

Policy Evaluation

2. When are Randomized Controlled Trial (RCT) useful?

Lab and Field Experiments

Survey experiments

João Pereira dos Santos

E-mail: joao.santos@iseg.ulisboa.pt

ISEG, IZA, GLO

Experiments in economics?

“Unfortunately, we can seldom test particular predictions in the social sciences by experiments explicitly designed to eliminate what are judged to be the most important disturbing influences. Generally, we must rely on evidence cast up by the “experiments” that happen to occur.”

– Milton Friedman (1953)

But since then...

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2002



Photo from the Nobel Foundation archive.

Daniel Kahneman



Photo from the Nobel Foundation archive.

Vernon L. Smith

To Daniel Kahneman "for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty" and Vernon L. Smith "for having established laboratory experiments as a tool in empirical economic analysis, especially in the study of alternative market mechanisms"

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2017



© Nobel Media AB. Photo: A. Mahmood
Richard H. Thaler

To Richard H. Thaler "for his contributions to behavioural economics"

And many more (as we have seen last class)!

Advantages of the Experimental Method

Avoid selection biases: we are comparing apples to apples, the comparison group that does not differ systematically from the treatment group at the outset of the program/evaluation

Test a theory: game theory offered testable predictions of behaviour. "We don't know what you should do. The whole purpose of running the experiments is to see what you will do. You should do whatever seems best to you."

Simple ($\neq$ Easy)! "Whispering in the ears of princes" (Roth)

Can be used to measure unobserved behaviour (e.g., discrimination, preferences,...)

Transparency: the role of the pre-analysis plan (PAP), a document produced at the design stage of an impact evaluation that sets out in advance how the researcher will analyze data + power calculations

What types of cause-and-effect questions can randomized evaluations help to answer?

- How effective is a given program?
 - Who benefits most?
- How do different versions of a program compare to one another?
 - Which components work or do not work? How do these function together?
- How do program effects compare under different delivery mechanisms?
 - How to accurately target beneficiaries? How to increase program take-up?
- How cost-effective is a program?
 - How does it compare to other programs designed to accomplish similar goals?

We can use cost-benefit analysis using RCT results

What makes a good experiment?

Should an experiment replicate reality?

Should an experiment replicate a formal model?

- Often no

Goal:

- a design that offers the best opportunity to learn something useful and to answer the questions that motivate your research
- an experiment is judged by it's impact on our understanding

Answer depends on what you are testing, and who you are talking to, but... a good design!

- simple compared to reality and simpler than relevant models
- designed to test specific hypotheses
- tests or controls for alternative hypotheses

There are many, many decisions you need to make on your design... and it is very difficult/ impossible to change them afterwards!

Direct vs indirect experimental control

- Direct experimental control: Control vs. Treatment
 - Test hypothesis by changing one variable at a time to avoid confounds
 - Only change variables which are directly relevant to the hypothesis being tested, otherwise holding the environment fixed
- Indirect experimental control: Uncontrolled factors? controlled via randomization
 - By randomly assigning subjects to treatments, we can eliminate subjects differing attitudes as a cause of differences between treatments. This relies on the law of large numbers, implying that a large sample may be necessary.

Within- or Between-subject design?

- Within-subject: participants make decisions in all treatments (panel vs. cross section)
- Between-subject: different participants make decisions in each treatment

Advantage vs. disadvantage of within-subject design:

- Advantage: each subject is its own control. Need not worry about having different characteristics of participants in each treatment (often easier to get significance)
- Disadvantage: order effect / fatigue

Design Choices

- One round versus many rounds?
- Pay one round or all rounds?
- Use language that is neutral?
- Train participants or test them before you use them as participants in your experiment?

Design Choices: Lab experiment?

- Advantages:
 - More control: less distractions, typically it is easier to get strict instructions followed when experiments are run in the lab and students may follow difficult instructions more easily
 - Small incentives are often more meaningful
 - More transparency: The subject pool (undergraduates) is well understood. In the field you may worry you use a subject pool prone to some bias, that is then attributed to the experiment
 - More replicable: Lab experiments are very easy (and cheap) to replicate. This may make us more comfortable with surprising results (Remember: We want to protect ourselves from fooling ourselves!)
 - Market Design: the lab allows us to generate many markets.

