

# Protecting the Vulnerable:

## The Design and Implementation of Effective Safety Nets



**December 2 - 13, 2002**  
**Washington, DC**

**The World Bank**

*Who gains from the program?  
How much do they gain?*

***Assessing policy and program  
impacts on poverty***

Martin Ravallion

## **Part 1: The evaluation problem**

- Concepts
- Alternative methods
- Recurrent problems in practice

## **Part 2: Examples from recent World Bank evaluations**

2.1 A workfare program

2.2 Wage subsidy + training

2.3 Dynamic tests of a safety net

# *Part 1. The Evaluation Problem*

To assess impact we need to measure the gains to participants at given levels of welfare, e.g., “income” gains to the “poor.”

The gain is the difference between participant’s welfare level with the program and that without it.

However, while a post-intervention welfare indicator is observed, its value in the absence of the program is not, i.e., it is a counter-factual.

The essential problem in evaluation is one of missing data on the counter-factual of what would have happened in the absence of the intervention

Naïve comparisons are still common:

- participants vs non-participants
- reflexive comparisons (before vs after)

**However it is known that such methods can be very deceptive....**

## *Benefits from rural roads?*

	Without road (n=56)	With road (n=44)	% increase (t-test)
<b>Case 1</b>	<b>1.29</b>	<b>2.41</b>	<b>87% (2.29)</b>
<b>Case 2</b>	<b>1.29</b>	<b>1.98</b>	<b>54% (2.00)</b>

Mean incomes in villages with and without a road  
(\$/day/person)

	Without road (n=56)	With road (n=44)	% increase (t-test)
<b>Case 1: Road yields 20% income gain</b>	<b>1.29</b>	<b>2.41</b>	<b>87% (2.29)</b>
<b>Case 2: Road yields no income gain</b>	<b>1.29</b>	<b>1.98</b>	<b>54% (2.00)</b>

Mean incomes in villages with and without a road  
(\$/day/person)

# Tools to fill in the missing data

***Randomization:*** Only a random sample is allowed to participate. “Randomized out” group is the counterfactual.

***Matching:*** Match participants to non-participants from a larger survey. The matches are chosen on the basis of similarities in observed characteristics.

***Propensity-score matching:*** Match on the basis of the probability of participation.

### ***Double difference:***

- Collect baseline data on non-participants and (probable) participants before the program.
- Compare with data after the program.
- Subtract the two differences, or use a regression with a dummy variable for participant.

### ***Matched double difference:***

- Match participants and non-participants based on observed characteristics
- Doing a double difference eliminates any time-invariant bias due to miss-matching, selection bias, omitted variables etc

***Instrumental variables:*** Use variables that influence participation -- but do not affect outcomes given participation -- to identify the exogenous variation in outcomes due to the program. The counter-factual is then identified.

***Simulation methods:***

- Theory driven; economic model identifies key parameters
- Calibration to survey data
- Simulate impacts; sensitivity tests to alternative theoretical assumptions

# Examples from World Bank evaluations

2.1 Workfare program → matching

2.2 Wage subsidy + training → randomization +  
instrumental variables

2.3 Dynamic tests of a safety net → simulation  
methods

# Recurrent problems in practice

## **Survey data and analysis take too long**

- Weak feedback into program implementation.
- Results come too late to make a difference

## **Project monitoring has little or no evaluative content**

- Plenty of data on inputs,
- but little on performance relative to a relevant counter-factual

## Recurrent problems cont.,

### **Programs have to be put in place quickly**

Neither randomization or baseline survey are feasible.

### **Allowing for unobservables**

- Latent variables may jointly influence program participation and outcomes
- This biases the results
- Finding valid instrumental variables is often difficult
- Exclusion restrictions especially problematic

## *Part 2. Examples from recent World Bank impact evaluations*

### **2.1 An example using propensity score matching methods: Argentina's Trabajar Programs**

With financial and technical support from the World Bank, the Government of Argentina introduced Trabajar 2 in May 1997 (expanded and reformed version of Trabajar 1).

