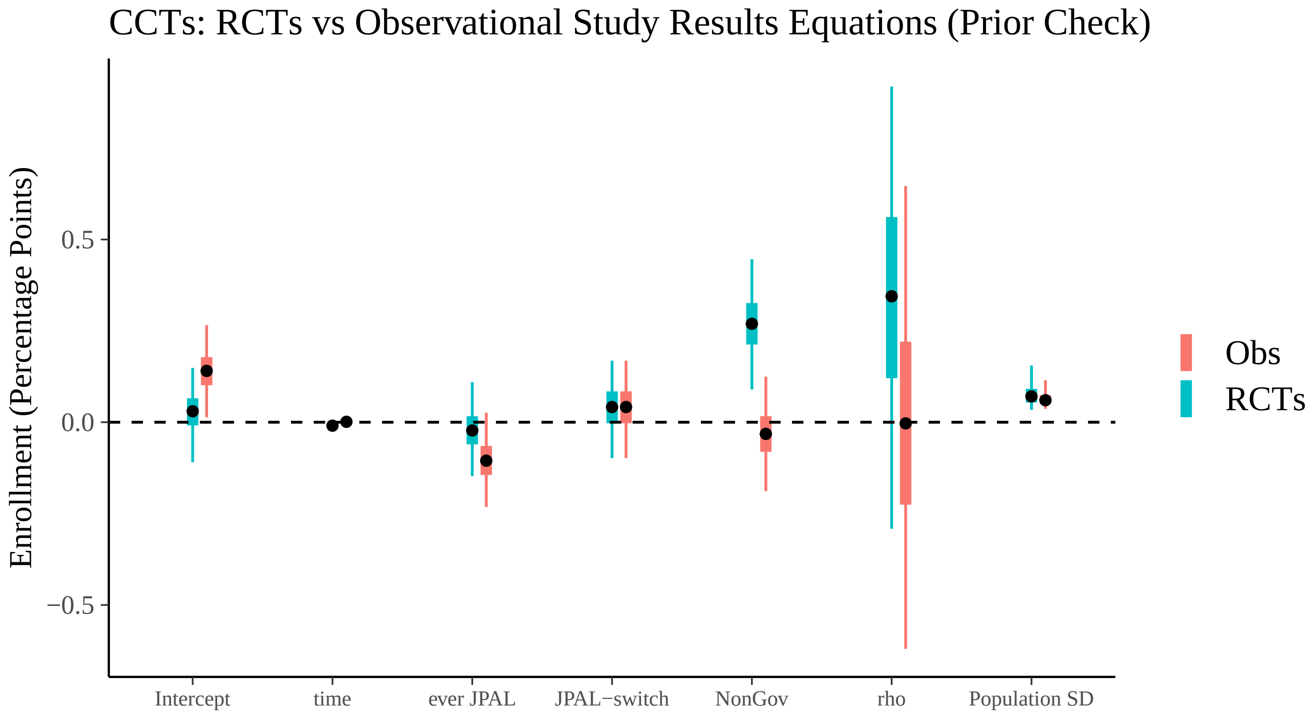

Figure 4: Conditional Cash Transfers Results with prior enforcing positive site selection

Table 5: Microcredit Studies

Paper	Country	TE (SE)	RCT?
Banerjee et al (2015)	India	14.07 (12.48)	1
Cotler and Woodruff (2008)	Mexico	12.94 (10.78)	0
Montgomery (2006)	Pakistan	77 (112.12)	0
Takahashi, Higashikata and Tsukada (2010)	Indonesia	0 (0)	0
Tarozzi et al (2015)	Ethiopia	8.31 (6.98)	1
Attanasio et al (2015)	Mongolia	-0.54 (0.59)	1
Crepon et al (2015)	Morocco	19.99 (12.99)	1
Angelucci et al (2015)	Mexico	0 (4.96)	1
Augsburg et al (2015)	Bosnia	35.32 (28.45)	1
Karlan and Zinman (2011)	Philippines	70.32 (80.44)	1
Edgecomb and Garber (1998)	Honduras	379.32 (216.75)	0
McNelly and Dunford (1999)	Bolivia	4.26 (7.29)	0

Notes: Effects and standard errors are in USD PPP over an interval of 2 weeks. Values are rounded to 2 significant figures, zeros are not literal zeroes but very small values.

