

Monitoring and Evaluation for Results

The Practice of Impact Evaluation

The Practice of Impact Evaluation

Evaluation of social programs, projects, or policies:

A systematic assessment to determine the relevance, efficiency, effectiveness, impact, and sustainability of a planned, ongoing, or completed intervention.

- Needs assessment
- Evaluability assessment and design assessment
- Process evaluation, implementation evaluation
- Tracer studies, mid-term evaluations, etc.
- Impact evaluation (e.g., “rigorous”) and meta-evaluation

What is impact evaluation?

Impact evaluations assess the specific outcomes attributable to a particular intervention or program. They do so by comparing outcomes where the intervention is applied against outcomes where the intervention does not exist.

An appropriate comparison represents what would have happened in the absence of the intervention. By establishing a valid comparison of outcomes for these two groups, an impact evaluation seeks to provide direct evidence of the extent to which an intervention changes outcomes.

Development Impact Evaluation (DIME) Initiative, World Bank (2007)

What is results-based monitoring?

- Results monitoring is a continuous process of collecting and analyzing information to compare how well a project, program or policy is performing against expected results. (Sometimes referred to as results-based management, management for results, and performance measurement.)
- Typical features
 - No need for comparison groups.
 - No inherent causality (but predicated on a results framework or logic model)
 - Performance targets are clear and measurable.
 - Distinct from, but complementary to, an impact evaluation

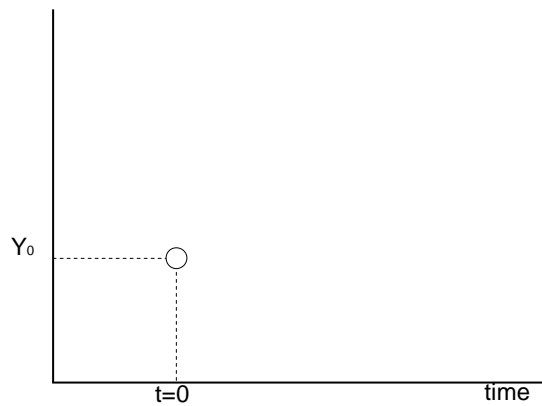
Types of Impact Evaluations

- Note that almost all types of impact evaluations have a comparison group and baseline data.
- The most widely used quantitative impact evaluation designs
 - See handout
- When baseline data do not exist, refer to Bamberger's work on *reconstructing baseline data*

When might it be appropriate to pursue impact evaluations?

- Similar projects, programs, or policies but with conflicting or divergent results
- **Allocating resources across** alternative investment options
- Identify issues around a problem (e.g., drug use, dropout rates)
- Pilot projects (however, if scientific basis already exists to justify cause-effect relationships, a “rigorous” impact evaluation may be unnecessary.)
 - If other evidence of effectiveness is good, and potential benefits are large, how much more evidence is necessary?
 - Difference between results monitoring and impact evaluation
- Increase transparency, credibility, or accountability
- Discussion: There are many times when it is inappropriate to conduct an impact evaluation. Examples?

