What Can Ex-Participants Reveal about a Program's Impact?

Martin Ravallion
Emanuela Galasso
Teodoro Lazo
Ernesto Philipp

ABSTRACT

We propose a method for estimating the mean impact of an assigned social program when it is not feasible to do a pre-intervention baseline survey but it is feasible to track ex-participants. In our triple-difference estimator, measured outcome changes are compared between continuing participants and matched ex-participants, after netting out the outcome changes for a matched comparison group who never participated. With sufficient followup observations one can test the joint conditions required for correctly identifying the gains to current participants. We apply the method to a workfare program in Argentina. Significant impacts on participants' current incomes are revealed.

Martin Ravallion and Emanuela Galasso are with the Development Research Group at the World Bank. Teodoro Lazo and Ernesto Philipp are with the Trabajar project office of the Ministry of Labor, Government of Argentina.

The work reported in this paper is part of the ex-post evaluation of the World Bank's Social Protection III Project in Argentina. The authors' thanks go to staff of the Trabajar project office in the Ministry of Labor, Government of Argentina, who have helped in countless ways, and to the Bank's Manager for the project, Polly Jones, for her continuing support of the evaluation effort, and many useful discussions. The authors also benefited from comments from Hide Ichimura, Jyotsna Jalan, two referees, and seminar participants at Boston University, Columbia University, University College London, Yale University, the Network on Inequality and Poverty of the Latin American and Caribbean Economics Association, and the World Bank. Support from the Evaluation Thematic group of the World Bank's Poverty Reduction and Economic Management Network is gratefully acknowledged. The data used in this article can be obtained from the authors. These are the views of the authors, and need not reflect those of the Government of Argentina or the World Bank.

Correspondence: Martin Ravallion, World Bank, 1818 H Street NW, Washington DC, 20433 USA; <mravallion@worldbank.org>

[Submitted February 2002; accepted September 2003]
ISSN 022-166X E-ISSN 1548-8004

I. Introduction

It is well recognized that single difference comparisons of outcome measures between participants and nonparticipants in a social program can give severely biased estimates of impact. In attempting to reduce this bias, comparisons are often confined to observationally similar ("matched") units, and the Propensity Score Matching (PSM) method of Rosenbaum and Rubin (1983) has attracted recent interest as a flexible means of balancing observed covariates between the two groups. However, the problem of selection bias remains, that is, there may be latent differences between the two groups in characteristics that jointly influence participation and outcomes; selection bias violates the conditional independence assumption underlying PSM.

A popular approach for addressing this problem is the difference-in-difference (DD) estimator, obtained by comparing mean outcome changes over time between treatment and comparison groups, relative to the outcomes observed for a pre-intervention baseline. This eliminates all separable time-invariant bias. However, there is still a bias in DD estimators when the subsequent outcome changes are a function of initial conditions that also influence program placement.[1] Thus it is still important to ensure that the treatment and comparison groups are similar. It has been argued that combining PSM with DD can greatly reduce (but not eliminate) the bias found in other nonexperimental evaluations (Heckman, Ichimura, and Todd 1997; Heckman et al. 1998; Heckman and Smith 1999; Smith and Todd 2001).

However, what if one does not have a pre-intervention baseline? This is quite common for safety-net interventions that have to be set up quickly, in response to a macroeconomic or agro-climatic crisis. There is no time to do a baseline survey.

Suppose instead that we follow up samples of participants and nonparticipants over time, post-intervention, and that some participants become nonparticipants. What can we then learn about the program's impacts? Can we still identify the mean gain to current participants—the classic "treatment effect on the treated" as it is called in the evaluation literature?

The approach proposed here is to compare the observed income changes between those who leave the program ("leavers") and those who do not ("stayers"), with these two groups matched by propensity scores derived from their observed characteristics. Since there may also be economy-wide changes that have nothing to do with the program and may have different implications for leavers versus stayers in the absence of the program, we also track income changes for a matched comparison group of nonparticipants. Thus our estimation method requires matching both initial participants with nonparticipants as well as matching leavers with stayers. We then calculate a triple-difference, namely the difference between the DD for matched stayers and leavers.

While this approach is feasible without a baseline survey, it brings its own problems. Firstly, while differencing over time can eliminate bias due to latent (time-invariant) factors, there remains a potential bias due to any selective retrenchment

1. Jalan and Ravallion (1998) show that this can seriously bias evaluations of poor-area programs targeted on the basis of geographic characteristics that also influence the growth process.

from the program based on unobservables. We argue that the direction of bias can be determined under plausible assumptions. Secondly, there may well be a post-program "Ashenfelter's Dip," namely when earnings drop sharply at retrenchment, but then recover.[2] As in the preprogram dip, this is a potential source of bias in assessing the longer-term impact, although (as with the preprogram version) to the extent that the dip entails a welfare change it can still be relevant to assessing the short-term impact of a safety-net intervention. And the post-program dip is of interest in assessing the dynamics of recovery from retrenchment. Thirdly, there is the problem that past participation may bring current gains to those who leave the program. Assuming these lagged gains are positive, the net loss from leaving the program will be less than the gain from participation relative to the counter-factual of never participating. By exploiting further followup surveys, we offer a test for the joint conditions needed to identify the mean treatment effect on the treated, allowing for selection bias in who leaves and for lagged income gains.

As an application, we study a workfare program, which imposes work requirements on welfare recipients. The welfare outcomes of cutting such a program will depend in part on labor market conditions facing the participants. A high level of income replacement after retrenchment might suggest that unemployment is not a serious poverty problem. But even when there is high unemployment, there are other ways that retrenched workers might recover the lost income. Possibly the work experience on the program will help them find work, including self-employment. Or possibly private transfers will help make up for the loss of public support. Tracking ex-participants after their retrenchment and measuring their income replacement may thus provide important clues to understanding the true impact of a workfare program.

The specific program we study is Argentina's Trabajar Program. This aims to provide work to poor unemployed workers on approved subprojects of direct value to poor communities. The subprojects cannot last more than six months, though a worker is not prevented from joining a new project if one is available. In earlier research on the same program, Jalan and Ravallion (2003) estimated the counter-factual income of current participants if they had not participated using the mean income of a comparison group of nonparticipants, obtained by PSM. For the present study, we designed a survey of a random sample of current participants, and returned to the same households six months later, and then 12 months later. In addition to natural rotation, there was a very sharp contraction in the program's aggregate outlays after the first survey.

The following section outlines our evaluation method in theoretical terms. Section III describes the program and the data for its evaluation. Section IV presents our results, while some conclusions can be found in Section V.

II. Impact Estimation Method

People who join a social program are likely to differ from those who do not, and people who leave the program are potentially different to those who stay.

2. "Ashenfelter's dip" refers to the bias for inferring long-term impacts that can arise when there is a preprogram earnings dip (Ashenfelter 1978).

