

Researching a World Bank Lending Operation

***Assessing Social Protection:
Lessons from Ex-Post Evaluations
of Argentina's Trabajar Program***

Researching a World Bank Lending Operation

***Assessing Social Protection:
Lessons from Ex-Post Evaluations
of Argentina's Trabajar Program***

- 1: What is the incidence of gains to participants?*
- 2: Do they recover quickly from retrenchment?*
- 3: How can they be helped to find regular work?*

1. J. Jalan and M. Ravallion, “Estimating the Benefit Incidence of an Anti-Poverty Program by Propensity-Score Matching”

2. M. Ravallion, E. Galasso, T. Lazo and E. Philipp, “Do workfare participants recover quickly from retrenchment?”

3. E. Galasso, M. Ravallion and A. Salvia, “Assisting the transition from workfare to work: A randomized experiment in Argentina”

*The evaluation problem:
to measure the gains to participants
at given levels of welfare*

*The evaluation problem:
to measure the gains to participants
at given levels of welfare*

- The gain = participant's welfare level with the program **minus** welfare without the program.

*The evaluation problem:
to measure the gains to participants
at given levels of welfare*

- The gain = participant's welfare level with the program **minus** welfare without the program.
- 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 evaluation problem:
to measure the gains to participants
at given levels of welfare*

- The gain = participant's welfare level with the program **minus** welfare without the program.
- 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.

Tools to fill in the missing data

Randomization:

Randomized out group identifies counterfactual

Tools to fill in the missing data

Randomization:

Randomized out group identifies counterfactual

Matching:

Counterfactual=Matched non-participants from a larger survey; matched by observed characteristics.

Tools to fill in the missing data

Randomization:

Randomized out group identifies counterfactual

Matching:

Counterfactual=Matched non-participants from a larger survey; matched by observed characteristics.

Double difference:

Same counterfactual but time-invariant matching errors are eliminated by differencing over time

Tools to fill in the missing data

Randomization:

Randomized out group identifies counterfactual

Matching:

Counterfactual=Matched non-participants from a larger survey; matched by observed characteristics.

Double difference:

Same counterfactual but time-invariant matching errors are eliminated by differencing over time

Instrumental variables:

A valid IV identifies the exogenous variation in outcomes due to the program.

Argentina's Trabajar Program: Introduced in 1997 (pilot in 1996) with support from the World Bank

The program aims to reduce poverty by:

- **providing short-term work at relatively low wages, and**
- **locating the work in poor areas.**

Essential features of the program:

- Public information (provincial MOL offices)
- Local groups (municipalities and NGOs) are the sponsoring agencies.
- Proposals are first vetted for technical feasibility by trained staff.
- Sub-projects chosen by a points system until the budget is exhausted. About half are funded.
- Low wage rate

Assignment of responsibilities

- The center pays the labor cost; sponsoring agencies must come up with the rest
- Non-poor areas have an advantage in co-financing
- Project office in MOL manages budget, operational guidelines, monitoring, TA, handles payments to workers and disbursements.
- Center first decides the allocation across provinces (distribution of poor unemployed) Some flexibility to re-allocate based on demand.
- Provinces have effective control over final allocation (esp., for co-financing)

The points system rewards proposals:

- from poor areas,
- considered to yield larger public benefits
- from areas which have had relatively low participation
- that undercut the wage rate
- with higher share of labor costs in total project costs
- from sponsoring agencies that have a good track record of sub-project completion

Wage rate is set low for (1) self-targeting and (2) deterrence

- Wage rate set according to survey data on earnings distribution for full-time work
- Wage rate = $2/3$ of mean for the poorest decile in GBA. Max wage = min wage
- Same wage structure for the whole country, though incentives to undercut
- No other restrictions on eligibility
- Workers receive monthly payments directly by check
- World Bank disburses against payments to workers

Sub-projects

- Physical and social infrastructure:
 - pre-schools/crèches
 - health-posts
 - drainage and sanitation, water tanks
 - local roads
 - parks.
- 25 workers per project on average; mean duration of 5 months.
- 2/3 proposed by municipalities; 20% by NGOs; rest by provinces, national programs.