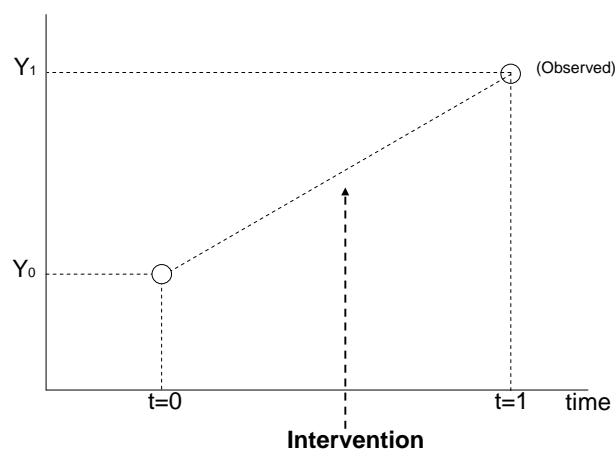
We observe an outcome indicator



WORLD BANK INSTITUTE
Promoting knowledge and learning for a better world

7

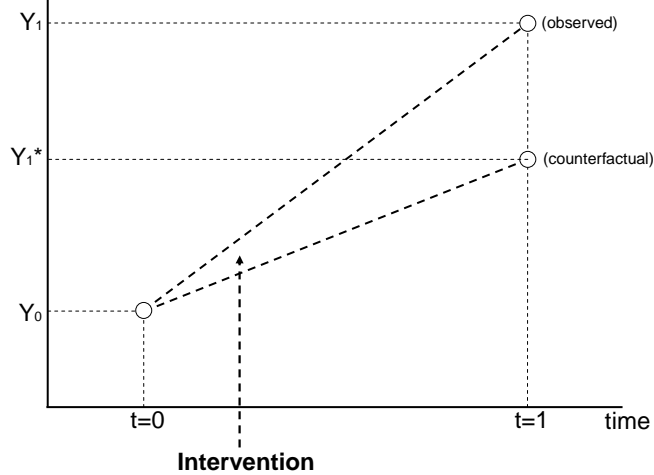
And its value rises after the intervention:



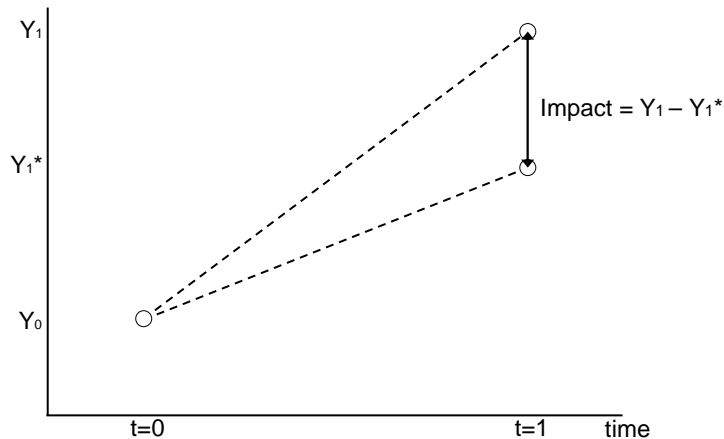
WORLD BANK INSTITUTE
Promoting knowledge and learning for a better world

8

Having the “ideal” counterfactual...



Allows us to estimate the true impact



The Basics of Program Evaluation

- Program evaluation is a set of techniques used to determine if an intervention 'works'.
- To determine the causal effect of the treatment we need knowledge of counterfactuals, that is, what would have happened in the absence of the intervention?
- The true counterfactual is not observable: Observational studies, or associations, will not estimate causal effects
- The key goal of all program/impact evaluation methods is to construct or "mimic" the counterfactual.

Continuation

- The control is intended to estimate what would have happened to the project population if the intervention had not occurred
- The difference between the change in t_1 for the project and t_1 for the control group is an estimation of impact
- This is called the Change Score

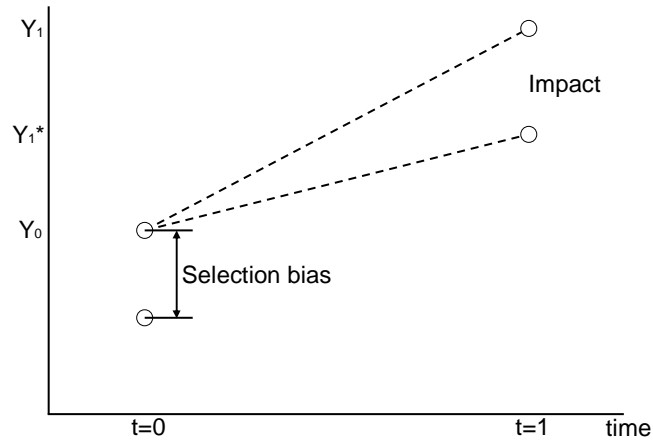
Continuation

- The validity of the change score as an estimate of impact depends on how closely the control group matches the treatment group on key characteristics relevant to the impacts being studied.

