Using Multisite Instrumental Variables to Estimate Treatment Effects and Treatment Effect Heterogeneity

Christopher Runyon

Background

Multisite randomized controlled trials (MSTs) are an appealing design for testing educational programs at scale because they "can be regarded as a 'fleet' of randomized experiments" (Raudenbush and Bloom, 2015, p. 477) or a "planned meta-analysis" (Bloom et al., 2017, p. 818). One popular MST design is a block-randomized controlled trial, where individuals within each block are randomized to treatment and control groups, like when students within each of several schools are randomly assigned to a new educational program (Gerber and Green, 2012). The results across multiple sites can be collected to estimate the average effect of the program and the treatment effect heterogeneity across sites (Bloom, 2005; Raudenbush and Bloom, 2015). Understanding how effects of an educational program naturally vary across sites is an important step toward better understanding in what populations and contexts the treatment is most effective and what features lead to effective programs (Raudenbush and Bloom, 2015; Zvoch, 2016; Reardon and Stuart, 2017; Olsen, 2017).

Estimating average treatment effects is straightforward when all individuals comply with their random treatment assignment, but this does not always occur. When this randomization is broken, a common analytical approach is to analyze the groups based on treatment assignment (the intent-to-treat effect; ITT). However, this estimand is the effect of being *assigned* to the treatment, and not the treatment *receipt*. A second, less-common method is analyze the groups as treated, although it is unclear what the treatment effect estimand represents in this case.

One proposed strategy to estimate the causal effects of program participation under imperfect compliance is to use instrumental variables estimation (IV; e.g., Angrist et al., 1996; Bloom, 1984, 2005). An instrumental variable is a variable that is related to the outcome of interest *only* through the treatment. In the context of an MST, treatment assignment is an instrument for treatment receipt. The IV estimand is the effect of the treatment for those that complied with their treatment assignment, the Complier Average Treatment Effect (CATE). The CATE is already being using to estimate treatment effects in educational research in the presence of noncompliance (Clark et al., 2015; Kim et al., 2011; Lynch and Kim, 2017; Bettinger and Baker, 2014; Borman et al., 2009; Boatman and Long, 2018). The CATE is also recognized as a valid measure of causal treatment effects by the Institute of Educational Sciences (Schochet and Chiang, 2009) and the most recent version of the What Works Clearinghouse (WWC) Standards Handbook now includes standards for reporting the CATE in randomized controlled trials with imperfect compliance (Version 4.0, Section IID; U.S. Department of Education, 2017).

There is a small (but growing) body of literature on investigating the performance of IV methods for estimating the CATE and CATE heterogeneity in MSTs. Raudenbush et al. (2012) proposed three estimators that use IV to estimate the CATE and CATE heterogeneity in MSTs. Reardon et al. (2014) reported on the performance of using IV to estimate treatment effects in MSTs when a key assumption is violated. But there has not yet been a thorough examination of these estimators in recovering the CATE and CATE heterogeneity.

Method

The present simulation study compares the ability of 5 different estimators to recover the CATE and CATE heterogeneity across a host of simulation conditions that resemble well-known educational programs (Weiss et al., 2017). The five estimators are:

- **AsTreated** - A multilevel model where the treatment and effect groups are defined by their observed treatment status.

- **ITT** - A multilevel model where the treatment groups are defined by their treatment assignment; this method is implemented as described in Bloom et al. (2017).

- **IV1** - The second multisite IV estimator described in Raudenbush et al. (2012), which uses an MST IV approach that pools dataacross sites to obtain a global estimate of the treatment effect and it variability.

- **IV2** - The third multisite IV estimator described in Raudenbush et al. (2012), which uses precision weighting of the IV estimates obtained at each site, treating the estimated within- and across-site variances as known.

- **IV3** - One of the bias-correcting IV estimators described in Reardon et al. (2014) that closely resembles the third estimator described in Raudenbush et al. (2012), except that shrunken estimates from the first-stage IV equation are used in the second-stage IV equation.

