Randomized Control Trials and Policy Evaluation

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

Toulouse School of Economics (TSE)

M2 PPD-EEP-EEE

Spring 2026

Part 3: Design and Implementation Issues

1. Sample size and statistical power (AI Section 7)
2. Non-compliance (IR Ch 23,24)
3. Spillovers (AI Section 11)
4. Sample Attrition (DGK Section 6.4)
5. Multiple outcomes (DGK Section 7.2)

Sample Size and Statistical Power

Power Calculations for Randomized Experiments

- These are intended to be carried out prior to the experiment
  - ⇒ But you can also do that ex post to bound minimum detectable effect
- Assess if proposed experiment has chances of finding reasonable effect size
- 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

- Size: probability of rejecting the null hypothesis when it is in fact true

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

- Power: probability of rejecting the null hypothesis when it is fact false

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

- 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 average null and alternative hypotheses

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

- Under the alternative hypothesis we have that

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

T-Stats and Rejection Probability

- Under the alternative hypothesis, the asymptotic t-statistics is

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

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

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

Sample Size Given Treatment Effect τ

- Rejection probability $\geq \beta$ under the alternative hypothesis, hence

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

- This implies that

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

- Required sample size for a given effect size τ is thus

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

Minimum Detectable Effect (τ)

- Use standardized effect sizes

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

- Benchmark with other effect sizes of similar interventions

  $\Rightarrow$ E.g. minimum effect size for test scores: $0.2 \cdot SD$

- Assess what effect size would make the program cost effective

  $\Rightarrow$ The "bang for the buck" (if the program were to be scaled up)

Allocation of Treatment across Units

- If no differential cost of treatment, 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

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

- This can be extended to more than one treatment

  $\Rightarrow$ You may need a larger sample size than for each treatment separately

Power Calculations: Example

- Experiment assigning unemployed individuals into job training
- $\alpha = 0.05$ and $\beta = 0.8$
- SD of labor earnings is 6000 \$
- $\gamma = 0.5$
- $\tau = 1/6 \times SD(\text{earnings}) = 1000$ \$
  - $\Rightarrow N = \frac{(\Phi^{-1}(0.8)+\Phi^{-1}(0.975))^2}{0.167^2 \cdot 0.5^2} = 1,302$, with 651 treated and 651 controls
  - $\Rightarrow$ The larger the MDE the smaller N (e.g. $\tau = 2000$ \$ implies $N = 282$)

Data Collection Strategies and Statistical Power

- Data collected before the program is implemented
  - $\Rightarrow$ Existing dataset from the same or a similar population
  - $\Rightarrow$ Data from own pilot survey or experiment (baseline survey)
- The number of repeated samples vs larger cross-section
  - $\Rightarrow$ Depends on auto-correlation of the outcome variable

Power Calculations under Clustered Randomization

- Recall regression model for unit-level analysis

$$Y_i^{\text{obs}} = \alpha + \tau \bar{W}_g + \underbrace{\nu_g + \omega_i}_{\epsilon_{ig}}$$

$\Rightarrow$ ν_g is common shock at cluster-level, i.i.d across clusters with variance σ_ν^2

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

- G clusters $N_g = N$, $\forall g = 1, ..., G$. The OLS variance of τ is

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

Power Calculations under Clustered Randomization

- Under complete randomization the variance is

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

- Loss in precision due to cluster-level vs. unit-level randomization is

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

$\Rightarrow$ Trade-off between number of individuals per group and number of groups

$\Rightarrow$ When intra-class correlation is large, N matters less than G

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)

Improving Precision Ex Post

- Use covariates to increase statistical power
  - $\Rightarrow$ OLS without or with interaction terms
  - $\Rightarrow$ Machine Learning (ML) methods such as random forests or lasso
- Linear estimators entail efficiency losses (i.e. higher variance)
  - $\Rightarrow$ By assuming linearity in $E(Y(w) \mid X)$
  - $\Rightarrow$ Variance reduction using ML tools is proportional R^2

Flexible Regression Adjustment

- Flexible (non-parametric) model for fitting conditional expectation functions
- Sample splitting (k-fold cross-fitting) to avoid over-fitting bias
  - $\Rightarrow$ For each fold $k \in 1, ..., K$ and each $W \in \{0, 1\}$, fit $\hat{m}_w^{(-k)}(X)$
  - $\Rightarrow$ For each unit i in fold k, predict $\hat{m}_{w,i} = \hat{m}_w^{(-k)}(X)$
  - $\Rightarrow$ Calculate residuals for each w, $\hat{\epsilon}_{w,i} = Y_i - \hat{m}_{w,i}$
  - $\Rightarrow$ $ATE = \frac{n_1}{n} \sum_{i:W_i=1} \hat{\epsilon}_{1,i} - \frac{n_0}{n} \sum_{i:W_i=0} \hat{\epsilon}_{0,i}$

