Non-experimental Methods (non-technical track)

Hannah Uckat

Economist, Development Impact, World Bank

European Commission

WORLD BANK GROUP

Reminder: The idea of an impact evaluation (IE)

- Identify the **causal effect** of an intervention
  - For example, what is the impact of subsidized loans on business employment?
- What is a causal effect?
  - Changes in outcomes of interest (e.g., employment) that are **exclusively** explained by the intervention (e.g., subsidized loans).
- How to establish the causal link in an IE?
  - We need to find a valid **counterfactual**, so that we can *compare what happened to what would have happened* in the absence of the intervention.

Reminder: In search of a counterfactual

Problematic counterfactuals

- Comparisons
- Before – After
- Participants – Non-participants

Causal impact under certain assumptions & with limitations

- Non-experimental methods
  - Difference in differences
  - Regression discontinuity

Today!

Causal impact

(last session)

- Experimental methods (randomized trials)

Reminder: Randomized Controlled Trials

- **Random assignment of treatment** is considered the gold standard.
    - Relies on few assumptions
    - Less data is needed – but more planning
    - Easy to explain

- What if random assignment is not possible?
    - For example, large infrastructure projects (roads, irrigation) or sensitive policies (taxes)
    - There are **non-experimental methods of evaluation** (difference-in-differences, regression discontinuity design)
    - Each of these relies on **key assumptions that we cannot test**.
    - However, under these assumptions, we can evaluate programs that cannot be randomized.

Experimental approaches in 2019…

And non-experimental approaches in 2021!

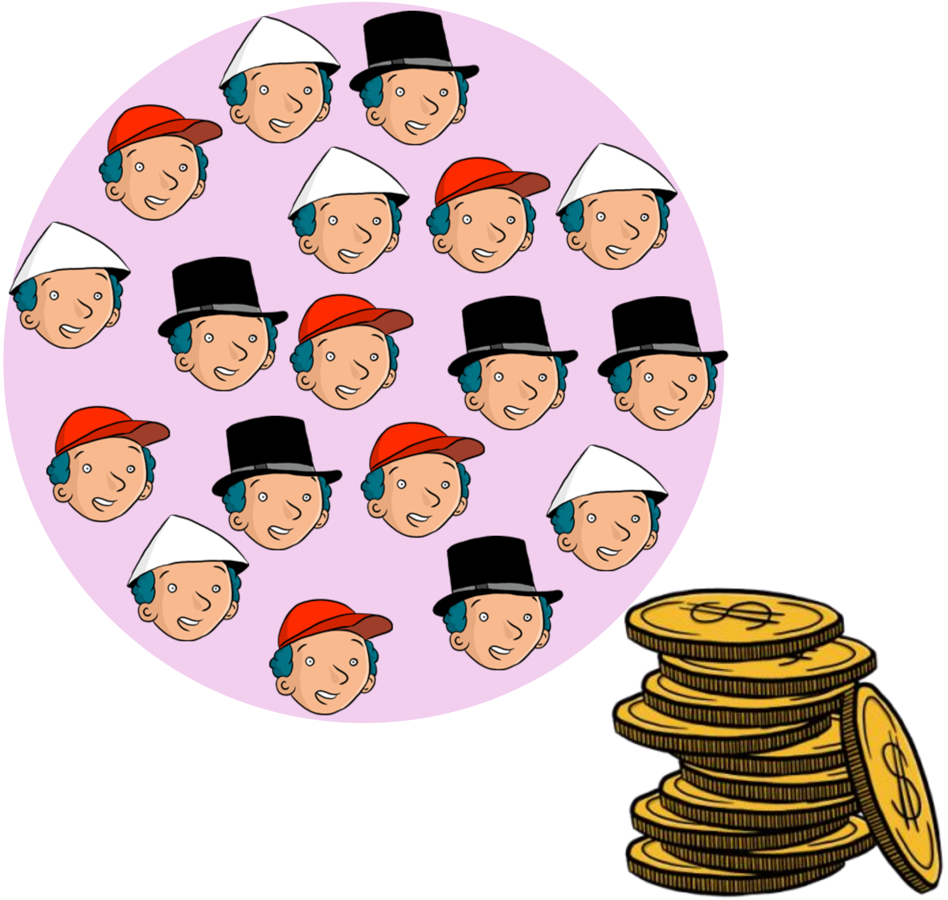
Non-experimental methods

1. Difference-in-differences (Diff-in-diff)
2. Regression Discontinuity Design (RDD)
3. Mix and match

Non-experimental methods

1. **Difference-in-differences (Diff-in-diff)**
2. Regression Discontinuity Design (RDD)
3. Mix and match