Design Choices: or Field experiment?

- Advantages:
 - Sexier, it happens in a real context
 - Subject pool is the subject of interest: use politicians to study legislative bargaining, use villages in Africa to study deworming efforts (Duflo and co-authors), use Uber, Amazon, etc. users to study design changes on consumption
 - Sometimes you want to test if a change would have a sizeable effect when many other things happen as well. (The lab is not so great to estimate parameters, the lab may be more useful to study treatment effects)

To know more: <https://www.povertyactionlab.org/page/handbook-field-experiments>

- Lab vs. Field:
 - unclear which is cheaper: you want really large samples

Methodological norms: Monetary incentives

- Advantages:
 - Subjects make more effort / pay more attention leading to less noise and more consistent choices
 - Equalises marginal gains across subjects
- Disadvantages:
 - Its expensive and limits stakes (barriers to entry)
 - Clear evidence that monetary incentives leads to differences in behaviour
e.g., more risk aversion and less generous behaviour with monetary gains

Methodological norms: No deception

You cannot lie... But you don't need to say the whole truth.
Researchers can use abstract instead of concrete wording

- Advantages:

- Subjects believe the instructions and do not try to outguess the experimenter
- Does not impose an externality (contamination) on other experiments/ researchers

- Disadvantages:

- Makes it harder to study situations where lying is important (response: lying game)
- Clear evidence that monetary incentives leads to differences in behaviour
e.g., more risk aversion and less generous behaviour with monetary gains

Unit of observation and level of randomization

At which level should a study randomize? What are the outcomes we care about, and at what level are we able to measure them?

- Randomizing at the individual level
 - e.g., people, patients, or students
- Randomizing at the group/cluster level
 - e.g., villages, clinics, or schools
 - Outcomes can still be measured at the individual level

Balance & Stratification

Balance = the treatment and control groups are comparable on certain key characteristics

- Can check using balance tests (differences of means). If you find that some variables are unbalanced, consider the number of imbalances and their magnitude

One way to achieve balance (at least for certain observables):
Stratification = dividing the sample into subgroups (also known as strata) that share certain characteristics (e.g., age, gender, etc.) and then randomizing within each subsample

How to “detect an effect” ?

To be sure that the measured program effect is due to the program itself and not due to natural variation or random chance

Trade-off the risks of false positives and false negatives:

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

We do this by conducting a thought experiment (hypothesis testing)

Statistical significance

Ask: How likely would it be to observe this outcome due to random chance alone (natural variation in participants)?

- If it is reasonably unlikely (5% probability or less), then we conclude that the result was statistically significant
 - This is what we mean when we say “detect an effect”

Statistical significance is about avoiding a false positive (concluding your program had an impact when it did not)

A statistically significant result is unlikely to have been produced by chance

Statistical power

Ask: If there were a true underlying effect, how likely would we be able to detect it in this experiment?

The statistical power of an evaluation is the probability of detecting an impact when there actually is one

In other words, statistical power is the likelihood of avoiding a false negative (concluding there is no impact when there actually is one)

By convention, we aim for 80% power

- This means that we expect that 80% of the time we will be able to detect an effect if there is one
- 20% of the time we will falsely conclude there is no impact of our program

Absence of evidence or evidence of absence?

If we do not have a statistically significant result, there are two interpretations:

1. There is no effect of our program (true negative!)
2. There is an effect, but we don't have enough statistical power to observe it (false negative!)

Without adequate power, an evaluation may not teach us much

Failure to find a statistically significant effect can be misinterpreted as the failure of the program, rather than the failure of the evaluation

Is there evidence of sharks?

Absence of evidence



Source: Wikimedia Commons

Evidence of absence



In the first case, we cannot conclude that there are no sharks under the surface of the water

Key inputs in determining statistical power

- Effect size: the minimum detectable effect (MDE) for a given power level
 - Accounting for the take up rate
- Sample size
 - Accounting for attrition
- Variation in the outcome
- Proportion of the sample in the program group
- Unit of randomization (i.e. clustering)

Challenge: Units within clusters are not independent of one another

Power calculation formula (solving for sample size)

$$N = (1.96 + 0.84)^2 \cdot \frac{1}{P(1 - P)} \cdot \frac{\overset{\sigma^2 = \text{variance}}{\sigma^2}}{MDE^2}$$

N = sample size

Considering standard values for significance and power levels
($\alpha = 5\%$ and power set to 80%)

P = Proportion of the sample in the treatment group

MDE is the smallest effect size that can be detected given the other inputs. (This might factor in participation or “take up rate”)

More detail: <https://www.povertyactionlab.org/resource/power-calculator>

How to improve power?

