

Evaluation

Randomized Controlled Trials (RCT)

Jean-Noël Senne

Université Paris-Saclay

Problématiques Contemporaines du Développement

M2EIED

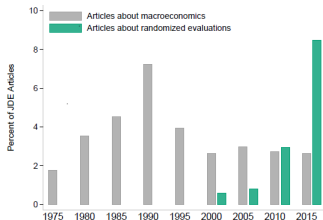
Université de Pau et des Pays de l'Adour

2023 - 2024

The success of RCTs

- **Randomized experiments** have become increasingly popular over the past 20 years
- RCTs are now widely used in various fields of economics (and social sciences) → Many “randomistas”

Abstracts of 2,695 articles from JDE



Nobel prize 2019



⇒ Why are RCTs considered as the “golden standard” for evaluation methods?

Plan

The experimental ideal

The problems of experiments

The experimental approach

- **Random assignment to treatment** : treatment T (participation to a program, intervention) is literally determined by a random draw
 - Treatment group ($T = 1$)
 - Control group ($T = 0$)
 - Very unlikely situation in the real world : individuals self-select into treatment due to their characteristics and the potential benefits they expect from treatment...
- ⇒ RCTs generate randomness to **eliminate the selection bias** !
- Simplest way to identify causal effects
→ make likely the **(strong) independance assumption**

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

Why does randomization work ?

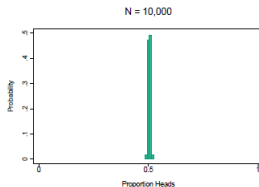
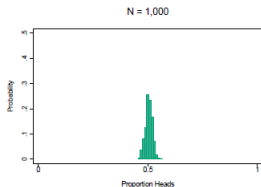
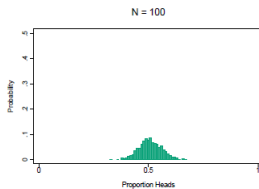
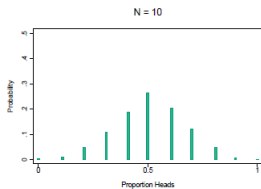
- Randomization works not by eliminating *individual difference* but rather by ensuring that the *mix of individuals* being compared is the same
- ⇒ The distribution of both observable and unobservable characteristics is **statistically identical** in the treatment and control groups
- Treatment and control groups are **random samples** of the population
- ⇒ **Expected outcomes** are the same in the absence of treatment

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

- By chance, omitted variables is still likely to matter in **small samples**...
- ... but the **law of large numbers** tells us that a sample average can be brought as close as we like to the population average just by enlarging the sample

The law of large numbers

- Let's flip a coin : what proportion of heads are you likely to get with N draws ?



⇒ The *sample average* gets closer to the *population average* (0.5) as $N \rightarrow +\infty$!

Estimating causal effects

- If treatment is random, the control group *perfectly imitates* the **unobservable counterfactual** :

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

- The difference in means estimator gives the **average causal effect** :

$$\begin{aligned}
 &\textbf{Difference in group means} \\
 &= E(Y_i | T_i = 1) - E(Y_i | T_i = 0) \\
 &= E(Y_{1i} | T_i = 1) - E(Y_{0i} | T_i = 0) \\
 &= \underbrace{E(Y_{1i} | T_i = 1) - E(Y_{0i} | T_i = 1)}_{\text{ATT}} + \underbrace{E(Y_{0i} | T_i = 1) - E(Y_{0i} | T_i = 0)}_{\text{Selection bias} = 0} \\
 &= \underbrace{E(Y_{1i}) - E(Y_{0i})}_{\text{ATE}}
 \end{aligned}$$

Regression equivalent

- The regression counterpart is a simple **OLS regression** :

$$Y_i = \alpha + \beta T_i + \epsilon_i$$

$$\beta_{OLS} = E(Y_i | T_i = 1) - E(Y_i | T_i = 0)$$

⇒ β_{OLS} is **unbiased** since, out of randomness, $E(\epsilon_i | T_i) = 0$

- OLS estimation is often used because it allows to **add controls** :

$$Y_i = \alpha + \beta T_i + X_i \gamma + \epsilon_i$$

- **But why ?**

- ▶ Randomization may fail (in small samples) or is done conditionnal on some observable characteristics (conditional random assignment)
- ▶ Additional controls **increase precision** if they have substantial explanatory power over Y_i (see next slide)
- ▶ With interactions, it allows to estimate **heterogenous treatment effects** (observable heterogeneity)