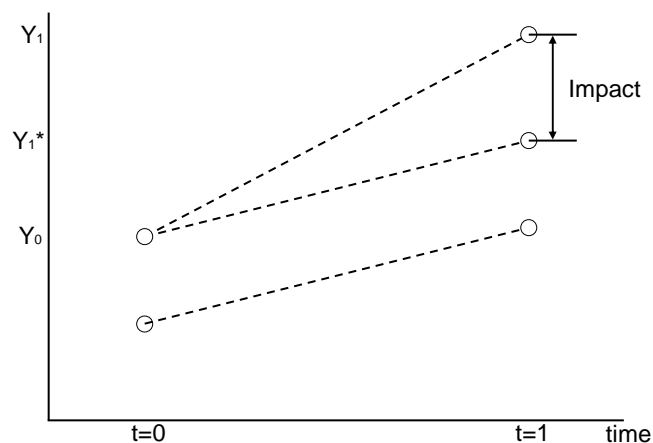
Special challenges in constructing comparison groups

- Selection bias
 - Project communities or individuals are purposefully selected to target
 - Groups most likely to be successful
 - The poorest and most vulnerable groups (many of which may be less likely to be successful)
 - Participants are self-selected
 - Often people/communities who apply are most likely to be successful

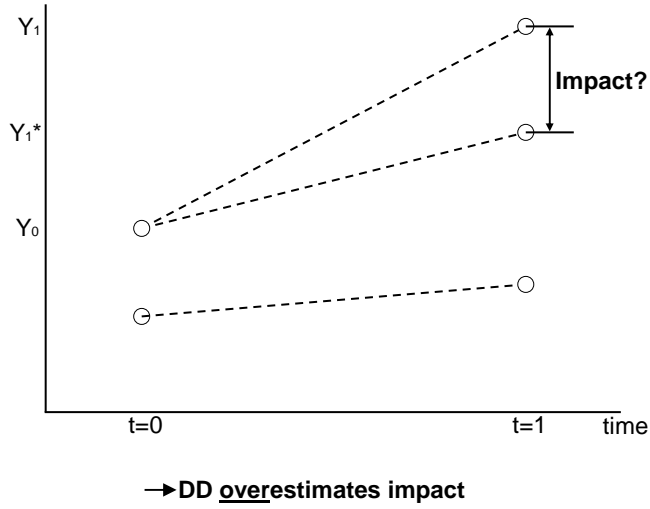
Selection bias



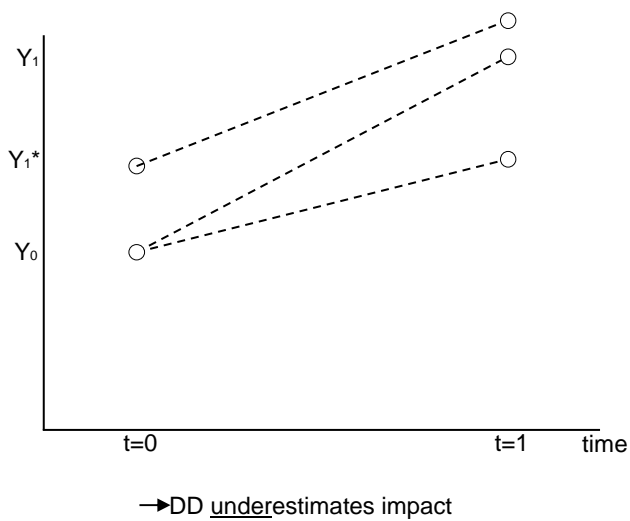
Difference-in-difference (“diff-in-diff”) requires that the bias is additive and constant over time



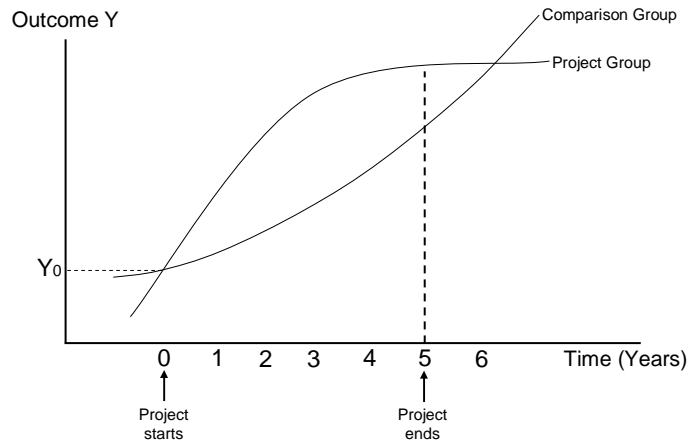
The method fails if the comparison group is on a different trajectory



Or...



