Econometrics of Policy Evaluation: Randomization

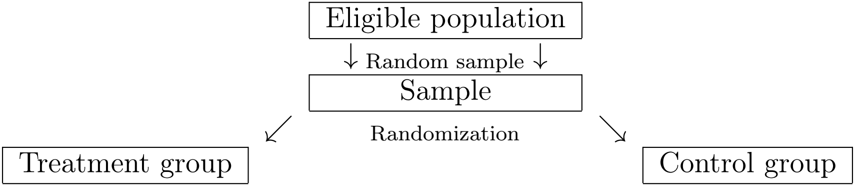
Cristian Huse

Randomized Trials

- How do researchers learn about counterfactual states of the world in practice?
- In many fields, evidence about counterfactuals is generated by randomized trials or experiments
    - e.g. medical research
- Under certain conditions, randomized trials ensure that outcomes in the comparison group really do capture the counterfactual for a treatment group
- Recall the "Fundamental Problem of Causal Inference"? Randomized trials are the most powerful tool for creating a valid counterfactual to solve this problem.

The Goal of Randomization: Internal and External Validity

- Statisticians recommend a formal two-stage randomization model:
    1. A **random sample** of units is selected from a defined population
    2. This sample of units is **randomly assigned** to treatment and comparison groups



Eligible population

↓ Random sample ↓

Sample

Randomization

Treatment group

Control group

Why Two Stages of Randomization?

- To achieve two types of validity:
  - **Internal validity** – i.e. ensure that the observed effect on the outcome is due to the treatment rather than to other confounding factors
  - **External validity** – i.e. ensure that the results in the sample will represent the results in the population within a defined level of sampling error
- **Random assignment** is the key to **internal validity**.
  - It ensures the comparison group is a valid counterfactual, solving the selection bias problem.
- **Randomly sampling** from a larger population is about **external validity**.
  - It allows us to generalize our findings from the experimental sample back to that population.

Why Two Stages of Randomization?

- **Example:**
  - A vaccine trial might randomly assign the treatment among a sample of 18-65 year olds. The results are internally valid for that group.
  - Whether they are externally valid for people over 65 is a separate question.

How Randomization Achieves the Indep. Assumption

- By assigning treatment with a coin flip, we mechanically break any link between a unit's characteristics (both observed and unobserved) and their treatment status.
- Therefore, by design, the treatment and control groups will, **on average**, look identical in every respect except for the treatment itself. This forces the selection bias term to be zero.
- Why "on average"?

Why "on average"?

- When
  1. You randomly draw a sample of units from a large population; and
  2. The number of units you draw gets large

- The average of any characteristic of your sample will tend to become closer to the expected value ("the sample mean converges to the population mean")

- Thus, in large samples, the **Law of Large Numbers** ensures that two-stage randomized trials meet the Independence Assumption and the estimator

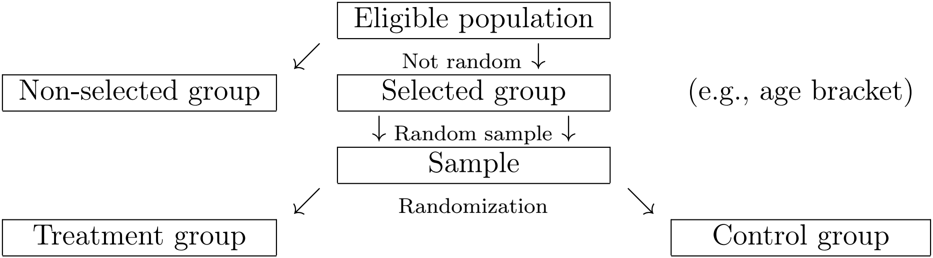
$$\delta = E[Y_i(1) - Y_i(0)] = \left[\bar{Y}(1)|D = 1\right] - \left[\bar{Y}(0)|D = 0\right]$$

consistently estimates the Average Treatment Effect (ATE)[1]

[1] We might use either P or D to denote the program of interest

Randomization vs. Sample Selection

- What if the randomization takes place on a selected group of units?



- The treatment effect is on that selected group of units!
  - **Example:** In the STAR experiment, which randomized class sizes for students, small classes had a positive effect. This result has high internal validity. However, the experiment was conducted only in public schools with a certain demographic profile. We can't be certain the same effect would hold for students in wealthy private schools or in a different country. The effect is only guaranteed for the selected group.