The data-generating parameters are the number of sites, within-site sample size, the degree of noncompliance, treatment effect magnitude, treatment effect heterogeneity, and the magnitude of bias induced by the noncompliance (i.e., selection bias). The exact parameter values are enumerated in Table 1. The R syntax for the data-generating model and implementation the five estimators can be found in Appendices A and B, respectively. The performance of the five estimators in recovering the treatment effect and treatment effect heterogeneity is evaluated by examining the bias and relative parameter bias across 500 replications of each of the simulation conditions[1].

[1]More replications across a larger set of data-generating parameters, including treatment allocation variance, compliance variance, and the control-group-treatment-effect correlation are currently underway and will be complete by 2020.

Results

Table 2 reports the relative parameter bias for both the CATE and CATE hetergeneity estimates for a selected set of simulation conditions. In these simulation conditions, all three multisite IV estimators are able to recover the CATE well. The ITT consistently underesimates the CATE, whereas the "AsTreated" method consistently overestimates the CATE. Of the three IV estimators, IV1 is able to best recover the treatment effect heterogeneity across the conditions, with IV2 and IV3 often underestimating the variability in treatment effects.

Figure 1 shows a more nuanced understanding the amount of bias exhibited by the estimators as a function of the number of sites and number of individuals at a site. While the "AsTreated" and ITT method exhibit the expected behavior, all estimators display a large amount of variability in estimating the average CATE in small sample conditions, such as when there are only 20 sites and 20 participants at each site (which may resemble a study across 20 classrooms). While increasing the number of sites reduces the variance in this estimate, increasing the number of participants per site seems to better reduce the variability in the CATE estimate (e.g., comparing 20 sites with 100 individuals per site vs. 100 sites with 20 individuals per site).

Figure 2 shows the effects of compliance on CATE heterogeneity recovery for each estimator in small-sample conditions[2]. The ITT underestimates variability as the true variability increases and compliance decreases. All three IV methods are able to better recover the CATE heterogeneity as the true heterogeneity increases. These results suggest that the IV methods introduced by Raudenbush et al. (2012) and Reardon et al. (2014) can be effective in recovering the CATE, with the second IV MST method introduced by Raudenbush et al. (2012) being the most effecting in recovering CATE heterogeneity. Best recommendations do vary by simulation condition; more complete and pointed recommendations will be presented.

[2]The odd shape is due to variances only having positive values, and thus a limit on its bias.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Bettinger, E. P. and Baker, R. B. (2014). The effects of student coaching: An evaluation of a randomized experiment in student advising. *Educational Evaluation and Policy Analysis*, 36(1):3–19.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation review*, 8(2):225–246.

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches*. Russell Sage Foundation.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., and Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, 10(4):817–842.

Boatman, A. and Long, B. T. (2018). Does remediation work for all students? how the effects of postsecondary remedial and developmental courses vary by level of academic preparation. *Educational Evaluation and Policy Analysis*, 40(1):29–58.

Borman, G. D., Benson, J. G., and Overman, L. (2009). A randomized field trial of the fast forword language computer-based training program. *Educational Evaluation and Policy Analysis*, 31(1):82–106.

Clark, M. A., Gleason, P. M., Tuttle, C. C., and Silverberg, M. K. (2015). Do charter schools improve student achievement? *Educational Evaluation and Policy Analysis*, 37(4):419–436.

Gerber, A. S. and Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Kim, J. S., Capotosto, L., Hartry, A., and Fitzgerald, R. (2011). Can a mixed-method literacy intervention improve the reading achievement of low-performing elementary school students in an after-school program?: Results from a randomized controlled trial of read 180 enterprise. *Educational Evaluation and Policy Analysis*, 33(2):183–201.

Lynch, K. and Kim, J. S. (2017). Effects of a summer mathematics intervention for low-income children: A randomized experiment. *Educational Evaluation and Policy Analysis*, 39(1):31–53.

Olsen, R. B. (2017). Evaluating educational interventions when impacts may vary across sites. *Journal of Research on Educational Effectiveness*, 10(4):907–911.

Raudenbush, S. W. and Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, 36(4):475–499.

Raudenbush, S. W., Reardon, S. F., and Nomi, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of research on Educational Effectiveness*, 5(3):303–332.

