Measuring the Effectiveness of Non-Randomized Policy Interventions with Heterogeneous Treatment Effect

Melinda Vigh*

and

Chris Elbers

Vrije Universiteit Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands

January 31, 2017

Abstract

Social and development programs usually have heterogeneous treatment effects and are often targeted at those with the largest expected benefits (selection on gains). This paper demonstrates the estimation of relevant treatment effect measures in such situations using panel data and no instruments. We discuss the necessary conditions and assumptions for the identification of the Average Treatment Effect, the Average Treatment Effect on the Treated and the Total Program Effect using the Correlated Random Slopes model. The use of the method is illustrated using simulated data and on a large-scale water and sanitation intervention in Mozambique. We find evidence of selective intervention placement on the incentivized outcome variable but not on non-incentivized outcome variable.

Keywords: Selection effect; Correlated random coefficients

*Corresponding author, E-mail: m.vigh@vu.nl

1 Introduction

Social and development programs are rarely implemented following an (intended or accidental) randomization design. In addition, the effects of the programs are often heterogeneous in the population, and implementers may target the program on groups most in need or where they expect the highest benefits of the program. Quantifying the effectiveness of such programs is not straightforward. However, it provides useful information for policy-makers both in terms of planning and accountability.

We demonstrate how the Correlated Random Slopes (CRS) model of Wooldridge (2005, 2010) can be used to estimate the Average Treatment Effect (ATE), the Average Treatment Effect on the Treated (ATET) and a generalized version of the Total Program Effect (TPE) of Elbers and Gunning (2014) in situations with (non-randomized) policy interventions where panel data was collected in multiple survey rounds, and instrumental variables are unavailable or of little help.

In particular, the proposed method can be used in situations when the beneficiaries with the highest treatment effects or gains in outcomes are more likely to receive the program due to self-selection or selective intervention placement (targeting). Heckman et al. (2006) called this phenomena "essential heterogeneity" or "selection on gains."

TIn the selection on gains context, the Correlated Random Slopes model removes the effects of selectivity (correlation between the treatment effect and intervention variable) by essentially exploiting information present in the timing of the intervention delivery. This information is summarized in the time mean of the intervention variable: beneficiaries with higher expected treatment effect receive the intervention earlier, which results in a higher value for the time-mean of the intervention variable for this group. Hence, this identification strategy requires that we have at least three observations over time, and that the intervention is rolled out in different locations between the follow-up surveys. The method can be used to estimate treatment effects using living standard surveys when the intervention is rolled out gradually, and it can even

be extended to repeated cross-section surveys in cases when the intervention was targeted at village or regional level.

We show how in the context of our model the identification of ATE can be attained at the cost of strong assumptions. In comparison, the Total Program Effect of Elbers and Gunning (2014) measures the impact of the intervention (or a complex program) in the total program area inclusive of the effects of selectivity, while the ATET measures the average treatment effect on the beneficiaries also including the effects of selectivity. The estimation of the TPE and ATET need less strong assumptions as they do not require the elimination of the selection effects. They provide a measure of the actual impacts of the program on the beneficiary population already reached (provided that a representative sample or sampling weights were used for the estimation), which is of interest to the policy makers.[1]

Using simulated data from the normal distribution, we show that the Correlated Random Slopes regression provides an unbiased estimate of ATE when the correlation between the heterogeneous treatment effect and the intervention variable stems from selection on the expected gains of the beneficiaries, even when the expected gains/treatment effects are based on noisy observation. However, when other sources of correlation are introduced in the model, notably time-varying treatment effects, the estimate of ATE is no longer unbiased. In these cases, the estimates of ATET (and TPE) are still unbiased.

In order to illustrate the importance of accounting for correlated heterogeneous treatment effects, we present a simplified analysis of the effectiveness of the One Million Initiative in Mozambique. This example is a realistic scenario for the use of the proposed method. The sanitation intervention of the One Million Initiative was implemented by local NGOs, who were allowed to select the location of the interventions. They were also additionally rewarded for working with communities where the intervention was successful (all households switched to using a latrine). Community participation was also a crucial

[1] Note, however, that if selection on gains occurs then the size of ATET changes as the intervention is rolled out. Hence, the estimates of ATET and TPE are specific to a time period.

element of the interventions. Therefore, we expect quite some heterogeneity in the treatment effects, and that the implementing NGOs took this heterogeneity into account when deciding on the locations of the interventions. As a consequence, the sanitation interventions were likely to be subject to selection on gains.

We find that selective placement of the sanitation intervention was targeted at the incentivized outcome variable (latrine ownership) but not on non-incentivized outcomes (hand-washing practices). The Average Treatment Effect on the Treated is positive and significant for both variables. However, after controlling for selectivity, we find that the Average Treatment Effect of the CLTS interventions on latrine ownership is not significantly different from zero.

Our findings contribute to the literature on showing the importance of accounting for heterogeneous treatment effects in policy evaluations. We investigate the situation when heterogeneous treatment effects are utilized to allocate the interventions to those with the highest expected gains. In this setting, the results of Wooldridge (2005, 2010) and Elbers and Gunning (2014) are applied to estimate treatment effects of interest (ATE, ATET and TPE) using panel data methods without instrumental variables. The estimation procedure can be implemented using standard fixed effects estimation methods. We demonstrate the use of this method on simulated data and also on a large-scale policy intervention. In addition, we show that the estimating equation of the Total Program Effect of Elbers and Gunning (2014) can be nested in the Correlated Random Slopes model using a general specification of the random coefficients following Chamberlain (1980). This approach allows us to generalize the TPE for more than two time periods.

2 Identification strategy

We are interested in the Average Treatment Effect, Average Treatment Effect on the Treated and the Total Program Effect of Elbers and Gunning (2014). Denoting the outcome variable of interest with the intervention as Y_h^1 and without

the intervention as Y_h^0, and the intervention variable as D_h for each beneficiary (household) h, we can write the above treatment effects as

$$ATE = E(Y_h^1 - Y_h^0) \tag{1}$$

$$ATET = E(Y_h^1 - Y_h^0 | D_h = 1) \tag{2}$$

$$TPE = E\left((Y_h^1 - Y_h^0)D_h\right) \tag{3}$$

For the program beneficiaries, we do not observe the outcome without the intervention (counterfactual). When the interventions are randomly assigned, Y_h^0 can be estimated by the outcomes of the population that did not receive the treatment assuming that both groups have the same distribution. Then, $ATE = E(Y_h^1|D_h = 1) - E(Y_h^0|D_h = 0)$. However, using observational data (without random assignment), more assumptions are required in the form of an assumed data generating process or regression model.

2.1 Regression model

Consider the following model with beneficiary-specific treatment effect (β_h):[2]

$$Y_{ht} = \alpha_t + D_{ht}\beta_h + X_{ht}\theta + \eta_h + \varepsilon_{ht} \tag{4}$$

where α_t denotes the time-varying but common trend between the intervention arms.[3] Further, X_{ht} stands for household characteristics with constant coefficients (θ) and η_h is a heterogeneous intercept or fixed effect. The intervention variable, D_{ht}, has value zero before h received the intervention, and it can have binary, discrete or continuous values after the delivery of the intervention for h.[4] Note that D_{ht} can represent a single intervention or a vector of intervention

[2] Here, we do not discuss the setting where the treatment effect is allowed to change over time.

[3] It is an essential modelling assumption of estimating treatment effects by difference-in-difference that the (counterfactual) trend in the outcomes is the same for all intervention arms. Without this assumption, it would not be possible to distinguish between the effect of time-trend and the effect of the intervention.

[4] For a difference-in-difference specification, the first observation of D_{ht} is standardized as $D_{h1} = 0$ denoting that the first observations were measured before the interventions were rolled out. In later

arms. The outcome variable, Y_{ht}, can also be binary, discrete or continuous. In this paper, we discuss the estimation using linear least squares methods. Wooldridge (2010) also discusses implementations for non-linear models.

In addition, we assume that the program implementers have some additional information about the expected size of the treatment effect on the potential beneficiaries, which are unobserved by the researchers. In the absence of randomization, the program implementers are able to implement the program first at those locations that have the highest expected impact (selection on gains). This setting has three important implications: first, the intervention placement will be correlated with the heterogeneous treatment effect. Second, the location and timing of the interventions will carry additional information for the researchers that we can exploit in the estimation procedure. Third, the treatment effect will decrease over time as the intervention is rolled out to beneficiaries in order of the size of the expected impact.

We are interested in the estimation of model (4) because policy interventions are often targeted at beneficiaries most in need or with the highest potential benefit. For example, in the case of the One Million Initiative, we have good reasons to believe that the implementing NGOs placed the sanitation interventions at locations where they expected the treatment effects to be the highest.

In the case of model (4), we can write the treatment effects of interest as $ATE = E(\beta_h)$, $ATET = E(\beta_h|D_{hT} = 1)$ and $TPE = E(\Delta D_{hT}\beta_h)$ with $\Delta D_{hT} = D_{hT} - D_{h1}$.

