Non-Experimental Evaluation Methods: Advanced Track

Enhancing Cohesion Policy Impact for Jobs, Skills, and Competitiveness Workshop

Berlin, May 2025

Caio Piza – Senior Economist at DIME/WB

Goal of any impact evaluation

- Identify *causal effects*
- Causal effects?
  - Changes in the outcomes of interest (e.g. sales and profits) that are **exclusively** explained by the intervention (e.g. training program, access to finance etc.)

How do we establish causation in an IE?

Need to find the **counterfactual**

So we can compare

WHAT HAPPENED	WITH	WHAT WOULD HAVE HAPPENED IN THE ABSENCE OF THE INTERVENTION

Non-Experimental Methods

1. Difference-in-differences (Diff-in-Diff)
   - Diff-in-Diff with matching
   - Diff-in-Diff with staggered treatment

2. Regression discontinuity design (RDD)
   - Local randomization
   - RDD with random assignment
   - Related method: interrupted time-series (before-and-after with high frequency data)

Case: subsidized credit program was launched to ease MSMEs' access to working capital during the Covid-19 pandemic crisis.

- Eligibility criteria: registered firms with up to 10 employees
- A subset of firms in the target population applies to the credit line whereas others don't.
- Question: What's the *causal impacts* of a subsidized credit program on MSMEs' profits?

Non-Experimental Methods

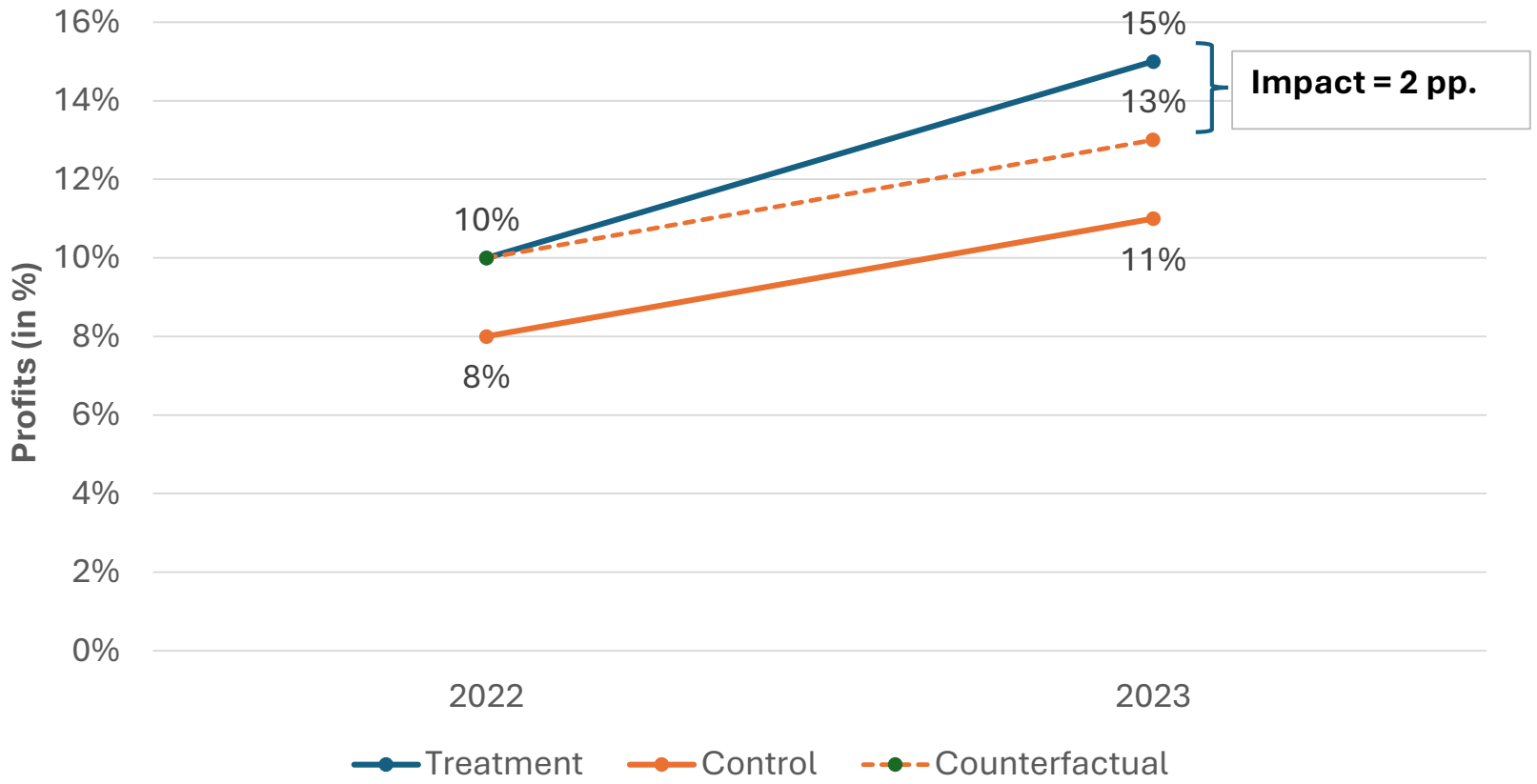
1. Difference-in-differences (Diff-in-Diff)
   - ➢Diff-in-Diff with matching
   - ➢Diff-in-Diff with staggered treatment
2. Regression discontinuity design (RDD)
   - ➢Local randomization
   - ➢RDD with random assignment
   - ➢Related method: interrupted time-series (before-and-after with high frequency data)

How can one evaluate this?
Diff-in-Diff (DiD)

- The canonical case (2x2): two groups (participants and non-participants) before and after the program.