Table 6: Microcredit Bayesian Selection Model Results

	Selection.Coef	95% CI	RCT.Coefs	95% CI	Obs.Coefs	95% CI
Intercept	0.24	(-1.09,1.59)	0.29	(-1.55,2.09)	0.21	(-1.49,1.91)
time	0.38	(0.06,0.84)	-0.3	(-1.25,0.67)	-0.24	(-1.13,0.62)
ever JPAL	0.09	(-1.3,1.49)	0.25	(-1.54,2.07)	0.19	(-1.56,1.9)
JPAL-switch	-0.13	(-1.5,1.24)	0	(-0.19,0.2)	0	(-0.19,0.2)
Gov	-0.68	(-2.34,0.94)	-0.01	(-1.96,1.94)	-0.01	(-2.02,2.01)
rho			0.03	(-0.56,0.62)	-0.02	(-0.59,0.56)
Population SD			0.94	(0.12,2.54)	0.89	(0.1,2.47)
LATE (Non-Gov)	0.02	(-2.44,2.49)				
LATE (Gov)	0.06	(-3.7,3.82)				

Notes: N = 12, Hamiltonian Monte Carlo Iterations = 5,000, Number of Chains = 4, R-hat criterion indicating good chain mixing lies below 1.001 in all cases..

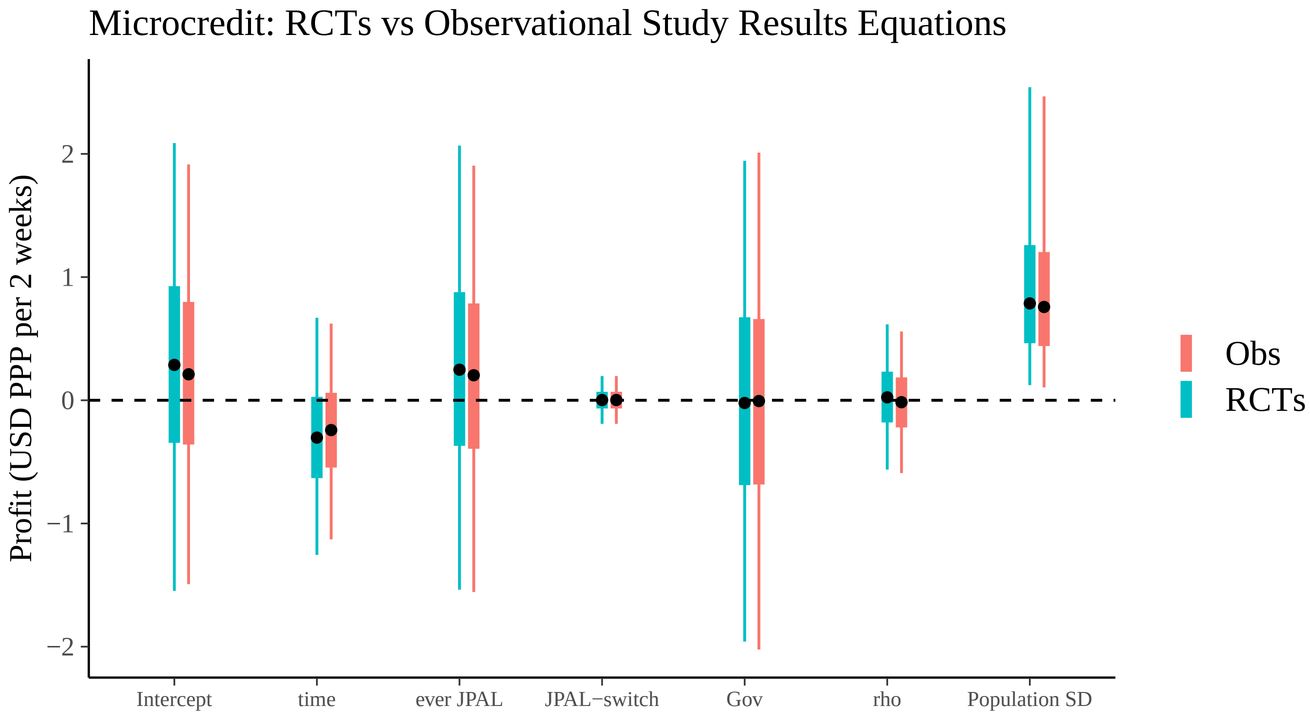

Figure 5: Microcredit Bayesian Selection Model Results

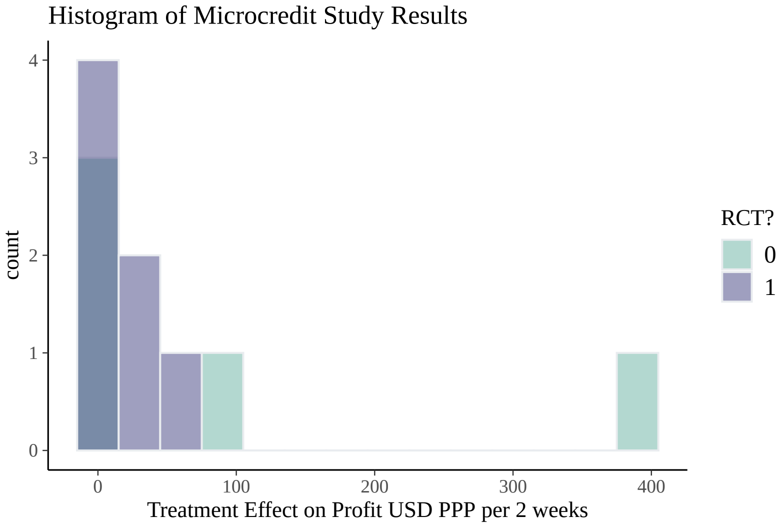

Figure 6: Histogram of Microcredit Study Results, Split by Design

Table 7: Microcredit Bayesian Selection Model Results with Time Plausibly Exogenous

	Selection.Coef	95% CI	RCT.Coefs	95% CI	Obs.Coefs	95% CI
Intercept	0.24	(-1.1,1.6)	-0.01	(-1.61,1.64)	0.07	(-1.54,1.64)
ever JPAL	0.1	(-1.33,1.51)	-0.03	(-1.63,1.63)	0.06	(-1.55,1.64)
time	-0.13	(-1.49,1.24)	-0.01	(-0.21,0.18)	-0.01	(-0.2,0.18)
JPAL-switch	0.38	(0.06,0.85)	0	(-0.2,0.2)	0	(-0.2,0.19)
Gov	-0.69	(-2.36,0.9)	-0.01	(-1.93,1.93)	0.01	(-1.95,2.01)
rho			0.03	(-0.55,0.6)	0.01	(-0.58,0.57)
Population SD			0.91	(0.12,2.48)	0.85	(0.11,2.37)

Notes: N = 12, Hamiltonian Monte Carlo Iterations = 5,000, Number of Chains = 4, R-hat criterion indicating good chain mixing lies below 1.001 in all cases. We do not report a LATE because enforcing time as plausibly exogenous means this is no longer a Differences in Differences IV strategy and we do not encourage its interpretation in this way: it is a robustness check on the general patterns in the results from the proper model.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *113*(2), 231–263.