We can reformulate (4) by splitting the coefficient of the treatment effect into its mean and the deviation around the mean: $\beta_h = \beta + b_h$ such that $E(b_h) = 0$. After regrouping the terms, we have

$$Y_{ht} = \alpha_t + D_{ht}\beta + X_{ht}\theta + \eta_h + [D_{ht}b_h + \varepsilon_{ht}] \tag{5}$$

It is easy to see that $ATE = \beta$ in this case. If the terms in square brack-

periods, $D_{ht} = 0$ for the comparison group and non-zero for the beneficiaries that have already been reached by the intervention.

ets are uncorrelated with the regressors D and X, i.e. $E(b_h|D,X) = 0$ and $E(\varepsilon_{ht}|D,X) = 0$, we can consistently estimate β using fixed effects regression on:

$$Y_{ht} = \alpha_t + D_{ht}\beta + X_{ht}\theta + \eta_h + e_{ht} \tag{6}$$

where all terms in the square brackets have been pushed into the error term $e_{ht} = D_{ht}b_h + \varepsilon_{ht}$.

However, if we expect that the locations of the intervention (D) were selected based on where they would achieve the highest potential effect (high b_h), then $E(b_h|D) \neq 0$. In order to circumvent the problem caused by endogenous intervention placement, we treat the heterogeneous effect b_h as omitted variable, and include its expected value conditional on regressors D and X in the regression following Mundlak (1978), Chamberlain (1982, 1984) and Wooldridge (2005, 2010). We are interested in controlling for the component of the heterogeneous effect b_h that is correlated with the regressors D and X, and leave the uncorrelated random component as part of the error term.

Notice that when selection on gains occurs, the value of the intervention variable D contains valuable information about the conditional expectation of the treatment effect, $E(\beta_h|D)$, based on the valuation of the implementing organization. We assume that the implementers are able to observe relevant characteristics of the target population that are unobservable to the researchers, and use this information to place the interventions in an optimal way.

We calculate the conditional expectation of b_h by allowing the heterogeneous parameters to depend on the regressors in a parametric way. We approximate b_h as a linear combination of the group means of regressors D and X over time ($\bar{D}_h$ and $\bar{X}_h$):

$$E(b_h|D,X) = (\bar{D}_h - \mu_{\bar{D}_h})\xi + (\bar{X}_h - \mu_{\bar{X}_h})\psi \tag{7}$$

where $\bar{X}_h = 1/T \sum_{t=1}^{T} X_{ht}$ is the household mean of variable X over time, and $\mu_{\bar{X}_h} = E(\bar{X}_h)$ is the expectation of the household means, which is estimated

by the sample mean over all households assuming a representative sample: $\hat{\mu}_{\bar{X}_h} = 1/H \sum_{h=1}^{H} \bar{X}_h$. The variables for D are defined similarly. The demeaning of the right-hand side terms in (7) ensures that $E(b_h) = 0$.[5] Importantly, the method requires that the regressors X and D are strictly exogenous conditional on the homogeneous (α, β and θ) and heterogeneous regression coefficients (b_h and η_h).

Because the assumptions made about the distribution of the heterogeneous parameters in (7) are crucial for the identification of β, it is useful to spend some time discussing their implications. Assuming a linear approximation in (7) implies monotonicity in the heterogeneous parameter, i.e. $E(b_h|\bar{D}', X) \geq E(b_h|\bar{D}'', X)$ or $E(b_h|\bar{D}', X) \leq E(b_h|\bar{D}'', X)$ for all $\bar{D}' > \bar{D}''$. This means that the interventions are implemented at locations in the order of the expected size of the treatment effects. If this assumption is not reasonable for the data, the identification of β fails based on (7).

Regarding the structure assumed in (7), note that Chamberlain (1980) assumed a more general structure for the distribution of the heterogeneous parameters using a linear combination of the regressors ($Z = \{D, X\}$) for all time periods: $E(b_h|Z) = Z_{h1}\phi_1 + Z_{h2}\phi_2 + \cdots + Z_{hT}\phi_T$. Regression (7) uses the group mean of the regressors over time, which is a special case of the general formulation, where each time period receives the same weight ($\phi_1 = \cdots = \phi_T = 1/(T+1)$). Another special case is the specification of Elbers and Gunning (2014), who used the changes in the regressors to predict the heterogeneous parameters ($\phi_1 = -1, \phi_T = 1$). These two sets of assumptions have quite different implications for the interpretation of the model: using means of regressors over time, one assumes that the heterogeneity is driven by the (relative) mean level of household characteristics, while using the changes in the regressors, one assumes that the heterogeneity is driven by changes in the household characteristics. For example, if we expect that (for the researchers) unobserved community or household characteristics drive the selection process and not the

[5] Wooldridge (2010) points out that when using unbalanced panel sample, the means above have to be calculated over the sample used for estimation.

changes in characteristics, then the natural choice is to use the group means of the regressors, especially the interventions, to predict the value of the heterogeneous parameters as done in (7).

Assuming that (7) is correct and substituting into (5), we get the Correlated Random Slopes model

$$\begin{aligned} E(Y_{ht}|D,X) =& \alpha_t + D_{ht}\beta + X_{ht}\theta + D_{ht} \otimes (\bar{D}_h - \mu_{\bar{D}_h})\xi + \\ & D_{ht} \otimes (\bar{X}_h - \mu_{\bar{X}_h})\psi + E(\eta_h|D,X) + E(\varepsilon_{ht}|D,X) \end{aligned} \tag{8}$$

where $D \otimes X$ denotes all interaction terms between D and X. Assuming that $E(\varepsilon_{ht}|D,X,\eta_h) = 0$ or that no selection occurred on time-varying unobservable variables, we can estimate (8) by fixed effects estimation.

Regression (8) allows us to test the presence of correlated random slopes. A Wald-test (or Likelihood Ratio-test) of $H_0 : \boldsymbol{\xi} = 0$ tests for selection on gains in the intervention placement (correlation between the treatment effect and the intervention), while $H_0 : \boldsymbol{\psi} = 0$ tests the heterogeneity of the treatment effect conditional on the mean of observed household specific variables (Wooldridge, 2010).

2.2 Treatment effects of interest

Fixed effects estimation of (8) results in consistent estimates for β (ATE) provided that the conditions and assumptions for ATE in Table 1 are satisfied.[6]

These rather strong assumptions can be weakened when estimating the Total Program Effect as shown in the last column of Table 1. Elbers and Gunning (2014) argue that for policy-makers the total effect of the program inclusive of the effects of selection may be more relevant. In the current setting, using a population representative sample, the generalised version of TPE ($E(\Delta D_{hT}\beta_h)$) can be calculated as the change in the conditional expectation of the terms involving D_{ht} in equation (8):

[6] Wooldridge (2005) shows that β in (8) estimates the ATE in the eligible population when the model is correctly specified.

Table 1: Assumptions required for identifying ATE, ATET and TPE

Assumptions	**ATE**	**TPE & ATET**
A1. $E(\varepsilon_{ht}\|D, X) = 0$	Yes	Yes
A2. D and X are strongly exogenous†	Yes	No
A3. $E(b_h\|\bar{D}', X) \geq (\leq) E(b_h\|\bar{D}", X)$ for all $\bar{D}' > \bar{D}"$	Yes	No
A4. Common trend α_t for all h	Yes	Yes
Further conditions		
C1. $T \geq 2$	Yes, if D is not binary	Yes
C2. $T \geq 3$ and D is continuously rolled out	Yes, if D is binary	No

Note: The table shows whether the above 3 assumptions and 2 futher conditions are required for identifying ATE, TPE & ATET. The notation is based on equation 5.
† conditional on the regression coefficients.

$$TPE_T(\beta_h) = \overline{\Delta D_{hT}}\beta + \overline{\Delta D_{hT} \otimes (\bar{D}_h - \mu_{\bar{Q}_h})}\xi + \overline{\Delta D_{hT} \otimes (\bar{X}_h - \mu_{\bar{X}_h})}\psi \qquad (9)$$

where $\overline{\Delta D_{hT}} = 1/H \sum_{h=1}^{H}(D_{hT} - D_{h1})$ or the mean of the change in the intervention variable over the households, and similarly for the other expressions involving $\overline{\Delta D_{hT}}$.

Similarly, we can also calculate the Average Treatment Effect on the Treated, which in the current setting will also include the effects of selectivity. Hence, it is not directly comparable to the ATET calculated in RCTs or other settings where the heterogeneity of the treatment effect (β_h) is independent of the intervention placement (D_{ht}). Here, ATET is calculated as

$$ATET(\beta_h) = \beta + \frac{\overline{(\Delta D_{hT} \otimes (\bar{D}_h - \mu_{\bar{D}_h})}\xi + \overline{\Delta D_{hT} \otimes (\bar{X}_h - \mu_{\bar{X}_h})}\psi}{\overline{\Delta D_{hT}}} \qquad (10)$$

The delta method[7] can be used to calculate the standard error of both treatment effects. For example, the variance of TPE is calculated as $Var(TPE(\beta_h)) =$

[7] On the delta method see, for example, section 5.6 of Davidson and MacKinnon (2004), section 7.2.8. of Cameron and Trivedi (2005) or section 3.5.2 of Wooldridge (2002).

$g'V(\beta,\xi,\psi)g$, where $g = (\overline{\Delta D_{hT}}, \overline{\Delta D_{hT} \otimes (\bar{D}_h - \mu_{\bar{Q}_h})}, \overline{\Delta D_{hT} \otimes (\bar{X}_h - \mu_{\bar{X}_h})})$ is a column vector and $V(\beta,\xi,\psi)$ is the partition of the variance-covariance matrix from estimating (8).

Finally, it is important to point out that while the identification of the ATE requires that the model is correctly specified (in particular, the heterogeneity parameters in equation 7), the estimates for ATET and TPE are also valid when the model is misspecified (see Table 1). This is because their estimates depend on the value of $E(\Delta D_{hT}\beta_h))$ rather than β. Hence, we do not necessarily need to have a consistent estimate for ATE in order to estimate ATET and TPE.

