VARIETIES OF THE COUNTERFACTUAL

Jesse Lecy

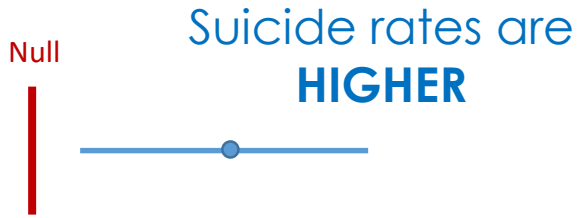
THE THREE TYPES OF COUNTERFACTUALS

Pre-Post with comparison group

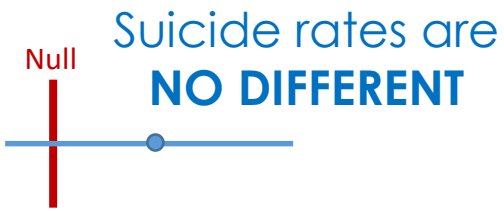
Post-test only design

Reflexive design (pre-post only)

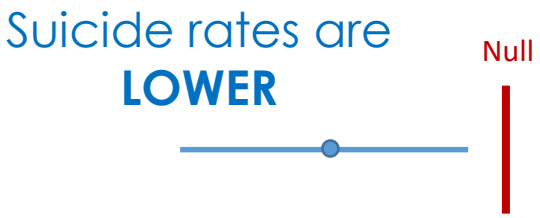
Assume you have selected a reasonable comparison.



Null Hypothesis:
Population Average



Null Hypothesis:
All HS Students
OR All Californians



Null Hypothesis:
All Suburban HS Students

You have created your groups, tested for equivalence and attrition bias, decided on the type of measure you will use in the study (ITT or TOT).

Figure 4.3 Steps in Randomized Assignment to Treatment

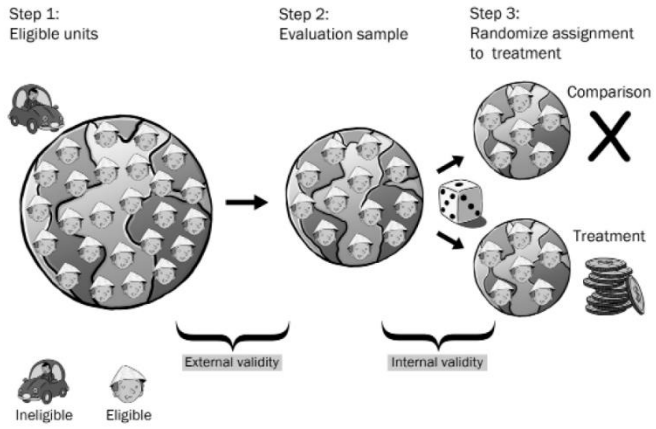


TABLE 2
Background Characteristics of Students in Treatment and Control Groups
(Total numbers of cases in parentheses)

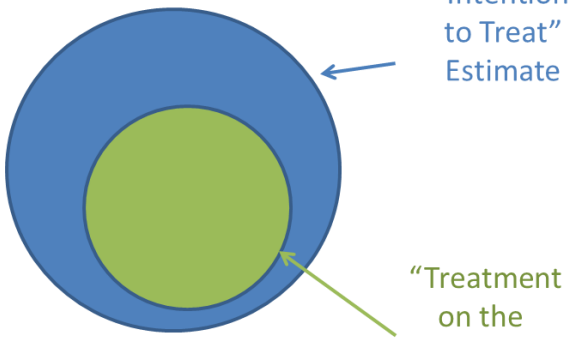
Characteristic	All students in the study: Choice students	All students in the study: Control students	All students in the study: p value[a]	All students with scores three or four years after application: Choice students	All students with scores three or four years after application: Control students	All students with scores three or four years after application: p value[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

TABLE 2
Background Characteristics of Students in Treatment and Control Groups
(Total numbers of cases in parentheses)

Characteristic	All students in the study: Choice students	All students in the study: Control students	All students in the study: p value[a]	All students with scores three or four years after application: Choice students	All students with scores three or four years after application: Control students	All students with scores three or four years after application: p value[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.



Now how do we actually calculate program impact?

$$Y = b_0 control + b_1 treat + b_2 time + b_3 controls + e$$

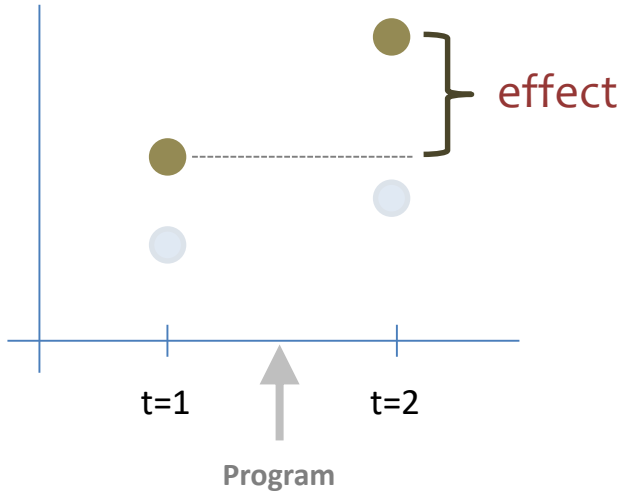
$b_?$ = program impact?

THE THREE TYPES OF COUNTERFACTUAL ESTIMATES

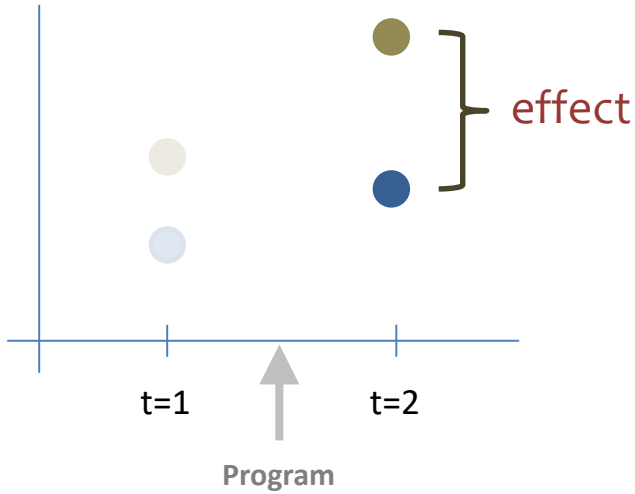
3 VALID COUNTERFACTUALS:

Each variety of counterfactual has a different formula for the program effect estimate.

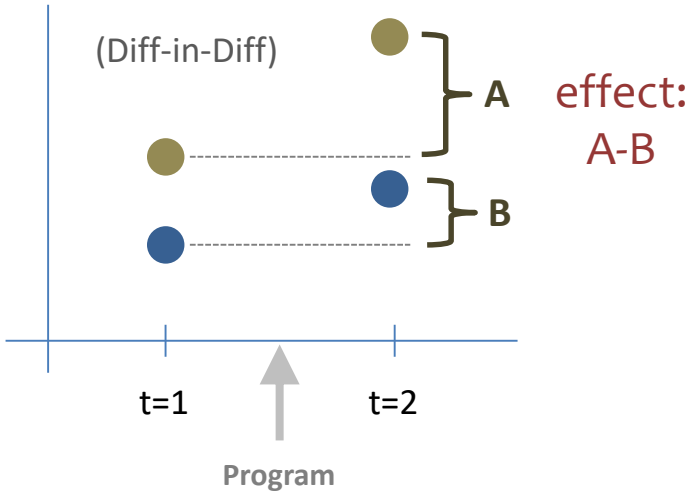
PRE-POST REFLEXIVE
Estimator
effect = T2-T1



POST-ONLY
Estimator
effect = T2-C2



PRE-POST W COMPARISON
Estimator
effect = (T2-T1) – (C2-C1)



Treatment Groups, **T1**=before, **T2**=post-program measure

Control Groups, **C1**=before, **C2**=post-program measure

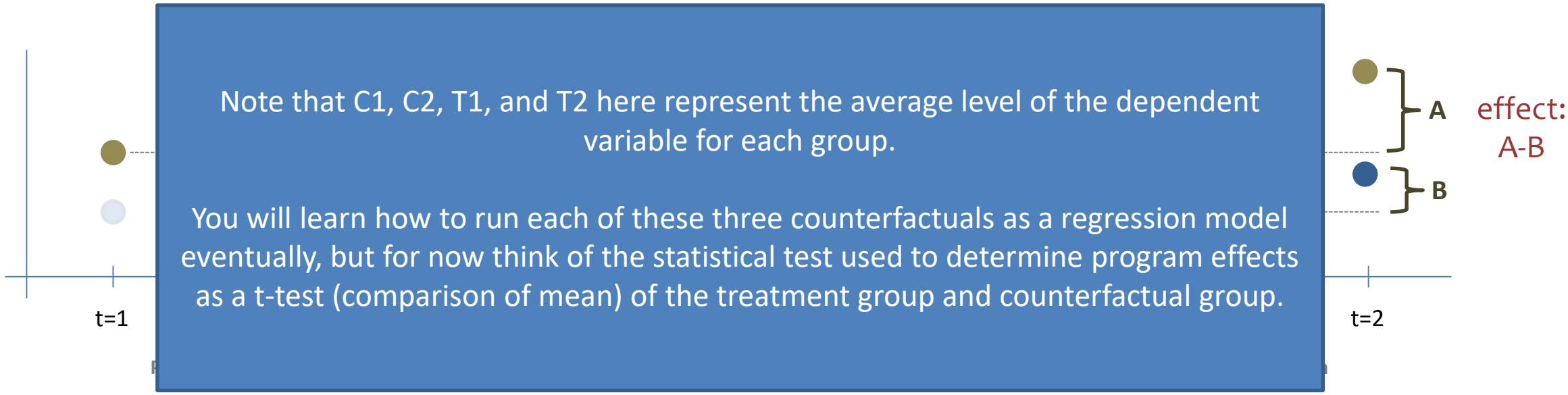
3 VALID COUNTERFACTUALS:

Each variety of counterfactual has a different formula for the program effect estimate.

PRE-POST REFLEXIVE
Estimator
effect = T2-T1

POST-ONLY
Estimator
effect = T2-C2

PRE-POST W COMPARISON
Estimator
effect = (T2-T1) – (C2-C1)



Note that C1, C2, T1, and T2 here represent the average level of the dependent variable for each group.

You will learn how to run each of these three counterfactuals as a regression model eventually, but for now think of the statistical test used to determine program effects as a t-test (comparison of mean) of the treatment group and counterfactual group.