Ali, M. M., N. N. Mikhail, and M. S. Haq (1978). A Class of Bivariate Distributions Including the Bivariate Logistic. *Journal of Multivariate Analysis 8*, 405–412.

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics 130*(3), 1117–1165.

Andor, M., J. Peters, A. Gerster, and C. M. Schmidt (2020). Social Norms and Energy Conservation Beyond the US. *Journal of Environmental Economics and Management 103*(102351).

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–2794.

Arellano, M. and S. Bonhomme (2017). Quantile selection models. *Econometrica 85*(1), 1–28.

Athey, S., R. Chetty, and G. Imbens (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

Bandiera, O., G. Fischer, A. Prat, and E. Ytsma (2021). Do women respond less to performance pay? Building evidence from multiple experiments. *American Economic Review: Insights, forthcoming*.

Bjorklund, A. and R. Moffitt (1987). The Estimation of Wage Gains and Welfare Gains in Self-Selection Models. *The Review of Economics and Statistics 69*(1), 42–49.

Blundell, R. and M. Costa Dias (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources 44*(3), 565–641.

Brinch, C., M. Mogstad, and M. Wiswall (2017). Beyond LATE with a discrete instrument. *Journal of Political Economy 125*(4).

Chliova, M., J. Brinckmann, and N. Rosenbusch (2015). Is microcredit a blessing for the poor? a meta-analysis examining development outcomes and contextual considerations. *Journal of business Venturing 30*(3), 467–487.

