Randomized Control Trials and Policy Evaluation

Matteo Bobba
matteo.bobba@tse-fr.eu
Office: T.353

Toulouse School of Economics (TSE)

M2 PPD/ERNA/EEE, Winter 2025

Part 3: Design and Implementation Issues

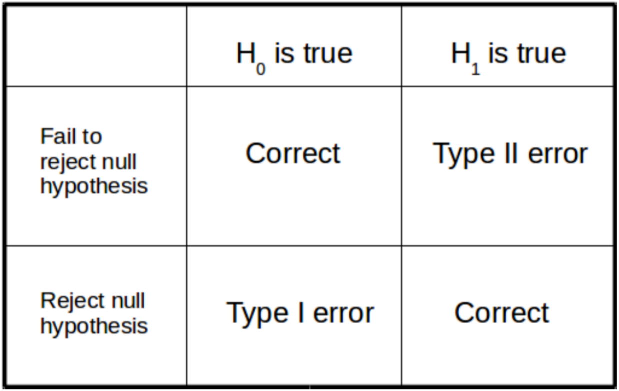
1. Sample size and the power of experiments (AI Section 7 & DGK Section 4)
2. Non-compliance (IR Ch 23,24 & DGK Section 6.2)
3. Spillovers (AI Section 11 & DGK Section 6.3)
4. Attrition and multiple outcomes (DGK Sections 6.4,7.2)

Sample Size and the Power of Experiments

Power Calculations for Randomized Experiments

- These are intended to be carried out **prior** to any experiment
- The idea is to assess whether or not the proposed experiment has a reasonable chance of finding effect sizes that one might possibly expect
- Two ways of thinking about power calculations
  - ⇒ Find sample size given pre-specified effect size
  - ⇒ Find effect size given pre-specified sample size

Type I and II Errors



	H_0 is true	H_1 is true
Fail to reject null hypothesis	Correct	Type II error
Reject null hypothesis	Type I error	Correct

Notation

- The **size** of the test is the probability of rejecting the null hypothesis when it is in fact true (false positive)

  $\Rightarrow$ $P(\text{Type I Error}) \leq \alpha = 0.05$

- The **power** of the test is the probability of rejecting the null when it is fact false (true positive)

  $\Rightarrow$ $1 - P(\text{Type II Error}) \geq \beta = 0.80$

- True average treatment effect is $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$

- Proportion of treated units: $\gamma = \sum_i W_i/N$

- Conditional variance of outcome is $\sigma_t^2 = \sigma_c^2 = \sigma^2$

Hypothesis Testing

- The parameter of interest in RCTs is the difference in means of outcomes between a **hypothetical population** that is treated and a population that is untreated

$$\begin{aligned} H_0 &: \mathbb{E}[Y_i(1) - Y_i(0)] = 0 \\ H_a &: \mathbb{E}[Y_i(1) - Y_i(0)] \neq 0 \end{aligned}$$

- We test the null hypothesis by comparing the means of a randomly chosen **sample**

$$T = \frac{\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}$$

- Given random sampling, same chances of over or under estimating the "true" (population) means

Power Calculations

- Under the alternative hypothesis we have that

$$\frac{\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} - \tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}} \approx \mathcal{N}(0,1)$$

- The implied t-statistics is approximately normal

$$T \approx \mathcal{N}\left\{\frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}, 1\right\}$$

- We reject the null hypothesis if $T > t_\alpha$

$$P\left\{\mid T \mid > \Phi^{-1}(1-\alpha/2)\right\} \approx \Phi\left\{-\Phi^{-1}(1-\alpha/2) + \frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}\right\}$$

Power Calculations

- We want the rejection probability to be at least β given that the alternative hypothesis is true, hence

$$\beta = \Phi\left\{-\Phi^{-1}(1-\alpha/2) + \frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}\right\}$$

- This implies that

$$\Phi^{-1}(\beta) = -\Phi^{-1}(1-\alpha/2) + \frac{\tau\sqrt{N}\sqrt{\gamma(1-\gamma)}}{\sigma}$$

- The required sample size for a given effect size τ is thus

$$N = \frac{\left(\Phi^{-1}(\beta) + \Phi^{-1}(1-\alpha/2)\right)^2}{\left(\tau^2/\sigma^2\right)\gamma(1-\gamma)}$$

Power Calculations: Example

- Imagine you are considering the design of an experiment assigning unemployed individuals into job training
- $\alpha = 0.05$ and $\beta = 0.8$
- SD of labor earnings is 6000 $
- $\gamma = 0.5$
- $\tau = 1/6 \times SD($earnings$)$ =1000 $
- $N = \frac{(\Phi^{-1}(0.8)+\Phi^{-1}(0.975))^2}{0.167^2 \cdot 0.5^2} = 1,302$, with 651 treated and 651 controls