Implementing Randomization

- In words: *your aim is to select individuals into treatment and control groups "correctly"*
- Intuition: lottery
- Examples
  - If you want to assign 50% of the sample to treatment and comparison: flip a coin for each person
  - If you want to assign 40% of the sample to the treatment group, then roll a die for each person. A 1 or a 2 is treatment; a 3, 4, 5 or a 6 is comparison
- In practice:
  - This is done with statistical software. For example, you would generate a random number for each participant and assign everyone with a number below 0.4 to the treatment group.

Caveats

(This connects with “Empirical Challenge” slides in Causality)

- Even if you randomize correctly, things can go South...
- **Non-Compliance & Attrition (Violates Randomization):**
  - If people don’t stick to their assigned group (non-compliance) or if they drop out of the study at different rates (differential attrition), the groups are no longer perfectly random, and selection bias can creep back in.
- **Hawthorne & John Henry Effects (Violates SUTVA):**
  - These are behavioral effects where the control group is contaminated by the experiment itself.
  - This violates the SUTVA assumption (discussed in the last lecture) because the outcome of the control group is affected by the existence of the treatment group.

Level of Randomization

- In practice, how should you run your "lottery"?
- District, village, family, child?
- Level of randomization determines the power of the trial:
  - More units in the treatment and comparison groups ⇒ Estimate of difference between T and C becomes more precise
- The lower the level, the more potential for contamination, and the harder to administer the program
  - **Example:** Imagine a deworming program in schools. If we randomize at the student level, treated students may reduce the overall parasite load in the classroom, partially treating the control students ("spillover" or "contamination"). To avoid this, we can randomize at the school level. This solves the spillover but leaves us with fewer units (schools vs. students), which reduces our statistical power.
- **In General:** Choose the lowest level that is still administratively feasible

Randomized vs. Non-Randomized Trials

- **Randomized experiments:**
  - By breaking the link between characteristics and treatment, randomization allows us to estimate causal effects with minimal assumptions.
- **Non-randomized methods:**
  - The rest of this course is about what to do when randomization isn't possible.
  - These "quasi-experimental" methods rely on clever research designs and strong, testable assumptions to try and replicate what an experiment does by design.

Regression Equation

- In the potential outcome framework discussed above, the observed outcome for a unit is either one of two potential outcomes
- The observed outcome for unit i can be written as the following average of the two potential outcomes:

$$Y_i = p_i Y_i(1) + (1 - p_i) Y_i(0) = Y_i(0) + (Y_i(1) - Y_i(0)) p_i$$

- This can be rewritten as the following regression

$$Y_i = \alpha + \delta P_i + \epsilon_i$$

where

  - $\alpha (= Y_i(0))$ is the average outcome for the control group, i.e. when $P_i = 0$
  - $\delta (= Y_i(1) - Y_i(0))$ is the difference in average outcomes between the treated and non-treated subjects
  - the error term ϵ_i captures any other individual factors that may affect the relationship between the program and the outcome of interest

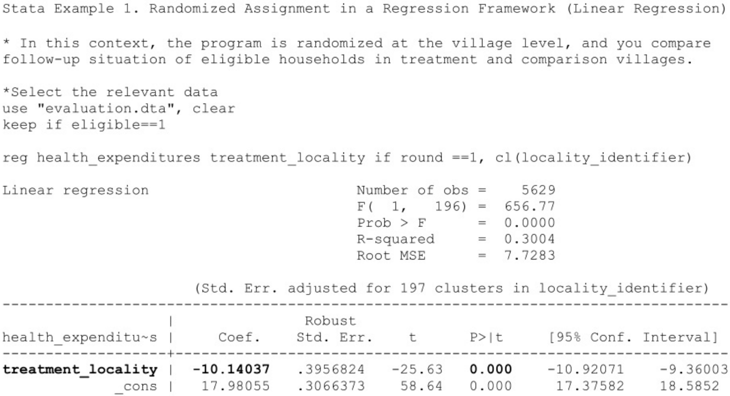
Equivalence Result

- The above regression is equivalent to testing the difference in average outcomes between treated and control groups
  $E\left(Y_i|P_i=1\right)=\alpha+\delta+E(\epsilon_i|P=1)$
  $E\left(Y_i|P_i=0\right)=\alpha+E(\epsilon_i|P=0)$
- Taking the difference gives

$$E\left(Y_i|P_i=1\right)-E\left(Y_i|P_i=0\right)=\delta+E(\epsilon_i|P=1)-E(\epsilon_i|P=0)$$