Coe, R., J. Njoloma, and F. Sinclair (2019). Loading the dice in favour of the farmer: Reducing the risk of adopting agronomic innovations. *Experimental Agriculture 55*(S1), 67?83.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012). Plausibly exogenous. *The Review of Economics and Statistics 94*(February), 260–272.

de Chaisemartin, C. and X. D'Haultfoeuille (2018). Fuzzy Differences-in-Differences. *Review of Economic Studies 85*, 999–1028.

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia : Evidence from an Unusual Policy Experiment. *American Economic Review 91*(4), 795–813.

Gelman, A., J. B. Carlin, H. S. Stern, D. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis.*

Gui, G. (2020). Combining observational and experimental data using first-stage covariates. *arXiv preprint arXiv:2010.05117*.

Hartman, E., R. Grieve, R. Ramsahai, and J. S. Sekhon (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 757–778.

Heckman, J. and R. Pinto (2014). Causal analysis after haavelmo. *Econometric Theory 31*, 1–37.

Heckman, J. and E. Vytlacil (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement 5*(4), 475–492.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica 47*(1), 153–161.

Hsiang, S. M., M. Burke, and E. Miguel (2013). Quantifying the influence of climate on human conflict. *Science 341*(6151).

Hudson, S., P. Hull, and J. Liebersohn (2017). Interpreting Instrumented Difference-in-Differences. *Mimeo*.

Hull, P. (2018). Estimating hospital quality with quasi-experimental data. *Available at SSRN 3118358*.

Imbens, G. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and D. B. Rubin (1997). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *Review of Economic Studies 64*(4), 555–574.

JPAL (2021). About us.

Kahn-Lang, A. and K. Lang (2020). The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications. *Journal of Business and Economic Statistics 38*(3), 613–620.

Katz, L. F., J. R. Kling, and J. B. Liebman (2001). Moving to opportunity in Boston: Early Results of a Randomized Mobility Experiment. *Quarterly Journal of Economics* (May), 607–653.

Kline, P. and C. R. Walters (2019). On Heckits, LATE, and Numerical Equivalence. *Econometrica 87*(2), 677–696.

Kowalski, A. E. (2016). Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Experiments. *NBER Working Paper 22363*.

Laajaj, R., K. Macours, C. Masso, M. Thuita, and B. Vanlauwe (2020). Reconciling yield gains in agronomic trials with returns under african smallholder conditions. *Scientific reports 10*(1), 1–15.

Manski, C. F. (1994). The Selection Problem. In C. A. Sims (Ed.), *Advances in Econometrics, Sixth World Congress Vol. 1*, pp. 143–170. Cambridge University Press.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics 11*(1), 57–91.

Miller, M., J. Reichelstein, C. Salas, and B. Zia (2015). Can you help someone become financially capable? a meta-analysis of the literature. *The World Bank Research Observer 30*(2), 220–246.

Mogstad, M., A. Santos, and A. Torgovitsky (2018). Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters. *Econometrica 86*(5), 1589–1619.

Olsen, R. J. . (1980). A Least Squares Correction for Selectivity Bias. *Econometrica 48*(7), 1815–1820.

Peysakhovich, A. and A. Lada (2016). Combining observational and experimental data to find heterogeneous treatment effects. *arXiv preprint arXiv:1611.02385*.

Reid, J. L., M. E. Fagan, and R. A. Zahawi (2018). Positive site selection bias in meta-analyses comparing natural regeneration to active forest restoration. *Science advances 4*(5), eaas9143.

Roy, A. (1951). Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers 3*, 135–146.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics 6*(4), 377–401.

Toomet, O. and A. Henningsen (2008). Sample selection models in R: Package sample-Selection. *Journal of Statistical Software 27*(7), 1–23.

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association 18*(6), 3045–3089.

Vytlacil, E. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica 70*(1), 331–341.

Wang, S. and D. Y. Yang (2021). Policy Experimentations in China: the Political Economy of Policy Learning. *Working Paper*.

