



Virtual Training Workshop on Counterfactual Impact Evaluation (CIE)

C4ED – EUTF
September 2021

Welcome to the Training Workshop on Counterfactual Impact Evaluation (CIE)

The material of this workshop was produced with the financial support of the European Union. Its contents are the sole responsibility of C4ED and do not necessarily reflect the views of the European Union

Day 2 Agenda

09:00 – 09:30	Discussion of assignment using Experimental Methods
9:30 – 10:30	Session 4A: Introduction to Quasi-Experimental Methods: DiD, Matching, IV, RDD
20 min	Break
10:50 – 12:00	Session 4B: Introduction to Quasi-Experimental Methods: DiD, Matching, IV, RDD + Assignment of applied exercise.
45 min	Lunch Break
12:45 – 13:05	Discussion of assignment using Quasi-Experimental Methods
13:05 – 14:05	Session 5A: Setting the expectations right – timelines, data needs (data sources and sample size) and budget
15 min	Break
14:20 – 15:00	Session 5B: Setting the expectations right – timelines, data needs (data sources and sample size) and budget
15:00 – 15:30	Q&A

Communication during the training

- post your questions in the chat room
- like questions of others, so we know they are particularly relevant or urgent
- Carolin will read out all questions, which will be answered in between topics or at end of sessions
- use the longer breaks to ask more questions
- suggest improvements if you can't follow or disagree (we are open to criticism and constructive suggestions for improvement)
- more feedback and questions (especially for the Q&A session):
Send an email to Zahra Sharafi (z.sharafi@c4ed.org) or Dr. Giulia Montresor (g.montresor@c4ed.org)

Recap

Experimental impact evaluation

- Experiments use a counterfactual framework to ensure that units in T and C groups are on average statistically identical in (un) observable characteristics through random assignment of the intervention
- Experiments are not always feasible:
 - Randomization may not be socially or politically acceptable
 - Randomization may not be feasible
 - IE is designed only after implementation starts

Quasi-experimental IE methods

Quasi-experimental impact evaluation

- Construct the counterfactual by making a set of assumptions that help to establish comparability between T and C groups
- Assumptions are not testable



Center for Evaluation
and Development

How impact has often been measured..

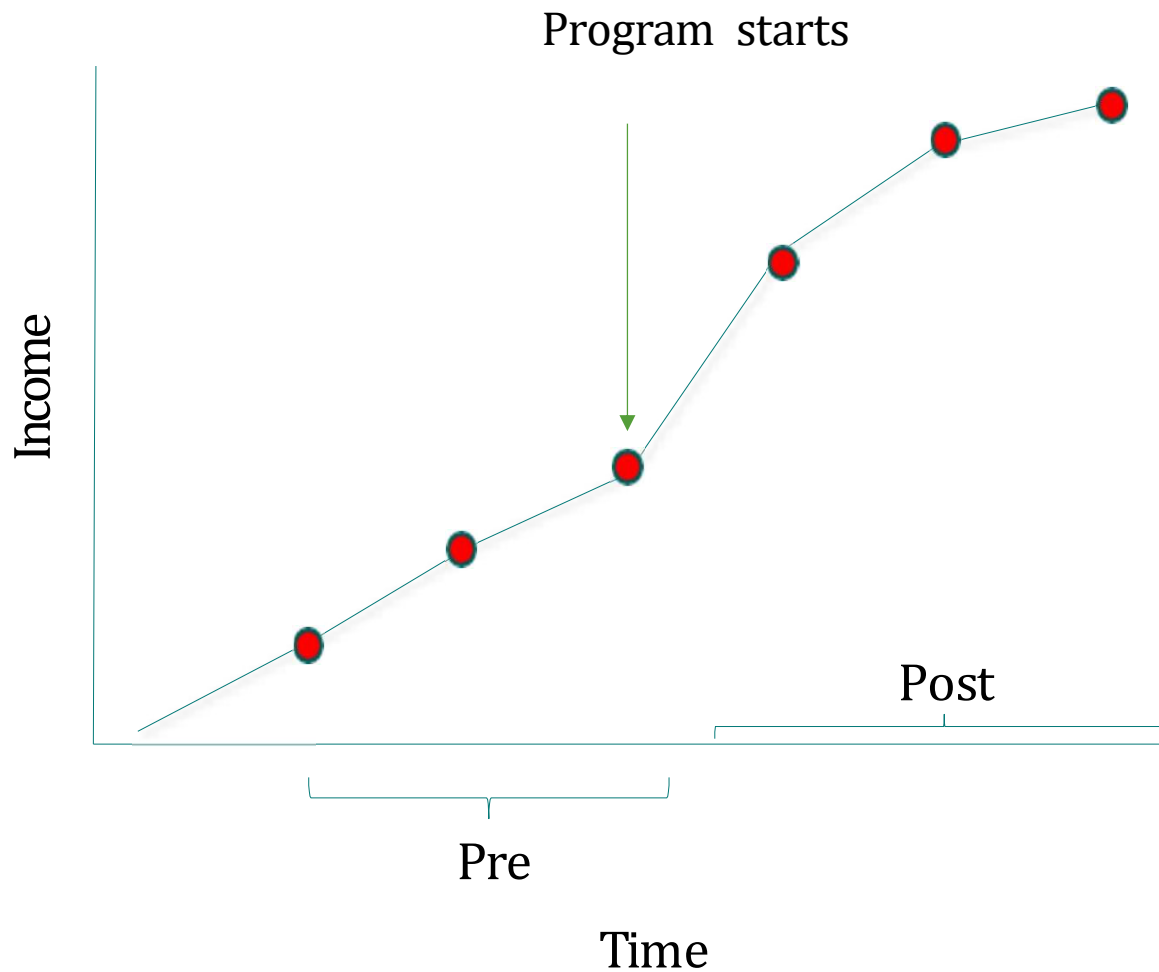
BEFORE – AFTER COMPARISON

Before-After Comparison

- You may compare the outcome level of participants BEFORE and AFTER intervention
- You need: data before and after the intervention

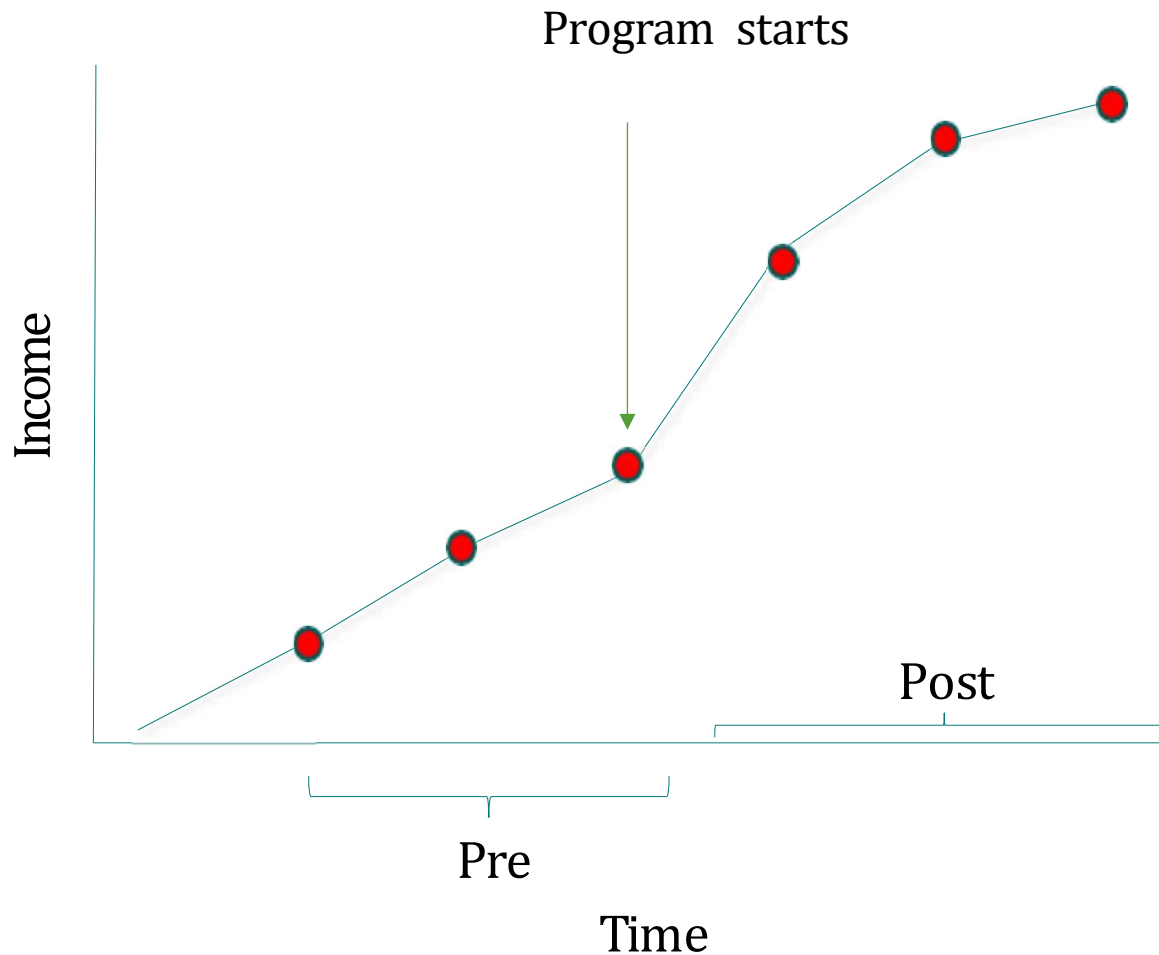
Before-After Comparison

Program: vocational training program for youth
Treatment group: group of enrolled youth



Before-After Comparison

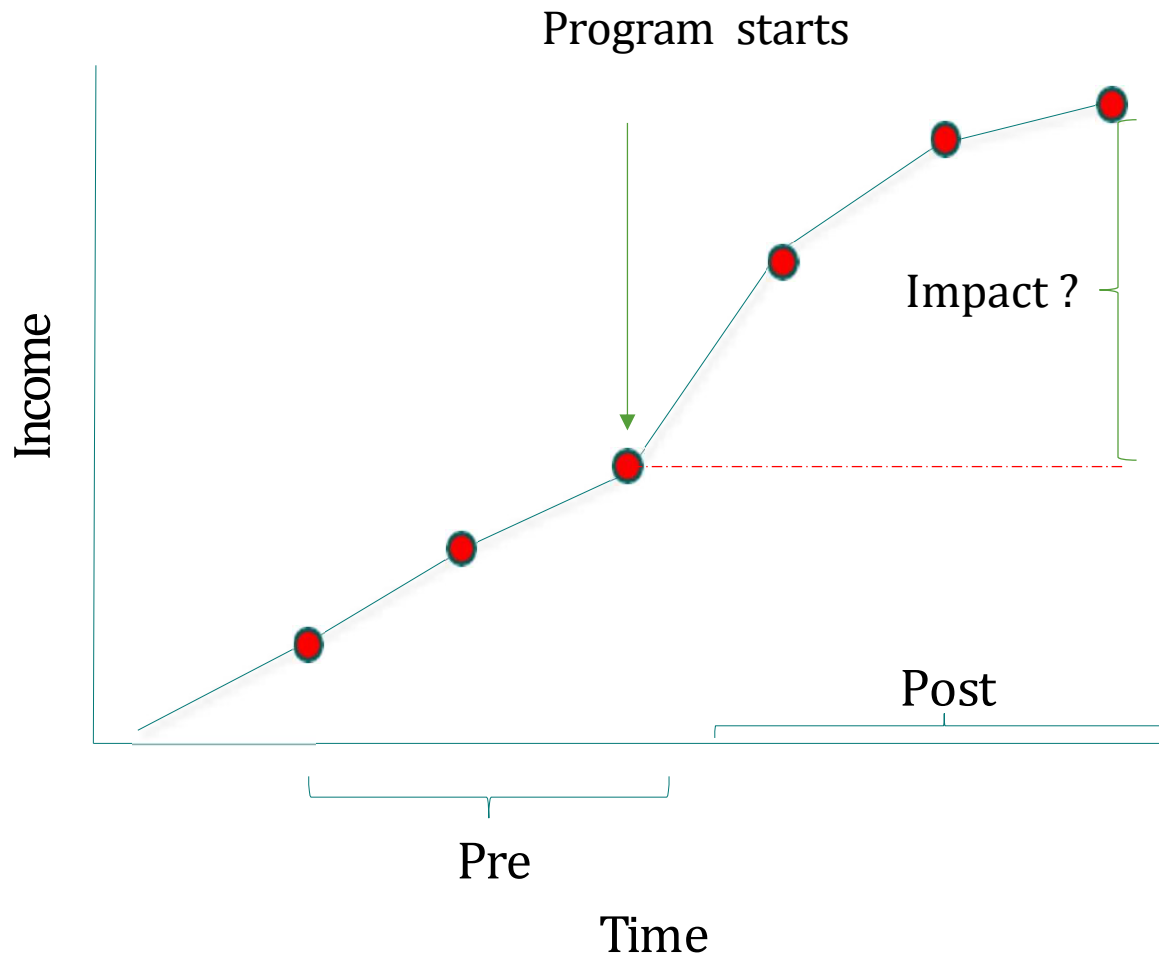
Program: vocational training program for youth
Treatment group: group of enrolled youth



Counterfactual:
What would have
happened in the
absence of the
program?

Before-After Comparison

Program: vocational training program for youth
Treatment group: group of enrolled youth



Counterfactual:
What would have
happened in the
absence of the
program?

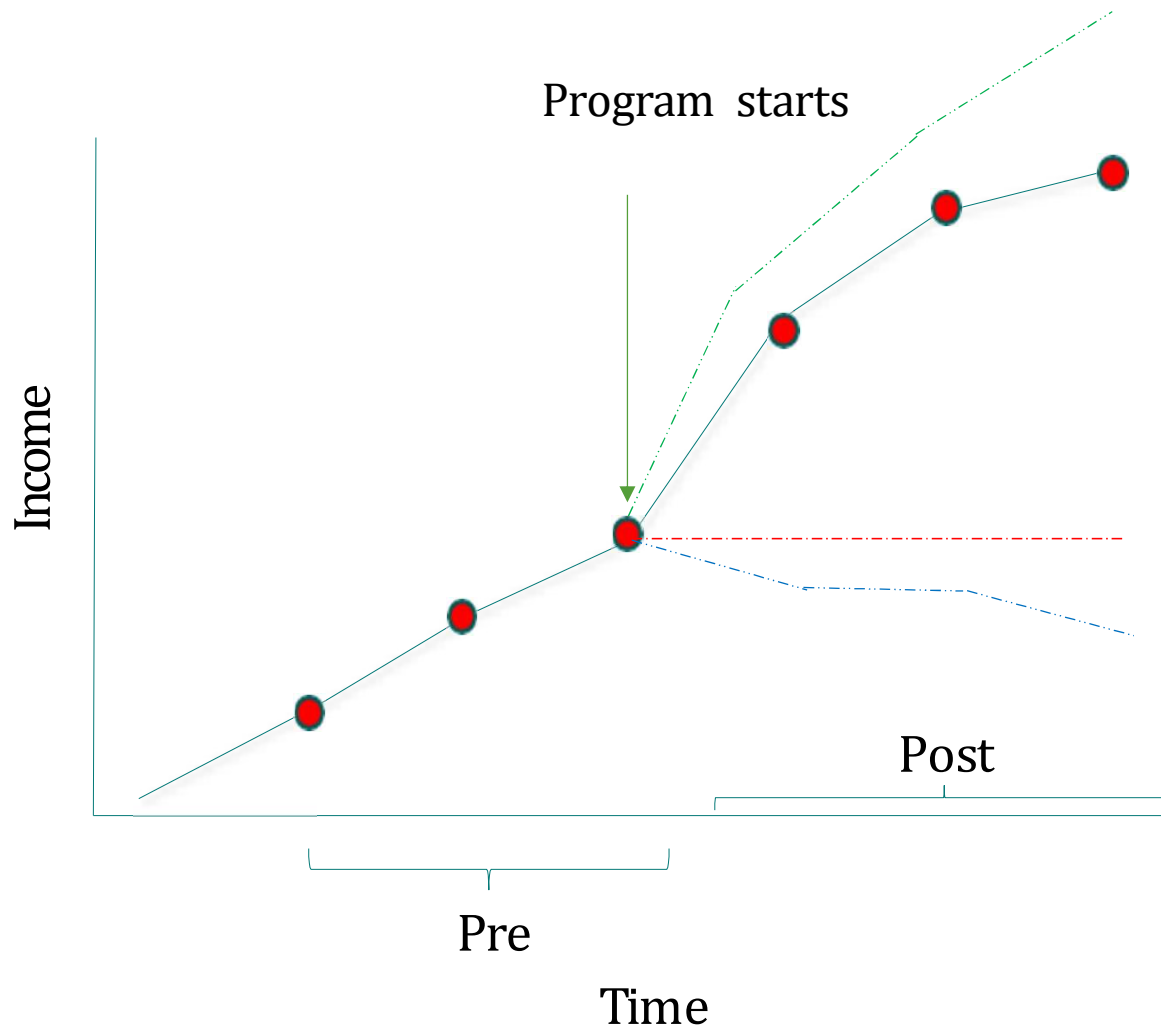




Before-After Comparison

- It is likely that – because of **external** factors which affect income (macro-economic, geographical, climatic, etc.) – the level of income would not have stayed the same **in the absence of** the training program
 - the baseline outcome level is likely not to be a good estimate of the counterfactual

Before-After Comparison



Counterfactual:
What would have
happened in the
absence of the
program?