Reardon, S. F. and Stuart, E. A. (2017). Editors' introduction: Theme issue on variation in treatment effects. *Journal of Research on Educational Effectiveness*, 10(4):671–674.

Reardon, S. F., Unlu, F., Zhu, P., and Bloom, H. S. (2014). Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *Journal of Educational and Behavioral Statistics*, 39(1):53–86.

Schochet, P. Z. and Chiang, H. (2009). Technical methods report: Estimation and identification of the complier average causal effect parameter in education rcts. ncee 2009-4040. *National Center for Education Evaluation and Regional Assistance*.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, and What Works Clearinghouse (2017). Procedures and standards handbook (version 4.0).

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., and Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, 10(4):843–876.

Zvoch, K. (2016). Intent-to-treat and treatment take-up effects of summer school on the literacy outcomes of struggling early readers. In *Treatment Fidelity in Studies of Educational Intervention*, pages 88–107. Routledge.

Table 1: Enumeration of Simulation Conditions

Parameter	Simulation Conditions
Number of Sites	200, 100, 50, 20
Number of Simulees Per Site	200, 100, 20
Degree of Within-Site Size Variability	10%, 20%
Compliance	1, 0.9, 0.75
Treatment Effect Size	0, .3, .7
Treatment Effect Heterogeneity	0, 0.1, 0.25
Magnitude of Selection Bias Constant	.1, .25, .5

Table 2: Relative Parameter Bias of the Treatment Effect Estimate and Treatment Effect Heterogeneity Estimate.

					Estimator									
					AsTreat		ITT		IV1		IV2		IV3	
SiteVar	Comp	Tx	TxSD	SB	PE	SD	PE	SD	PE	SD	PE	SD	PE	SD
Panel 1: Site Size Variance Varies														
0.1	0.9	0.3	0.1	0.1	1.067	0.940	1.100	0.820	1.000	1.040	1.000	0.920	1.000	0.940
0.2	0.9	0.3	0.1	0.1	1.070	0.920	0.903	0.800	1.003	1.020	1.003	0.900	1.003	0.920
Panel 2: Compliance Varies														
0.1	1	0.3	0.1	0.1	1.000	0.950	1.000	0.950	1.000	1.010	1.000	0.960	1.000	0.960
0.1	0.9	0.3	0.1	0.1	1.070	0.950	0.903	0.800	1.003	1.020	1.003	0.900	1.003	0.920
0.1	0.75	0.3	0.1	0.1	1.167	0.960	0.750	0.660	1.000	1.110	1.003	0.880	1.000	0.920
Panel 3: Treatment Effect Size Varies														
0.1	0.9	0	0.1	0.1	-	0.940	-	0.830	-	1.030	-	0.950	-	0.950
0.1	0.9	0.3	0.1	0.1	1.070	0.920	0.903	0.800	1.003	1.020	1.003	0.900	1.003	0.920
0.1	0.9	0.7	0.1	0.1	1.026	0.930	0.898	0.810	0.997	1.010	0.999	0.870	0.997	0.910
Panel 4: Treatment Effect Heterogeneity Varies														
0.1	0.9	0.3	0	0.1	1.067	-	0.897	-	0.997	-	1.000	-	0.997	-
0.1	0.9	0.3	0.1	0.1	1.070	0.920	0.903	0.800	1.003	1.020	1.003	0.900	1.003	0.920
0.1	0.9	0.3	0.25	0.1	1.063	0.992	0.897	0.884	0.997	0.984	0.997	0.980	0.997	0.984
Panel 5: Selection Bias Magnitude Varies														
0.1	0.9	0.3	0.1	0.1	1.070	0.920	0.903	0.800	1.003	1.020	1.003	0.900	1.003	0.920
0.1	0.9	0.3	0.1	0.25	1.167	0.900	0.900	0.800	1.000	1.020	1.000	0.880	1.000	0.910
0.1	0.9	0.3	0.1	0.5	1.340	0.910	0.907	0.830	1.007	1.050	1.010	0.910	1.006	0.950

Note. PE RPB = Treatment effect point estimate relative paramter bias. SD RPB = Treatment effect heterogeneity point estimate relative parameter bias. The other data-generating parameters not varied in this table are the number of sites (100), site size (100), the correlation between the control group and treatment effect across sites (0.0), allocation = 50%, no variance of allocation across sites, and no variance in compliance across sites.