Flexible Regression Adjustment: Example

- The impact of Medicaid on emergency room visits
- Covariates: gender, age, prior health, and education + prior ER visits

Table 7: Variance Reduction for OHIE

	SM	LRA	FRA
ER Visits	0.0132	0.0143	0.0139
	(0.0085)	(0.0079)	(0.0077)
Medicaid Take-Up	0.172	0.159	0.150
	(0.0063)	(0.0062)	(0.0062)
LATE	0.0892	0.0902	0.0870
	(0.0496)	(0.0498)	(0.0482)

$\Rightarrow$ FRA improves std.err. by about 2-3% relative to the next best alternative

$\Rightarrow$ For similar power, researchers could reduce sample sizes by about 5-6%

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 similar treatment (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 drop-out and get treatment without being assigned to it

Offered and Observed Treatment

- Let $Z_i \in \{0, 1\}$ be the randomly assigned treatment offer
- $W_i(z) \in \{0, 1\}$ potential treatment and $W_i^{\text{obs}} = W_i(Z_i)$ observed treatment
  - $\Rightarrow$ Full compliance: $W_i(0) = 0, W_i(1) = 1$
  - $\Rightarrow$ One-sided non-compliance: $W_i(0) = 0, W_i(1) \in \{0, 1\}$
  - $\Rightarrow$ Two-sided non-compliance: $W_i(0) \in \{0, 1\}, W_i(1) \in \{0, 1\}$

Potential and Observed Outcomes

- Potential outcomes under non-compliance 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

- As-treated (or blind) analysis: units are compared by accepted treatment

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

- Truncated analysis: units who do not comply with offered treat are dropped

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

Intention-to-treat (ITT) Estimand

- Outcomes are compared by the offered treatment ($Z \perp \{Y(z,w)\}$)

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

- Differences in averages of realized outcomes by treatment assignment

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

- Or regress Y_i^{obs} on Z_i and a constant term (plus eventual covariates)

ITT Analysis: Inference

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

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

$\Rightarrow$ Where:

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

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

$\Rightarrow$ Var-cov easily generalize to covariates, heterosk-robust, and cluster-robust

ITT Analysis: Example

- Effect of vitamin A supplements on infant mortality in Indonesia

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

$\Rightarrow \overline{Y}_0^{obs} = 0.9956$, $\widehat{\sigma}_0^2 = 0.0797^2$, $\overline{Y}_1^{obs} = 0.9962$, $\widehat{\sigma}_1^2 = 0.0616^2$

$\Rightarrow \widehat{\tau}^{itt} = 0.0026$ and std.err. $= 0.0009$, hence $CI^{0.95}(\tau^{itt}) = (0.0008, 0.0044)$

ITT Analysis: Drawback

- The ITT effect combines partly the direct and indirect effect of treatment
  - $\Rightarrow$ E.g. The psychological effect of assignment of the supplements
- Poor external validity since non-compliance likely depends on the context
- The causal effect of taking the treatment may be more policy-relevant

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)

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

$\Rightarrow$ Treat offer (weakly) increases incentives to take the treatment

$\Rightarrow$ Cannot be perverse: individuals would do the opposite of their assignment

Possible Compliance Status with/out Monotonicity

		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)

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

$\Rightarrow$ IV regression of Y_i on W_i using Z_i as the excluded instrument

LATE: Example

- Vietnam draft priority based on random ordering of birth dates within cohorts

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)

$\Rightarrow \widehat{\tau}^{\text{itt}} = -0.213 \ (\widehat{s.e.} = 0.20)$

$\Rightarrow \widehat{\tau}^{w} = 0.1460 \ (\widehat{s.e.} = 0.0108)$

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

LATE: Possible violations of the exclusion restriction

- Never takers: $Y_i(0,0) \neq Y_i(1,0)$

  $\Rightarrow$ Dodging the draft if assigned will likely involve differences in earnings

- Always takers: $Y_i(0,1) \neq Y_i(1,1)$

  $\Rightarrow$ If accepting means a different tasks, then differences in later earnings

- Compliers and defiers: $Y_i(0,w) \neq Y_i(1,w)$

  $\Rightarrow$ Effect on earnings is attributed to serving in the military and not to the draft

LATE: Possible violation of the monotonicity assumption

- Defiers: $W_i(1) < W_i(0)$

  $\Rightarrow$ Some would be willing to volunteer if not drafted but resist the serve if drafted

- Consider another example: the effect of fertility on labor supply

  $\Rightarrow$ IV is sex composition of a family's existing children

  $\Rightarrow$ $W_i(1) \geq W_i(0)$ requires all families to have same preference for sex mix