- Key identifying assumption in DiD design: parallel trends
  - **Parallel trends**: the time trend of the comparison group's outcomes of interest informs the *counterfactual* - what would have happened to the treatment group in the absence of the treatment.
  - *The self-selection is driven by **time-invariant unobserved characteristics** of the firm (or firm fixed effects).*

Illustration of the parallel-trends assumption



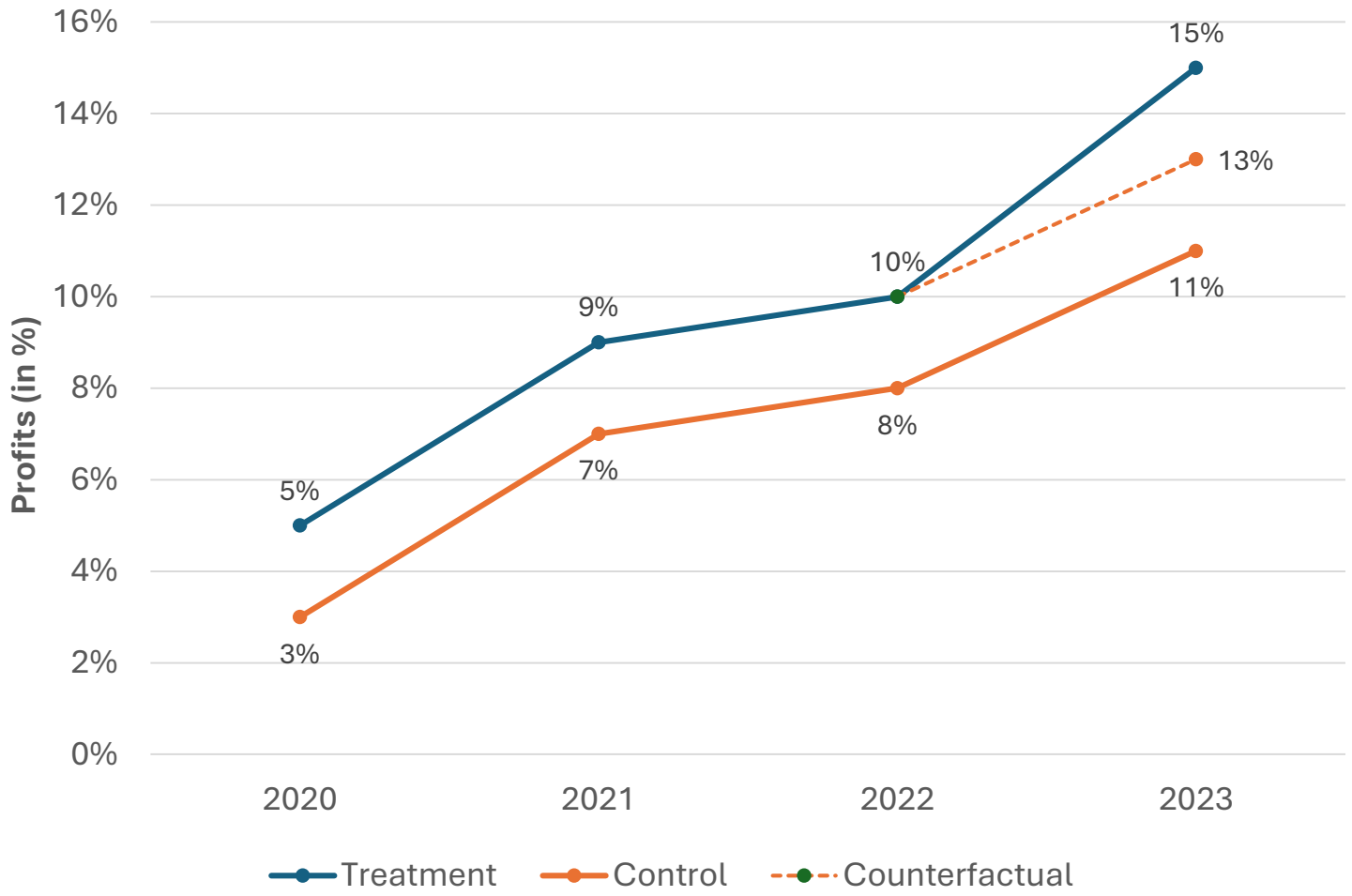
How can one evaluate this?
Diff-in-Diff (DiD)

- The canonical case (2x2): two groups (participants and non-participants) before and after the program.

- Key identifying assumption in DiD design: parallel trends
  - **Parallel trends**: the time trend of the comparison group's outcomes of interest informs the *counterfactual* - what would have happened to the treatment group in the absence of the treatment.
  - *The self-selection is driven by **time-invariant unobserved characteristics** of the firm (or firm fixed effects).*

  - Is the parallel-trend assumption plausible in the present case?
  - Is the *strict exogeneity assumption* likely to hold? In words: is the treatment assignment based on past realizations of the outcome variable?

With historical (admin) data, one can test the plausibility of the parallel-trend assumption



Non-Experimental Methods

1. Difference-in-differences (Diff-in-Diff)
   - ➢Diff-in-Diff with matching
   - ➢Diff-in-Diff with staggered treatment
2. Regression discontinuity design (RDD)
   - ➢Local randomization
   - ➢RDD with random assignment
   - ➢Related method: interrupted time-series (before-and-after with high frequency data)

Diff-in-Diff with matching

- *This is useful when historical (panel) data exist, and pre-trends **ARE NOT** parallel*

- What is the intuition of **matching techniques**?
- The intervention targets firms with characteristics one can observe (e.g., firm size, sector, firm age, sales…) – this is called ***selection in observed characteristics***.
  - Firms' participation decision is based on observed characteristics **ONLY**
  - Non-participating firms with similar observed characteristics (or conditional probability to participate) will generate a valid counterfactual.

Diff-in-Diff with matching

- *In practice…*

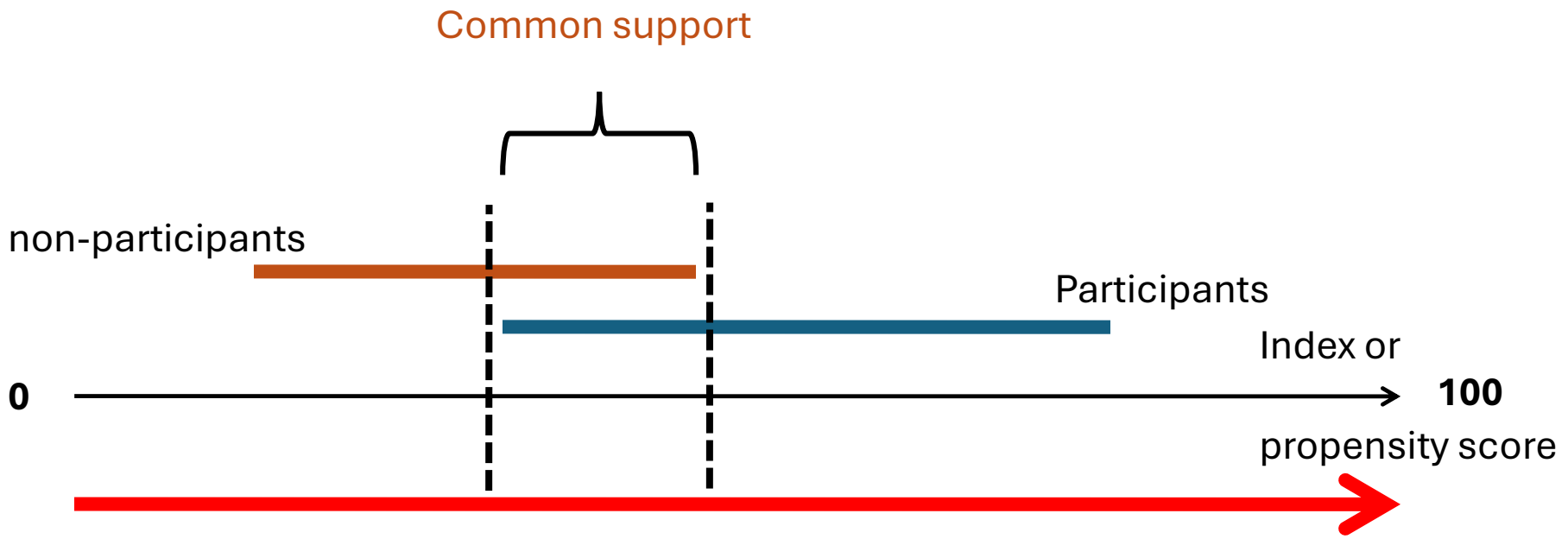
- Estimate a probability model (e.g., logit or probit) using a vector of observed characteristics of the firm. The predicted conditional probability is called *propensity score*.