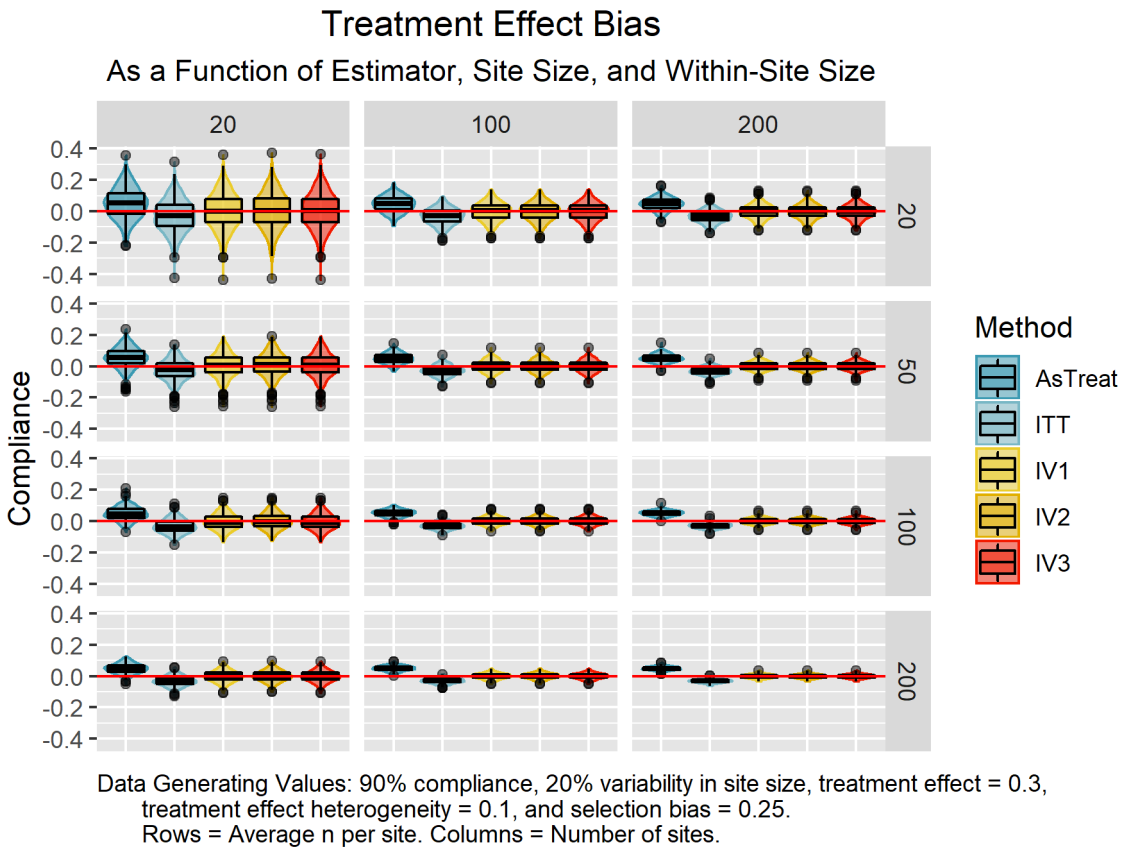

Figure 1: Treatment Effect Bias.