More precision on precision

- If treatment assignment is truly random, conditioning on X_i does not affect **point estimates**...

$$(1) \quad Y_i = \alpha + \beta T_i + \epsilon_i$$

$$(2) \quad Y_i = \alpha + \beta T_i + X_i \gamma + \nu_i$$

⇒ X_i and T_i are **(mean-)independent** : $E(X_i | T_i) = E(X_i)$

⇒ Models (1) and (2) yield the **same estimate** $\hat{\beta}_{OLS}$ (*Frish-Waugh theorem*)

- ... but it reduces **standard errors** !

$$V(\hat{\beta}_{OLS}) = \sigma^2 (TT')^{-1} \quad \text{where } \sigma^2 \text{ is the residual variance } V(\epsilon_i)$$

$$V(\epsilon_i) = V(X_i \gamma + \nu_i) > V(\nu_i)$$

⇒ Model (2) estimates $\hat{\beta}_{OLS}$ with **more precision**

Randomization in practice

How do we implement an RCT step by step? Let

- 1 Define the **population of interest** (sample N not necessarily random)
- 2 **Randomly divide** the sample into two treatment and control groups
- 3 Collect **baseline data** (before the treatment is implemented)
- 4 Check the randomization (**balance check**)
- 5 Process monitoring (treatment implementation)
- 6 Collect **endline data** (after treatment is implemented)
- 7 Run your **impact evaluation analysis** (OLS regressions)

⇒ So that's all? Easy, isn't it?

⇒ **Why don't we randomize everything?**

Back to the hospital

Let's reconsider the hospital example, with potential outcomes :

	$Y_{0,i}$	$Y_{1,i}$
Sick	$z - s$	$z - s + b - c$
Not sick	z	$z - c$

Let's note :

- ▶ z = individual i 's health if she doesn't get sick
- ▶ s = reduction in health associated with sickness
- ▶ b = benefit a sick person receives from going to the hospital
- ▶ c = reduction in health from going to the hospital
- ▶ λ = share of sick people in the population

(We can reasonably assume that $b > c > 0$)

⇒ How can we estimate the **causal effect** of going to the hospital ?

Hospital without random assignment

- What happens **without random assignment** ?
- Only sick people go to the hospital (healthy don't)
- What do we learn from a naive comparison of means ?

$$\begin{aligned}
 \text{Difference in means} &= E(Y_i | T_i = 1) - E(Y_i | T_i = 0) \\
 &= E(Y_{1i} | T_i = 1) - E(Y_{0i} | T_i = 0) \\
 &= \underbrace{(z - s + b - c)}_{E(Y_{1i} | T_i = 1)} - \underbrace{z}_{E(Y_{0i} | T_i = 0)} \\
 &= (b - c) - s
 \end{aligned}$$

- ⇒ **Difference in means = ATT** (effect of hospitalization on those who choose to go to the hospital) plus a **selection bias** (people choose to go because they are sick)

Hospital with random assignment on the entire population

- Suppose (absurdly) that we **randomize who goes to the hospital**
- Randomization breaks the link between sickness and going to the hospital (eliminates the selection bias)
- Let S_i is a dummy for being sick :

$$P(S_i = 1 | T_i = 1) = P(S_i = 1 | T_i = 0) = P(S_i = 1) = \lambda$$
- What do we learn now from a comparison of means?

Difference in means

$$\begin{aligned}
 &= E(Y_i | T_i = 1) - E(Y_i | T_i = 0) \\
 &= E(Y_{1i} | T_i = 1) - E(Y_{0i} | T_i = 0) \\
 &= \underbrace{\lambda(z - s + b - c) + (1 - \lambda)(z - c)}_{E(Y_{1i} | T_i=1)} - \underbrace{\lambda(z - s) + (1 - \lambda)z}_{E(Y_{0i} | T_i=0)} \\
 &= \lambda b - c
 \end{aligned}$$

- ⇒ **Difference in means = ATE** (effect of hospitalization on the population)
- ⇒ But... is it really what we are really interested in ?