And there can be no a priori presumption that the same factors determine the first selection process as the second. Our strategy for estimating a program's impact without either randomized assignment or a pre-intervention baseline survey is to compare measured outcome changes between those who stay in the program and those who leave. To allow for economy-wide changes that have nothing to do with the program (such as changes in the overall unemployment rate) we also net out the outcome changes for an observationally similar comparison group of people who never participated. In this section we spell out conditions under which this triple-difference estimator identifies the current gains to participants.[3]

The gain in the outcome measure for a given program participant at a given date is defined as the difference between the actual outcome measure while participating and its value if the program did not exist. The gain from the program at date t can be defined as:

$$(1) \quad (G_t | D_t = 1) \equiv (Y_t^{T*} - Y_t^{C*} \mid D_t = 1)$$

where Y_t^{T*} is the true value of the outcome variable, Y_t^{C*} is the true value of the counter-factual outcome for the participant, and D_t is an indicator for actual participation, taking the value 1 for participants and 0 for nonparticipants. However, both the outcome variable under treatment and the counter-factual value are measured with error. We can write the measured variables as:

$$(2.1) \quad Y_t^T = Y_t^{T*} + \eta^T + \varepsilon_t^T$$

$$(2.2) \quad Y_t^C = Y_t^{C*} + \eta^C + \varepsilon_t^C$$

where η^i $(i = T,C)$ are time-invariant error components (such as due to selection bias) and ε_t^i $(i = T,C)$ are zero-mean time-varying error terms (allowing for time varying selection bias and measurement errors).

We focus on $T = 2$, though we consider an extension to $T = 3$. In the following exposition we assume that an estimate of Y_t^C is available for an observationally similar comparison group. In our empirical work we will use propensity score matching to clean out observable heterogeneity prior to using our triple difference estimator to deal with selection bias based on unobservables. The Appendix reviews the methods we use for matching, which are reasonably standard.

The identifying assumption in all double-difference studies is that the selection bias into the program is both additively separable from outcomes and time-invariant:

$$(3) \quad E[(Y_t^C - Y_{t-1}^C) | D_t = 1] = E[(Y_t^C - Y_{t-1}^C) | D_t = 0]$$

Under this assumption, the overall difference-in-difference can be written as:

$$(4) \quad DD = E[(Y_t^T - Y_{t-1}^T) | D_t = 1] - E[(Y_t^C - Y_{t-1}^C) | D_t = 0]$$
$$= E[G_t - G_{t-1} | D_t = 1]$$

In the standard DD setup with two time periods ($T = 2$), Period 1 precedes the intervention and $G_1 = 0$. Then DD gives the mean current gain to participants Time 2:

3. The only prior use of a "triple-difference" method that we know of in the literature is Gruber (1994) though the application is different; Gruber included interaction effects between time and location (as well as separate time and location effects) in modeling the earnings effects of labor laws in the United States.

(5) $DD = E(G_2 | D_2 = 1)$

However, in our case, the program is in operation in Period 1. The scope for identification arises from the fact that some participants at Date 1 subsequently drop out of the program. Our triple-difference estimator is the difference between the double difference for stayers and leavers; for $T = 2$ this can be defined as:[4]

$$(6)\quad DDD = E[(Y_2^T - Y_2^C) - (Y_1^T - Y_1^C) | D_2 = 1, D_1 = 1] \\ - E[(Y_2^T - Y_2^C) - (Y_1^T - Y_1^C) | D_2 = 0, D_1 = 1]$$

This can also be written in the form:

$$(7)\quad DDD = [E(G_2 | D_2 = 1, D_1 = 1) - E(G_2 | D_2 = 0, D_1 = 1)] \\ - [E(G_1 | D_2 = 1, D_1 = 1) - E(G_1 | D_2 = 0, D_1 = 1)]$$

The first term in square brackets on the RHS is the net gain to continued participation in the program, given by the difference between the gain to participants in Period 2 and the gain to those who dropped out. Notice that there may be some gain to leavers from past participation ($E(G_2 | D_2 = 0, D_1 = 1) \neq 0$). For example, participants may have learned a skill that raises their future earnings after leaving the program. Thus $E(G_2 | D_2 = 1, D_1 = 1) - E(G_2 | D_2 = 0, D_1 = 1)$ gives the loss to those who leave the program, allowing for the possibility that leavers may benefit from past participation. The second term on the RHS of Equation 5 is the selection bias arising from any effect of the gains at Date 1 on participation at Date 2.

It is readily verified from Equation 5 that *DDD* consistently identifies the mean gain to participants $E(G_2 | D_2 = 1, D_1 = 1)$ if the following two conditions hold:

Condition 1: There is no selection bias in who leaves the program, that is,

(8) $E(G_1 | D_2 = 1, D_1 = 1) = E(G_1 | D_2 = 0, D_1 = 1)$

Condition 2: There are no current gains to nonparticipants, that is,

(9) $E(G_2 | D_2 = 0, D_1 = 1) = 0$

A third survey round allows a joint test of these two sufficient conditions for interpreting *DDD* as an estimate of the gains to current participants in Period 2. Suppose one decomposes the aggregate estimate of *DDD* at Time 2 according to whether or not a person leaves the program in the third survey round. Assuming that Conditions 1 and 2 hold, the value of *DDD* can be exactly decomposed as:

$$(10)\quad DDD = E(G_2 | D_3 = 1, D_2 = 1, D_1 = 1)\Pr(D_3 = 1 | D_2 = 1, D_1 = 1) \\ + E(G_2 | D_3 = 0, D_2 = 1, D_1 = 1)\Pr(D_3 = 0 | D_2 = 1, D_1 = 1)$$

If there is also no selection bias in who drops out at $t = 3$, then:

(11) $DDD = E(G_2 | D_3 = 1, D_2 = 1, D_1 = 1) = E(G_2 | D_3 = 0, D_2 = 1, D_1 = 1)$

4. Suppose that we do not have a comparison group of nonparticipants; instead, we just calculate the double difference for stayers versus leavers (that is, the gain over time for stayers less that for leavers). It is evident that this will only deliver an estimate of the current gain to participants if the counter-factual changes over time are the same for leavers as for stayers. More plausibly, one might expect stayers to be people who tend to have lower prospects for gains over time than leavers in the absence of the program. Then the simple *DD* for stayers versus leavers will underestimate the impact of the program.

In other words, if our two sufficient conditions for *DDD* to be a valid estimate of the gain to current participants hold, and there is no selection bias at $t = 3$ then there should be no difference in the estimate of gains to participants in Period 2 according to whether or not they drop out in Period 3. We will test this implication. If it fails to hold then there could either be lagged gains or selection bias (at $t = 2$ and/or $t = 3$) or some combination of the two.

III. The Program and Data

In response to a sharp increase in the measured unemployment rate, the Government of Argentina greatly expanded and redesigned its Trabajar Program in May 1997, with financial and technical support from the World Bank. The Trabajar Program aims to provide short-term work at relatively low wages on socially useful projects in poor areas. The projects are proposed by local (governmental and non-governmental) organizations with priority given to proposals that are likely to benefit poor areas, according to ex-ante assessments.[5] Workers cannot join the program unless they are recruited to an approved project. The projects last a maximum of six months, but a worker is not prevented from switching to a new project on the same basis.

The wage rate was initially set at \$200 per month, which was cut to \$160 in 1999 at the time of an overall contraction in outlays. The aim in setting the wage rate was to ensure good targeting performance, and to help ensure that workers would take up regular work when it became available. Earnings data from the October 2000 Permanent Household Survey for Argentina indicate that 95 percent of workers in full-time jobs (35 hours or more per week, whether formal or informal sector) earned more than the prevailing wage rate on the Trabajar program.