- With randomization, $E(\epsilon_i|P=1)=E(\epsilon_i|P=0)$ because, on average, all other factors (ϵ_i) are balanced across the two groups. This is why the selection bias term disappears and δ gives us the causal effect.
- In practice: regress outcome of interest on a binary treatment variable which is equal to 1 if unit receives treatment (additional covariates are often included)

Randomized Assignment: Linear Regression



```
Stata Example 1. Randomized Assignment in a Regression Framework (Linear Regression)

* In this context, the program is randomized at the village level, and you compare
follow-up situation of eligible households in treatment and comparison villages.

*Select the relevant data
use "evaluation.dta", clear
keep if eligible==1

reg health_expenditures treatment_locality if round ==1, cl(locality_identifier)

Linear regression                                    Number of obs =    5629
                                                     F(  1,    196) =  656.77
                                                     Prob > F      =  0.0000
                                                     R-squared     =  0.3004
                                                     Root MSE      =  7.7283

                      (Std. Err. adjusted for 197 clusters in locality_identifier)
------------------------------------------------------------------------------------
                   |               Robust
health_expenditu~s |      Coef.   Std. Err.      t     P>|t     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
treatment_locality |  -10.14037   .3956824    -25.63   0.000    -10.92071    -9.36003
             _cons |   17.98055   .3066373     58.64   0.000     17.37582    18.5852
```

Testing Baseline Balance in a Regression Framework

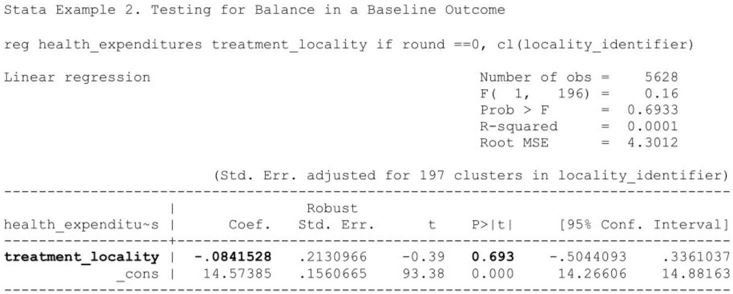
- Independence of potential outcomes is one of the most crucial assumptions to ensure that the difference in average outcomes between program beneficiaries and non-beneficiaries provides a consistent estimate of the average treatment effect

- While this assumption cannot generally be verified, some falsification tests can be implemented to identify cases when it does not hold

- Falsification 1: assume pre-program data is available. If so, program participation should have no effect on pre-program outcomes, i.e. shouldn't reject $\delta = 0$ in the following regression

$$Y_i^{pre} = \alpha + \delta P_i + \epsilon_i$$

Testing for Balance in a baseline Outcome



```
Stata Example 2. Testing for Balance in a Baseline Outcome

reg health_expenditures treatment_locality if round ==0, cl(locality_identifier)

Linear regression                                      Number of obs =     5628
                                                       F(  1,   196) =     0.16
                                                       Prob > F      =   0.6933
                                                       R-squared     =   0.0001
                                                       Root MSE      =   4.3012

                        (Std. Err. adjusted for 197 clusters in locality_identifier)
------------------------------------------------------------------------------------
                   |               Robust
health_expenditu~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
treatment_locality |  -.0841528    .2130966    -0.39   0.693    -.5044093    .3361037
             _cons |   14.57385    .1560665    93.38   0.000     14.26606    14.88163
------------------------------------------------------------------------------------
```

- The coefficient on treatment_locality is not statistically insignificant (p=0.693), meaning there was no pre-existing difference in health expenditures. The randomization was successful.