⇒ The larger the MDE the smaller N (e.g. τ =2000 $ implies $N = 282$)

Power Calculations under Clustered Randomization

- Recall regression model for unit-level analysis

$$Y_i^{\text{obs}} = \alpha + \tau \bar{W}_g + \underbrace{\nu_j + \omega_i}_{\epsilon_i}$$

⇒ ν_j is common shock at cluster-level, i.i.d across clusters with variance σ_ν^2

⇒ ω_i is usual error term, i.i.d across individuals with variance σ_ω^2

- Assume G clusters of equal size $N_g = N$, $\forall g = 1, ..., G$. The variance of the OLS estimator of τ is

$$\frac{N\sigma_\nu^2 + \sigma_\omega^2}{\gamma(1-\gamma)NG}$$

- Under complete randomization the variance is

$$\frac{\sigma_\nu^2 + \sigma_\omega^2}{\gamma(1-\gamma)NG}$$

Power Calculations under Clustered Randomization

- Given sample size NG, the loss in precision due to cluster-level vs. unit-level randomization is
$$1 + (N-1)\frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\omega^2}$$
- Trade-off between number of individuals per group and number of groups which depends on the intra-class correlation $\rho = \sigma_\nu^2/(\sigma_\nu^2 + \sigma_\omega^2)$

⇒ Precision varies proportionally with number of clusters G

⇒ Nb of obs. per cluster affects precision much less, especially when ρ is large

Intra-class Correlation: Examples From Education Studies

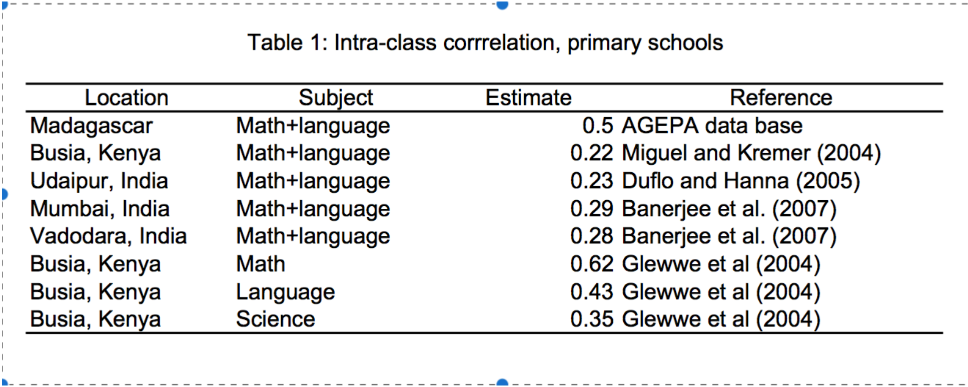
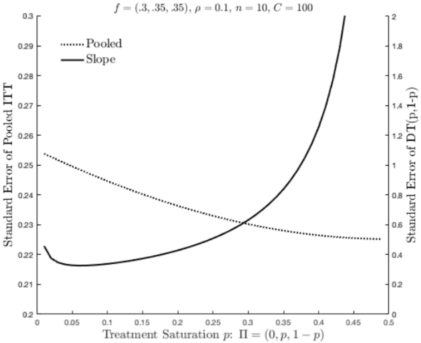


Table 1: Intra-class corrrelation, primary schools

Location	Subject	Estimate	Reference
Madagascar	Math+language	0.5	AGEPA data base
Busia, Kenya	Math+language	0.22	Miguel and Kremer (2004)
Udaipur, India	Math+language	0.23	Duflo and Hanna (2005)
Mumbai, India	Math+language	0.29	Banerjee et al. (2007)
Vadodara, India	Math+language	0.28	Banerjee et al. (2007)
Busia, Kenya	Math	0.62	Glewwe et al (2004)
Busia, Kenya	Language	0.43	Glewwe et al (2004)
Busia, Kenya	Science	0.35	Glewwe et al (2004)

Randomized Saturation Designs: Baird et al (ReStat, 2018)



⇒ Power trade-off: choosing the set of saturations and the share of clusters to assign to each saturation

Minimum Detectable Effect

1. Use standardized effect sizes

$$\frac{\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}}{\sigma}$$

2. Benchmark with other effect sizes of interventions with similar objectives
   - E.g. minimum effect size for test scores: $0.2 \cdot SD$

