The Evaluation Problem
Introduction

Jean-Noël Senne

Université Paris-Saclay

Problématiques Contemporaines du Développement
M2EIED
Université de Pau et des Pays de l'Adour
2023 - 2024

Who am I ?

Jean-Noël Senne

- Associate Professor, University Paris-Saclay
- Research Associate, IRD-DIAL

Research interests

- Development Economics, Migration, Education, Health, Africa
- Microeconomics, Applied Econometrics

Contact

- Mail : jean-noel.senne@universite-paris-saclay.fr
- Web : https://sites.google.com/site/jeannoelsenne/

What should you know ?

Organization

- 20 hours
  - Session 1 : Fri. 01/01 (4 hours) ; Mon. 05/02 (2,5 hours)
  - Session 2 : Mer. 21/02 (2 hours) ; Thu. 22/02 (2 hours) ; Fri. 23/02 (4 hours)
  - Session 3 : Fri. 16/03 (4 hours)

Evaluation

- Presentation of an article

Material

- Lecture slides + references

Textbooks

- Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics. Princeton university press.
- Angrist, J. D., & Pischke, J. S. (2014). Mastering'metrics: The path from cause to effect. Princeton university press.
- Givord, P. (2014). Méthodes économétriques pour l'évaluation de politiques publiques.. Economie prevision, (1), 1-28.

Online course

- Mastering Econometrics, Marginal Revolution University (by J. Angrist)

What are we going to do together ?

We'll cover various and popular techniques in econometrics for the analysis of (impact) evaluation, focusing not only on why they work, but also on the data and assumptions they require

1. *Chapter 1*. Introduction to Evaluation
2. *Chapter 2*. Randomized Controlled Trials (RCT)
3. *Chapter 3*. Matching Models (MM)
4. *Chapter 4*. Advanced Instrumental Variables (IV)
5. *Chapter 5*. Difference-in-Differences (DID)
6. *Chapter 6*. Regression Discontinuity Design (RDD)

Plan

What is impact evaluation ?

What makes evaluation an (econometric) problem ?

What empirical methods can solve this problem ?

What does evaluation mean ?

Impact evaluation seeks to answer **2 different questions** :

1. **What happens to a group of people who are affected by a particular "common experience" ?**
   - Public policy interventions (e.g trainings, labor costs, class size, school building, vaccination...)
   - Shocks to living conditions (e.g price shocks, natural disasters, conflicts...)
   - Particular choices (e.g dropping out of school, marital arrangements, investment in a business...)

⇒ You need a *tracer study*

2. **Does this "common experience" change things (and by how much) ?**
   - Does training help people get a job ?
   - Do price shocks decrease employment ?
   - Does longer education improve earnings ?

⇒ You need an *impact evaluation method*

⇒ You want to know *ex post* whether the **changes in outcomes** were **causally due** to the experience (or whether they would have happened anyway)

Why do we evaluate ?

You want to measure the **causal impact** of an intervention (or a shock/choice)
$\Rightarrow$ **Theory of change** (anticipation of the effects)

$$\underbrace{\text{Inputs} \rightarrow \text{Activities} \rightarrow \text{Outputs}}_{\text{Intervention}(X)} \longrightarrow \underbrace{\text{Intermediate outcomes} \rightarrow \text{Final outcome}}_{\text{Results}(Y)}$$

But why ?

- To improve knowledge (on various economic, social, political issues)
- To mesure the effectiveness of a public policies
- To inform policy makers
- To design (or improve) public policies

Why is evaluation difficult ?

You want to answer **conterfactual questions** about an *alternative* scenario you don't observe...

- *Would people be employed if they had not been trained ?*
- *Would employment be higher if prices were lower ?*
- *Would people earn less if they had not gone to school ?*

⇒ The purpose of all evaluation methods is to *'mimic'* this **unobservable conterfactual**...

⇒ ... dealing with standard **econometric issues** to identify causal effects :

- Confounding factors
- **Selection** and **endogeneity** bias

⇒ and particular attention to :

- **The assumptions** (and data) required for robust identification
- **The heterogeneity** of the effects

Do hospitals make people healthier ?

Your health status is : excellent, very good, fair, or poor ?

	Hospital	**No Hospital**	**Difference**
Health status	3.21	3.93	-0.72^{***}
	(0.014)	(0.003)	
Observations	7,774	90,049	

A simple comparison of means suggests that going to the hospital makes people worse off....