- The variables included in the estimation of the propensity scores should help predict BOTH the outcome variable(s) and the participation decision.

- The propensity-score matching (PSM) technique allows one to compare outcomes of firms that have similar predicted probabilities (estimated propensity scores).

Matching…

- Illustration



In practice, the way one computes the matched sample matters to reduce bias and increase precision!

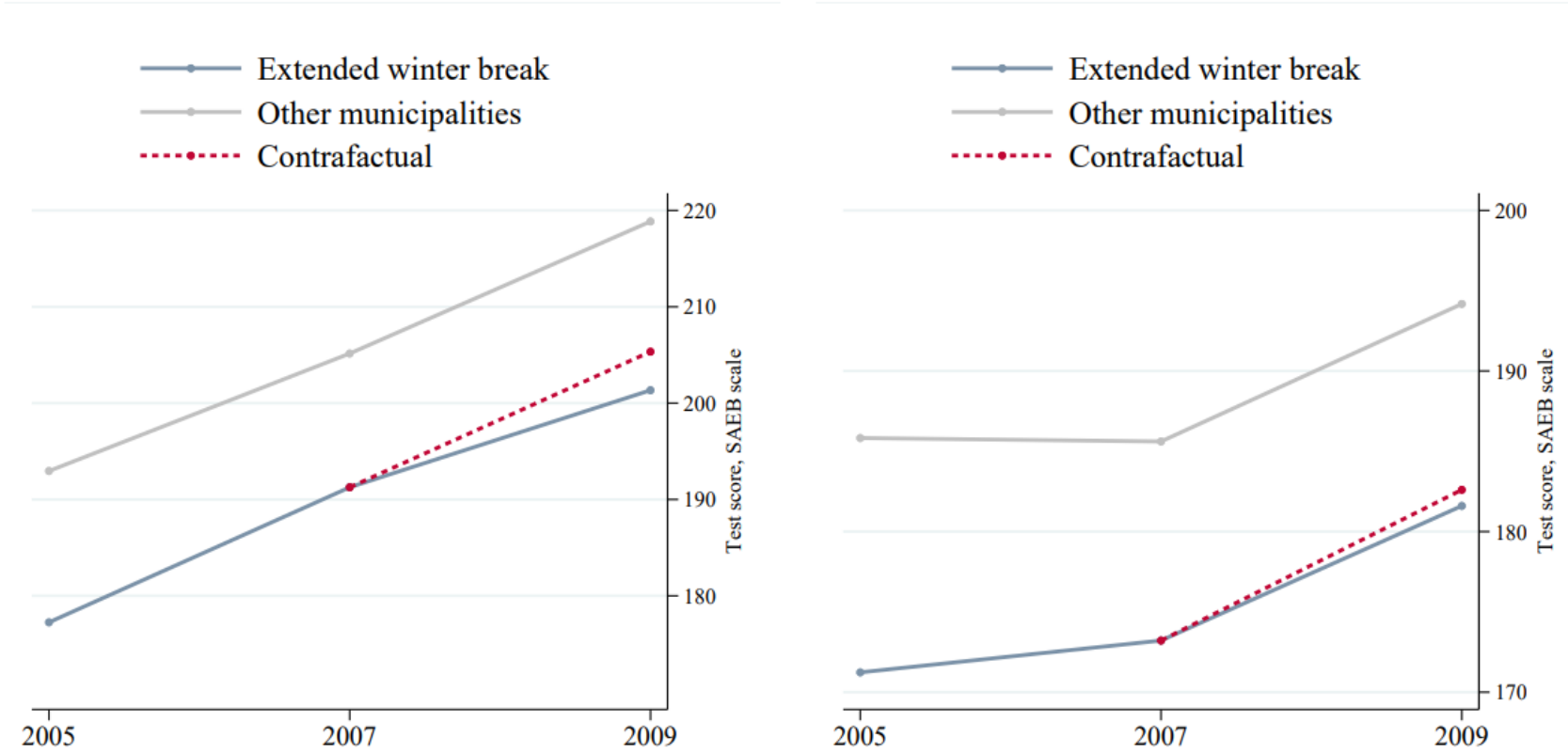
- The PSM is one way to obtain a matched sample. There are other ways, such as Inverse-probability weighting (IPW) techniques.
- Abadie (2005): IPW-DiD
- Sant'Anna and Zhao (2020): IPW-DiD and Doubly-robust DiD are superior to PSM-DiD, and DR-DiD is more efficient/precise than the IPW-DiD.

Example: *Learning when schools shutdown: impacts of H1N1 outbreak on learning loss and learning gaps (WB Policy Research Working Paper)*

- In June/2009, H1N1 was declared a pandemic by WHO.
- In Brazil, at least 55,000 people were diagnosed and 2,200 died.
- In July, São Paulo Health State Department recommended the extension of children's winter break for 2-3 weeks.
- In Brazil, state and local governments share the provision of primary and secondary education:
- Most of the 645 municipalities in São Paulo state have schools run by both state and local governments.
- The winter break was extended in all schools run by the state and in all municipal schools in 13 municipalities.
- This measure affected almost 70% of the students and over half of the public schools (51.5%).