3. Assess what effect size would make the program cost effective
   - The "bang for the buck" (if the program were to be scaled up)
   - The experiment may be of intrinsic interest irrespective of the policy implications

Residual Variance & Intra-Class Correlation

- Data collected before the program is implemented
  - Historical data from the same or a similar population (e.g. HH survey, admin data, research papers)
  - ⇒ Data from own pilot survey or experiment (baseline survey)

- The number of repeated samples (McKenzie, 2012)
  - The more repeated obs. per individual the lower the residual variance of the outcome
  - ⇒ Depends on auto-correlation of the outcome variable

Allocation of Treatment across Units

- If no differential cost, MDE is minimized for $\gamma = 0.5$
- Otherwise, $\min MDE$ s.t. $N(1-\gamma)C_c + N\gamma C_t \leq B$, which gives

$$\frac{\gamma}{1-\gamma} = \sqrt{\frac{C_c}{C_t}}$$

- Analogously, we can derive an expression for the minimum total cost, $C^\star$, required in order to achieve a power of $1-\beta$ with a given value of MDE

⇒ With more than one treatment, you may need a larger sample size than for each treatment separately

Other Practical Considerations

- Imperfect compliance and attrition should be taken into account when determining the required sample size (see next class)
- Use covariates to increase power
  - Ex-ante: stratified randomization
  - Ex-post: add control variables
  - Choosing which variables to control must in principle be specified in advance to avoid the risk of specification searching
  - Use statistical criteria to choose covariates (e.g. machine learning tools)

Sample Code: Power Calcs

```
sampsi 0 0.1, sd(1) alpha(0.05) power(0.90) ratio(1) pre(0)

sampsi 0 0.1, sd(1) alpha(0.05) n(1000) ratio(1) pre(0)

sampsi 0 0.1, sd(1) alpha(0.05) power(0.90) ratio(1) pre(1)
r01(0.5) method(change)

sampsi 0 0.1, sd(1) alpha(0.05) power(0.90) ratio(1) pre(1)
r01(0.5) method(ancova)
```

Sample Code: Power Calcs with Clustering

```
clustersampsi, mu1(0) mu2(.1) rho(1) alpha(0.05) beta(0.8) m(1)
[replicate sampsi]

clustersampsi, mu1(0) mu2(.1) rho(0.5) alpha(0.05) beta(0.8) m(20)

clustersampsi, mu1(0) mu2(.1) rho(0.5) alpha(0.05) beta(0.8) k(60)
```

Sample Code: Power Calcs with Clustering

```
loneway y id_var
local icc = r(rho)
xtsum y, i(id_var)
local clusters =  r(n)
local clustersize = int(_N/`clusters')
clustersampsi, mu1(0) mu2(.1) rho(`icc') k(`clusters')
[too few clusters]
clustersampsi, mu1(0) mu2(.1) rho(`icc') m(`clustersize')
clustersampsi, mu1(`cmean' ) mu2(`tmean') sd1(`sd') sd2(`sd')
rho(`icc') m(`clustersize')
```

Non-Compliance

Defining (Non-)Compliance

- Some units assigned to treatment may end up not taking the treatment
  - E.g. don't enroll in job training
- Some units assigned to control may still take the treatment or another treatment similar to the one under study
  - E.g. access to other training courses
- These are **one-sided** or **two-sided** compliance issues
  - One-sided if it is only possible to drop-out of the treatment
  - Two-sided if there are both possibilities of dropping-out and getting the treatment (or a similar one) without being assigned to it

Treatment Assignment and Potential Treatment

- Let $Z_i \in \{0, 1\}$ be the randomly assigned treatment assignment
- Let $W_i(z) \in \{0, 1\}$ denote the potential treatment and $W_i^{\text{obs}} = W_i(Z_i)$ the realized value of the treatment
- Perfect compliance: $W_i(0) = 0, W_i(1) = 1$
- One-sided non-compliance: $W_i(0) = 0, W_i(1) \in \{0, 1\}$
- Two-sided non-compliance: $W_i(0) \in \{0, 1\}, W_i(1) \in \{0, 1\}$

Potential and Observed Outcomes

- Potential outcomes are defined as:

$$Y_i(z, w)$$

- Realized outcomes are, accordingly

$$Y_i^{\text{obs}} = Y_i(Z_i, W_i(Z_i)) = \begin{cases} Y_i(0,0) & \text{if } Z_i = 0, W_i(0) = 0 \\ Y_i(0,1) & \text{if } Z_i = 0, W_i(0) = 1 \\ Y_i(1,0) & \text{if } Z_i = 1, W_i(1) = 0 \\ Y_i(1,1) & \text{if } Z_i = 1, W_i(1) = 1 \end{cases}$$