$\Rightarrow$ What's wrong ?

Plan

What is impact evaluation ?

What makes evaluation an (econometric) problem ?

What empirical methods can solve this problem ?

The evaluation problem - Rubin's causal model

- You want to evaluate the causal effect of a **treatment (T)** on some **outcome (Y)** that may be impacted by the treatment
- Each individual i can get the treatment ($T_i = 1$) or not ($T_i = 0$)
- For each individual i, there are **2 potential outcomes** :
  - Y_{1i} = value of i's outcome if she **does** get the treatment
  - Y_{0i} = value of i's outcome if she **doesn't** get the treatment

⇒ **The causal effect** of the treatment is :

$$\Delta_i = Y_{1i} - Y_{0i}$$

⇒ **Fundamental problem of causal inference** : you never observe both potential outcomes for the same individual !

⇒ **Identification problem** (many identification issues can be thought of this way !)

The evaluation problem - Rubin's causal model

- Indeed, each individual i is either treated or untreated
- For each individual i, you only observe the **realised outcome** Y_i :

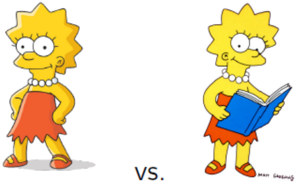
$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & \text{if} \quad T_i = 1 \quad \textit{(treated)} \\ Y_{0i} & \text{if} \quad T_i = 0 \quad \textit{(untreated)} \end{cases} \\ &= Y_{0i}(1 - T_i) + Y_{1i} T_i \\ &= Y_{i0} + \underbrace{(Y_{1i} - Y_{0i})}_{\text{impact}} T_i \end{aligned}$$

$\Rightarrow$ **Missing counterfactual** data problem !

$\Rightarrow$ What is the **right counterfactual ?** $\rightarrow$ i.e the outcome that would have been observed without (or with) treatment

The ideal counterfactual

- What is the impact of giving Lisa a textbook ?



- In an ideal world (for researchers), you would clone the *treated* Lisa

⇒ Impact = Lisa's score with a book - Lisa clone's score without a book

- In the real world, you either observe Lisa with a book or without...

⇒ What is a **relevant counterfactual** for Lisa ?

False counterfactuals

2 types of 'naive' but (typically) **wrong counterfactuals** :

1. **Pre-treatment vs. post-treatment** comparisons

$\rightarrow$ (before/after treatment)

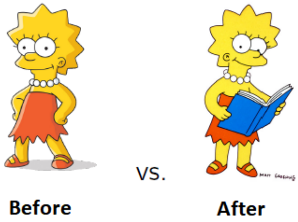
2. **Treated vs. non-treated** comparisons

$\rightarrow$ (with/without treatment)

$\Rightarrow$ Extremely strong (and often unreasonable...) assumptions are required for these impact evaluation approaches to be credible

False counterfactual 1 : Before/After

- You could compare Lisa's score **before and after** giving her a book



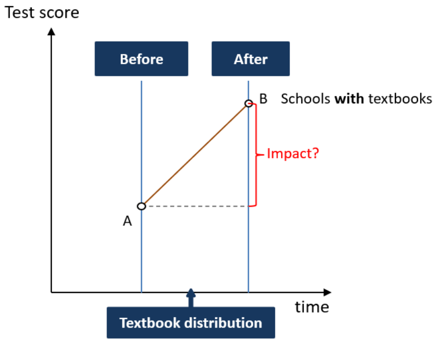
⇒ Impact = Lisa's score before - Lisa's score after

⇒ But this naive estimator is likely to be **biased !**

- Maybe Lisa's score would have improved anyway **(time trend)**
- Or other **counfounding factors** have made Lisa's score improved

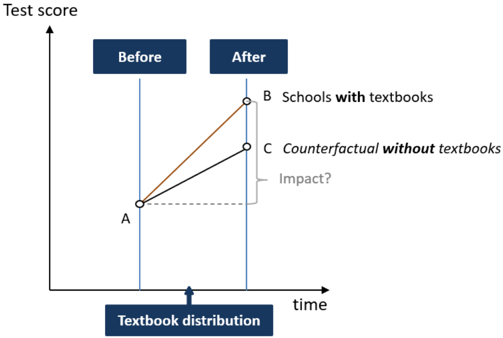
False counterfactual 1 : Before/After

- You want to evaluate the impact of a textbook distribution program on school performance
- You have data on schools before and after the implementation of the program
- You compare pupils' test scores **before and after the program**



⇒ Is this the causal impact of the program ?