Example: *Learning when schools shutdown: impacts of H1N1 outbreak on learning loss and learning gaps (WB Policy Research Working Paper)*

Figure 1: Students' proficiency in locally-managed schools, fifth-grade (2005-2009)



- The paper compares the learning outcomes of different cohorts of 5th graders before and after the school shutdown episode.
- The DiD compares learning outcomes in municipal schools in 13 municipalities that extended the winter break vs. municipalities that did not.

Example: *Learning when schools shutdown: impacts of H1N1 outbreak on learning loss and learning gaps (WB Policy Research Working Paper)*

- **Extension**: The paper uses a triple difference-in-differences design to leverage the within municipalities variation across state and municipal schools.
  - Group 1: municipalities where state schools closed but municipal schools didn't. (1st DiD explores variation across municipal and state school and time)
  - Group 2: municipalities where both state and municipal schools closed (2nd DiD explores variation across municipal and state schools and time - **placebo**).

- The triple difference estimates the effects taking the difference between the two DiD above. In the present case, it shows the impacts of the policy on learning in state schools.

Example: *Learning when schools shutdown: impacts of H1N1 outbreak on learning loss and learning gaps (WB Policy Research Working Paper)*

- What are the main advantages of the triple difference over the DiD in this example?!
  1. It accommodates idiosyncratic shocks at municipal level.
  2. The identification strategy relies on a weaker assumption than the standard parallel trends. *The triple diff requires is that any difference in learning trends across municipal and state schools in Group 2 would be what one would observe in Group 1 had the municipal schools closed* (the *counterfactual*).

Table 3: Impact of the school shutdowns on students' learning, fifth-grade (2007-2009)

Estimated decrease in Math and Portuguese Proficiency, SAEB scale																
	Math								Portuguese							
	DiD	DiD	DiD	TD	TD	TD	TD	TD	DiD	DiD	DiD	TD	TD	TD	TD	TD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
H1N1	-3.26**	-3.27**	-2.75**	-4.35***	-4.25***	-4.38***	-4.29***	-4.25***	-0.76	-0.78	-0.54	-3.49***	-3.40***	-3.51***	-4.29***	-2.73***
sd	(1.18)	(1.22)	(1.23)	(0.96)	(0.97)	(0.96)	(0.97)	(1.06)	(0.86)	(0.91)	(0.95)	(0.81)	(0.81)	(0.81)	(0.97)	(0.90)
Wild-bootstrap p-value	0.0400	0.0260	0.0410	0.0000	0.0000	0.0000	0.0000	0.0001	0.3920	0.4150	0.6060	0.0000	0.0000	0.0000	0.0000	0.0025
95% CI	[-5.6,-0.9]	[-5.7,-0.9]	[-5.2,-0.3]	[-6.2,-2.5]	[-6.1,-2.4]	[-6.3,-2.5]	[-6.2,-2.4]	[-6.3,-2.2]	[-2.5,0.9]	[-2.6,1.0]	[-2.4,1.3]	[-5.1,-1.9]	[-5.0,-1.8]	[-5.1,-1.9]	[-6.2,-2.4]	[-4.5,-1.0]
N. schools	3912	3912	3912	5329	5329	5329	5329	5329	3912	3912	3912	5329	5329	5329	5329	5329
Adj. R2	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.6
Proficiency - Treatment Group before the school shutdowns (2007)																
Mean	193.38	193.38	193.38	195.3	195.3	195.3	195.3	195.3	175.44	175.44	175.44	177.88	177.88	177.88	195.3	177.88
Sd	15.44	15.44	15.44	16.09	16.09	16.09	16.09	16.09	15.4	15.4	15.4	15.46	15.46	15.46	16.09	15.46
ATT est (in sd)	-0.21	-0.21	-0.18	-0.27	-0.26	-0.27	-0.27	-0.26	-0.05	-0.05	-0.03	-0.23	-0.22	-0.23	-0.27	-0.18

Non-Experimental Methods

1. Difference-in-differences (Diff-in-Diff)
   - ➢Diff-in-Diff with matching
   - ➢Diff-in-Diff with staggered treatment

2. Regression discontinuity design (RDD)
   - Local randomization
   - RDD with random assignment
   - Related method: interrupted time-series (before-and-after with high frequency data)

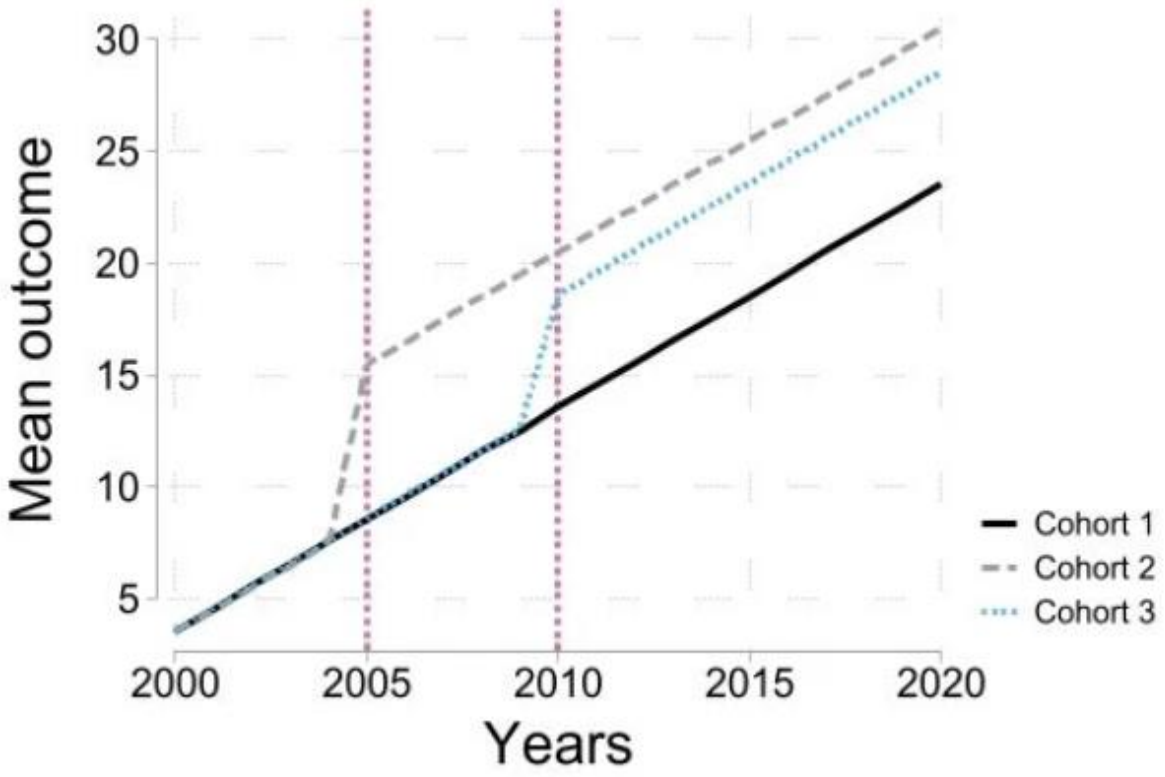
Diff-in-Diff with Staggered Design

- The canonical DiD design works well when all firms targeted by an intervention join the program at the same time.