Treatment Groups, **T1**=before, **T2**=post-program measure

Control Groups, **C1**=before, **C2**=post-program measure

Recall from Unit on Dummy Variable Models

(basic set-up for a comparison of group means in regression)

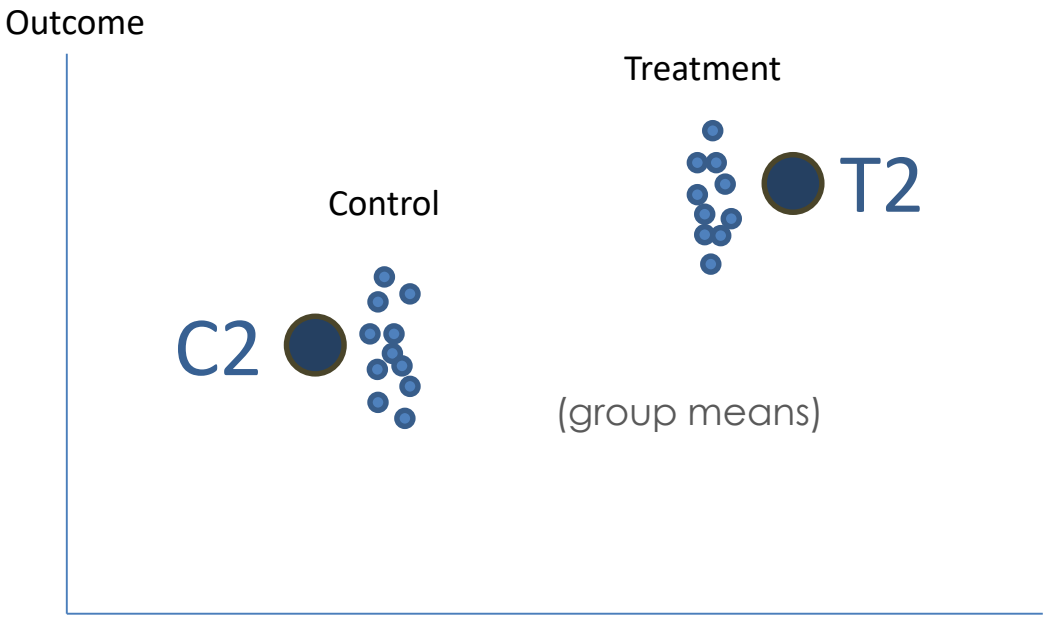
b_0 will measure Y-bar of the omitted group C2

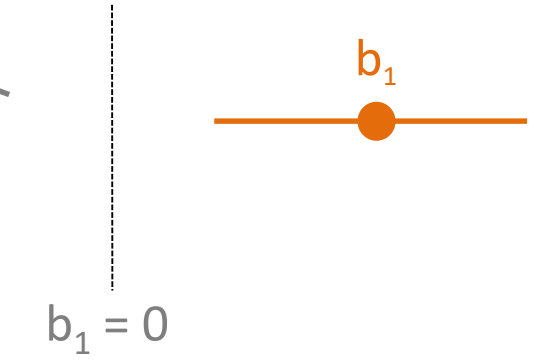
$b_0 + b_1$ represents the Y-bar of T2

$$Y = b_0 + b_1(\text{TreatDummy}) + e$$

$$b_1 = \text{MEAN}_{\text{treat}} - \text{MEAN}_{\text{control}}$$

$$b_1 = \text{T2} - \text{C2}$$

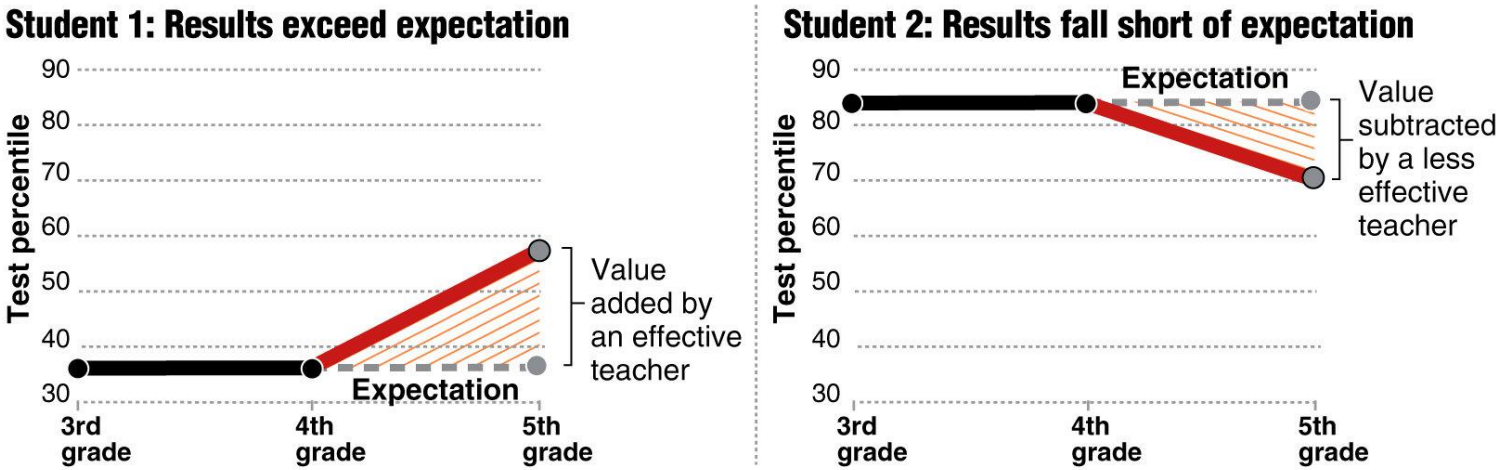




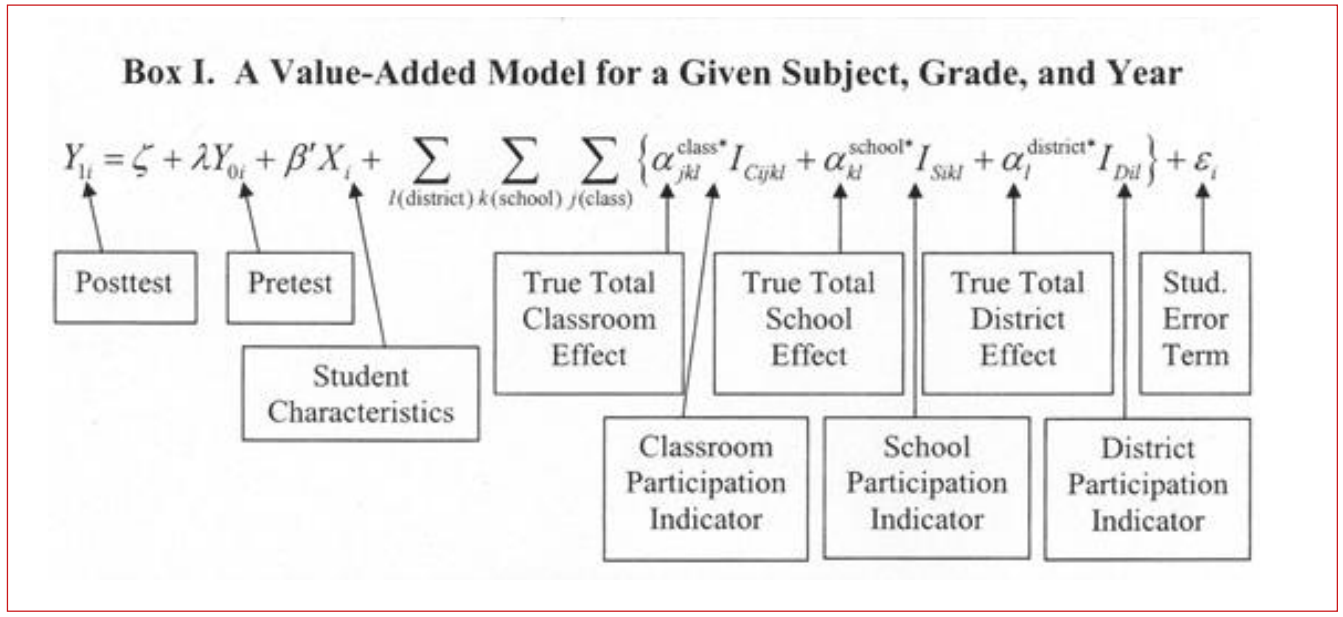
The default hypothesis test in the regression then uses b_1 to **test for a meaningful difference between T2 and C2**

What is 'value added'?

"Value added" rates teachers based on their students' progress on standardized tests year after year. The difference between a student's expected growth and actual performance is the "value" a teacher adds or subtracts during the year.



Source: California Standards Tests, Los Angeles Unified School District, Los Angeles Times reporting
Graphic: Los Angeles Times



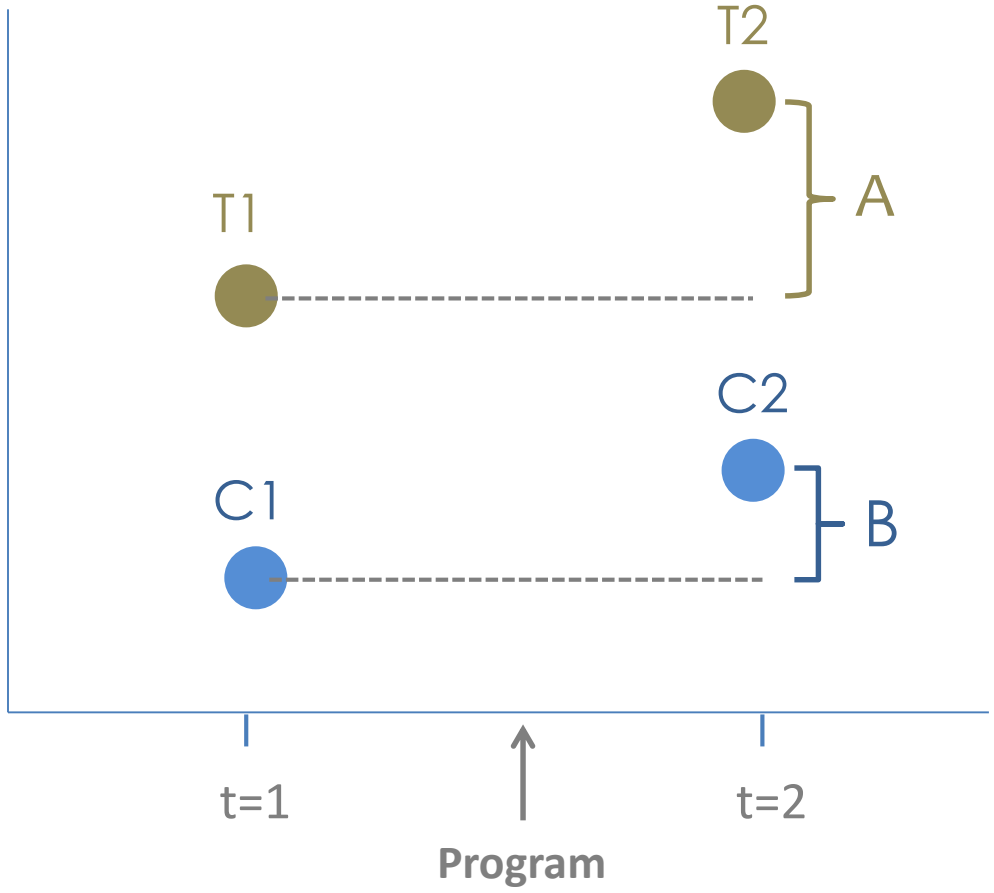
When we have lots of groups, lots of time periods, and lots of control variables at different levels our models will eventually look like this.

