United Nations University
World Institute for Development Economics Research

Session 9A:
Ex-Post Evaluation of TEs II – Concepts & Methods

Amina Ebrahim

18 September 2023 | Accra, Ghana

Follow-up Technical Meeting on Tax Expenditures

UNU WIDER

Monitoring & Evaluation Fundamentals

Monitoring	Evaluation
Ongoing data collection	Time-specific
Internal	Can be internal or external
Informs ‘real-time’ decision making	Is the policy effective, impactful, efficient, relevant?
	Informs current and future policy-making

UNU WIDER

Impact evaluation

1. Studies the existence and magnitude of impact
2. Aims to identify the causal relation
3. Uses scientific methods

- Impact evaluation is an essential tool for evidence-based policy-making
- Evidence-based policymaking means using empirical knowledge and research supported facts in planning and implementing policies
- Help tracing the **accountability**, **efficiency** and **impact** of policies
- Results can be useful for Redesigning, comparing and improving policies

UNU WIDER

Basic components of impact evaluation

Question(s) to be answered

Outcome of interest

Outcomes we are interested in investigating → Impact indicators

Evaluation design

Refers to the type of evaluation methodology that is used

UNU WIDER

Components of impact evaluation:

1. Questions to be answered

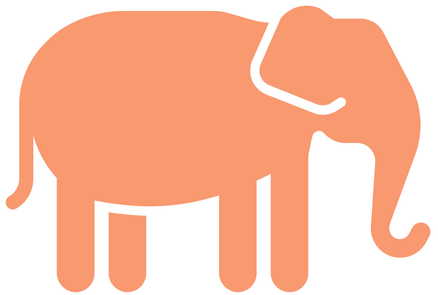
- Depend on the policy design and information needs.
- Examples:
  - Is the policy change increasing tax revenue?
  - What are the distributional impacts of the tax schedule?
  - Is the tax deduction impacting investment?

UNU WIDER

Components of impact evaluation

2. Outcomes of interest

- Outcomes we are interested in investigating → Impact indicators
- Examples:
  - Number of registered taxpayers
  - Amount of tax revenue
  - Sales → business activity
  - Number of formal employees



UNU WIDER

Components of impact evaluation

3. Evaluation design

- Impact evaluation uses scientific methods (Statistics and Econometrics) to estimate average impacts.

UNU WIDER

What is an indicator?

To evaluate policy = to measure change.

We need indicators

Indicators are signals of change

- What are we measuring

Examples: Taxpayer registration campaign

- More taxpayers → increase in taxpayer registration
- More taxpayers → increase in revenue collection

An indicator is a quantitative or qualitative factor or variable

- provides a simple and reliable means to measure achievement or to reflect the changes
- In principle, an indicator is anything that you can measure, and that reflects a change

Quantitative indicators

Measures of quantity common in research conducted with tax data

Expressed as numbers (units, proportions, rates of change, %, prices etc)

Information that can be counted and aggregated

UNU WIDER

What can evaluations tell us?

- Impact evaluations are used to verify the **existence** and *magnitude* of impacts of a particular intervention on outcomes of interest (impact indicators).
- Focuses on identifying causal relations between Tax Incentive implementation and changes in outcomes of interest.
- In the absence of the project, such changes would not have been observed.
- Scientific methods (statistics and econometrics) are used to estimate the average impacts of the TE.

How do we use the results?

Impact evaluation results can be useful inputs for:

- Redesigning and/or improving a TE: Innovations, scaling up, shutting down
- Comparing alternative versions of a TE (which % is best?)
- Comparing alternative TE (is there a better TE?)

UNU WIDER

Limitations of non-rigorous but commonly used methods: An example

Consider the following situation:

- Program: Introduction of a new Tax Incentive for small firms
- Objective: increasing productivity (reduce credit constraints), increase tax compliance
- Question to be answered: How did the TE impact productivity and revenue collection?

Some commonly used methods to answer the question “what was the impact of the TE on productivity?” are:

- Compare TE beneficiaries before and after the TE start;
- Compare TE beneficiaries to non-beneficiaries that did not receive the TE.