When results are not additive and time invariant



Some commonly used approaches to constructing comparison groups

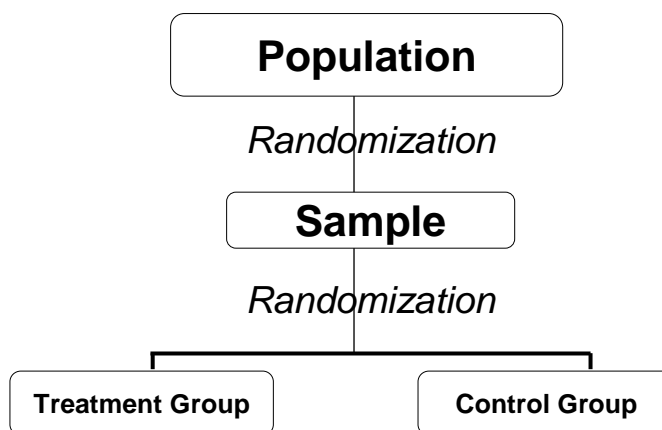
- Randomized allocation of treatment
- Pipeline
- Matching on observables
 - Judgmental matching (qualitative methods for matching, but including natural experiments, e.g., twins; geographic isolation)
 - Propensity score matching

Randomization

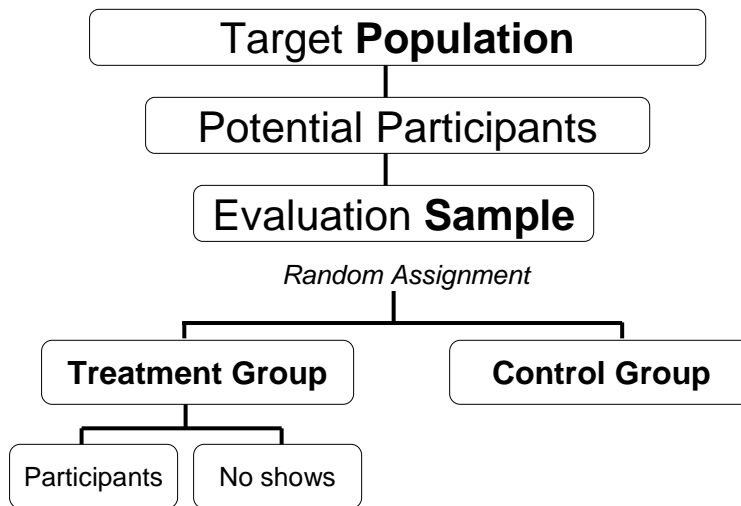
In the context of social programs, randomization involves:

- Equal probability of units' selection into the program
- Random assignment (in deciding where to put the program) and random selection (in deciding who participates in the program) to dispel selection bias

Setting up a randomized evaluation



Setting up a randomized evaluation



Based on Orr (1999)

Challenges

- Attrition (high dropout rates from control or treatment groups)
- Control group may be unwilling to cooperate or respond
- Hawthorne effects (threat to internal validity)
- Partial equilibrium analysis (threat to external validity). It may be difficult to extrapolate results from a pilot experiment to a national program
- Ethics and fairness of random assignment/selection
- Political feasibility of random assignment and selection for social programs

Pipeline approach

- What is the “pipeline” approach? (Exploits project/program design)
 - Applicants who have not yet received program form the comparison group
 - Assumes exogenous assignment among applicants
 - Reflects latent selection into the program
- Challenges:
 - Similar to those for randomized controlled trials discussed earlier

Propensity score matching

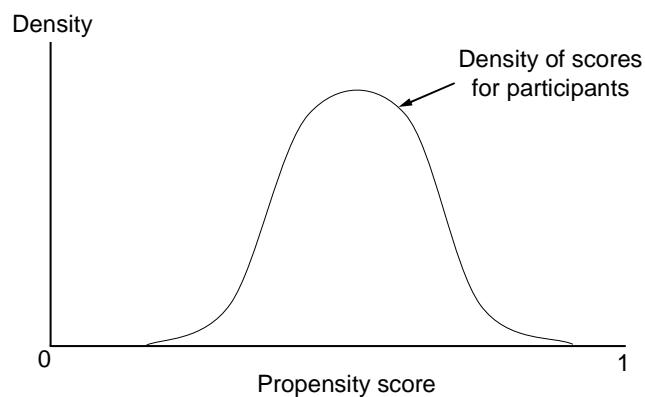
What is “propensity score matching” or PSM?