Naive Estimands

1. As-treated (or blind) analysis, where units are compared by treatment received, rather than assigned:

$$\tau^{\mathsf{at}} = \frac{1}{N}\sum_{i=1}^{N}[Y_i(Z_i,1) - Y_i(Z_i,0)]$$

2. Per-protocol (or truncated) analysis, where units who do not comply with their assigned treatment are simply dropped from the analysis

$$\tau^{\mathsf{pp}} = \frac{1}{N_c}\sum_{i:W_i(0)=0,W_i(1)=1}[Y_i(1,1) - Y_i(0,0)]$$

Intention-to-treat (ITT) Analysis

- The receipt of the treatment is ignored, and outcomes are compared by the assignment to the treatment ($Z \perp\!\!\!\perp \{Y(z,w)\}_{(z,w)\in\{0,1\}^2}$)

$$\tau^{\text{itt}} = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1,W_i(1)) - Y_i(0,W_i(0))]$$

- We can estimate τ^{itt} using differences in averages of realized outcomes by treatment assignment

$$\widehat{\tau}^{\text{itt}} = \overline{Y}^{\text{obs}}_{Z_i=1} - \overline{Y}^{\text{obs}}_{Z_i=0}$$

- As usual, $\widehat{\tau}^{\text{itt}}$ can also be obtained by regressing Y_i^{obs} on Z_i and a constant term

ITT Analysis: Inference

- The sampling variance for $\widehat{\tau}^{\text{itt}}$ is

$$\widehat{\mathbb{V}}(\widehat{\tau}^{\text{itt}}) = \frac{\widehat{\sigma}_0^2}{N_0} + \frac{\widehat{\sigma}_1^2}{N_1}$$

- Where:

$$\widehat{\sigma}_0^2 = \frac{1}{N_0 - 1} \sum_{i:Z_i=0} \left(Y_i(0, W_i(0)) - \overline{Y}_0^{\text{obs}} \right)^2 = \frac{1}{N_0 - 1} \sum_{i:Z_i=0} \left(Y_i^{\text{obs}} - \overline{Y}_0^{\text{obs}} \right)^2$$

$$\widehat{\sigma}_1^2 = \frac{1}{N_1 - 1} \sum_{i:Z_i=1} \left(Y_i(1, W_i(1)) - \overline{Y}_1^{\text{obs}} \right)^2 = \frac{1}{N_1 - 1} \sum_{i:Z_i=1} \left(Y_i^{\text{obs}} - \overline{Y}_1^{\text{obs}} \right)^2$$

ITT Analysis: Example

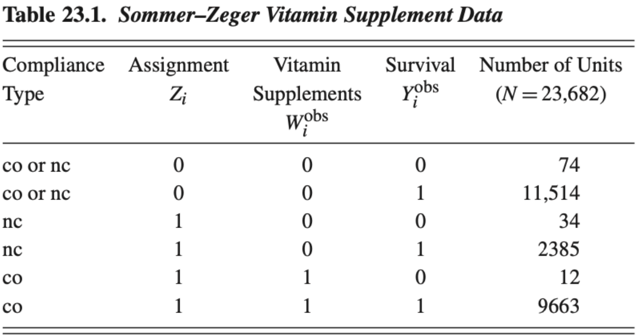


Table 23.1. ***Sommer–Zeger Vitamin Supplement Data***

Compliance Type	Assignment Z_i	Vitamin Supplements W_i^{obs}	Survival Y_i^{obs}	Number of Units ($N = 23{,}682$)
co or nc	0	0	0	74
co or nc	0	0	1	11,514
nc	1	0	0	34
nc	1	0	1	2385
co	1	1	0	12
co	1	1	1	9663

- $\overline{Y}_0^{\text{obs}} = 0.9956$, $\widehat{\sigma}_0^2 = 0.0797^2$, $\overline{Y}_1^{\text{obs}} = 0.9962$, $\widehat{\sigma}_1^2 = 0.0616^2$
- $\widehat{\tau}^{\text{itt}} = 0.0026$ and $\widehat{\mathbb{V}}(\widehat{\tau}^{\text{itt}}) = 0.0009^2$, hence $CI^{0.95}(\tau^{\text{itt}}) = (0.0008, 0.0044)$

ITT Analysis: Drawback

- The ITT effect combines partly the direct effect of taking the treatment and the indirect effect through the assignment
  - E.g. The biological effect of taking the supplements, and the psychological effect of assignment to take the supplements on actually taking them