Hypothetical example



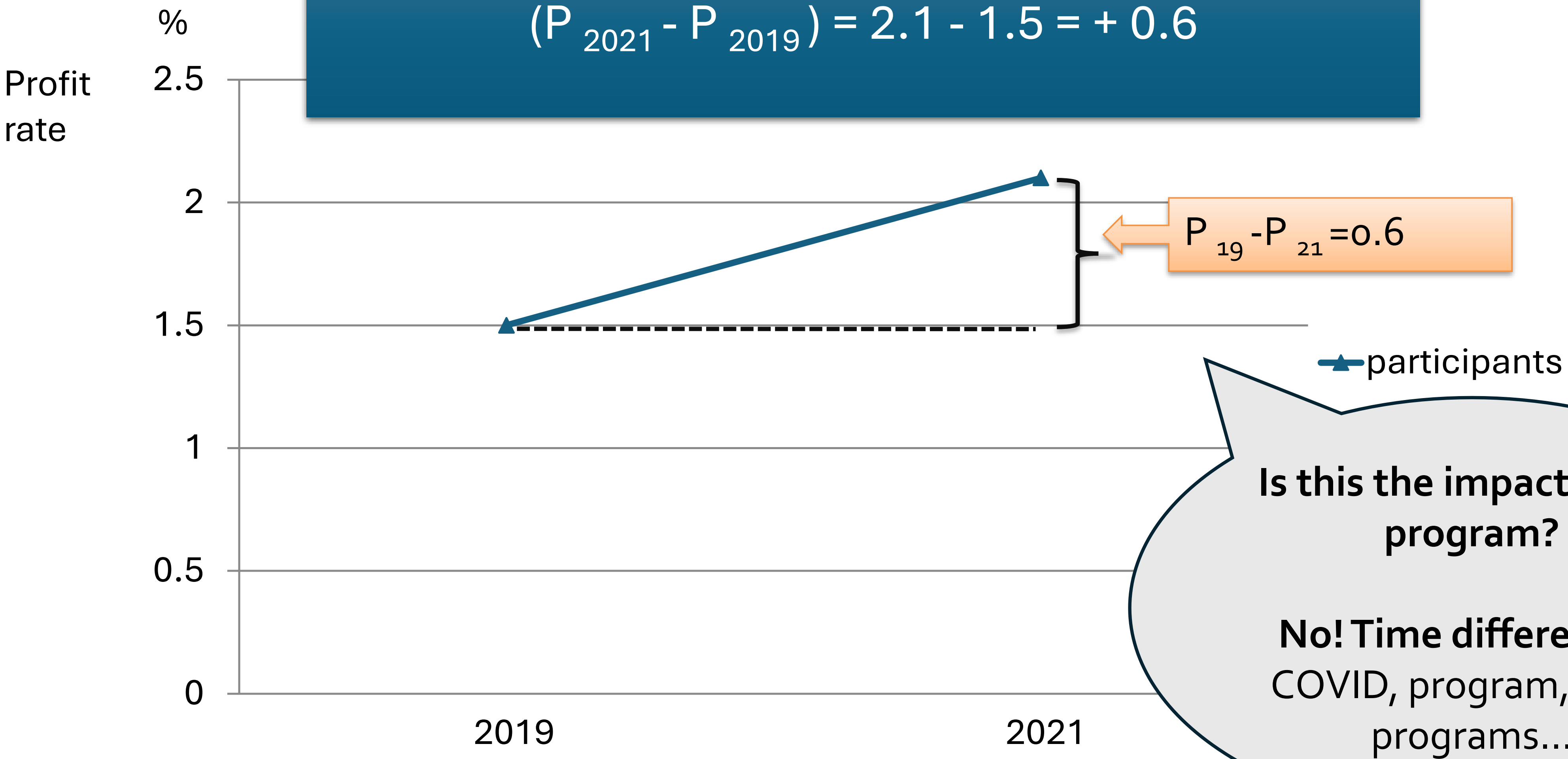
An MSME support agency wants to increase the profitability of businesses and provided them with subsidized loans in 2020.

Their question is:

What is the impact of a **subsidized loan** on the **profit rate** of companies?

Compare participants before and after? Problematic!

$(P_{2021} - P_{2019}) = 2.1 - 1.5 = +0.6$

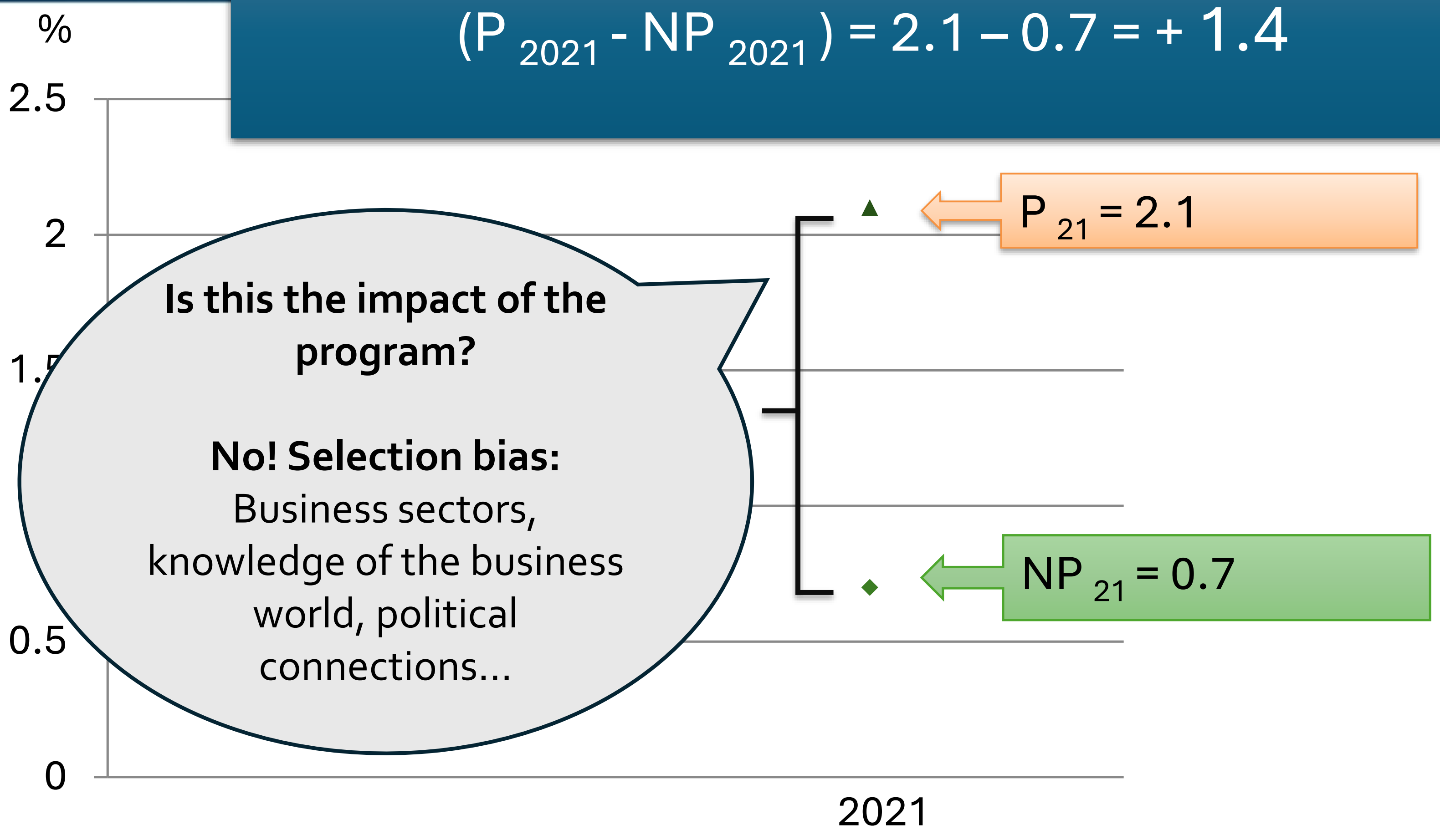


Is this the impact of the program?