Testing Baseline Balance in a Regression Framework

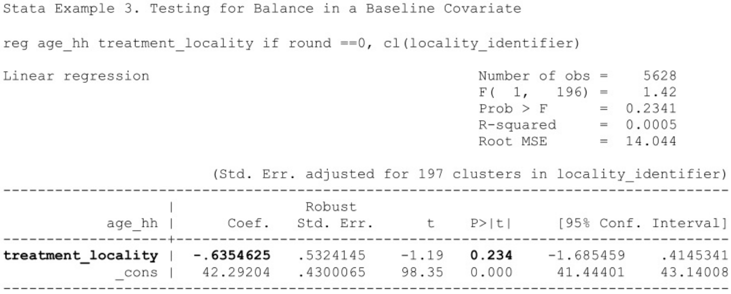
- Falsification 2: to check that characteristic X_i is similar across treated and non-treated, **pre-program**, can estimate

  $$X_i^{pre} = \alpha + \delta P_i + \epsilon_i$$

  i.e. replace Y_i^{pre} with X_i^{pre}. Not rejecting the null $\delta = 0$ suggests that characteristic X_i is balanced across treatment and control groups

  - Note: This can be done for potentially all characteristics

Example: Testing for Balance in a Baseline Covariate



```
Stata Example 3. Testing for Balance in a Baseline Covariate

reg age_hh treatment_locality if round ==0, cl(locality_identifier)

Linear regression                                      Number of obs =    5628
                                                       F(  1,    196) =    1.42
                                                       Prob > F      =  0.2341
                                                       R-squared     =  0.0005
                                                       Root MSE      =  14.044

                    (Std. Err. adjusted for 197 clusters in locality_identifier)
------------------------------------------------------------------------------
             |               Robust
      age_hh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
treatment_locality |  -.6354625   .5324145    -1.19   0.234    -1.685459    .4145341
       _cons |   42.29204   .4300065    98.35   0.000     41.44401    43.14008
------------------------------------------------------------------------------
```

- The coefficient on treatment_locality is not statistically insignificant (p=0.234), meaning there was no pre-existing difference in age.

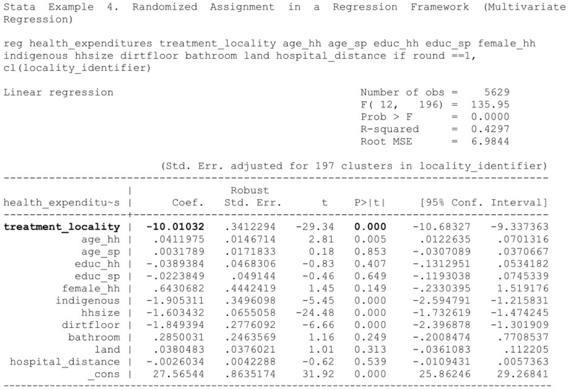
Multivariate Regression

- Characteristics of units can be controlled for in the above regression

$$Y_i = \alpha + \delta P_i + \gamma X_i + \epsilon_i$$

- Why?
  1. **Adjusting for "Bad Luck" in Randomization:** While randomization ensures balance on average, in any single experiment, we might get unlucky and have a key variable (like age) be slightly different between groups. Including this variable as a control adjusts for this chance imbalance.
  2. **Increasing Statistical Precision:** This is the most common reason. Control variables that are strong predictors of the outcome (e.g., baseline test scores predicting follow-up scores) explain a large portion of the outcome's variance. This reduces the "noise" (ϵ_i) in the regression, which shrinks the standard error of our treatment effect and makes our estimate more precise.

Randomized Assignment: Multivariate Regression



```
Stata  Example  4.  Randomized  Assignment  in  a  Regression  Framework  (Multivariate
Regression)

reg health_expenditures treatment_locality age_hh age_sp educ_hh educ_sp female_hh
indigenous hhsize dirtfloor bathroom land hospital_distance if round ==1,
cl(locality_identifier)

Linear regression                                      Number of obs =    5629
                                                       F( 12,    196) =  135.95
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4297
                                                       Root MSE      =  6.9844

                        (Std. Err. adjusted for 197 clusters in locality_identifier)
------------------------------------------------------------------------------------
             |               Robust
health_expenditu~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
treatment_locality |  -10.01032   .3412294   -29.34   0.000    -10.68327   -9.337363
            age_hh |   .0411975   .0146714     2.81   0.005     .0122635    .0701316
            age_sp |   .0031789   .0171833     0.18   0.853    -.0307089    .0370667
           educ_hh |  -.0389384   .0468306    -0.83   0.407    -.1312951    .0534182
           educ_sp |  -.0223849    .049144    -0.46   0.649    -.1193038    .0745339
         female_hh |   .6430682   .4442419     1.45   0.149    -.2330395    1.519176
        indigenous |  -1.905311   .3496098    -5.45   0.000    -2.594791   -1.215831
            hhsize |  -1.603432   .0655058   -24.48   0.000    -1.732619   -1.474245
         dirtfloor |  -1.849394   .2776092    -6.66   0.000    -2.396878   -1.301909
          bathroom |   .2850031   .2463569     1.16   0.249    -.2008474    .7708537
              land |   .0380483   .0376021     1.01   0.313    -.0361083     .112205
 hospital_distance |  -.0026034   .0042288    -0.62   0.539    -.0109431    .0057363
             _cons |   27.56544   .8635174    31.92   0.000     25.86246    29.26841
------------------------------------------------------------------------------------
```

