

# Running field experiments: Practical, theoretical, and ethical considerations

Mauricio Romero

# Running field experiments: Practical, theoretical, and ethical considerations

Biases from Study Effects

Lab Experiments in the Field

The Critique of RCTs

# Running field experiments: Practical, theoretical, and ethical considerations

Biases from Study Effects

Lab Experiments in the Field

The Critique of RCTs

## Placebo Effects

- ▶ Main goal of **double-blinded** trials is to eliminate placebo effects
- ▶ In social science: Usually not possible to blind subjects to their treatment status
- ▶ Might want to blind them to the existence of the experiment
- ▶ Placebo effects that arise from receipt of the treatment, as long as they would be experienced in the real implementation of the program, are generally thought to be a valid part of treatment effects in social science
- ▶ Example: “I work harder because someone was willing to trust me with a microfinance loan”

## Survey Effects

- ▶ Being surveyed can change the salience of certain activities (e.g., hand-washing, spousal abuse, savings)
- ▶ If the survey intended to capture outcomes is altering the impact of the treatment, this bias would not be observed in the actual implementation of the program
- ▶ Some ways of measuring survey treatment effects:
  - ▶ Randomize 2 survey groups: 'high-frequency' / 'low-frequency'
  - ▶ If "treatment\*survey frequency" interactions are significant the high-frequency impacts may be miss-estimated
  - ▶ Block-randomize a group (T and C) that will only be surveyed at endline
  - ▶ Treatment effect unbiased by survey effects in endline-only group (but no baseline)

## Observer Effects

- ▶ Not always possible to hide “treatment” status from enumerators
- ▶ Many questions require a certain degree of subjectivity
- ▶ Enumerators may affect responses differentially in response to the treatment
- ▶ Hidden monitoring devices allow you to measure actual behavior free of survey and observer effects
- ▶ But **this is a form of deception**

## Hawthorne and John Henry Effects

- ▶ **Hawthorne:** treatment group responds differently to the treatment because of being studied (e.g., social desirability bias)
- ▶ **John Henry:** control group responds differently to the treatment because of the inequality with the treatment created by the experiment
- ▶ These effects are harder to deal because they are an inherent part of the study
- ▶ The ideal information structure is usually that individuals in the treatment understand they are getting the program, but neither treatment nor control is aware that they are participating in an experiment
- ▶ Is this ethical? Is this deception?

# Running field experiments: Practical, theoretical, and ethical considerations

Biases from Study Effects

Lab Experiments in the Field

The Critique of RCTs



# Running field experiments: Practical, theoretical, and ethical considerations

Biases from Study Effects

Lab Experiments in the Field

The Critique of RCTs

## Trust Game

- ▶ Player 1 is given a sum of money, and can either pocket it or send some component to player 2
- ▶ The amount sent to Player 2 is triple
- ▶ Player 2 can then decide whether or not to send any of that sum back to Player 1, without multiples
- ▶ The subgame perfect equilibrium is that Player 2 shouldn't send anything back to Player 1, and therefore Player 1 shouldn't send anything to Player 2
- ▶ This game is usually interpreted to measure Player 1's 'Trust' and Player 2's 'Trustworthiness'

## Dictator Game

- ▶ Player 1 is given a sum of money and asked how much they want to give to Player 2
- ▶ Not really a 'game' since there is no strategic interaction
- ▶ Should see zero contributions and you typically see half of the money given.

## Ultimatum Game

- ▶ Player 1 suggests a division of the money
- ▶ Player 2 can either accept or reject the division
- ▶ If player 2 accepts, the money is divided according to the proposal, and if Player 2 rejects, then neither player gets anything
- ▶ Typically see a 50-50 split proposed (not the sub-game perfect equilibria)
- ▶ Typically see offers less than 20% rejected (even though this is irrational for Player 2)

## Public Goods Game

- ▶ A group of players (typically between 2 and 10) are given a sum of money, which they can either pocket or contribute to a common pool
- ▶ The amount of money in the pool is then doubled, divided evenly, and returned to the players
- ▶ Game is typically iterated at least 10 times, creating a dynamic Prisoner's Dilemma
- ▶ One-shot equilibrium is to free-ride on the group and contribute nothing
- ▶ Overall winnings will be largest for groups which sustain cooperation over time, and a collapse at the end
- ▶ Ability to sanction members at a cost (sometimes just as simple as a frowning face to indicate displeasure)
- ▶ Usually games played with socioeconomically heterogeneous groups have lower contributions than with homogenous groups