- However, there are several policies with a non-random rollout, i.e., participating firms enter the program in different points in time (e.g., every week, month, quarter etc.). In these cases, the canonical DiD can lead to biased estimates because some treated units are used as comparison group.

- This scenario creates challenges one needs to consider when using the DiD design:
  - (i) what's the proper comparison group? and
  - (ii) how to account for dynamic effects (e.g., treatment effects vary across cohorts and/or treatment effects grows over time).

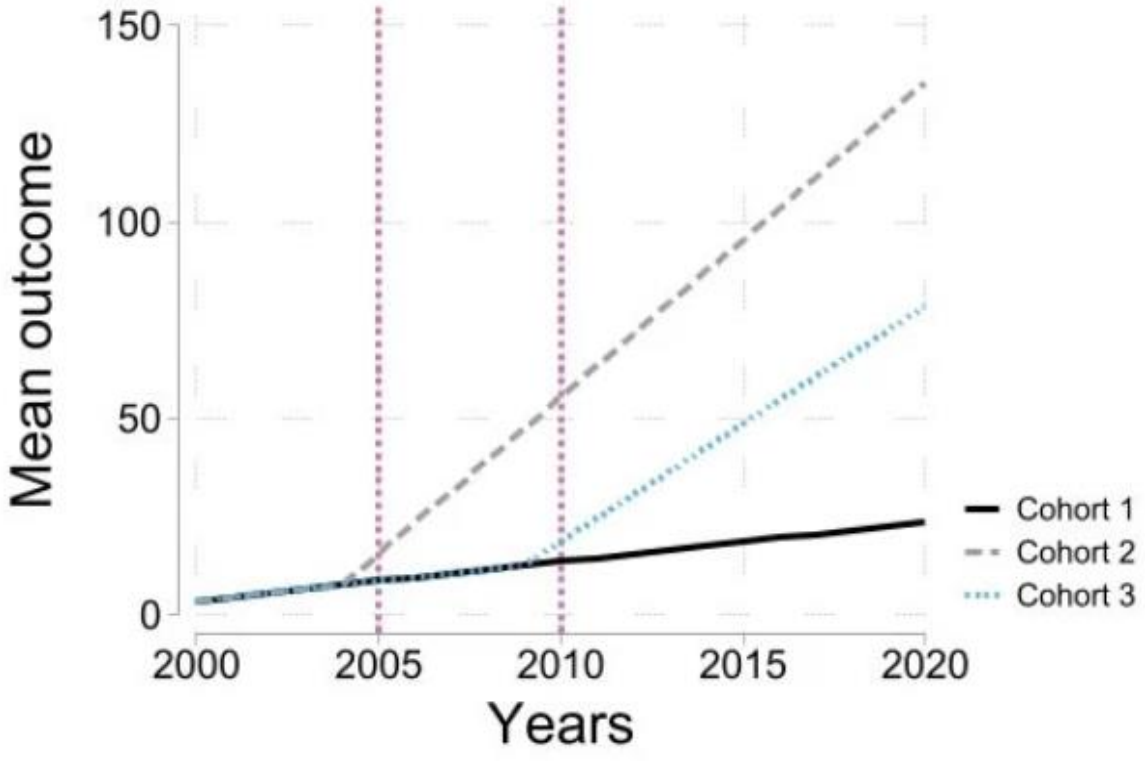
Staggered DiD with constrant treatment effects



- Notice that once the treatment kicks in, the outcome variable of the treatment group changes levels but not the slope. However, **the treatment effect is not the same across treated cohorts**.

- In this case, the DiD can be estimated comparing different cohorts:
- Cohorts 2 vs. 1 before and after 2005
- Cohorts 1 vs. 3 before and after 2010
- Cohorts 2 vs. 3 before and after 2005 (but up to 2010).

Staggered DiD with dynamic treatment effects



- Once the treatment kicks in, the outcome variable of the treatment group changes the slope but not the level (the treatment effect grows over time). As before, the **treatment effect** is not constant across treated cohorts.

- Similarly, the DiD can be estimated comparing different cohorts:
- Cohorts 2 vs. 1 before and after 2005
- Cohorts 1 vs. 3 before and after 2010
- Cohorts 2 vs. 3 before and after 2005 (but up to 2010).

Diff-in-Diff with Staggered Design

- Wooldridge (2021) shows that the standard pooled OLS (or RE estimator) can be specified flexibly to account for both types of treatment effects heterogeneities.

  '... *there is nothing inherently wrong with TWFE as an estimation method. The problem is that is it often applied to a model that is too restrictive.*' (Wooldridge, 2021: p.34)

- For a review, check Roth et al. (2023) *What's trending in difference-in-differences? A synthesis of the recent econometrics literature?*

- *Good starting point:* https://blogs.worldbank.org/en/impactevaluations/new-synthesis-and-key-lessons-recent-difference-differences-literature

Non-Experimental Methods

1. Difference-in-differences (Diff-in-Diff)
   - ➢Diff-in-Diff with matching
   - ➢Diff-in-Diff with staggered treatment

2. Regression discontinuity design (RDD)
   - ➢Local randomization
   - ➢RDD with random assignment
   - ➢Related method: interrupted time-series (before-and-after with high frequency data)

RDD

- Powerful method if there exists:
  - ✓ A continuous eligibility index: the running variable should be smoothly distributed around the threshold (e.g., McCrary density test)
    - ➢ An imperfect manipulation of the running variable (by the applicants)
  - ✓ A clear-cut (**arbitrary**) eligibility cut-off so that **the observed (an unobserved!) characteristics are smoothly distributed around the threshold**
    - ➢ Run RD regressions using X variables as dependent variable.

Important: The causal estimates are valid only for those subjects/firms close to the cut-off point.