A Proofs not included in the main text

A.1 Proposition 3

Proof. By Assumption 6.1

$$\begin{aligned}&P(D_{jc1}^{JPALswitch=1}=e|everJPAL_c=1)-P(D_{jc1}^{JPALswitch=0}=e|everJPAL_c=1)\\&=P_{11}-(P_{01}+P_{10}-P_{00}).\end{aligned}$$

By Assumption 6.3 the response in treatment (research design) choice to the JPAL-IPA office being operational at $t=1$ can be only be weakly towards the treatment (experimental design). So the above is equal to the fraction of studies changing their research design as a result of the JPAL office opening, which we refer to as complying studies. We let p_c refer to the fraction of such studies in the cross section of studies in contexts where $everJPAL_c=1$ at time 1.

$$\begin{aligned}&P_{11}-(P_{01}+P_{10}-P_{00})\\&=P(D_{jc1}^{JPALswitch=1}\neq D_{jc1}^{JPALswitch=0}|everJPAL_c=1)=p_c.\end{aligned}\qquad(36)$$

For shares of the other two possible responses to the JPAL-IPA office opening, let:

$$\begin{aligned}&P(D_{jc1}^{JPALswitch=1}=D_{jc1}^{JPALswitch=0}=e|everJPAL_c=1)=p_a\\&P(D_{jc1}^{JPALswitch=1}=D_{jc1}^{JPALswitch=0}=o|everJPAL_c=1)=p_n.\end{aligned}$$

We now turn to identifying the expected outcome (results) that would have been observed at $t=1$ had the JPAL office not been operational in contexts with $everJPAL_c=1$:

$$\begin{aligned}&E[R_{jc1}|fix(JPALswitch_{c1})=0,everJPAL_c=1]\\&=E[R_{jc1}^{o}|D_{jc1}^{JPALswitch=0}=o,everJPAL_c=1](p_n+p_c)+E[R_{jc1}^{e}|D_{jc1}^{JPALswitch=0}=e,everJPAL_c=1]p_a\\&=(\bar{R}_{01}^{o}+\bar{R}_{10}^{o}-\bar{R}_{00}^{o})(1-(P_{01}+P_{10}-P_{00}))+(\bar{R}_{01}^{e}+\bar{R}_{10}^{e}-\bar{R}_{00}^{e})(P_{01}+P_{10}-P_{00}).\end{aligned}$$

Replacement of $E[R_{jc1}^{d}|D_{jc1}^{JPALswitch=0}=o,everJPAL_c=1]$ with $\bar{R}_{01}^{d}+\bar{R}_{10}^{d}+\bar{R}_{00}^{d}$ follows from Assumption 6.2. We now subtract $E[R_{jc1}|fix(JPALswitch_{c1})=0,everJPAL_c=$

1] from the observed $E[R_{jc1}|everJPAL_c = 1]$:

$$\begin{aligned}
&E[R_{jc1}|everJPAL_c = 1] - E[R_{jc1}|fix(JPALswitch_{c1}) = 0, everJPAL_c = 1] \\
&\quad = (E[R_{jc1}^{e,JPALswitch=1}|everJPAL_c = 1, D_{jc1}^{JPALswitch=1} \neq D_{jc1}^{JPALswitch=0}]p_c \\
&\qquad\qquad + E[R_{jc1}^{e,JPALswitch=1}|everJPAL_c = 1, D_{jc1}^{JPALswitch=1} = D_{jc1}^{JPALswitch=0} = e]p_a \\
&\qquad\qquad + E[R_{jc1}^{o,JPALswitch=1}|everJPAL_c = 1, D_{jc1}^{JPALswitch=1} = D_{jc1}^{JPALswitch=0} = o]p_n) \\
&\qquad - (E[R_{jc1}^{o,JPALswitch=0}|everJPAL_c = 1, D_{jc1}^{JPALswitch=1} \neq D_{jc1}^{JPALswitch=0}]p_c \\
&\qquad\qquad + E[R_{jc1}^{e,JPALswitch=0}|everJPAL_c = 1, D_{jc1}^{JPALswitch=1} = D_{jc1}^{JPALswitch=0} = e]p_a \\
&\qquad\qquad + E[R_{jc1}^{o,JPALswitch=0}|everJPAL_c = 1, D_{jc1}^{JPALswitch=1} = D_{jc1}^{JPALswitch=0} = o]p_n) \\
&\quad = (E[R_{jc1}^{e}|everJPAL_c = 1, D_{jc1}^{JPALswitch=1} \neq D_{jc1}^{JPALswitch=0}] \\
&\qquad\qquad - E[R_{jc1}^{o}|everJPAL_c = 1, D_{jc1}^{JPALswitch=1} \neq D_{jc1}^{JPALswitch=0}])p_c
\end{aligned}$$

where the last equality follows from Assumption 6.4. Dividing by p_c as identified in Equation (36) yields the result. □

A.2 Lemma 1

To help in deriving the parametric LATEs in Equations (26) and Proposition 4, we first prove a lemma regarding the LATE in our DD-IV setting which nests the uniform and Gaussian cases.