- Note that the coefficient on treatment_locality (-10.01) is very similar to the bivariate estimate (-10.14), confirming that randomization worked. However, the standard error is now smaller (0.34 vs 0.40), and the R-squared is higher (0.43 vs 0.30). This shows the increase in precision from adding controls.

Overview

- Barrera-Osorio and Linden (2009) evaluate an RCT (Randomized Control Trial) in Colombia
- Program activities:
  - Re-furbish computers donated by private firms and install them in public schools
  - Train teachers in the pedagogic uses of computers with the help of a local university
- 2006: 97 schools were subject to a randomization
  - 48 of them received computers
  - 49 did not receive computers
- 2008: Follow-up survey
- Note the two-part treatment: it raises the question of what delivers the effect, if it exists.

Evaluate Question

- Step 1 – Evaluation question:
  - What is the impact of intervention P on variable of interest Y?
- Program
  - $P = 1$ if school selected for installation of computers + teacher training
- Variables of interest
  - Student learning (**final outcome**)
  - Classroom practice (intermediate outcome)
  - Number of teachers trained (**output**)
  - Number of working computers in schools (**output**)

Verify Balance

- Step 2 – Verify balance:
  - Goal: Check that characteristics are balanced between treatment and control groups (std. errors in parentheses)
  - Verify balance of test scores and demographic variables

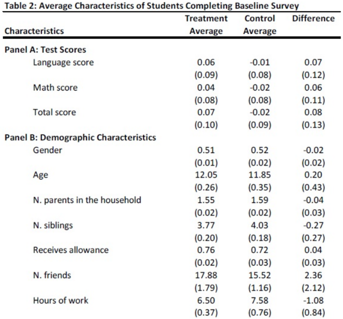


Table 2: Average Characteristics of Students Completing Baseline Survey

Characteristics	Treatment Average	Control Average	Difference
Panel A: Test Scores			
Language score	0.06	-0.01	0.07
	(0.09)	(0.08)	(0.12)
Math score	0.04	-0.02	0.06
	(0.08)	(0.08)	(0.11)
Total score	0.07	-0.02	0.08
	(0.10)	(0.09)	(0.13)
Panel B: Demographic Characteristics			
Gender	0.51	0.52	-0.02
	(0.01)	(0.02)	(0.02)
Age	12.05	11.85	0.20
	(0.26)	(0.35)	(0.43)
N. parents in the household	1.55	1.59	-0.04
	(0.02)	(0.02)	(0.03)
N. siblings	3.77	4.03	-0.27
	(0.20)	(0.18)	(0.27)
Receives allowance	0.76	0.72	0.04
	(0.02)	(0.03)	(0.03)
N. friends	17.88	15.52	2.36
	(1.79)	(1.16)	(2.12)
Hours of work	6.50	7.58	-1.08
	(0.37)	(0.76)	(0.84)

- Nothing significant here, which is exactly what we want to see. This gives us confidence that the treatment and control groups were comparable before the program began.

Estimate Impact

- Step 3 – Estimate impact:
  - In words: *Compare the average Y for the treatment group with the average Y for the comparison group*
  - In an OLS regression (to get standard errors)

    $$Y_{ij} = \beta_0 + \beta_1 T_j + \epsilon_{ij}$$

    cluster s.e.'s at the school level – why?
  - Because randomization was at the school level.
  - Students within the same school are not independent observations. Clustering standard errors adjusts for this.