Regression Discontinuity Design

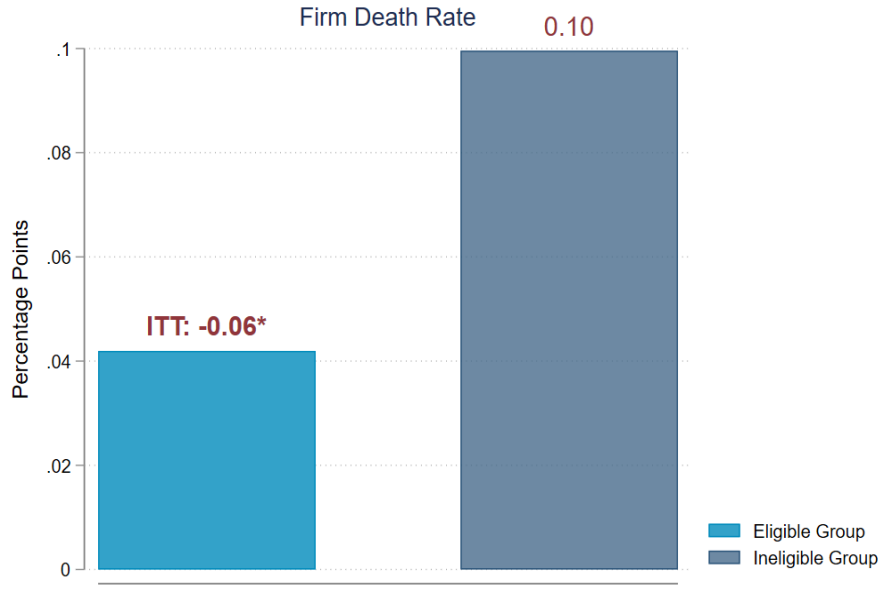
- Sharp RD design (or sharp RDD): participation is a deterministic function of the eligibility rule. The probability of treatment jumps from 0 to 1 around the threshold.
  - ITT = ATT.

- Fuzzy RD design (or fuzzy RDD): participation is a probabilistic function of the eligibility rule
  - ITT < LATE (local average treatment effect) – similar to IV (2SLS)

- Implication:
  - sharp RD designs are more powerful (more precision) and consequently requires smaller sample sizes.
  - Fuzzy RDD: similar to a ‘local experiment’ with imperfect compliance.

Example: Impact of a subsidized credit line during Covid-19: a fuzzy RD design

- Program launched in May 2020 in São Paulo state.
- Firms above a credit score threshold would be eligible to borrow.
- Registered firms with credit score above 500 were eligible to borrow.



Example: Subsidized credit line during Covid-19

- Important steps to estimate impacts:
  1. Check if the running var is smoothly distributed around the threshold.
  2. Define the bandwidth size to estimate impacts.
     - There are different algorithms to select the optimal bandwidth. I personally like Imbens and Kalyanaraman (2012) – minimizes the MSE.
     - If the sample size in the optimal bandwidth size is relatively small, one may want to use a wider window size. However, the wider the bandwidth size the more biased the estimate is likely to be. *Trade-off between bias and variance*.
  3. Check if the covariates are balanced around the threshold using the selected bandwidth size.
  4. Estimate the treatment effects (ITT and/or LATE) parametrically or non-parametrically.

Example 1: Out of the Shadows? Revisiting the Impact of the Brazilian SIMPLES Program on Firms' Formalization Rates (Piza, JDE 2018)

- The Program: a tax reform in Nov 1996.
- The system combined 6 different federal taxes and one social contribution into one monthly-based rate
- Two eligibility criteria:
  1. Annual revenue (different thresholds for micro and small firms)
  2. Sectors: retail trade, manufacturing, transportation, civil construction and other services that do not require a professional with a regulated occupation

Example 1: Out of the Shadows? Revisiting the Impact of the Brazilian SIMPLES Program on Firms' Formalization Rates (Piza, JDE 2018)