⇒ An ITT analysis may have poor external validity since non-compliance likely depends on the context

- The causal effect of taking the treatment may be more policy-relevant than the causal effect of assigning individuals to take the treatment

Local Average Treatment Effects (LATE)

- An alternative approach is to incorporate non-compliance in the analysis
- Consider all the possible patterns of compliance behavior

$$C_i = \begin{cases} c & \text{if } W_i(0) = 0, W_i(1) = 1 \\ d & \text{if } W_i(0) = 1, W_i(1) = 0 \\ a & \text{if } W_i(0) = 1, W_i(1) = 1 \\ n & \text{if } W_i(0) = 0, W_i(1) = 0 \end{cases}$$

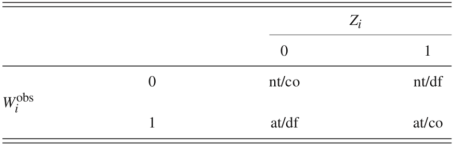
LATE: Assumptions

A1 Exclusion restriction (no direct effect of the assignment on outcomes)

$$Y_i(z,w) = Y_i(z',w) = Y_i(w), \forall z, z', w.$$

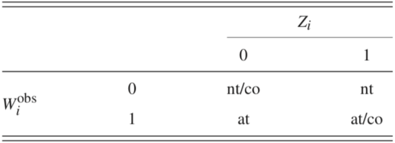
A2 Monotonicity (no defiers, only for two-sided noncompliance settings)

$$W_i(1) \geq W_i(0)$$



		Z_i	
		0	1
W_i^{obs}	0	nt/co	nt/df
	1	at/df	at/co

Compliance Status



		Z_i	
		0	1
W_i^{obs}	0	nt/co	nt
	1	at	at/co

Compliance Status with Monotonicity

LATE: Definition

- Under A1-A2 we can identify the ATE for compliers (LATE)

$$\tau^{\text{late}} = \frac{1}{N_c} \sum_{i:W_i(0)=0, W_i(1)=1} [Y_i(1) - Y_i(0)] = \frac{\frac{1}{N}\sum_{i=1}^{N} [Y_i(W_i(1)) - Y_i(W_i(0))]}{\frac{1}{N}\sum_{i=1}^{N} [W_i(1) - W_i(0)]}$$

⇒ This can be consistently estimated in an IV regression of Y_i on W_i using Z_i as the excluded instrument (Wald estimator)

LATE: Example

- The Vietnam draft lottery: random assignment by drawing the 365 days of the year in a certain order

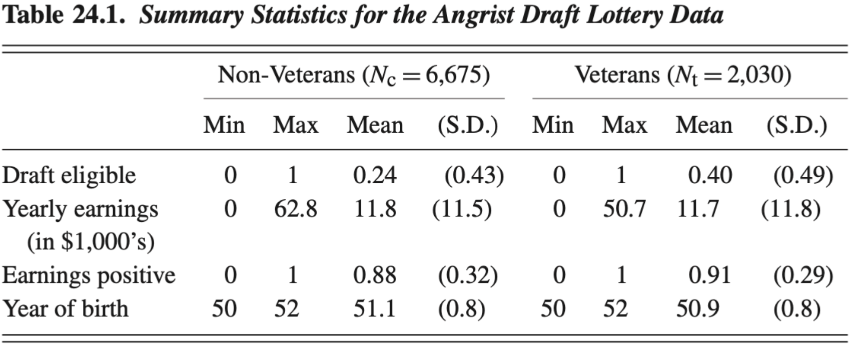


Table 24.1. ***Summary Statistics for the Angrist Draft Lottery Data***

	Non-Veterans ($N_c = 6{,}675$)				Veterans ($N_t = 2{,}030$)			
	Min	Max	Mean	(S.D.)	Min	Max	Mean	(S.D.)
Draft eligible	0	1	0.24	(0.43)	0	1	0.40	(0.49)
Yearly earnings (in $1,000's)	0	62.8	11.8	(11.5)	0	50.7	11.7	(11.8)
Earnings positive	0	1	0.88	(0.32)	0	1	0.91	(0.29)
Year of birth	50	52	51.1	(0.8)	50	52	50.9	(0.8)

LATE: Example (cont'd)