3 Illustrations

3.1 The data generating process

In order to demonstrate the use of the proposed method, we use a simple example with a data generating process (DGP) that follows the intuition of the heterogeneous treatment effect arising from the logic of selection on gains as described in the previous section:

$$Y_{ht} = \beta_h D_{ht} + \theta X_{ht} + \eta_h + \varepsilon_{ht} \tag{11}$$

$$\beta_h = \beta + \gamma(\bar{X}_h - \mu_{\bar{X}}) + \nu_h \tag{12}$$

$$\tilde{\beta}_h = \beta_h + n_h \tag{13}$$

$$D_h = \begin{cases} (0,1,1) & \text{if } \tilde{\beta}_h > B_2 \\ (0,0,1) & \text{if } B_3 < \tilde{\beta}_h \leq B_2 \\ (0,0,0) & \text{if } \tilde{\beta}_h \leq B_3 \end{cases} \quad \text{for } t = (1,2,3) \tag{14}$$

The variables in the outcome equation (11) follow that of regression (4). In (12), the heterogeneous treatment effect (β_h) is randomly determined, independently of the intervention. The program implementers observe a noisy measure of the tretament effect ($\tilde{\beta}_h$ in equation 13), and decide based on this value

where to implement the interventions D_{ht}. Following (14), the interventions are first implemented at those households where the intervention is expected to have the highest impact based on the noisy observation. The thresholds B_2 and B_3 determine the share of households receiving the intervention before period $t = 2$ and before period $t = 3$. The noise to signal ratio in (13) (σ_n^2/σ_ν^2) and thresholds B_2 and B_3 in (14) determine the correlation (ρ) between the treatment effect β_h and the intervention placement $\bar{D}_h$.

The DGP also includes an exogeneous variable X, which may affect the outcome directly (θ in equation 11) but also through its effect on the size of the treatment effect (γ in equation 12). For the case of simplicity, we use time-constant $\bar{X}_h$, hence is not identified when using fixed effects estimation. Therefore, we set $\theta = 0$. We use the standard normal distribution to generate $\bar{X}_h$.

In the following, we fix $\beta = 1$, $\theta = 0$, $\varepsilon_{ht} \sim N(0,1)$, $\nu_h \sim N(0,1)$ and $B_2 = 2$, and investigate the effect of changes in the other parameters (σ_n^2, γ, B_3) on the performance of the fixed effect and correlated random slopes estimators of the ATE (β), ATET ($E(\beta_h|D_{Th} = 1)$) and Total Program Effect ($TPE = E(\beta_h D_{Th})$).[8]

3.2 Estimating equations

We use the within transformation to eliminate the fixed effect η_h in the outcome equation (11), and estimate the following two regression equations with $T = 3$ periods and a sample of $H = 1000$ households:

$$FE: Y_{ht} = \beta_{FE} D_{ht} + e_{ht} \tag{15}$$

$$CRS: Y_{ht} = \beta_{CRS} D_{ht} + \xi_{CRS} D_{ht}(\bar{D}_h - \mu_{\bar{D}}) + \psi_{CRS} D_{ht}(\bar{X}_h - \mu_{\bar{X}}) + u_{ht} \tag{16}$$

From the fixed effect regression (15) we have $ATET(FE) = \hat{\beta}_{FE}$ and the overall effect of the program can be calculated as $\hat{\beta}_{FE}\overline{D_{hT}}$.

[8] All variables are generated independently from the normal distribution.

The treatment effects using the correlated random slope regression (16) are calculated as:

$$ATE(CRS) = \hat{\beta}_{CRS} \tag{17}$$

$$ATET(CRS) = \hat{\beta}_{CRS} + \hat{\xi}_{CRS}\frac{\overline{D_{hT}(\bar{D}_h - \mu_{\bar{D}})}}{\overline{D_{hT}}} + \hat{\psi}_{CRS}\frac{\overline{D_{hT}(\bar{X}_h - \mu_{\bar{X}})}}{\overline{D_{hT}}} \tag{18}$$

$$TPE(CRS) = \hat{\beta}_{CRS}\overline{D_{hT}} + \hat{\xi}_{CRS}\overline{D_{hT}(\bar{D}_h - \mu_{\bar{D}})} + \hat{\psi}_{CRS}\overline{D_{hT}(\bar{X}_h - \mu_{\bar{X}})} \tag{19}$$

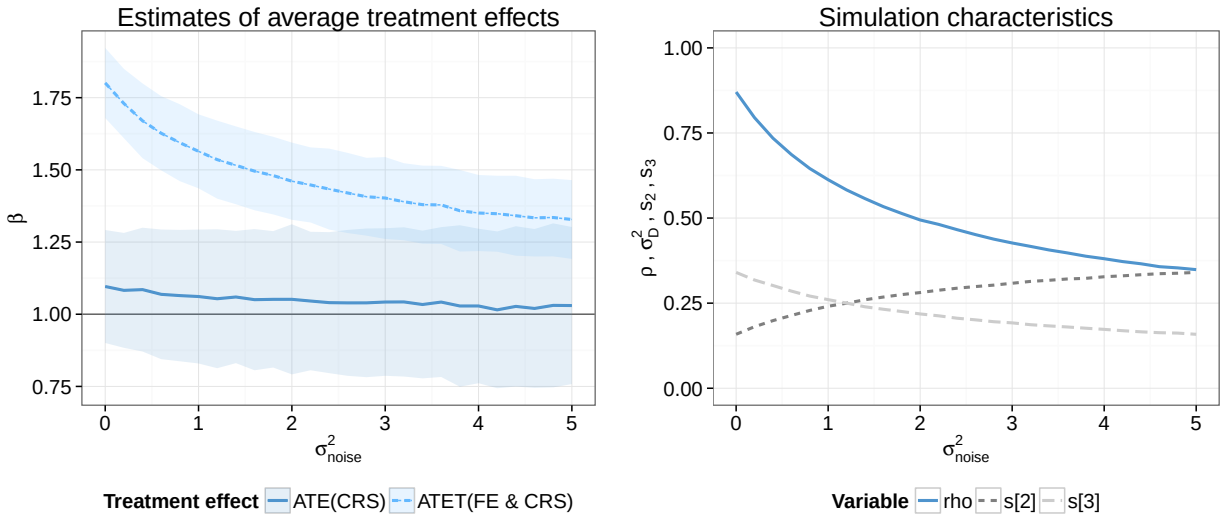
In the next subsections, we investigate the distribution of these treatment effect estimates using $R = 500$ replications as we change the parameters that determine the program allocation (D_{ht}) and the treatment effect heterogeneity (β_h).

3.3 Noisy inference

First, we investigate how the magnitude of noise in the implementer's observation of the treatment effect ($\tilde{\beta}_h$ in equation 13) affects the treatment assignment and the treatment effect estimates. For simplicity, we set $\gamma = 0$, thereby excluding the exogeneous regressor $\bar{X}_h$ from the DGP (and also from the CRS regression (16) with $\psi_{CRS} = 0$). In addition, we also fix the implementer's thresholds for intervention delivery at $B_2 = 2$ and $B_3 = 1$.

The first thing to notice on the right side of Figure 1 is the high correlation (ρ) between the treatment effect and the intervention placement (solid line). This is influenced by the observations of the implementers in equation (13). The correlation decreases as the noise-to-signal ratio (σ_n^2/σ_ν^2) in the implementers' observations increases. It also changes the share of people exposed to treatment in period 2 (s_2) and 3 (s_3), but the total share of population with the intervention remains stable close to 50%.

The left side figure shows that the CRS regression estimates the ATE ($\beta = 1$) reasonably as the DGP is still in line with the identifying assumption of (7). The FE and CRS estimates of ATET are identical in this case. The ATET deviates from ATE due to the large selection effect present in the DGP. However, as the noise



$\beta = 1$, $\gamma = 0$, $\sigma_e^2 = 1$, $\sigma_\nu^2 = 1$, $B_2 = 2$, $B_3 = 1$, H = 1000 , T=3 and R = 500

Note: For CRS and FE estimates, 95 percent confidence interval included around mean

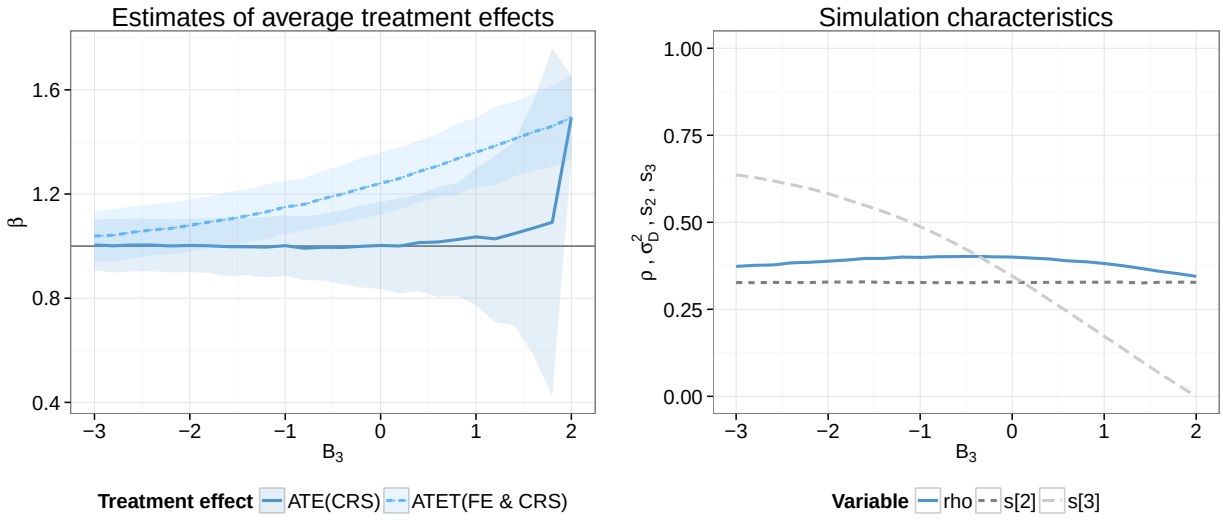
Figure 1: The effect of noisy inference

in the implementer's observations increases, her ability to select households with the highest treatment effect decreases. As a consequence, the ATET also decreases closer to the ATE. Notice also that the standard error of the estimate of ATE is larger than that of ATET, and it increases as the correlation between the intervention variable and the treatment effect decreases.