Are these methods capable of estimating the impact of the program on productivity?

Limitations of non-rigorous but commonly used methods

Consider possibility 1: Compare TE beneficiaries before and after the program.

Suppose that:

- Before: Average productivity of TE beneficiaries was 5 days/month producing a leather bag;
- After: Average productivity of TE beneficiaries was 3 days/month producing a leather bag;

We can say that there was an **increase** in productivity among beneficiaries over time, but we cannot guarantee that this increase was exclusively due to the TE:

- Other simultaneous factors may have contributed to this observed productivity increase (e.g. new soft leather introduced, reduced electricity cuts, etc).

Limitations of non-rigorous but commonly used methods

Consider possibility 2: Compare program beneficiaries to non-beneficiaries who did not apply for the program.

Suppose that:

- TE beneficiaries: Average productivity was 5 days/month producing a leather bag;
- TE non-beneficiaries: Average productivity was 5 days/month producing a leather bag;

We can say that non-beneficiaries are the same, but we cannot guarantee that the program had *zero or negative impact*;

- Non-beneficiaries may be different from beneficiaries in relevant ways (e.g. higher productivity before, different firm sizes, different access to technologies and inputs).

Limitations of non-rigorous but commonly used methods

Possibility 1. Comparing program beneficiaries before and after the program will only give us the impact if productivity was not affected by any *other factors* besides the intervention.

- This is rare in practice.

Possibility 2. Comparing program beneficiaries to non-beneficiaries will only give us the program impact if non-beneficiaries are equal to beneficiaries except for intervention participation.

- We must have information about beneficiaries and non-beneficiaries to investigate whether this comparison is adequate.
- It is not always possible to guarantee that these two groups are similar in all relevant characteristics, as some might not be observable and will be correlated to both program participation and outcomes (e.g. firm's motivation, firm size, firm sector).

Causality

Impact evaluations are used to identify ***causal relationships*** between TE participation and changes in outcomes of interest.

- Main question: what is the impact (causal effect) of TE participation on the outcome of interest?

***Causality* implies that:**

- We can attribute observed changes in the outcome of interest to TE participation.
- In the absence of the TE, such changes in the outcome of interest would not have been observed.

More about causality

To attribute changes in the outcome of interest to TE participation, we would like to observe participants in ***two different states*** (having participated and not having participated) ***at the same moment in time*** (after the intervention).

The counterfactual corresponds to the outcome of interest that would have been observed for participants if they had not participated in the TE.

The impact of participation would be given by comparing these two states.

- This comparison would eliminate all other factors that might have affected the outcome of interest so that any observed difference could be interpreted as the project impact.

What is the Counterfactual?

Example: to estimate the impact of a Wage subsidy program on firm employment we would like to compare beneficiary firm in two different states after the intervention has taken place:

Impact**:**

Firm employment if participated

Real

vs.

Firm employment if no participation

Not real

Counterfactual

UNU WIDER

However, *it is impossible to observe two different states simultaneously*.

Selection of a comparison group

To estimate the counterfactual, we will use a ***comparison group*** that mimics what would have happened to participants without the TE.

An adequate comparison group is a key element for any impact evaluation!

UNU WIDER

Selection of a comparison group

The ideal comparison group is composed by non-participants with the same characteristics of participants, except for TE participation.

- **Analogy:** clones.
- If participants and non-participants in the comparison group differ only in terms of TE participation, then any observed difference in the outcome of interest between them after the intervention can be attributed to project participation, i.e., the impact.

The comparison group is usually also called the ***control group***.

UNU WIDER

Selection of a comparison group

To estimate the impact, we will always compare two groups:

- ***Treatment group***: composed of TE participants;

- ***Control group***: composed of non-participants that are as similar as possible to participants, except for TE participation. They should not have been affected by the intervention, neither directly nor indirectly.