## **The program aims to reduce poverty in two ways:**

Firstly, by providing short-term work at relatively low wages, it aims to self-select unemployed workers from poor families.

Secondly, the scheme tries to locate the projects in poor areas.

# The evaluation problem

- Randomization was not an option
- Nor was it possible to delay the program so as to do a baseline survey
- However, the statistics office was doing a new national survey six months after the program started
- The statistics office agreed to add on a survey of program participants

# Theory of score matching

- Ideally we would match on the entire vector  $X$  of observed characteristics.
- However, this is practically impossible.  $X$  could be huge.
- Rosenbaum and Rubin show that matching can be done using

***Propensity score = Estimated probability of participation given  $X$***

## Steps in score matching:

- 1:** You need representative, highly comparable, surveys of the non-participants and participants.
- 2:** Pool the two samples and estimate a logit model of program participation.
- 3:** Create the predicted values from the logit regression (“propensity scores”).

## Steps in score matching:

- 4: Some of the non-participant sample may have to be excluded at the outset because they have a propensity score which is outside the range (typically too low) found for the treatment sample.

Failure to assure “common support” can be an important source of bias in observational studies

## Steps in score matching:

**5:** For each participant you find the non-participant sample that has the closest propensity score (“nearest neighbor”). 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. Various weighting schemes,

# Measure of impact

$$\bar{G} = \sum_{j=1}^P (Y_{j1} - \sum_{i=1}^{NP} W_{ij} Y_{ij0}) / P$$

Various weighting schemes:

 Nearest k neighbors

 Kernel-weights

# *How does Propensity Score Matching compare to an experiment?*

- PSM is the observational analogue of an experiment in which placement is independent of outcomes
- The difference is that a pure experiment does not require the untestable assumption of independence conditional on observables.
- But PSM requires good data.

## *How does PSM perform relative to other methods?*

- In comparisons with results of a randomized experiment on a US training program, Heckman et al. and Dehejia and Wahba found that PSM can achieve a good approximation
- Much better than the non-experimental regression-based methods studied by Lalonde for the same program.

# Questions to be addressed by the Trabajar evaluation:

- *How income-poor are the participants?*
- *What are their net income gains?*
- *What non-income factors influence participation? Politics? “Social capital”*
- *Is there a gender bias? 15% of participants in the first six months were female. Why?*
- *Other forms of bias? Are the old given preference over the young?*

**The participation regression suggests that participants are more likely to be:**

.... poor, as indicated by their housing, neighborhood, schooling, and their subjective perceptions of welfare and expected future prospects