  $\Rightarrow$ $W_i(1) < W_i(0)$ if there are families with stopping rule for two male children

LATE: Summary

- LATE is ATE for those who moved from being untreated to being treated
- If exclusion restrictions do not hold (IV-Wald$\neq$LATE), use ITT instead
- If monotonicity does not hold (IV-Wald$\neq$LATE), use ITT instead

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

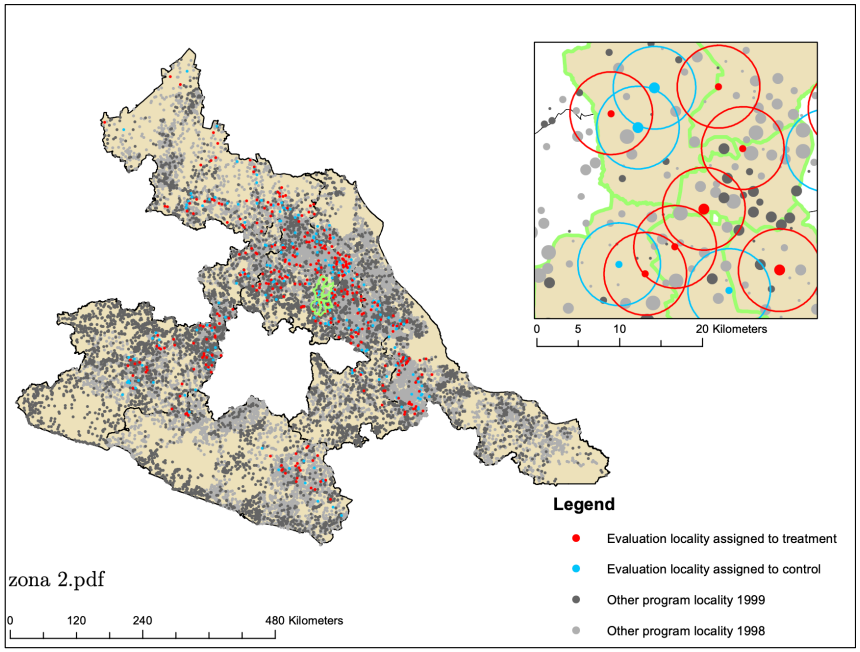
Quantifying (Local) Spillovers Ex Post

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

- $N_{d,i}^W$: 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
  - $\Rightarrow$ β_1: ATE
  - $\Rightarrow$ $\beta_2 \overline{N_{d,i}^W}$: average spillover effect at distance d from unit i
  - $\Rightarrow$ This works if spillovers are local and experimental sample sufficiently "dense"

Example: Spatial Spillovers in *Progresa*

- Cross-village spillovers that operate between treated villages



Geographic Locations of *Progresa* Villages

	(1)	(2)
Own Village Treated	0.097^{***}	0.081^{***}
	(0.014)	(0.025)
Actual Treatment Frequency		Villages
# Treated in 0–5km	0.029^{*}	-0.020
	(0.015)	(0.023)
(# Treated in 0–5 km) × Treat		0.078^{**}
		(0.033)

Program Spillovers across Villages

Quantifying Spillovers Ex Ante

- Relax SUTVA within clusters, but maintain it across clusters
  - $\Rightarrow$ ATE = direct treatment effect + within-cluster spillovers
  - $\Rightarrow$ How can we separate the two?

$\Rightarrow$ Two variants of clustered RCTs to quantify spillovers

- $\Rightarrow$ Partial population design
- $\Rightarrow$ Randomized saturation design

Partial Population Design

- Many programs have a clear target population
  - How do these programs affect untreated people nearby?
- Collect data on outcomes and covariates for those sub-populations
  - E.g. social networks in micro-credit programs
- Randomize treatment at a broader level

$\Rightarrow$ Within-cluster (non-random) program assignment: $P_{ig} \in \{0, 1\}$

$$\begin{aligned} ITE &= E(Y_i(1) - Y_i(0) \mid P_{ig} = 0) \\ &= E(Y_i \mid W_g = 1, P_{ig} = 0) - E(Y_i \mid W_g = 0, P_{ig} = 0) \end{aligned}$$

Partial Population Design: Example

- A scholarship program in Kenya
  - ⇒ Scholarship awarded to highest scoring 15% girls in six grade at district level
  - ⇒ Randomization at the school level
  - ⇒ 56% of program schools had at least one winner and 5.5 winners on average
- Program raised test scores by for girls (ATE=0.19 SD)
  - ⇒ Positive within-school spillovers on boys (ITE=0.08 SD)
  - ⇒ and for girls with low baseline test scores (ITE=0.12-0.13 SD)