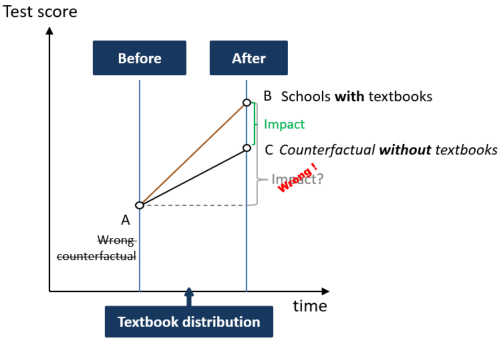
False counterfactual 1 : Before/After



⇒ Yes if there is no time trend and/or nothing (other than textbook distribution) happened over the period, but...

⇒ ... what if test scores would have improved anyway ? Or have improved (at least partly) for other reasons that occured during the same period ?

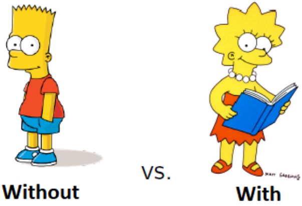
False counterfactual 1 : Before/After



⇒ If you could observe the *right* counterfactual (what would have been the test score without textbooks), you would conclude that the causal impact of the programme is lower

False counterfactual 2 : With/Without

- You could compare Lisa's score with a book to another child's score without a book



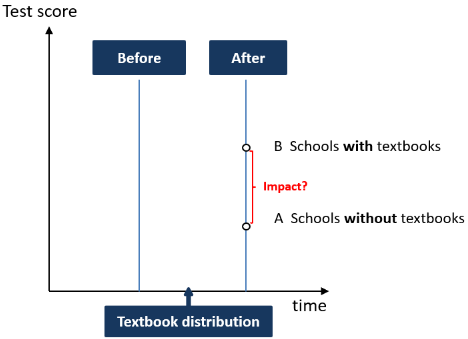
$\Rightarrow$ Impact = Lisa's score with a book - Bart's score without a book

$\Rightarrow$ But this naive estimator is also likely to be **biased !**

- Certainly Lisa's score would have been better than Bart's score **even without a book**
- Maybe Lisa **expect higher benefits** than Bart from reading books

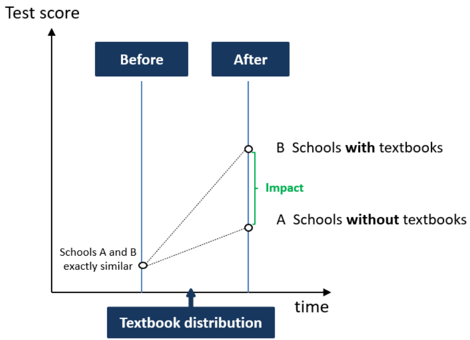
False counterfactual 2 : With/Without

- You compare pupils' test score between schools with and without a book



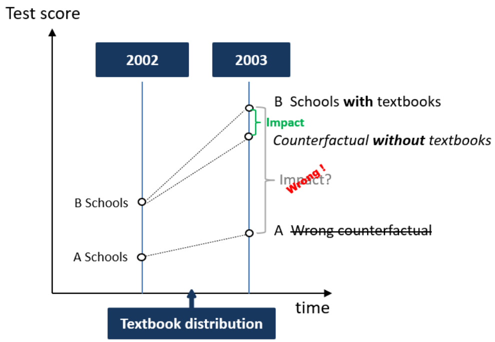
⇒ Is this the causal impact of the program ?

False counterfactual 2 : With/Without



⇒ Yes if schools with/without textbooks were initially similar, but...

False counterfactual 2 : With/Without



⇒ ... but what if schools with textbooks had initially better test scores than school without textbooks ? Or if schools with textbooks expected higher gains from the program ?