- Increase the sample size
 - Increase the number of units or clusters or reduce attrition
 - Conduct individual-level random assignment when possible
- Increase the effect size
 - Increase the intensity of the treatment or take-up/compliance
- Simplify the design
 - Reduce the number of treatment arms or the number of hypotheses you test as the study needs to be powered for the smallest MDE among the intended comparisons
- Reduce variation in the outcome
 - Stratify the randomization to ensure baseline balance on important observables
 - Control for covariates (especially baseline measures of the outcome) to reduce residual variance - it should give similar point estimates!

Practical tips

- Perform power calculations early in the design phase (before the program is implemented)
- Don't panic about the number of assumptions required
 - Power calculations should be considered a rough guide in the decision of whether to carry out the study and provide an estimate of how large the sample should be
- Conduct sensitivity analyses to test how power changes with changes to any critical assumptions
 - Create “best case” scenarios and “worst case” scenarios and evaluate those
 - If the best case scenario MDE is unrealistically high or requires an unrealistically large sample size, consider how to tweak the design to increase power
 - If sufficient power cannot be achieved (e.g., do we have enough schools?), an RCT might not be the best way forward

Threats to validity

During the conception phase, we design an evaluation that enables us to answer our research questions

But the implementation phase of the evaluation is also extremely important: many things can go wrong!

- Spillovers
- Attrition
- Evaluation-driven effects
- Partial compliance

Spillovers

Remember our discussion about SUTVA

Spillovers can be positive or negative and therefore...

Spillovers can cause impact to be underestimated or overestimated

Channels through which spillovers occur include:

- physical,
- informational/behavioral,
- marketwide/general equilibrium

Physical Spillover

Example: Cash transfer program

A member of the treatment group receives a cash transfer and gives some of the money to friends or relatives who are assigned to the control group



Behavioural/Informational Spillover

Example: Handwashing promotion campaign

Control group imitates neighbors' hygiene practices or learns about the health benefits of handwashing



Marketwide/General Equilibrium Effects Spillover

Example: Access to microcredit for entrepreneurs

Control group entrepreneurs are in competition with treatment group for recruiting from a limited number of workers in the local economy



What can be done about spillovers?

Measure spillovers

- Build plans to collect data on spillovers into the experimental design (to account for network effects)
- Measure spillovers in the analysis phase

Avoid spillovers

- Choose level of randomization wisely, and randomize at a higher level if concerned about spillovers (e.g., school, village)
- Incorporate spatial buffers between treatment and control units (don't interview friends, neighbours,... of treated/ control units)

Think about trade-offs...

Unit of Randomization: Student?



Unit of Randomization: Classroom

We call groups of units "clusters"
Randomization at the group level:
"cluster randomized trial"



Unit of Randomization: School

We call groups of units "clusters"
Randomization at the group level:
"cluster randomized trial"



Attrition

Attrition occurs when study group members leave the study and data on their outcomes cannot be collected

- It may be a problem depending on how much of the study sample we lose
- It is a problem if the type of people who leave is correlated with the treatment
- Common drivers of attrition include mobility or migration, motivation, and mortality

E.g.: consider the impact of microcredit on business profits:

Without attrition

Before Treatment		After Treatment	
T	C	T	C
 2,000	 2,000	 2,200	[absent]
 2,500	 2,500	 2,700	 2,500
 3,000	 3,000	 3,200	 3,000
Avg.			
	2,500	2,700	2,750
Difference:			
	0	Difference:	-50

With attrition: What if the most disadvantaged households migrated to other regions?

Before Treatment		After Treatment	
T	C	T	C
 2,000	 2,000	 2,200	[absent]
 2,500	 2,500	 2,700	 2,500
 3,000	 3,000	 3,200	 3,000
Avg.			
	2,500	2,700	2,750
Difference:		Difference:	
	0		-50

What can be done about attrition?