No! Time differences: COVID, program, other programs…

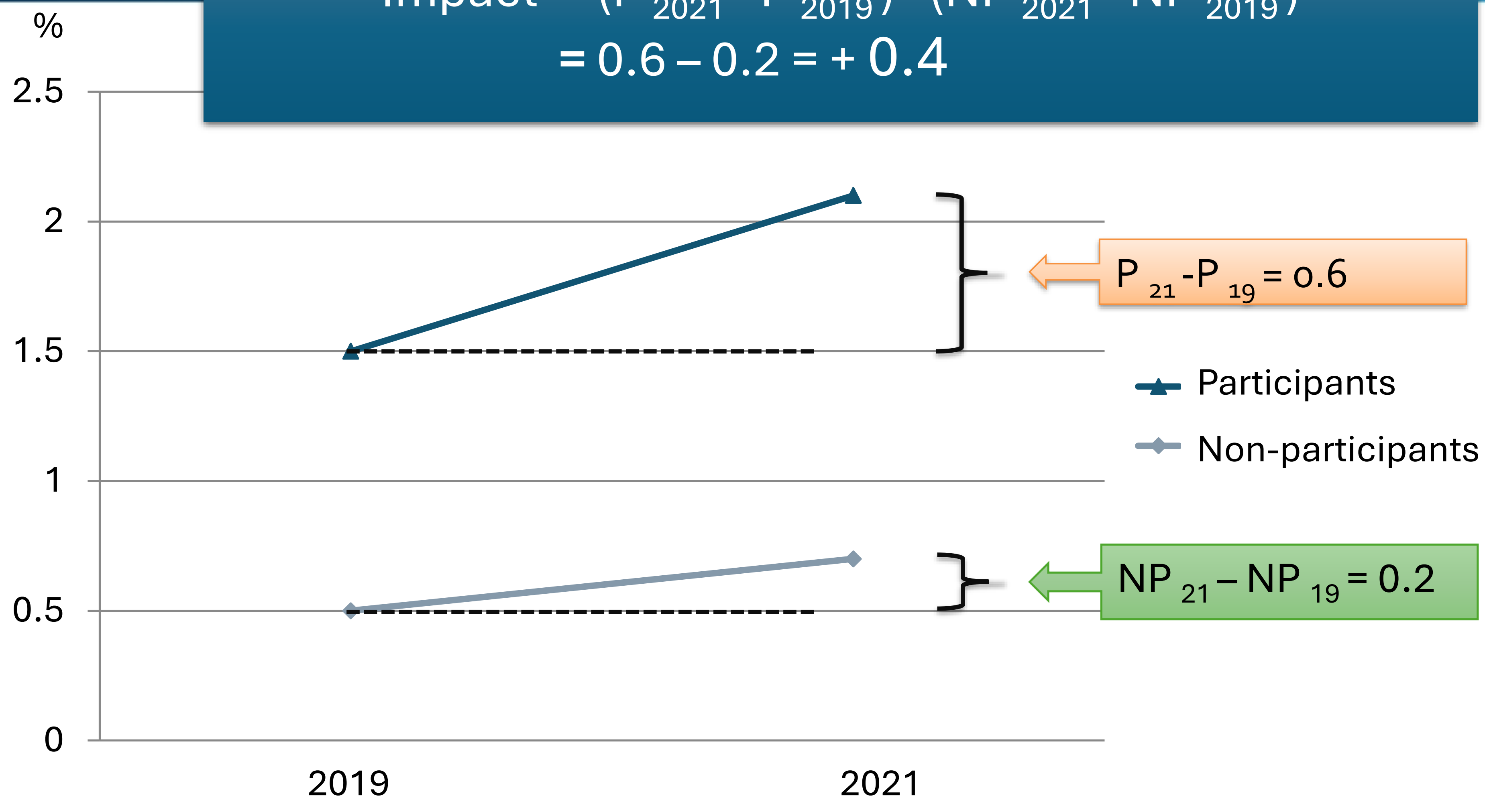
Compare participants and non-participants after? **Problematic!**

$$(P_{2021} - NP_{2021}) = 2.1 - 0.7 = +1.4$$



Difference-in-differences: Combine the two approaches!

$$\text{Impact} = (P_{2021} - P_{2019}) - (NP_{2021} - NP_{2019})$$
$$= 0.6 - 0.2 = +0.4$$