But not this semester.

Difference-in-Difference Estimator

Pre-Post w Comparison
Estimator

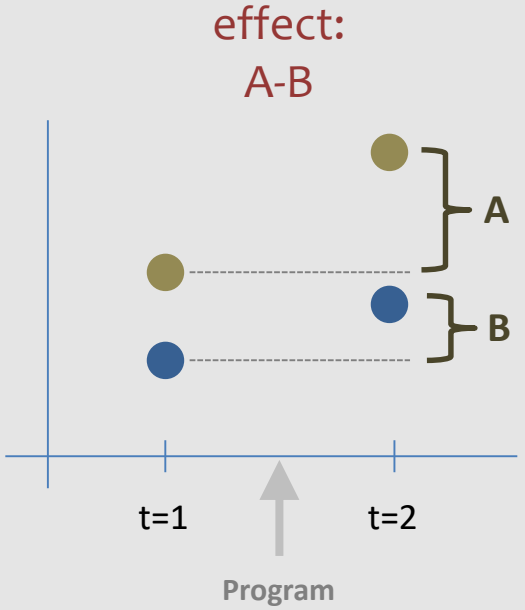
effect = **(T2-T1) – (C2-C1)**



effect:
A-B
(Diff-in-Diff)

Accounts for initial difference in groups and secular trends

Difference-in-Difference Estimator

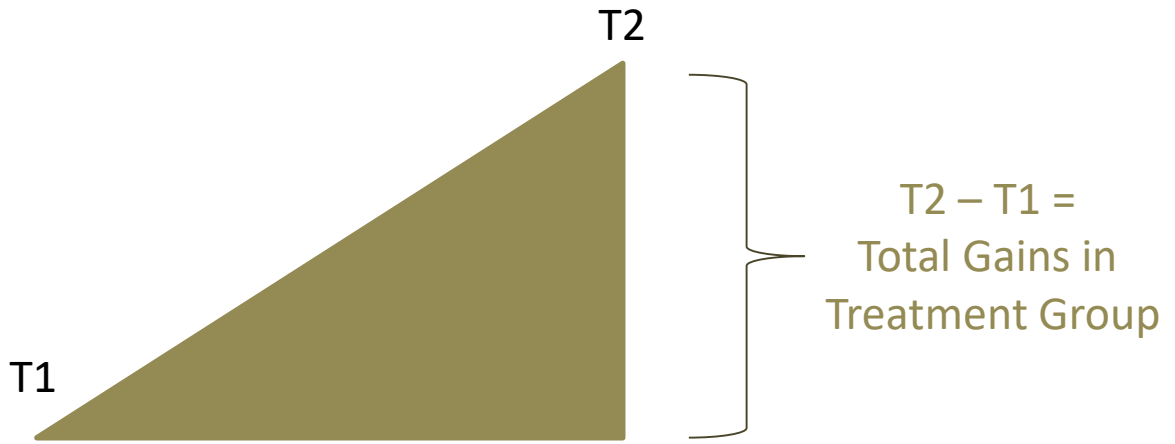


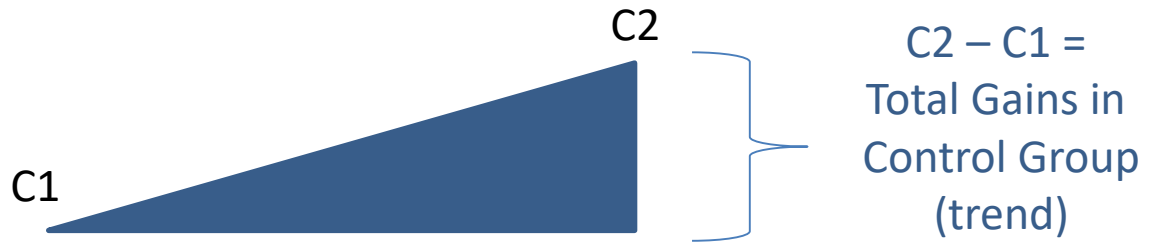
Pre-Post w Comparison

$$\text{effect} = (T2\text{-}T1) - (C2\text{-}C1)$$

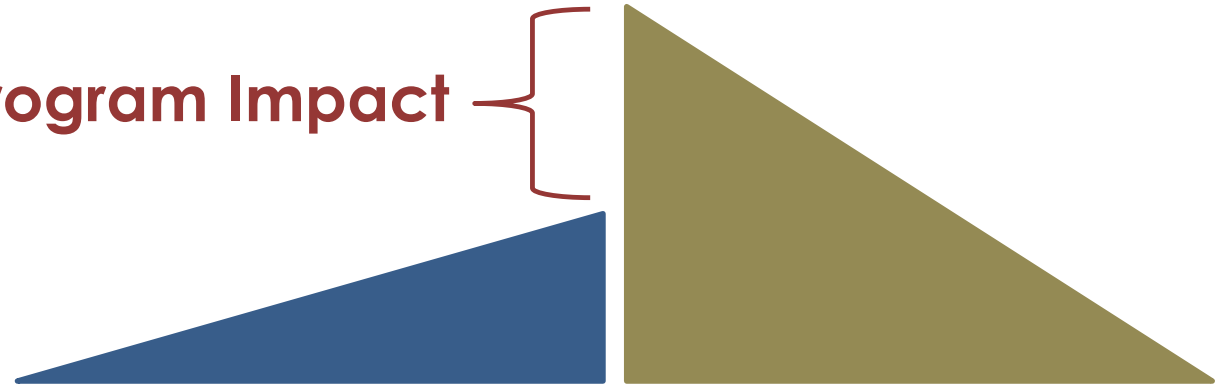
↑ Gains during the program period by treatment group

↑ Gains independent of the program (secular trends)

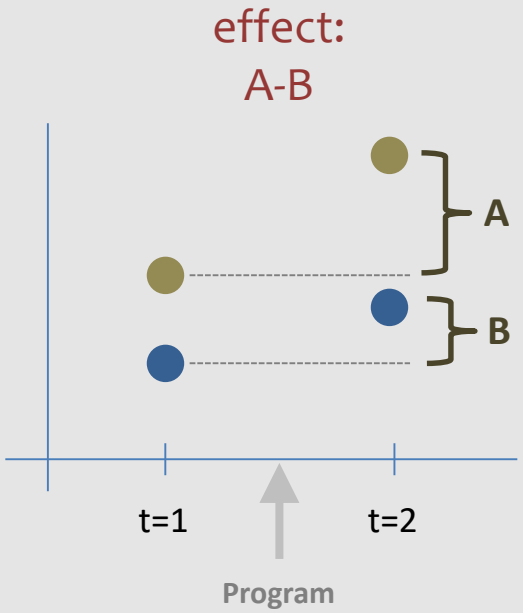




(T2-T1)-(C2-C1) = **Program Impact**



Difference-in-Difference Estimator



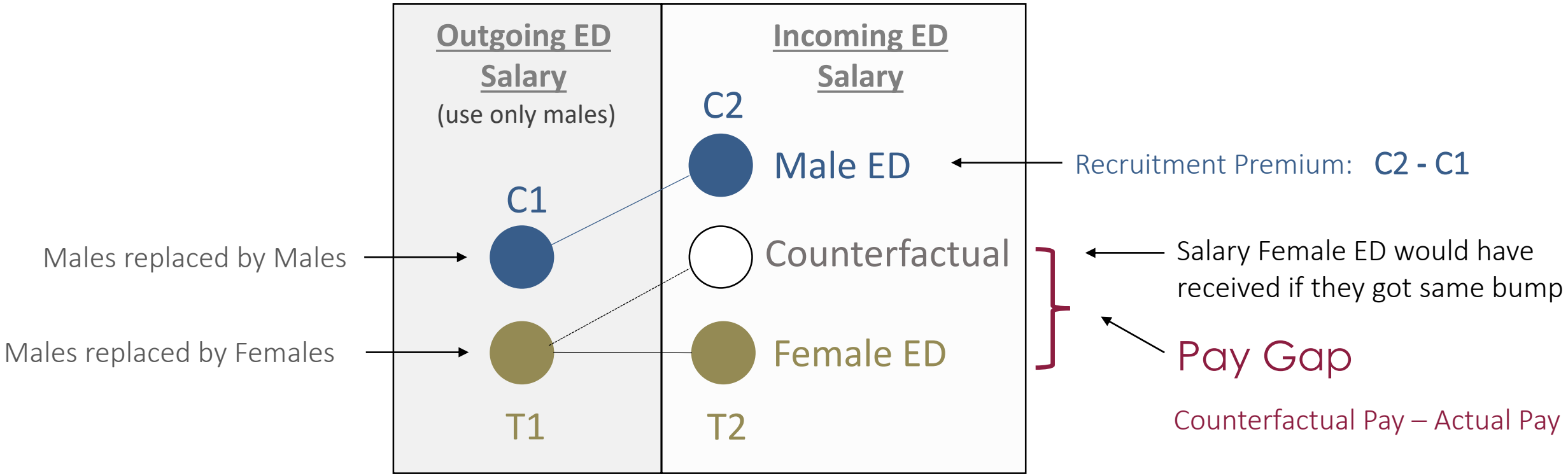
$$effect = (T2-T1) - (C2-C1)$$

Two important things to note:

(1) The groups prior to the treatment are not identical. The difference in difference is robust because it does not require equivalent treatment and "control" (comparison) groups.