Figure 1 – Proportion of formal firms before and after SIMPLES

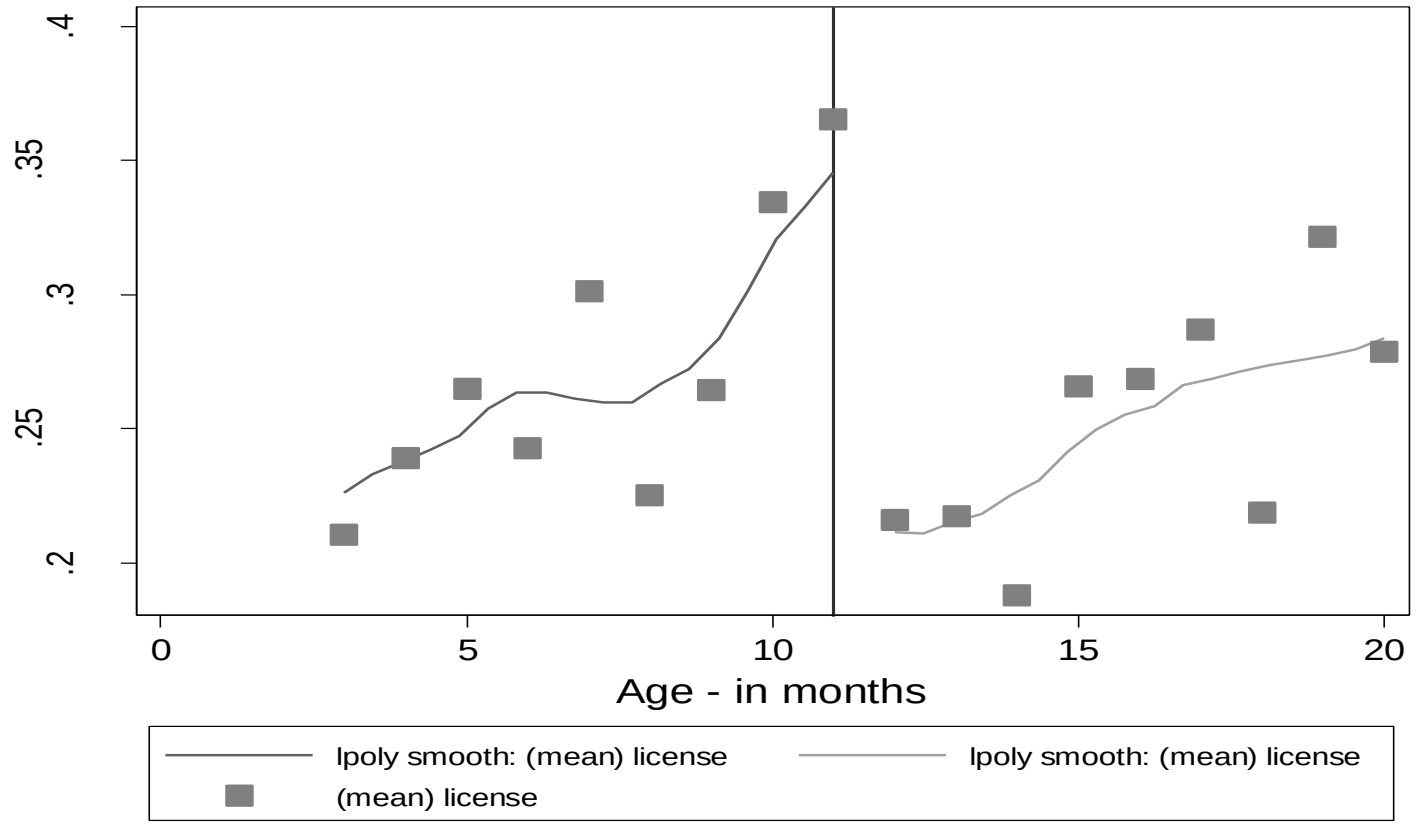


Table 1
First-stage regression: Effect of SIMPLES on formalization rates.

RD Estimates	3-Month	4-Month	5-Month	6-Month	7-Month	8-Month
Cutoff = November 1996 (original)						
D	0.13***	0.099**	0.099**	0.11***	0.10***	0.12***
	(2.63)	(2.3)	(2.57)	(3.03)	(3.27)	(3.92)
N	1399	1664	2012	2315	2860	3236

Note: ***, **, * Statistically significant at 1, 5, and 10 percent. These are spline linear regressions in which y is regressed on a constant, a dummy that is 1 for firms created after the threshold and 0 otherwise (D), the assignment variable defined in months (Z) and an interaction term between the "after" dummy and the assignment variable (DZ) to allow for different trends in each side of the threshold. T-statistics in parentheses with standard errors clustered at forcing variable level. Estimates obtained with linear probability model.

Validity Tests of the RD Design

McCrary density test for the manipulation of the assignment variable (*time in business*)

Figure 3 – McCrary Density Test for the Manipulation of the Forcing Variable (cutoff = Nov 1996)

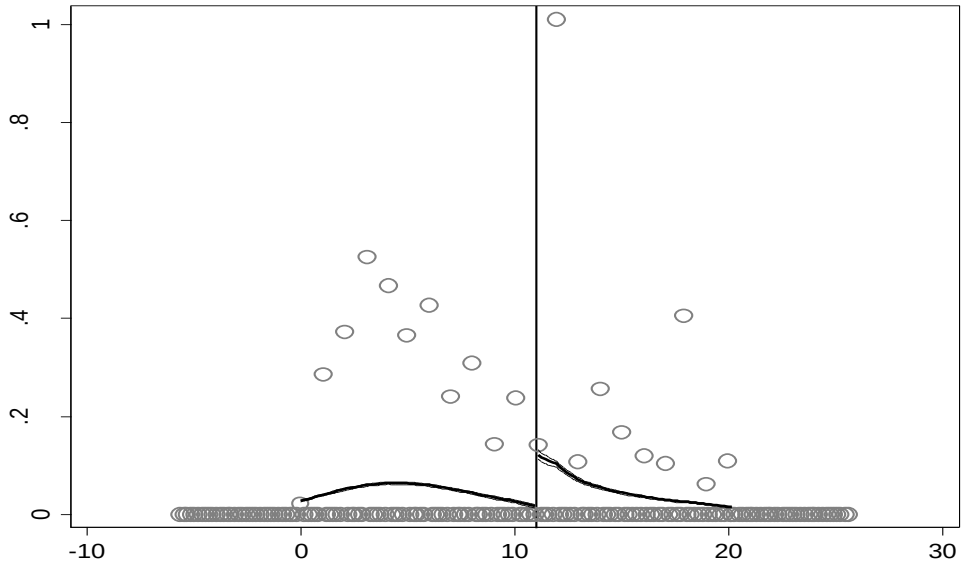
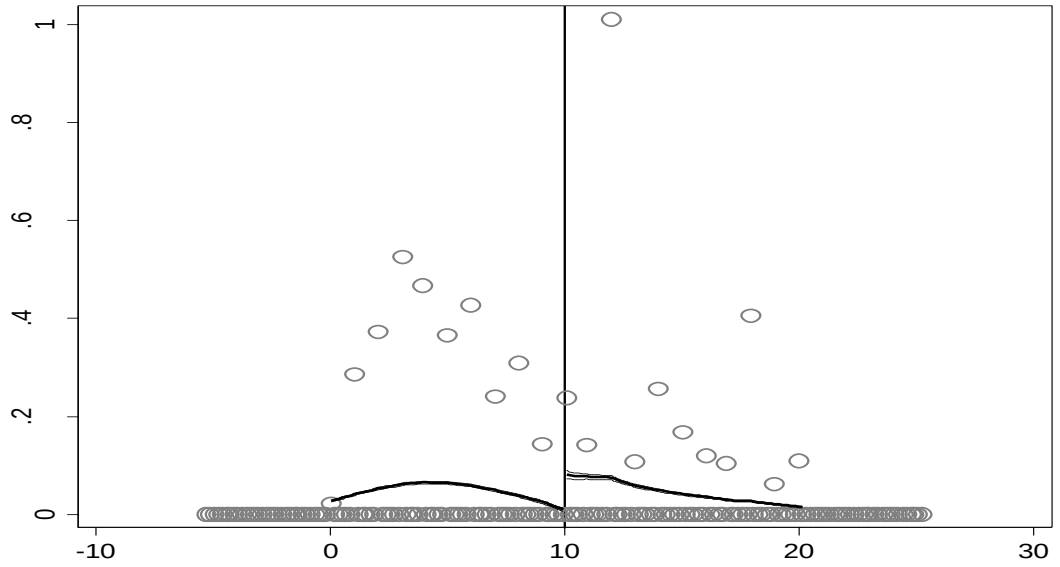


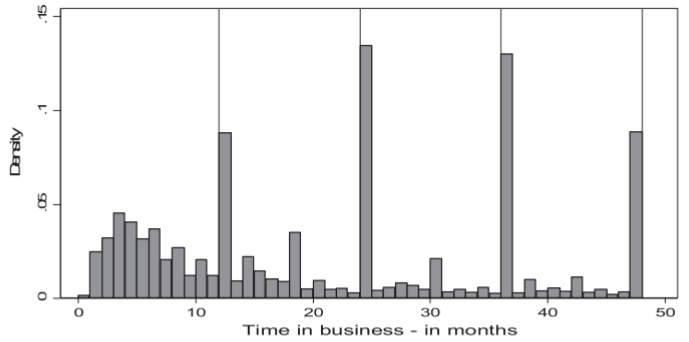
Figure 4 – McCrary Density Test for the Manipulation of the Forcing Variable (cutoff = Dec 1996)



