

Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching

Jyotsna JALAN

Indian Statistical Institute, New Delhi 110016, India (jjalan@isid.ac.in)

Martin RAVALLION

The World Bank, Washington, DC 20433 (mravallion@worldbank.org)

We apply recent advances in propensity-score matching (PSM) to the problem of estimating the distribution of net income gains from an Argentinean workfare program. PSM has a number of attractive features in this context, including the need to allow for heterogeneous impacts, while optimally weighting observed characteristics when forming a comparison group. The average direct gain to the participant is found to be about half the gross wage. Over half of the beneficiaries are in the poorest decile nationally, and 80% are in the poorest quintile. Our PSM estimator is reasonably robust to a number of changes in methodology.

KEY WORDS: Argentina; Impact evaluation; Poverty alleviation; Workfare.

Antipoverty programs often require that participants must work to obtain benefits. Such “workfare” programs have been turned to in crises, such as due to macroeconomic or agroeconomic shocks, in which a large number of poor able-bodied people have become unemployed. Typically, the main aim is to raise the current incomes of poor families hurt by the crisis. (On the arguments and evidence on this class of interventions, see Ravallion 1991, 1999; Besley and Coate 1992; and Lipton and Ravallion 1995.)

To assess the distributional impact of such a program, we need to measure the income gains to participants conditional on preintervention income, where the income gain is the difference between household income with the program and that without it. This conditional impact estimate is commonly referred to as a program’s “benefit incidence.” The “with” data can be collected without much difficulty. But the “without” data are fundamentally unobserved since an individual cannot be both a participant and a nonparticipant of the same program. This is a well-known and fundamental problem in all causal inferences (Holland 1986).

An assumption that is sometimes made in benefit incidence analysis is that the gross wages paid are an adequate measure of the income gains to participants. (See, for example, the various assessments of the cost effectiveness of workfare programs reviewed in Subbarao et al. 1997.) This assumption would be reasonable if labor supply to a workfare program came only from the unemployed. But that is generally not the case in practice. Moreover, even if a participating worker were unemployed at the time she joined the program, there is an opportunity cost of participation. Joining the program will leave less time for job search. There are likely to be effects on time allocation within the household. For example, Datt and Ravallion (1994) find that other family members took up the displaced productive activities when someone joined a workfare program in rural India. Such behavioral responses will reduce foregone income, although we can still expect it to be positive. Without taking proper account of foregone incomes, we cannot know the true incidence of program benefits.

This article estimates the income gains from a workfare program, and how those gains vary with preintervention incomes. To draw a statistical comparison group to workfare participants from a larger contemporaneous and comparable survey of nonparticipants, we apply recent advances in propensity-score matching (PSM) methods, following Rosenbaum and Rubin (1983). Matching methods have been quite widely used in evaluations, but there have as yet been few economic applications of matching based on the propensity score. Some exceptions are Heckman, Ichimura, and Todd (1997), Dehejia and Wahba (1998, 1999), Hotz, Imbens, and Mortimer (1999), Lechner (1999), and Lopez (1999).

We study the Trabajar Program, an antipoverty program of the Government of Argentina, supported by a World Bank loan and technical assistance. A number of features of this setting lend themselves to PSM methods. As is common in a crisis, other evaluation methods requiring randomization or a baseline (preintervention) survey were not feasible. However, it was possible to do a postintervention survey in which the same questionnaire was administered to both the participants and the nonparticipants, and in a setting in which it was plausible that both groups came from the same economic environment. The Trabajar participant could be identified in the larger survey.

Furthermore, using kernel density estimation techniques, we are able to ensure that participants are matched with the nonparticipants over a common region of the matching variables. Any remaining bias in the matching estimator can thus be attributed to unobserved characteristics. The design of the program can be expected to entail considerable rationing of participation according to observables. The sample of nonparticipants is very likely to include people who wanted to

participate, but were unable to do so due to, say, nonavailability of the program. While our application is well suited to above cases, bias due to unobservables cannot be ruled out.

A further advantage of PSM methods for this problem is that they lend themselves naturally to studying the heterogeneity of program impact. This is of obvious interest for an antipoverty program, in which knowledge of the distribution of impacts conditional on preintervention incomes is crucial to judging the program's success.

The following section discusses the evaluation problem and our methods. Section 2 describes the Trabajar program and our data. Section 3 presents the results, and offers an economic interpretation. Section 4 concludes.

1. ESTIMATING THE BENEFIT INCIDENCE OF A WORKFARE PROGRAM

In assessing the gains from a workfare program, the workers' earnings are naturally the main focus, and that will be the case here. However, it should be noted that earnings net of foregone income are only one of the potential benefits. There could also be risk benefits from knowing that the program exists. There may well also be benefits from the outputs, depending on (among other things) how well targeted the workfare projects are to poor areas.

We first outline a simple model of self-targeting which provides arguments for workfare, pointing to the importance of foregone incomes. We then describe the matching method we use to estimate foregone incomes.

1.1 The Problem

The following rudimentary model has the essential features necessary to characterize the "self-targeting" argument often made in favor of workfare (Ravallion 1991). The model assumes that foregone income from accepting a workfare job is $F(Y)$, a smoothly increasing function of preintervention income Y , scaled to lie between 0 and 1. Foregone income increases with preintervention income due to differences in education, experience, and so on that are naturally correlated with both earnings and family income. The workfare program offers a wage W , with $F(0) < W < F(1)$. Workers are only concerned about the net wage gain, that is, the work alternatives are judged to be the same in other respects.