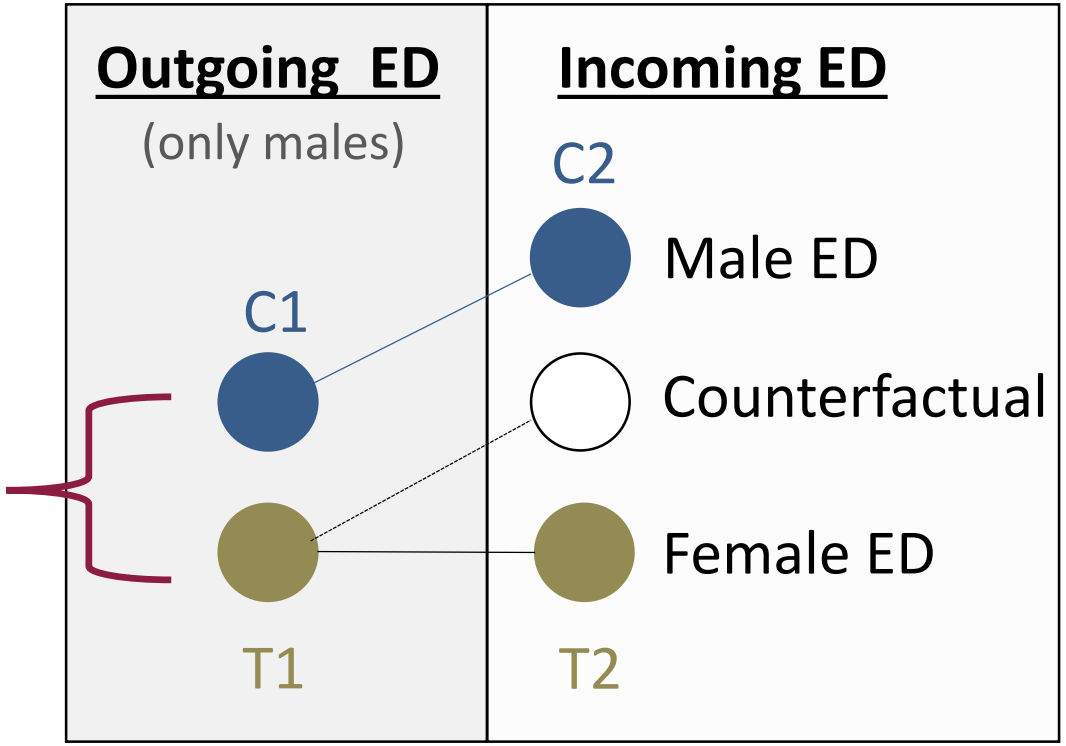
(2) The comparison group measures the changes we would expect to observe in the treatment group over time independent of the treatment. This assumption must be true for the diff-in-diff estimator to be unbiased.

Example: Estimating the gender pay gap for Executive Directors



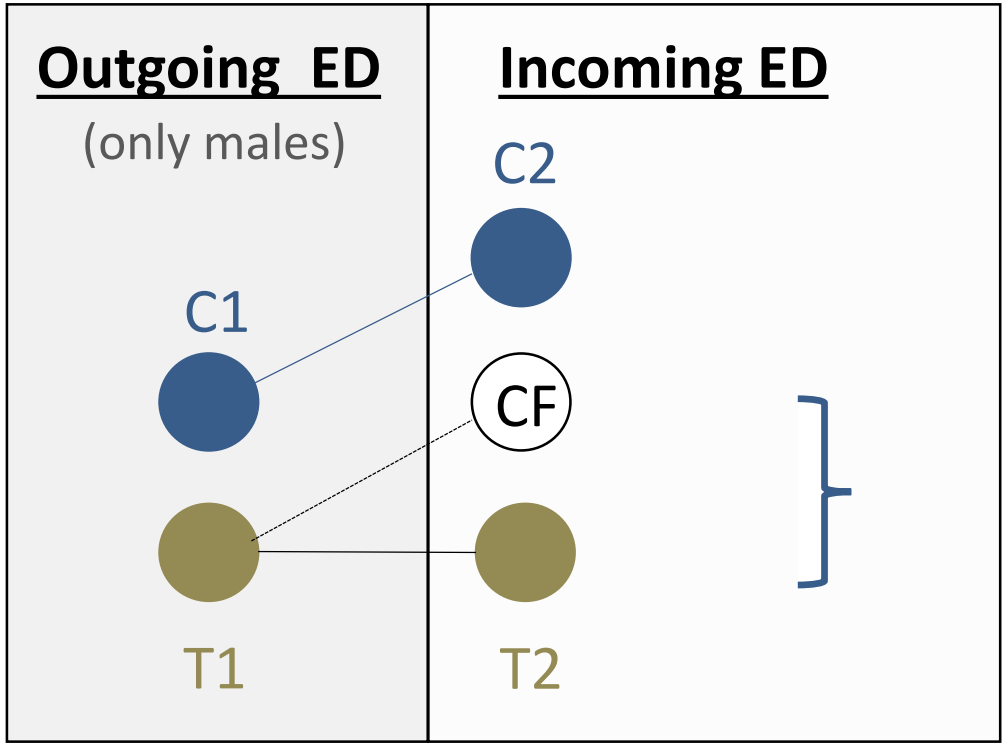
Counterfactual Pay: Baseline Pay (T1) + Typical (male) Recruitment Premium (C2 – C1)

Note that in the diff-in-diff the groups do not have to be identical prior to the treatment (new hire)

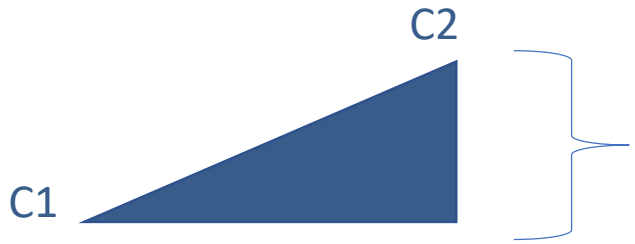


This is the major strength of the estimator.

Also note that we never compare C2 to T2 (the two post-treatment groups)

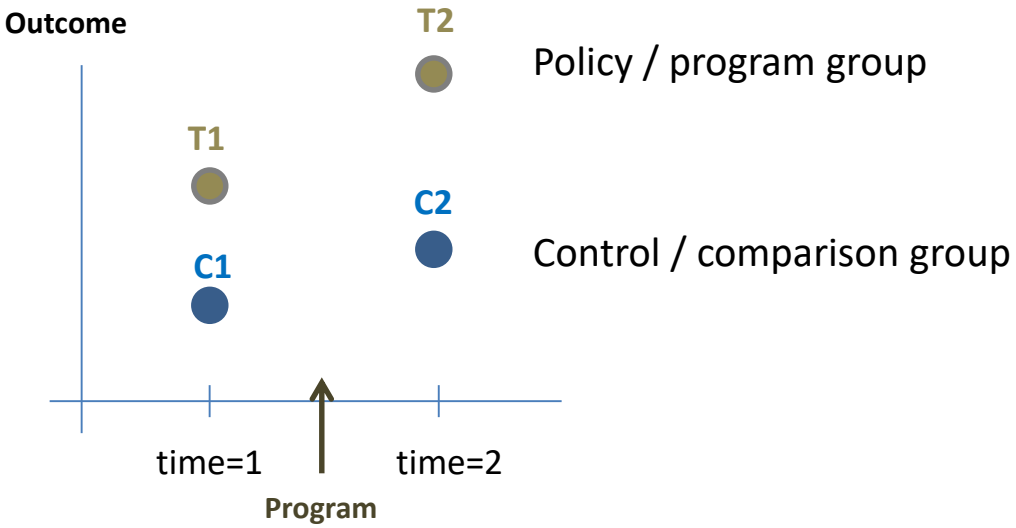


The proper counterfactual is how much would female EDs be getting paid if they were given recruitment premiums as incoming male EDs.



The comparison group is there to capture the trend, which is then used to construct the counterfactual.

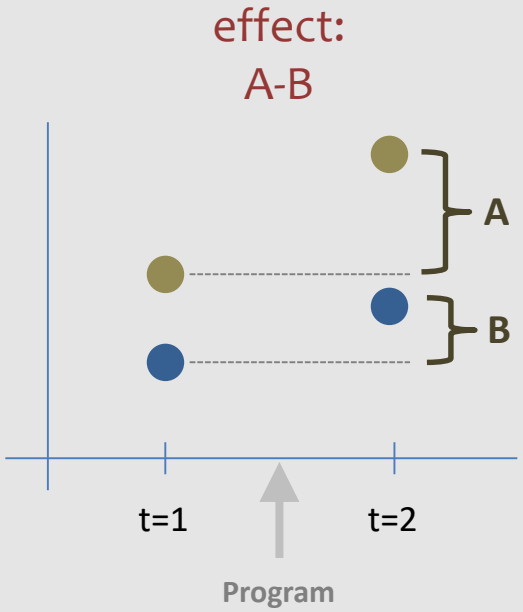
"Control" Versus "Comparison" Groups



The term "control group" typically refers to a group created through randomization, so they are assumed to be identical. We sometimes use the terminology "comparison group" to differentiate the these cases. The control groups represents the counterfactual, the comparison group does not (it just captures trend in this case).