3.4 Changing thresholds

In the following, we fix the noise-to-signal ratio at σ_n^2/σ_ν^2 =4. In Figure 2 we investigate the effect of changes in the lower threshold for inclusion in the intervention in period 3 (B_3 in equation 14), while keeping the threshold for treatment in period 2 constant at B_2 =2.

The right panel in Figure 2 shows that the share of population receiving the intervention in period 2 is constant s_2=33%, while in period 3 the share goes from s_3=64% when $B_3 = -3$ to s_3 =0% when $B_3 = B_2 = 2$. Hence, when $B_3 = -3$ there are only very few households that do not receive the intervention, while when $B_3 = 2$ there are no households that receive the intervention in period 3.



$\beta = 1$, $\gamma = 0$, $\sigma^2_e = 1$, $\sigma^2_v = 1$, $B_2 = 2$, $\sigma^2_{noise} = 4$, $H = 1000$, T=3 and $R = 500$

Note: For CRS and FE estimates, 95 percent confidence interval included around mean

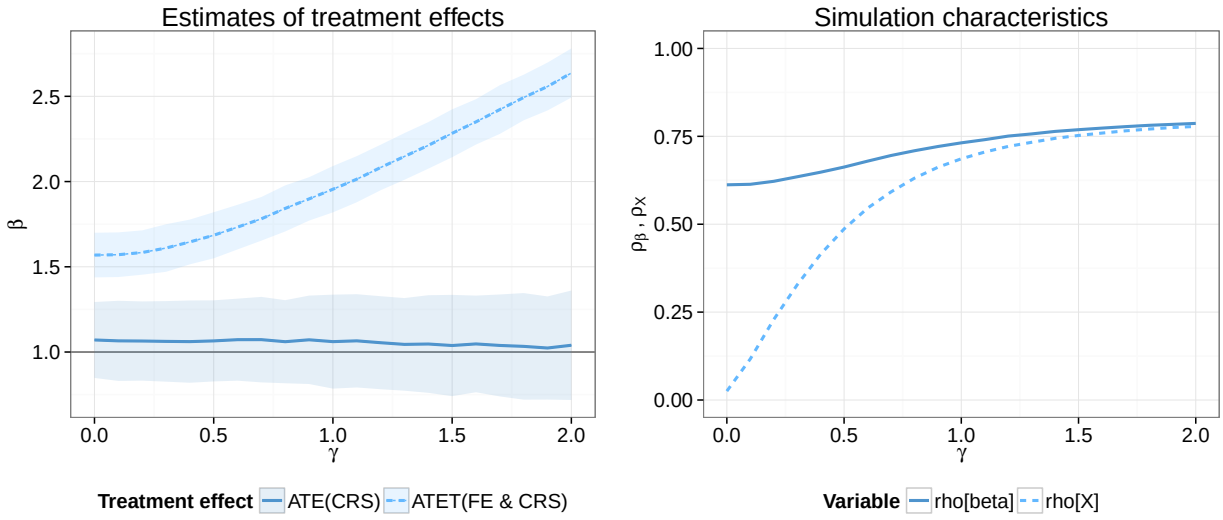
Figure 2: The effect of changing thresholds

When the whole population receives the intervention, there is no selection effect, and the ATET approaches the ATE (left panel of Figure 2). The selection effect and the ATET increases as more households are systematically excluded from the intervention as a result of increasing the threshold for inclusion for period 3 (B_3).

The CRS estimate of ATE remains unbiased until s_3 approaches close to zero. However, its standard error increases as s_3 diminishes. At the extreme of $s_3 = 0$, the estimate the CRS model collapses to the FE estimate because data from period 3 provide no information about the treatment effect (see the discussion in Appendix B).

3.5 Selection on observables

Now, we turn to discussing the effect of selection on observables. We fix B_3 =1 and σ^2_n/σ^2_ν =1, and investigate the effect of changes in γ in equation (12). As mentioned above, X_h is observable to both the implementer and the researcher and it is time-constant. Hence, θ in the outcome equation (11) is not identified using fixed effects estimation. Regression equations (15) and (16) remain valid



$\beta = 1$, $\sigma^2_e = 1$, $\sigma^2_v = 1$, $\sigma^2_{noise} = 1$, $B_2 = 2$, $B_3 = 1$, H = 1000 , T=3 and R = 500

Note: For CRS and FE estimates, 95 percent confidence interval included around mean

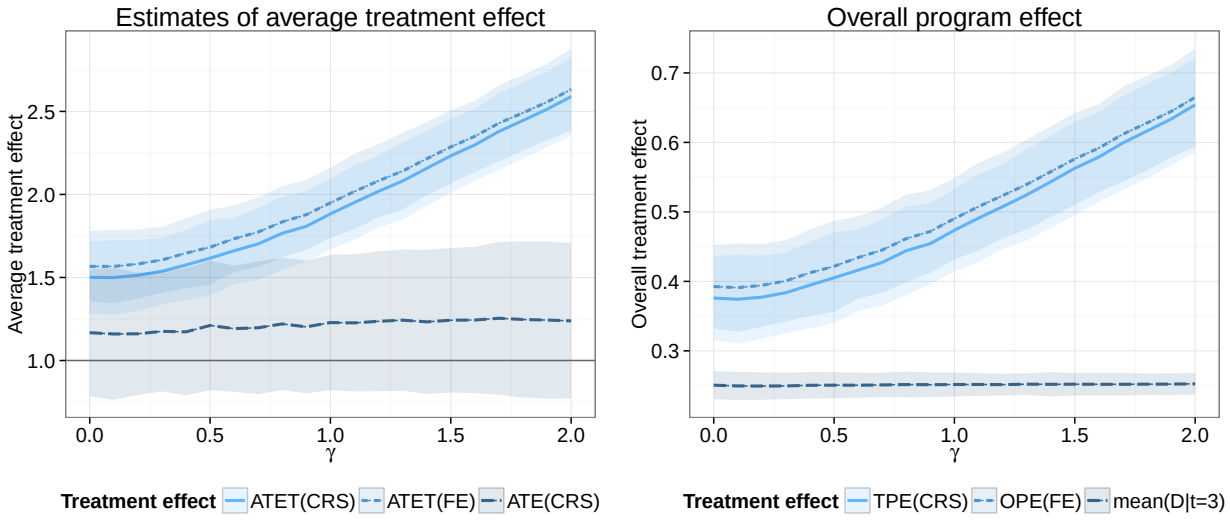
Figure 3: The effect of observable treatment effect heterogneity

for the estimation of the FE and CRS regressions.

Figure 3 displays the simulation results using iR replications as γ increases from 0 to 2. The exogeneous variable $\bar{X}_h$ follows the standard normal distribution. The right panel of the figure shows that the correlation between the intervention and the treatment effect (ρ or ρ_β here) is quite high due to the low noise-to-signal ratio. As γ increases from zero, so does the correlation between the intervention and the control variable $\bar{X}_h$ (ρ_X), which quickly approaches the size of ρ_β. The variance of the treatment effect (β_h) also increases with γ ($var(\beta_h) = \sigma^2_\nu + \gamma^2\sigma^2_{\bar{X}}$). As a consequence, a larger share of the population is exposed to the intervention in period 2, but the overall share of population receiving the intervention remains unchanged around 1% (not shown). The ATET also increases with γ (left panel) as a further consequence. The left panel of the figure shows that the CRS estimate of ATE remains unbiased.

3.6 Non-binary treatment

Now, we modify the DGP (11)-(14) by imposing a non-binary treatment variable. For simplicity, we define the intensity of the treatment using the uniform



$\beta = 1$, $\sigma_e^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{noise}^2 = 1$, $B_2 = 2$, $B_3 = 1$, $H = 1000$, T=3 and R = 500

Note: For CRS and FE estimates, 95 percent confidence interval included around mean

Figure 4: The effect of non-binary treatment and observable treatment effect heterogneity

distribution as $i_h \sim U(0,1)$. Hence, (14) becomes

$$D_h = \begin{cases} (0, i_h, i_h) & \text{if } \tilde{\beta}_h > B_2 \\ (0, 0, i_h) & \text{if } B_3 < \tilde{\beta}_h \leq B_2 \text{ for } \quad t = (1,2,3) \\ (0, 0, 0) & \text{if } \tilde{\beta}_h \leq B_3 \end{cases} \tag{14N}$$

This could be though of, for example, as different levels of subsidized co-payments or the distance to a distribution center. First, we treat the household specific intervention intensity independent from the household specific treatment variable. All other model parameters are as in section 3.5.