Lemma 1. *Let*

$$\begin{aligned}
&D_{jct} = e \iff \mathbb{1}\{\pi_0 + \pi_1 t + \pi_2 JPAL_c + \pi_3 everJPAL_c * t \geq J(U_{jct})\} \\
&E[R_{jct}^d|everJPAL, t, U] = \alpha_d + \beta_{1d} t + \beta_{2d} everJPAL + \gamma_d (J(U_{jct}) - \mu_J) \\
&U_{jct} \perp\!\!\!\perp t, everJPAL_c.
\end{aligned} \tag{37}$$

Then

$$\begin{aligned}
LATE =&(\bar{R}_{01}^e + \bar{R}_{10}^e - \bar{R}_{00}^e) - (\bar{R}_{01}^o + \bar{R}_{10}^o + \bar{R}_{00}^o) \\
&- \left(\frac{DID_R^e}{DID_\lambda^e}(\lambda_e(P_{01}) + \lambda_e(P_{10}) - \lambda_e(P_{00})) - \frac{DID_R^o}{DID_\lambda^o}(\lambda_o(P_{01}) + \lambda_o(P_{10}) - \lambda_o(P_{00}))\right) \\
&+ \left(\frac{DID_R^e}{DID_\lambda^e} - \frac{DID_R^o}{DID_\lambda^o}\right)\left(\frac{P_{11}\lambda_e(P_{11})) - (P_{01} + P_{10} - P_{00})\lambda_e(P_{01} + P_{10} + P_{00})}{DID_P}\right)
\end{aligned} \tag{38}$$

where the terms $\lambda_e(p), \lambda_o(p), DID_\lambda^d, DID_P$, and DID_R^d are defined in the statement of Proposition 4.

Proof. Analogously to Equations 20, 21, 22, and 23, we can write:

$$\bar{R}^e_{00} = \alpha_e + \gamma_e \lambda_e(P_{00}),$$

$$\bar{R}^e_{10} = \alpha_e + \gamma_e \lambda_e(P_{10}) + \beta_{1e}, \tag{39}$$

$$\bar{R}^e_{01} = \alpha_e + \gamma_e \lambda_e(P_{01}) + \beta_{2e}, \tag{40}$$

$$\bar{R}^e_{11} = \alpha_e + \gamma_e \lambda_e(P_{11}) + \beta_{1e} + \beta_{2e}. \tag{41}$$

γ_e is identified from DID^e_R:

$$\bar{R}^e_{11} - \bar{R}^e_{01} - (\bar{R}^e_{10} - \bar{R}_{00}) = \gamma_e(\lambda_e(P_{11}) - \lambda_e(P_{01}) - (\lambda_e(P_{10}) - \lambda_e(P_{00})))$$

so

$$\gamma_e = \frac{\bar{R}^e_{11} - \bar{R}^e_{01} - (\bar{R}^e_{10} - \bar{R}_{00})}{\lambda_e(P_{11}) - \lambda_e(P_{01}) - (\lambda_e(P_{10}) - \lambda_e(P_{00}))}.$$

γ_e, β_{1e}, and β_{2e} are then identified as in Equation 24, and using Equations (39) and (40), respectively. We can do the same for γ_o using:

$$\bar{R}^o_{00} = \alpha_o + \gamma_o \lambda_o(P_{00}),$$

$$\bar{R}^o_{10} = \alpha_o + \gamma_o \lambda_o(P_{10}) + \beta_{1o},$$

$$\bar{R}^o_{01} = \alpha_o + \gamma_o \lambda_o(P_{01}) + \beta_{2o},$$

$$\bar{R}^o_{11} = \alpha_o + \gamma_o \lambda_o(P_{11}) + \beta_{1o} + \beta_{2o}.$$