## Microfinance Game

- ▶ Places players in groups and then gives them sum of money which they can either pocket or invest in a risky project with a positive expected return
- ▶ In the event that the project does not pay off, group members must then decide whether to cover each others' loans
- ▶ If the aggregate group loan is paid off then the game continues to the next round
- ▶ Game tests a complex combination of risk aversion and reciprocity-based altruism
- ▶ Relevant for circumstances in which individuals' exposure to investment risk is a function of the quality of mutual insurance in the group

## Discounting Game

- ▶ Players are offered a choice between receiving a sum of money today and a greater sum of money in the future
- ▶ By varying the amount offered in the future you can estimate the distribution of the rate of time preferences
- ▶ For a given individual you have to present with multiple choices to get anything other than a binary measure of 'Players discount rate higher than X?'
- ▶ Executing the Discounting game requires returning at a later date to make the subsequent payments
- ▶ if there is any doubt in players' minds about the credibility of this promise then the game measures joint effect of credibility and discounting
- ▶ Varying the timeframe and the payoffs allows one to test for hyperbolicity (i.e., present bias): does this individual display discount rates that are dependent upon the period of time over which the discounting problem is being considered?

## Risk Aversion Game

- ▶ Players are given a sum of money which they can either pocket or hazard on a risky project with positive expected returns
- ▶ Dice rolled to determine actual outcome on the spot
- ▶ By ratcheting up the risk and expected return simultaneously you can estimate higher-order moments of risk aversion



## Habyarimana, Humphreys, Posner, Weinstein. “Why Does Ethnic Heterogeneity Undermine Public Goods Provision?”

- ▶ Try to understand the result, found in many contexts, that ethnically diverse places are less good at providing public goods. Develop three hypotheses:
  1. Preferences: members of the same ethnic group want the same things, or have other-regarding preferences and so want good things for each other
  2. Technology: co-ethnics either share linguistic advantages, or are more able to locate each other within social networks
  3. Strategy Selection: w/multiple equilibria, co-ethnics believe that the other will play cooperatively and therefore Nash solutions are self-fulfilling

## Habyarimana, Humphreys, Posner, Weinstein. “Why Does Ethnic Heterogeneity Undermine Public Goods Provision?”

- ▶ Try to understand the result, found in many contexts, that ethnically diverse places are less good at providing public goods. Develop three hypotheses:
  1. Preferences: members of the same ethnic group want the same things, or have other-regarding preferences and so want good things for each other
    - ▶ Stated preferences over public good provision: no differences across ethnicities
    - ▶ Dictator game: offerer is anonymous, recipients are known: no discrimination
  2. Technology: co-ethnics either share linguistic advantages, or are more able to locate each other within social networks
  3. Strategy Selection: w/multiple equilibria, co-ethnics believe that the other will play cooperatively and therefore Nash solutions are self-fulfilling

## Habyarimana, Humphreys, Posner, Weinstein. “Why Does Ethnic Heterogeneity Undermine Public Goods Provision?”

- ▶ Try to understand the result, found in many contexts, that ethnically diverse places are less good at providing public goods. Develop three hypotheses:
  1. Preferences: members of the same ethnic group want the same things, or have other-regarding preferences and so want good things for each other
  2. Technology: co-ethnics either share linguistic advantages, or are more able to locate each other within social networks
    - ▶ Players had to describe to someone else who can't see what they're doing how to put together a puzzle. Tests linguistic, cultural ability to converse. No differences
    - ▶ Select 'targets' who were people unknown to the runners, runners sent out with a name and the town in which the target lives. Co-ethnics better at finding each other
  3. Strategy Selection: w/multiple equilibria, co-ethnics believe that the other will play cooperatively and therefore Nash solutions are self-fulfilling

## Habyarimana, Humphreys, Posner, Weinstein. “Why Does Ethnic Heterogeneity Undermine Public Goods Provision?”

- ▶ Try to understand the result, found in many contexts, that ethnically diverse places are less good at providing public goods. Develop three hypotheses:
  1. Preferences: members of the same ethnic group want the same things, or have other-regarding preferences and so want good things for each other
  2. Technology: co-ethnics either share linguistic advantages, or are more able to locate each other within social networks
  3. Strategy Selection: w/multiple equilibria, co-ethnics believe that the other will play cooperatively and therefore Nash solutions are self-fulfilling
    - ▶ Dictator game, both offer & recipient known. Significantly more offered to co-ethnics