Figure 4 shows the simulation results for the same scenario as in Figure 3 but with a non-binary intervention variable as in equation (14N). The left panel of the figure shows the average treatment effects (ATE and ATET), and the right panel the overall treatment effects (TPE and the overall treatment effect from the FE regression) and mean value of the non-binary intervention variable in period 3. The latter is lower compared to the binary case.

Unlike the previous examples with the binary intervention variable, here,

the ATET from the FE and CRS regressions differ from one another, albeit not significantly in the current setting. In addition, the CRS estimate of ATE appears to be more biased compared to the binary case.

3.7 Non-binary correlated treatment

Here, we further modify the intervention assignment (14) to allow the intervention intensity to be correlated with the treatment effect:

$$\bar{D}_h = \begin{cases} \left((1-\lambda)i_h + \lambda\Phi(\beta_h)\right)(0,1,1) & \text{if } \tilde{\beta}_h > B_2 \\ \left((1-\lambda)i_h + \lambda\Phi(\beta_h)\right)(0,0,1) & \text{if } B_3 < \tilde{\beta}_h \leq B_2 \text{ for} \quad t = (1,2,3) \\ (0,0,0) & \text{if } \tilde{\beta}_h \leq B_3 \end{cases} \tag{14C}$$

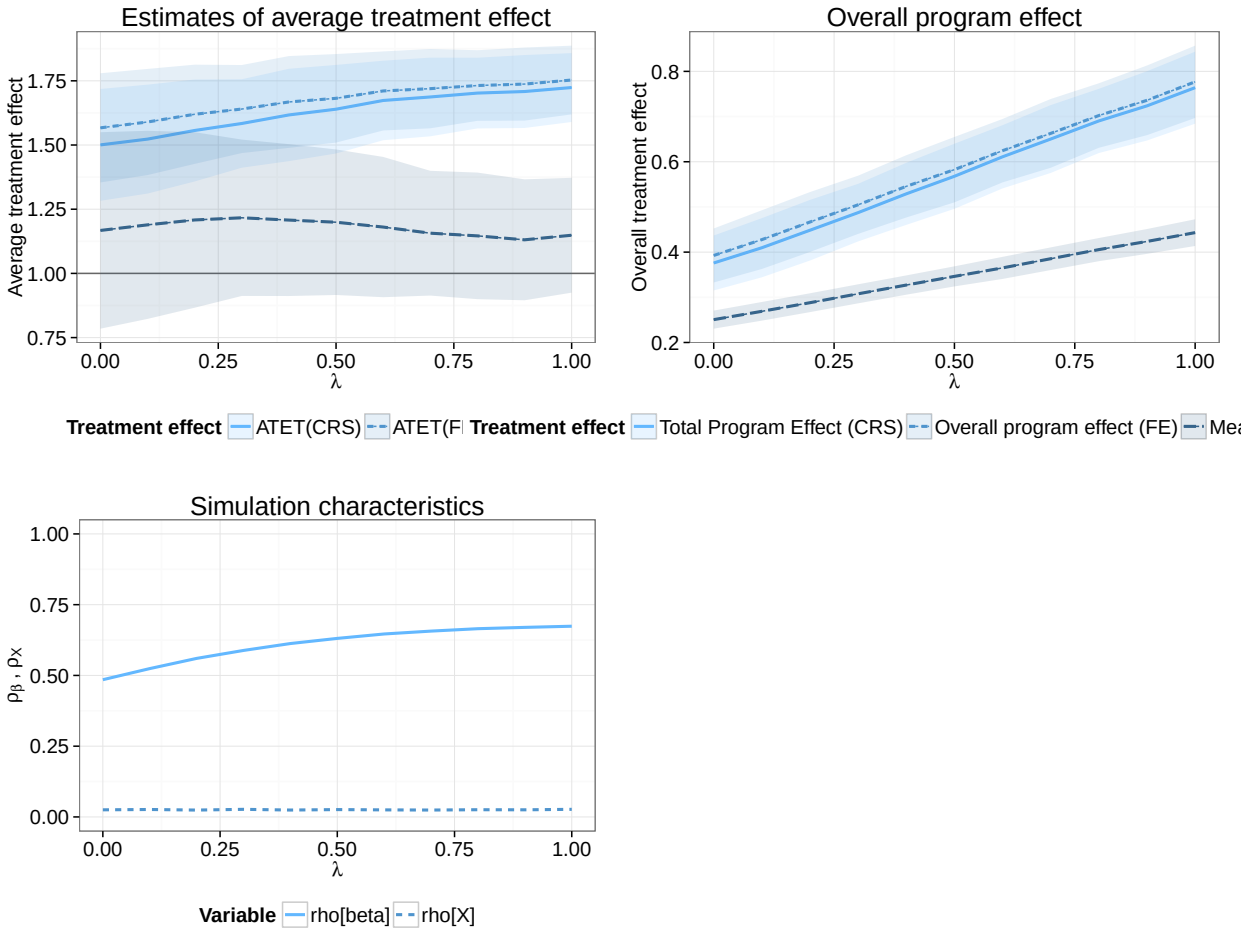
where $\Phi(\beta_h)$ is the cummulative density of β_h and it accounts for the correlation between the intervention intensity and β_h. The cummulative density is used here because it provides a value between (0,1) similarly to $i_h \sim U(0,1)$. $\lambda \in (0,1)$ controls the size of correlation between the intervention intensity and β_h; the correlation increases with λ.

Following up on the previous analogues, this artifical example could be interpreted as providing larger subsidized co-payments or locating the distribution center closer to those for whom the treatment effect would be higher.

Figure 5 displays the simulation results as a function of λ, while setting $\gamma = 0$ for simplicity. The results show that the ATET increases as the correlation between the intervention intensity and treatment effect increases with λ. The mean value of the intervention variable also increases with λ (right panel, dark dashed line), leading to increases in the total/overall program effect.

3.8 Time-varying treatment effects

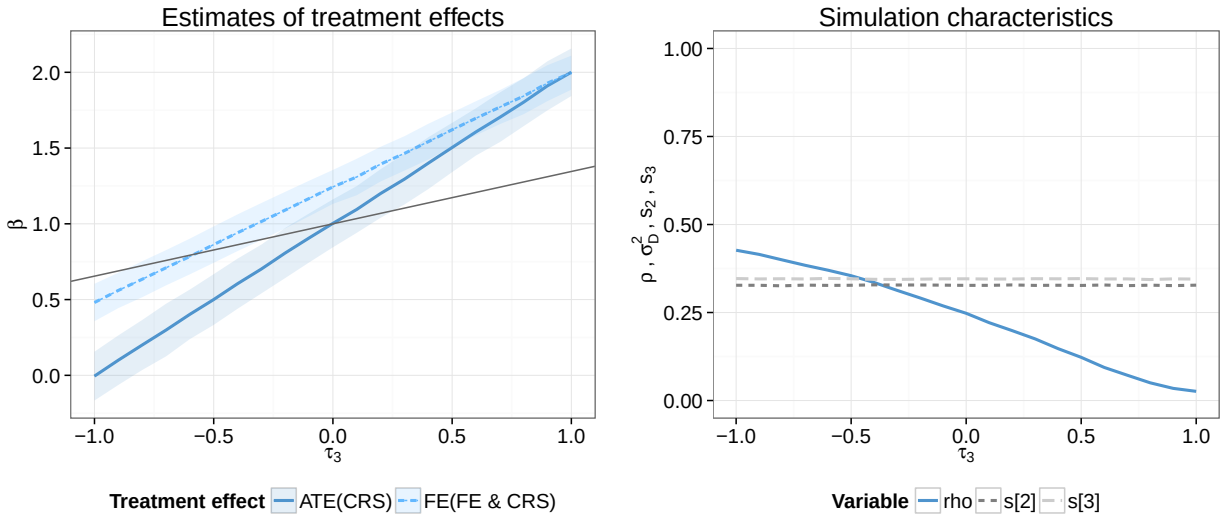
In this section, we investigate how misspecification in the regression model affects the estimates. A particularly relevant source of misspecification is when



$\beta = 1$, $\sigma_e^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{noise}^2 = 1$, $\gamma = 0$, $B_2 = 2$, $B_3 = 1$, $H = 1000$, T=3 and $R = 500$

Note: For CRS and FE estimates, 95 percent confidence interval included around mean

Figure 5: Observable treatment effect heterogneity and the performance of FE and CRS estimators



$\beta = 1$, $\sigma^2_e = 1$, $\sigma^2_v = 1$, $\sigma^2_{noise} = 4$, $B_2 = 2$, $B_3 = 0$, H = 1000 , T=3 and R = 500

Note: For CRS and FE estimates, 95 percent confidence interval included around mean

Figure 6: The effect of time-varying treatment effects

the treatment effects are not constant over time. To demonstrate this case, we modify the outcome equation (11) as

$$Y_{ht} = (\beta_h + \tau_t)D_{ht} + \eta_h + \varepsilon_{ht} \tag{11T}$$

Hence, now the actual treatment effect also depends on the time when it is implemented. We keep the rest of the DGP as in (12)-(14) with binary treatment effect.

We are specifically interested in looking at the effect of τ_3 while keeping $\tau_2 = 0$. We fix the other model parameters at $B_2 = 2$, $B_3 = 0$, $\sigma^2_n = 4$, $\beta = 1$, $\gamma = 0$, $\sigma^2_\varepsilon = 1$ and $\sigma^2_\nu = 1$.