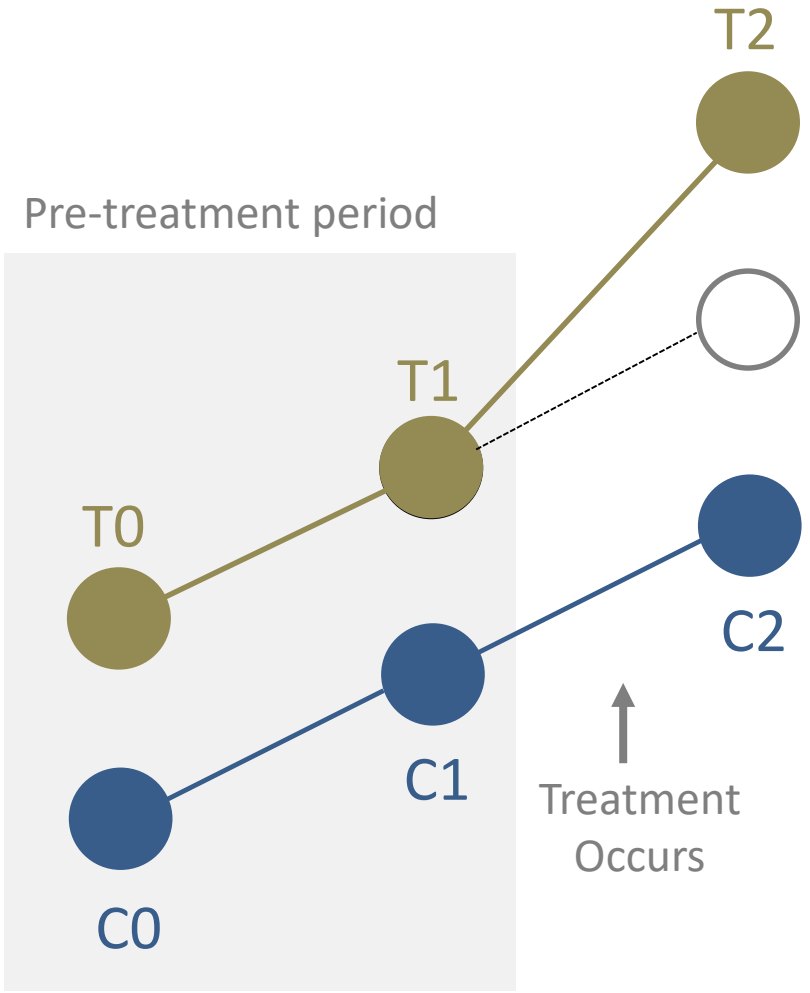
Single Difference:	Difference in Difference:
Effect = T2 – T1	Effect = (T2 – T1) – (C2 – C1)
	trend
Effect = T2 – C2	

Difference-in-Difference Estimator



Parallel lines test:

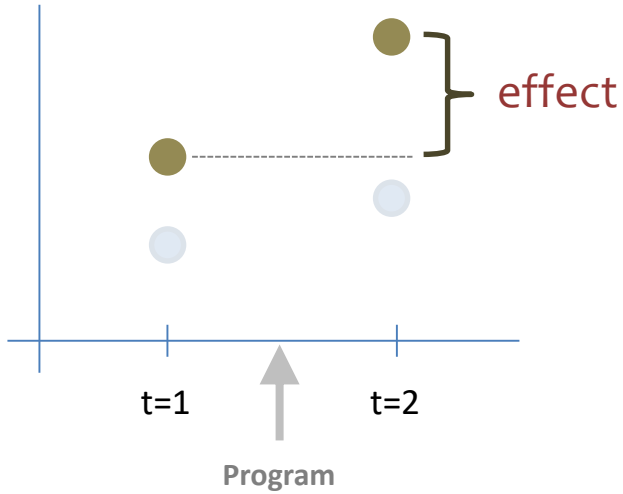
$$T1-T0 = C1-C0$$



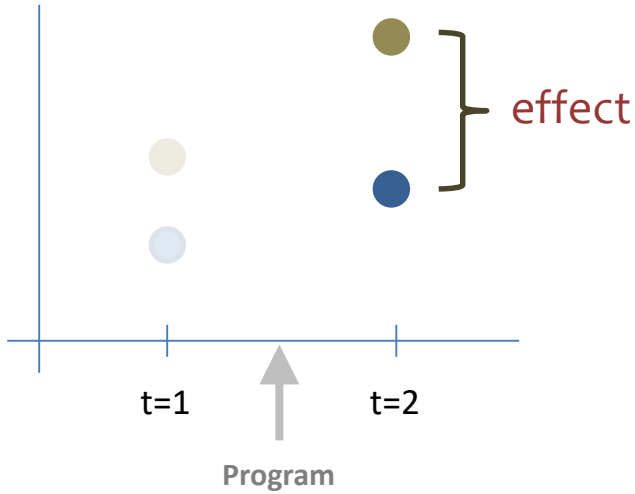
Parallel lines assumption: for the diff-in-diff counterfactual to be valid C2-C1 has to accurately capture the secular trend (gains independent of the treatment). The test of this is comparing rates of change in treatment and comparisons groups prior to the intervention. They should b the same.

2 REMAINING COUNTERFACTUALS:

PRE-POST REFLEXIVE
Estimator
effect = T2-T1



POST-ONLY
Estimator
effect = T2-C2



PRE-POST W COMPARISON
Estimator
effect = (T2-T1) – (C2-C1)

Validity of the reflexive estimator:

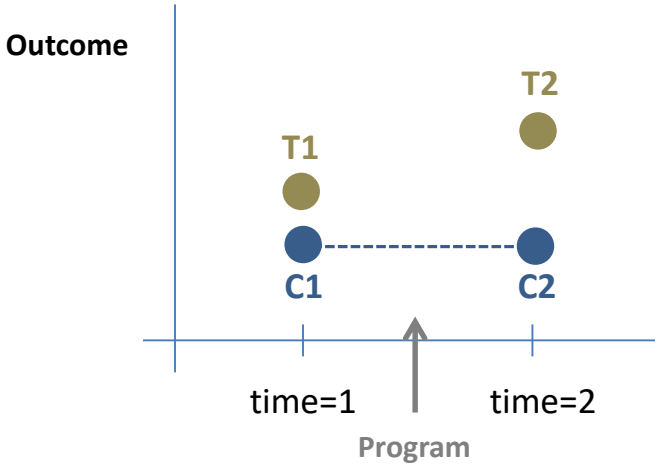
Reflexive (Pre-Post)

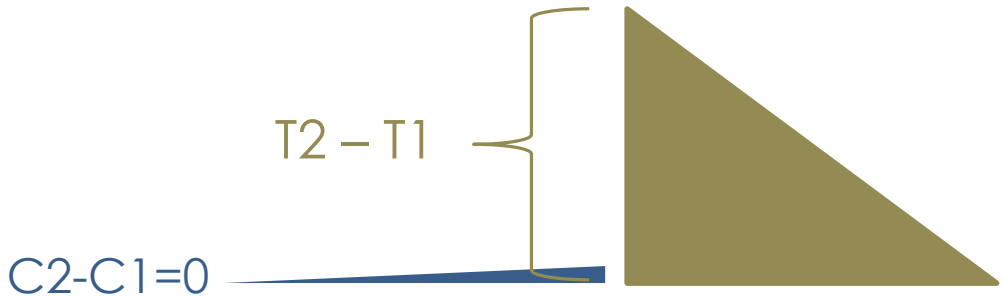
$(T2 - T1) - (C2 - C1) = T2 - T1$

IFF

$C2 - C1 = 0$
(no trend)

In some special cases we expect there to be no change independent of the treatment. When this is true the reflexive estimator T2-T1 is appropriate and unbiased.





b_0 will measure Y-bar of the omitted group **T1**

$b_0 + b_1$ represents the Y-bar of **T2**

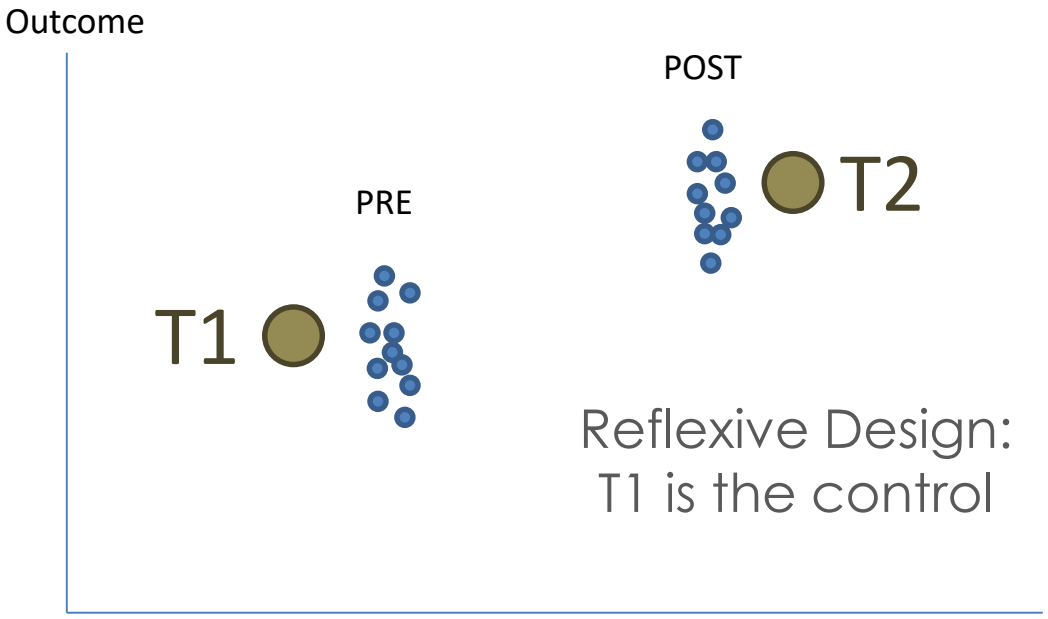
$$Y = b_0 + b_1(\text{TreatDummy}) + e$$

$$b_1 = \text{MEAN}_{\text{treat}} - \text{MEAN}_{\text{control}}$$

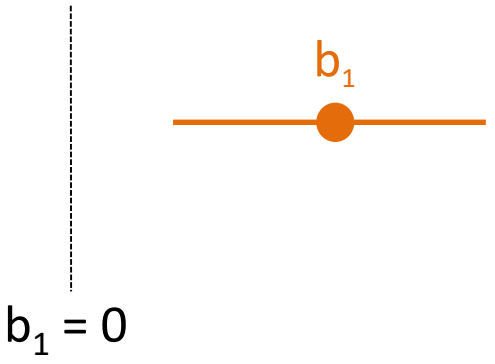
$$b_1 = T2 - T1$$

The default hypothesis test in the regression then uses $\mathbf{b_1}$ to test for a meaningful difference between T2 and T1

In reflexive models T1 is the counterfactual or control group, otherwise the regression set-up is the same

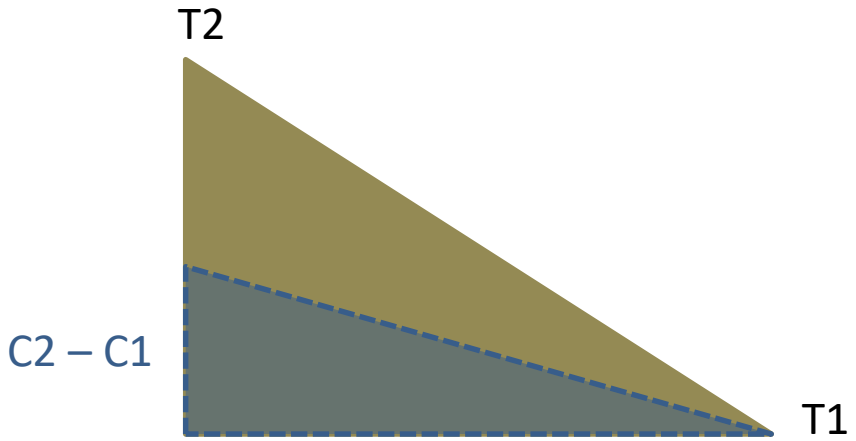


Reflexive Design:
T1 is the control



The comparison group is necessary to capture any changes that we expect during the study period independent of the treatment group. If we fail to account for those changes we will incorrectly attribute all change to the program and over-state program impact.