Selection of a comparison group

Treatment and control groups should be similar in at least 3 ways:

1. On average, characteristics of treatment and control groups should be the same;
2. Treatment and control groups should react to the TE in the same way;
3. Treatment and control groups cannot be differentially exposed to other interventions during the evaluation period.

If these 3 conditions are met, any differences in the outcome of interest between the two groups can be attributed to project participation, i.e. can be interpreted as the causal impact of the TE.

If these conditions are not met, impact estimation will be *biased*:

Difference between groups = impact + other factors (e.g. previous differences)

Bias

The true impact of project participation will not be estimated.

UNU WIDER

Selection of a comparison group

In any impact evaluation, we will have to select an ***adequate*** *control group*.

- With an adequate control group, we can *eliminate selection bias* and correctly estimate the impacts.

- In the next modules of the course, we will explore different possibilities for selecting a control group and for impact estimation.
  In all cases, the main point will be the elimination of the selection bias to estimate the causal effects of interventions.

How can we select an adequate control group?

- The *key* to control group selection will be ***participation selection criteria***.
- As we will discuss next, participation selection criteria will determine possible control groups and the most adequate method to estimate impacts.

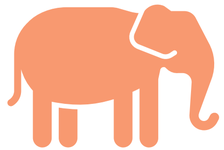
UNU WIDER

Difference-in-Differences: When to Use It

- In many cases, the assignment rules of the program are not as clear or as well defined as discussed in the previous approaches.

- If that is the case, approaches such as experimental evaluation, regression discontinuity design, and instrumental variables will not be feasible.

- If baseline **data** is available, one possibility is to use *difference-in-differences*.



Source: Gertler et al (2011).

UNU WIDER

Difference-in-Differences: When to Use It

The ***difference-in-differences*** approach requires stronger assumptions and is only feasible if baseline data is available for both participants and non-participants.

- Baseline data must be from a pre-intervention period.

This method does not require specifying treatment assignment rules.

In this method, we compare the changes in outcomes (impact indicators) over time (before and after) between a treatment group (participants of the intervention) and a control group (non-participants of the intervention).

- This is the reason why the method is called *difference-in-differences*.