Figure 6 shows the results. For this example, it is not as straightforward to calculate the ATE as it changes over time. In the figure on the left, we show the line that corresponds to ATE weighted by the share of household exposed to the interventions in period 2 and 3 ($E(ATE) = \beta + s_3\tau_3$) as a reference line. The estimates of ATE are severely biased. The CRS regression underestimates ATE when $\tau_3 < 0$ and overestimates it when $\tau_3 > 0$. We observe that as the cor-

relation between the treatment effect ($\beta + \tau_t$) and the intervention placement decreases with increasing the idiosyncratic treatment effect (τ_3), the estimates of the CRS regression converge to the FE estimates. When the correlation coefficient (ρ) becomes negative, the bias of the CRS estimate exceeds the bias of FE regression.

Hence, in case the CRS regression model is misspecified, i.e. the source of correlation is not only the selection on gains, its estimate of the ATE will be biased. However, the estimates of the ATET remain valid as discussed in section 2.

When only three rounds of data are available, the CRS regression model is just identified. Therefore, it is not possible to test whether the assumptions of the CRS model (equation 7) are valid, or that there are other sources of heterogeneity in the DGP, for example, time-varying treatment effects. Hence, in the real world, it is essential to have a good understanding of the implementation of the program, and we need additional information to argue for the validity of one or the other set of identifying assumptions.

4 Example of the One Million Initiative

To demonstrate the use of the proposed method on actual data, we turn to analyzing the effectiveness of a sanitation intervention on latrine ownership and proper hand-washing in Mozambique. We use this example because it represents a realistic scenario for the use of the proposed method.

4.1 The One Million Initiative

The One Million Initiative (OMI) was implemented between 2006-2013 in 18 districts of Manica, Sofala and Tete provinces of Mozambique as a cooperation between UNICEF and the governments of Mozambique and the Netherlands. The program was to reach one million people in poor rural areas (40% of the population in the program area) and to provide them with means to use

safe and sustainable drinking water and hygienic sanitation facilities. Specifically, OMI carried out community water supply interventions by creating new boreholes and training water committees on maintenance, and community-based sanitation and hygiene education using the Community-Led Total Sanitation approach (CLTS). The interventions were gradually implemented between 2008 and 2013. Community participation was an essential component of the interventions.

Here, we only look at the effectiveness of the sanitation interventions (CLTS). The aim of CLTS was to eradicate open defecation by triggering communties to build latrines for themselves though awareness raising.[9] No subsidies were provided to promote latrine construction, which were almost always traditional latrines built from locally available material. The intervention also promoted proper hand-washing after defecation.

The CLTS interventions were implemented by local NGOs in each program district. The NGOs communicated the program to the communities and decided on the location of the CLTS intervention in agreement with the communities. Annual targets were set for the number of triggered communities. The triggered communities could apply to certify that they were free of open defecation during the annual Open Defecation Free (ODF) communities evaluation campaign. The NGOs were financially rewarded for the number of ODF communities in their district. Therefore, it was in their best interest to implement the CLTS intervention in communities that had a high likelihood of success.

Regarding the water supply interventions, the communities had to request these through the NGOs. However, the district government decided on the location of these. In practice, there was a substantial overlap between the two types of interventions. For further information about the interventions, see Vigh et al. (2016).

[9] See Kar and Chambers (2008) for more information about the implementation of Community Led Total Sanitation.

Table 2: Number of communities in treatment arms

	Early (2008-2010)	Late (2010-2013)	Total
CLTS	23	11	34
Comparison	22	22	22
Total	45	33	56

4.2 Data

For the evaluation of the One Million Initiative, survey data was collected in three rounds: August-October 2008 (baseline), August-October 2010 (midline) and July-August 2013 (endline). In each round, data was collected at 1600 households in 80 communities. These communities were not a priori assigned to intervention and control arms. Instead, the implementing organizations decided on the intervention locations and then reported these to UNICEF and the research team. Here, we use the data only on the 34 communities that received the CLTS intervention either between the baseline and midline (Early) or between the midline and endline (Late),[10] and 22 communities that did not receive any CLTS and water supply interventions (Comparison) (Table 2). Among the CLTS intervention communities, 23 also received the water supply intervention.

In general, the CLTS intervention was more effective when implemented together with the water suppply intervention. However, here we simplify the analysis by classifiying all communities that received the CLTS intervention in the same intervention arm without reference to the water supply intervention. The results are qualitatively similar when we separate the treatment arms by the presence of the water supply intervention.

In the regression analysis, we use data on 1066 households that were in-

[10] We omitted 7 communities that received the water supply intervention before the midline survey and the CLTS intervention only after the midline survey, because the timing of the interventions could interfere with the effectiveness of the CLTS intervention. The 17 communities that received only the water supply intervention are also omitted from the sample.

terviewed in at least two survey rounds in the 56 communities included in the data analysis (N=2951). The tables in Appendix A summarise the community and household characteristics.

4.3 Outcome variables

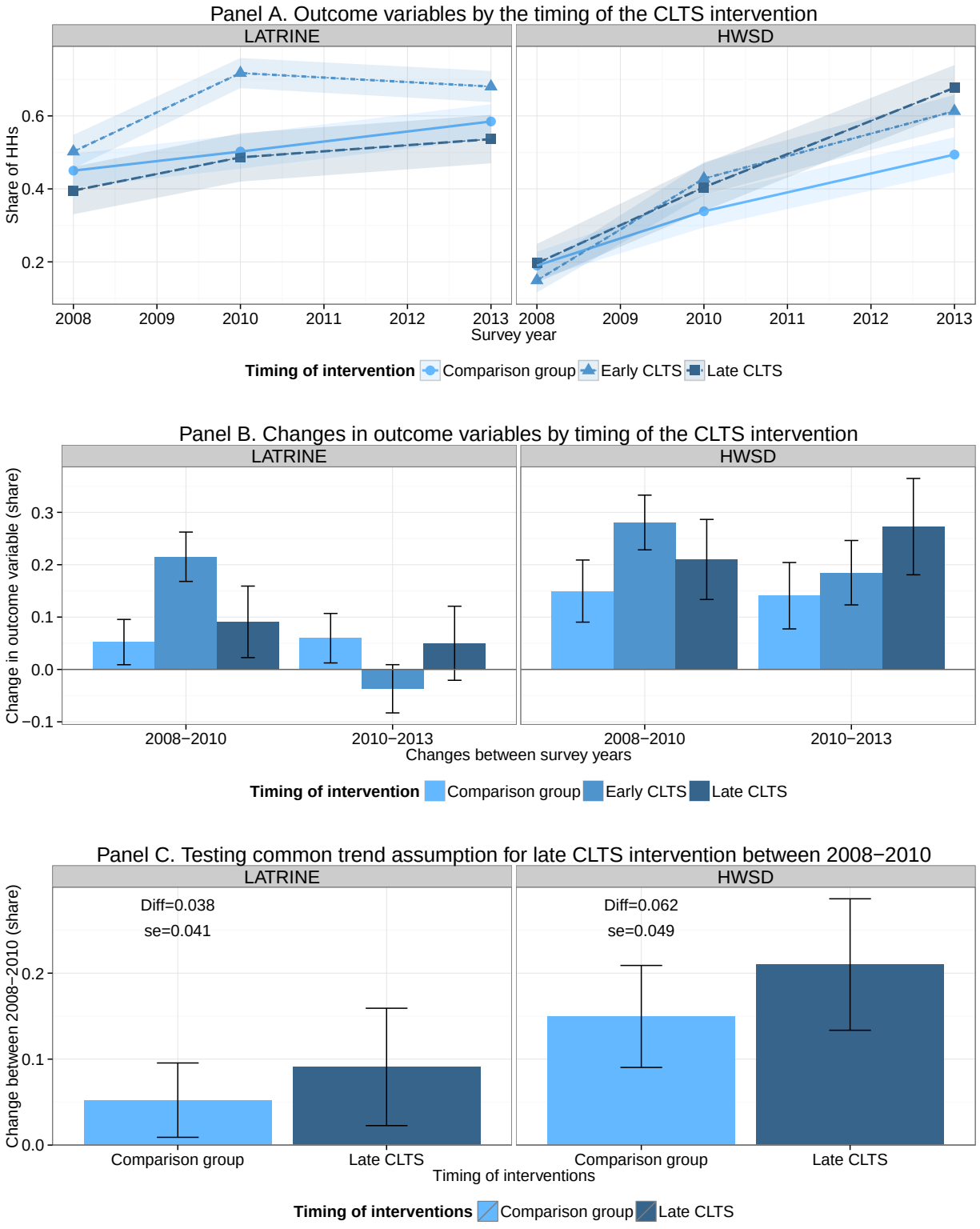
We present two outcome variables that were expected to be most affected by the CLTS intervention: latrine ownersip (LATRINE) and whether adults wash their hands with soap or ash after defecation (HWSD).[11] Both of these variables are defined at the household level. Latrine ownership was verified by interviewers and hand-washing is self-reported.