⇒ If you could observe the *right* counterfactual (what would have been the test score without textbooks), you would conclude that the causal impact of the programme is smaller

Evaluation parameters of interest

- You actually want to estimate the **average causal effect** of the treatment
- 2 evaluation parameters in particular :
  1. **ATT** *(Average Treatment Effect on the Treated)*

     $$ATT = E(Y_{1i} - Y_{0i} | T_i = 1)$$

  2. **ATE** *(Average Treatment Effect)*

     $$ATE = E(Y_{1i} - Y_{0i})$$

- But you only observe...

  $$E(Y_{1i} | T_i = 1)$$
  $$E(Y_{0i} | T_i = 0)$$

- ... and not counterfactuals

  $$E(Y_{1i} | T_i = 0)$$
  $$E(Y_{0i} | T_i = 1)$$

$\Rightarrow$ **Identification problem**

Selection bias

- The 'naive' estimator (i.e the difference in *observed* means between treated and untreated) is typically a **biased estimator of ATT** :

$$\textbf{Difference in means} = E(Y_i|T_i = 1) - E(Y_i|T_i = 0)$$
$$= E(Y_{1i}|T_i = 1) - E(Y_{0i}|T_i = 0)$$

- Adding in $\underbrace{-E(Y_{0i}|T_i = 1) + E(Y_{0i}|T_i = 1)}_{=0}$, we get :

Difference in means

$$= \underbrace{E(Y_{1i}|T_i = 1) - E(Y_{0i}|T_i = 1)}_{\text{ATT}} + \underbrace{E(Y_{0i}|T_i = 1) - E(Y_{0i}|T_i = 0)}_{\text{Selection bias}}$$

- **Why is this bias likely ?**
  - There may be systematic differences between treated and untreated individuals, *even in the absence of treatment*
  - Comparative advantages : individuals choose to be treated when the expect the treatment to make them better off (i.e $T_i = 1$ if $Y_{1i} - Y_{0i} > c$, where c is the cost)

Identification of ATT

Independance assumption

The potential outcome Y_0 is independant of treatment assignment :

$$Y_0 \perp T$$

$$E(Y_0|T=1) = E(Y_0|T=0) = E(Y_0)$$

- Treated are similar (on average) to untreated = **No selection**

$$\begin{aligned} ATT &= \overbrace{E(Y_{1i}|T_i=1)}^{\text{Observed}} - \overbrace{E(Y_{0i}|T_i=1)}^{\text{Unobserved}} \\ &= \overbrace{E(Y_{1i}|T_i=1)}^{\text{Observed}} - \overbrace{E(Y_{0i}|T_i=0)}^{\text{Observed}} \\ &= \textbf{Difference in group means} \end{aligned}$$

⇒ The counterfactual for the treated group is the *observed* outcome for the untreated group

Identification of ATE

(Stronger) Independance assumption

Both potential outcomes (Y_0, Y_1) are independant of treatment assignment :

$$(Y_0, Y_1) \perp T$$

$$E(Y_0|T=1) = E(Y_0|T=0) = E(Y_0)$$
$$E(Y_1|T=1) = E(Y_1|T=0) = E(Y_1)$$

- Treated are similar (on average) to untreated = **No selection** + Treatment effect is similar = **Homogenous treatment effect**

$$\begin{aligned} ATE &= \overbrace{E(Y_{1i})}^{\text{Unobserved}} - \overbrace{E(Y_{0i})}^{\text{Unobserved}} \\ &= \overbrace{E(Y_{1i}|T_i=1)}^{\text{Observed}} - \overbrace{E(Y_{0i}|T_i=0)}^{\text{Observed}} \\ &= \textbf{Difference in group means } (= ATT) \end{aligned}$$

$\Rightarrow$ The counterfactuals for both groups are the *observed* outcome for the other group

Selection as an endogeneity issue

- Each potential outcome is a random variable (specific to each person) :

$$Y_{0i} = \beta_0 + \epsilon_{0i}$$
$$Y_{1i} = (\beta_0 + \epsilon_{0i}) + \beta_1$$

- Simple linear regression model :

$$\begin{aligned} Y_i &= Y_{i0} + (Y_{1i} - Y_{0i})T_i \\ &= \beta_0 + \beta_1 T_i + \epsilon_{0i} \end{aligned}$$

- OLS estimation of the treatment effect (β_1) :

$$\begin{aligned} \beta_{1OLS} &= E(Y_i|T_i = 1) - E(Y_i|T_i = 0) \\ &= \beta_1 + \underbrace{E(\epsilon_{0i}|T_i = 1) - E(\epsilon_{0i}|T_i = 0)}_{\text{endogeneity bias}} \end{aligned}$$