- Possible violations of the exclusion restriction
  - ⇒ **Never takers**: $Y_i(0,0) = Y_i(1,0)$. Dodging the draft if assigned (i.e. by moving to Canada) will likely involve large differences in later earnings
  - ⇒ **Always takers**: $Y_i(0,1) = Y_i(1,1)$. If being assigned and accepting means a different allocation to tasks in the military from what would have happened when applying voluntarily, this might imply differences in later earnings
  - ⇒ **Compliers and defiers**: $Y_i(0,w) = Y_i(1,w)$. The effect on earnings is attributed to serving in the military and not to the draft
- Possible violation of the monotonicity assumption
  - ⇒ Some individuals who would be willing to volunteer if they are not drafted but would resist the serve if drafted

LATE: Example (cont'd)

- $\widehat{\tau}^{\text{itt}} = -0.213 \ \ (\widehat{s.e.} = 0.20)$
- $\widehat{\tau}^{w} = 0.1460 \ \ (\widehat{s.e.} = 0.0108)$
- $\widehat{\tau}^{\text{late}} = \frac{\widehat{\tau}^{\text{itt}}}{\widehat{\tau}^{w}} = -\frac{0.21}{0.1460} = -1.46 \ \ (\widehat{s.e.} = 1.36)$

LATE: Summary

- The LATE parameter is the average treatment effect for those who have been moved from being untreated to being treated
  - E.g. those who would not have served in the military without the draft but entered because of the assignment

⇒ If exclusion restrictions do not hold, IV-Wald$\neq$LATE

⇒ If monotonicity does not hold, IV-Wald$\neq$LATE (IV-Wald measures the treatment effect for individuals who are moving in and out of the treatment without distinguishing them)

Spillovers

Taxonomy of Spillovers

1. Externalities
   - Physical: e.g. disease transmission in health applications
   - Behavioral: e.g. peer effects (learning, imitation, social norms, etc)
2. Equilibrium effects
   - Local: e.g. displacement effects in job training programs
   - Global: e.g. college tuition policies and returns to college

Spatial Spillovers in Standard RCT Designs

- Miguel and Kremer (2004) proposed a way to estimate the size and geographic scope of spillovers

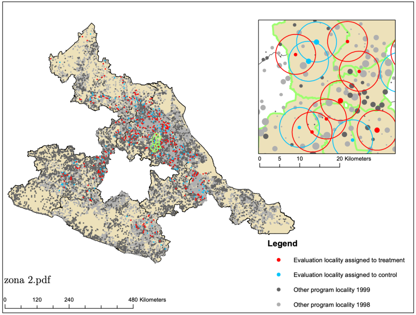
$$Y_i = \alpha + \beta_1 W_i + \beta_2 N_{d,i}^W + \beta_3 N_{d,i} + \epsilon_i$$

- $N_{d,i}^W$: number of units assigned to treatment at distance d from unit i
- $N_{d,i}$: total number of units at distance d from unit i
- β_1: ATE
- $\beta_2 \overline{N_{d,i}^W}$: average spillover effect at distance d from unit i

$\Rightarrow$ This works under specific circumstances (local spillovers and experimental sample sufficiently “dense”)

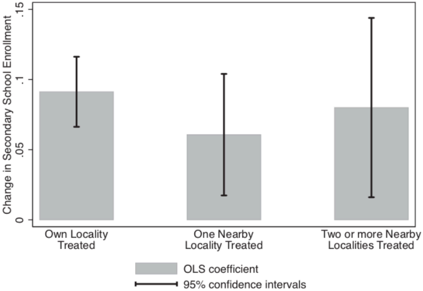
Example: Spatial Spillovers in *Progresa*

- Bobba and Gignoux (2019) finds evidence of cross-village spillovers that operate within treated villages



0 5 10 20 Kilometers
Legend
● Evaluation locality assigned to treatment
● Evaluation locality assigned to control
● Other program locality 1999
● Other program locality 1998
zona 2.pdf
0 120 240 480 Kilometers

Geographic Locations of *Progresa* Villages



Program Spillovers across Villages

Attrition

Sample Attrition

- Attrition occurs when outcomes cannot be measured for some study participants who were part of the original sample
    1. Individuals drop-out of the program and/or cannot be found (e.g. out-migration, death, etc)
    2. Participants refuse to be interviewed or refuse to answer some of the questions

⇒ Non-random attrition can undermine the comparability of the treatment and control group (selection bias)

- This may occur even when attrition rates are similar in treat and control

- Random attrition reduces sample size, reducing statistical power
    - Factor-in expected attrition rate when performing ex-ante power calculations

Attrition Ex ante

- Avoid resentments of the control group
    - Enlarge the unit of the randomization (e.g. village/municipality)
- Data collection strategies to track participants over time
    - Pilot data collection and procedures
    - If participants drop-out, go find them at home (e.g., the Balsakhi program)
    - Collect tracking info in the survey
    - Intensive follow-up for a random sub-sample of the attritors