Before-After Comparison

- **Assumptions:**
 - The level of the outcome of interest would not have changed without the intervention
 - There are no other factors (than the program) that affected the outcome over time
- **Difficulty:** You do not know what would have happened without the intervention, because you do not have any comparison group

The use of a comparison group

... You increase the credibility of results by measuring the counterfactual with the use of a comparison group...



Center for Evaluation
and Development

The choice of a comparison group

What “**RIGOR**” in impact evaluation boils down to is finding
the **best possible comparison group**

A good comparison group

- Has the same characteristics (on average) as the treatment group
- Is not exposed to the program
- Would react similarly to the program as the treatment group (if it were to participate)



Center for Evaluation
and Development

How impact has often been measured..

SIMPLE DIFFERENCE



Simple Difference



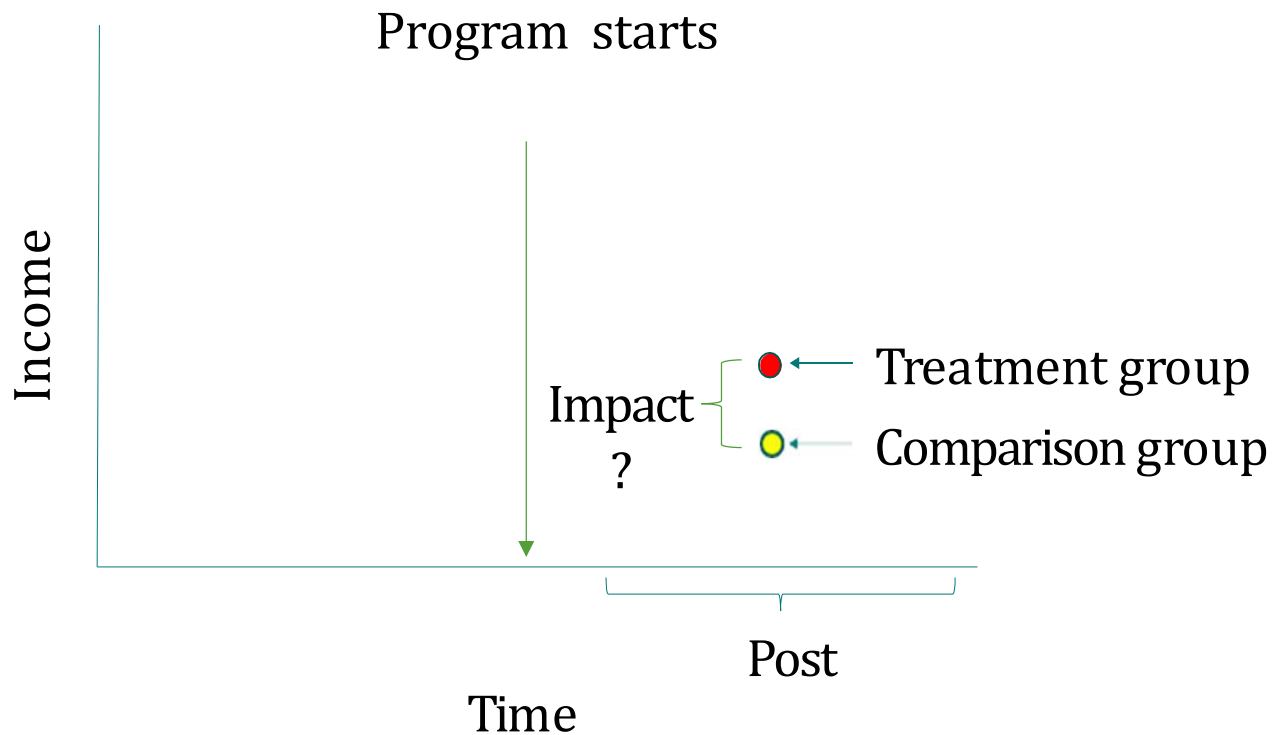
- You may compare (at least) two groups , i.e. participants vs non-participants , at one point in time
- You need data for at least two groups at one point of time

Simple Difference

Program: vocational training program for youth

Treatment group: group of enrolled youth

Comparison group : group of youth who, despite being eligible, chose not to enroll

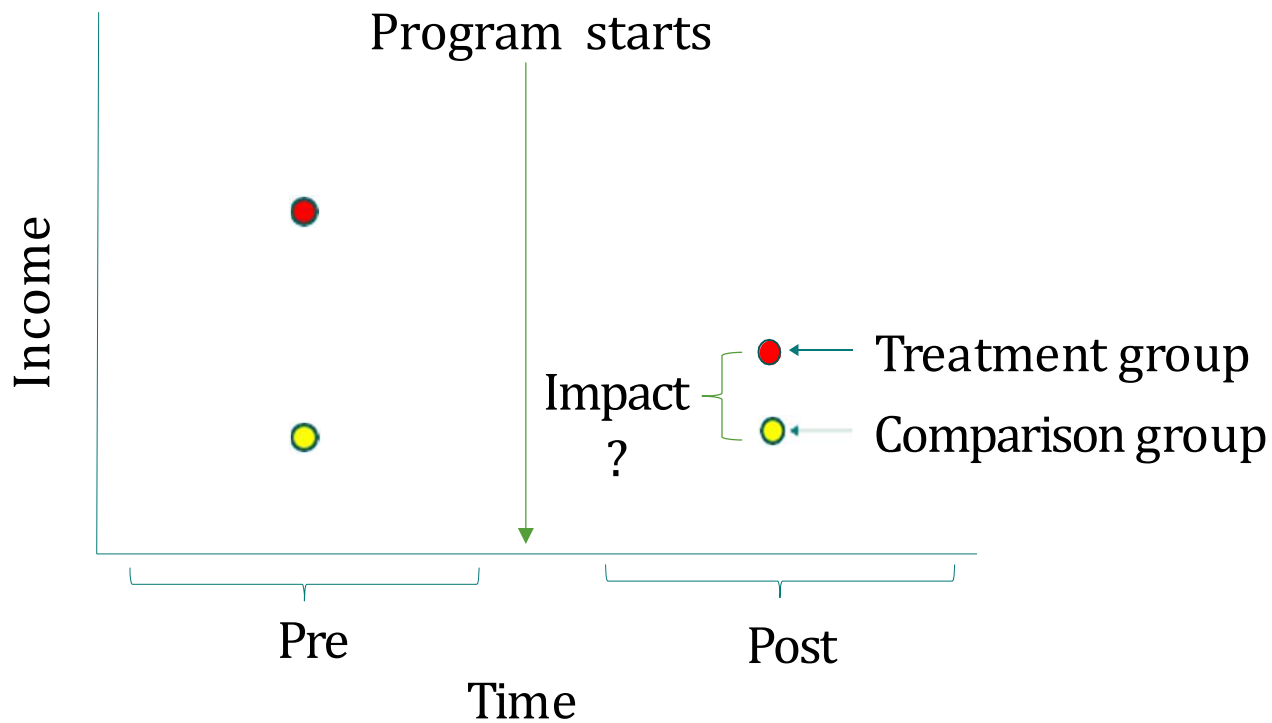


Simple Difference

Program: vocational training program for youth

Treatment group: group of enrolled youth

Comparison group : group of youth who, despite being eligible chose not to enroll



Simple Difference

- The comparison group from the previous example does not provide a good estimate of the counterfactual
- If you observe a difference in income post-training between the two groups, you would not be able to disentangle whether it is due to the training or underlying differences in motivation, skills and other factors that exists between the two groups
- This creates bias in the estimate of impact, aka ***selection bias***

Simple Difference

- Youth who chose to participate may be highly motivated and expect a higher return to training
- Youth who chose not to enroll may be discouraged and not expect benefits from the training

→ It is likely that these two groups would have behaved quite differently in the labour market and would have earned different levels of income **even in the absence of** the vocational training program

Quasi-experimental methods

1. Difference-in-difference

2. Matching

Quasi-experimental
methods

**3. Instrumental
variables (IV)**

**4. Regression
discontinuity
design (RDD)**

Difference-In-Difference

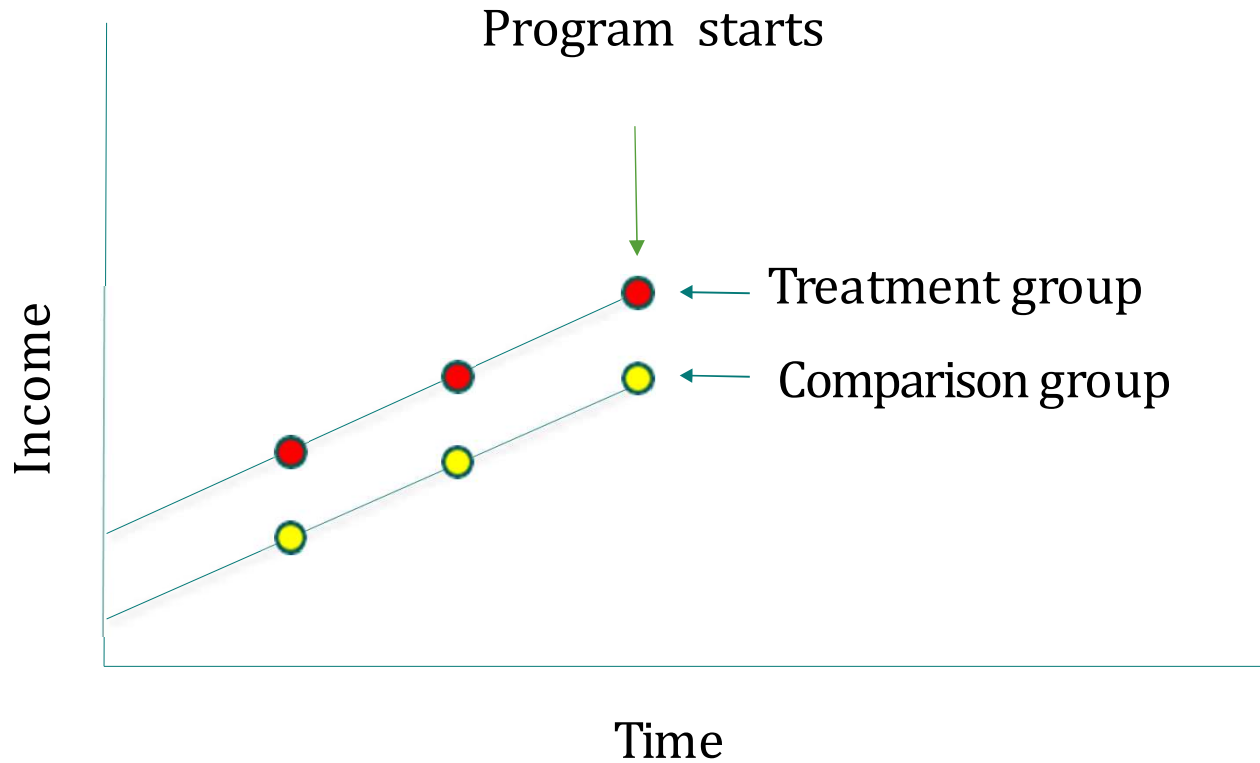
- Program: vocational training program for youth
 - Some district boards decide to adhere to the program, other do not
 - Simply comparing income for youth between enrolled and non-enrolled districts will be problematic (selection bias)
 - Idea: What if we combine the two methods discussed before ,
 - Difference before - after can remove bias from external factors that are **constant** over time
 - Difference between groups – controls for different baseline conditions
- hence, Difference – in – Difference

Difference-In-Difference

- Compare (at least) two groups, i.e. participants vs non-participants, over time
- **Assumptions:**
 - Two groups that are comparable in the outcome of interest and have the *parallel trends*
 - Same growth trends observed before program starts are used as suggestive evidence that this may hold in the absence of the program