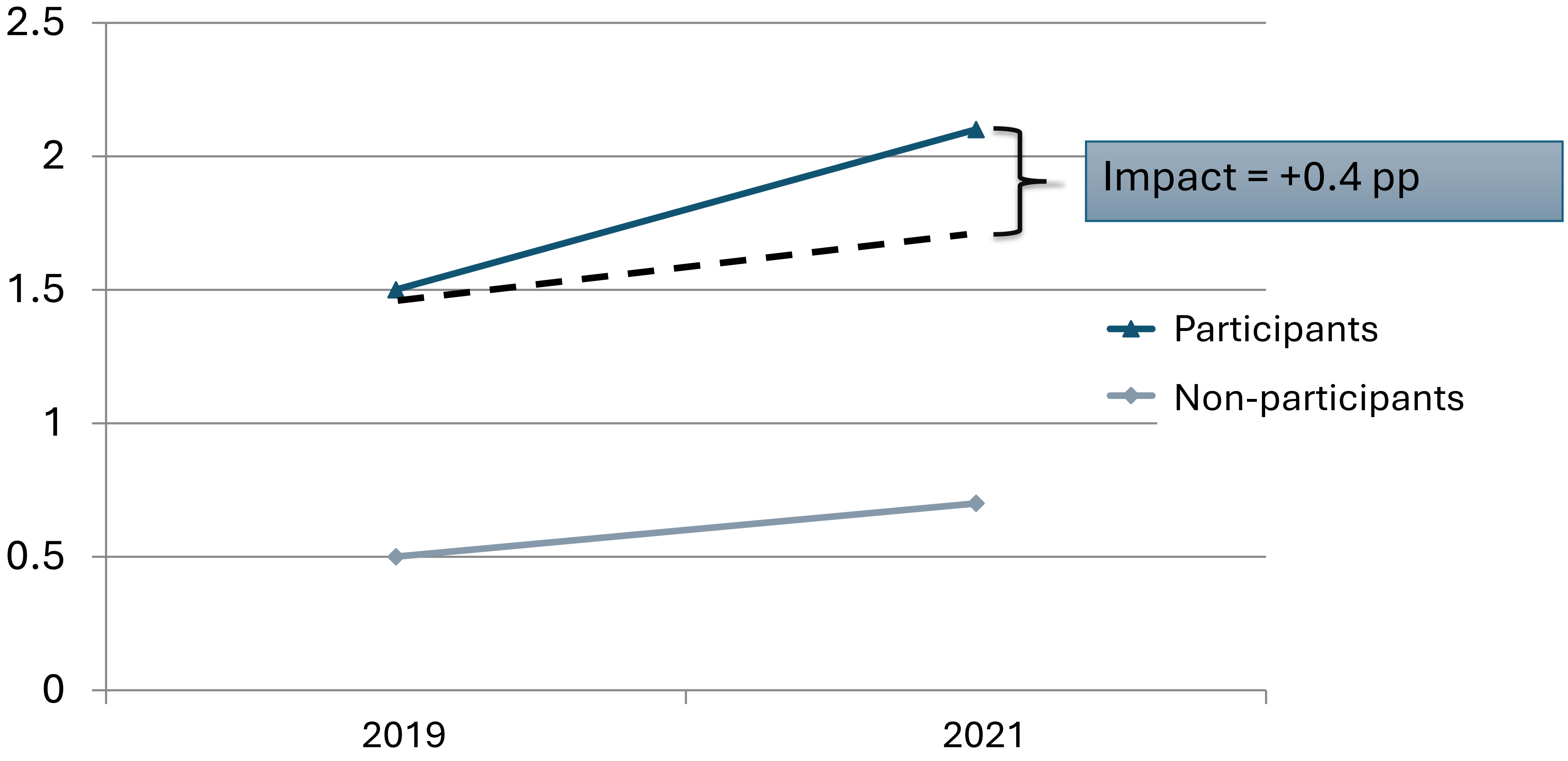
With a table

	Profit rate		
	2019	2021	**Difference** (2021-2019)
Participants (P)	1.5%	2.1%	0.6 pp
Non-participants (NP)	0.5%	0.7%	0.2 pp
Difference (P-NP)	1.0 pp	1.4 pp	**0.4 pp**

What does Difference-in-Differences do?

- **Idea**: Combine the time dimension (of the *before-and-after analysis*) with the selection dimension (of the *participants/non-participants analysis*).
- Difference-in-differences **acknowledges that program beneficiaries may be different** from non-beneficiaries.
- **Key assumptions:** Difference-in-differences assumes that outcomes change over time for only one of two reasons
  1. Events that affect beneficiaries and non-beneficiaries the same (the common trend assumption)
  2. The program itself (which only affects beneficiaries)

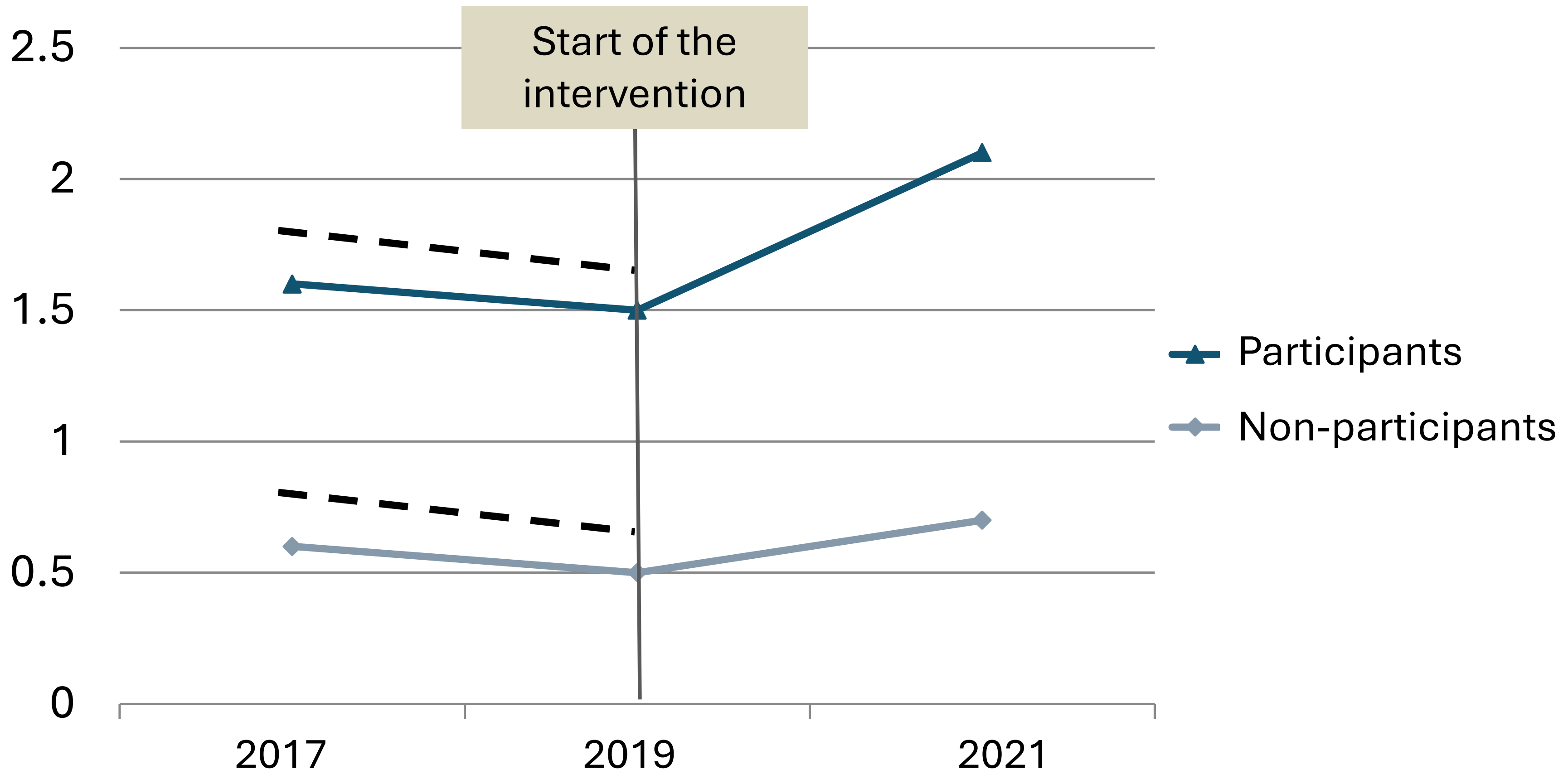
Key assumption: Common time trend in the absence of the intervention



What does the analysis of our example imply?