The data collection for this study began with a survey in May/June 1999 of Trabajar participants in the main urban areas of three provinces—Chaco (Gran Resistencia), Mendoza (Gran Mendoza) and Tucuman (Gran Tucaman—Tafi Viejo). These provinces were chosen as representing the range of labor markets found in Argentina. The families of 1,500 randomly chosen Trabajar workers were interviewed, spread evenly among the three provinces. This was a simple random sample from the list of all beneficiaries at the time. The households of participants were the units for interviewing.

The survey of participants was chosen to coincide with the twice-yearly Permanent Household Survey (PHS). This is an urban survey focusing on employment and incomes, though it also includes questions on education and demographics. We calculate individual income from questions on income from work (wages, bonuses, self-employment income, Trabajar earnings) and from nonlabor sources (pension, rents, dividends, fellowships, food coupons, private transfers). Provincial capitals or other urban centers with at least 100,000 inhabitants are included in the PHS. The survey is conducted twice a year, around May and October. The PHS sample size is set to achieve (with 95 percent confidence) a standard error of 1 percent

5. Ravallion (2000) examines the program's ex-post performance in targeting poor areas.

in the unemployment rate within each urban conglomerate. In large conglomerates, a random sample of geographic units is chosen, within which a random sample of households is drawn. In smaller conglomerates, a one-stage random sample is used.

The PHS is our source of the comparison group for initial participants. For program participants, the same interview questionnaire was used as for the PHS, though with some extra questions added. By using common survey instruments and interviewers we hope to avoid the matching bias that can arise when the surveys of participants and nonparticipants are not comparable (Heckman, Ichimura and Todd 1997; Heckman et al. 1998). To help improve matching, we also selected separate comparison groups from each province, to help ensure that the individuals belong to the same local labor markets.[6]

A followup survey of the same Trabajar participants was done in October/ November 1999, to coincide with the next round of the PHS, and similarly in May/June 2000.[7] The PHS has a rotating panel design with one quarter replaced each round, so it was possible to form a panel for the comparison group. Our matches were constrained to only include those who would be followed up. Naturally this limits the matching options, particularly so by the second followup survey, by which time only half of the original sample is resurveyed.

The PHS is a far shorter survey instrument than that used by Jalan and Ravallion (2003) for their single difference estimate of the impact on incomes of Trabajar participation. Since there are fewer observables in the data, the matching is unlikely to be as good. Results in the literature suggest that single-difference PSM estimates can be unreliable when the data available do not include important determinants of participation (Heckman et al. 1997, 1998; Smith and Todd 2001). However, here we have the advantage that we can follow up participants over time, exploiting the rotating panel design of the PHS. Thus, although we cannot expect that our single difference PSM estimates will be as reliable as in Jalan and Ravallion, we can eliminate the time-invariant errors due to miss-matching arising from violations of the conditional independence assumption.

The followup surveys indicate a sharp contraction in participation; 49 percent of the Trabajar workers interviewed in the first survey had left the program by the second survey (Table 1). Only 16 percent of the original Trabajar workers were employed on the program by the second followup survey. This contraction in employment on the program did not appear to stem from a "pull" effect from the rest of the economy. There was little sign of economic recovery between the surveys. The overall unemployment rate increased in one of the provinces (Chaco) and fell, but not greatly, in the other two; see Table 2, which also gives unemployment rates for the second follow up survey, six months later, and for six months prior to the first survey.

The large number of initial participants leaving the program appears instead to stem from aggregate program contraction. In normal times, there is a rotation process arising from the fact that subprojects do not last longer than six months. When a subpro-

6. Heckman et al (1998) find that the mismatch due to different questionnaire and different labor markets amounts to half of the selection bias in their analysis.

7. We dropped 25 participants who left between Rounds 1 and 2 and rejoined for Round 3. A fourth survey was done six months later. More than 90 percent of the initial Trabajar workers had left the program by the fourth wave. There were too few continuing participants to facilitate further analysis.

Table 1
Trabajar Participation Rates Across Survey Rounds

	May 1999 (Baseline Survey)	October 1999 (First Follow-up Survey)	May 2000 (Second Follow-up Survey)
Total interviewed	1,459	1,332	1,291
Participants (percent of total interviewed)	1,459 (100 percent)	632 (47.4 percent)	212 (16 percent)
Chaco	504	149 (34 percent)	17 (4 percent)
Mendoza	474	285 (63 percent)	146 (32 percent)
Tucuman	481	198 (44 percent)	49 (11 percent)
Percent nonparticipant who are employed		49.0 percent	60.5 percent

Source: Authors' calculations from the Trabajar sample.

Table 2
Unemployment Rates in the Selected Provinces and Nationally

Percent of the Labor Force Unemployed (Urban Areas)	October 1998	May 1999 (Baseline Survey)	October 1999 (First Followup Survey)	May 2000 (Second Followup Survey)
Chaco	11.3	9.5	12.4	10.4
Mendoza	5.7	7.6	6.8	9.8
Tucuman	14.9	19.2	15.9	19.9
All urban areas	12.4	14.5	13.8	15.4

Source: Trabajar Project Office, Ministry of Labor.

ject ends, its beneficiaries are not automatically incorporated in another project. In the country as a whole, 45 percent of Trabajar workers participated in only one project (46 percent in Chaco, 52 percent in Mendoza and 51 percent in Tucuman). On top of this designed rotation, there was a severe contraction in aggregate outlays on the program starting at the end of 1999. This was an outcome of overall fiscal austerity, to keep Argentina within macroeconomic targets. Aggregate spending on the program by the central government in the first five months of 2000 was only 29 percent of its level in the last five months of 1999. Since the central government only covered the wage payments to Trabajar workers, this meant an equi-proportionate reduction in Trabajar employment. Existing subprojects were completed, but the number of new projects approved shrank

Table 3
Average Wage Rate for Trabajar Projects in the Selected Provinces

Wage Rate ($ per Month)	May 1999 (Baseline Survey)	October 1999 (First Follow-up Survey)	May 2000 (Second Follow-up Survey)
Chaco	194.9	183.8	165.5
Mendoza	200.0	192.8	168.9
Tucuman	195.9	195.4	165.5

Source: Trabajar Project Office, Ministry of Labor.

sharply in the latter part of 1999, to bring down the center's outlays.[8] As already noted, the wage rate was also cut; Table 3 gives the sample mean wages by survey round.

The aggregate cuts to the program made it less likely that past participants would find another subproject to join. A large new workfare program, the Emergency Employment program took up some of the slack in 2000. This was not in operation by the time of the second survey (first followup survey), but it was by the third survey. While our impact estimates using the first and second surveys are not likely to be affected by this new program, this is not true of the results using the third survey. We return to this point.