- PSM uses statistical methods to assign a score to each participant and non-participant based on observable characteristics. These scores - called propensity scores - are subsequently used to identify a control group.
- Typically used when there is no comparison group, although can be useful also for matching at baseline

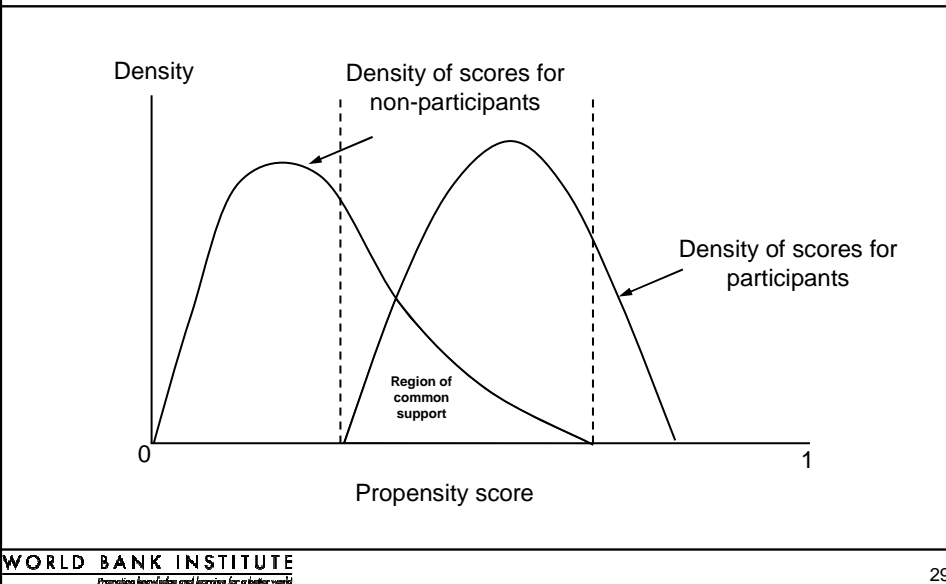
Steps in Score Matching

1. Representative & highly comparable surveys of non-participants and participants.
2. Pool the two samples and estimate a logit (or probit) model of program participation.
3. Restrict samples to assure **common support** (important source of bias in observational studies)
4. For each participant find a sample of non-participants that have similar propensity scores.
5. Compare the outcome indicators. The difference is the **estimate of the gain** due to the program for that observation.
6. Calculate the mean of these individual gains to obtain the average overall gain.

Distribution of propensity scores



Distribution of propensity scores: Participants versus non-participants



Challenges and limitations to PSM

- Requires large samples and good data
- Assumes no selection bias based on unobservable characteristics.
- Finding the “region of common support” may be a problem, in that there is only a small subset of participants from the “treated” and “untreated” groups with similar propensity scores

Other challenges to rigorous impact evaluation

- **Contagion** (e.g., spillover effects from “treatment” to “control” sites; introduction of another programs in either a “control” or a “treatment” site that may affect relevant outcomes there)
- **Budget, time and political constraints**, including administrative issues (e.g., timeliness of reporting results)
- There might be **different results depending on when the “before” or “after” data are collected**
 - For example, effects of interventions may not have had sufficient time to manifest themselves; or “before” data were collected too late.

Other challenges to “rigorous” impact evaluation

- **Technical sophistication vs. relevance**
 - May be unable to address relevant issues (e.g., the change processes, sustainability of impacts, unintended consequences, and so on).
 - May be difficult to change research instruments once study commences; some rigidity of procedures
 - Little, if any, information on contextual factors (motivation, self-confidence, political/social factors, etc.) influencing outcomes
 - May not be applicable for all social interventions (e.g., no control group; difficulties with collecting data from target beneficiaries on sensitive topics (e.g., drug addicts) or from remote villagers)

Sources

This presentation draws heavily on work done by World Bank sources, including:

- Martin Ravallion, DEC
- Emmanuel Skoufias and Markus Goldstein, PREM
- Sebastian Martinez and Paul J. Gertler, HDN
- Howard White, IEG
- Development Impact Evaluation (DIME) Initiative