UNU WIDER

Difference-in-Differences: Example

Consider the following situation

Program: Introducing a new Income Tax rule

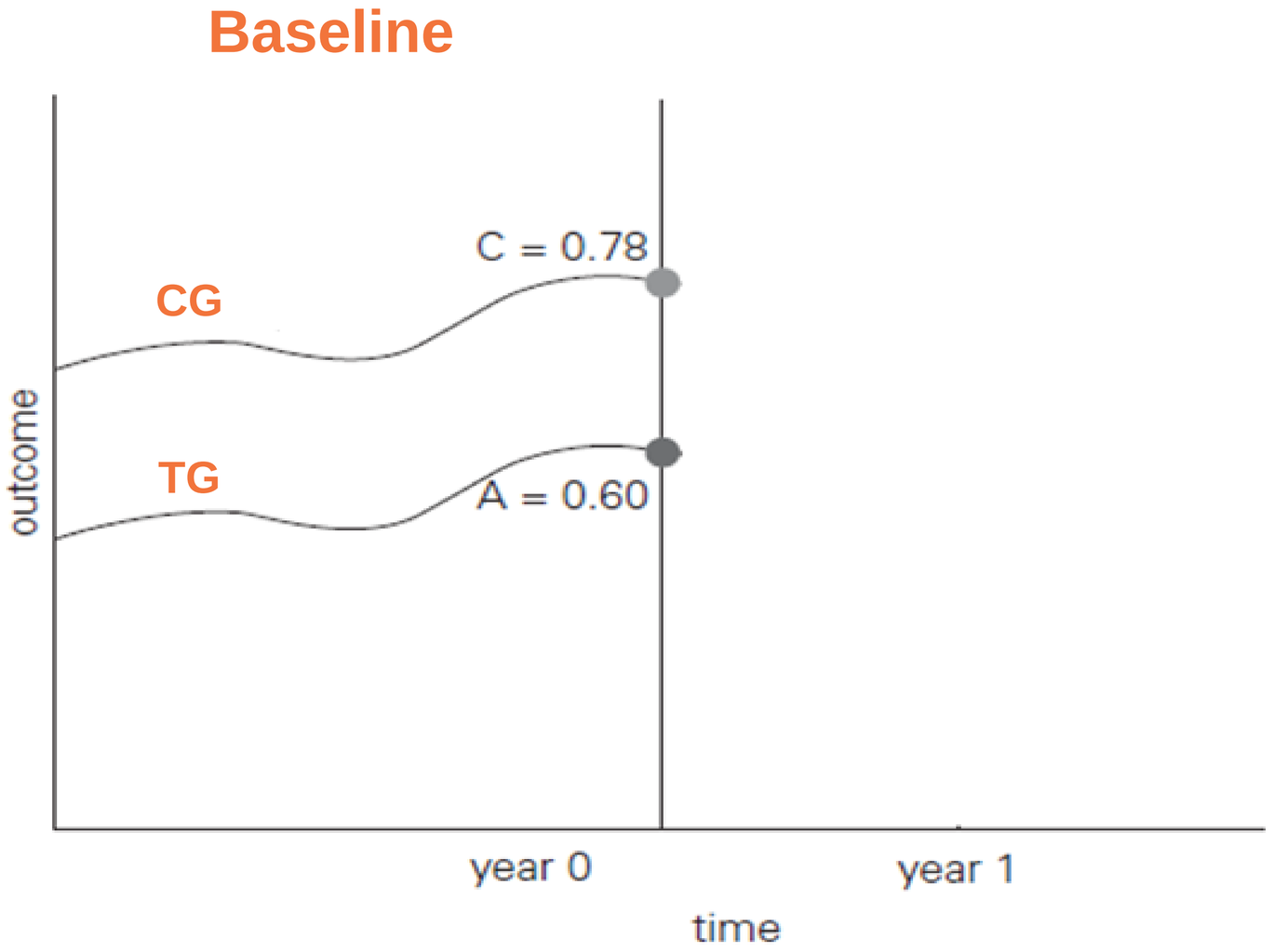
Objective: increasing government revenue by increasing tax rates for richer individuals

Eligible units: Individuals with income different from zero

Selection to participate: Income higher than a specific threshold

Impact evaluation objective: estimating the impact of the program on government revenues