Validity Tests of the RD Design

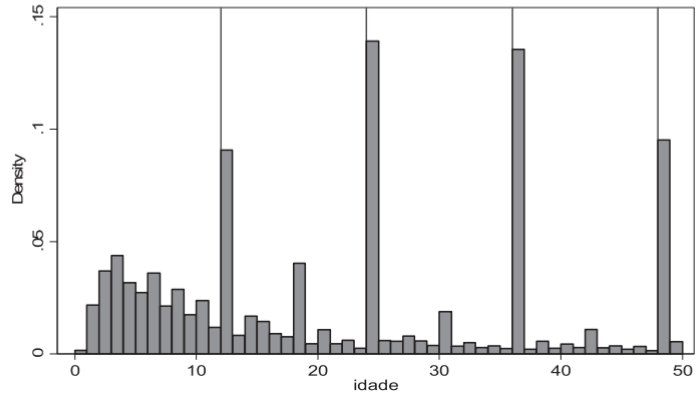
- This example shows that the running variable IS NOT continuously distributed around the threshold due to manipulation or rounding.
- In cases like this, the standard RDD estimates won't be valid.
- What to do?

a



Note: The reference lines indicate firms 12, 24, 36 and 48 months in business respectively.

b



Note: The reference lines indicate firms 12, 24, 36 and 48 months in business respectively.

Validity Tests of the RD Design

- *Data heaping in RD Designs* (Barreca et al. 2015)
  - If the heap in the forcing var is not random – i.e., it predicts the outcome of interest – then the RD estimates will be biased
  - One ad-hoc way of dealing with the bias ('donut-RD approach'): drop the obs in the vicinity of the cutoff point -- the heap data
  - This approach is useful to check robustness as '*it has the potential to highlight misspecification in any RD design*.' (Barreca et al. 2015, p. 8)
  - A more efficient approach is to use a dummy variable for the heaps and add the dummies in the regressions as controls.

Table 1
First-stage regression: Effect of SIMPLES on formalization rates.

RD Estimates	3-Month	4-Month	5-Month	6-Month	7-Month	8-Month
Cutoff = November 1996 (original)						
D	0.13***	0.099**	0.099**	0.11***	0.10***	0.12***
	(2.63)	(2.3)	(2.57)	(3.03)	(3.27)	(3.92)
N	1399	1664	2012	2315	2860	3236
Cutoff = November 1996 (original but excluding firms created in Oct 1996)						
D	0.059	0.13	0.11*	0.079	0.076*	0.095**
	(1.48)	(1.80)	(2.10)	(1.60)	(1.84)	(2.34)
N	744	1009	1357	1660	2205	2581
Cutoff = November 1996 (original with a dummy for firms created in Oct 1996)row						
D	0.059	0.13	0.11*	0.079	0.076*	0.095**
	(1.50)	(1.82)	(2.11)	(1.60)	(1.85)	(2.35)
N	1399	1664	2012	2315	2860	3236

Note: ***, **, * Statistically significant at 1, 5, and 10 percent. These are spline linear regressions in which y is regressed on a constant, a dummy that is 1 for firms created after the threshold and 0 otherwise (D), the assignment variable defined in months (Z) and an interaction term between the "after" dummy and the assignment variable (DZ) to allow for different trends in each side of the threshold. T-statistics in parentheses with standard errors clustered at forcing variable level. Estimates obtained with linear probability model.

Non-Experimental Methods

1. Difference-in-differences (Diff-in-Diff)
   - ➢Diff-in-Diff with matching
   - ➢Diff-in-Diff with staggered treatment

2. Regression discontinuity design (RDD)
   - Local randomization
   - RDD with random assignment
   - Related method: interrupted time-series (before-and-after with high frequency data)

Example of Local Randomization

- The main difference between the standard RDD (continuity-based approach) and the local randomization design is the selection mechanism.

- The LR design estimate treatment effects using a simple difference in means using the smallest bandwidth size possible.

JOURNAL ARTICLE

The Short- and Longer-Term Effects of a Child Labor Ban Get access >

Caio Piza ✉, André Portela Souza ✉, Patrick M Emerson ✉, Vivian Amorim ✉

The World Bank Economic Review, Volume 38, Issue 2, May 2024, Pages 351–370, https://doi.org/10.1093/wber/lhad036

Published: 08 November 2023 **Article history** ▾

- The bandwidth is selected based on a series a balance tests. The test begins with a narrow bandwidth size and stops when balance is no longer observed. The optimal bandwidth size is the largest bandwidth size that ensures a balanced sample.

Non-Experimental Methods

1. Difference-in-differences (Diff-in-Diff)
   - ➢Diff-in-Diff with matching
   - ➢Diff-in-Diff with staggered treatment

2. Regression discontinuity design (RDD)
   - ➢Local randomization
   - ➢RDD with random assignment
   - ➢Related method: interrupted time-series (before-and-after with high frequency data)

RDD with random assignment (based on Karlan and Zimmerman, 2010)

- Example: A credit program is offered to firms with credit score >50 points (the points range from 0-100).

- Say that 10,000 firms apply to the credit line and 6,000 score above 50. Among the 4,000 with score < = 50, say that 800 have a credit score in the interval [45-50].

- If the lender is willing to expand access to credit but is concerned with the risk of doing so, it could expand the program at the margin.

RDD with random assignment

- Example: The credit program is offered to firms with credit score >50 points (the points range from 0-100).

- Expansion of credit at the margin: randomly assign 300/800 to become eligible to borrow and then assess their repayment rates.

- RCT sample: score = [45, 49]: 300 vs. 500 – impact of credit on marginally riskier firms
- RDD 1 sample: score = [45, 55] – local impact of credit on firms
- RDD 2 sample: 300 vs. firms with score = [50, 55] – impact of credit on repayment rates (useful analysis for the lender)

Takeaways

- Before-and-after AND participants vs. non-participants: **not good methods to measure causal impacts**

- Diff-in-diff and RDD can provide reliable estimates for the impact of an intervention but
  - Rely on (sometimes strong) assumptions;
  - Require historical (admin) data and definitely more data than any RCT.
  - Need to be carefully implemented, particularly in a prospective evaluation.