Findings

- Step 3 – Findings:

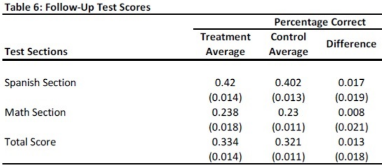


Table 6: Follow-Up Test Scores

	Percentage Correct		
Test Sections	Treatment Average	Control Average	Difference
Spanish Section	0.42	0.402	0.017
	(0.014)	(0.013)	(0.019)
Math Section	0.238	0.23	0.008
	(0.018)	(0.011)	(0.021)
Total Score	0.334	0.321	0.013
	(0.014)	(0.011)	(0.018)

Estimate Impact

- Step 4 – Estimate impact, multivariate:
  - In words: *Compare the average Y for the treatment and comparison groups, and add controls for baseline characteristics*
  - In a multivariate regression:

$$Y_{ij} = \beta_0 + \beta_1 T_j + \beta_2 X_{ij} + \epsilon_{ij}$$

    where X_{ij} is a vector of baseline characteristics for student i in school j

Findings

- Step 5 – Findings:

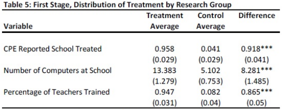


Table 5: First Stage, Distribution of Treatment by Research Group

Variable	Treatment Average	Control Average	Difference
CPE Reported School Treated	0.958	0.041	0.918***
	(0.029)	(0.029)	(0.041)
Number of Computers at School	13.383	5.102	8.281***
	(1.279)	(0.753)	(1.485)
Percentage of Teachers Trained	0.947	0.082	0.865***
	(0.031)	(0.04)	(0.05)

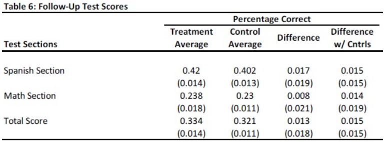


Table 6: Follow-Up Test Scores

	Percentage Correct			
Test Sections	Treatment Average	Control Average	Difference	Difference w/ Cntrls
Spanish Section	0.42	0.402	0.017	0.015
	(0.014)	(0.013)	(0.019)	(0.015)
Math Section	0.238	0.23	0.008	0.014
	(0.018)	(0.011)	(0.021)	(0.019)
Total Score	0.334	0.321	0.013	0.015
	(0.014)	(0.011)	(0.018)	(0.015)

- **Main result:**
  - Little effect on students' test scores, robust across grade levels, subjects, and gender
- **Why? (What is the mechanism?)**
  - Computers not incorporated into educational process
  - This is outside the model and requires institutional knowledge/further investigation after getting the results – but ideally should be understood early!

- Randomized assignment is the most robust method for estimating counterfactuals; it is considered the gold standard of impact evaluation. Some basic tests should still be considered to assess the validity of this evaluation strategy in a given context
    - **Are the baseline characteristics balanced?** Compare the baseline characteristics of the treatment group and the comparison group
    - **Has any non-compliance with the assignment occurred?** Check whether all eligible units have received the treatment and that no ineligible units have received the treatment. If noncompliance has occurred, you will need to use the instrumental variable method (see [G], chapter 5)
    - **Are the numbers of units in the treatment and comparison groups sufficiently large?** If not, you may want to combine randomized assignment with difference-in-differences (see [G], chapter 7)
    - **Is there any reason to believe that outcomes for some units may somehow depend on the assignment of other units?** Could there be an impact of the treatment on units in the comparison group (i.e., a violation of the SUTVA assumption) (see [G], chapter 9)?

Take-aways

- In randomized assignment, each eligible unit has the same probability of being selected for treatment
    - This ensures equivalence between the treatment and comparison groups in both observed and unobserved characteristics
- An evaluation is internally valid if it provides an accurate estimate of the counterfactual through a valid comparison group
- An evaluation is externally valid if the evaluation sample accurately represents the population of eligible units
    - The results of the evaluation can then be generalized to the population of eligible units

References

- Gertler et al (2016). Impact Evaluation in Practice, 2nd. Edition. Washington, DC: Inter-American Development Bank and World Bank
  - Chapter 4
- Gertler et al (2016). Impact Evaluation in Practice, 2nd. Edition, Technical Companion (Version 1.0). Washington, DC: Inter-American Development Bank and World Bank.
  - p. 5-10
- Barrera-Osorio, F. and L. Linden (2009): “The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia”, World Bank Policy Research Working Paper WPS4836.