Program Impact?



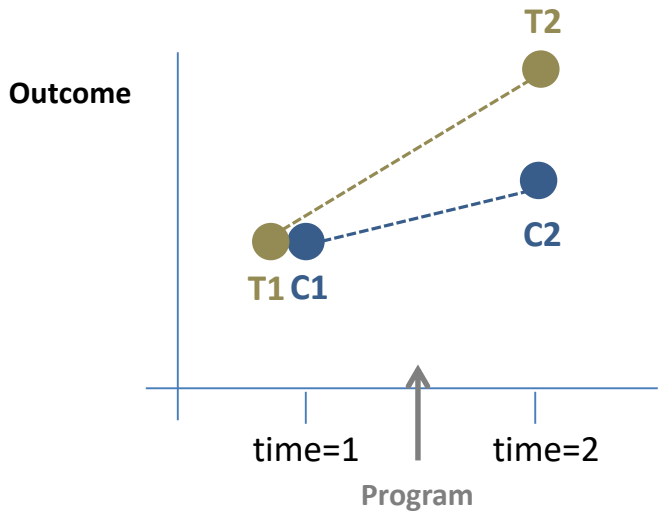
Validity of the post-test only estimator:

Post-Only

$(T2 - T1) - (C2 - C1) = T2 - C2$

IFF

$C1 - T1 = 0$
(equivalent at time 1)



If we have confidence that the two groups are identical prior to the treatment, then mathematically T2-C2 will still account for gains independent of the treatment. This condition is necessary for the post-test only estimator to be unbiased.

In experimental design, this is usually accomplished through randomization or lottery.

Observational students typically use matching models to create equivalent groups.

Recall from Unit on Dummy Variable Models

(basic set-up for a comparison of group means in regression)

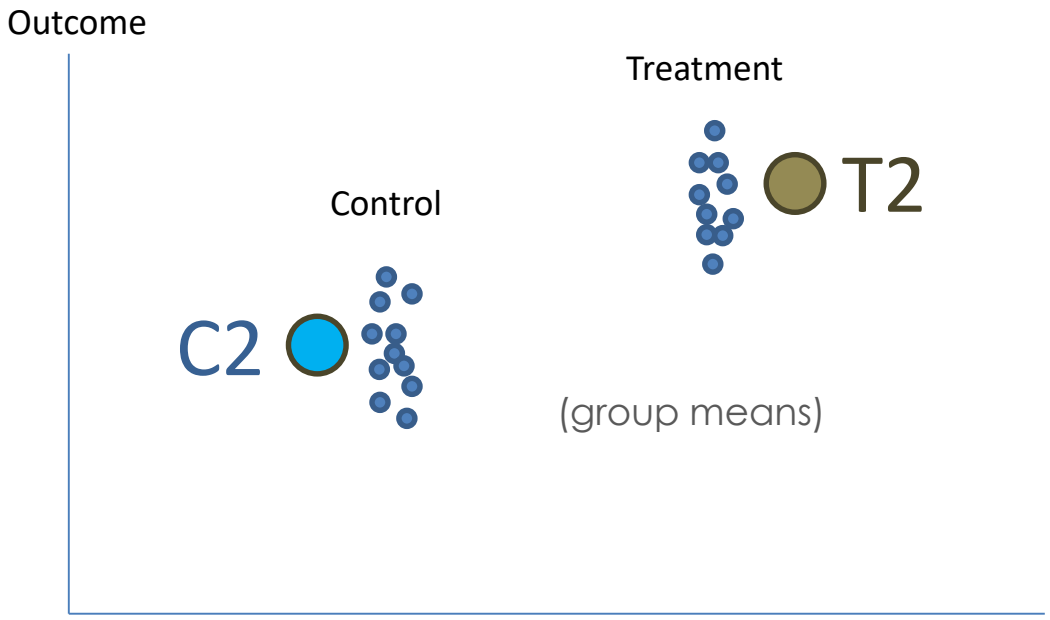
b_0 will measure Y-bar of the omitted group C2

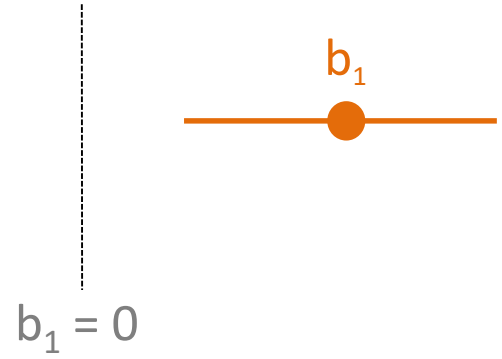
$b_0 + b_1$ represents the Y-bar of T2

$$Y = b_0 + b_1(\text{TreatDummy}) + e$$

$$b_1 = \text{MEAN}_{\text{treat}} - \text{MEAN}_{\text{control}}$$

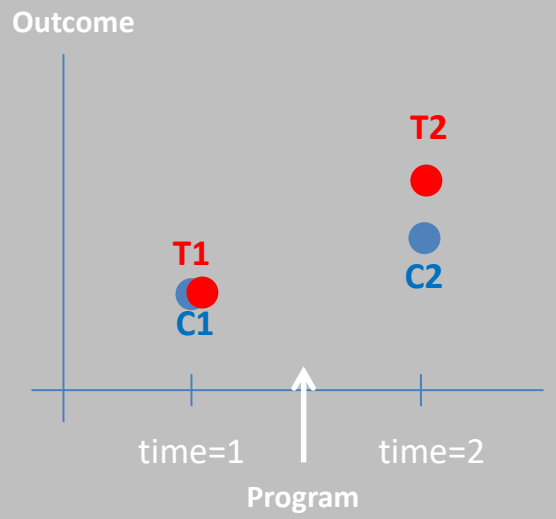
$$b_1 = \text{T2} - \text{C2}$$

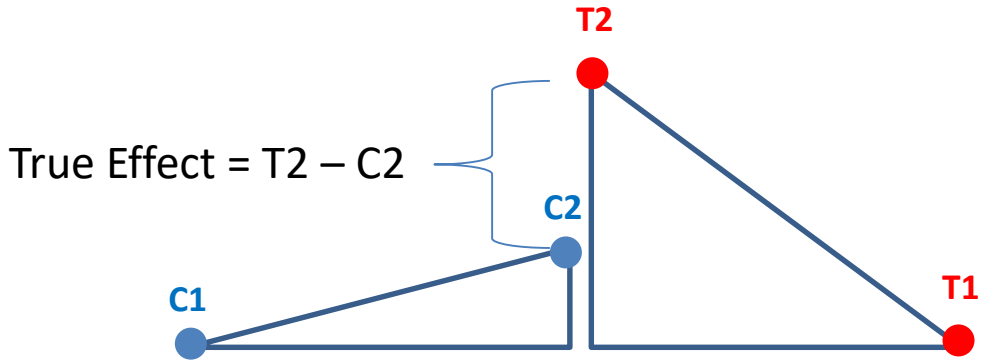




The default hypothesis test in the regression then uses b_1 to **test for a meaningful difference between T2 and C2**