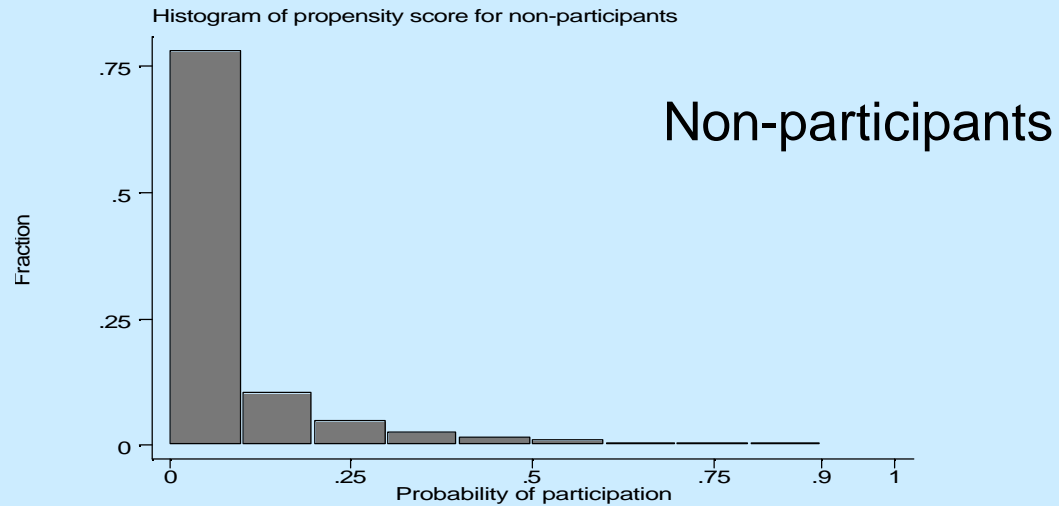
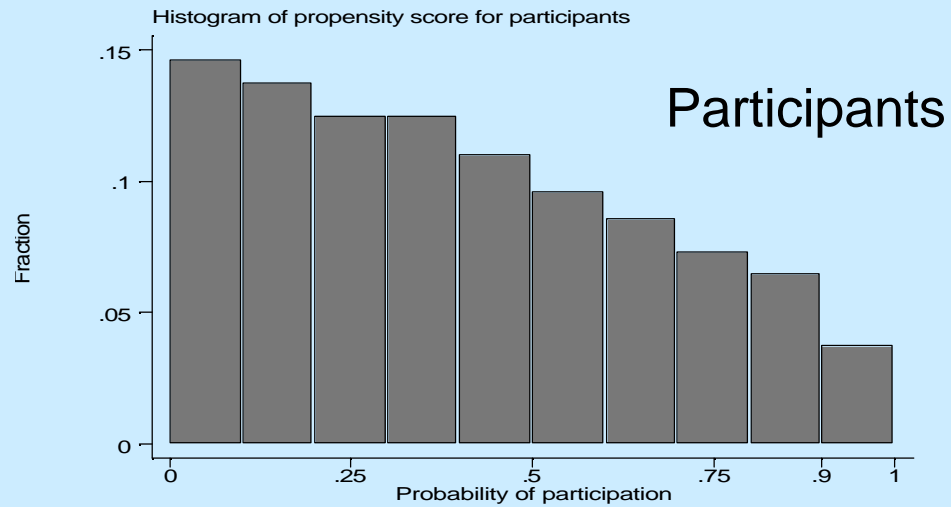
.... males who are head of households and married

.... longer-term residents of the locality rather than migrants from other areas;

.... members of political parties and neighborhood associations.

Social and political connections influence the likelihood of being recruited into a sub-project proposal. But the effect is not large.

# Propensity score distributions



# Estimated gains from Trabajar

The average gain is about half of the mean Trabajar wage.

80% of Trabajar participants have a pre-intervention income (income - net gain from the program) that puts them in the poorest 20% nationally.

Over half of the participants are in the poorest decile nationally.

# Bias in non-behavioral incidence

Standard incidence numbers underestimate how poor the participants would be without the program; over-estimate net gains.