Randomized Saturation Design

- A variant of clustered randomization where
  - $\Rightarrow$ Assign each cluster to a treatment saturation, $S_g = \sum_{i \in g} W_{ig} \in [0, 1)$
  - $\Rightarrow$ Assign each individual to a treatment status $W_{ig} = \{0, 1\}$ according to S_g
- Potential outcomes vary by own treatment and local saturation of treatment

$$Y_i(W_{ig}, S_g)$$

- Correlation between the treatment statuses of individuals in the same cluster
  - $\Rightarrow$ Optimal combination of clustered and stratified designs

Randomized Saturation Design: Estimands

- Direct and indirect treatment effects

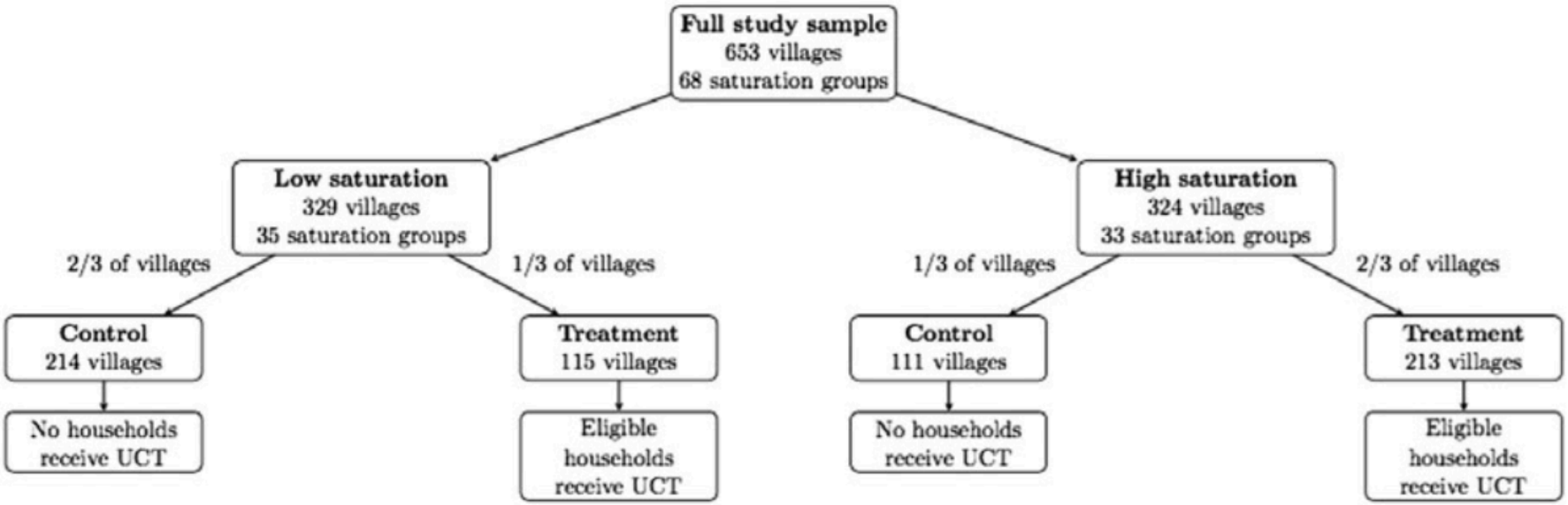
$$\begin{aligned} ATE(s) &= \mathbb{E}(Y_i(1,s) - Y_i(0,0)) = \\ &\quad + \mathbb{E}(Y_i|W_{ig}=1, S_g=s) - \mathbb{E}(Y_i|W_{ig}=0, S_g=0) \\ ITE(s) &= \mathbb{E}(Y_i(0,s) - Y_i(0,0)) = \\ &= \mathbb{E}(Y_i|W_{ig}=0, S_g=s) - \mathbb{E}(Y_i|W_{ig}=0, S_g=0) \end{aligned}$$

- Total Policy Effect

$$\begin{aligned} TPE(s) &= \mathbb{E}(Y_i|S_g=s) - \mathbb{E}(Y_i|S_g=0) = \\ &= s \times ATE(s) + (1-s) \times ITE(s) \end{aligned}$$