(Note : $ATE \neq ATT$ due to fundamentally heterogenous treatment effects)

Hospital with random assignment on the sick

- Suppose that we **randomize who goes to the hospital among the sick**
- Randomization still breaks the link between sickness and going to the hospital (eliminates the selection bias)...
- $P(S_i = 1 | T_i = 1) = P(S_i = 1 | T_i = 0) = 1$
- ... but focuses on the **relevant counterfactual**
- What do we learn now from a comparison of means?

$$\begin{aligned}
 \text{Difference in means} &= E(Y_i | T_i = 1) - E(Y_i | T_i = 0) \\
 &= E(Y_{1i} | T_i = 1) - E(Y_{0i} | T_i = 0) \\
 &= \underbrace{(z - s + b - c)}_{E(Y_{1i} | T_i = 1)} - \underbrace{(z - s)}_{E(Y_{0i} | T_i = 0)} \\
 &= b - c
 \end{aligned}$$

- ⇒ **Difference in means = ATT** (effect of hospitalization on the sick)
- ⇒ But... is it really the ideal experiment?

Hospital with random assignment of access

- It might be better to **randomize access to the hospital**
- Let Z_i be an indicator of random assignment to a treatment group whose access to the hospital is facilitated (control group cannot use the hospital).
- Randomization breaks the link between sickness and access to the hospital

...

$$P(S_i = 1|Z_i = 1) = P(S_i = 1|Z_i = 0) = P(S_i = 1) = \lambda$$

- But people choose whether or not to go to the hospital
- **Imperfect compliance**

$$P(T_i = 1|Z_i = 1) = \lambda \neq P(T_i = 1|Z_i = 0) = 0$$

- ⇒ **Treatment take-up is endogenous** : only sick people will choose to go to the hospital
- ⇒ Back to **selection...** ? Yes, but you handle it by randomizing the likelihood to go !

Hospital with random assignment of access

- What do we learn now from a comparison of means?

Difference in means

$$\begin{aligned}
 &= E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0) \\
 &= \underbrace{\lambda E(Y_{1i} | T_i = 1) + (1 - \lambda) E(Y_{0i} | T_i = 0)}_{E(Y_i | Z_i = 1)} - \underbrace{E(Y_{0i} | T_i = 0)}_{E(Y_i | Z_i = 0)} \\
 &= (\lambda(z - s + b - c) + (1 - \lambda)z) - (\lambda(z - s) + (1 - \lambda)z) \\
 &= \lambda(b - c)
 \end{aligned}$$

⇒ **Difference in means = ITT** (intention-to-treat, i.e. effect of access to hospitalization on the sick)

⇒ $ITT = ATT \times \text{compliance}$

- In the end, 3 experiments yields 3 different average treatment effects ...

⇒ **What can we actually learn from RCTs?**

Plan

The experimental ideal

The problems of experiments

Implementation issues

While their approach is simple, RCTs raise issues of (operational) feasibility :

- ▶ They are **theoretically unfeasible** in some contexts (some events *cannot* be randomized)
 - Fundamentally Unidentified Questions : **FUQ!** (*Angrist & Pischke*)
- ▶ They are **practically unfeasible** in other contexts (some events *are hard* to randomize)
- ▶ They raise **ethical issues**
 - Should we deny treatment to someone it might help ?
 - Institutional Review Board/Ethical committee must approve all experiments with human subjects
- ▶ They rely on **good collaboration** between the research and operational teams (who may have different objectives) and have high **financial costs**
- ▶ Mix up of treatment and control groups and **imperfect compliance**
 - People cannot be compelled to participate to the experiment
 - Within the experiment, people cannot be compelled to take the treatment

⇒ **How can we solve those problems ?**

The variety of experimental designs

There is a variety of **experimental designs** to evaluate various types of programs with various forms of interventions