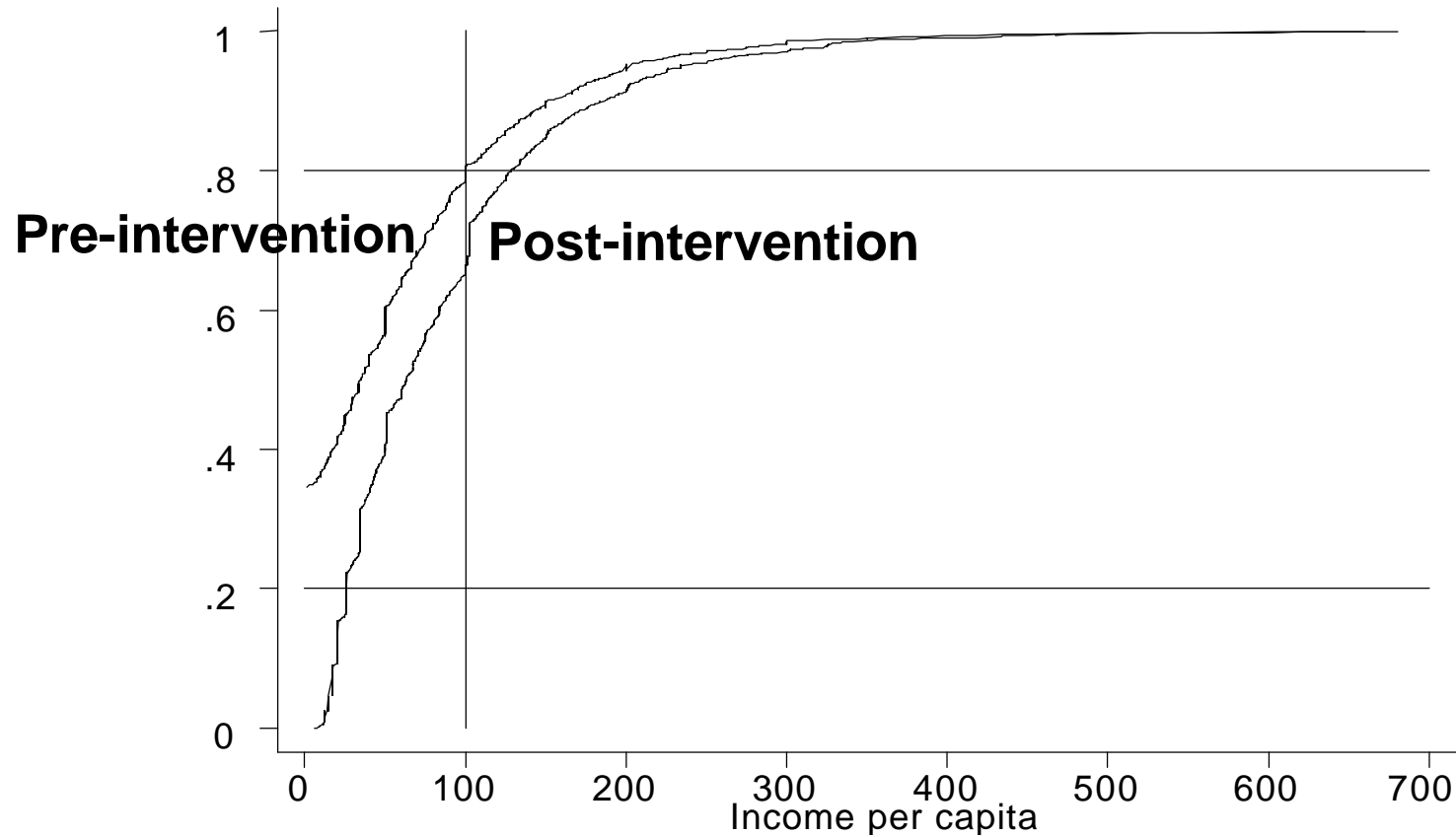
This bias is most notable is amongst the poorest 5%:

- while the non-behavioral analysis suggests that 40% of participant households are in the poorest 5%,
- the estimate factoring in foregone incomes is much lower at 10%.

# Incidence of income gains

Fractiles formed from the national income Distribution	Transfer benefit =wage	Factoring in foregone
Ventile 1	38.8	10.3
Ventile 2	21.3	42.4
Decile 2	18.5 (78.6)	26.8 (79.5)
Decile 3	9.5	10.9
Decile 4	5.8	6.4
Decile 5	1.9	2.0
Deciles 5-10	4.1	1.3

# *Impacts on poverty amongst participants*



# Single difference matching can be heavily biased when data are weak

## 1. *Single-diff. matching using a lighter survey instrument gave implausible results*

Latent heterogeneity due to lighter survey instrument (esp., missing social data)

## 2. *However, tracking individuals over time allows a matched double difference*

- addresses some of the limitations of single-difference on weak data
- allows us to study the dynamics of recovery

## 2.2 An experiment in helping poor people get off workfare

- Concerns about impact of Active Labor Market Programs. Inconclusive non-experimental evaluations
- A randomized evaluation of supplementary programs to assist the transition from the Trabajar Program to regular work.