Other evaluation components

■ Poor-area targeting

- Spending map regressed on poverty map
- Design changes greatly enhanced poor-area targeting
- But targeting worsens with program contraction

■ Engineering assessments (*ex-ante* + *ex-post*)

■ Social assessments

- Trabajar was almost unique amongst public programs in reaching poor, remote rural areas;
- with high concentrations of minority groups;
- but concerns about gender bias; “a program for men”

1. What is the incidence of direct benefits to workers?

- Randomization was not an option
- Nor was it possible to delay the program 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: If pre-intervention outcomes are independent of participation given X then matching can be done using only the probability of participating given X :

$$P = P(X)$$

predicted value is the “propensity score”

- Key implication of conditional independence (“strong ignorability”):

$$E(G | X) = E[G | P(X)]$$

- This reduces the potentially high dimensional problem to a single dimension, provided $P(X)$ is known.
- Relying on conditional dependence means that data quality is crucial to PSM (more later)

How does PSM compare to an experiment?

- PSM is the observational analogue of an experiment in which placement is independent of pre-intervention outcomes
- The difference is that a pure experiment does not require the conditional independence
- A true experiment balances both observables and unobservables

How does PSM compare to regression-based methods?

- PSM does not make any assumptions about the model determining outcomes, beyond conditional independence
- PSM does not use all the data; valid comparisons cannot be made outside the region of common support.
- IVE makes an alternative conditional independence assumption, namely the exclusion restriction.

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.
- Smith and Todd question robustness of the Dehejia and Wahba results

Steps in propensity-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”).
- 4:** Check for common support
- 5:** Find best matches in terms of PS
- 6:** Compare the outcome indicators. The difference is the estimate of the gain due to the program
- 7:** Calculate the mean of these individual gains to obtain the average overall gain.

The mean impact estimator

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

The mean impact estimator

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

Various weighting schemes:

- Nearest k neighbors
- Kernel-weights:

$$W_{ij} = K_{ij} / \sum_{j=1}^P K_{ij}$$

$$K_{ij} = \frac{K[P(X_i) - P(X_j)]}{\sum_{j=1}^P K[P(X_i) - P(X_j)]}$$

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;
- well-connected: members of political parties and neighborhood associations (though the effect is not large.)

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%.

Distribution of direct income gains from the Trabajar program

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

2. Do participants recover quickly from retrenchment?

- What happens to Trabajar participants after they leave the program?
- Do retrenched workers recover the lost income from the program? How quickly?
- What can be learnt about the program's impact by tracking leavers and stayers over time?

Evaluation issues

- Two sources of selection bias:
 1. decision to join the program
 2. decision to stay or drop out
- There are observed and unobserved characteristics that affect both participation and income in the absence of the program
- *Past* participation can bring *current* gains for those who leave the program

Matching + double difference

- Propensity Score Matching of participants and non-participants based on observed characteristics
- Double Difference (DD): Difference in gains over time between participants and non-participants. This eliminates bias due to miss-matching (if time-invariant)

Double-Matched Triple Difference

- Match participants with a comparison group of non-participants
- Match leavers and stayers
- Compare gains to continuing participants with those who drop out

Triple Difference (DDD) =

DD for stayers - DD for leavers

Income of a Trabajar worker: $Y_{it}^T = Y_{it}^* + G_{it}$

Single difference: $E[Y_{it}^T - Y_{it}^C]$

Double difference: $E[\Delta(Y_{it}^T - Y_{it}^C)] = \Delta G_{it}$

Triple difference: $E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 1] - E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 0]$

“stayers”

“leavers”

$$E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 1] - E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 0] =$$

$$E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 1] - E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 0] =$$

$$[E(G_{i2} | D_{i2} = 1) - E(G_{i2} | D_{i2} = 0)] \quad \text{net gain from participation}$$

$$- [E(G_{i1} | D_{i2} = 1) - E(G_{i1} | D_{i2} = 0)] \quad \text{selection bias}$$

$$E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 1] - E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 0] =$$

$$[E(G_{i2} | D_{i2} = 1) - E(G_{i2} | D_{i2} = 0)] \quad \text{net gain from participation}$$

$$- [E(G_{i1} | D_{i2} = 1) - E(G_{i1} | D_{i2} = 0)] \quad \text{selection bias}$$

Sign of the selection bias?

If leavers have lower gains from participation then DDD underestimates the gain from the program

Test for whether DDD identifies gain to current participants

- Joint conditions for DDD to estimate gain to participants:
 - no current gain to ex-participants;
 - no selection bias in who leaves
- Third round of data allows a test: mean gains in round 2 should be the same whether or not one drops out in round 3

Data

- Sample of 1500 Trabajar participants in 3 provinces (Chaco, Mendoza and Tucuman);
- Tracked over time (6/12/18 months) from May 1999
- Comparison group from a national household survey
- Administered the same questionnaire
- Rotating panel (1/4 of households replaced each round)
- Drop-out due to:
 - rotation (sub-projects last 6 months)
 - cuts to the number of new projects approved
 - selection bias?