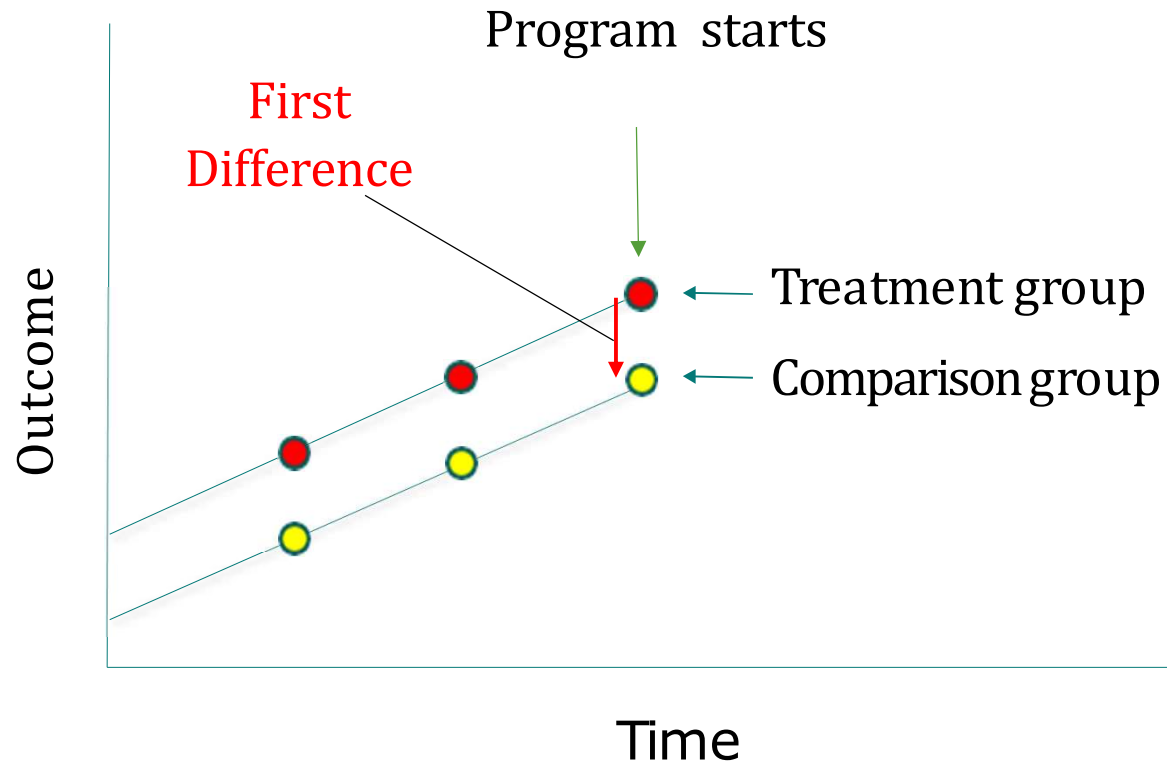


Difference-In-Difference



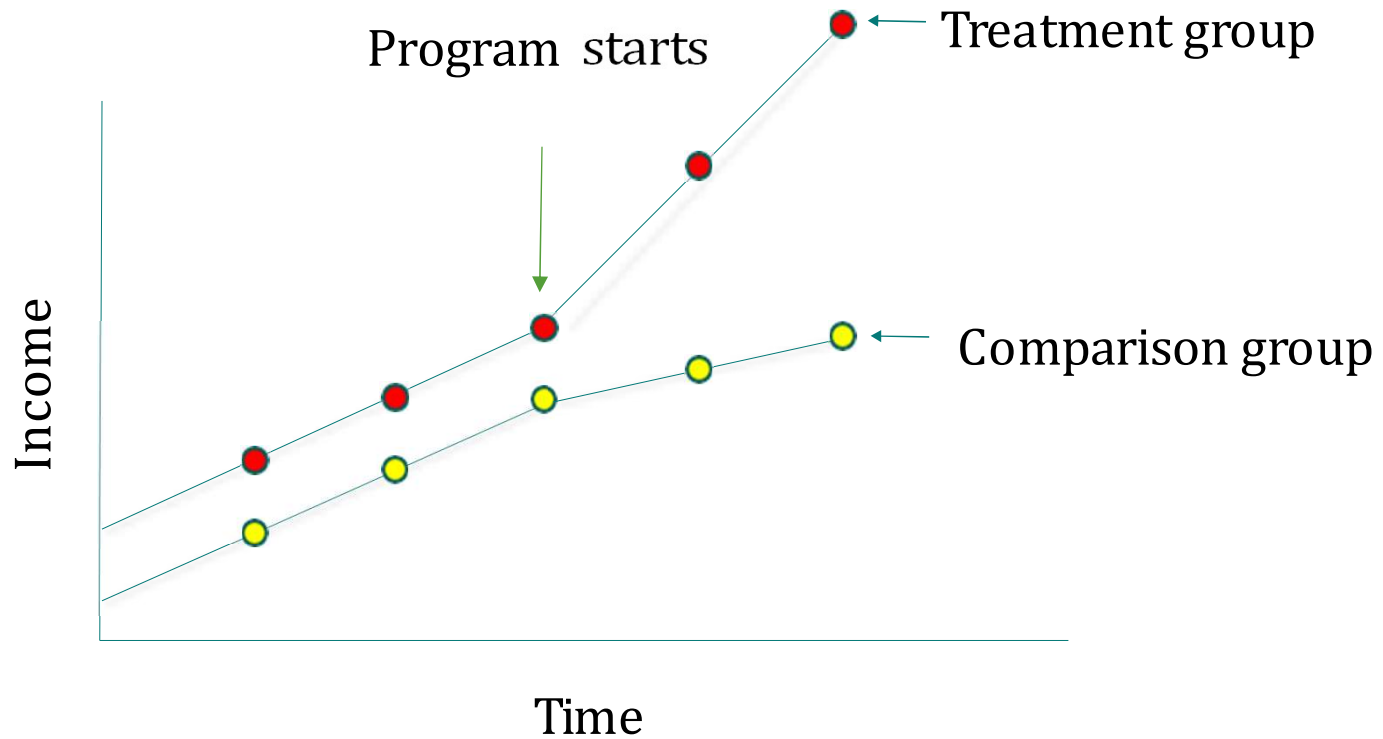


Difference-In-Difference



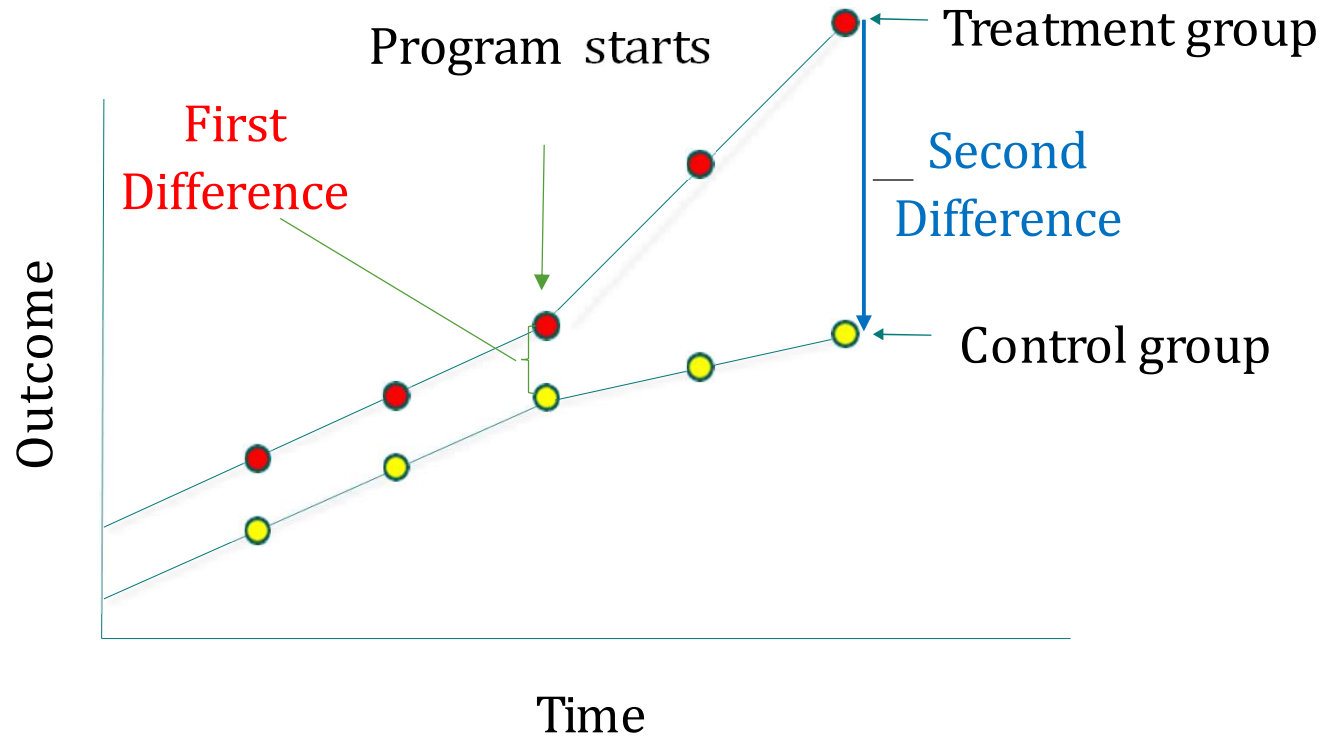


Difference-In-Difference



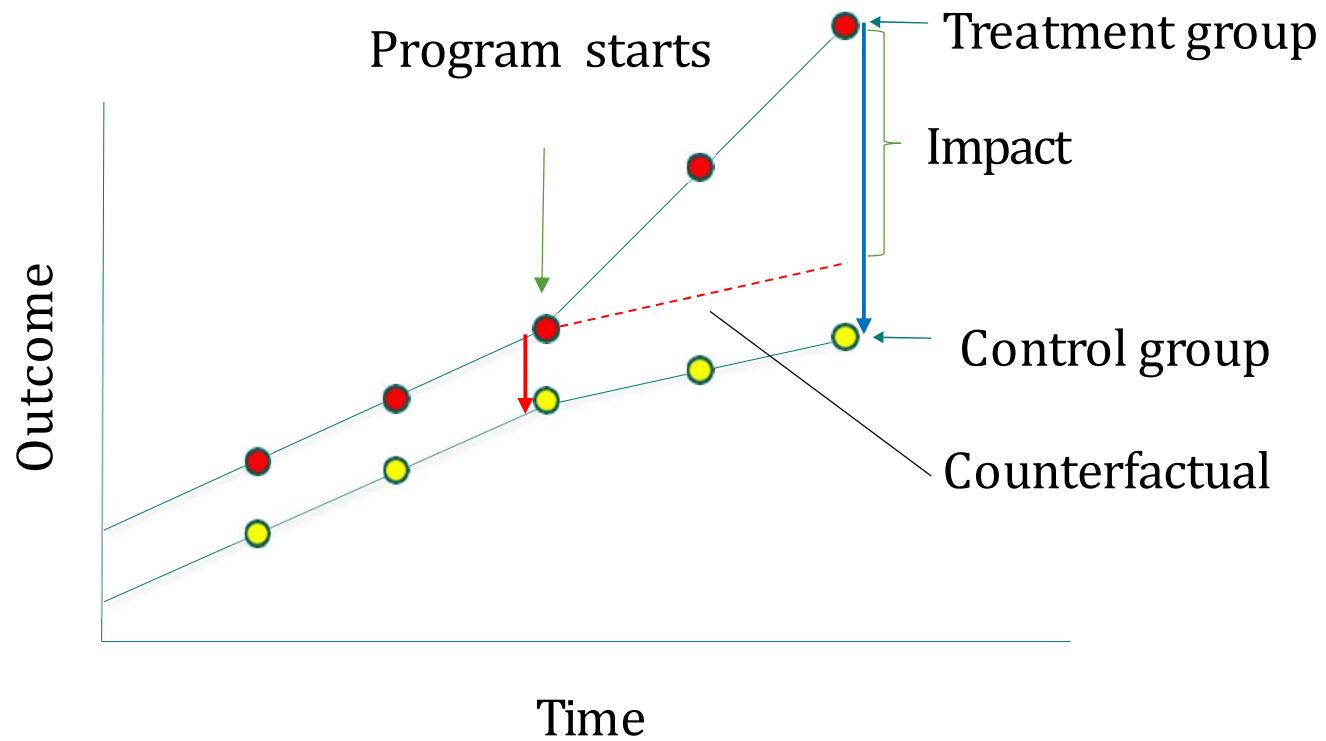


Difference-In-Difference



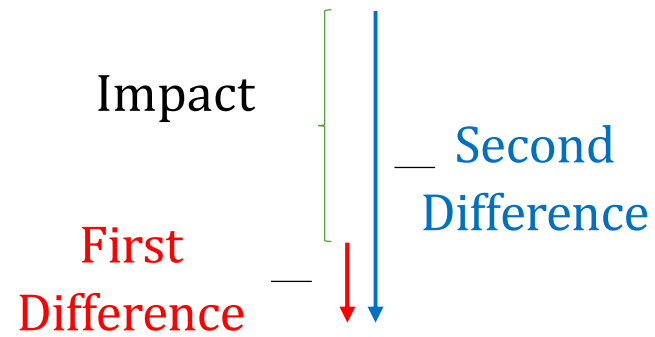


Difference-In-Difference



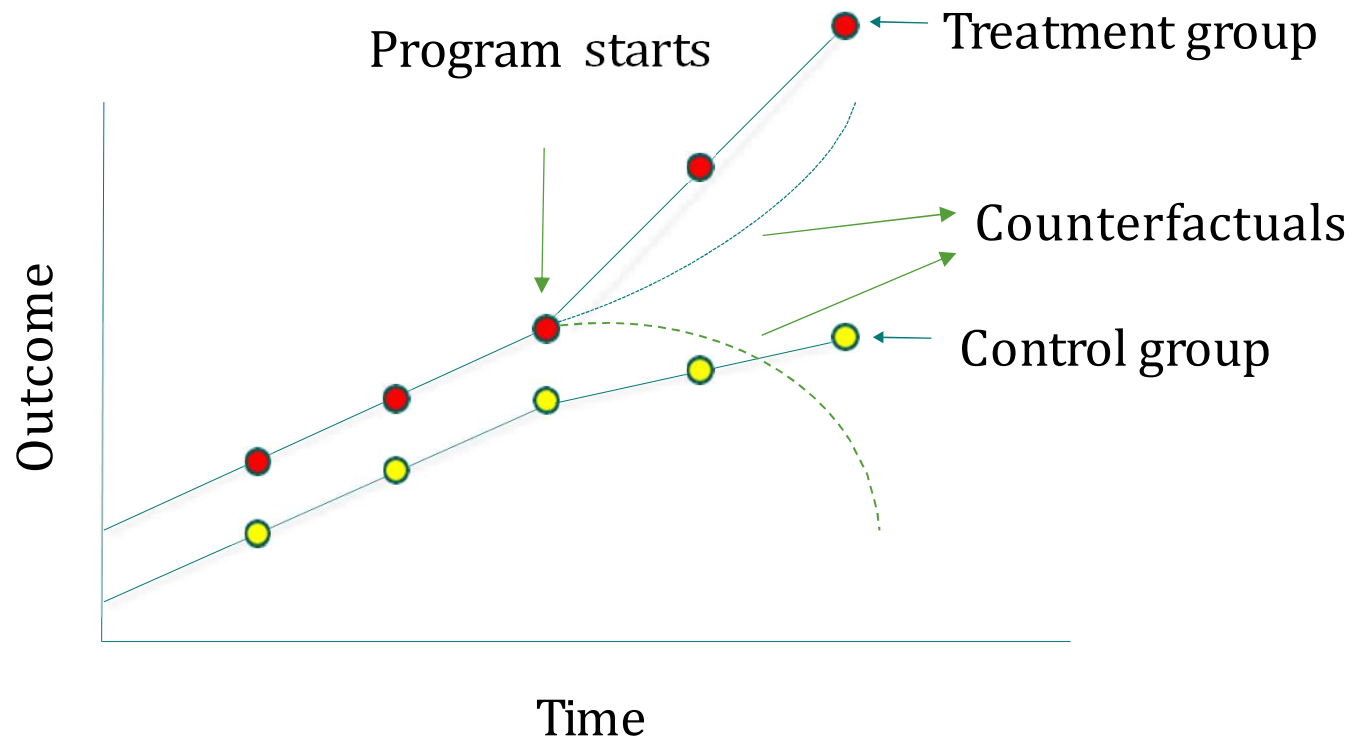


Difference-In-Difference





Difference-In-Difference



Case study: DiD

Evaluation of PROGRESA (Schultz, 2004)

- The first conditional cash transfer program
- (3-year) Educational grant to eligibly-poor mothers of a child enrolled in school: conditional on child's attendance of 85% of school days
- Schultz applies DiD to compare school enrollment of children between PROGRESA localities (Treatment) and non-PROGRESA localities (Controls)

DiD: Case Study

Outcome= Cumulative years of enrolment after first grade	Before Program	After Program	Difference
Treatment group (Progresa localities)	6.80	6.95	0.15
Comparison group (Non-progresa localities)	6.66	6.14	-0.52
Difference	0.14	0.81	0.67

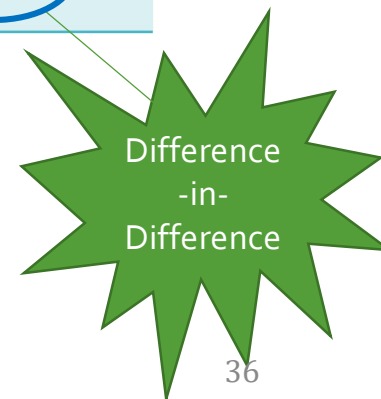


Source: Results from Schultz (2004) – Table 7


DiD: Case Study


Outcome= Cumulative years of enrolment after first grade	Before Program	After Program	Difference
Treatment group (Progresa localities)	6.80	6.95	0.15
Comparison group (Non-progresa localities)	6.66	6.14	-0.52
Difference	0.14	0.81	0.67

Simple cross-comparison



DiD: Case Study

Outcome= Cumulative years of enrolment after first grade	Before-after comparison 		
	Before Program	After Program	Difference
Treatment group (Progresa localities)	6.80	6.95	0.15
Comparison group (Non-progresa localities)	6.66	6.14	-0.52
Difference	0.14	0.81	0.67

Simple cross-comparison 

Difference
-in-
Difference

Difference-In-Difference

- **Assumptions:**

- Parallel growth trends → time-invariant (un-) observed differences
- No time-varying differences e.g. due to shocks

- **Difficulty:**

- Proving that the treatment and comparison group would have followed the same growth trend in absence of the program
- Necessary to argue that these assumptions hold. This can be very difficult to do without actual evidence.
- Is the comparison group really comparable? Across all relevant characteristics?

Quasi-experimental methods

1. Difference-in-difference

2. Matching

Quasi-experimental
methods

3. Instrumental
variables (IV)

4. Regression
discontinuity
design (RDD)



Matching

- Program: vocational training program for youth
 - Treatment group: group of enrolled youth
 - Comparison group : **group of non-enrolled youth**
-
- Idea of matching :
 - For each individual in T group, match an individual from the comparison group
 - Finding a good match for each program participant requires approximating as good as possible the characteristics that explain youth's enrollment in the program
 - Estimate impact as the difference in outcome between matched T and C group



Matching

- Program: vocational training program for youth
- Treatment group: group of enrolled youth
- Comparison group : **group of non-enrolled youth**

Exact matching

Treated units				Untreated units			
Age	Gender	Months unemployed	Secondary diploma	Age	Gender	Months unemployed	Secondary diploma
19	1	3	0	24	1	8	1
35	1	12	1	38	0	1	0
41	0	17	1	58	1	7	1
23	1	6	0	21	0	2	1
55	0	21	1	34	1	20	0
27	0	4	1	41	0	17	1
24	1	8	1	46	0	9	0
46	0	3	0	41	0	11	1
33	0	12	1	19	1	3	0
40	1	2	0	27	0	4	0

Propensity Score Matching can be used if exact matching is not feasible

Propensity Score Matching (PSM)

- Idea of Propensity Score matching:
 - Use a set of **observed characteristics to** estimate the **probability of program participation** of an individual (aka «**propensity score**»)
 - Then match treatment and comparison group with similar propensity score values

Propensity Score Matching (PSM)

- **Assumptions:**

- Deep understanding of the observable covariates that drive participation in an intervention
- Matching only on observable characteristics (assumes no unobserved differences!)
- Common support (requires substantial overlap between the propensity scores program participants and non-participants)

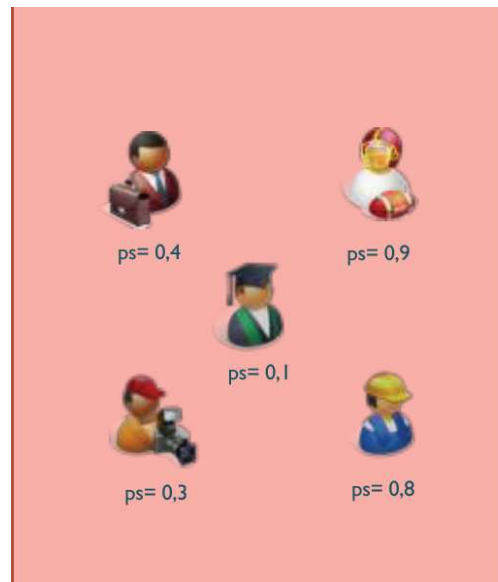
- **Difficulty:**

- Large(r) comparison group sample needed
- Lack of “common support” can affect external validity

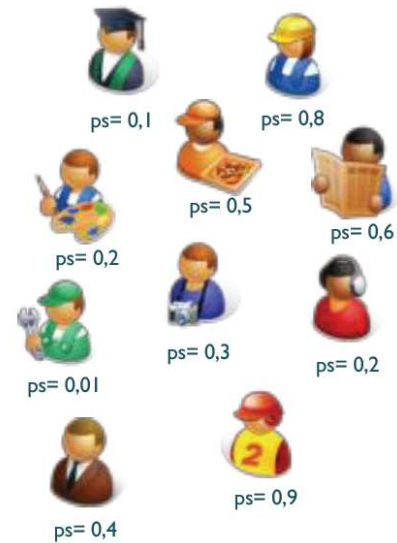
Propensity Score Matching (PSM)

- Many matching algorithms are available
- 1-1 matching example:

Participants



Others



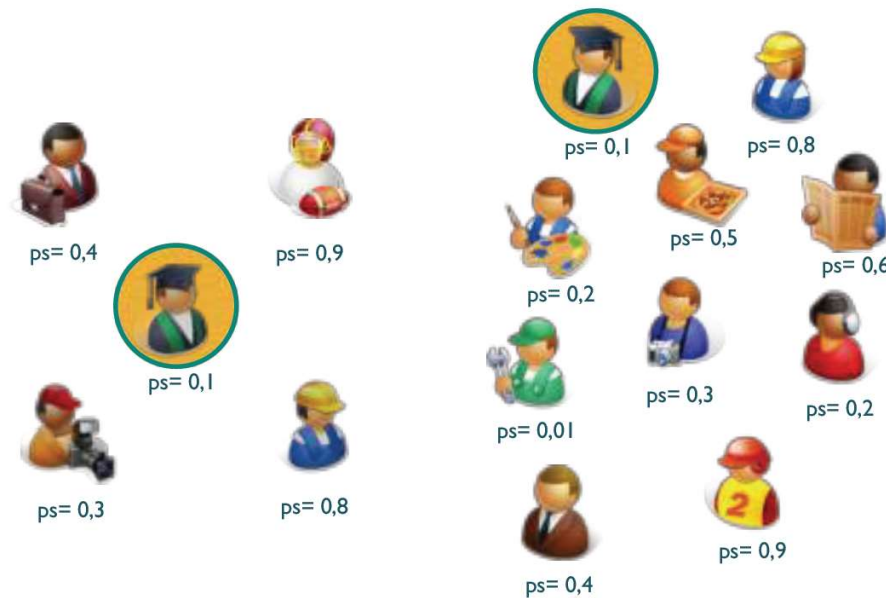
Source: Image adjusted from Trycinski (2011)

Propensity Score Matching (PSM)

- Many matching algorithms are available
- 1-1 matching example:

Participants

Others



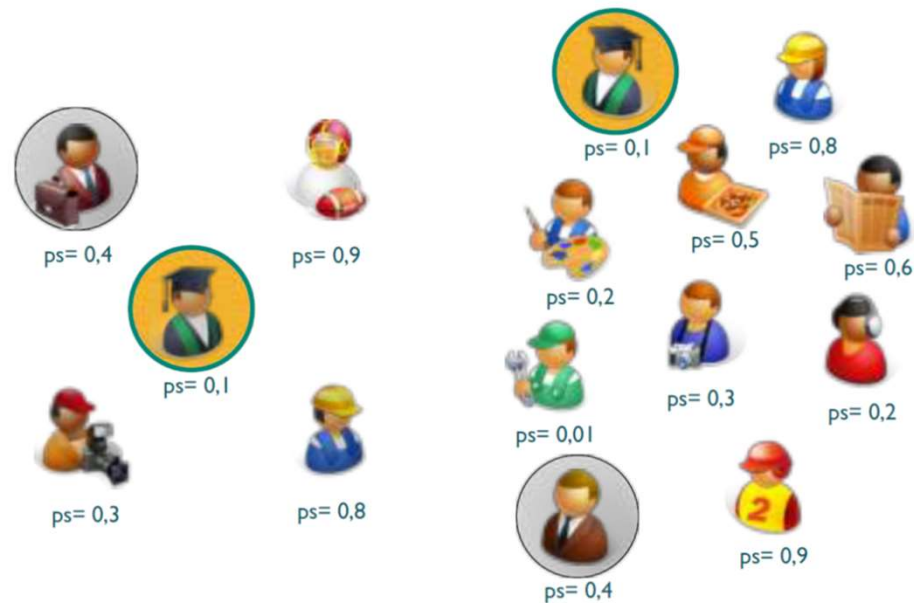
Source: Image adjusted from Trycinski (2011)

Propensity Score Matching (PSM)

- Many matching algorithms are available
- 1-1 matching example:

Participants

Others

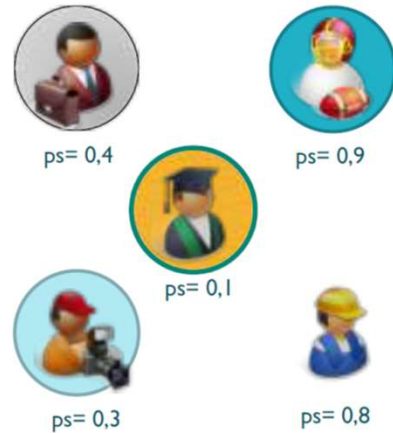


Source: Image adjusted from Trycinski (2011)

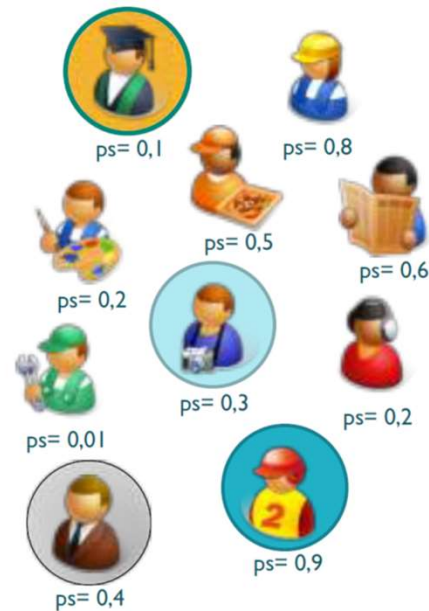
Propensity Score Matching (PSM)

- Many matching algorithms are available
- 1-1 matching example:

Participants

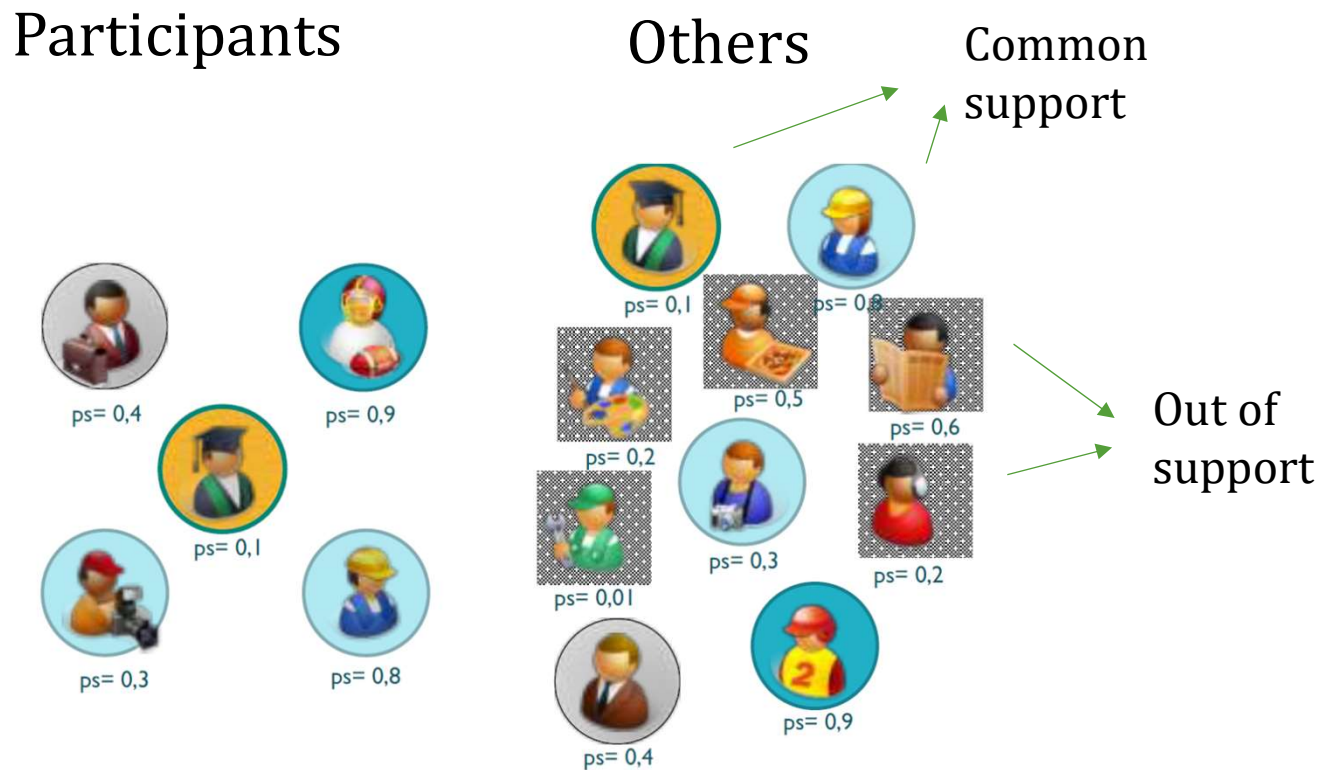


Others



Propensity Score Matching (PSM)

- Many matching algorithms are available
- 1-1 matching example:

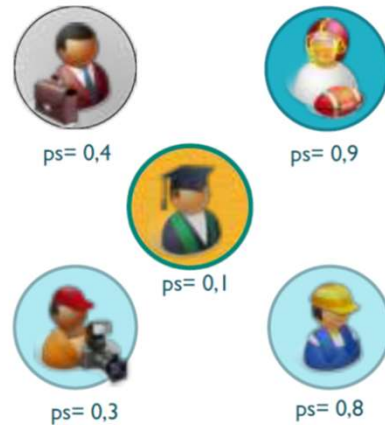


Source: Image adjusted from Trycinski (2011)

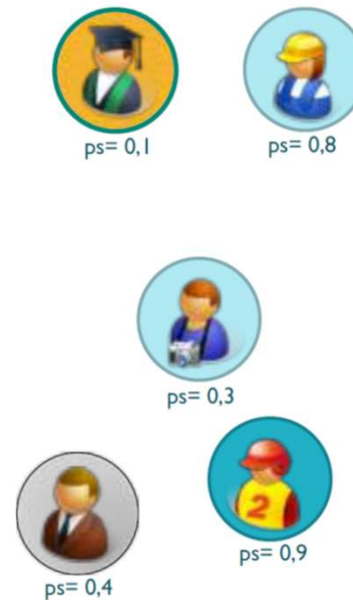
Propensity Score Matching (PSM)

- Many matching algorithms are available
- 1-1 matching example:

Participants



Others



Case study: PSM

Evaluation of the Integrated Food Security Program (IFSP) in Ethiopia (Abebaw, Fentie and Kassa 2010)

- **Program:** environmental rehabilitation, promotion of agriculture and livestock, infrastructure construction and maintenance in Ahmara region
- **Problem:** Non-random program placement based on vulnerability criteria of kebele (village) and households
 - Treatment households are significantly different from non-treatment households, in the same kebele and across kebele

Propensity Score Estimation

Variables	Specification (3) Coefficient (std. error)
SEX	0.239 (0.606)
PAGE	0.015 (0.0175)
PEDU	0.201 (0.368)
PFAMNO	0.053 (0.115)
PLAND	-0.261 [*] (0.147)
PTLU	-0.069 (0.057)
PHOUSTAT	1.688 ^{***} (0.539)
PHHDURAB	-0.015 ^{***} (0.004)
PMKTDIST	-0.084 ^{**} (0.039)
PEXTDIST	-0.147 [*] (0.075)
PFAMNO squared	
PAGE squared	
PLAND squared	
PTLU squared	
PHHDURAB squared	
PMKTDIST squared	
PEXTDIST squared	
Constant	1.735 [*] (0.967)
Sample size (N)	200
Pseudo-R ²	0.19
LR χ^2 value	34.78 ^{***}
Log-likelihood	-111.82

***, ** and * stand for significance at the 1%, 5% and 10% levels, respectively.

Source: Abebaw, Fentie and Kassa (2010)

- Dependent variable =1 if HH participates in the program, 0 otherwise
- Matching variables measured pre-program :
 - Gender of household head
 - Age of household head
 - Education of household head
 - Number of household members
 - Land holdings
 - Livestock holdings
 - Housing type
 - Durable goods value
 - Distance to market
 - Distance from development agent's office
- Note: matching on pre-program OR time-invariant variables is highly recommended, as it avoids to bring in bias (behaviour is affected by program attendance!) ⁵¹

Common support assessment

Before matching



Source: Abebaw, Fentie and Kassa (2010)

- Scores between 0.02 and 0.95 in the comparison (Non-IFSP) group
 - Scores between 0.17 and 0.96 in the treatment (IFSP) group
- exclusion of 23 households

Case study: PSM

Impact estimates

	Specification (3)
Household food calorie intake	694.96*** (4.77)
Balancing property satisfied	Yes
Common support imposed	Yes
Number of observations	
IFSP households	99
Non-IFSP households	79

*** stands for significance at 1% level. In parentheses are bootstrapped standard errors (50 replications). Bandwidth=0.25.

Source: Abebaw, Fentie and Kassa (2010)

- Many matching algorithms are available
 - Kernel algorithm: match each treated individual to all comparison ones within a bandwidth, weighted by closeness of pscore value
- Effect is the difference in outcome means between matched units

Result: + 696 kilo calories per day per adult equivalent unit, about 30% more than in comparison group

Quiz?

Quiz 1

Go to
www.menti.com

Enter the code
9156 9718



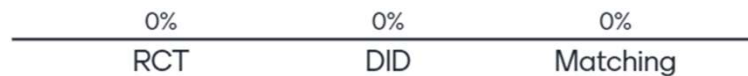
Or use QR code

Quiz?


Go to www.menti.com and use the code 9156 9718

Which method among the following only works when baseline outcome data is available?

 Mentimeter



Press ENTER to show correct

 Voting is closed

 Results are hidden

18


Quasi-experimental methods

1. Difference-in-difference

2. Matching

Quasi-experimental
methods

3. Instrumental
variables (IV)

4. Regression
discontinuity
design (RDD)

Instrumental Variables (IV)

- IV does not create a comparison group but uses a **regression framework** to estimate the impact of an intervention (from either cross-sectional or panel data)
- IV counteracts selection bias, especially how **unobservable characteristics can bias impact estimates**
- If such unobservable characteristics are correlated with the outcome and program participation, estimates of program impact will be **biased**

Instrumental Variables (IV)

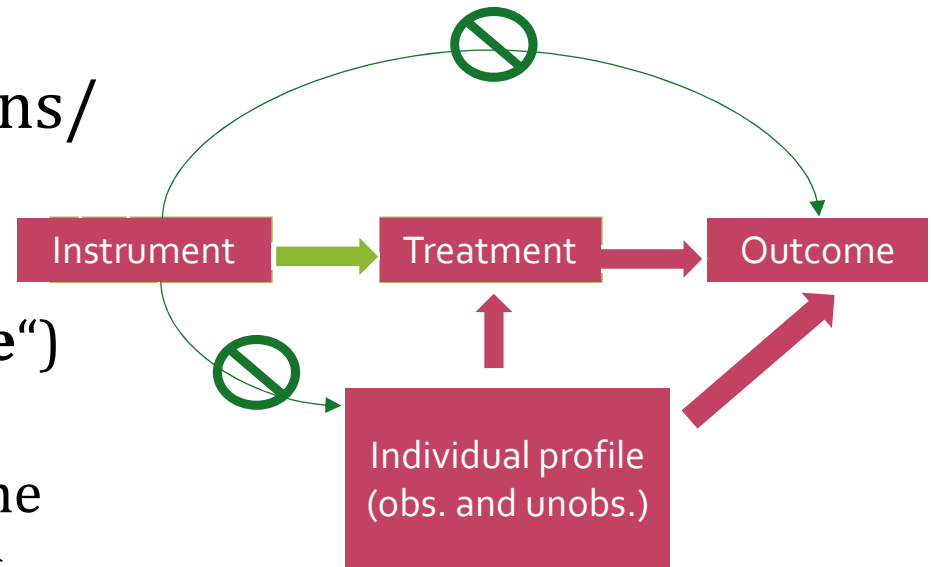
- This approach uses an **additional variable (the IV)** that is **highly correlated with program participation**, but is **not correlated with unobservable characteristics** affecting outcomes
- It uses this additional IV variable to **'clean'** the treatment variable by separating out and discarding the part of the treatment that is affected by **unobservable characteristics**

Instrumental Variable (IV) Estimations

The following assumptions/
conditions have to be met:

1) The instrument affects the
likelihood of treatment („**Relevance**“)

2) The instrument does not affect the
outcome through any other channel
than the treatment (“**Exclusion**”)



→ Then we are able to isolate the impact on outcomes from
the bias caused by the influence of unobservables

Case study: IV

Evaluation of a business training on microenterprise profits in Tanzania (Berge et al., 2011)

- Target: microenterprise clients of microcredit loan groups
- Program: entrepreneurship training sessions + business grants
- Design: random assignment of microcredit loan groups to business training sessions.
- Actual program participation:
 - Training attendance rate: 70% → partial compliance



Selection bias!