- In our example, the program had a positive effect on the profit rate.
- Is the hypothesis of a common/parallel time trend plausible?

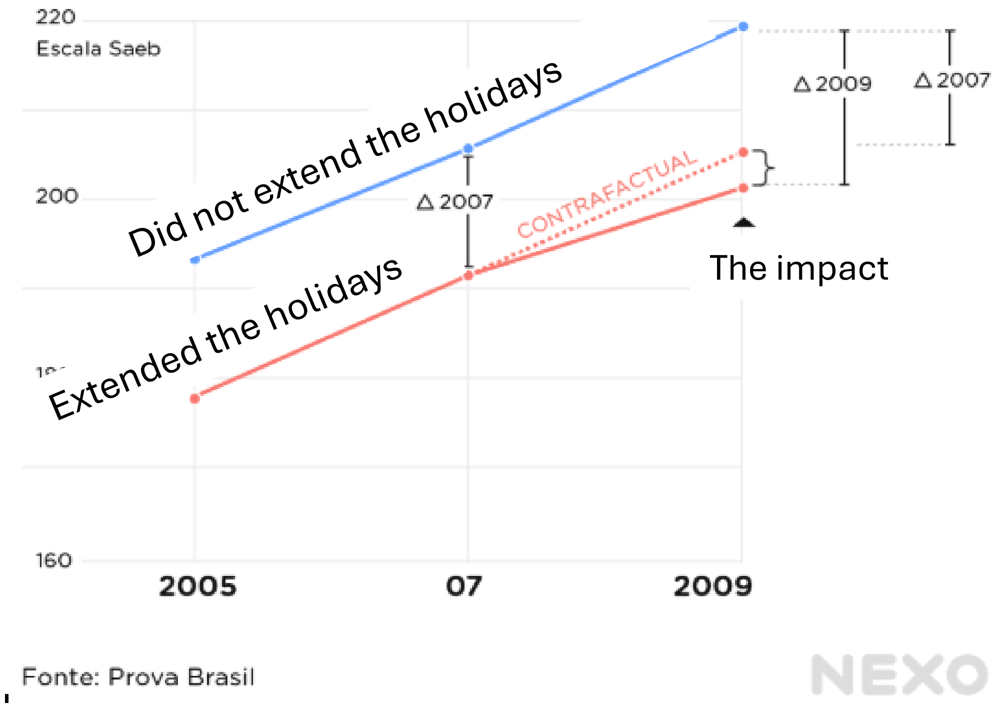
Analyze the plausibility of this hypothesis with historical data if possible



We need to make a compelling case for the assumption of common time trends

- We can’t know whether trends would have been the same.
- We need to provide evidence showing that control and treatment groups behaved similarly before intervention start
  - E.g. using administrative data
- The assumption is more likely to hold if the similarity at baseline and selection is based on criteria other than our outcome indicator of interest
  - Often not the case: targeted as certain groups, those not targeted may not be the best comparison

Application: The impact of school closures on learning



Source: Amorim et al. 2024

- In 2009, during the **H1N1 flu pandemic**, some municipalities in the state of São Paulo, Brazil, decided to extend school holidays for 3 weeks.
- **Comparison** between student learning in municipal schools that remained closed for 3 weeks (treatment group) and that in schools that did not remain closed (control group)
- **The impact:** Closing schools for 3 weeks reduced learning by about 2 months.

Summary: Difference-in-differences method

- **Idea:**
  - Compares differences in outcomes between participants and non-participants in the program *over time*

- **Identification hypothesis:**
  - “Parallel/common trends” in the absence of the program

- **The counterfactual**
  - *Change over time for <u>non-participants</u>* in the program is the counterfactual for *<u>participants'</u> change over time*

- Under the common trend assumption, diff-in-diff can produce unbiased estimates of the causal effect.

Summary: Difference-in-differences method

1. Need data on outcomes before and after the program was implemented
   - Ideally, this includes historical data for some results to analyze the common trend
2. Need a comparison group
   - Who did not receive the program (at the same time)
   - Who are comparable (e.g. similar in many characteristics, could be expected to have similar outcomes)
3. Need many units in treatment and comparison groups
   - We can't draw credible comparisons between (say) just two/ten/twenty companies
4. Need more advanced methods if there are multiple periods, units receive treatment at different times, and impacts vary for different units.
   - See his blog and this blog for a non-technical discussion.

Non-experimental methods

1. Difference-in-differences (Diff-in-diff)
2. **Regression Discontinuity Design (RDD)**
3. Mix and match.

Regression Discontinuity (RDD) Method

Many programs select using **an index or score** :

Anti-poverty programs	Target households below the poverty line
Pensions	Target the population above a certain age
Scholarships	Intended for students with high test scores
Agriculture	Fertilizer intended for small farmers (< a certain number of hectares)
Tax regimes	More favorable tax regime for smaller firms (e.g. < a revenue threshold).