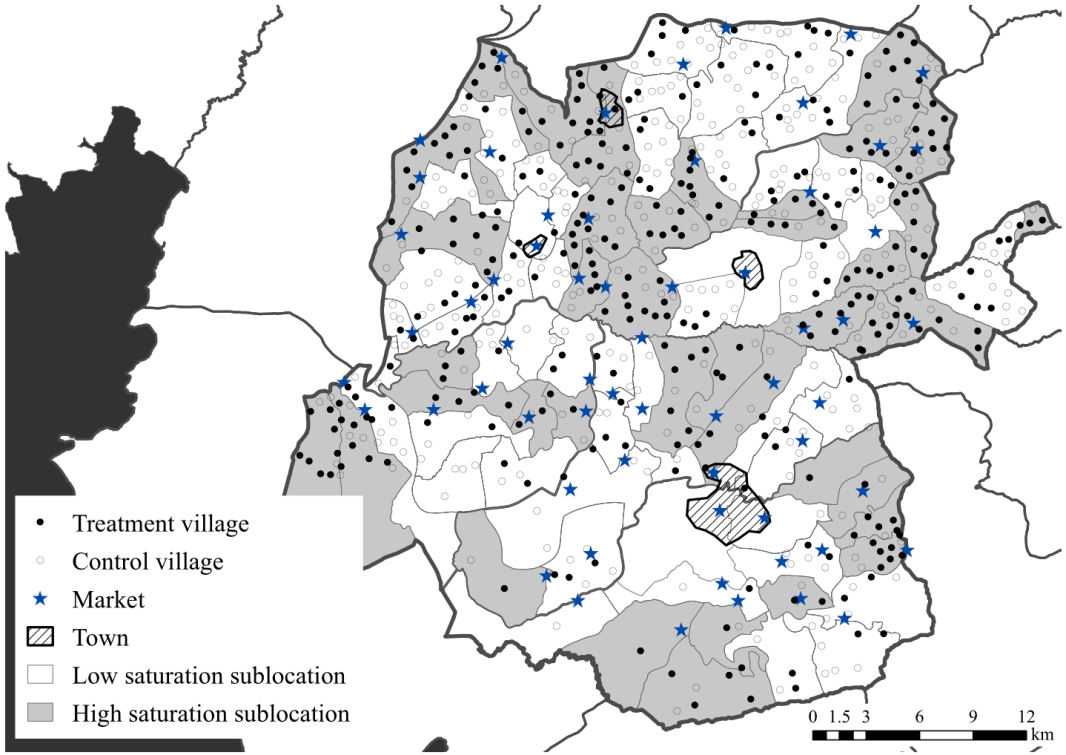
Randomized Saturation Design: Example

- One-time cash transfers of about USD 1,000 to over 10,500 poor households



Randomized Saturation Design: Example

- Variation across and within areas in the share of treated villages



Randomized Saturation Design: Example

- If spillover/equilibrium effects were localized, then

$$y_{ivs} = \alpha_1 W_v + \alpha_2 \text{HighSat}_s + \delta_1 y_{ivs,t=0} + \epsilon_{ivs}$$

- Because economic interactions do not respect geographic boundaries

$$y_{ivs} = \alpha + \beta \text{CashPC}_v + \sum_{r=2}^{R} \beta_r \text{CashPC}_{v,r} + \delta_1 y_{ivs,t=0} + \epsilon_{ivs}$$

$\Rightarrow$ $\text{CashPC}_v = f(\#\text{ Elig HHs}_v)$ so use W_v and $(\text{Share Elig HHs})^w_{v,r}$ as IV

$\Rightarrow$ $TPE = \hat{\beta}(\bar{\text{CashPC}}_v \mid i \in Elig) + \sum_{r=2}^{R} \hat{\beta}_r(\bar{\text{CashPC}}_{v,r} \mid i \in Elig)$

$\Rightarrow$ Similar specification for non-recipients HHs (without W_v)

Randomized Saturation Design: Example

	(1)	(2)	(3)	(4)
	Recipient Households		Non-Recipient Households	
	𝟙(Treat Village) Reduced Form	Total Effect IV	Total Effect IV	Control, Low-Saturation Mean (SD)
Panel A: Expenditure				
Household expenditure, annualized	293.59	338.57	334.77	2536.01
	(60.11)	(109.38)	(123.20)	(1933.51)
Non-durable expenditure, annualized	187.65	227.20	317.62	2470.69
	(58.59)	(99.63)	(119.76)	(1877.23)
Food expenditure, annualized	72.04	133.84	133.30	1578.05
	(36.96)	(63.99)	(58.56)	(1072.00)
Temptation goods expenditure, annualized	6.55	5.91	−0.68	37.07
	(5.79)	(8.82)	(6.50)	(123.54)
Durable expenditure, annualized	95.09	109.01	8.44	59.41
	(12.64)	(20.24)	(12.50)	(230.83)

Attrition

Sample Attrition

- Attrition occurs when outcomes cannot be measured for some participants
  - $\Rightarrow$ $S_i = 0 \rightarrow$ Drop-outs and/or cannot be found (e.g. out-migration, death, etc)
  - $\Rightarrow$ Participants refuse to be interviewed or refuse to answer some of the questions
- Can undermine the comparability of the treatment and control group
  - $\Rightarrow$ This may occur even when attrition rates are similar in treat and control
  - $\Rightarrow$ Even if random, attrition reduces sample size and statistical power