Case study: IV

- Random assignment can be used as instrument for actual program participation
 - ✓ Predicts training attendance
 - ✓ Is assumed to affect profits only via training attendance and to be uncorrelated with unobserved characteristics (e.g. lack of motivation) driving non-compliance
- IV estimates impact for **compliers**, i.e. *those microcredit clients that were randomly assigned to trainings and effectively attended them*

Case study: IV

Impact estimates on profit margin and sales

	(1) Profit Margin ITT	(2) Profit Margin ATET	(3) Sales ITT	(4) Sales ATET
Training	-0.014 (0.028)	-0.015 (0.031)	0.257** (0.123)	0.288** (0.137)
Grant	-0.004 (0.016)	-0.004 (0.015)	0.038 (0.073)	0.036 (0.072)
Training*Female	0.003 (0.033)	0.003 (0.037)	-0.262* (0.157)	-0.295* (0.177)
Female	-0.013 (0.024)	-0.013 (0.024)	0.044 (0.110)	0.044 (0.109)
Sum Female	-0.010 (0.018)	-0.012 (0.021)	-0.006 (0.089)	-0.007 (0.103)
Observations	494	494	494	494

Note: The table reports regressions of profit margin and sales on treatment status, all regressions controlling for gender and covariates. Covariates include age, gender, education, number of businesses, PRIDE branch, PRIDE loan size, marketing index, religion, and the lagged dependent variable. Cluster-robust standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Source: Berge et al. (2011) - Table 4B

Impact of treatment assignment
(average impact on compliers and
non-compliers) aka “Intent To Treat”

Impact of actual treatment on
compliers (IV), aka “Average
Treatment Effect on the Treated”

Instrumental Variables (IV)

- **Main Assumptions:**

- IV is highly correlated with actual treatment (Relevance)
- IV affects the outcome of interest only via treatment (Exclusion)

- **Difficulty:**

- Finding a valid IV, especially outside experimental studies, is very difficult

Quasi-experimental methods

1. Difference-in-difference

2. Matching

Quasi-experimental
methods

3. Instrumental
variables (IV)

4. Regression
discontinuity
design (RDD)

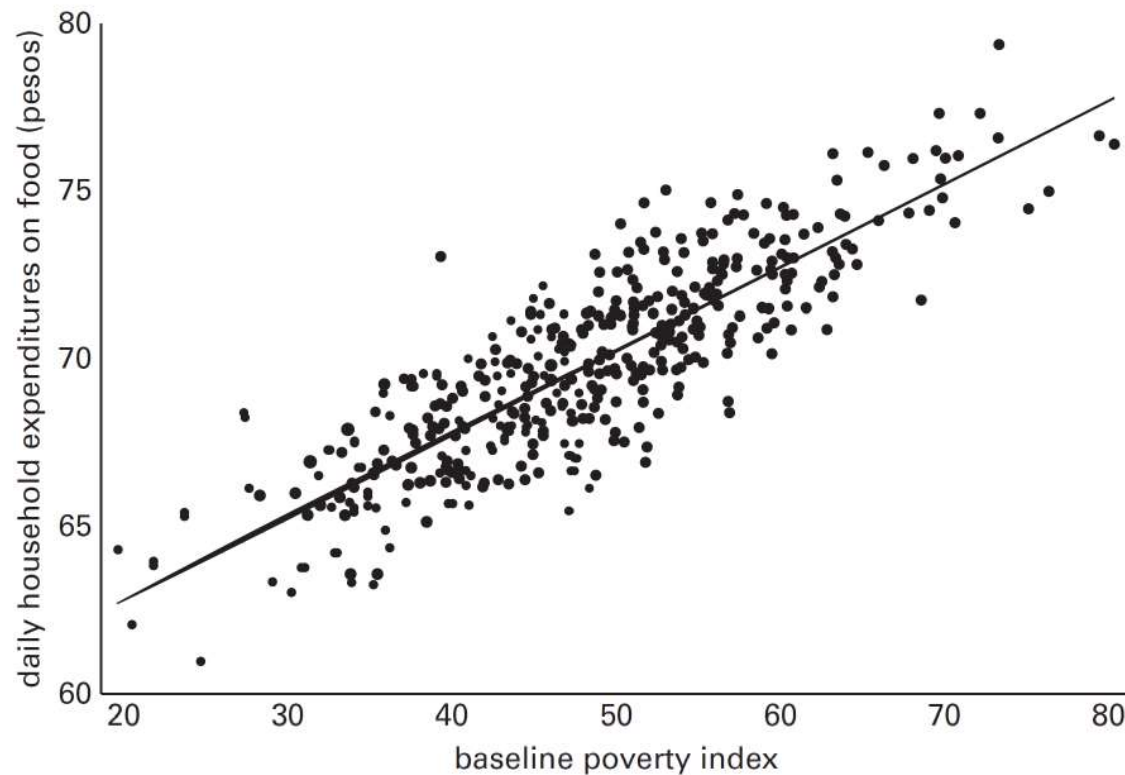
Regression Discontinuity Design

- Program participation is sometimes based on a transparent rule with a clear-cut threshold, e.g.:
 - University admission based on test score > 80 th percentile
 - Legal rules applying to enterprises with number of employees > 50
 - cash transfers to households with poverty score $< 50\%$
 - Left side of a geographical border

Regression Discontinuity Design

- Program: Household cash transfer
- Eligibility rule: households below 50% poverty score (aka *running variable*)

Household expenditure in relation to poverty (Pre-intervention)

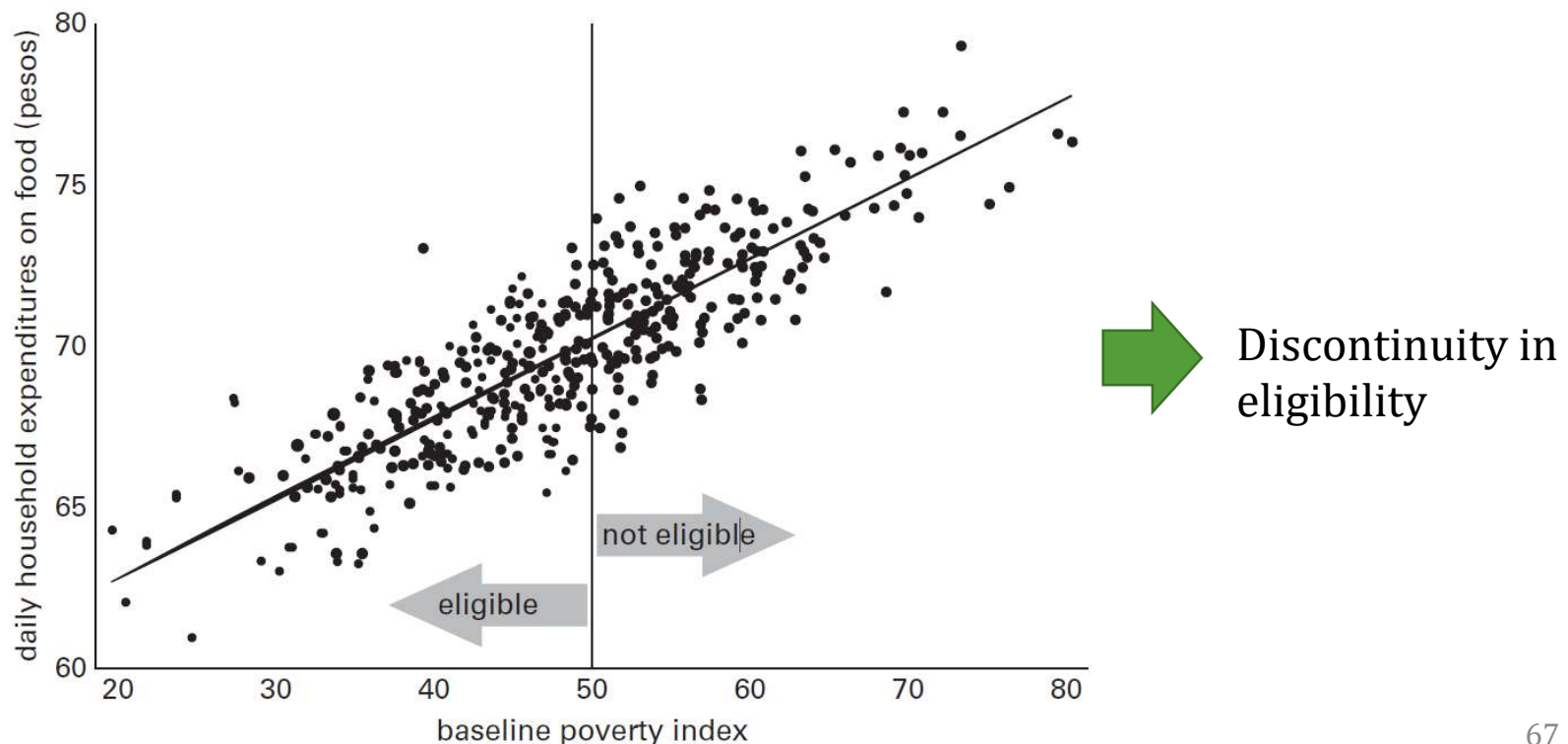


Source: Gertler et al., 2011

Regression Discontinuity Design

- Program: Household cash transfer
- Eligibility rule: households below 50% poverty score (aka *running variable*)

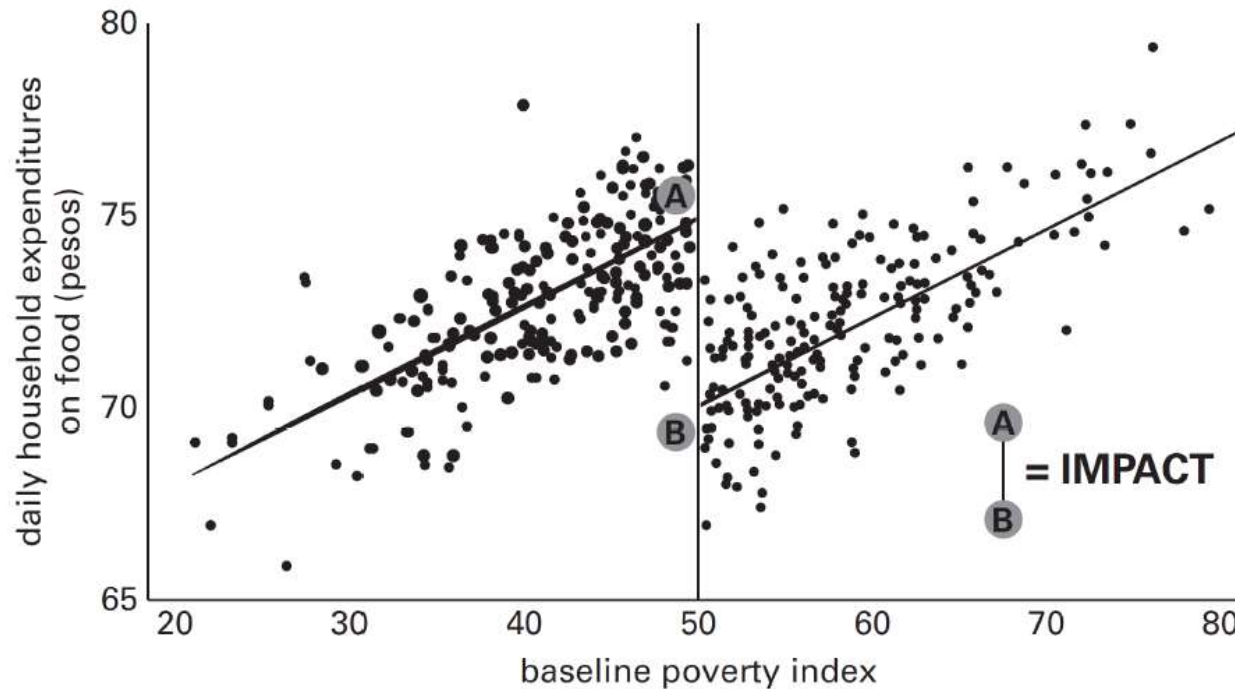
Household expenditure in relation to poverty (Pre-intervention)



Regression Discontinuity Design

- Program: Household cash transfer
- Eligibility rule: households below 50% poverty score (aka running variable)

*Household expenditure in relation to poverty (**Post-intervention**)*

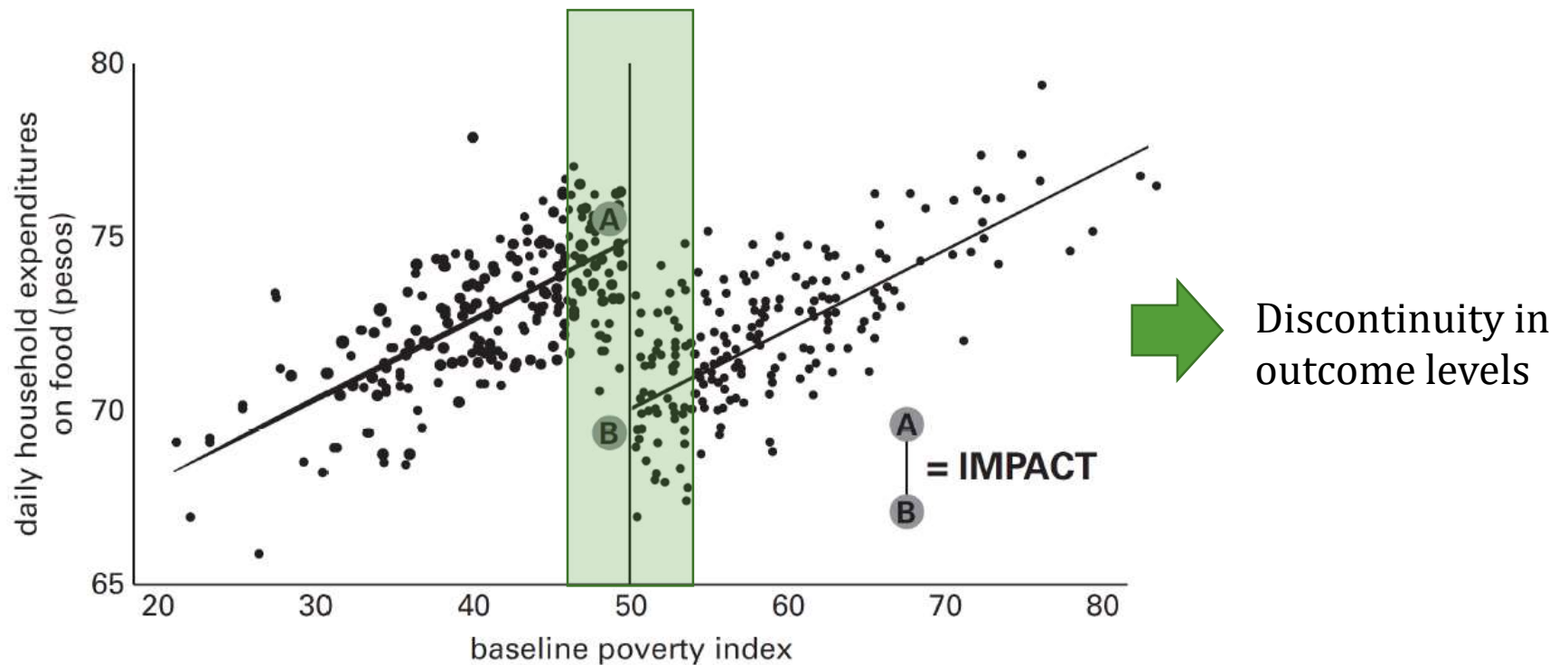


Discontinuity in
outcome levels

Regression Discontinuity Design

- Program: Household cash transfer
- Eligibility rule: households below 50% poverty score (aka running variable)

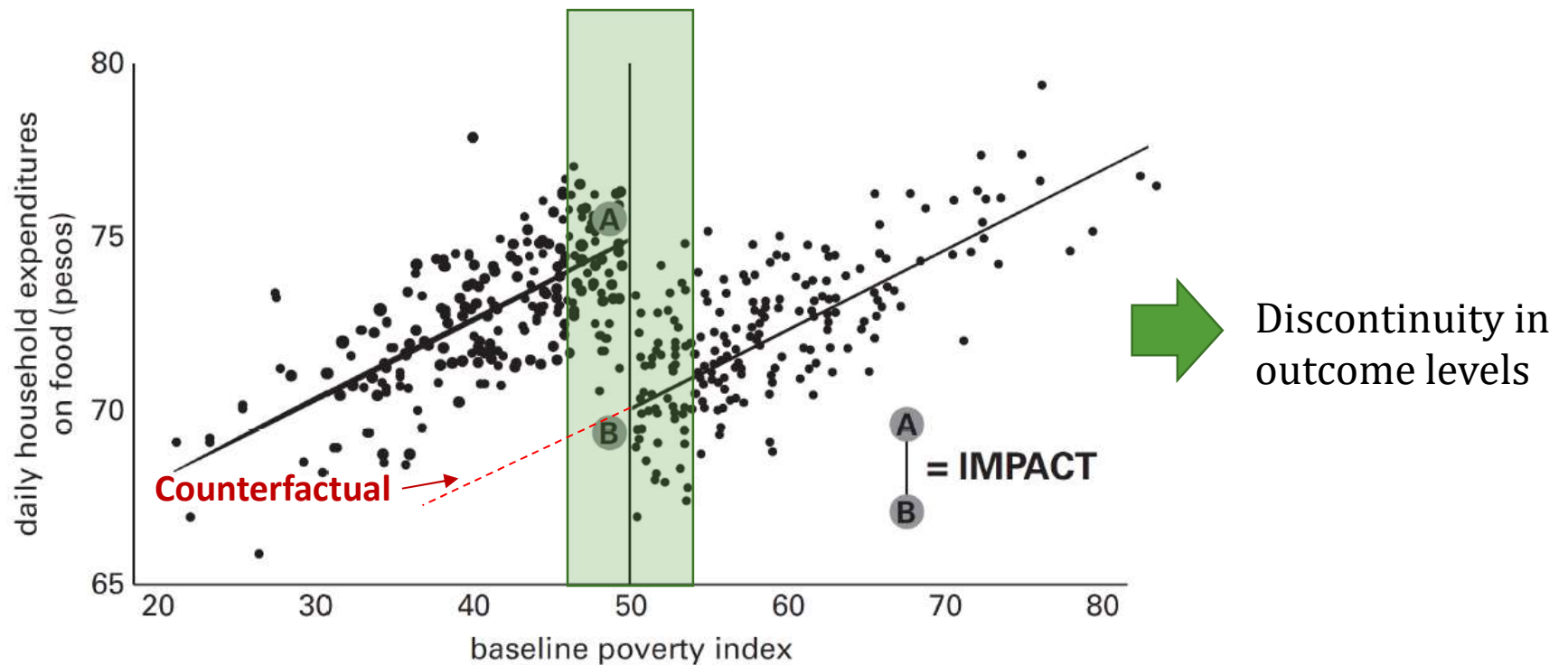
Household expenditure in relation to poverty (Post-intervention)



Regression Discontinuity Design

- Program: Household cash transfer
- Eligibility rule: households below 50% poverty score (aka running variable)

Household expenditure in relation to poverty (*Post-intervention*)



Regression Discontinuity Design

- People with a similar score/rank are comparable in observable and unobservable characteristics
- Measures the impact as difference in outcome levels for the individuals around the cut-off, e.g.:
 - just above the cut-off (comparison group) vs just below the cut-off (treatment group)

Regression Discontinuity Design

- **Assumptions:**

- Continuous running variable
- A sufficient number of observations exist in a bandwidth around the cut-off
- Potential participants are not able to precisely manipulate their score
 - Graphical analysis of individuals distribution based on running variable
- No other factors/programs should generate discontinuity around the cut-off for eligibility

Regression Discontinuity Design

- **Difficulties:**

- Results can be sensitive to the choice of the bandwidth around the cut-off
 - Estimate using different bandwidths
- Results only valid for units around the cut-off: **local** impact estimate.
 - We don't know if the program maybe improved the income of the very poor much more
- In many cases, the compliance with program participation may be partial
 - Use of IV!