Figure 7 shows the changes in the two outcome variables for communities with the CLTS intervention and the comparison group. Panel A shows the development in the share of households owning a latrine (left) and washing hands with soap after defecation (right). The outcomes are shown separately for the Early and Late CLTS intervention groups and the comparison group. Panel B shows the corresponding changes in the outcome variables between the survey periods for the three groups. For an effective intervention, the outcome is expected to increase (more than in the comparison group) from 2008 to 2010 for the Early intervention group, and from 2010 to 2013 for the Late intervention group. In other periods, the intervention groups should follow the trend of the comparison group. The outcome on hand-washing follows the described pattern. Note that there is also a positive and siginificant trend in the comparison group for both outcome variables.[12]

The changes in latrine ownership are significantly smaller (diff=-16.5pp, p=0.0001) after the Late CLTS intervention compared to the Early intervention group. We argue that this pattern is the one we would expect if the implementing NGOs selected the location of the CLTS interventions based on where they

[11] Almost all latrines are owned and used by a single household. Therefore, latrine use is almost perfectly correlated with ownership.

[12] The positive trend in the comparison group can partially be attributed to other simultaneously running sanitation interventions and by increasing wealth levels.



Note: 95 percent confidence interval included around mean

Difference calculated controlling for HH size and wealth index. Std.error clustered at community level.

Figure 7: Changes in outcome variables and the timing of the CLTS interventions

expected the largest treatment effect in terms of the communities becoming open defecation free, as they were incentivized to do.

Our identification strategy requires that the comparison and intervention groups follow the same trend prior to the intervention (common trend). In Panel C of Figure 7, we test this assumption for the late CLTS intervention group on the changes of outcomes from 2008 to 2010. The late CLTS intervention group had a somewhat higher trend than the comparison group for both outcome variables. However, the difference is not significantly different from zero (p=0.352 and p=0.208) when controlling for observable household characteristics (household size and wealth index). Based on this finding, we cannot reject the validity of the common trend assumption. However, we are not able to test, and must assume that the same is true for the early CLTS intervention group.

4.4 Regression model

Because the CLTS intervention was implemented at the community level (c), we adjust the formulation of (8) to include heterogeneous treatment effect at the community level ($\beta_{c(h)}$) instead of at the household level (h). We estimate the following model:

$$\begin{aligned} E(Y_{ht}|D,X) =& \alpha_t + D_{c(h)t}\beta + X_{ht}\theta + D_{c(h)t} \otimes (\overline{D}_{c(h)} - \mu_{\overline{D}_c})\xi + \\ & D_{c(h)t} \otimes (\overline{X}_{c(h)} - \mu_{\overline{X}_c})\psi + E(\eta_h|D,X) + E(\varepsilon_{ht}|D,X) \end{aligned} \tag{20}$$

where $D_{c(h)t}$ shows whether there have been a CLTS intervention in the community up until period t in community c of household h. X_{ht} consists of time-varying household specific control variables (household size and wealth index).

The demeaned variables, $\overline{D}_{c(h)}, \overline{X}_{c(h)}$, are now defined at the community level. Hence, $\overline{X}_{c(h)} = 1/(TH_c)\sum_{t=1}^{T}\sum_{h=1}^{H_c} X_{ht}$ is the community mean of X over time, and $\mu_{\overline{X}_c} = E(\overline{X}_{c(h)})$ is the expectation of the community means in the population.

Table 3: Estimation results for latrine ownership

	Regression model		
	FE	CRS	CRS
	(1)	(2)	(3)
β:CLTS	0.0827***	0.0341	0.0331
	(0.0248)	(0.0351)	(0.0354)
ξ:CLTS		0.2396**	0.2250*
		(0.1203)	(0.1197)
Mean at baseline	0.471	0.471	0.471
Observations	2951	2951	2951
Adjusted R^2	0.039	0.040	0.043
ψ estimated	No	No	Yes
Prob($\psi = 0$)			0.964

Note: *p<0.1; **p<0.05; ***p<0.01
All regressions control for HH size, wealth index and time dummies.
HH FE regression results with standard errors corrected for clustering at community level.
Sample includes HHs participating in at least 2 survey rounds.
In (3) ψ contains HH size and wealth index.

4.5 Results

The CLTS implementing NGOs were only rewarded based on latrine ownership (if all households use a latrine) but no incentives were provided based on hand-washing. Therefore, we expect to find that selective intervention placement affected the treatment effects of latrine ownership more than of hand-washing.

Table 3 shows the estimation results for latrine ownership. The first column shows the results of estimating the standard fixed effects model ($\xi = 0$ and $\psi = 0$). Using the standard difference-in-difference specification, we would conclude that CLTS had a positive and significant effect on latrine ownership (coef=8.3pp, p=0.001). However, when we control for the correlation between the treatment effect and the intervention placement (column 2), the size

Table 4: Estimation results for hand-washing with soap

	Regression model		
	FE	CRS	CRS
	(1)	(2)	(3)
β:CLTS	0.1470***	0.1283***	0.1295***
	(0.0319)	(0.0466)	(0.0465)
ξ:CLTS		0.0930	0.0860
		(0.1449)	(0.1453)
Mean at baseline	0.180	0.180	0.180
Nr. observations	2925	2925	2925
Adjusted R^2	0.118	0.118	0.118
ψ estimated	No	No	Yes
Prob($\psi = 0$)			0.664

Note: *p<0.1; **p<0.05; ***p<0.01
All regressions control for HH size, wealth index and time dummies.
HH FE regression results with standard errors corrected for clustering at community level.
Sample includes HHs participating in at least 2 survey rounds.
In (3) ψ contains HH size and wealth index.

of the treatment effect is halved and insignificant (coef=3.4pp, p=0.332). At the same time, we observe a positive and significant correlation between the intervention variable and the treatment effect (p=0.047) as evidence for selective intervention placement, which is in line with the incentive structure. The last column of the table shows that the treatment effect on latrine ownership is not correlated with observable community characteristics (Prob($\psi = 0$)=0.964 jointly for average household size and wealth index).

Turning to the effectiveness of the CLTS intervention on hand-washing, Table 4 shows that the standard difference-in-difference method estimates the treatment effect at 14.7 percentage points (p=0.0000). Controlling for selective intervention placement, the estimated treatment effect changes only little (coef=12.8pp) as the correlation between the treatment effect and the

intervention placement is insignificant (Prob($\xi = 0$)=0.521). Other observable community characteristics are also uncorrelated with the treatment effect (Prob($\psi = 0$)=0.664).

Table 5: Treatment effects for latrine ownership and hand-washing

	Treatment effects				
	β_{FE}	ATE (β_{CRS})	$\beta_{FE} - \beta_{CRS}$	ATET	TPE
	(1)	(2)	(3)	(4)	(5)
LATRINE: CLTS	0.0827***	0.0331	0.0497	0.0834***	0.0223**
	(0.0248)	(0.0354)	(0.0433)	(0.0247)	(0.0109)
HWSD: CLTS	0.1470***	0.1295***	0.0176	0.1481***	0.0560***
	(0.0319)	(0.0465)	(0.0564)	(0.0316)	(0.0145)

Note: *p<0.1; **p<0.05; ***p<0.01
ATE, ATET and TPE is based on model with $\psi = 0$.
TPE is calculated at the aggregate level.
Standard errors corrected for clustering at community level.

Table 5 summarizes the treatment effect estimates based on the regression results. The results of the first two column have been discussed above. The estimates of the Average Treatment Effect are not significantly different from the standard difference-in-difference estimates (see column 3). However, one would draw different conclusions about the effectiveness of the CLTS intervention on latrine ownership based on the ATE and the traditional difference-in-difference estimates.

Policy makers may be more interested in the actual effects of the program than the hypothetical case when everyone or only a random sample received the intervention (ATE). Applying sampling weights to achieve a population representative distribution of the explanatory variables, the Total Program Effect shows that the CLTS intervention achieved an additional 2.2 percentage point increase in latrine ownership and 5.6 percentage point increase in proper hand-washing after defecation in the population of the One Million Inititative pro-

gram area. These effects incorporate the effect of selective intervention placement and the expanse of the intervention delivery. As they measure the impact of the interventions at the aggregate level, they are not comparable to the average treatment effects reported in Table . However, TPE can easily be used to quantify the benefits of the program, and compare it to the costs of the implementation of the CLTS intervention.

The Average Treatment Effect on the Treated in column 4 also includes selection effects. The estimates are close to the standard difference-in-difference estimates. However, they are weighted by the distribution of observable community characteristics (including the intervention placement) and their correlation with the treatment effect. In contrast, the fixed effects estimates are weighted by the variance-covariance matrix of the regressors.

Summarizing the results, we find evidence of selective intervention placement on the incentivized outcome (latrine ownerhip) but not on the non-incentivized outcome (hand-washing). For the former outcome, the conclusions about the general effectivess of the CLTS intervention are upward biased when not taking into account the effects of selective intervention placement (selection on gains).

5 Discussion

The estimation results in the application suggest strong evidence for strategic intervention placement in terms of latrine ownership but not for hand-washing practices. This makes sense from the perspective of CLTS implementing NGOs, who were incentivized on the overall level of latrine ownership. If the incentives are correctly aligned, then strategic intervention placement is also desirable in terms of the total effect of the program.

Given selective intervention placement, when can we use the Correlated Random Slopes model of (8) to estimate the average treatment effect? Using the CRS model, we control for heterogeneity of the treatment effect among the beneficiaries by using the variation in the timing of the interventions and

the differences in the distribution of the heterogneous parameter, i.e. we compare the distribution of the treatment effect between the beneficiary population that received the intervention before and after the first follow-up survey. Importantly, the estimation method assumes that the treatment effects change monotonously (and linearly) over time, meaning that, in expectation, the best (or worst) performers are selected first, then the second best (worst), etc. Hence, the distribution of the treatment effect among the beneficiaries that received the intervention later should lie between that of the early beneficiaries and the actual treatment effect (i.e., $E_{Early}(\beta_c) > E_{late}(\beta_c) > E(\beta)$). Therefore, it can be useful to plot the data and estimate the time-varying heterogeneous treatment effect model of first. These would give guidance whether the correlated random slopes model makes sense given the data.