IV. Results

Recall that there are two matches that need to be done, one for selecting the comparison group of nonparticipants in the first survey and one for balancing covariates between subsequent leavers and stayers. The PSM methods we have used are reasonably standard, and are described in more detail in the Appendix. Table 4 gives the logit regressions for constructing the propensity scores for the two stages of matching. Recall that we have a narrower database for the first stage matching than in Jalan and Ravallion (2003). However, in modeling whether an initial participant drops out in the second round we can make use of a somewhat richer set of questions that we could add for the Trajabar sample. The extra questions on participation in neighborhood associations and indicators of whether the selection into the program was due to personal contacts (with various actors) allow us to measure the importance of social networks to program participation, as found to be important in Jalan and Ravallion (2003). Moreover, we have information on the workers' labor force histories (prior to joining Trabajar). The evaluation literature has recently emphasized changes in labor force status as an important determinant of participation in training programs.[9]

8. The rate of new projects approved was 439 per month in the period February to June 2000, as compared to 872 per month in February-November 1999 (the peak summer months of December and January have unusually low project approval rates, so we dropped them from this calculation).

9. See for example Dehejia and Wahba (1999, 2002), Heckman and Smith (1999), Heckman, Lalonde, and Smith (1999).

Table 4
Logit Regressions for Program Participation

	Chaco				Mendoza				Tucuman			
	First Matching		Second Matching		First Matching		Second Matching		First Matching		Second Matching	
	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Common variables												
Aged 25–29	0.205	0.20	–0.444	0.38	0.024	0.23	0.285	0.41	0.584	0.21**	–0.410	0.42
Aged 30–39	–0.571	0.23**	–0.277	0.44	–0.638	0.24**	–0.427	0.48	–0.080	0.24	0.000	0.48
Aged 40–49	–1.117	0.28**	0.815	0.63	–0.649	0.27**	0.904	0.53*	0.220	0.26	0.134	0.57
Aged 50–54	–1.089	0.32**	0.577	0.60	–1.165	0.29**	0.502	0.56	–0.265	0.29	1.071	0.69
Male	1.687	0.20**	–0.745	0.48	1.329	0.17**	0.832	0.50*	0.838	0.18**	0.553	0.42
Head of the household	–0.089	0.27	0.079	0.51	0.373	0.29	–0.124	0.49	–0.229	0.28	–0.463	0.54
Spouse of the head	–0.438	0.36	–1.405	0.76*	–0.170	0.36	0.807	0.99	–0.438	0.34	0.068	0.80
Married	0.352	0.22	0.290	0.38	–0.277	0.23	–0.112	0.41	–0.256	0.22	0.057	0.43
PS not completed	–0.524	0.30*	0.735	0.70	1.000	0.28**	0.263	0.77	0.024	0.26	0.393	0.62
PS completed	–0.537	0.27**	0.526	0.65	0.741	0.25**	1.005	0.71	–0.238	0.22	0.507	0.51
SS not completed	–1.158	0.27**	0.505	0.61	–0.232	0.26	0.928	0.68	–0.795	0.22**	1.012	0.48**
SS completed	–0.955	0.30**	0.590	0.69	–0.636	0.31**	1.080	0.80	–0.677	0.28**	0.720	0.62
House is a villa	–0.322	0.26	–2.310	0.81**	0.468	0.28*	–1.658	1.36	0.442	0.38	0.183	1.18
House is an apartment.	–0.765	0.44*	–0.088	0.60	–0.535	0.23**	0.288	0.46	0.420	0.21**	0.592	0.48
1 room	1.717	0.30**			2.187	0.32**	0.046	0.70	1.570	0.34**	–1.347	0.70*
2 rooms	1.141	0.25**	0.532	0.52	1.657	0.29**	–0.389	0.60	1.634	0.26**	–0.465	0.60
3 rooms	0.315	0.22	0.381	0.45	1.137	0.28**	0.029	0.57	0.932	0.24**	–0.627	0.57
4 rooms	0.270	0.23	0.578	0.48	0.714	0.29**	0.110	0.61	0.337	0.26	0.174	0.56
Bathroom in the household	–0.170	0.26	0.110	0.70	1.115	0.28**	1.408	0.61**	1.285	0.31**	–0.262	0.62

Table 4 (*continued*)

	Chaco				Mendoza				Tucuman			
	First Matching		Second Matching		First Matching		Second Matching		First Matching		Second Matching	
	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Own only land	−0.407	0.20**	−0.244	0.33	0.675	0.28**	0.077	1.17	−1.810	0.26**	0.084	0.50
Renting	−1.415	0.28**	0.351	1.11	−1.089	0.24**	−0.723	0.55			−1.001	0.61
Walls—de Mampostería	−1.183	0.32**	−2.043	1.16*	0.350	0.18*	−0.545	0.43	−1.891	0.19**	−0.563	0.44
Fraction nonmigrants	1.390	0.24**	1.289	0.55**	−0.163	0.22	−0.087	0.51	0.465	0.25*	−1.472	0.66**
Extended family	0.411	0.18**	−0.375	0.37	0.305	0.18*	−0.158	0.36	−0.060	0.19	0.082	0.44
Fraction children 6–12 attending school	−1.685	0.40**	0.143	1.02	−0.807	0.46**	0.535	0.97	0.632	0.43	−0.365	0.98
Fraction children 13–18 attending school	−0.330	0.18*	0.222	0.36	−0.227	0.19	−0.239	0.35	−0.423	0.18**	0.000	0.40
Fraction members 0–5	−5.579	1.12**	0.490	1.73	−1.687	0.87*	0.760	1.97	−6.302	0.99**	2.318	1.86
Fraction members 6–14	−3.019	1.13**	−1.849	1.85	−0.975	0.95	0.569	2.18	−6.662	0.99**	2.110	1.96

Fraction members 15–64	−3.878	1.01**	−0.748	1.47	−1.850	0.78**	0.387	1.70	−5.194	0.88**	1.420	1.61
Household size	0.166	0.04**	−0.018	0.08	0.032	0.04	−0.100	0.08	0.113	0.04**	−0.091	0.08
Extra variables for Trabajar participants												
Participated in neighborhood associations			−0.714	0.46			−0.686	0.40*			−0.245	0.43
Entered in Trabajar due to personal contacts with												
Municipality officials			−0.939	0.48**			−0.006	0.34			1.298	0.62**
Union leaders			0.059	0.47			0.692	0.39*			−0.327	0.42
Former Trabajar workers			−0.800	0.46*			−0.502	0.45			−1.055	0.51**
Dirigentes barriales/ others			0.154	0.41			0.280	0.38			−1.047	0.47**
Previously employed as												
Temporary worker			0.391	0.38			0.849	0.49*			−0.254	0.44
Permanent worker			−0.448	0.36			0.440	0.52			−0.504	0.48
Constant	2.314	1.14**	2.567	2.08	−2.750	0.91*	−3.536	2.08	3.020	0.96	1.298	1.97
Number of observations	2,023		359		2,615		352		1,827		302	
Log likelihood	−824.8		−197.0		−952.1		−197.6		−795.8		−175.4	

Table 4 (*continued*)

	Chaco				Mendoza				Tucuman			
	First Matching		Second Matching		First Matching		Second Matching		First Matching		Second Matching	
	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Pseudo R2	0.268		0.155		0.221		0.128		0.238		0.157	
F-test joint significance basic specification (*p*-value)			*0.064*				*0.284*				*0.253*	
F-test joint significance new variables (*p*-value)			*0.008*				*0.044*				*0.002*	

Note: (1) 1st stage matching of participants with nonparticipants using Trabajar and PHS samples; (2) 2nd stage matching of leavers and stayers using Trabajar sample. PS, SS stand for primary school and secondary school respectively. Reported standard errors have been corrected for heteroskedasticity using the White formula; * significant at 10 percent; ** significant at 5 percent.