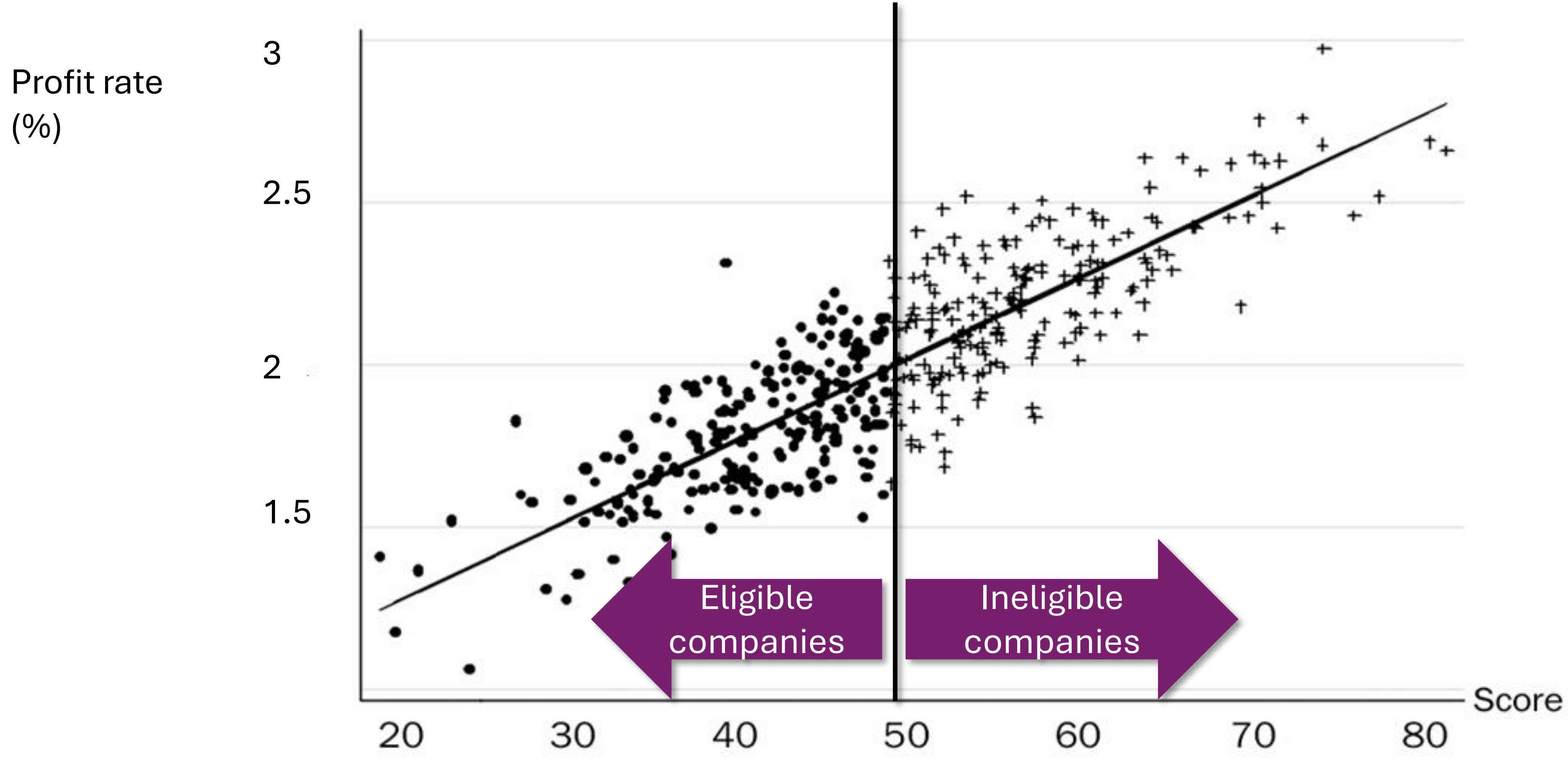
Hypothetical example

Intuition:

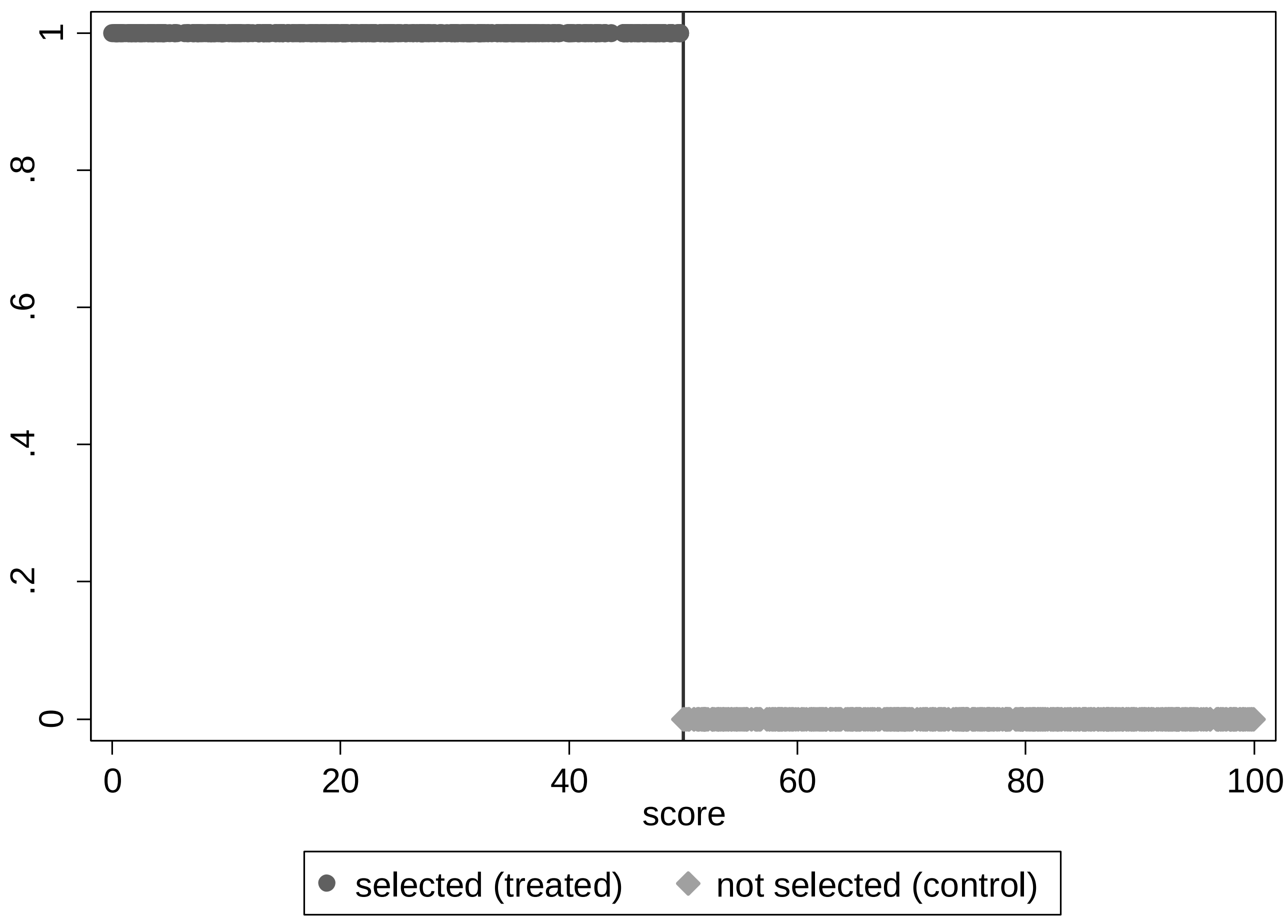
- A group of experts evaluates the expression of interest of all companies wishing to benefit from a subsidized loan
- The score ranges from 0 to 100
- The program aims to help businesses most in need. Therefore, the program is aimed at companies with a score <= 50.
- **Idea**: After the intervention, **compare the** profits of companies with a **score slightly below 50** (eligible for the subsidized loan)