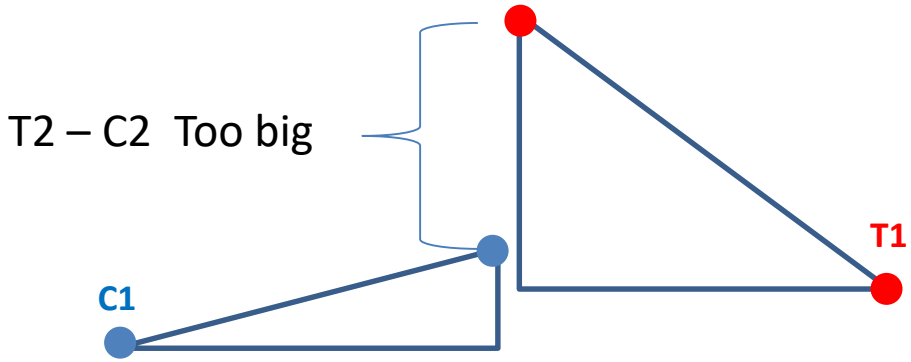
Post-Test Only Measure





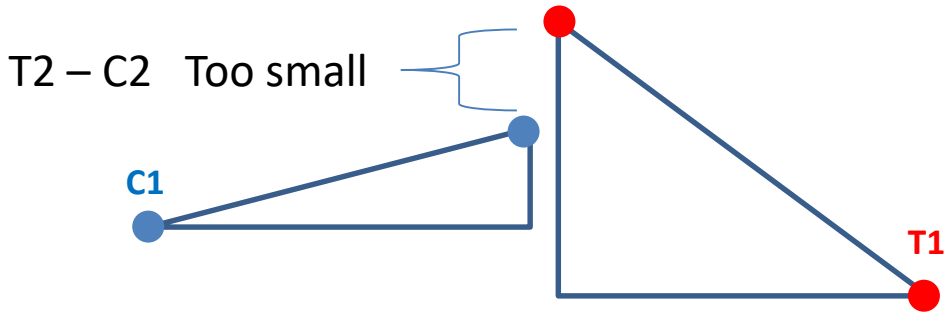
If C1 = T1

Measured effect accurately represents program impact



C1 < T1 biased

Measured effect overstates program impact

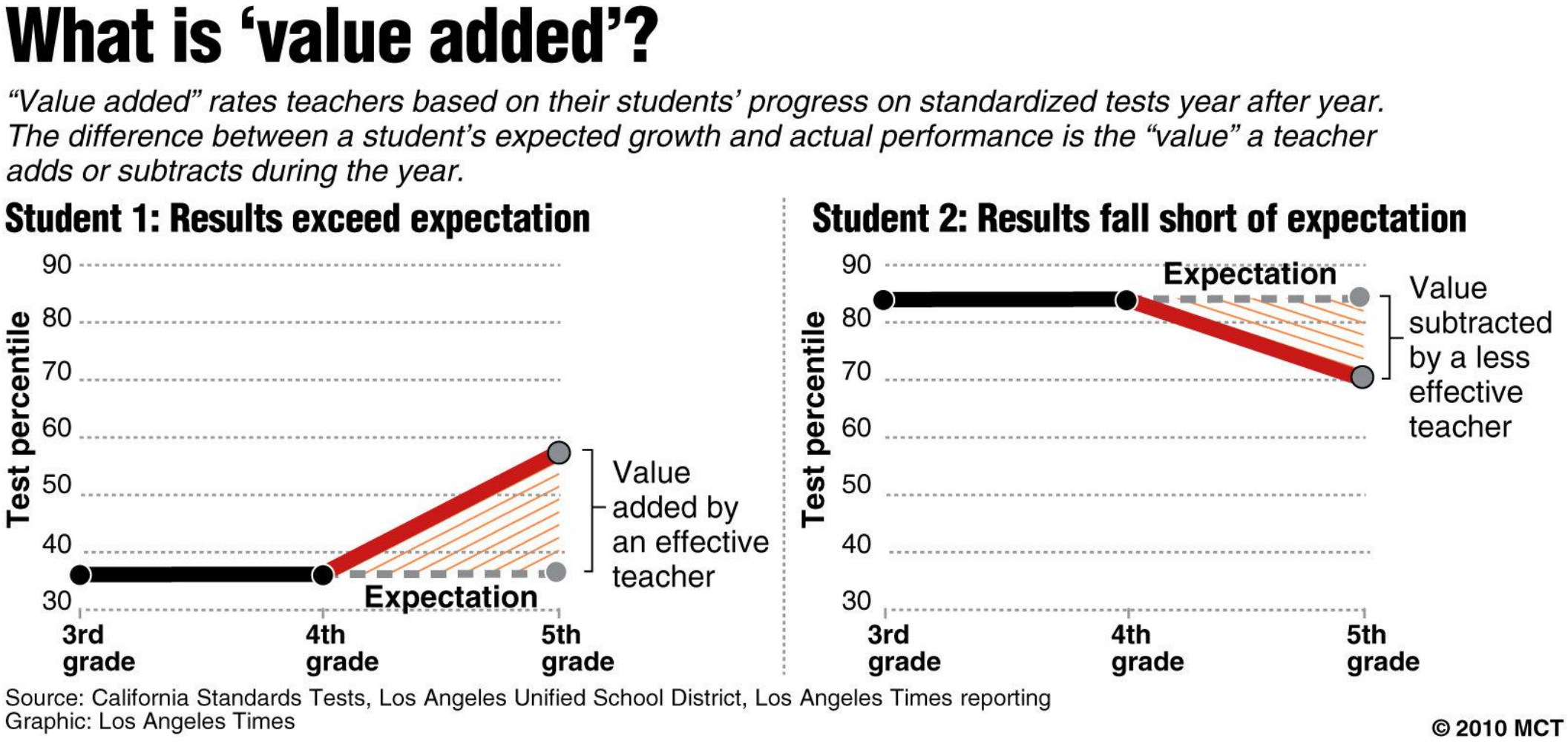


C1>T1 biased

Measured effect understates program impact

EXAMPLES

Example from education policy – which estimator is this?

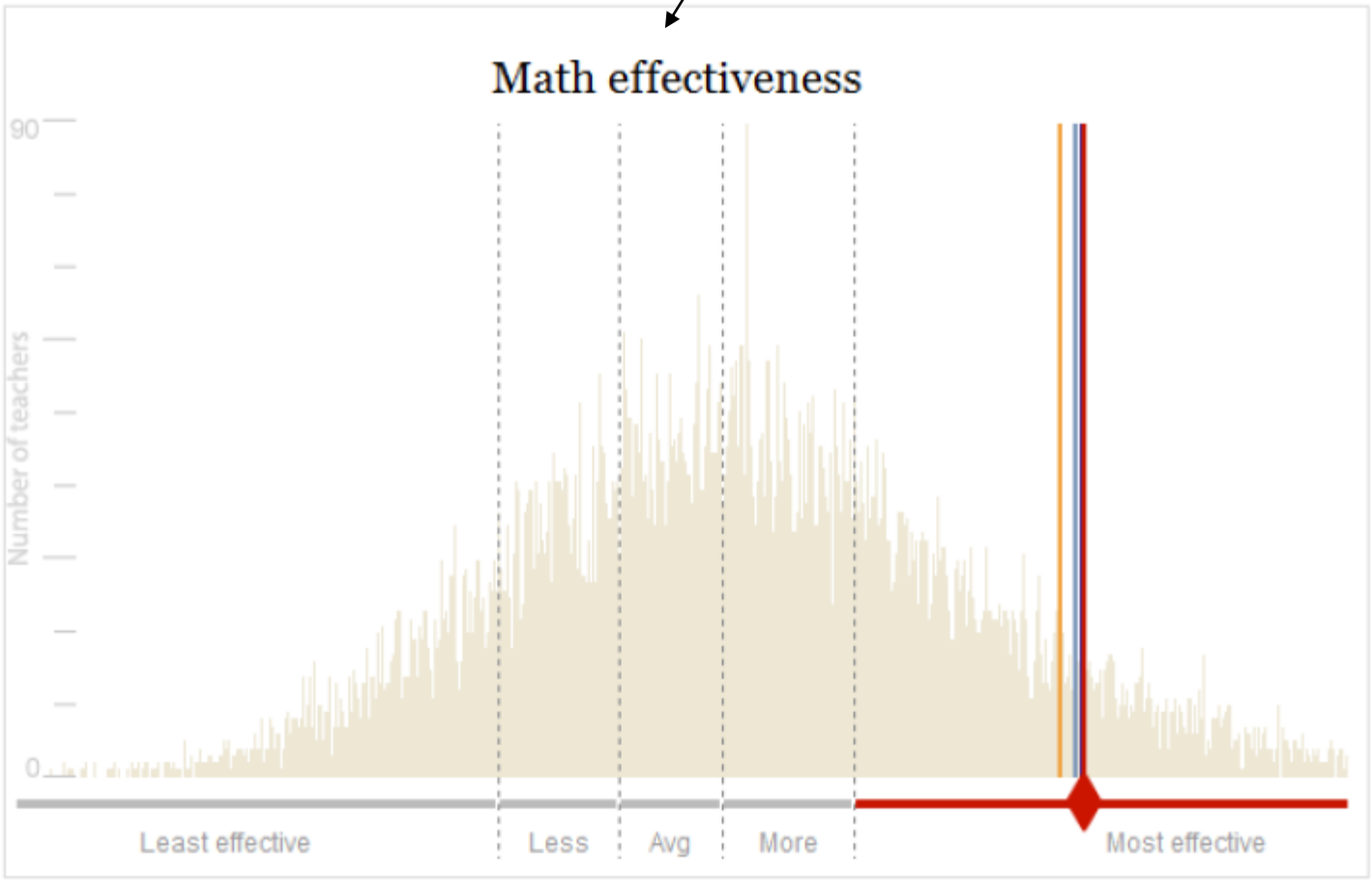


COMPARING TEACHER "EFFECTS"

Change in student performance from the beginning of the academic year to the end of the academic year (measured in percentile units).

Reflexive measure because our notion of "impact" comes from comparing performance of the same students over time, not comparing one classroom to another at a point in time.

Note, because we use percentiles for performance (tests will be different at the end of each grade so it is hard to compare raw scores) there is an element of a difference-in-difference design. But percentiles require that the average change across all groups (the secular trend or expectation of gain independent of the treatment) has to be zero across all groups because of the math. But if it is a sub-population there could be a trend, in which case the diff-in-diff might be appropriate.



REVISIT THE MICRO-FINANCE EXAMPLE

Microfinance example of bias from selection INTO a study group

Number of each "type" of person in the study

	NOT Entrepreneurial	Entrepreneurial
No Loan	30	15
Takes a Loan	20	35

Average weekly income after loan period

	NOT Entrepreneurial	Entrepreneur
No Loan	$10	$20
Takes a Loan	$10	$20

Income not impacted by the loan

You are more likely to take a loan if you know you are good at business

Takes Loan?

The loan appears to have an impact!

$$NO: \quad \frac{30 \cdot \$10 + 15 \cdot \$20}{45} = \$13.33$$

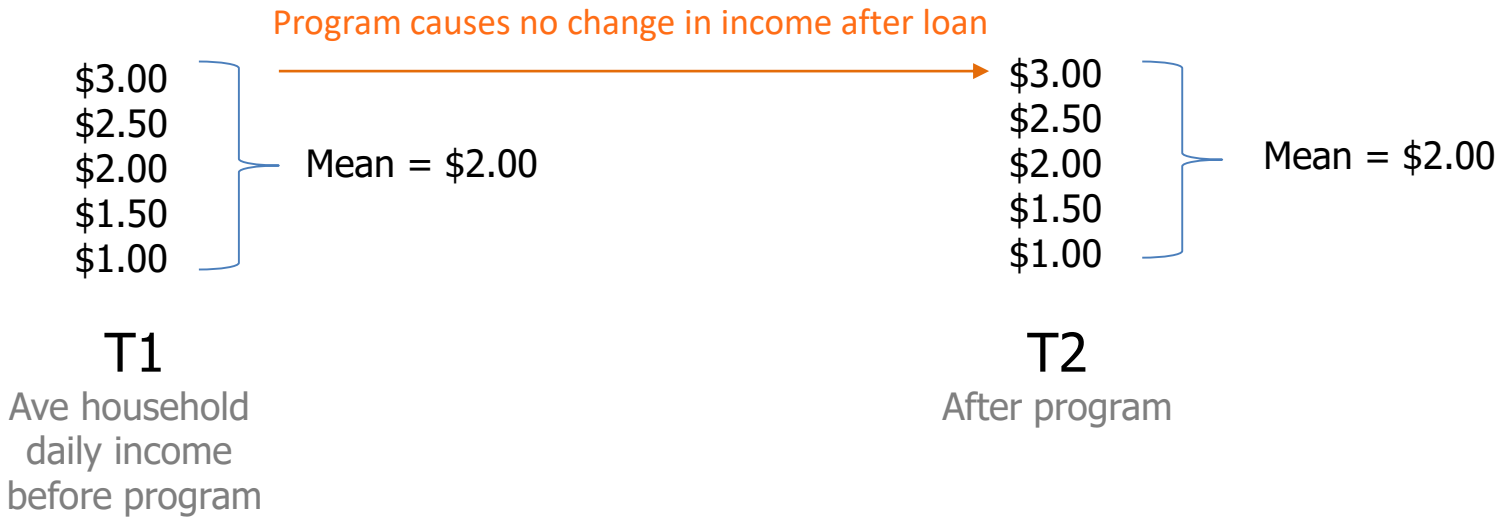
$$YES: \quad \frac{20 \cdot \$10 + 35 \cdot \$20}{55} = \$16.37$$

Which counterfactual estimator are we using here?