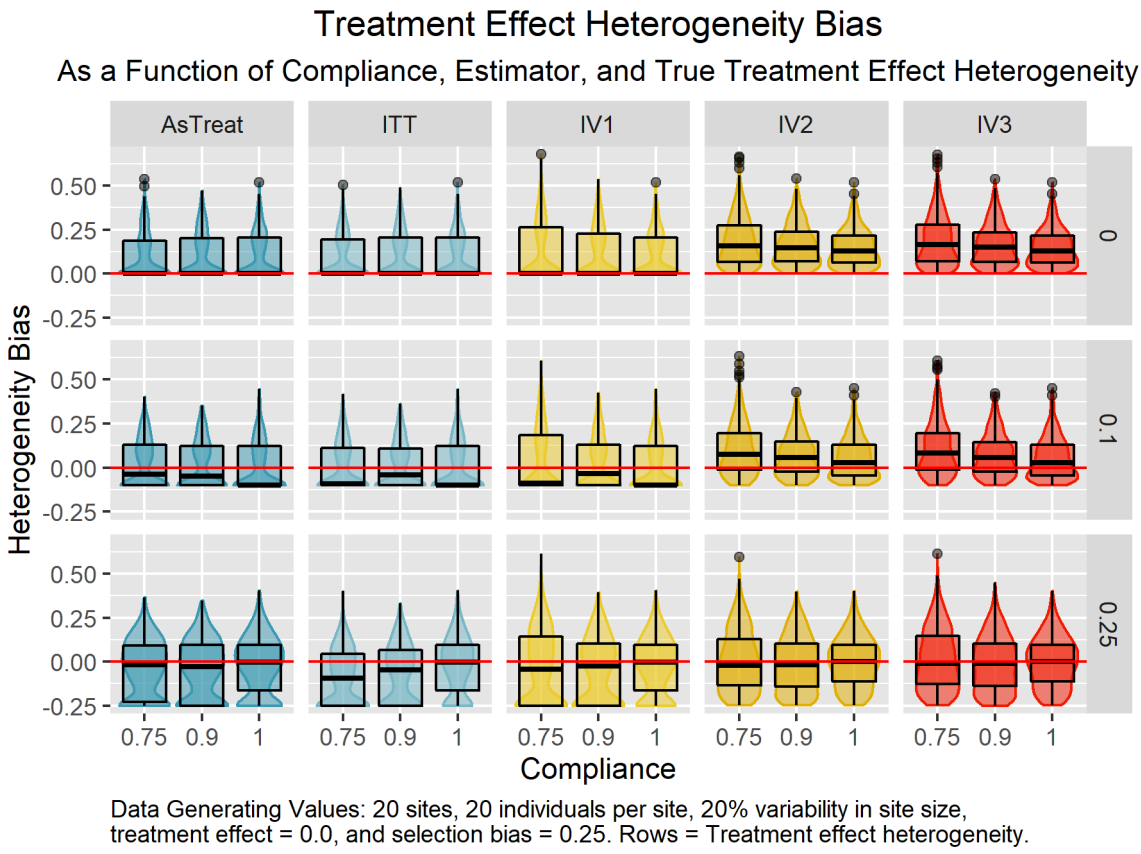

Figure 2: Treatment Effect Heterogeneity Bias.