Case study: RDD

Evaluation of a development versus an humanitarian model of refugee assistance (MacPherson and Sterck, 2019)

Set- up:

- Kenya – large influx of South Sudanese refugees
- Two camps in Turkana county: Kakuma (1991) and Kalobeyei (2016)
- The opening of Kalobeyei camp in 2016 creates a discontinuity in UNHCR camp assignment rule
 - Compare refugees who come **before and after a cut-off date**
- Sample: Households registered between February 2015 to August 2017

Case study: RDD

Kakuma (comparison)

vs

Kalobeyi (treatment)

Humanitarian model based on
care and maintenance via:

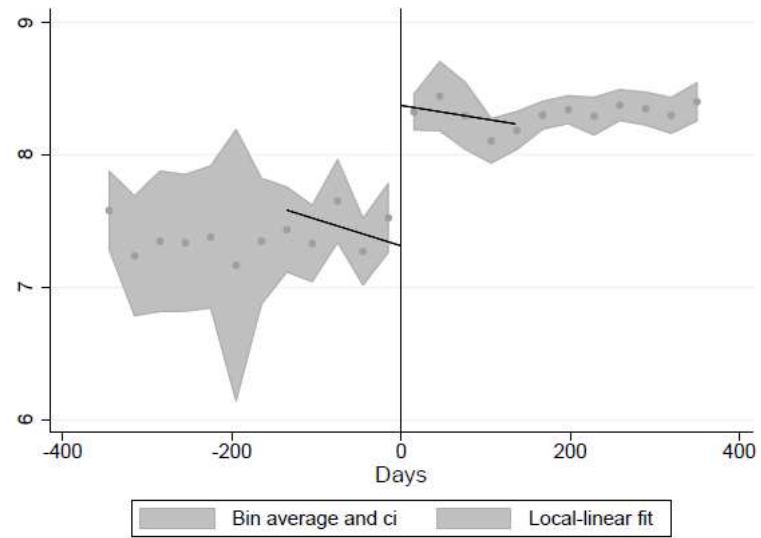
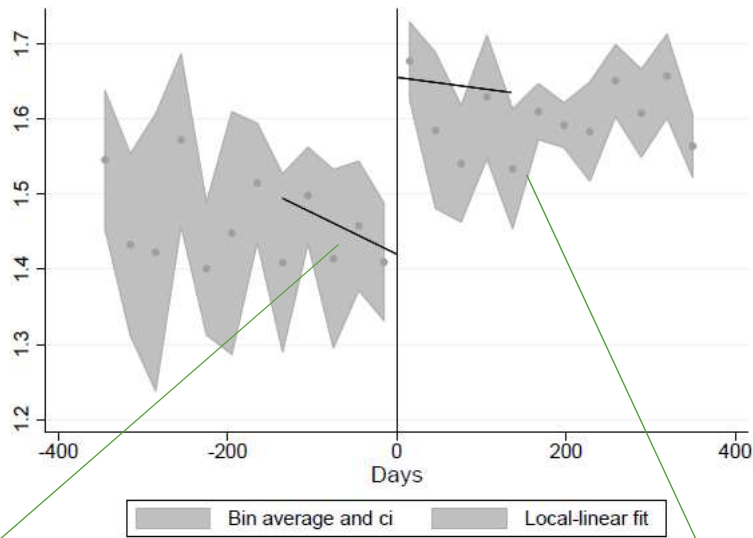
- in-kind food transfers

Development model forstering self-
reliance via:

- cash assistance
- kitchen gardens

Case study: RDD

The Kalobeyei effect



Comparison group
(Kakuma)

(a) Dietary variety (log)

(b) Daily calories per adult equivalent (log)

Source: MacPherson and Sterck (2019) - excerpt from figure 3.b

Treatment group
(Kalobeyei)

Recap: Quasi-experimental methods

Method	Description	Assumptions	Remarks
Difference in Difference (DiD)	Measures the before vs after outcome change for T and C group. Then subtracts the two to find the change in outcome for T as compared to C group. Accounts for constant differences between T and C group over time.	<ul style="list-style-type: none"> Parallel trends: Outcomes for T and C groups would have experienced the same growth trends in the absence of programme There are no unobserved shocks differently affecting T and C groups and affecting the outcome 	<ul style="list-style-type: none"> Needs baseline data Use of historical data pre-program to provide suggestive evidence that parallel trends assumption holds
Matching	Each participant in the T group is matched with a similar non-participant in the C group. Then outcome levels between the matched T-C individuals are compared.	<ul style="list-style-type: none"> No unobserved differences between T and C groups correlated with outcome and program participation There is sufficient common support in the probability of participation (propensity score) between T and C group 	<ul style="list-style-type: none"> Requires a deep understanding of the selection into programme participation Large C group needed Should match on characteristics unaffected by treatment (ideally baseline)
Instrumental Variables (IV)	Does not create a comparison group but uses a regression framework for impact estimation. It counteracts unobservable variable bias.	<ul style="list-style-type: none"> The IV highly predicts treatment The IV affects the outcome only through the treatment 	<ul style="list-style-type: none"> Finding a valid IV is very difficult Effects are “local”: only valid for individuals who are affected by the instrument (e.g. compliers to random assignment in RCT)
Regression Discontinuity Design (RDD)	Programme eligibility depends on some clear-cut rule based on cut-off such that individuals can be ranked. Outcome levels for individuals just below (e.g. T group) and above (e.g. C group) the cut-off are compared sometime after the program started	<ul style="list-style-type: none"> Continuous running variable Sufficient number of individuals observed close to the cut-off Participants cannot manipulate their programme eligibility No other factor creates discontinuity in outcomes apart from the programme 	<ul style="list-style-type: none"> Large sample around the cut-off needed Effects are “local”: only valid for individuals around the cut-off Use of IV in case of partial compliance

Assignment

- A recent national survey in Fantasia Land has revealed that the employment rate for female youth is 55%.
- The government has decided to engage with local no-profit organizations to implement a training program to promote employability and entrepreneurship of female youth. More vulnerable districts will be targeted in a first phase (2022-2024), where vulnerability is defined by a governmental committee on the basis of poverty level, average unemployment, food-security, and emigration rates.
- After this first phase, the program may be scaled-up to the national level.
- Eligible program participants will be unemployed female youth (18-35 years of age) living in the vulnerable districts identified by the program. The program has not yet started.

The government asks for your help: They want to learn about the impact of the program on employment and entrepreneurship for female youth.

Please suggest a **quasi-experimental** impact evaluation method that could be used. Explicitly describe how you would identify the treatment and comparison group.



45 min break

Assignment

- A recent national survey in Fantasia Land has revealed that the employment rate for female youth is 55%.
- The government has decided to engage with local no-profit organizations to implement trainings to promote employability and entrepreneurship of female youth. More vulnerable districts will be targeted in a first phase (2022-2024), where vulnerability is defined by a governmental committee on the basis of poverty level, average unemployment, food-security, and emigration rates.
- After this first phase, the program may be scaled-up to the national level.
- Eligible program participants will be unemployed female youth (18-35 years of age) living in the vulnerable districts identified by the program. The program has not yet started.

The government asks for your help: They want to learn about the impact of the program on employment and entrepreneurship for female youth.

Please suggest a **quasi-experimental** impact evaluation method that could be used.

- Explicitly describe how you would identify the treatment and comparison group
- What are the limitations and strengths of the method(s)?

Assignment - Discussion

Treatment group: participating unemployed female youth (18-35 years of age) in program districts

A) DiD

Comparison group :



Assignment - Discussion

Treatment group: participating unemployed female youth (18-35 years of age) in program districts

A) DiD

Comparison group : unemployed female youth (18-35 years of age) in non-program districts

Assignment - Discussion

Treatment group: participating unemployed female youth (18-35 years of age) in program districts

A) DiD

Comparison group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

Assignment - Discussion

Treatment group: participating unemployed female youth (18-35 years of age) in program districts

A) DiD

Comparison group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

• Assumptions:

- Parallel trends: Outcomes for T and C groups would have experienced the same growth trends in the absence of programme
- There are no unobserved shocks differently affecting T and C groups and affecting the outcome

Assignment - Discussion

Treatment group: participating unemployed female youth (18-35 years of age) in program districts

A) DiD

Comparison group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

- Assumptions:

- Parallel trends: Outcomes for T and C groups would have experienced the same growth trends in the absence of programme
- There are no unobserved shocks differently affecting T and C groups and affecting the outcome

- Data:

Assignment - Discussion

Treatment group: participating unemployed female youth (18-35 years of age) in program districts

A) DiD

Comparison group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

- **Assumptions:**

- Parallel trends: Outcomes for T and C groups would have experienced the same growth trends in the absence of programme
- There are no unobserved shocks differently affecting T and C groups and affecting the outcome

- **Data:** Baseline (BL) and Endline (EL)

Assignment - Discussion

Treatment group: participating unemployed female youth (18-35 years of age) in program districts

A) DiD

Comparison group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

- **Assumptions:**

- Parallel trends: Outcomes for T and C groups would have experienced the same growth trends in the absence of programme
- There are no unobserved shocks differently affecting T and C groups and affecting the outcome

- **Data:** Baseline (BL) and Endline (EL)

B) DiD + Matching: Matching can be combined with DD to match participants with most similar non-participants on the basis of relevant characteristics explaining the selection into the program.



Center for Evaluation
and Development

Assignment - Discussion

C) Matching

Assignment - Discussion

C) Matching

Control group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

- Assumptions:

Assignment - Discussion

C) Matching

Control group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

- Assumptions:
 - No unobserved differences between T and C groups correlated with outcome and program participation
 - There is sufficient common support in the probability of treatment (propensity score) between T and C group
- Data:

Assignment - Discussion

C) Matching

Control group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

- Assumptions:
 - No unobserved differences between T and C groups correlated with outcome and program participation
 - There is sufficient common support in the probability of treatment (propensity score) between T and C group
- Data: Ideally Baseline (BL) + Endline (EL)

Assignment - Discussion

C) Matching

Control group : unemployed female youth (18-35 years of age) in non-program districts

Alternative comparison group: unemployed female youth (18-35 years of age) in program districts

- Assumptions:
 - No unobserved differences between T and C groups correlated with outcome and program participation
 - There is sufficient common support in the probability of treatment (propensity score) between T and C group
- Data: Ideally Baseline (BL) + Endline (EL)

D) RDD ?

Is there a cut-off rule?



Center for Evaluation
and Development

SETTING EXPECTATIONS RIGHT



- **Timeline**
- **Data quality and research ethics**
- **Data needs and sources**
- **Budget**