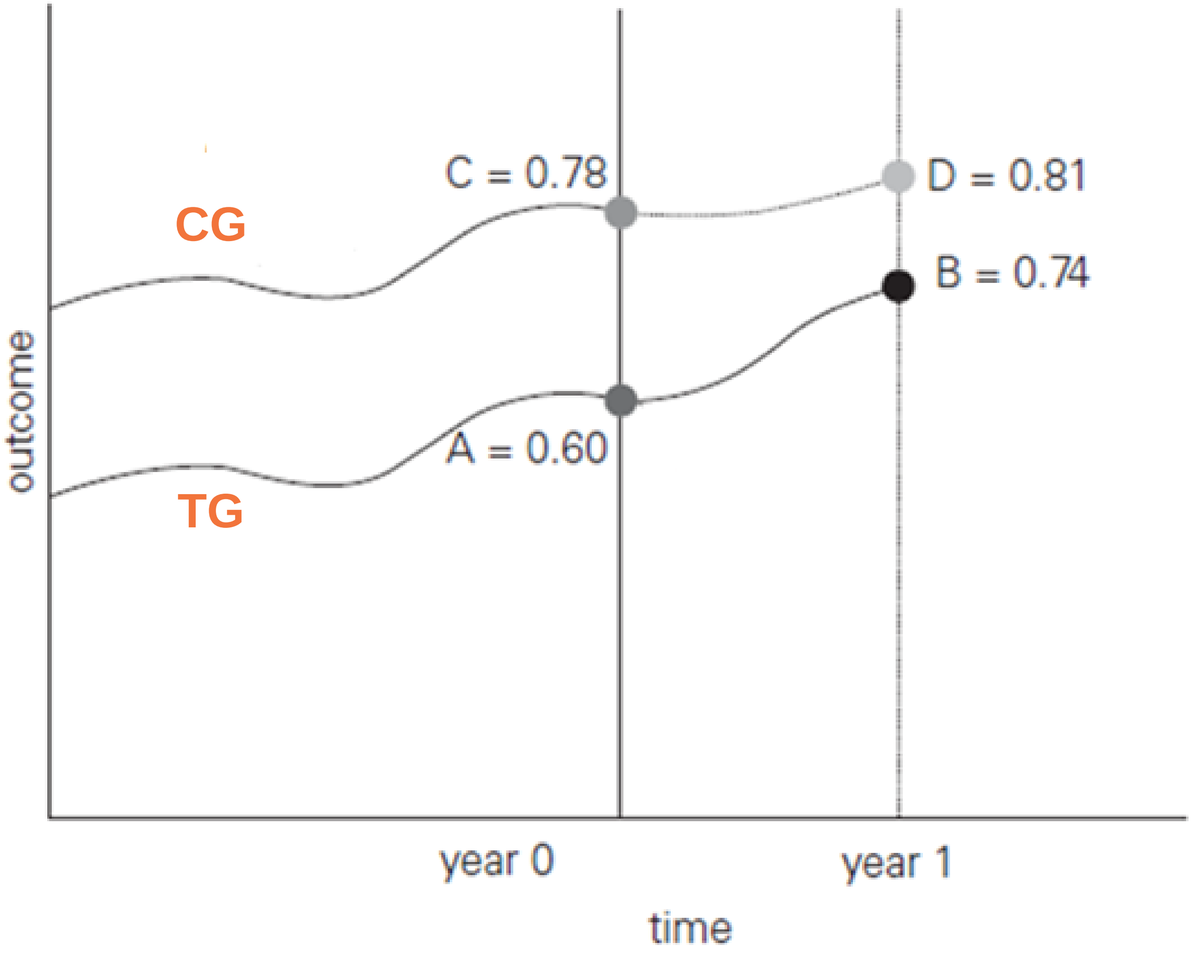
Difference-in-Differences: Example



Source: adapted from Gertler et al (2011).