Appendix A: Data-Generating Syntax

```
###################
## Full-Data DGM ##
###################

multisite_data_dgm <- function(nsites, npersite,
                               site_var, rho,
                               compliance, comp_var,
                               allocation_var,
                               TxEff, TxHetero, SBias,
                               condNum){

  sitevals <- generate_site_params(nsites, npersite,
                                   site_var, rho,
                                   compliance, comp_var,
                                   allocation_var,
                                   TxEff, TxHetero, SBias,
                                   condNum)

  site_list <- split(sitevals, sitevals$site)
  full_data <- map_dfr(site_list, within_site_dgm)
  return(full_data)
}

####################
## Site-Level DGM ##
####################

generate_site_params <- function(nsites, npersite,
                                 site_var, rho,
                                 compliance, comp_var,
                                 allocation_var,
                                 TxEff, TxHetero, SBias,
                                 condNum){

  # Make site-level values first.
  sitevals <- data.frame(site = as.factor(1:nsites))

  # Number of simulees at each site
  sitevals$sitesize <- case_when(site_var != 0 ~ round(runif(nsites,
                                 npersite - ((npersite*site_var)/2),
                                 npersite + ((npersite*site_var)/2))),
                                 site_var == 0 ~ npersite)
```

```
  # Compliance level at each site
  if(comp_var != 0){
    sitevals$compliance_s <- runif(nsites,
              compliance - ((compliance*comp_var)/2),
              compliance + ((compliance*comp_var)/2))
  }else{
    sitevals$compliance_s <- compliance
  }


  # Creating unbalanced treatment allocation if desired
  if(allocation_var != 0){
    sitevals$apct <- runif(nsites,
                           0.5 - ((0.5 * allocation_var)/2),
                           0.5 + ((0.5 * allocation_var)/2))
  }else{
    sitevals$apct <- 0.5
  }

  # Additional DGM parameter
  # 0, positive correlation, negative correlation?
  rho <- rho
  deltaInts <- mvtnorm::rmvnorm(nsites, mean = c(1, TxEff),
                        sigma = matrix(c(1, rho * TxHetero,
                        rho * TxHetero, TxHetero^2), 2, 2))

  sitevals$delta_s <- deltaInts[,2]
  sitevals$stage2int <- deltaInts[,1]

  # Amount to add to AT and NT within-site
  sitevals$biasdeg <- SBias
  sitevals$condNum <- condNum
  return(sitevals)
}

######################
## Within-Site DGM ##
######################

within_site_dgm <- function(data){

  wndata <- data.frame(simulee = 1:data$sitesize)
  wndata$site <- data$site
```

```r
# Setting up the treatment allocation / assignment
ntreat <- with(data, round(sitesize * apct))
wndata$Zi <- c(rep(1, ntreat), rep(0, data$sitesize - ntreat))

noncomp_num <- with(data, sitesize - round(compliance_s * sitesize))
even <- ((noncomp_num %% 2) == 0)

if(even){
  NT <- noncomp_num/2
  AT <- noncomp_num/2
}else{
  NT <- trunc(noncomp_num/2)+1
  AT <- trunc(noncomp_num/2)
}

comps <- data$sitesize - noncomp_num
strata_vec <- c(rep("NT", times = NT),
                rep("AT", times = AT),
                rep("C", times = comps))

wndata$strata <- sample(strata_vec, length(strata_vec))

wndata$Di <- with(wndata,
                  ifelse(strata == "NT", 0,
                  ifelse(strata == "AT", 1,
                  ifelse(strata == "C", Zi, NA))))

wndata$strata <- as.factor(wndata$strata)

# Outcome is a function of the intercept, treatment receipt, and error
s2e <- rnorm(data$sitesize, 0, 1)
s2int <- data$stage2int
delta <- data$delta_s

wndata$Y_is <- s2int + delta*wndata$Di + s2e

# Explicitly adding in confounding, as the additional effect is a result
# of unmodeled covariates. The effect size of the unmodeled covariate
# is biasdeg. This correlates potential outcomes with the strata.
wndata$Y_is <- with(wndata,
                    ifelse(strata == "AT", Y_is + data$biasdeg,
                    ifelse(strata == "NT", Y_is - data$biasdeg,
                                  Y_is)))
```

```r
  wndata$ZiCenter <- with(wndata, Zi - mean(Zi))
  wndata$DiCenter <- with(wndata, Di - mean(Di))
  wndata$condNum <- wndata$condNum

  return(wndata)
}

####################################
# Specifying DGM parameter values #
####################################

design_factors <- list(nsites = c(200, 100, 50, 20),
                       npersite = c(200, 100, 20),
                       site_var = c(0.10, 0.20),
                       rho = c(0, .3), # unreported in current study
                       compliance = c(1, 0.9, 0.75),
                       comp_var = c(0, 0.10), # unreported in current study
                       allocation_var = c(0, 0.10), # unreported in current study
                       TxEff = c(0, 0.3, 0.7),
                       TxHetero = c(0, 0.1, 0.25),
                       SBias = c(0.1, 0.25, 0.5))

simConds <- purrr:::cross_df(design_factors)
simConds <- subset(simConds, !((compliance == 1) & (comp_var == .10)))
simConds$condNum <- 1:nrow(simConds)
```