# Experiment design

- A random sample of 850 Trabajar workers
- 280 got nothing; they formed the control group.
- The rest got a voucher that entitled them to a wage subsidy, received by any private-sector employer who hired that worker into a regular job. Subsidy= $\frac{3}{4}$  min.wage for 18 months.
- For 300 the voucher came with skill training; but 90 did not take this up.
- After a baseline survey, there were three follow-up surveys of all workers at six month intervals, spanning 18 months.
- Experiment was kept secret

## *What impact on employment?*

- By the final survey round the proportion of voucher recipients getting a private sector job was 14% versus 9% for the control group.
- This difference is statistically significant (5% level).
- The gains were confined to the young (under 30) and women

## *What impact on incomes?*

- There was no significant income gain for voucher recipients (for either total family income or labor earnings of the Trabajar participant).
- It appears that voucher recipients took up private sector jobs in the expectation of a higher and/or more stable stream of future incomes.

# *Did the training have impact?*

- No significant extra impact from the training.
- However, there could be bias due to endogenous compliance
- For example, if low skilled workers with little prospect of employment expect gains from training then underestimate impact
- Still no impact of training using 2SLS with assignment as the IV for treatment

# Endogenous take-up of training: Instrumental variables estimator

$D = 1$  if treated, 0 if control

$Z = 1$  if assigned to treatment, 0 if not.

$$D_i = Z_i \mathbf{p}_1 + \mathbf{h}_{1i} \quad \text{Compliance regression}$$

$$Y_i = Z_i \mathbf{p}_2 + \mathbf{h}_{2i} \quad \text{Outcome regression} \\ \text{("intention to treat effect")}$$

$$\frac{\hat{\mathbf{p}}_2}{\hat{\mathbf{p}}_1}$$

**2SLS estimator** (=ITT deflated  
by compliance rate)

# *Puzzle 1: No impact on incomes?*

- There was no significant income gain for voucher recipients (for either total family income or labor earnings of the Trabajar participant).
- It appears that voucher recipients took up private sector jobs in the expectation of a higher and/or more stable stream of future incomes.

## *Puzzle 2: low take-up by employers*

- Take up of the wage subsidy by firms amongst those who got a private job was low (just 3). (consistent with US experience)
- Hidden costs of take-up: social charges for registering the worker; severance pay; spillover to other workers

## *Supply-side effects?*

- Possibly those receiving the voucher were more confident in approaching potential employers,
- or possibly the latter took the voucher as some sort of indicator of the applicant's quality as a prospective worker.

# *The wage subsidy was cost-effective*

- It appears that the impact of the voucher was not through the access to a wage subsidy.
- Low subsidy take-up by employers
- So don't judge impact of a wage subsidy by its take-up rate
- Government saved 5% of its workfare wage bill for an outlay on subsidies = 10% of that saving
- Caveats on scaling up

## 2.3 Dynamic tests of the safety net

Conventional impact measures are static. Same poverty rate of 50% over time could mean same people are poor or completely different people are poor.

**Panel data allow us to construct the joint distribution over time**

<b>Persistently poor: Poor in both years</b>	<b>Escaped poverty: Poor in the first period, but not in second</b>	<b>Poor in first period</b>
<b>Fell into poverty: Not poor in the first period, but poor in second</b>	<b>Persistently non-poor: Not poor in either period</b>	<b>Not poor in first period</b>
<b>Poor in second period</b>	<b>Not poor in second period</b>	<b>Panel population</b>

# Measuring performance of the safety net

- **PROT ("Protected") = Change in proportion who fell into poverty.**
- **PROM ("Promotion") = Change in proportion who escaped poverty.**

## Examples: Social assistance in Hungary and Russia

- Fixed effects model used to estimate the consumption impact of transfers

$$\Delta C_i = \Delta \mathbf{a} + \mathbf{b} \Delta T_i + \Delta X_i \mathbf{g} + X_{i0} \Delta \mathbf{g} + \mathbf{e}_i$$

- Simulations of the joint distribution, with and without changes in transfers

# **Hungary: Higher social assistance reduced transient poverty**

- This was mainly due to better protection from poverty rather than promotion
- And it was mainly due to higher outlays rather than better targeting

## *How did Russia's safety perform in the 1998 crisis?*

- Widespread deterioration in welfare.
- Both gainers and losers at all levels.
- The safety net helped, but fell far short of what was needed to protect living standards
- Even without better targeting, a modest expansion of the safety net could have prevented an increase in poverty