Which counterfactual estimator would be unbiased?

Microfinance example of bias from selection OUT OF a study group

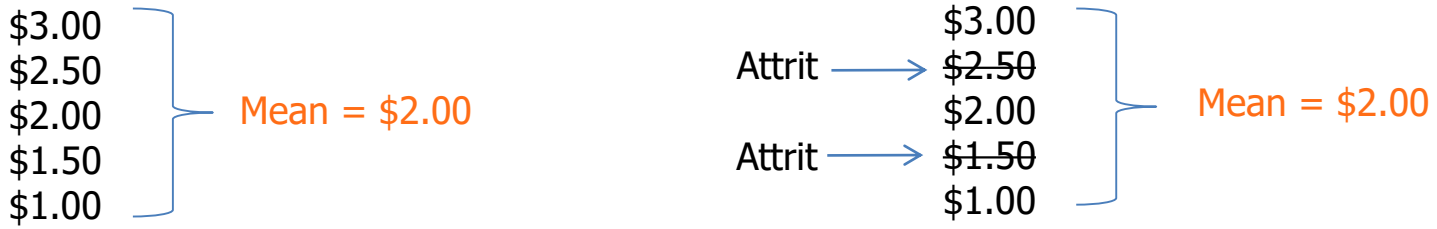
Reflexive design



$$T2 - T1 = 0$$

causal estimate is unbiased

Random Attrition Example



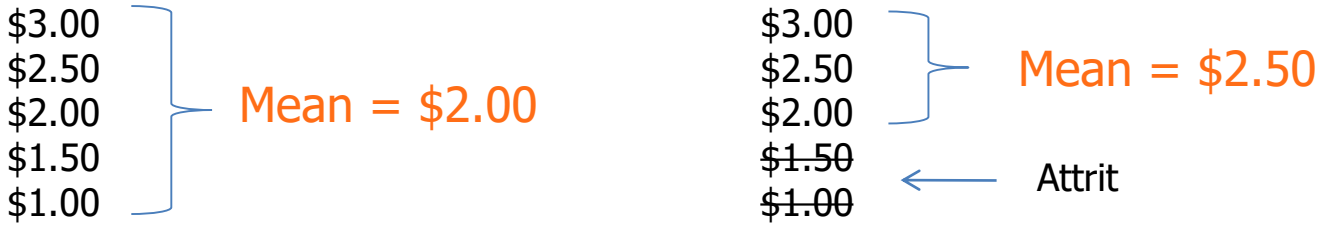
T2 - T1 = 0

Impact study accurately represents program effects

Program is not determined to be effective (no change)

Which estimator is used here?

Non-Random Attrition Example



$$T2 - T1 \neq 0$$

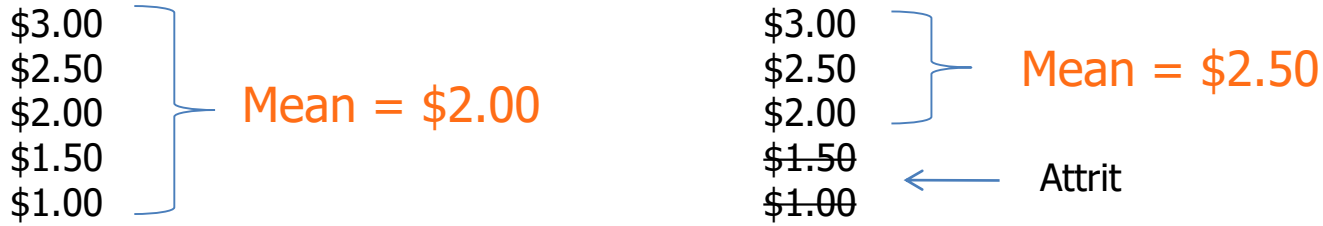
We over-estimate program effects

Program appears to be effective

Which estimator is used here?

Can we use a different one?

Non-Random Attrition Example



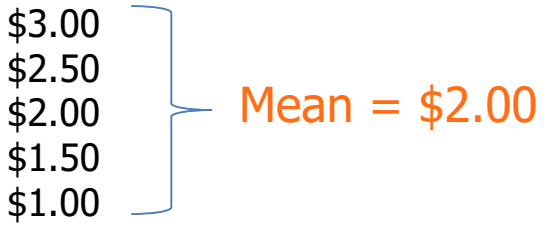
Reflexive → $T2 - T1 \neq 0$

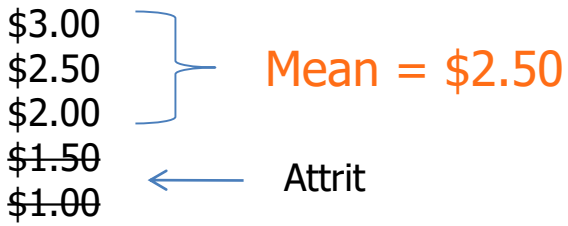
We over-estimate program effects

Program appears to be effective

Which effect measure is used here? Intention to treat, or treatment on the treated?

Non-Random Attrition Example





$$T2 - T1 \neq 0$$

We over-estimate program effects

Program appears to be effective

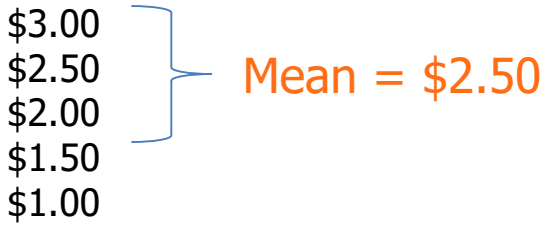
Can we use a better treatment measure?

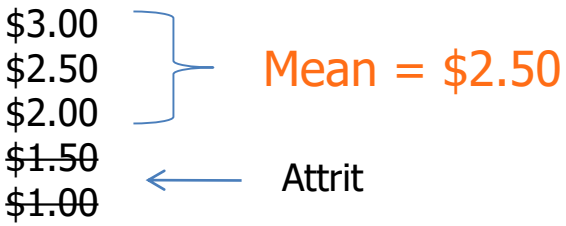
Which effect measure is used here? Intention to treat, or treatment on the treated?

Non-Random Attrition Example

Remove attriters from T1

T2 – T1 = 0





This is not about ITT vs TOT! To calculate those we need data on both groups in T2. Rather this is about how we use data when attrition is present

PREVIEW OF CPP 525: ADVANCED REGRESSION

COUNTERFACTUALS IN PRACTICE: QUASI-EXPERIMENT METHODS

1) **Difference-in-Difference: Effect = (T2-T1) – (C2-C1)**

    1) Difference-in-Difference Analysis
    2) Fixed Effects Models

2) **Reflexive Design: Effect = T2 – T1**

    1) Pre-Post Design with only a Treatment Group
    2) Time Series Regression

3) **Post-Test Only Design: Effect = T2 – C2**

    1) True Experiments (T2-C2)
    2) Post-Test Only Comparison, No Baseline (T2-C2)
    3) Propensity Score Matching
    4) Regression Discontinuity Design

Box I. A Value-Added Model for a Given Subject, Grade, and Year

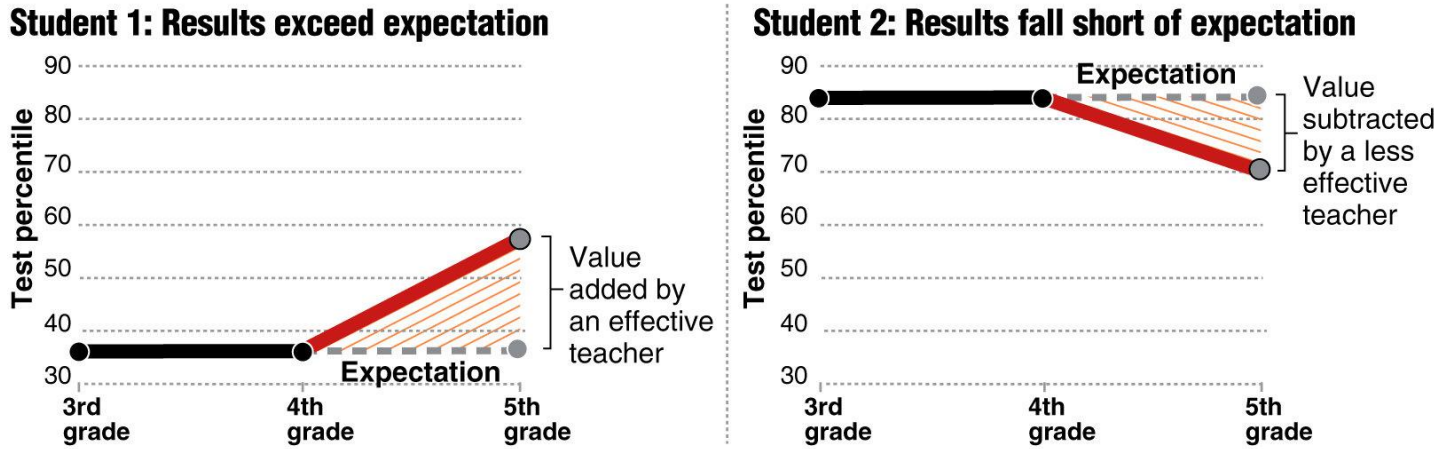
$$Y_{1i} = \zeta + \lambda Y_{0i} + \beta' X_i + \sum_{l(\text{district})} \sum_{k(\text{school})} \sum_{j(\text{class})} \left\{ \alpha_{jkl}^{\text{class}^*} I_{Cijkl} + \alpha_{kl}^{\text{school}^*} I_{Sikl} + \alpha_{l}^{\text{district}^*} I_{Dil} \right\} + \varepsilon_i$$

- Posttest
- Pretest
- Student Characteristics
- True Total Classroom Effect
- True Total School Effect
- True Total District Effect
- Stud. Error Term
- Classroom Participation Indicator
- School Participation Indicator
- District Participation Indicator

These will start to make more sense.

What is 'value added'?

"Value added" rates teachers based on their students' progress on standardized tests year after year. The difference between a student's expected growth and actual performance is the "value" a teacher adds or subtracts during the year.



Source: California Standards Tests, Los Angeles Unified School District, Los Angeles Times reporting
Graphic: Los Angeles Times