We find that the additional variables for the Trabajar sample are jointly significant in explaining who drops out of the program. However, the first logit regression (used to determine the comparison group of nonparticipants from the national sample) still has far higher predictive ability, as indicated by the pseudo R^2. The lack of observable correlates of which individuals dropped out adds weight to the a priori arguments in Section III that this was not due to "pull" factors, which one would expect to be correlated with observables (such as education). The identified determinants of program participation in the first regression seem plausible.

We performed a balancing test (Hotelling *T*-squared test) for the covariates: conditional on the deciles of the predicted propensity score, we tested for difference in the means of the *X*'s used in the estimation of the propensity score (Dehejia and Wahba 1999; Smith and Todd 2001). We did not find any systematic patterns of significant differences between the *X*'s in the treatment and control samples after conditioning on the propensity score.

Of the original sample of 1,459 Trabajar participants, we restrict the sample to those workers aged 15–65. We had to drop 264 observations because satisfactory matches were not available in the PHS for both survey rounds, or key data were missing.[10] The 1,195 Trabajar participants were then matched with 1868 distinct individuals in the PHS (allowing up to five matches, and with replacement). After forming the panel across the two surveys, we ended up with a sample of 1,018 Trabajar participants who could be matched satisfactorily and followed up in the second round. Figure 1 gives the distribution of the log of the odds ratios for Trabajar samples and the PHS samples by province. The vertical lines give the regions of common support. Note that we are mainly losing observations from the PHS sample with low probabilities of participating in Trabajar.

Given that there is likely to be a discouraged worker effect (that is, some participants are normally inactive in the labor force) we choose not to confine the selection of comparison households to those who were deemed to be active in the labor force in the PHS. We found that one-third of the selected controls were inactive, though 97 percent of the Trabajar workers were matched to at least one person active in the labor force. As a check, we reestimated the propensity score model confining the sample to active workers. There were very few changes in the significance of the covariates, and the region of common support was very similar.

In the second survey, 520 of the Trabajar participants from the first round dropped out of the program. After matching on the basis of the propensity scores based on Table 4 (Column 2), we had 419 stayers matched with 400 leavers.

Table 5 gives the estimates of *DDD*. (All mathematical expectations in Section II are replaced by sample means.) Our estimate of the income gain to stayers from their participation in the program (net of the income gain attributed to past participation) is \$140 per month—about three-quarters of the mean gross wage in October 1999 (Table 3). Table 6 repeats the calculations of Table 5, but this time no matching is done in the second survey. The results are very similar, consistent with our expectation that the bulk of the people dropping out of the program were doing so involuntarily. For if withdrawal from the program had been voluntary then we would expect it to be correlated with observed correlates and (hence) that the second-stage matching would make a noticeable difference to our results.

10. Thirty-six were outside the region of common support, 137 did not satisfy the maximum absolute difference in propensity scores of 0.00001, and 64 did not have at least one match in both survey rounds.

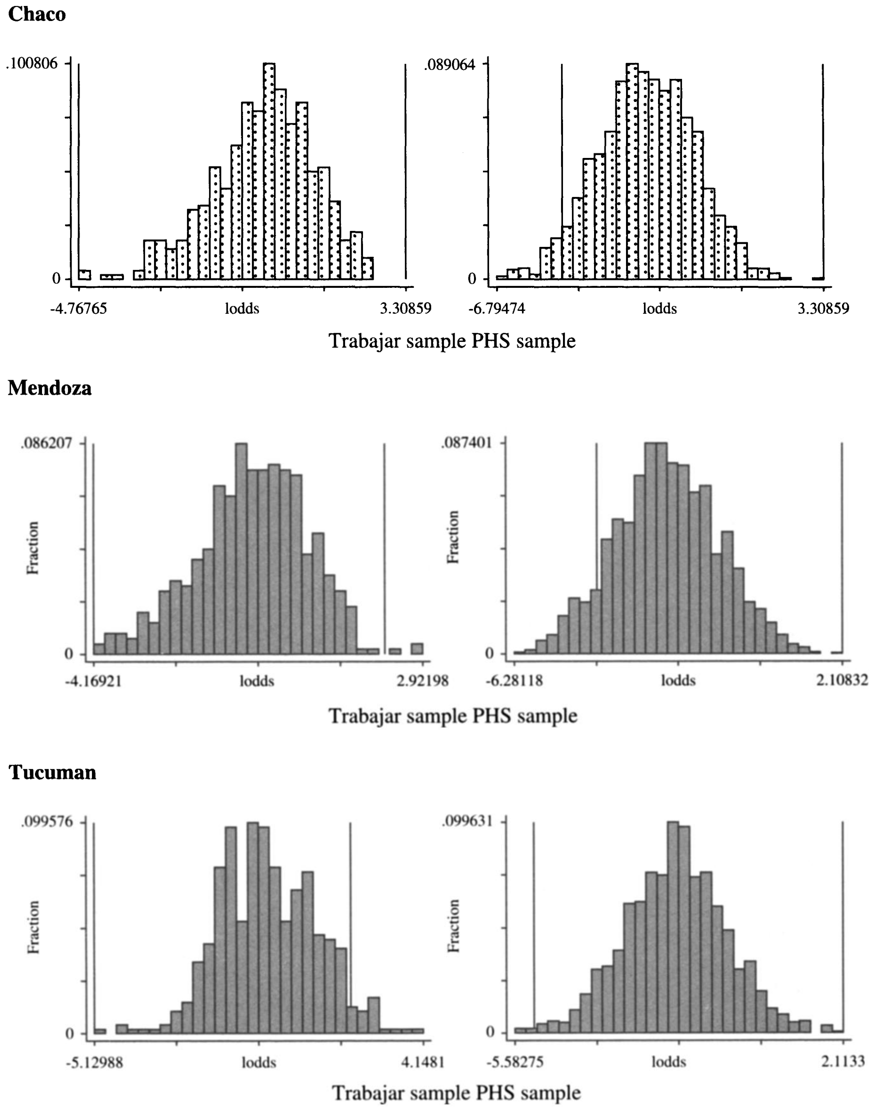

Figure 1

Log of the odds ratio for first-stage matching

Note: The vertical bars delimit the region of common support.

Twenty percent of the participants lived in households in which other members were also participating. There is a concern that these households may invalidate the assumption of no interference between units.[11] So, as a check, we repeated the analysis dropping those households. The sign and significance of variables in the propensity score model did not change in any important way. On excluding households with more than one participants the *DDD* estimate rose only slightly, to $146 (with a standard error of 13).

It is notable how poorly the single difference estimator performs in the first survey, with no significant positive impact indicated (Tables 5 and 6). There is clearly a large bias due to latent heterogeneity in the single difference estimator in our data. However, the single difference estimator comparing stayers and leavers in the second round appears to do much better; indeed, it gets closer to the *DDD* estimate than the double difference estimate for program leavers.