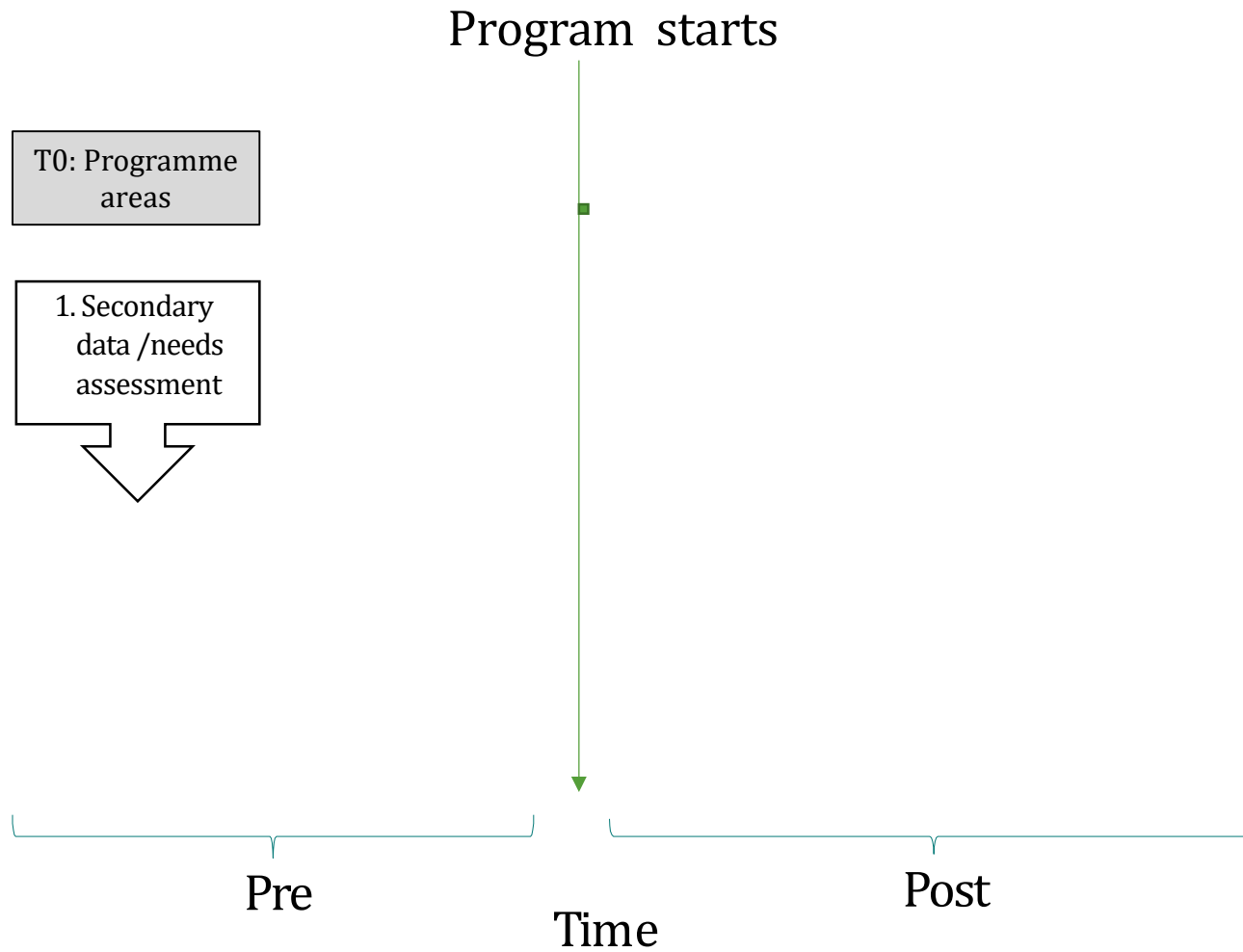


Center for Evaluation
and Development

TIMELINE

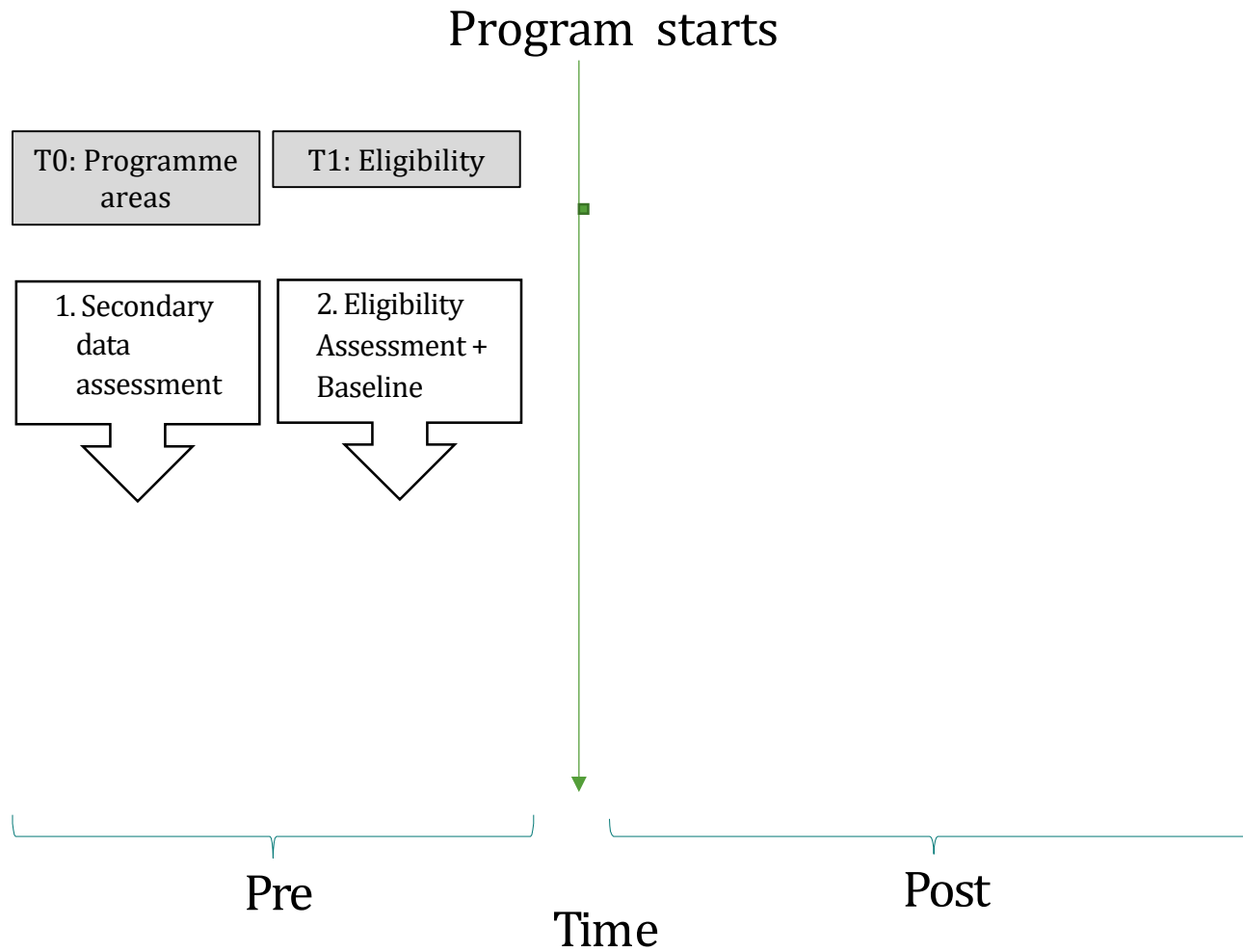


Timeline



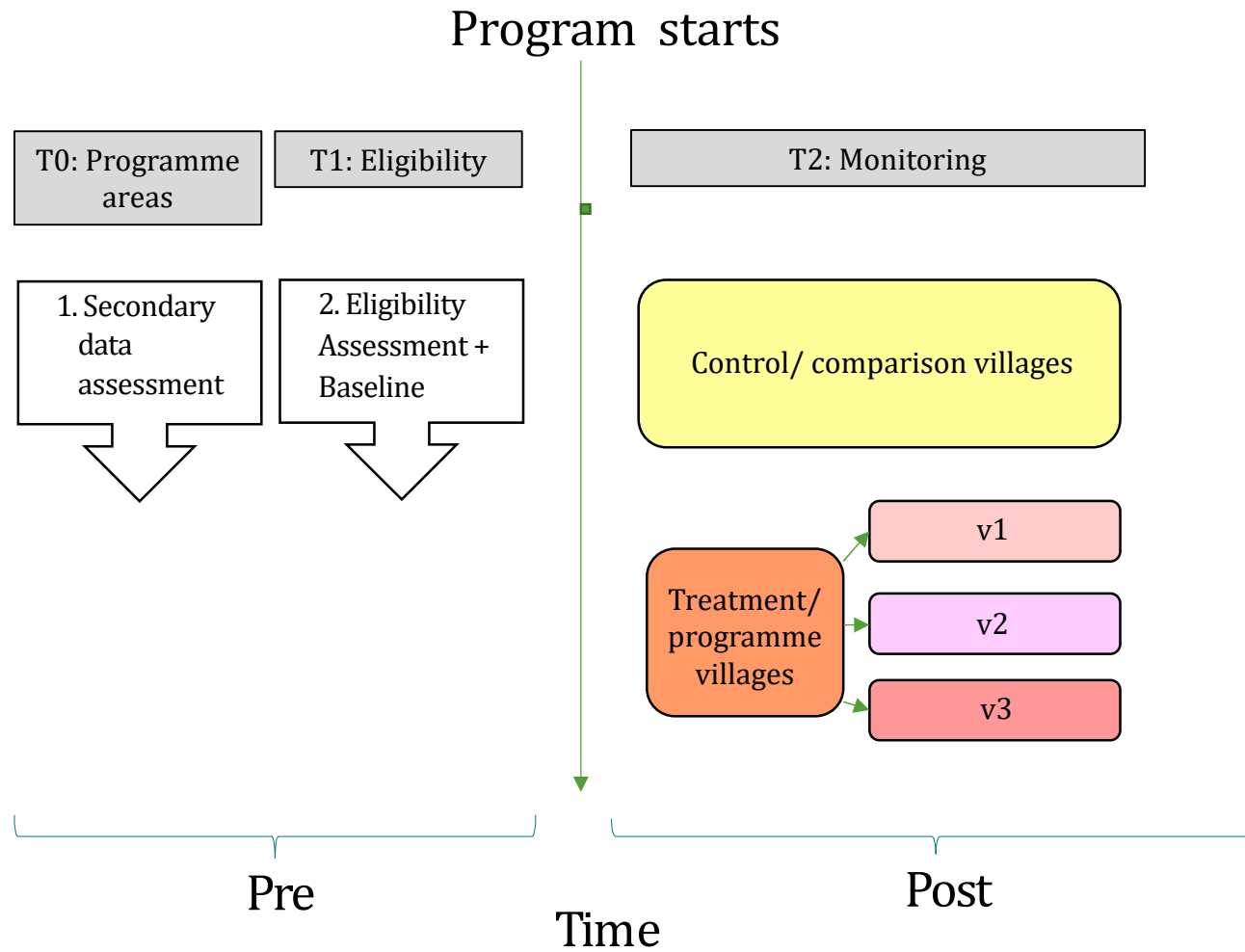


Timeline

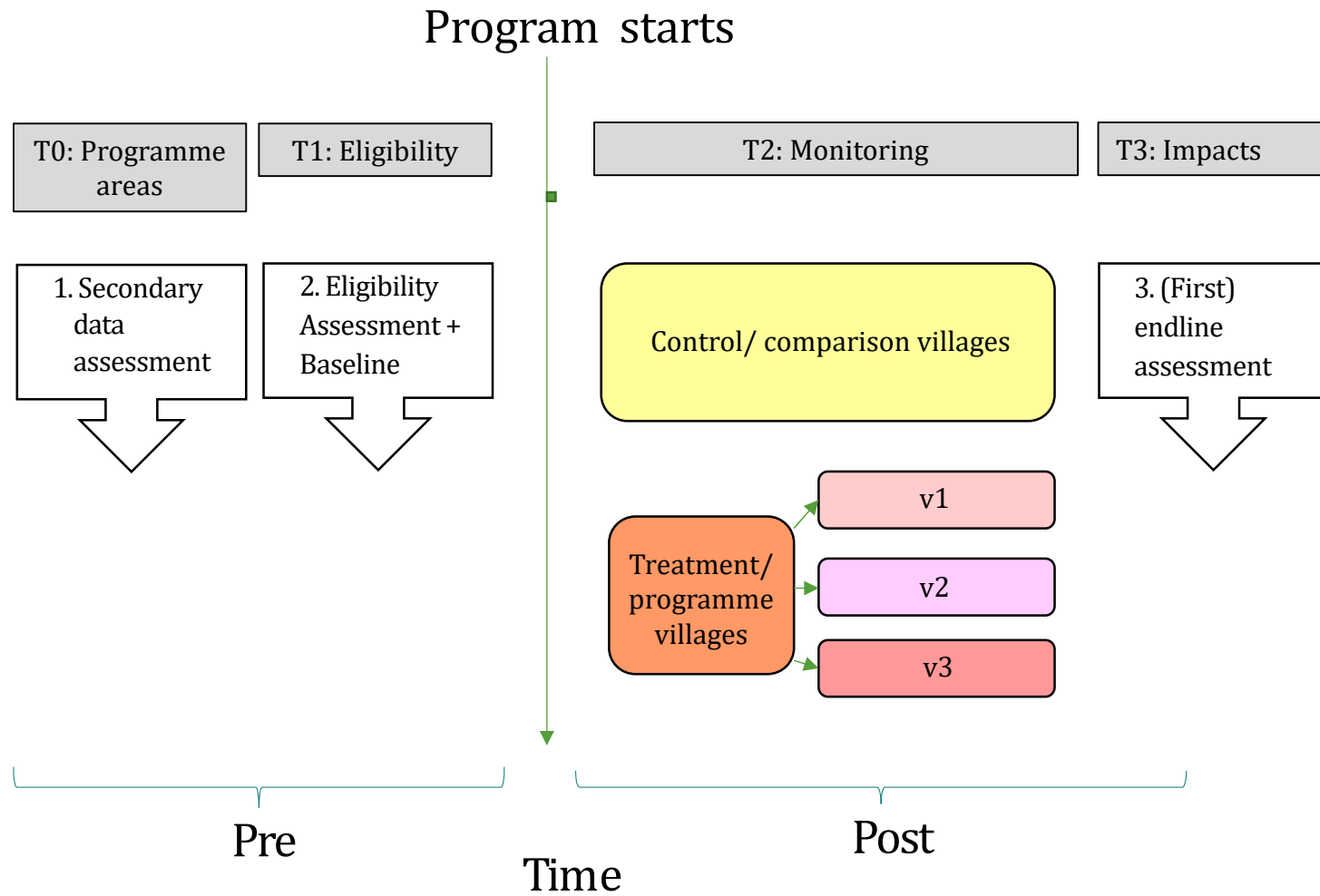




Timeline



Timeline



General remarks:

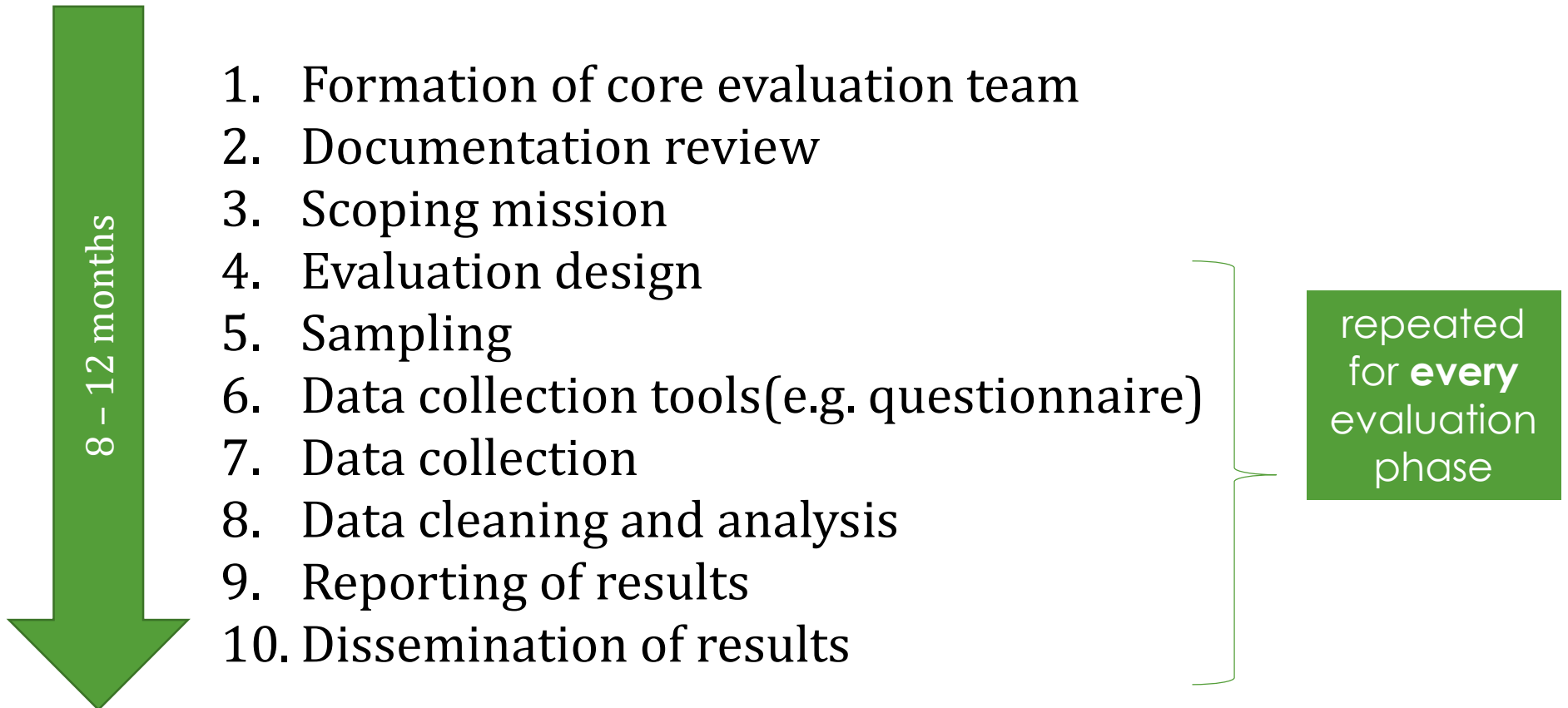
- Impact Evaluation and project implementation are *intertwined*
- Robust impact evaluation is planned in the *beginning* of the project, before start of project implementation
- Evaluation phases:
 - I. **Baseline** (if needed): before project implementation
 - II. **Midline** (optional)
 - III. **Endline**: Reasonable timing for estimation of impact



Timeline

- Decision for baseline and midline depends on the *selected evaluation design* as well as project interests and resources
 - RCT → baseline data collection is highly desirable but not strictly necessary
 - DiD → baseline data collection is mandatory
- Should be determined *together* with an IE specialist

Timeline of one evaluation phase



Timeline – Example for Baseline

Baseline -Year 2022 - Months													
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	
Preparation of Scoping Mission	■	■											
Scoping mission		■											
Desk review		■	■										
Writing of IE design report			■	■									
Preparation of survey tools				■	■	■							
Preparation data collection				■	■	■							
Pre-test and training						■							
Data collection						■	■	■					
<i>Project Implementation to start (earliest)</i>													
Data cleaning								■	■				
Data analysis									■	■			
Writing of IE Baseline report										■	■	■	
Dissemination of findings												■	■

Possible hitches and glitches

1. Foreseeable challenges



- Ethical clearance and local research permissions
- Procurement takes time
- Holiday/festivals/elections
- Missing/incomplete data

1. Plan sufficient time for activities !!
2. Local knowledge for timing is important !!
3. Get contact information of respondents !!

Possible hitches and glitches

2. Unforeseeable challenges



- Natural disasters, pandemics, local conflict
- Delays in project implementation
- Change in project team/contact person of local partner

4. Be prepared for changes and include buffer !!
5. Be flexible and innovative !!
6. Get documentation for everything !!



Center for Evaluation
and Development

DATA QUALITY AND RESEARCH ETHICS

Strategies to improve data quality

- **Improve** how to collect, store, and manages data over the course of the program
- Consider (sector-wide) **guidelines related to the ownership, protection and security of data** (define internal institutional framework for data governance)
- Leverage data science innovations from the private sector
- Consider data quality checks

High relevance

- **The World Development Report for 2021 is on Data for Development**

The report will

- **influence** research and practice;
- spur a discussion amongst relevant actors for harnessing the value of data for the poor and establish best practices for policy making

[Link to report](#)

Pre-Analysis Plan

- Pre-analyses Plans (PAPs) have become a prominent tool to promote data and research ethics over the last decade
- PAP sets out in advance how the researcher will analyze data: research hypotheses, indicators, measurement, IE method, sampling strategy, strategies for data cleaning, attrition, estimation and statistical inference.

Suggested readings:

[WB blog post on PAP checklist: link](#)

[Paper: , Olken \(2015\), Promises and Perils of Pre-Analysis Plans](#)

Pre-Analysis Plan

A pre-analysis plan is voluntarily developed by the researchers in order to :

- show commitment against “data-mining” and cherry-picking either positive or negative statistically significant results

Idea: If a researcher can choose which results to report, it is easy to see how results can be manipulated.

- refine the analysis strategy

How does it work?

- In an ideal scenario the PAP corresponds to writing the report before seeing the results (endline data)
- However, in most cases, the data itself may reveal interesting patterns that are worth exploring, which is why the document is (morally) binding in its major analysis yet not binding in further ones as long as those are well documented.

Pre-Analysis Plan

- **Shortly before endline data-collection begins** the document is logged (safely secured/ archived) online (AEA RCT Registry website; RIDIE- 3IE website) and might be later requested by the scientific community for reference.

For whom is a Pre-Analysis plan useful?