Using the definition of the LATE and the identification results above, we can write it as:

$$
\begin{aligned}
LATE =& \bar{R}^e_{00} - \bar{R}^o_{00} \\
&- \left(\frac{DID^e_R}{DID^e_\lambda} \lambda_e(P_{00}) - \frac{DID^o_R}{DID^o_\lambda} \lambda_o(P_{00}) \right) \\
&+ (\bar{R}^e_{10} - \bar{R}^e_{00}) - (\bar{R}^o_{10} - \bar{R}^o_{00}) \\
&- \left(\frac{DID^e_R}{DID^e_\lambda} (\lambda_e(P_{10}) - \lambda_e(P_{00}) - \frac{DID^o_R}{DID^o_\lambda} (\lambda_o(P_{10}) - \lambda_o(P_{00})) \right) \\
&+ (\bar{R}^e_{01} - \bar{R}^e_{00}) - (\bar{R}^o_{01} - \bar{R}^o_{00}) \\
&- \left(\frac{DID^e_R}{DID^e_\lambda} (\lambda_e(P_{01}) - \lambda_e(P_{00}) - \frac{DID^o_R}{DID^o_\lambda} (\lambda_o(P_{01}) - \lambda_o(P_{00})) \right) \\
&+ \left(\frac{DID^e_R}{DID^e_\lambda} - \frac{DID^o_R}{DID^o_\lambda} \right) \left(\frac{P_{11}\lambda_e(P_{11})) - (P_{01} + P_{10} - P_{00})\lambda_e(P_{01} + P_{10} + P_{00})}{DID_P} \right) \\
=& (\bar{R}^e_{01} + \bar{R}^e_{10} - \bar{R}^e_{00}) - (\bar{R}^o_{01} + \bar{R}^o_{10} + \bar{R}^o_{00}) \\
&- \left(\frac{DID^e_R}{DID^e_\lambda} (\lambda_e(P_{01}) + \lambda_e(P_{10}) - \lambda_e(P_{00})) - \frac{DID^o_R}{DID^o_\lambda} (\lambda_o(P_{01}) + \lambda_o(P_{10}) - \lambda_o(P_{00})) \right) \\
&+ \left(\frac{DID^e_R}{DID^e_\lambda} - \frac{DID^o_R}{DID^o_\lambda} \right) \left(\frac{P_{11}\lambda_e(P_{11})) - (P_{01} + P_{10} - P_{00})\lambda_e(P_{01} + P_{10} + P_{00})}{DID_P} \right).
\end{aligned}
$$

□

A.3 Equation (26)

In this Appendix, we show that the model laid out in Equations (15), (16), (17), combined with Assumption 7 yield the LATE given in Equation (26).

Proof. From Equations (15), (16), (17) and assuming the set of Equations (37) with $J(\cdot) = I(\cdot)$, where $I(\cdot)$ is the identity function, so Lemma 1 tells us the LATE takes the form

$$
\begin{aligned}
LATE =& (\bar{R}^e_{01} + \bar{R}^e_{10} - \bar{R}^e_{00}) - (\bar{R}^o_{01} + \bar{R}^o_{10} + \bar{R}^o_{00}) \\
&- \left(\frac{DID^e_R}{DID^e_\lambda} (\lambda_e(P_{01}) + \lambda_e(P_{10}) - \lambda_e(P_{00})) - \frac{DID^o_R}{DID^o_\lambda} (\lambda_o(P_{01}) + \lambda_o(P_{10}) - \lambda_o(P_{00})) \right) \\
&+ \left(\frac{DID^e_R}{DID^e_\lambda} - \frac{DID^o_R}{DID^o_\lambda} \right) \left(\frac{P_{11}\lambda_e(P_{11})) - (P_{01} + P_{10} - P_{00})\lambda_e(P_{01} + P_{10} + P_{00})}{DID_P} \right).
\end{aligned}
\tag{42}
$$

Given $U(\cdot) = I(\cdot)$, $\lambda_e(p) = \frac{p-1}{2}$ and $\lambda_o(p) = \frac{p}{2}$. This linearity in p means that:

$$
\lambda_d \left(\sum_{l=1}^{L} p_l \right) = \sum_{l=1}^{L} \lambda_d(p_l). \tag{43}
$$

Furthermore

$$DID_P = (P_{11} - P_{01}) - (P_{10} - P_{01}),$$

so Equation 42 simplifies to the Wald Estimator form

$$LATE = \frac{[P_{11}\bar{R}^e_{11} + (1 - P_{11})\bar{R}^o_{11}] - [P_{CF}\bar{R}^e_{CF} + (1 - P_{CF})\bar{R}^o_{CF}]}{(P_{11} - P_{01}) - (P_{10} - P_{00})}.$$

□

A.4 Proposition 4

Proof. The form of the LATE follows immediately from Lemma 1. Without linearity of the $\lambda_d(\cdot)$ functions, Equation 43 does not hold and Equation 42 does not simplify. □