When we broke down the results by province, we found that the losses from retrenchment are smaller for areas with less tight local labor markets: our estimate of *DDD* is lowest for Mendoza, where the unemployment rate is also lowest (Table 2).[12]

Table 5

Triple-Difference Estimates Net of Income Gains

	Trabajar Participants in Round 2 ($D_{i2}=1$) $N=419$			Matched Trabajar Nonparticipants Round 2 ($D_{i2}=0$) $N=400$		
	Trabajar Participants in Round 1 ($D_{i1}=1$)	Matched Nonparticipants Round 1 ($D_{i1}=0$)	Single Difference	Trabajar Participants in Round 1 ($D_{i1}=1$)	Matched Nonparticipants Round 1 ($D_{i1}=0$)	Single Difference
	$\overline{Y}_t^T=$	$\overline{Y}_t^C=$		$\overline{Y}_t^T=$	$\overline{Y}_t^C=$	
$t=1$	228.9	282.7	−53.8	223.6	294.4	−70.8
	(3.8)	(13.2)	(14.0)	(2.9)	(12.6)	(12.9)
$t=2$	228.4	277.3	−48.8	83.0	288.8	−205.8
	(4.1)	(13.3)	(13.9)	(6.2)	(12.0)	(13.9)
Single difference	$\overline{\Delta Y}_t^T=$	$\overline{\Delta Y}_t^C=$		$\overline{\Delta Y}_t^T=$	$\overline{\Delta Y}_t^C=$	
	−0.5	−5.4		−140.7	−5.6	
	(4.5)	(7.4)		(6.2)	(8.1)	
Double difference	$[\Delta(\overline{Y}_2^T-\overline{Y}_2^C) \mid D_{i2}=1]=$ 4.9 (8.3)			$[\Delta(\overline{Y}_2^T-\overline{Y}_2^C) \mid D_{i2}=0]=$ −135.1 (10.6)		
Triple difference	$[\Delta(\overline{Y}_2^T-\overline{Y}_2^C) \mid D_2=1]-[\Delta(\overline{Y}_2^T-\overline{Y}_2^C) \mid D_2=0]=$ 140.0 (13.4)					

Note: Standard errors in parentheses.

11. We are grateful to a referee for pointing this out.
12. The working paper version in Ravallion et al. (2002) gives the provincial breakdown.

Table 6
Triple-Difference Estimates with Only First-Stage Matching

	Participants in Round 2 ($D_{i2} = 1$) $N = 498$			Nonparticipants Round 2 ($D_{i2} = 0$) $N = 520$		
	Participants in Round 1 ($D_{i1} = 1$)	Matched Nonparticipants Round 1 ($D_{i1} = 0$)	Single Difference	Participants in Round 1 ($D_{i1} = 1$)	Matched Nonparticipants Round 1 ($D_{i1} = 0$)	Single Difference
	$\overline{Y}_t^T =$	$\overline{Y}_t^C =$		$\overline{Y}_t^T =$	$\overline{Y}_t^C =$	
$t = 1$	228.1	286.7	−58.4	225.4	286.1	−60.6
	(3.4)	(12.4)	(13.0)	(2.9)	(10.6)	(10.9)
$t = 2$	228.2	278.4	−50.2	83.4	281.6	−198.2
	(3.7)	(12.0)	(12.5)	(5.6)	(9.9)	(11.8)
Single difference	$\overline{\Delta Y}_t^T =$ 0.10 (4.7)	$\overline{\Delta Y}_t^C =$ −8.2 (7.2)		$\overline{\Delta Y}_t^T =$ −142.0 (5.5)	$\overline{\Delta Y}_t^C =$ −4.4 (6.8)	
Double difference	$[\Delta(\overline{Y}_2^T - \overline{Y}_2^C) \mid D_{i2} = 1] =$ 8.3 (7.9)			$[\Delta(\overline{Y}_2^T - \overline{Y}_2^C) \mid D_{i2} = 0] =$ −137.6 (9.1)		
Triple difference	$[\Delta(\overline{Y}_2^T - \overline{Y}_2^C) \mid D_2 = 1] - [\Delta(\overline{Y}_2^T - \overline{Y}_2^C) \mid D_2 = 0] =$ 145.9 (12.1)					

Note: Standard errors in parentheses.

To check for household behavioral responses, we recalculated the triple differences at the household level, decomposing the overall income change into changes in Trabajar income, income of other household members, and other nonwage income. We found no sign of any spillover effect on the earnings of other household members (Ravallion, et al. 2002).

Table 7 gives the results when we look at income replacement over 12 months. Naturally we cannot do this over the same samples as before, and (in particular) the sample of continuous stayers is greatly reduced. The third survey round does indicate a sizable recovery of income for leavers from the second round; the treatment group's mean income dips from \$228 per month in the first survey to \$86 in the second, but rebounds to \$138 in the third (Table 7). This could reflect in part the income gains from the new Emergency Employment Program introduced in 2000 (Section II). However, the magnitudes involved suggest that other income sources are very likely to have been involved. The special Trabajar module has a question on whether individuals participated in another temporary program (other than Trabajar); 5 percent of the Trabajar sample report doing so in the third survey, as opposed to 2 percent in the second survey. However, it is readily verified that for the new program to account fully for the recovery it would need to have employed 70 percent of ex-Trabajar participants.

Table 7
Triple-Difference Estimates over 12 Months (Two Followup Rounds)

	Stayers in Rounds 2 and 3 ($D_{i2}=1$; $D_{i3}=1$) $N=118$			Leavers in Rounds 2 and 3 ($D_{i2}=0$; $D_{i3}=0$) $N=424$		
	Treatment	Comparison	Single Difference	Treatment	Comparison	Single Difference
	$\overline{Y}_t^T=$	$\overline{Y}_t^C=$		$\overline{Y}_t^T=$	$\overline{Y}_t^C=$	
$t=1$	228.0	285.7	−57.7	227.5	272.2	−44.6
	(8.1)	(27.9)	(29.6)	(3.4)	(13.7)	(13.8)
$t=2$	241.3	314.9	−73.7	85.5	276.4	−190.9
	(8.4)	(32.9)	(33.0)	(6.2)	(12.3)	(14.1)
$t=3$	219.5	289.0	−69.5	138.4	267.3	−128.9
	(6.8)	(25.3)	(25.8)	(7.0)	(18.6)	(14.5)
Single difference	$\overline{\Delta Y}_2^T=$ 13.3	$\overline{\Delta Y}_2^C=$ 29.3		$\overline{\Delta Y}_2^T=$ −142.0	$\overline{\Delta Y}_2^C=$ 4.2	
	(9.2)	(20.8)		(6.2)	(9.3)	
	$\overline{\Delta Y}_3^T=$ −21.7	$\overline{\Delta Y}_3^C=$ −25.9		$\overline{\Delta Y}_3^T=$ 52.9	$\overline{\Delta Y}_3^C=$ −9.0	
	(9.1)	(18.0)		(7.4)	(16.5)	
Double difference	$[\Delta(\overline{Y}_2^T-\overline{Y}_2^C) \mid D_{i2}=1, D_{i3}=1]=$ −16.0			$[\Delta(\overline{Y}_2^T-\overline{Y}_2^C) \mid D_{i2}=0, D_{i3}=0]=$ −146.3		
	(22.0)			(11.9)		
	$[\Delta(\overline{Y}_3^T-\overline{Y}_3^C) \mid D_{i2}=1, D_{i3}=1]=$ 4.2			$[\Delta(\overline{Y}_3^T-\overline{Y}_3^C) \mid D_{i2}=0, D_{i3}=0]=$ 62.0		
	(20.0)			(16.5)		
Cumulative gain	−11.8			−84.3		
	(22.4)			(15.7)		
Triple difference	$[\Delta(\overline{Y}_2^T-\overline{Y}_2^C) \mid D_2=1, D_3=1]-[\Delta(\overline{Y}_2^T-\overline{Y}_2^C) \mid D_2=0, D_3=0]=$ 130.3					
	(25.3)					
	$[\Delta(\overline{Y}_3^T-\overline{Y}_3^C) \mid D_2=1, D_3=1]-[\Delta(\overline{Y}_3^T-\overline{Y}_3^C) \mid D_2=0, D_3=0]=$ −57.8					
	(33.1)					
Cumulative gain				72.5		
				(32.0)		