- It is not a document for ethical or governmental clearance, though it can be used as such.
- Transparency and commitment of an impact evaluation.



DATA NEEDS AND SOURCES



PRIMARY DATA COLLECTIONS

**(collected directly by researchers from beneficiaries
and main sources)**

Collecting the right data

The CART principles



Credible

Collect high quality data and analyze the data accurately



Actionable

Commit to act on the data you collect



Responsible

Ensure the benefits of data collection outweigh the costs



Transportable

Collect data that generate knowledge for other programs

Source: Gugerty, Mary Kay and Dean Karlan (2018)

Collect high quality data and analyze them accurately

This is possible when data are:

- Valid: should capture the essence of what they are seeking to measure.
- Reliable: the same data collection procedure should produce the same data
- Unbiased: there should not be systematic differences between how someone answers a question and the true answer

Commit to act on the data you collect

- Do you have a plan on how to use the data?
- Only collect data that you will use
 - Is there a specific action that you will take on the findings?
 - Do you have the resources and the **commitment to take that action?**

- Set up the right systems to handle the data you collect

Ensure the benefits of data collection outweigh the costs

- Collecting too much data is inefficient
- Too little data or not collect data on about what took place is not responsible
 - Lack of data could hide flaws of a program and lead to wrong decision making

Ensure the benefits of data collection outweigh the costs

- Collecting too much data is inefficient
- Too little data or not collect data on about what took place is not responsible
 - Lack of data could hide flaws of a program and lead to wrong decision making

Trade-offs:

- Data collection methods: are there cheaper/ more efficient methods without lowering quality?
- Resource use: Is the budget justified given the expected results compared to the rest of program budget?
- Use of respondents' time: Does the information sought justify the time asked to respondents?
- Is the timing right for an impact evaluation? How much do we expect to learn? Will future decisions be influenced by the results?

Transportable

Collect data that generate knowledge for other programmes

- The goal is to generate lessons that can help design/invest in effective programmes and policies
- Need of an underlying theory to explain the findings: Can your ToC be replicated?
 - Clear and complete ToC will help generating similar work, assessing whether your ToC may be expected to work in other contexts



SAMPLE SIZE

Sample size

- Evaluation Question: What is the impact or causal effect of a program on the outcome of interest?
 - In other words: Is the measured program impact different from zero?
- So, how large should the sample be? What is the minimum sample size required to conduct a study that will convincingly answer the policy questions of interest?
 - The larger the sample, the more precise the estimate BUT ...
 - Collecting more data is costly !
- Key practical concern for IE: **trade-off** between the cost of data collection and the precision of estimates
 - Power calculations are a tool to inform this trade-off and define sample size required for IE

Power Calculations: what goes into the formula?

1. Minimum Effect Size (MES) or Minimum Detectable Effect (MDE)

What is the level of impact below which an intervention should be considered unsuccessful?

Intuition: harder to detect small impacts than large impacts

Small impacts → requires large **sample size**

2. Baseline Information

Baseline (average) value of the outcome of interest

Baseline standard deviation of the outcome of interest

3. Statistical Precision

Precision with which we can measure the MDE, given the sample size.

2 components: **significance level** and **statistical power**

Power Calculations

1. Where can we find information on baseline outcome values? Is such information available for your project?

- Previous studies of similar projects in similar settings
- Secondary data (e.g. nationally representative surveys)
- Project data (e.g. feasibility study? Primary data?)

2. How can we determine the MDE? Is there a clearly defined expected impact for your project?

- Previous studies of similar projects in similar settings
- Policy objectives → what is considered an acceptable MDE for the program to be considered successful?
- Economic analysis/ feasibility studies

Power Calculations

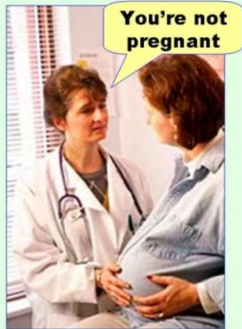
3. Statistical precision

Table 7.1: Possible Errors in Estimating Impact

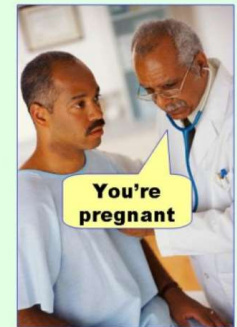
	Find No Significant Impact	Find a Significant Impact
Intervention has no impact	No error (correct conclusion)	Type I error (False positive)
Intervention has an impact	Type II error (False negative)	No error (correct conclusion)

Source: White and Raitzer (2017)

Type II error
(false negative)



Type I error
(false positive)



- **Significance Level** = likelihood of Type I error
 - Popular value in social sciences = 5%
- **Statistical Power** = probability of detecting an impact when it exists in reality → effectively 1 minus likelihood of Type II error
 - Popular value in social sciences = 80%

Power Calculations – Trade-offs

WARNING

- Trade-offs in power calculations are **non-linear**

Example:

- Baseline income \$1,000, standard deviation 1,000
- 5% significance and 80% power
- Expected effect: +50% (i.e. +\$500)
 - Required sample size: **128** (64 in treatment group, 64 in comparison group)
- Expected effect: +25% (i.e. +\$250)
 - Required sample size: **506** (253 per group)

MDE divided by 2, but required sample size almost **quadrupled** !!!

Why do you need a large sample size?

Other things being equal, you need a higher sample size...

- if you want to capture a small MDE
- if you anticipate imperfect take-up
- if you anticipate high attrition
- if you have no baseline values
- if your outcome shows high variance
- if you need to cluster your implementation level
- if units (e.g. people) in a cluster (e.g. village) are very similar
- if you use a quasi-experimental evaluation method (e.g. PSM)

Quiz?

Quiz 3

Go to
www.menti.com

Enter the code
2876 9433



Or use QR code

Quiz?

Go to www.menti.com and use the code 2876 9433


Which one of the following statement is NOT true? You need a larger sample size

 Mentimeter

0%
If you want to capture a small effect size

0%
If you anticipate imperfect take up

0%
If you anticipate low attrition

 Voting is closed

 Results are hidden

Press ENTER to show correct

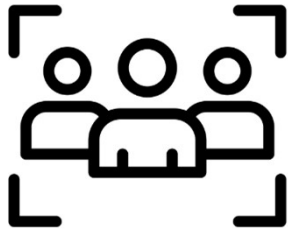




Center for Evaluation
and Development

MIXED-METHODS

Mixed methods: Why qualitative research in evaluation?



Understand the views, experiences and motivations of beneficiaries, implementers and stakeholders in greater depth.



Understand the processes and mechanisms by which impacts occur – How and why?

- Investigate if a project had any unintended (both positive and negative) consequences



Questions about meaning and motivation examine how a particular behavior or action is understood, or how people make sense of their circumstances.

Suggested reading: [Link to Mixing qualitative and quantitative methods: a conversation \(worldbank.org\)](https://www.worldbank.org)

Main methods



Key Informant Interviews with people (e.g. community leaders, program staff) who have particularly informed perspectives on an aspect of the program being evaluated

In-depth interviews with participants to learn about about their experiences and expectations related to the program,



Focus group discussions are group interview designed to explore people's attitudes about aspects being evaluated



Observations: systematic observations to understand phenomena , especially hidden ones (e.g. child labour)

Examples of qualitative contributions

1

Causes and consequences of child labor in ET:

- Role of urbanization / modernization
- Peer experiences, independence, city life

2

Water project in Benin:

- Misunderstandings due to indirect communication via 'middle-men'

3

Gender-based violence in DRC: willingness of men to change day-to-day behavior, if maintain feeling of authority / respect in household, control over certain decisions

4

Business in Ghana: Women investing without profit maximization due to gender roles and sake of marital relationship, etc.

Sampling for qualitative research

- Qualitative analyses typically require much smaller sample size than quantitative analyses
 - The goal of qualitative researchers is to attain saturation, which occurs when adding more participants to the study does not result in additional perspectives of information
- Sampling is usually non-random (e.g. via snowball sampling)
 - The focus is NOT on generalizable results or on detecting causal impact BUT on describing perceptions, potential mechanisms, triangulation



Center for Evaluation
and Development

SECONDARY DATA COLLECTIONS

(administrative data)

From Primary to Secondary Data



Secondary data sources

- Administrative records (anonymized but disaggregated)
- Private sector data (e.g. on consumption)
- Geo-referenced data

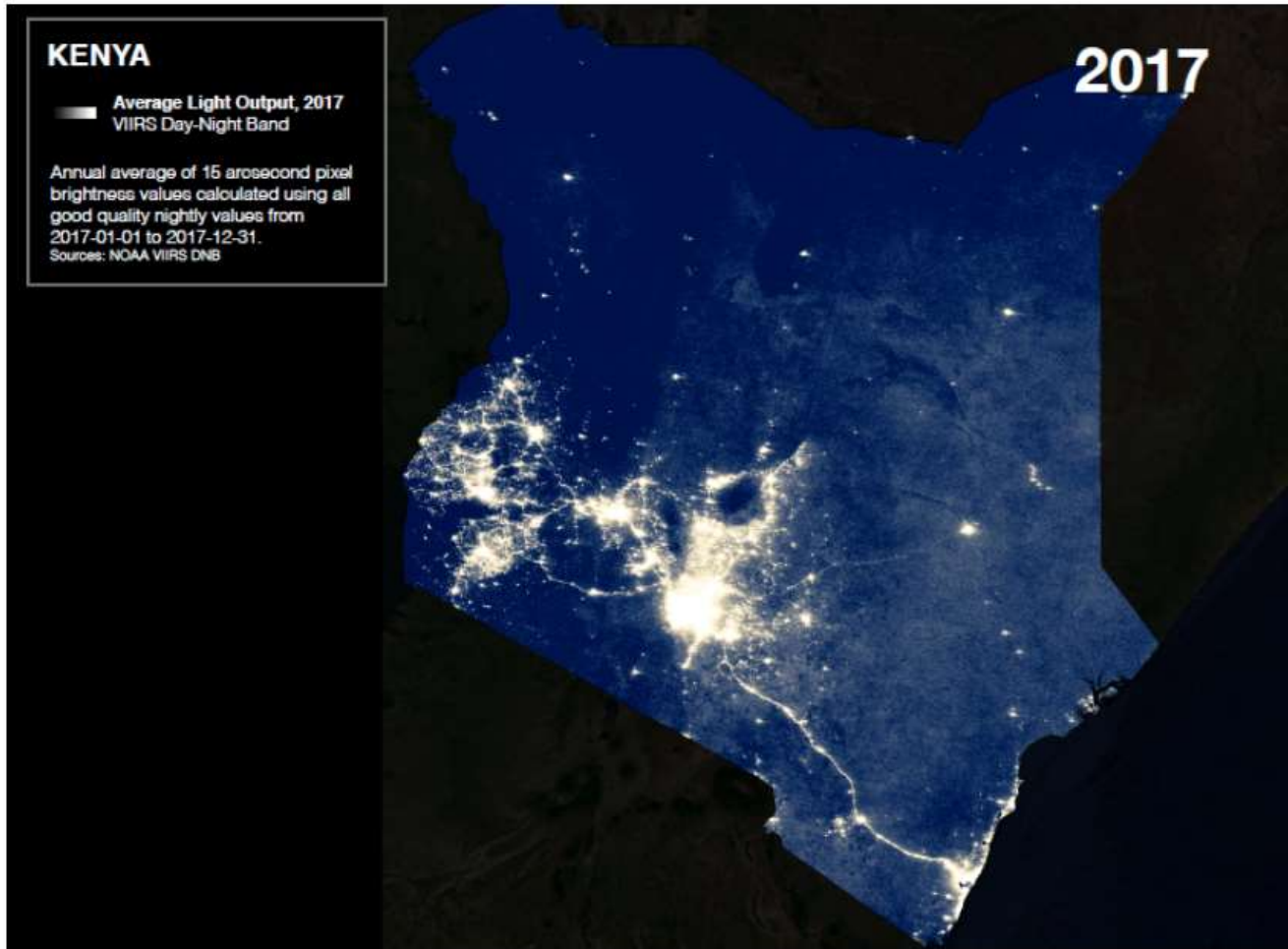
Secondary data sources

- Administrative records (anonymized but disaggregated)
- Private sector data (e.g. on consumption)
- Geo-referenced data
- Internal monitoring data !

Book recommendation: [Link to J-PAL's new book on administrative data sources \(online free access\)](#)



Example: Secondary data



Source: [Link to World bank blog «Innovations in satellite measurements for development»](#)

Example: Secondary data



Source: [Link to World bank blog «Innovations in satellite measurements for development»](#)

Strategies to improve data utilization

- **Identify** secondary, administrative data sources (MIS data; private sources, such as mobile phones, electronic transactions, and satellites)
- **Build** secondary data-bases where non-existent to avoid frequent, expensive data-collections
- **Improve** usability of data stored (formats; linkage/ IDs) and connect databases
- **Allow** for the usage of these data sets in creative and innovative ways (ex-post evaluations; nudging/ nimble evaluations); Share data with researchers
- **Apply** advanced technical methods for data-collection, analysis and experiments to generate greater learning
- **Extend** partnerships

Benefits of secondary data

- **Faster** than time-consuming surveys
- **Less expensive** data
- **Bigger** administrative data sets
- More **precise impact** estimates
- More **accurate** with less attrition
- More **inclusive** than survey data

Monitoring data

Monitoring generally involves tracking progress with respect to previously identified plans or objectives, using data easily captured and measured on an ongoing basis.

Impact evaluation **should not** proceed without solid data on implementation.

Monitoring: Purpose

Monitoring is carried out for a variety of different purposes, generally having little to do with impact evaluation.

For example:

- Internal use by project managers to identify if the project is on target or not (e.g. what services actually are being provided; who is being served)
- Address donor demands for reporting and accountability
- Serve as an **early warning system**, and in the case of negative or unexpected findings may suggest the need to consider a change in approach while the project or program is still underway

Example: Monitoring

Monitoring could allow tracking:

- **which** new treatment arms are introduced.
- **which** new treatment arms are introduced **together** (combinations).
- **whether** the randomization protocol /IE design is followed.
- **when** new treatment arms (top-ups) are introduced. The identification of the exact timing would allow to measure the exposure intensity (in terms of duration) to the treatment.



Center for Evaluation
and Development

BUDGET

Budget

- Budget for IE (esp. data collection) should be estimated *realistically* and *earmarked* in the beginning of the project
- **Determining factors:**
 - Overall *living cost/price level* in a country
 - *Sample size* and numbers of evaluation points
 - *Transport*
 - *Security*
 - Number of *languages* spoken in project region
 - *Outsourcing* of data collection to an external firm

Budget items for data collection

Examples of budget items

Staff Cost	Field coordinator, supervisor, enumerator, moderator (qualitative), translator....
Training Cost	Training venue, catering, training stipend for participants, accomodation....
Transport	Car hire, fuel, driver, bus fare, motorcycle during training and data collection
Other	Tablets, incentives, printing of training material, communication/internet cost, venue for focus group discussions (qualitative)

Budget – External firm

Pro/Contra External Firm for Data Collection and/or Analysis



- Specialized firms usually produce higher data quality
- Frees time of the project team
- Often no choice since procurement is required and best practice
- Ensures independence of impact evaluation



- Procurement takes time
- Cost is usually higher (including for coordination)
- Less flexible and might be risky
- Still necessary to check data quality and analysis

For procurement:
important to have someone knowledgeable to judge
quality of technical proposals



Center for Evaluation
and Development



Thank you

- [Abebaw Ejigie, D., Fentie, Y. and Kassa, B. \(2010\), The impact of a food security program on household food consumption in Northwestern Ethiopia: A matching estimator approach, *Food Policy*, 35, issue 4, p. 286-293](#)
- [Berge, L., I. Oppedal, K. Bjorvatn, Tungodden, B. \(2011\) Human and Financial Capital for Microenterprise Development: Evidence from a Field and Lab Experiment. NHH Discussion Paper Sam 1 2011. Norwegian School of Economics, Bergen, Norway](#)
- [Gertler, P. J., Martinez, S., Premand, P., Rawlings L., Vermeersch, C. M. J. \(2011\) Impact Evaluation In Practice, First Edition, The World Bank.](#)
- [Gertler, P. J., Martinez, S., Premand, P., Rawlings L., Vermeersch, C. M. J. \(2016\) Impact Evaluation In Practice, Second Edition, The World Bank.](#)
- [Gugerty, M. K., Karlan, D. \(2018\) The Goldilocks Challenge: Right-Fit Evidence for the Social Sector, Oxford Scholarship Online.](#)
- [MacPherson, C. Sterck, O. \(2019\) Humanitarian vs. Development Aid for Refugees: Evidence from a Regression Discontinuity Design, CSAE Working Paper Series 2019-15, Centre for the Study of African Economies, University of Oxford.](#)
- [Schultz, T. \(2004\) School subsidies for the poor: evaluating the Mexican Progresa poverty program, *Journal of Development Economics*, 74\(1\): 199-250.](#)
- [Tan, B. \(2018\) Prioritizing the Learning Agenda: the CART Principles, Practitioners Forum, Adaptive Programming and Monitoring, Evaluation and Learning, Philippines.](#)
- [Trzcinski, R. \(2011\) Active labour market measures and entrepreneurship in Poland, Impact Evaluation Spring School, Hungary.](#)
- [White, H., & Raitzer, D. A. \(2017\). Impact evaluation of development interventions: A practical guide. Asian Development Bank.](#)