Design	Most useful when	Advantages	Disadvantages
Basic lottery	Program oversubscribed OK for some to get nothing	Familiar Easy to understand Easy to implement Can be implemented in public	Control group may not cooperate Differential attrition
Phase in	Expanding over time Everyone must receive treatment eventually	Easy to understand Constraint easy to explain Control comply as expect to benefit later	Anticipation of treatment may impact short run behavior Difficult to measure long term impact
Rotation	Everyone must get something at some point, not enough resources a year for all	More data points than phase in	Difficult to measure long term
Encouragement	Program has to be open to all comers When take up in general is low but can be impacted with incentive easily.	Can randomize at individual level even when program isn't	Measures impact of those who respond to the incentive Need big enough enducement to get change in take up Encouragement may have direct effect

- ⇒ **Encouragement designs** are easier to implement and more acceptable (often no choice...)
- ⇒ ... but they came **at a cost** (not sure who will get treated in the end)

Encouragement design principle

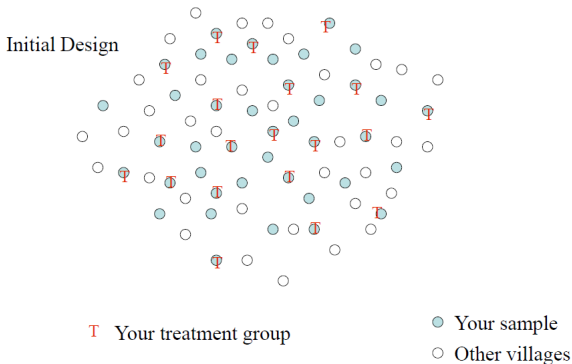
- It might be difficult (or hardly acceptable) to limit access to treatment to the control group
- The intervention is open to all comers but you **randomly provide an incentive** (informational or financial) for some of them to get treated

→ Encouragement design

- You don't need that all individuals react to the incentive, but you expect (and need) a **higher treatment take-up** in the group that receives the incentive
- Often relevant when natural take-up (without incentive) is low
- ⇒ **How can we estimate average causal effects in such settings?**

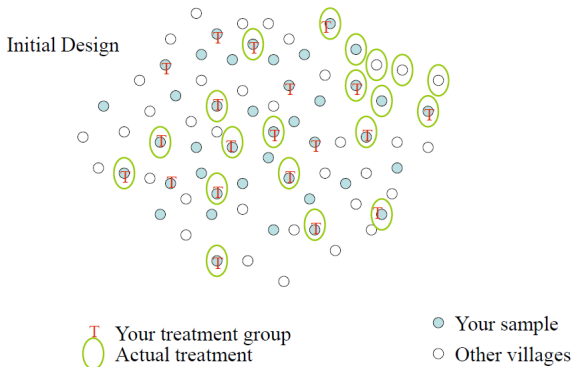
Intended *versus* actual treatment

Let's consider this **initial encouragement design** : you randomly assign across a sample of villages an incentive to get treated



Intended *versus* actual treatment

Your **actual treatment** might differ from your **intended treatment**...why?



- **Two-sided imperfect compliance :**

- Some villages that you intended to treat decided not to get treated
- Some control villages that you did not intend to treat decided to get treated

⇒ **Treatment take-up becomes endogenous again...**

Encouragement design in practice

- You randomly assign an **incentive** Z to two experimental groups :
 - ▶ **Intended treatment group** ($Z = 1$) receives an incentive to take treatment T (but some won't)
 - ▶ **Intended control group** ($Z = 0$) don't receive any incentive to take treatment T (but some will)
- ⇒ Treatment take-up is **endogenous** (individuals choose to get treated) ...
- ⇒ ... but you expect take-up to be higher when $Z = 1$ (**exogenous**)

- Two types of compliance :
 - ▶ **Perfect compliance**

$$P(T = 1|Z = 1) = 1$$

$$P(T = 1|Z = 0) = 0$$

- ▶ **Imperfect compliance**

$$0 < P(T = 1|Z = 0) < P(T = 1|Z = 1) < 1$$

⇒ How can we deal with imperfect compliance?