Matching participants with non-participants in first survey

A person is more likely to participate if:

- young; male; less educated
- lives in house with only 1 or 2 rooms
- is renting the house
- is in a large/extended family
- with a lower fraction of migrants
- and a low fraction of children attending school

Matching stayers and leavers

A person is less likely to drop out from Trabajar if:

- participating in neighborhood associations
- employed in past as a temporary worker
- entered Trabajar through personal contacts

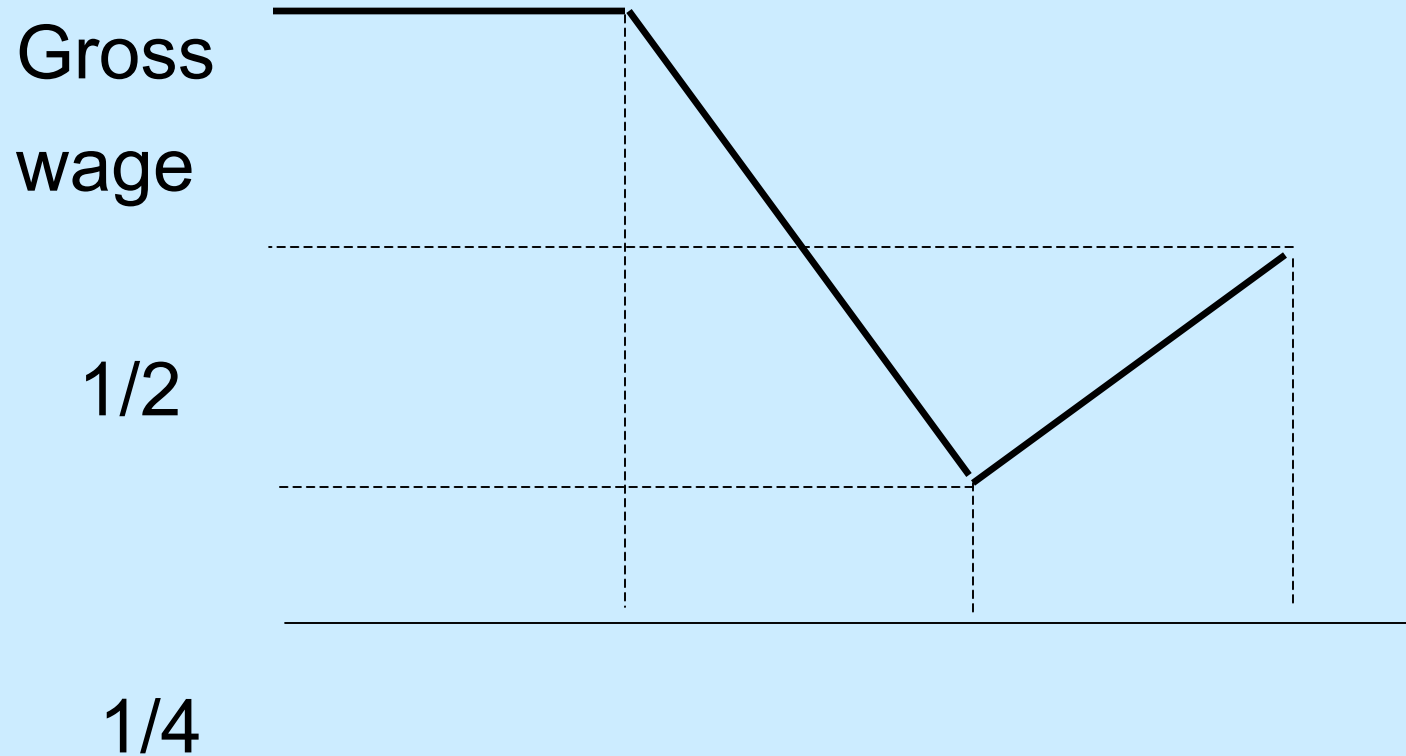
However, weak explanatory power for drop-outs;
consistent with exogenous rationing

Triple difference....