However, it is equally important to have a good understanding of the implementation of the program in order to develop alternative hypotheses why the treatment effect could change over time. When only three rounds of data are available, it is not possible to test whether the treatment effects are time-varying or the intervention placement was selective. Hence, we need additional information to argue for the validity of one or the other set of identifying assumptions.

6 Conclusion

In this paper, we focused on the correlation between the treatment effect and the intervention placement in terms of selection on gains. Using the results of Wooldridge (2005, 2010) and Chamberlain (1980) we discussed how the correlated random slopes model can be used to estimate both the Average Treatment Effect and a generalized version of the Total Program Effect of Elbers and Gunning (2014).

The estimation of ATE relies on strong assumptions. The simulation results showed that when these assumptions are satisfied, the CRS regression provides an unbiased estimate of ATE. However, when other sources of heterogeneity

are introduced into the data generating process, notably time-varying treatment effects, the estimates of ATE are biased. Therefore, it has to be carefully considered whether the assumptions of ATE are reasonable for the case at hand.

The ATET and the TPE rely on less strong assumptions as they estimate the treatment effect (in the total program area in case of TPE) inclusive of the effects of selectivity. The resulting treatment effect estimates are the primary interest of policy makers when evaluating the benefits of the program. In addition, for forward looking policy makers, the comparison of ATET to credible estimates of ATE can give guidance with respect to the expected impacts of expanding the program.

Using the One Million Initiative as example, we presented a case where selective intervention placement was present in the intervention design: the implementing NGOs of the CLTS interventions were incentivized to carry out the interventions in communities where they expected the effect on latrine ownership to be the highest. Indeed, we find evidence of selection on gains for this outcome variable. Controlling for selection effects our conclusions changed about the effectiveness of the interventions. However, looking at another relevant outcome variable of the CLTS intervention, which was not incentivized for the NGOs (hand-washing with soap after defecation), we find no effect of selective intervention placement.

Based on these results, we recommend testing for the presence of correlated heterogeneous treatment effects when more than two rounds of data is available. As a first step, one could test for the presence of time-varying treatment effects, which could be the first sign that the treatment effect heterogeneity (which is almost always present) is correlated with an underlying structural factor. Equally important is to test the validity of the common trend assumption for intervention groups with multiple pre-intervention survey rounds.

References

Cameron, A. Colin and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Chamberlain, G (1982), "Multivariate Regression Models for Panel Data." *Journal of Econometrics*, 1, 5–46.

Chamberlain, G (1984), "Panel Data." *Handbook of Econometrics*, Volume 2, 1248–1318.

Chamberlain, Gary (1980), "Analysis of Covariance with Qualitative Data." *The Review of Economic Studies*, 47, 225–238, URL http://www.jstor.org/stable/2297110.

Davidson, Russel and James G. MacKinnon (2004), *Econometric Theory and Methods*. Oxford University Press, New York.

Elbers, Chris and Jan Willem Gunning (2014), "Evaluation of Development Programs: Randomized Controlled Trials or Regressions?" *World Bank Economic Review*, 28, 432–445.

Heckman, James J, Sergio Urzua, and Edward Vytlacil (2006), "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics*, 88, 389–432, URL http://www.mitpressjournals.org/doi/pdf/10.1162/rest.88.3.389.

Kar, Kamal and Robert Chambers (2008), *Handbook on Community-Led Total Sanitation*. URL http://www.communityledtotalsanitation.org/sites/communityledtotalsanitation.org/files/cltshandbook.pdf.

Mundlak, Y. (1978), "On the Pooling of Time Series and Cross Section Data." *Econometrica*, 46, 69–85.

Vigh, Melinda, Chris Elbers, and Jan Willem Gunning (2016), "Effectiveness of the One Million Initiative in Mozambique."

Wooldridge, Jeffrey M. (2002), *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

Wooldridge, Jeffrey M. (2005), "Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models." *The Review of Economics and Statistics*, 87, 385–390.

Wooldridge, Jeffrey M. (2010), "Correlated random effects models with unbalanced panels." *Mimeo*.

A Appendix: Descriptive statistics

Table 6: Household characteristics in the regression sample

	N obs	Mean 2008	Mean 2010	Mean 2013
Household size	2,951	5.729	5.249	5.602
Nr. children under 5	2,951	0.911	0.857	0.746
Wealth index	2,951	-0.033	-0.007	0.080
Share with education (age>14)	2,587	0.466	0.490	0.589
Latrine ownership	2,951	0.471	0.594	0.636
Hand-washing with soap after def.	2,925	0.180	0.392	0.585
Use of improved water points	2,951	0.155	0.381	0.463

Table 7: Community characteristics at baseline in the regression sample

	N obs	Mean 2008	S.d. 2008
Prevalence of water related diseases (past 6 months)	56	0.297	0.159
Availability of IWP	56	0.357	0.483
Avail. EP2 school	56	0.321	0.471
Avail. health unit	56	0.214	0.414
No cholera	55	0.527	0.504
Population 1-500	56	0.214	0.414
Population 501-1000	56	0.304	0.464
Population 1001-1500	56	0.161	0.371
Population 1501-2000	56	0.089	0.288
Population >2000	56	0.214	0.414

B Appendix: The importance of the third survey round

In section 2, we stated that our estimation method relies on $T \geq 3$ for identifying the ATET in the presence of correlated heterogeneous treatment effects. In this subsection, we demonstrate under which conditions we are able to separate the effect of selectivity (or correlated heterogeneity) from the ATET using the outcomes on latrine ownership.[13] First, we discuss the intuition and then look at example regression results.

Recall from section 2 that in the estimation regression (8) the term $D_{ct} \otimes (\bar{D}_c - \mu_{\bar{D}_c})\xi$ and $D_{ct} \otimes (\bar{X}_c - \mu_{\bar{X}_c})\psi$ are used to control for the heterogeneity of the treatment effect (b_c).[14] Among the variables D and X, the intervention variables (D) are of interest to us, as these are the variables that are presumably correlated with the heterogeneity in the treatment effect. Hence, in this section we focus the discussion around $D_{ct} \otimes \bar{D}_c$ and $t \otimes \bar{D}_c$.[15]

Let us, first, examine these expressions for $T = 2$ given a single binary treatment variable and a true baseline at $t = 1$. Then, for the intervention group $D_{c_I} = (0, 1)$ at $t = (1, 2)$ resulting in $\bar{D}_{c_I} = (0.5, 0.5)$, and for the comparison group $D_{c_C} = (0, 0)$ resulting in $\bar{D}_{c_C} = (0, 0)$. Now, observe that $D_{ct} \otimes \bar{D}_c = 0.5 D_{ct}$.[16] Due to this multicollinearity, we cannot identify the coefficient of $D_{ct} \otimes \bar{D}_c$. Hence, without further information about the structure of the heterogeneous effects, we are not able to control for the correlation between intervention placement and the heterogeneous treatment effects.

What if we have an additional observation before or after the interventions

[13] We selected latrine ownership for the example because we expect selectivity to affect this outcome the most.

[14] Note that for convenience we removed the reference to the households in the subscripts. Hence, c replaces $c(h)$.

[15] We omit $D_{ct} \otimes \mu_{\bar{D}_c}$ because $\mu_{\bar{D}_c}$ is constant over the sample, and, therefore, this term is perfectly correlated with D_c.

[16] Which follows as, at $t = (1, 2)$, $D_c \otimes \bar{D}_c$ is $(0, 0.5)$ and $(0, 0)$ for the intervention and comparison groups, respectively.

were implemented? Assume that we conduct a second follow-up survey after all interventions have been completed before the second survey. Now we have $T = 3$ with $D_{c_I} = (0, 1, 1)$ and $D_{c_C} = (0, 0, 0)$ indicating whether an intervention happened before time t. Importantly, we do not have observations with $D = (0, 0, 1)$. Now, $D_c \otimes \bar{D}_c$ is $(0, 0.67, 0.67)$ and $(0, 0, 0)$ for the intervention and comparison groups, respectively. Hence, $D_{ct} \otimes \bar{D}_c = 0.67 D_{ct}$. As a result, the additional survey round does not allows us to control for the correlation between the intervention placement and the treatment effect.

Finally, let us assume that the intervention is rolled out continuously between the first and third round of data collection. Hence, we have observations with $D = (0, 0, 1)$. Now, $D_c \otimes \bar{D}_c$ is $(0, 0.67, 0.67)$ for the early intervention locations, $(0, 0, 0.33)$ for the later intervention locations and $(0, 0, 0)$ for the comparison group. Due to the differences in $\bar{D}_c$ for the early and late intervention groups, we are no longer able to express $D_c \otimes \bar{D}_c$ as a linear combination of the intervention variable, time dummies and community fixed effects. Hence, we are able to control for the selectivity on the treatment effect (selection on gains).

It is important to point out that the above considerations do not apply for multivalued (continuous or discrete) variables in X. For these variables, the coefficients of $D_{ct} \otimes \bar{X}_c$ and $t \otimes \bar{X}_c$ can also be identified when $T = 2$.