$\Rightarrow$ **Selection = endogeneity** :

$$\begin{aligned} & E(\epsilon_{0i}|T_i = 1) \neq E(\epsilon_{0i}|T_i = 0) \\ \Leftrightarrow \quad & E(Y_{0i}|T_i = 1) \neq E(Y_{0i}|T_i = 0) \end{aligned}$$

Additional issues : heterogenous treatment effects

- Each potential outcome is heterogenous (specific to each person) ...
- ... but the treatment effect Δ_i can also be **heterogenous**
- Let's define a more general model :

$$Y_{0i} = g_0(x_i) + \epsilon_{0i}$$
$$Y_{1i} = g_1(x_i) + \epsilon_{1i}$$

- Linear regression model :

$$\begin{aligned} Y_i &= Y_{i0} + (Y_{1i} - Y_{0i})T_i \\ &= g_0(x_i) + \underbrace{((g_1(x_i) - g_0(x_i)) + (\epsilon_{1i} - \epsilon_{0i}))}_{\text{heterogenous treatment effect}} T_i + \epsilon_{0i} \end{aligned}$$

⇒ The coefficient is individual-specific

⇒ Treatment effect $(Y_{1i} - Y_{0i})$ is **heterogenous** if treated an untreated do not have the same distribution... :
  - For x = **Observable heterogenity**
  - For ϵ = **Unobservable heterogenity**

⇒ **In general, ATE $\neq$ ATT**

(Note : Restrictions $\epsilon_{1i} = \epsilon_{0i}$ and $g_1(x_i) = g_0(x_i) + \beta_1$ leads to homogenous treatment effects, i.e. $ATE = ATT = \beta_1$. But this a strong assumption...)

Additional issues : internal validity

- **Internal validity** relates to the capacity of drawing causal inference from your estimation (i.e. the estimated impact can be arguably attributed to the treatment)

→ **Is the treatment (as good as) random ?**

- The **Stable Unit Treatment Value Assumption (SUTVA)**

  "The potential outcomes for any unit do not vary with the treatments assigned to other units" (Imbens & Rubin (2015))

- Threats to internal validity :
  - Existence of **spillovers**
  - Mix-up of treated and control groups
  - Imperfect randomization or compliance
  - Hawthorne and John Henry effects
  - **Unbalanced attrition** (between treated and control groups)
  - **Power issues** (in small samples)

Additional issues : external validity

- **External validity** relates to the capacity of extrapolating and generalizing results to other contexts (populations, periods, countries, etc.)

$\rightarrow$ **Is the sample random ?**

$\rightarrow$ How much can I learn from a single study ?

$\rightarrow$ How much can I learn without a model ?

- Threats to external validity
  - Non-representative sample
  - **Heterogenous treatment effects**
  - Contextual effects
  - Experiment specificity
  - **General equilibrium** effects

Plan

What is impact evaluation ?

What makes evaluation an (econometric) problem ?

What empirical methods can solve this problem ?

Building a counterfactual

- A robust (empirical) evaluation method should provide an answer to :
  - Selection issues (priority)
  - Heterogeneity issues (if possible)

  ... with peculiar attention to the internal/external validity of the results

- The choice of the relevant method will depend on :
  - The type of question
  - The data at hand
  - The assumptions they require

$\Rightarrow$ **Common feature : you need to build or find a counterfactual**, i.e a comparison or control group which has no systematic difference and is unaffected by the treatment

"The furious five" methods (Angrist)

- 2 big types of empirical evaluation methods :
  - **Experimental methods** *(treatment is random)*
    Idea : you randomly assign the treatment to create a control group which 'mimics' the counterfactual scenario
    → Randomized Controlled Trials **(RCT)** *(lecture 2)*
  - **Non-experimental methods** *(treatment is as-good-as-random)* (Natural experiments or Quasi-experiments)
    Idea : you argue that an (already existing) control group 'mimics' the counterfactual scenario
    → Matching models **(MM)** *(lecture 2)*
    → (Advanced) Instrumental Variables **(IV)** *(lecture 3)*
    → Difference-In-Differences **(DID)** *(lecture 4)*
    → Regression Discontinuity **(RDD)** *(lecture 5)*