Findings on impact

- The income losses for leavers are about $\frac{3}{4}$ wage after 6 months
- Loss is smaller in areas with lower levels of unemployment
- Over time (after 12 months) some losses are recovered to around less than $\frac{1}{2}$ wage
- Post-program Ashenfelter dip
- Joint test passes; DDD identifies gain to participants
- Yes, qualitative evidence of expected longer-term gains (jobs, skills, contacts)

Post-program dip



Implications for evaluation methods

1. *Single-diff. comparison of participants with matched nonparticipants can be highly misleading without good data:*

- Single-diff results are implausible in this case
- Latent heterogeneity due to lighter survey instrument (esp., missing social data)

Implications for evaluation methods

1. *Single-diff. comparison of participants with matched nonparticipants can be highly misleading without good data:*

- Single-diff results are implausible in this case
- Latent heterogeneity due to lighter survey instrument (esp., missing social data)

2. *However, tracking individuals over time:*

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

3. *Single diff, leavers vs. stayers does well*

3. How can participants be helped to leave the program?

Concerns about impact of Active Labor Market Programs. Inconclusive non-experimental evaluations

3. How can participants be helped to leave the program?

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.

Confluencia, in the province of Neuquen

- 1993: downsizing and privatization of the state-owned oil company
- 1998: the Trabajar participation rate was still unusually high; 28% of people living in poor households that included an unemployed worker; corresponding national figure was 5%.
- However, the incidence of poverty with unemployment was not unusually high; 3.5% in Confluencia versus 4.2% nationally

The randomized experiment

- 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

Impacts of training?

- 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 compliance with training: Instrumental variables estimator

$D = 1$ if treated, 0 if control

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

$$D_i = Z_i\pi_1 + \eta_{1i} \quad \textbf{Compliance regression}$$

$$Y_i = Z_i\pi_2 + \eta_{2i} \quad \textbf{Outcome regression}$$

(“intention to treat effect”)

Endogenous compliance with training: Instrumental variables estimator

$D = 1$ if treated, 0 if control

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

$$D_i = Z_i\pi_1 + \eta_{1i} \quad \text{Compliance regression}$$

$$Y_i = Z_i\pi_2 + \eta_{2i} \quad \text{Outcome regression}$$

(“intention to treat effect”)

$$\frac{\hat{\pi}_2}{\hat{\pi}_1}$$

$$\hat{\pi}_1$$

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

4. Monitoring poor-area targeting performance

- Track spending across local government areas within each province.
- Poverty maps based on large sample surveys or census data. (Poverty data do not include program participation.)

- Performance measured by the “targeting differential” given by the regression coefficient of spending on the poverty rate across local areas within each province.
- If there is horizontal equity within a province then the targeting differential measures the expected difference in spending between the poor and the non-poor for a given province

Three Targeting Differentials

- The interprovincial TD, T^P
- The national interdepartmental TD, T^D
- The province specific TD, T_j for province j.

Exact decomposition:

$$T^D = SS^P \cdot T^P + \sum SS_j \cdot T_j$$

total *between* *within*
 provinces *provinces*

where SS_j is j's share of the total sum of squared deviations from national mean, and SS^P is the between province share.

Findings for Trabajar Programs

Under Trabajar 1, the inter-province Targeting Differential was \$25, and significantly different from zero at the 5% level.

This changed dramatically in Trabajar 2; T^P rose to \$74 which is highly significant ($t=4.85$).

But what about the allocation within provinces?

Under Trabajar 1, T^D was \$41 ($t=4.29$). Trabajar 1 was targeted to poor areas, despite poor performance in reaching poor provinces. T^D for Trabajar 2 rose to \$80 ($t=10.33$).

So there was an improvement in performance at reaching the poor across the country as a whole.

How much of this was due to the improved performance in targeting poor provinces?

17% of the Targeting Differential for Trabajar 1 was due to the allocation between provinces; the rest was due to targeting within provinces.

36% of the improvement in targeting performance across departments was due to better targeting of poor provinces. The rest (64%) was due to better targeting of the poor within provinces.

The bulk of the gain was due to intraprovincial targeting of poor areas

- The provinces differed greatly in their success at reaching poor areas.
- Signs that poorer provinces were less able to improve targeting with a higher budget.

Conclusions on the rapid assessment method

- The reforms to the Trabajar program greatly enhanced the extent to which it is targeted to provinces with high incidence of unmet basic needs
- The reforms improved the center's targeting of poor provinces, which enhanced overall performance in reaching the poor

- But the bulk of the gains were from the (diverse) provincial performances
- Some simple tools for program monitoring can provide rapid evaluative feedback to policy makers
- This reduces the uncertainty about impact, and helps improve performance incentives