Implementation phase

- More intensive follow-up efforts with survey respondents
 - Account for follow-up costs in project planning and funding
 - For example: Tracking of respondents who moved to neighbouring areas

Analysis phase

- Use bounded estimates to mitigate the effects of attrition on impact estimates: take the percentage difference between treatment and control and drop the top percentile and bottom percentile from the group with less attrition to bound the estimates, creating worst case and best case scenarios

Evaluation-driven effects

These effects occur when respondents change their behaviour in response to the evaluation itself instead of the intervention

Common causes: salience of being evaluated, social pressure

These include observer-driven effects and enumerator effects:

- Hawthorne effects: Behaviour changes due to attention from the study or intervention
- Anticipation effects: Comparison group changes behaviour because they expect to receive the treatment later (particular concern for phase-ins)
- Resentment/demoralization effects: Comparison group resents missing out on treatment and changes behaviour
- Demand effects: Behaviour changes due to perceptions of evaluator's objectives
- Survey effects: Being surveyed changes subsequent behaviour

What can be done about evaluation-driven effects?

Evaluation design

- Use a different level of randomization
- Measure the evaluation-driven effects in a subset of the sample
 - Prime a subset of the sample by reminding them of the evaluation (e.g., Mummolo and Peterson 2019)
 - Supplement survey data with other measures of behavioural outcomes (e.g., Fearon, Humphreys, and Weinstein 2008)

Implementation phase

- Minimize salience of evaluation as much as possible
 - Do not announce phase-in
- Downside is that this can be useful to reduce attrition!
 - Make sure staff are impartial and treat both groups similarly
- E.g., do not share treatment assignment with data collection staff

Partial compliance

Noncompliance occurs when a unit's treatment assignment (to a treatment or comparison group) does not match their treatment status

A study sample can be split into three distinct groups (assuming non-defiers):

- Compliers: follow assignment
- Always-takers: Always take the treatment, even if assigned to the control group
- Never-takers: Always refuse the treatment, even if assigned to the treatment group

Potential Sources of Noncompliance: logistical or political challenges, when service providers may find it difficult to administer customized treatment alongside their other responsibilities

Noncompliance can lead to sample selection bias

What can be done about non-compliance?

Design phase

- Randomize at a higher level to enable providers to treat clusters the same

Implementation phase

- Prevent noncompliance, e.g., by making take up easy or by incentivizing take up, but this cannot always be done
- Monitor noncompliance to be aware if/when it happens

Analysis phase • Interpret it during analysis phase:

- ITT: estimates the overall effect of the intervention, admitting that noncompliance can happen (which can be the policy-relevant parameter of interest)
- LATE: estimates the effect of the intervention for those who comply with their assignment to treatment or control

Criticisms

Generalizable? They don't guarantee external validity

- Which is quite right, but it is not like they are less externally valid than other methods. . .
- And because they are internally valid:
 - compare them across contexts
 - run them in different contexts (meta analysis!)

From what works to why?

- test different versions of an intervention to help determine which components are necessary for it to be effective
- provide information on intermediate outcomes

Costs and Sample size: small samples make estimates imprecise, especially for long-term impacts

Ethics

- In many cases we don't know whether an intervention is cost-effective
 - It is also possible to conduct a RCT without denying access to the intervention. E.g., randomly select people to receive encouragement to enroll without denying any interested participants access

Review of criticisms by Angus Deaton <https://www.sciencedirect.com/science/article/pii/S0277953617307359>

Still, many questions CANNOT be randomized (too expensive, not morally acceptable,...)

PLACEBO CHRISTMAS



Ethical considerations

Respect: Individuals are autonomous agents capable of making their own decisions

- This requires informed consent for their participation
- Persons with diminished autonomy are entitled to additional protection (children, individuals with cognitive impairments,...)

“Do no harm”

- Do not administer a treatment that is known to be harmful
- Do not withhold a known benefit that would otherwise be available

Costs vs benefits: Minimize risks

- Potential adverse effects of the intervention on privacy
- Psychological burden of responding to surveys
- Physical and safety risks to staff (geopolitical risks)

Justice requires fairness in the allocation of risks and benefits

RCTs can be used to measure unobserved behaviour

A seminal paper published in the AER (2004) gave us a tool to measure discrimination

Note: IRB approval can be difficult, consider trade-offs!