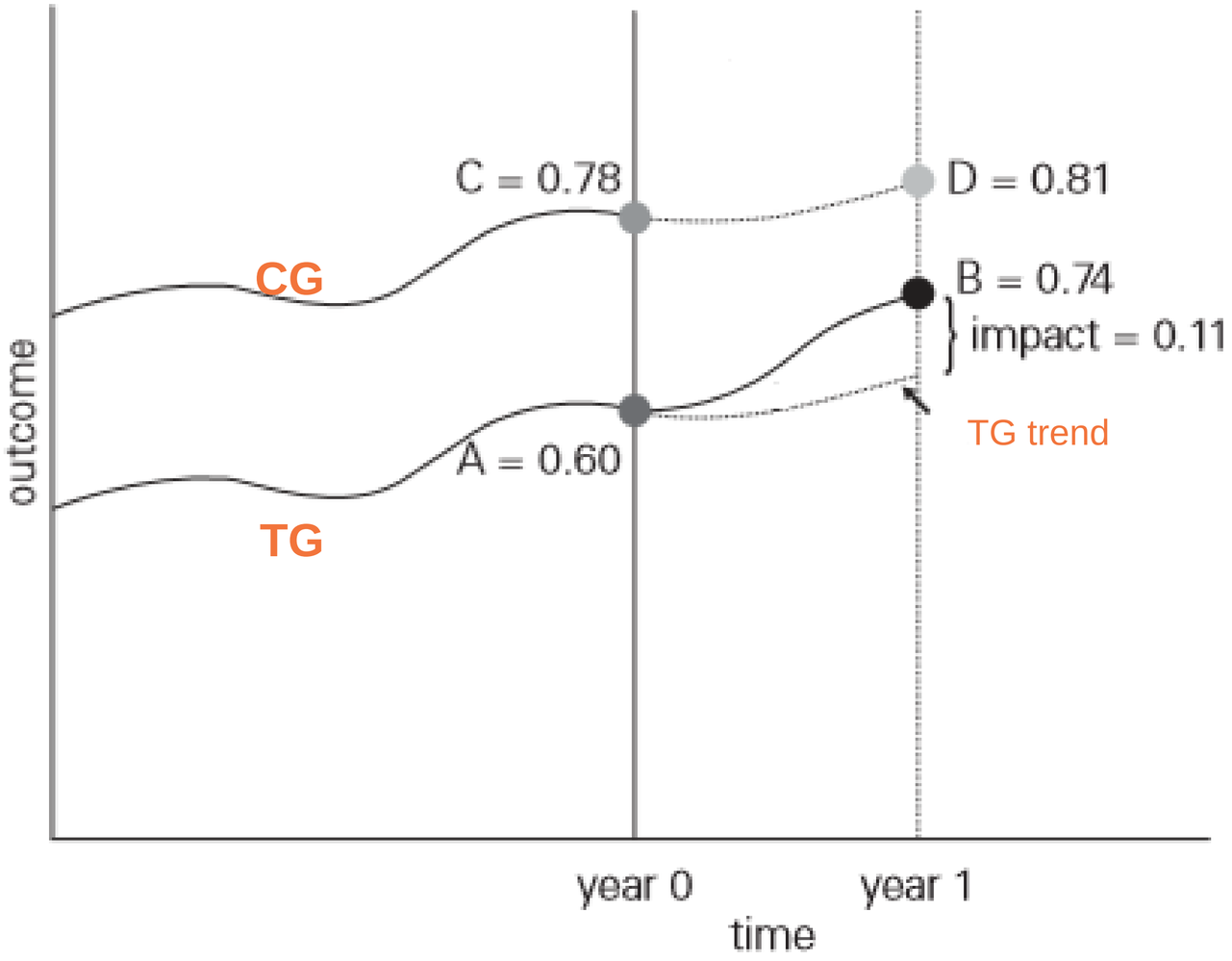
Difference-in-Differences: Example

Baseline and Endline



Source: adapted from Gertler et al (2011).

Difference in Difference: Example calculation



Impact calculation

Impact = (D-B) – (C-A)
= (0.81-0.74) – (0.78-0.60)
= 0.14 – 0.03
= 0.11

Positive Effect!

UNU WIDER

Difference-in-Differences: Assumptions

1

In the absence of the intervention, outcomes of treatment and control groups would display equal trends

If true, changes in outcomes over time for the control group represent the changes in outcomes that would have been observed for the treatment group in the absence of the intervention.

2

Treatment and control groups should not be affected in different ways by other factors at the same time as the intervention.

If this does not hold, impact estimation may be biased.

3

Group composition (treatment and control) should not change significantly over time (before and after the intervention).

There cannot be too much missing, otherwise, impact estimation may be biased.

UNU WIDER

Difference-in-Differences: Checking the validity

There is no way to test assumptions 1 and 2, but some common validity checks are:

- Check for equality of trends between groups before the TE introduction:

  This requires at least two observations before the intervention for both groups;

- Conduct “placebo” tests:
  - Using a fake treatment group that we are sure was not affected by the intervention;
  - Using a fake outcome variable that we are sure was not affected by the intervention;
  - Using different control groups to test the sensibility of the results.

Difference-in-Differences: Limitations

- The equal trends assumption must hold for the estimated impact to be unbiased. It is a strong assumption and its validity cannot be tested.
- If other simultaneous factors are present and affect the difference in trends between treatment and control groups, impact estimation using this approach will be invalid or biased (Gertler et al, 2010).

Recap

- Need to be clear about the indicators/outcomes
- Difference in Differences can offer a rigorous evaluation method
- Important to consider the comparison group carefully.

UNU WIDER

Thank you!

Amina Ebrahim
Research Fellow | UNU-WIDER
amina@wider.unu.edu

UNU WIDER