## Habyarimana, Humphreys, Posner, Weinstein. “Why Does Ethnic Heterogeneity Undermine Public Goods Provision?”

- ▶ Try to understand the result, found in many contexts, that ethnically diverse places are less good at providing public goods. Develop three hypotheses:
  1. Preferences: members of the same ethnic group want the same things, or have other-regarding preferences and so want good things for each other
  2. Technology: co-ethnics either share linguistic advantages, or are more able to locate each other within social networks
  3. Strategy Selection: w/multiple equilibria, co-ethnics believe that the other will play cooperatively and therefore Nash solutions are self-fulfilling
- ▶ Conclude that strategy selection mechanisms, and to some extent technology, are responsible for improved provision of public goods

# Running field experiments: Practical, theoretical, and ethical considerations

Biases from Study Effects

Lab Experiments in the Field

The Critique of RCTs

# Running field experiments: Practical, theoretical, and ethical considerations

Biases from Study Effects

Lab Experiments in the Field

The Critique of RCTs

## Specific quantitative criticisms

- ▶ Core point: you can never estimate the treatment effect on any specific individual purely with an RCT, only on the mean difference between two randomized groups
- ▶ RCTs provide information only over the difference in means, but we might be just as interested in other quantities such as the median difference; this cannot be recovered
- ▶ Also have no way of recovering the fraction of the sample for whom the treatment effect is positive, for example
- ▶ ‘Gold Standard’ value of RCTs does not extend to the analysis of heterogeneity, since with sufficient covariates I can always find significance on some subgroups. “Wisdom after the event”



## Philosophical Critique of RCTs as Learning Devices

- ▶ Much of the Randomista critique of observational work is that the world contains too many unobservable confounds to possibly control for
- ▶ Which of these confounders are constant across contexts? If few of them are, experiments have no external validity. While I believe one study in one place, I don't have any theory as to how it would operate in other contexts
- ▶ Nancy Cartwright: “the benefit that the conclusions follow deductively in the ideal case comes with great cost: narrowness of scope.”
- ▶ Conclusion: If you are a Randomista without theory, you have no structure within which to build real knowledge of the world

## Imbens has five basic points in his rejoinder to the criticism of RCTs

1. Fundamental purpose of the new empirical thrust in social science is to identify a causal effect cleanly. It is more useful to have an internally valid estimand on a selected sample than it is to have an externally valid but biased estimand
2. Observational techniques such as the Heckman selection system or matching have fundamentally failed to replicate experimental results (Lalonde)
3. There is an epistemological hierarchy of evidence, and RCTs are on top
4. While LATEs (from RCTs, RDD and IVs) do not estimate average treatment effects in the presence of heterogeneous treatment effects (or essential heterogeneity), they do provide internally valid impact estimates on (sometimes) well-defined populations
5. On external validity, replication helps with RCTs but does not necessarily help in observational studies if they all replicate the same bias!

## Methodological Ways to Overcome Shortcomings of RCTs:

- ▶ Pre-analysis Plans
  - ▶ In order to reap the 'gold standard' quality of experimental evidence, you must have pre-committed to hypotheses
  - ▶ Explicitly laying out the hypotheses that you will test once you have your data
- ▶ Randomization Routines as Commitment
  - ▶ Analysis of heterogeneity is a particularly difficult area for experimental analysis
  - ▶ "As Ye Randomize, So Shall Ye Analyze"
- ▶ Focus on Theory and Mechanisms:
  - ▶ Have a clearer idea of WHY a treatment had an effect in a given context
  - ▶ Test theory about this mechanism
  - ▶ Position to make statements about the external validity

## Metastudies:

- ▶ Each particular study has its own error term
- ▶ Model as a two-level error structure where it attempts to see whether the aggregate evidence is significant given the degree of variation in each specific study
- ▶ If indeed the world is so unknowable and variable that each study is fundamentally incomparable to another, the metastudy will return insignificant results
- ▶ Must have large number of experimental studies on very comparable interventions (so far only CCT and micro-credit)