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

By MARIANNE BERTRAND AND SENDHIL MULLAINATHAN*

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (JEL J71, J64).

Let us have a closer look at tables 3 (balance test) and 1 (main results): https://cos.gatech.edu/facultyres/Diversity_Studies/Bertrand_LakishaJamal.pdf

Social Economics Surveys and Experiments as a Key Research Tool

Large scale surveys that go in depth into people's minds and "listen to them."

Surveys have been used for a long time for statistics. Some variables are now better measured in high-quality admin data (like income, family situation, employment, etc.)

Yet, some things are invisible in data other than surveys (even great data!): perceptions, attitudes, knowledge, and views

In principle, one could specify complete structural model of these "invisible" factors, but requires many assumptions

You are creating the process that generates the data. You can create your own identifying variation

For the results to be reliable, it is critical that these surveys are well-designed, and deployed on appropriate samples

Stantcheva discusses how to recruit the sample, manage respondents' attention, and write these questions to mitigate biases in <https://www.annualreviews.org/content/journals/10.1146/annurev-economics-091622-010157>

Migration and redistribution

Alesina, Miano, and Stantcheva (2022 REStud)

Descriptive

How do people (mis)perceive immigration?

- Are perceptions of immigration, about the number, origin, religion, unemployment, education, poverty, correct amongst natives of the host countries?
- What are natives' views on immigration policies?
- Heterogeneity by political affiliation, work in high immigrant sector, income, education level...

Causal

What is the link between immigration and redistribution?

- Are perceptions of immigration and views about redistribution correlated?
- And do perceptions of immigrants “cause” preferences for redistribution?

Migration and redistribution

Large-scale surveys in 6 countries: France, Germany, Italy, Sweden, UK, and US, total of 22,500 respondents

Done through commercial survey companies in Nov 2017-Feb 2019.

Survey components:

Background info, perception of immigrants (number, origin, religion, hard work, economic conditions, support), policy preferences (redistribution + immigration)

Randomized treatments:

- Priming: “Order” treatment asks about immigration before redistributive policies
- Information (Facts) on 1) number, 2) origins of immigrants
- Anecdote on “hard-working” immigrant

Main Findings: Perceptions of Immigration Substantially and Systematically Wrong

Across countries and respondent characteristics:

- Stark overestimation of the number of immigrants
- Stark overestimation of share of Muslim
- Underestimation of immigrants education, employment, contribution to welfare state

People wrong about natives as well, but more so about immigrants

Larger misperceptions for respondents who are: i) in immigrant intensive, low-skill jobs, ii) without college, iii) female, and iv) right-wing

Perceived composition (not the number) of immigrants that differentiate natives' responses

- Left and right-wing equally misperceive % of immigrants, but right-wing believe immigrants have different characteristics

Main Findings: Effects of Priming

The group answering policy questions first has not been prompted to think about immigration at all. The other group has thought about immigration before answering policy questions

Finding: Just making people think about immigrants, before asking them questions on policies for redistribution (“order treatment”), makes them less likely to support redistribution

Paper available in <https://socialeconomicslab.org/research/publications/immigration-and-redistribution/>

Main Findings: Hard Facts vs. Narratives

Showing factual information on the share of immigrants and their origins has no effect, does not shift people's views on redistribution

Telling people a story about a “day in the life of a very hard-working immigrant” has positive impacts on support for redistribution. Why?

Because it counters the “free-rider” narrative which matters a lot for people's views

“Hard facts” do not work that well on the issue of immigration, “narratives” have a strong hold

Another example:
"Can ATMs get out the vote? Evidence from a nationwide field experiment"

It is not "rational" to vote

Can we use reminders to make people vote?

An unexploited method of voter mobilization: ATMs

with José Tavares and Pedro Vicente, published in the EER (2021)

Let me open the paper

https://www.sciencedirect.com/science/article/pii/S0014292121000441?casa_token=zjrvzS0Vo0AAAAAA:KIorcAYKXf5aduHY8ky1Tk0C0wrXTUKg4j-XkZ-dcit_5

Important to get the paper published: discuss cost-benefits