Attrition Ex Post

- Compare attrition rates across treatment and control groups
  - Compare baseline characteristics of attritors Vs. non-attritors
- If attrition is non-random then use treatment-effect bounds
  - Lee (2009) bounds rest on random assignment of treatment and monotonicity (treatment assignment can only affect attrition in one direction)
  - Trim lower or upper tails of distribution of outcome in treatment group by the differential attrition rate (share of non-attriters is equal in both groups)
  - Calculating group differences in mean outcome yields the lower and the upper bound for the treatment effect depending on the direction of the attrition bias

Attrition Ex Post: Lee Bounds

- Share of observations with observed outcome by group

$$\begin{aligned} q_T &= \frac{\sum_i 1(W_i{=}1, S_i{=}1)}{\sum_i 1(W_i{=}1)} \\ q_C &= \frac{\sum_i 1(W_i{=}0, S_i{=}1)}{\sum_i 1(W_i{=}0)} \end{aligned}$$

- Consider the case $q_T > q_C$. Then

$$q = \frac{q_T - q_c}{q_T}$$

and $(1-q)$ determine the quantiles at which the distribution of Y in the treatment group are trimmed

Attrition Ex Post: Lee Bounds (cont'd)

- The marginal (cutoff) values of Y that enter the trimmed means are

$$\begin{aligned} y_q^T &= G^{-1}_{Y|W=1,S=1}(q) \\ y_{1-q}^T &= G^{-1}_{Y|W=1,S=1}(1-q) \end{aligned}$$

- The upper bound and the lower bound are

$$\begin{aligned} \widehat{\theta}^{\text{upper}} &= \frac{\sum_i 1(W_i=1,S_i=1,Y_i \geq y_q^T)Y_i}{\sum_i 1(W_i=1,S_i=1,Y_i \geq y_q^T)} - \frac{\sum_i 1(W_i=0,S_i=1)Y_i}{\sum_i 1(W_i=0,S_i=1)} \\ \widehat{\theta}^{\text{lower}} &= \frac{\sum_i 1(W_i=1,S_i=1,Y_i \geq y_{1-q}^T)Y_i}{\sum_i 1(W_i=1,S_i=1,Y_i \geq y_{1-q}^T)} - \frac{\sum_i 1(W_i=0,S_i=1)Y_i}{\sum_i 1(W_i=0,S_i=1)} \end{aligned}$$

Attrition Ex Post: Lee Bounds (cont'd)

- Covariates that are determined before treatment can be used to tighten treatment-effect bounds
- Covariates that have some explanatory power for attrition $S_i \in \{0, 1\}$ are used to split the sample into cells
- Bounds are separately calculated for each cell
- A weighted average of cells' bounds is computed
- Lee (2009) shows that such averaged bounds are tighter than those that do not use any covariates

Attrition: Example

- Kremer et al. (2009) study a merit-based scholarship program in Kenya

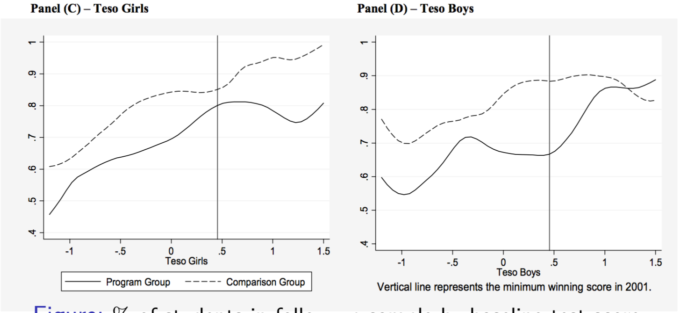


Figure: % of students in follow-up sample by baseline test score

- Lee bounds of the treatment effect in Teso district are very wide, ranging from -0.17 to 0.23

Multiple Hypothesis Testing

Beware of False Positives

- Different null hypotheses arise naturally for at least three reasons:
  1. When there are multiple outcomes of interest
  2. When the effect of a treatment may be heterogeneous across subgroups
  3. When there are multiple treatments of interest
- Standard inference considers each outcome separately

$\Rightarrow$ Multiple hypotheses lead to over-rejection of H_0 (no effect)

False Positives: Example

- Consider testing M null hypotheses simultaneously
- For each null hypothesis there is p-value $\sim U(0,1)$ when H_0 is true
- If all null hypothesis are true and that the p-values are independent, the probability of one or more false rejections is

$$P(\text{Type I Error}) = 1 - (1-\alpha)^M$$

⇒ This tends to one rapidly as M increases. E.g ($\alpha = 0.05$):

  - $P(\text{Type I} \mid M = 5) = 0.226$, $P(\text{Type I} \mid M = 10) = 0.401$, and $P(\text{Type I} \mid M = 100) = 0.994$

How can we Avoid False Positives Due to Multiple Hypothesis?