  .... with companies whose **score is barely above 50** (ineligible for subsidized loans).
- Whether a company falls just above or just below the threshold is "as good as random."

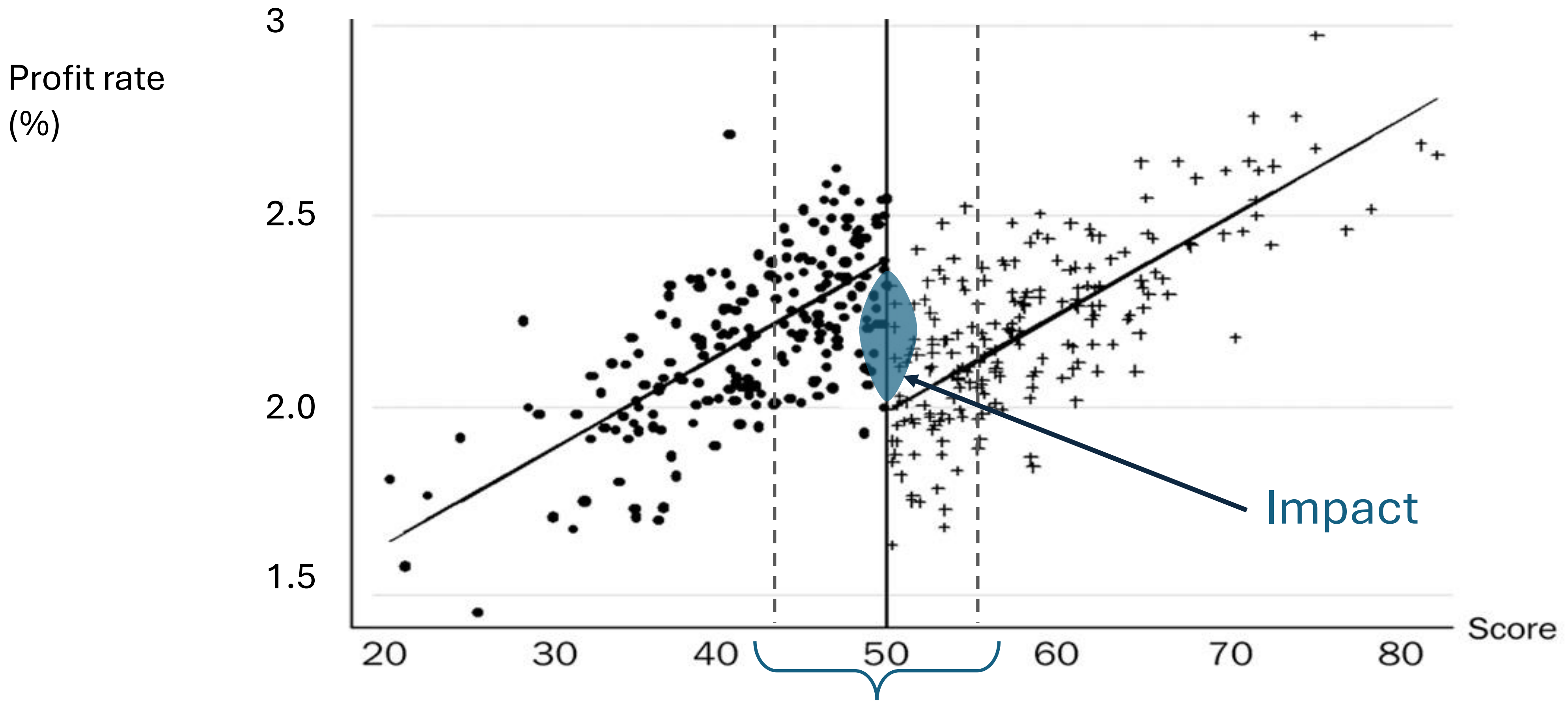
RDD: Profit rate before intervention



RDD: Probability of receiving treatment



RDD: Profit rate after intervention



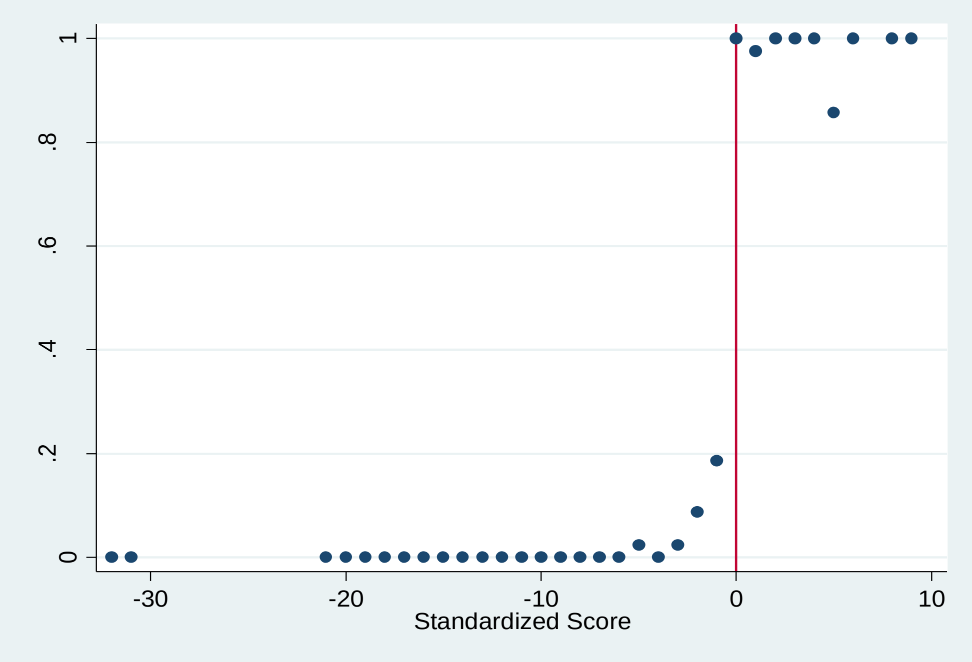
The size of this window is important!

Regression discontinuity design (RDD)

- **Idea:** RDD compares units just above the eligibility threshold to those just below.
- **Key assumption:** It is “as good as random” whether a unit falls just below or just above the threshold.
- RDD is an effective method if you have:
  - A continuous variable determining eligibility
  - A clearly defined eligibility threshold
  - No manipulation of eligibility
  - Large sample
- **Important**: The estimated causal impact is only valid for subjects who are close to the threshold defining eligibility for the program.
  - Is this the group you are interested in?

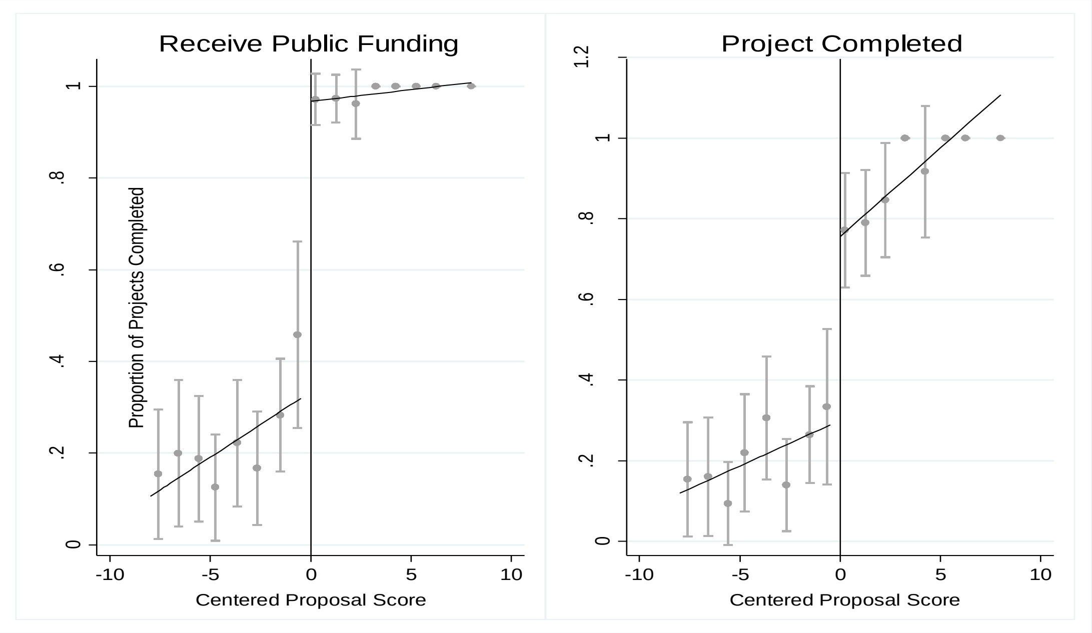
Application:
Effect of the Polish In-Tech program on innovation activities

Discontinuity in the probability of being funded at the eligibility threshold



Source: Bruhn and McKenzie (2016) - link to the article here

Funding resulted in additionality



Summary: Regression Discontinuity Design (RDD)

- **Fundamental hypothesis:**

  Units just above the threshold are comparable to those just below

- **RDD is based on understanding the selection process:**
  - With a clear selection rule and a simple and continuous quantifiable score, we know why some participants benefit, and others did not.
  - Program assignment is based on a threshold
  - Compare units around the threshold for evaluation

Summary: Regression Discontinuity Design (RDD)

- RD lends itself to evaluation when random allocation is not feasible:
  - Strategy applicable to any program that is based on a defined threshold