Appendix B: Estimator Implementation

```
library(dplyr)
library(metafor)
library(lme4)
library(tibble)

dataEval <- function(sitedata){

  ###############
  # As Treated #
  ###############
  treatSum <- lm(Y_is ~ 0 + site + site:Di, data = sitedata)
  trt_coefs <- grep(":Di", names(coef(treatSum)))
  siteTx <- coef(treatSum)[trt_coefs]
  siteV <- diag(vcov(treatSum))[trt_coefs]

  AsTreatREML <- rma.uni(yi = siteTx, vi = siteV)
  AsTreat_conv <- 1

  treat_pe <- as.numeric(AsTreatREML$beta)
  treat_sd <- sqrt(AsTreatREML$tau2)

  #######
  # ITT #
  #######
  ITT <- lmer(Y_is ~ 0 + Zi + site + (0 + Zi | site), data = sitedata)
  ITT_conv <- ifelse(is.null(ITT@optinfo$conv$lme4$code), 1, 0)
  ITT_pe <- as.numeric(fixef(ITT)['Zi'])
  ITT_sd <- unname(attributes(VarCorr(ITT)$site)$stddev['Zi'])

  ############################
  ## IV Method 1 - IVpooled ##
  ############################
  IV1_gamma <- lmer(Di ~ ZiCenter + (ZiCenter | site), data = sitedata)
  IV1_gamma_conv <- ifelse(is.null(IV1_gamma@optinfo$conv$lme4$code), 1, 0)
  gamma_fe <- as.numeric(fixef(IV1_gamma)['ZiCenter'])
  gamma_re <- unname(attributes(VarCorr(IV1_gamma)$site)$stddev['ZiCenter'])

  IV1_pe <- ITT_pe / gamma_fe

  IV1_sd <- sqrt((ITT_sd^2 - (ITT_pe^2 * gamma_re^2)) /
                   (gamma_fe^2 + gamma_re^2))
```

```
  ############################################################
  ## IV Method 2 - 2SLS with treatment * site interactions ##
  ############################################################
  IV2_s1 <- lm(Di ~ site + Zi:site, data = sitedata)
  sitedata$DiPred <- predict(IV2_s1)

  IV2_s2 <- lmer(Y_is ~ DiPred + (DiPred | site), data = sitedata)
  IV2_s2_conv <- ifelse(is.null(IV2_s2@optinfo$conv$lme4$code), 1, 0)

  IV2_pe <- as.numeric(fixef(IV2_s2)['DiPred']) # 0.231
  IV2_sd <- unname(attributes(VarCorr(IV2_s2)$site)$stddev['DiPred'])

  ######################################
  ## IV Method 3 - EBayes First Stage ##
  ######################################
  #Using equal variance for each group across all sites
  site_summary_info <-
    group_by(sitedata, site) %>% summarise(n = n())

  # Just assuming equal individual-level variances. Makes it easier.
  site_summary_info$siteInt <- fixef(IV1_gamma)[["(Intercept)"]] +
    ranef(IV1_gamma)$site[["(Intercept)"]]

  site_summary_info$EB_Beta <- fixef(IV1_gamma)[["ZiCenter"]] +
    ranef(IV1_gamma)$site[["ZiCenter"]]

  sitedata$DiEB <- site_summary_info$siteInt[sitedata$site] +
    sitedata$Zi*site_summary_info$EB_Beta[sitedata$site]

  IV3_s2 <- lmer(Y_is ~ DiEB + (DiEB | site), data = sitedata)
  IV3_s2_conv <- ifelse(is.null(IV3_s2@optinfo$conv$lme4$code), 1, 0)

  IV3_pe <- as.numeric(fixef(IV3_s2)['DiEB'])
  IV3_sd <- unname(attributes(VarCorr(IV3_s2)$site)$stddev['DiEB'])

  rdf <- tibble(
    condNum = sitedata$condNum[1],
    Method = c("AsTreat", "ITT", "IV1", "IV2", "IV3"),
    conv = c(AsTreat_conv, ITT_conv, IV1_gamma_conv, IV2_s2_conv, IV3_s2_conv),
    TxEff = c(treat_pe, ITT_pe, IV1_pe, IV2_pe, IV3_pe),
    EffSd = c(treat_sd, ITT_sd, IV1_sd, IV2_sd, IV3_sd)
  )
  return(rdf)
}
```