Attrition Ex ante

- Avoid resentments of the control group

  $\Rightarrow$ Enlarge the unit of the randomization (e.g. village/municipality)

- Data collection strategies to track participants over time

  $\Rightarrow$ Pilot data and survey protocols

  $\Rightarrow$ Collect tracking info + intensive follow-up for a random sub-sample

  $\Rightarrow$ Collect proxy information on those who have attrited

$$Y_i' = \begin{cases} Y_i & \text{if } S_i = 1 \\ Y_i^{\text{proxy}} & \text{if } S_i = 0 \end{cases}$$

Attrition Ex Post

1. Controlling for determinants of attrition
   - ⇒ Inverse probability weighting
   - ⇒ Heckman-selection model if credible "exclusion restriction"
2. Bounds on treatment effect
   - ⇒ Trim tails of distr. of treated outcomes by the differential attrition rate
   - ⇒ Difference in mean outcomes yields the lower/upper bound for ATE
   - ⇒ The other bound is estimated using the Manski-Lee approach

Inverse Probability Weighting ($Y_i \perp S_i \mid X_i$)

- Estimate the propensity score using Probit/Logit

$$\pi(X_i) = P(S_i = 1 \mid X_i) = \Phi(X_i'\beta)$$

- Define inverse probability weights as $\omega_i = \begin{cases} \frac{1}{\pi(X_i)} & \text{if } S_i = 1 \\ 0 & \text{if } S_i = 0 \end{cases}$

  $\Rightarrow$ Observed individuals who look like the missing ones receive more weight

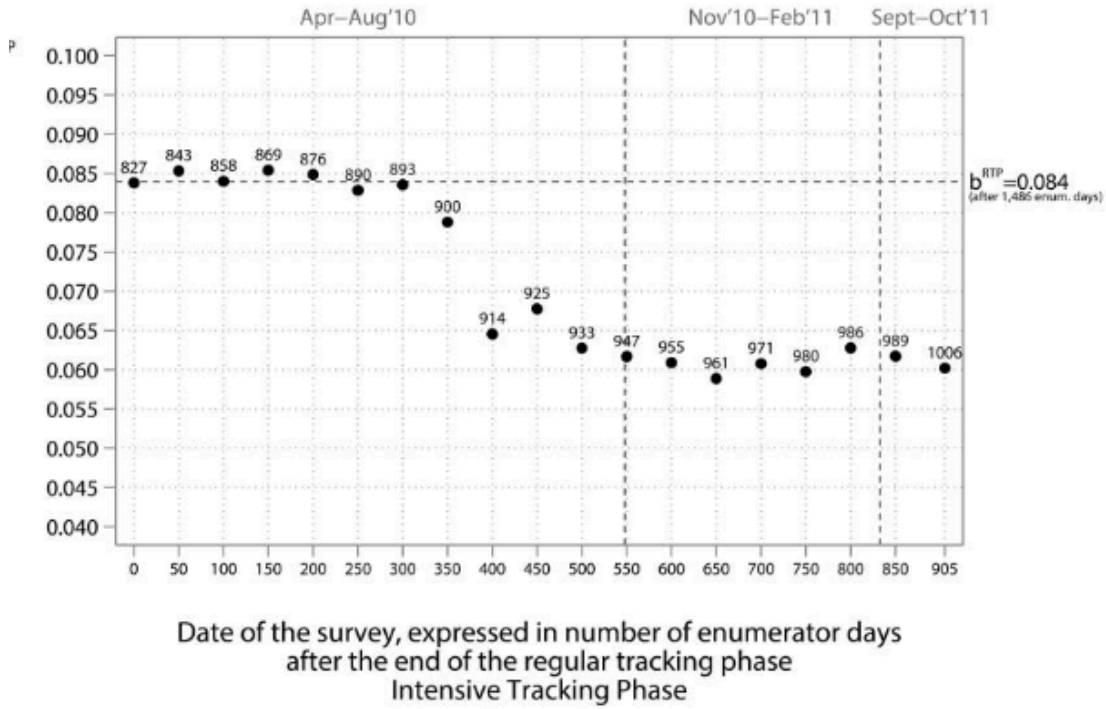
- Estimate ATE using weighted OLS

$$\min_{\alpha} \sum_{i=1} \omega_i (Y_i - \alpha W_i)^2$$

$\Rightarrow$ Stabilize weights to reduce variance: $\omega_i^{\text{stab}} = \frac{P(S_i=1)}{\pi(X_i)}$

IPW: Example

- Long term effect on employment of CCT program in Nicaragua (RPS)
- Modified version of IPW that overweights individuals that were hard to find



Lee Bounds

- Assume monotonicity: treat status affects attrition in just one direction
- 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

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

$\Rightarrow$ $q, (1-q)$: quantiles at which distr. of Y in the treatment group are trimmed