Note: Standard errors in parentheses; both matched comparison groups and matched leavers (as in Table 3).

The *DDD* estimate for round three indicates that about half of the second round loss is recovered. On aggregating over the two rounds, the net income gain to stayers is \$73 per month (= 130-58), representing slightly less than one half of the mean gross wage in round 3 (Table 2a).

What can we conclude about net income gains from the program? Recall that our *DDD* estimator also gives the net gain to current participation if there is no selection bias and no current gains to nonparticipants who had previously participated (Section III). Before we present a joint test of these conditions, it is of interest to reflect on some a priori considerations. Given that a large contraction was imposed on the program, selection bias might not be an important concern (Section III). What about lagged effects? One source of evidence can be found in the qualitative questions we added to the third round of the survey, on whether current or past Trabajar participants felt that the program had improved their earning opportunities outside the program. Table 8 summarizes the results. Among continuing participants, about half felt that the program improved their chances of getting a job; two-thirds felt that it gave them a marketable skill; about one-third felt that it expanded their contacts. These results are suggestive of lagged income gains to ex-participants from the program. However, there are two caveats. The expected gains may take some time to materialize, depending on the aggregate labor market conditions. Secondly, there may well be psychological biases in answering qualitative questions of this sort, such that participants overestimate the future gains beyond their current participation in the program.

Table 9 gives our results in testing the joint hypothesis of no current gains from past participation and no selection bias. Under the null, the gains from participating in the program in Period 2 should not depend on participation in Period 3. We are not able to reject the null hypothesis (p-value of 0.17).

Table 8
Perceived Gains From Past Participation from the Program

Length of exposure to the program N = 962	One Round $D_{i2} = 1; D_{i2} = 0; D_{i3} = 0$ $N = 464$ (48.2 percent)	Two Rounds $D_{i2} = 1; D_{i2} = 1; D_{i3} = 0$ $N = 339$ (35.2 percent)	Three Rounds $D_{i2} = 1; D_{i2} = 1; D_{i3} = 1$ $N = 136$ (14.1 percent)
	percent of respondents replying "yes"		
Expanded job opportunities			
Expected gains in $t = 2$	35.7	52.6	51.2
Expected gains in $t = 3$ all	33.7	37.9	48.7
Employed in $t = 3$	35.4	42.7	
Unemployed in $t = 3$	31.1	31.1	
Learned skills for other jobs			
Expected gains in $t = 2$	51.8	66.1	64.7
Expected gains in $t = 3$ all	45.0	53.5	61.5
Employed in $t = 3$	38.6	51.0	70.0
Unemployed in $t = 3$	55.0	57.1	
Expanded contacts for future			
Expected gains in $t = 2$	25.7	35.9	40.0
Expected gains in $t = 3$ all	21.9	22.9	32.6
Employed in $t = 3$	24.5	22.6	30.0
Unemployed in $t = 3$	17.8	23.3	

Note: Sample of Trabajar workers—special module in Period 3

Table 9

Joint Test of no Lagged Gains from Past Participation and no Selection Bias: Triple-Difference Estimates over Six Months

	Participants Round 2 (D_{i2} = 1)	Nonparticipants Round 2 (D_{i2} = 0)
Stayers Round 3 (D_{i3} = 1)	$N = 106$	$N = 20$
Single difference	$\overline{\Delta Y}_t^T = 14.2$ (10.2) $\overline{\Delta Y}_t^C = -7.5$ (19.5)	$\overline{\Delta Y}_t^T = -154.6$ (22.0) $\overline{\Delta Y}_t^C = 40.25$ (22.7)
Double difference	$[\Delta(\bar{Y}_2^T - \bar{Y}_2^C) \mid D_{i2} = 1] = 21.6$ (20.9)	$[\Delta(\bar{Y}_2^T - \bar{Y}_2^C) \mid D_{i2} = 0] = -194.8$ (33.7)
(A) Triple difference	$[\Delta(\bar{Y}_2^T - \bar{Y}_2^C) \mid D_2 = 1] - [\Delta(\bar{Y}_2^T - \bar{Y}_2^C) \mid D_2 = 0] = 216.5$ (50.3)	
Leavers Round 3 (D_{i3} = 0)	$N = 367$	$N = 531$
Single difference	$\overline{\Delta Y}_t^T = -2.6$ (4.5) $\overline{\Delta Y}_t^C = -7.9$ (7.8)	$\overline{\Delta Y}_t^T = -142.8$ (5.4) $\overline{\Delta Y}_t^C = -1.6$ (6.8)
Double difference	$[\Delta(\bar{Y}_2^T - \bar{Y}_2^C) \mid D_{i2} = 1] = 5.3$ (8.7)	$[\Delta(\bar{Y}_2^T - \bar{Y}_2^C) \mid D_{i2} = 0] = -141.2$ (9.0)
(B) Triple difference	$[\Delta(\bar{Y}_2^T - \bar{Y}_2^C) \mid D_2 = 1] - [\Delta(\bar{Y}_2^T - \bar{Y}_2^C) \mid D_2 = 0] = 146.5$ (13.0)	

t-test of equality H_0: $(A) = (B)$

Difference: 69.97
(52.0)
t-statistic: 1.34
p-value: 0.17

Note: Standard errors in parentheses. Under the joint null of "no lagged gains from past participation" and "no selection bias," the triple differences should be the same (see Equation 7).

However, the power of the test is limited by the small sample size for stayers in Period 3. Andrews (1989) provides approximation formulae for the (inverse) power function of the Wald test. For the two-sided test, the Andrews approximations at a probability of 0.5 is 102; our test statistic of 70 thus indicates low power with this sample size. So while we cannot reject the null of no bias, there is a high probability of failing to do so when in fact it is false.