1. Select one indicator in advance to be the primary outcome (PAP)
2. Collapse many indicators using an index
3. Directly adjust p-values by the number of tests we undertake

Summary Indexes

- A summary index is a weighted mean of several standardized outcomes
- The weights are calculated to maximize the amount of information captured in the index

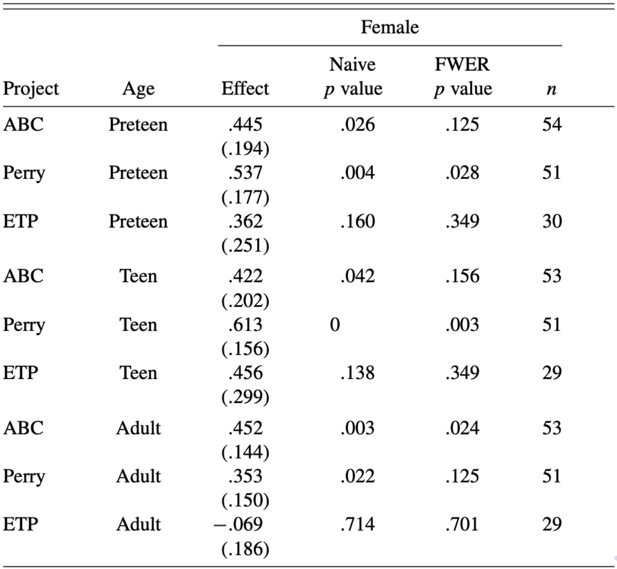
⇒ GLS-weighting procedure ensures that outcomes that are highly correlated receive less weight, while outcomes that are uncorrelated receive more weight

1. For all outcomes, switch signs where necessary so that positive direction always indicates a "better" outcome
2. Demean all outcomes and convert them to effect sizes by dividing each outcome by its control group standard deviation
3. Define J groupings of outcomes (domains). Each outcome $\tilde{y}_{jk}$ is assigned to one of these J areas (K_j outcomes in each domain j)
4. Create an index that is weighted average of $\tilde{y}_{jk}$ for individual i in area j weighted by the inverse of the covariance matrix of the transformed outcomes in area j

Adjust P-Values: Family-Wise Error Rate

- Suppose that a family of M hypotheses, $H_1, H_2, ..., H_M$, is tested, of which J are true ($J \leq M$)
- FWER is the probability that at least one of the J true hypotheses in the family is rejected
- Bonferroni correction: $p \times M$
- Westfall and Young (1993) step-down procedure:
  1. Sort outcomes $y_1, ..., y_M$ by increasing p-value
  2. Simulate the data under null hypothesis of no treatment effect
  3. Calculate $p_1^\star, ..., p_M^\star$
  4. Enforce original monotonicity: $p_r^{\star\star} = \min\{p_r^\star, p_{r+1}^\star, ..., p_M^\star\}$, where r denotes the original significance rank of the outcome
  5. Repeat (2)-(4) L times and record number of times S_r that $p_r^{\star\star} < p_r$
  6. Compute $p_r^{\text{fwer}} = S_r/L$

FWER-Adjusted P-Values: Example



		Female			
Project	Age	Effect	Naive *p* value	FWER *p* value	*n*
ABC	Preteen	.445 (.194)	.026	.125	54
Perry	Preteen	.537 (.177)	.004	.028	51
ETP	Preteen	.362 (.251)	.160	.349	30
ABC	Teen	.422 (.202)	.042	.156	53
Perry	Teen	.613 (.156)	0	.003	51
ETP	Teen	.456 (.299)	.138	.349	29
ABC	Adult	.452 (.144)	.003	.024	53
Perry	Adult	.353 (.150)	.022	.125	51
ETP	Adult	−.069 (.186)	.714	.701	29

Adjust P-Values: False Discovery rate

- FWER adjustment limits the probability of making *any* type I error
- We may be willing to tolerate some type I errors in exchange for greater power (FWER adjustments become increasingly severe as the number of tests grows)
- Alternative is to control for FDR, or the expected proportion of rejections that are type I errors
- Define V as the number of false rejections, and $t = V + U$ as the total number of rejections
- FWER is $P(V > 0)$, FDR is $E[Q = V/t]$

$\Rightarrow$ FDR requires less stringent p-value adjustments than FWER