Lee Bounds

- 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 and the lower bounds are

$$\widehat{\theta}^{\text{upper}} = \frac{\sum_i 1(W_i=1,S_i=1,Y_i \geq y_q^T)Y_i}{\sum_i 1(W_i=1,S_i=1,Y_i \geq y_q^T)} - \frac{\sum_i 1(W_i=0,S_i=1)Y_i}{\sum_i 1(W_i=0,S_i=1)}$$

$$\widehat{\theta}^{\text{lower}} = \frac{\sum_i 1(W_i=1,S_i=1,Y_i \leq y_{1-q}^T)Y_i}{\sum_i 1(W_i=1,S_i=1,Y_i \leq y_{1-q}^T)} - \frac{\sum_i 1(W_i=0,S_i=1)Y_i}{\sum_i 1(W_i=0,S_i=1)}$$

Lee Bounds with Covariates

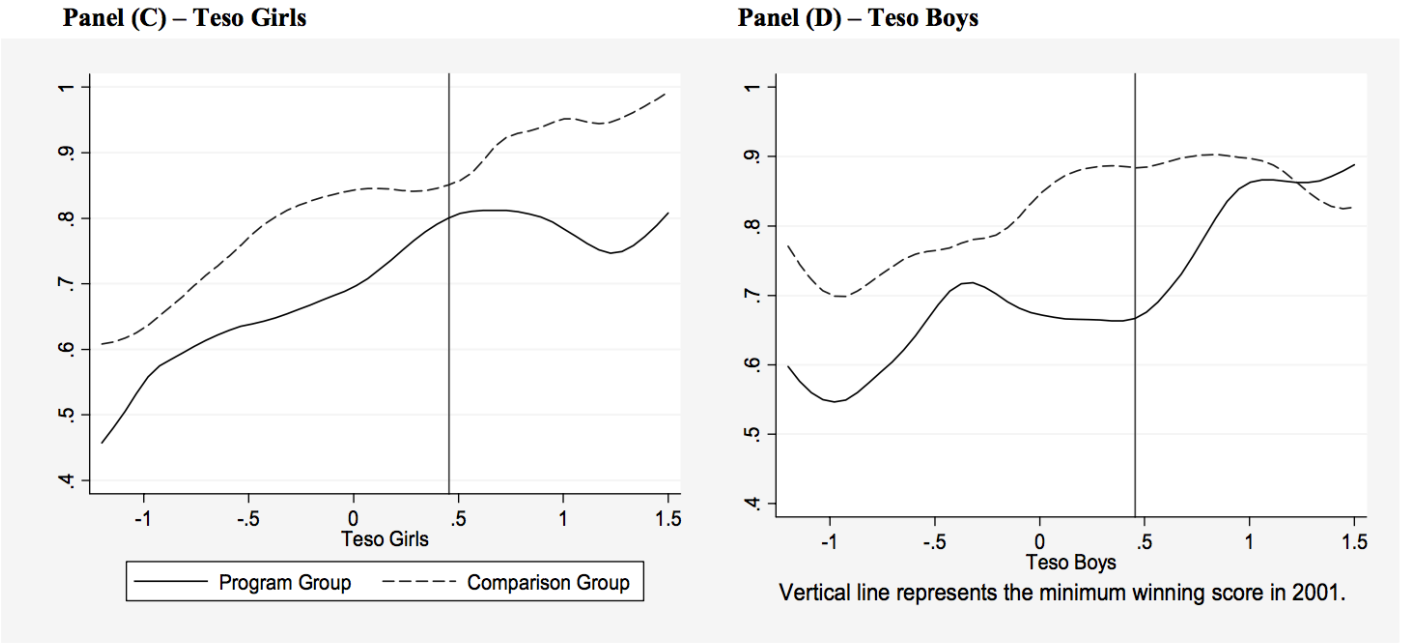
- Covariates can be used to tighten treatment-effect bounds
- Split the sample into cells J
- Bounds are separately calculated for each cell
- A weighted average of cells' bounds is computed

$$\omega_J = \frac{\sum_i 1(J_i = 1, S_i = 1, W_i = 0)}{\sum_i 1(S_i = 1, W_i = 0)} \text{ for each cell } J$$

$\Rightarrow$ Weights are prob. of cell membership for those that never suffer from attrition

Lee Bounds: Example

- Same merit-based scholarship program in Kenya that we saw earlier



⇒ Bounds 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:
  - $\Rightarrow$ When there are multiple outcomes of interest
  - $\Rightarrow$ When the effect of a treatment may be heterogeneous across subgroups
  - $\Rightarrow$ When there are multiple treatments of interest
- Standard inference considers each outcome separately
- Multiple hypotheses testing (MHT) 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 are true and p-values are independent, the prob. false rejections is

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

- This tends to one rapidly as M increases. E.g ($\alpha = 0.05$):
  - $\Rightarrow P(\text{Type I} \mid M = 5) = 0.226$
  - $\Rightarrow P(\text{Type I} \mid M = 10) = 0.401$
  - $\Rightarrow P(\text{Type I} \mid M = 100) = 0.994$

How can we Avoid False Positives Due to MHT?