It is of interest to compare our results with those of Jalan and Ravallion (2003). The latter paper used single-difference matching on a richer data set. Their estimated mean net gain to Trabajar participation was \$103 per month, rising to \$157 using nearest-neighbor matching. Our single-difference estimate gives an implausible result that deviates greatly from Jalan and Ravallion. This is not true of our *DDD* estimate in Table 5 of \$140. Aggregating over the three rounds, our estimated income gain to participants of \$73 is less than Jalan and Ravallion obtained (though it is roughly the same share of the Trabajar wage rate).[13] But this is what we would expect as long as

13. The estimated forgone income of participants as a proportion of the gross wage is higher than Datt and Ravallion's (1994) estimate for a workfare program in India, though the latter setting was arguably one in which unemployment (or at least underemployment) was higher.

participants are able in time to recover a greater amount of the income lost from retrenchment.

V. Conclusions

Social programs introduced during a crisis cannot reasonably be delayed to allow a baseline survey to be done. We have tried to see what can be learnt about impact by instead following up people who subsequently leave the program. In other words, our observations of participants' outcomes when not participating are *after* the program rather than before it.

We have proposed a triple-difference estimator of the income gains to participants in Argentina's Trabajar program. A random sample of participants was matched with a group of nonparticipants drawn from a comparable national survey. We then followed both samples over time, during a period of aggregate program contraction on top of designed rotation of program beneficiaries. Propensity score matching methods were used to balance observed covariates at two stages: between initial participants and the comparison group, and between those who left the program and those who stayed. Since we track outcomes over time, we can eliminate any separable time-invariant selection bias in the first matching. Such selection bias remains in the second stage matching, though we can sign the bias under plausible assumptions. The fact that there was a large centrally imposed contraction in program outlays as well as designed rotation helps reduce concerns about selection bias at the second-stage matching.[14]

We find that the estimated income losses to those who left the program were sizable, representing about three-quarters of the gross wage on the program within the first six months, though falling to slightly less than one-half over 12 months, indicating existence of a post-program version of "Ashenfelter's dip." Fully removing selection bias would probably yield even lower estimates of income replacement.

Interpreting our triple-difference estimate as a measure of the gains from the program requires two conditions: that there is no selection bias in who leaves the program, and there are no lagged income effects from past participation. On a priori grounds we find the selection bias argument implausible in our particular empirical setting, though it may be a more serious concern in other applications. On the other hand, the existence of lagged income effects is supported by qualitative questions in the survey we have used.

We have proposed a joint test of these two conditions, based on comparing the triple-difference estimate for those who left versus stayed in a third round of the survey. We cannot reject statistically the conditions required for using our triple-difference measure as an estimate of the gains to current participants, though

14. As a robustness check, we estimated the probability of dropping out between Waves 2 and 3 among participants (as in the logit regressions for program participation used to estimate the propensity score matching). There is no suggestion that the decision to continue participation in the program is correlated with observable characteristics such as age, gender of education.

in our particular application, the test has low power given how few participants remain in the third round.

While our results point to large losses from retrenchment and sizeable income gains to participants in the workfare program we have studied, one should be cautious in drawing conclusions for other settings. A key factor is likely to be the level of unemployment (notably among the poor) at the time the program is cut. If one cuts disbursements at a time of sufficiently rapid economic recovery, or in regions where recovery is under way, then the loss to workers is likely to be smaller than we have found.

From the point of view of evaluation design, our results suggest a tradeoff between the resources devoted to cross-sectional data collection for the purpose of single-difference matching, versus collecting longitudinal data with a lighter survey instrument. The light instrument we have used here was not able to deliver plausible single-difference estimates using propensity-score matching, when compared to prior estimates using richer data for the same program. However, it would appear that we have been able to satisfactorily address this problem by tracking households over time, even using the lighter instrument.

Appendix

Propensity Score Matching

In propensity score matching, an estimate of the counter-factual outcome for a given participants is obtained as a weighted mean of the measured outcomes for nonparticipants whose predicted probability of participation ("propensity score") is within some interval of the corresponding probability for that participant, given a vector of control variables, X_t. An important result due to Rosenbaum and Rubin (1983) essentially establishes that if there is no selection bias when matching according to X_t then there is no bias when matching exactly according to the propensity score of X_t.

More precisely, participants are matched to individuals who did not participate (or drop out) on the basis of the propensity score, $\Pr(D_i = 1|X_i)$. To estimate this we follow the standard practice in PSM applications of using the predicted values from a logit model to estimate the propensity score for each observation in the participant and the comparison-group samples. Matched pairs are constructed on the basis of how close the scores are across the two samples. The nearest neighbor to the ith participant is defined as the nonparticipant that minimizes $[\hat{\Pr}(D_i = 1|X_i) - \hat{\Pr}(D_j = 1|X_j)]^2$ over all j in the set of nonparticipants, where $\hat{\Pr}(D_k = 1|X_k)$ is the predicted propensity score for observation k. Matches are only accepted if $[\hat{\Pr}(D_i = 1|X_i) - \hat{\Pr}(D_j = 1|X_j)]^2$ is less than 0.00001 (an absolute difference less than 0.0032.) Notice that we can only include those observations on nonparticipants that share a common range of values of the propensity scores calculated for the participants (that is, the two groups share common support in the predicted propensity scores). In our empirical analysis, the comparison group for each participant was defined as the set of five nearest neighbors among nonparticipants.

References

Andrews, Donald W. K. 1989. "Power in Econometric Applications." *Econometrica* 57(5):1059–1090.
Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economic Studies* 60:47–57.
Datt, Gaurav, and Martin Ravallion. 1994. "Transfer Benefits from Public Works Employment." *The Economic Journal* 104:1346–69.
Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94:1053–62.
Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Review of Economics and Statistics* 84:151–61.
Gruber, Jonathan. 1994. "The Incidence of Mandated Maternity Benefits." *American Economic Review* 84(3):622–41.
Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64(4):605–54.
Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra E. Todd. 1998. "Characterizing Selection Bias using Experimental Data." *Econometrica* 66:1017–99.
Heckman, James J., Robert J. Lalonde, and Jeffrey A. Smith. 1999. "The Economics and Econometrics of Training Programs." In *Handbook of Labor Economics*, volume 3A, ed. Orley Ashenfelter and David Card. Amsterdam: North-Holland.
Heckman, James J., and Jeffrey A. Smith. 1999. "The Pre-Programme Earning Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies." *The Economic Journal* 109:313–48.
Jalan, Jyotsna, and Martin Ravallion. 1998. "Are There Dynamic Gains from a Poor-Area Development Program?" *Journal of Public Economics*. 67(1):65–86.
———. 2003. "Estimating the Benefit Incidence of an Anti-Poverty Program." *Journal of Business and Economic Statistics* 21(1):19–30.
Ravallion, Martin. 2000. "Monitoring Targeting Performance when Decentralized Allocations to the Poor are Unobserved." *World Bank Economic Review* 14(2):331–46.
Ravallion, Martin, Emanuela Galasso, Teodoro Lazo, and Ernesto Philipp. 2002. "Do Workfare Participants Recover Quickly from Retrenchment?" Policy Research Working Paper. World Bank. Washington D.C.
Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.
Rubin, Donald B., and Neal Thomas. 2000. "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates." *Journal of the American Statistical Association* 95:573–85.
Smith, Jeffrey A., and Petra E. Todd. 2003. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*. Forthcoming.
———. 2001. "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods." *American Economic Review* 91(2):112–18.