Intention to treat

- The difference in means between the intended treated and control groups estimates the **intention-to-treat effect** :

$$ITT = E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)$$

⇒ **ITT is unbiased** due to random assignment of $Z...$

⇒ ... but **is it the causal effect of interest** ?

- Let's note :

$$p_1 = P(T = 1|Z = 1)$$

$$p_0 = P(T = 1|Z = 0)$$

- And re-write the components of ITT :

$$\begin{aligned} E(Y_i|Z_i = 1) &= E(Y_{1i}|T_i = 1, Z_i = 1)p_1 + E(Y_{0i}|T_i = 0, Z_i = 1)(1 - p_1) \\ &= E(Y_{1i} - Y_{0i}|T_i = 1, Z_i = 1)p_1 \\ &\quad + E(Y_{0i}|T_i = 1, Z_i = 1)p_1 + E(Y_{0i}|T_i = 0, Z_i = 1)(1 - p_1) \\ &= E(Y_{1i} - Y_{0i}|T_i = 1)p_1 + E(Y_{0i}) \end{aligned}$$

Similarly :

$$E(Y_i|Z_i = 0) = E(Y_{1i} - Y_{0i}|T_i = 1)p_0 + E(Y_{0i})$$

From ITT to ATT

- We can now recover the ATT from the ITT :

$$\begin{aligned} ITT &= E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) \\ &= E(Y_{1i} - Y_{0i}|T_i = 1)(p_1 - p_0) \\ &= ATT \times (p_1 - p_0) \end{aligned}$$

⇒ We need to **estimate** $(p_1 - p_0)$

- We can estimate it from a **simple OLS regression** : $T_i = \pi_0 + \pi_1 Z_i + \nu_i$

$$\begin{aligned} \pi_{1OLS} &= E(T_i|Z_i = 1) - E(T_i|Z_i = 0) \\ &= p_1 - p_0 \end{aligned}$$

⇒ π_{1OLS} is **unbiased** since Z_i is randomly assigned

- We can then recover the ATT from an **instrumental variable estimation** !

$$ATT_{IV} = \frac{ITT}{(p_1 - p_0)} = \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(T_i|Z_i = 1) - E(T_i|Z_i = 0)}$$

⇒ Deviation from initial random assignment (imperfect compliance) is not an issue, since the incentive Z_i is a very robust instrumental variable !

The cost of imperfect compliance

- Warning! The instrumental variable estimator gives a **LATE (Local Average Treatment Effect)**
 - ▶ Local? Only those who get treated due to the incentive contribute to the estimation. The effect is estimated **on the compliers**
 - ▶ Need a sufficiently **high rate of compliance**
 - ▶ Treatment effects can be **heterogenous**

⇒ In general : **LATE** \neq **ATT** \neq **ATE** (unless constant treatment effect)

- With imperfect compliance, it can be shown that :

$$V(ATT_{IV}) = \frac{1}{\bar{Z}(1 - \bar{Z})} \frac{V(\epsilon)}{N} \frac{1}{\pi_1^2} > V(ATT_{OLS}) = \frac{1}{\bar{Z}(1 - \bar{Z})} \frac{V(\epsilon)}{N}$$

- ⇒ LATE estimation are **less precise** (tradeoff between bias and precision)
- ⇒ Lead to reconsider intrumental variable estimation (*see Lecture 3*)

Example : Deworming (Miguel & Kremer (2004))

- RCT on deworming medication in Kenya to increase school attendance
- Randomization was done at the school level
- Imperfect compliance
 - ▶ Some children in treated schools were absent when the doctor came and some refused to take the medication
 $P(T = 1|Z = 1) < 1$
 - ▶ Some children in untreated schools had siblings in treated schools and got access to the medication
 $P(T = 1|Z = 0) > 0$
- What causal effect can we measure?
 - ▶ The average **causal effect of the intervention**? → **ITT**
 - ▶ The average **causal effect of the medication**? → **ATT**

Example : Deworming (Miguel & Kremer (2004))