  - Possibility of exploiting multiple thresholds to improve external validity
  - Need a large sample
  - **The effect** is **causal** but **local** and therefore there is a **problem of generalization**
    - In our hypothetical example, RDD can answer the question "If we were to expand eligibility, what would be the impact of the subsidized loans on the newly eligible firms?"
    - RDD **cannot** answer "What is the impact of the subsidized loans on all firms that receive them?"

Non-experimental methods

1. Difference-in-differences (Diff-in-diff)
2. Regression Discontinuity Design (RDD)
3. **Mix and match**

Mix and match of methods

Get creative:

- Mix-and-match types of methods!
- Address relevant questions with relevant techniques

- For our hypothetical example of the impact of subsidized loans on profits:
  - Randomly assigning subsidized loans is politically not feasible.
  - We use an RDD based on the scoring variable used for assessing loan applications to analyze the impact of subsidized loans on profits for the marginal candidates …
  - …and pair this approach with an RCT that randomly assigns additional consultancy services.

Summary

- Before-after and participant vs. non-participant comparisons: **not good methods for measuring causal impacts**
- **Randomized controlled trials require minimal assumptions** and provide intuitive estimates, but are not always feasible
- **Difference-in-differences and regression discontinuity** methods can provide reliable estimates of the impact of an intervention but
  - are based on hypotheses (sometimes numerous!) and
  - must be implemented with care
- The most appropriate method **depends on the context and the available data.** Often, evaluating different parts of a program will require several different strategies.
- The results of impact evaluations are only valid if we use **rigorous methods**.

European Commission

WORLD BANK GROUP