Randomized Controlled Trials (RCT)

- Often considered as the *"golden standard"* for evaluation methods or as the **experimental ideal**
- This method is based on a **random assignment into treatment**
  $\rightarrow$ The treatment is *inherently* independant of the potential outcomes
- **Identification assumption :** (Strong) independance assumption

$$(Y_0, Y_1) \perp T$$

$\Rightarrow$ The **"naive" difference in means** between treated an untreated units is an unbiased estimator of the causal effect of interest :

$$\begin{aligned}
&E(Y_i|T_i = 1) - E(Y_i|T_i = 0) \\
=&E(Y_{1i}|T_i = 1) - E(Y_{0i}|T_i = 0) = ATT \\
=&E(Y_{1i}) - E(Y_{0i}) = ATE
\end{aligned}$$

Matching Models (MM)

- Control for **observable differences** between treated and untreated units $\simeq$ Regression
- This method is based on a **matching of each treated unit with a untreated "twin"** with similar observable characteristics
  $\rightarrow$ Conditional on a (large) set of X observable characteristics, the treatment is *as-good-as-random* (independant of potential outcomes)
- **Identification assumption :** Conditional Independence Assumption (CIA)

$$(Y_0, Y_1) \perp T | X$$

$\Rightarrow$ The **matched difference in means** between treated and untreated units is an unbiased estimator of the causal effect of interest

Intrumental variables (IV)

- Use an **exogenous source of variation in treatment** which generates a quasi-experimental scenario

- This method is based on a set of **exogenous variables** Z **(instruments)** which determine the treatment assignment, but are independent of the unobserved component of the potential outcomes
  $\rightarrow$ Based on Z, the treatment is *as-good-as-random*

- **Identification assumptions :** Relevance and excludability of the instruments

$$(\epsilon_0, \epsilon_1) \perp Z$$

  1. Relevance condition : $cov(Z, T) \neq 0$
  2. Exclusion restrition : $cov(Z, \epsilon) = 0$

$\Rightarrow$ The **instrumented difference in means** between treated and untreated units is an unbiased estimator of the causal effect of interest

Difference-In-Differences (DID)

- Combine **pre/post-treatment comparisons** in the evolution of the outcome between treated and untreated units
- This method is based on **panel data** assuming that changes in outcomes would have been similar between treated and untreated units in the absence of treatment
  $\rightarrow$ Conditional on time trends in Y_0, the treatment is *as-good-as-random*
- **Identification assumption :** Parallel trend assumption

$$E(Y_{0t'} - Y_{0t}|T = 1) = E(Y_{0t'} - Y_{0t}|T = 0)$$

$\Rightarrow$ The **pre/post-treatment difference in means** between treated and untreated units is an unbiased estimator of the causal effect of interest

Regression Discontinuity Design (RDD)

- Exploit **explicit rules (cutoffs)** for treatment assignment
- This method is based on a **discontinuity in treatment assignment** due to threshold rules on a *forcing* variable Z
  $\rightarrow$ Around a discontinuity value $\underline{Z}$, the treatment is *as-good-as-random*
- **Identification assumption :** Treatment discontinuity

$$T^+ = \lim_{Z \nearrow \underline{Z}} E(T|Z) \neq T^- = \lim_{Z \searrow \underline{Z}} E(T|Z)$$

$$E(Y_0) \text{ is continuous at } \underline{Z}$$

$\Rightarrow$ The **difference in means around the discontinuity** between treated and untreated units is an unbiased estimator of the causal effect of interest

References

Banerjee, Duflo (2017) *Handbook of Field Experiments, Volume 1 & 2, Elsevier North-Holland.*

Dhaliwal, Duflo, Glennester & Tulloch (2011) "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries : A General Framework with Applications for Education" mimeo, J-PAL, MIT.

Gretler, Martinez, Premand, Rawlings, Vermeersch (2011) *Impact evaluation in Practice*, World Bank

Imbens & Wooldridge (2009) "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, 47(1) :5-86.

Khandker, Koolwal & Samad (2010) *Handbook on impact evaluation : quantitative methods and practices*, World Bank

Ravallion (2006) "Evaluating Anti-Poverty Programs", *Handbook of Development Economics Volume 4*, edited by Robert E. Evenson and T. Paul Schultz, Amsterdam, North-Holland.