1. Ex-ante: select one indicator in advance to be the primary outcome (PAP)
2. Ex-post: collapse many indicators using an index
3. Ex-post: directly adjust p-values by the number of tests we undertake

The Rationale for Pre-registration

- Commitment device for the several design and implementation choices
  - $\Rightarrow$ Which randomized design?
  - $\Rightarrow$ How many individuals, clusters and survey rounds?
  - $\Rightarrow$ How many outcomes?
  - $\Rightarrow$ How treatment impacts vary with observables?
- Strict protocols to avoid ex-post data mining and p-hacking
  - $\Rightarrow$ PAP induce transparency before analyzing the data

The Rationale for Pre-registration: Example

- Reforming institutions by enhancing coordination/participation in community
  - $\Rightarrow$ Block grants and Village Development Committees
  - $\Rightarrow$ Detailed outcomes at both HH and village level
- No evidence of program impacts as pre-specified by the PAP
  - $\Rightarrow$ Two opposite interpretations according to selective treatment effects
  - $\Rightarrow$ Illustrate the risks of discretion in ex-post data analysis

The Pros and Cons of Pre-Analysis Plans

- PAP can encourage exploratory work: surprising findings (e.g. zero effects)
  - $\Rightarrow$ Forcing the researchers to think through their hypotheses beforehand
- Many research questions do not test a single hypothesis
  - $\Rightarrow$ Hypotheses are conditional on realizations of previous hypothesis tests
  - $\Rightarrow$ Pre-specifying the chain of every realization of the data may be over-whelming

Summary Indexes

- Equally weighted standardized averaging (SA)

  $\Rightarrow \frac{1}{K}\sum_{j=1}^{K}\frac{y_j-\bar{y}_j^c}{\sigma_{y_j}^c}$

- Principal Component Analysis (PCA)

  $\Rightarrow$ Uses the principal component(s) of the correlation matrix

  $\Rightarrow$ Weights chosen to maximize variance

- Inverse Variance Matrix weighting (IVM)

  $\Rightarrow$ Same as SA but weighted by inverse of the covariance matrix

  $\Rightarrow$ If an outcome is very noisy $\rightarrow$ low weight

  $\Rightarrow$ If two outcomes are highly correlated $\rightarrow$ they don't both get full weight

Adjust p-Values: Family-Wise Error Rate

- Family of M hypotheses, $H_1, H_2, ..., H_m$, is tested, of which $J \leq m$ are true
- FWER is the probability that at least one of the J true hypotheses is rejected
  - $\Rightarrow$ Bonferroni correction: $p \times m$
- Step-down procedure:
  - $\Rightarrow$ Resamples the data under the joint null using permutations or bootstrap
  - $\Rightarrow$ For each resample, record $\max(T_1^\star, T_2^\star, ..., T_m^\star)$
  - $\Rightarrow$ Adjusts p-values accordingly: $P(\max T^\star \geq T_{(1)})$
  - $\Rightarrow$ If largest test rejected, remove it and continue sequentially
  - $\Rightarrow$ Accounts for correlation across tests

FWER-Adjusted p-Values: Example

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

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 for greater power
- Control for expected proportion of rejections that are type I errors
- V: number of false rejections, $t = V + U$: total number of rejections
  - $\Rightarrow$ FWER is $P(V > 0)$, FDR is $E[Q = V/t]$
  - $\Rightarrow$ When some false hypotheses are correctly rejected, FDR $<$ FWER
  - $\Rightarrow$ FDR requires less stringent p-value adjustments than FWER

FDR q-Values: Example

			Female				
Outcome	Age	Project	Effect	CM	Naive p value	FDR q value	n
IQ	5	ABC	4.94 (3.58)	96.76	.176	.304	48
IQ	6.5	ABC	5.13 (3.35)	92.96	.134	.271	46
IQ	12	ABC	8.35 (2.75)	87.35	.004	.048	52
IQ	5	Perry	12.67 (4.30)	81.65	.004	.048	39
IQ	6	Perry	3.75 (3.21)	87.16	.241	.318	48
IQ	10	Perry	4.96 (3.45)	81.79	.173	.304	43
IQ	5	ETP	13.55 (6.09)	87.60	.015	.077	30
IQ	7	ETP	8.61 (6.69)	89.89	.118	.271	29
IQ	10	ETP	9.79 (5.73)	81.56	.067	.216	29