Ecole traitée Z=1			Ecole non traitée Z=0		
	a reçu médic. (T)	Présence (Y)		a reçu médic. (T)	Présence (Y)
Enfant 1	oui	24	Enfant 1	non	17
Enfant 2	oui	22	Enfant 2	non	14
Enfant 3	oui	18	Enfant 3	non	15
Enfant 4	oui	20	Enfant 4	oui	18
Enfant 5	non	15	Enfant 5	non	13
Enfant 6	oui	21	Enfant 6	non	19
Enfant 7	oui	19	Enfant 7	non	14
Enfant 8	non	14	Enfant 8	non	19
Enfant 9	oui	16	Enfant 9	non	20
Enfant 10	oui	18	Enfant 10	non	17
Moyenne	0,8	18,7	Moyenne	0,1	16,6

$$ITT = E(Y|Z = 1) - E(Y|Z = 0) = 18,7 - 16,6 = 2,1$$

$$ATT = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(T|Z = 1) - E(T|Z = 0)} = \frac{18,7 - 16,6}{0,8 - 0,1} = 3$$

More threats to internal validity

RCTs maximize the internal validity (capacity of drawing causal inference) but with some limitations :

- ▶ **Imperfect compliance** and randomization failures
 - You don't always get what (i.e. the average causal effect) you want
- ▶ Existence of **spillovers**
 - You need the **SUTVA assumption**
 - What is the appropriate unit of randomization ? (cluster-randomization makes sense when spillovers are anticipated)
- ▶ Treatment and control groups may behave differently (consciously or not) because they know they are being observed
 - **Hawthorne effect** : motivation increases among the treated group (to make the experiment work)
 - **John Henry effect** : motivation increases among the control group (to compensate its disadvantage)
- ▶ **Non-response or attrition bias**
 - Problem when it's correlated to treatment (additional selection issue)
 - The control group may not cooperate because it don't receive any treatment
- ▶ **Statistical inference** in small samples
 - Which sample size to credibly detect an effect that exists ?
 - See Gertler et al. (2011), chap. 11 on sample size calculation

Even more threats to external validity

RCTs have more serious issues of external validity (capacity of extrapolating the results) :

- ▶ The sample is often **non-representative**
- ▶ Experiments are **specific** to a given context
- ▶ **Limited duration**
 - What can we learn from a single study ?
 - Can we generalize the results to other contexts or periods ?
- ▶ **Heterogenous** treatment effects
- ▶ **Partial equilibrium** effects in small-scale experiments
 - What would happen if we would scale-up the intervention ?
 - What would be the general equilibrium effects ?
- ▶ Experiments often come **without a model** ("black-box")
 - How can we interpret the results (underlying mechanisms) ?
 - Frontier in this literature : combine experiments with structural models

So... are experiments worthless ?

- RCTs are a simple way to...
 - ▶ Solve the fundamental problem of evaluation = **selection**
 - ▶ Ensure a high degree of **internal validity** (despite limitations)
 - ▶ Provide a **benchmark** for the estimation of causal effects
 - but it comes at some costs
 - ▶ Only local average treatment effects (**LATE**) are estimated
 - ▶ Limited **external validity**
 - RCTs often end up being a (really) expensive way to buy a (really good) instrument !
 - The big question is : **can we achieve similar results with non-experimental data ?**
- The answer is often **yes !**

References

Banerjee, Duflo (2017) “An Introduction to the “Handbook of Field Experiments””, Chapter 1 in Banerjee, Duflo (Eds) Handbook of Field Experiments, Elsevier, North-Holland.

Banerjee, Duflo (2009) “The Experimental Approach to Development Economics” The Annual Review of Economics 1 :151-178.

Duflo, Glennerster et Kremer (2006) “Using Randomization in Development Economics Research : a Toolkit”, CEPR Discussion Paper Series No. 6059.

Gertler, Martinez, Premand, Rawlings, Vermeersch (2011) Impact evaluation in Practice, World Bank – Chapitre 11

Glennerster (2017) “The Practicalities of Running Randomized Evaluations : Partnerships, Measurement, Ethics, and Transparency”, Chapter 5 in Banerjee, Duflo (Eds) Handbook of Field Experiments, Elsevier, North-Holland.

Parienté (2008) “Analyse d'impact : l'apport des évaluations aléatoires”, Statéco, no. 103.