It is evident that, under these assumptions, only those workers with preintervention income less than $F^{-1}(W)$ will participate; the program will perfectly screen "poor" ($Y < F^{-1}(W)$) from "nonpoor" ($Y > F^{-1}(W)$). The schedule of gains is $G = W - F(Y)$ for $Y < F^{-1}(W)$ and $G = 0$ for $Y > F^{-1}(W)$, yielding postintervention incomes $Y + G$.

In this simple model, underestimating the foregone income will lead the evaluator to overestimate the impact on poverty. To see why, suppose that, in assessing the gains from the program, we use a biased estimate of foregone income, namely, $M(Y) < F(Y)$ for all Y . Then we will overestimate the gains for all Y up to $M^{-1}(W)$. The distribution of incomes under the biased estimate of foregone incomes must first-order dominate the actual distribution. So the error in assessing foregone

incomes will overestimate the impact on income poverty. This holds for a broad class of poverty measures (Atkinson 1987).

How can one estimate the foregone income? This is a counterfactual concept in that participants' incomes in the absence of the program are missing data. There are several methods one might adopt to assess the counterfactual, drawing on the literature on impact evaluation. One can do reflexive comparisons by collecting baseline data on probable (eligible) participants before the program was instituted. These data are then compared with data on the same individuals once they have actually participated in the program. Alternatively, potential participants are identified, and data are collected from them. However, only a random subsample of these individuals is actually allowed to participate in the program. Another possible approach is to use propensity-score matching methods, following Rosenbaum and Rubin (1983, 1985), Dehejia and Wahba (1998, 1999), Heckman et al. (1997) and Heckman, Ichimura, Smith, and Todd (1998). Here, the counterfactual group is constructed by matching program participants to nonparticipants from a larger survey such as the population census or an annual national budget survey. The matches are chosen on the basis of similarities in observed characteristics. We use matching methods on nonexperimental data to evaluate the impact of the program.

Since most countries now have a nationally representative socioeconomic survey instrument, the marginal cost of using PSM only includes the survey of program participants. The same survey instrument can then be taken to a sample of participants after the program has started, possibly with an extra module to cover specific questions related to the program. PSM estimates will be reliable, provided that participants and controls have the same distributions of unobserved characteristics. Failure of this condition to hold is often referred to as the problem of "selection bias" in econometrics or "selection on unobservables" (Heckman and Robb 1985). Secondly, the support for the comparison and the program participants should be equal. Finally, it is desirable that the same questionnaire be administered to both groups, and that participants and controls are from the same economic environment.

1.2 A Feasible Method of Estimating Benefit Incidence

Suppose that we have data on N participants in a workfare program, and another random sample of size rN ($r > 1$) from the population. The second set of data could be the population census or the national household survey that has information relevant to the participation decisions of the individuals. Using the two sets of data, we try to match the N program participants with a comparison group of nonparticipants from the population.

The two surveys must include information that helps predict participation in the program. Let X be the vector of such variables. Ideally, one would match a participant with a nonparticipant using the entire dimension of X , that is, a match is only declared if there are two individuals, one in each of the two samples, for whom the value of X is identical. This is impractical, however, because the dimension of X could be very high. Rosenbaum and Rubin (1983) show that matching can be performed conditioning on $P(X)$ alone rather than on

X , where $P(X) = \text{Prob}(D = 1|X)$ is the probability of participating conditional on X , the “propensity score” of X . If outcomes without the intervention are independent of participation given X , then they are also independent of participation given $P(X)$. This is a powerful result since it reduces a potentially high-dimensional matching problem to a single-dimensional problem.

The propensity score is calculated for each observation in the participant and the comparison-group samples using standard discrete choice parametric or semiparametric models. Since some studies show that the impact estimator is robust to the choice of the discrete choice model, we use standard parametric likelihood methods to compute the propensity scores (Todd 1995). Propensity-score matching (PSM) then uses the estimated $P(X)$'s or a monotone function of it to select comparison subjects.

We used the odds ratio $p_i = P_i/(1 - P_i)$, where P_i is the estimated probability of participation for individual i to construct matched pairs on the basis of how close the scores are across the two samples. The nearest neighbor to the i th participant is thus defined as the nonparticipant who minimizes $[p(X_i) - p(X_j)]^2$ over all j in the set of nonparticipants, where $p(X_k)$ is the predicted odds ratio for observation k . Sometimes, nearest neighbors may be far apart in terms of the distance metric between the propensity scores of the treated and comparison subjects. So we match using the “propensity-score caliper,” defined by

$$C[P(X_i)] = \{P(X_j) \mid \|P(X_i) - P(X_j)\| < \varepsilon\} \quad (1)$$

for ε arbitrarily small.

In their comparisons of nonexperimental methods of evaluating a training program with a benchmark experimental design, Heckman et al. (1997) find that failure to compare participants and controls at common values of matching variables is the single most important source of bias—considerably more important than the classic econometric problem of selection bias due to differences in unobservables. To ensure that we are matching only over common values of the propensity scores, we estimated the density of the scores for the nonparticipants at 100 points over the range of scores. We use a biweight kernel density estimator, and the optimal bandwidth value suggested by Silverman (1986). Once we estimate the density for the nonparticipants, we exclude those nonparticipants for whom the estimated density is equal to zero. We also exclude 2% of the sample from the top and bottom of the nonparticipant distribution.

The mean impact estimator of the program is given by

$$\bar{G} = \frac{\sum_{j=1}^P \left(Y_{j1} - \sum_{i=1}^{NP} W_{ij} Y_{ij0} \right)}{P} \quad (2)$$

where Y_{j1} is the postintervention household income of participant j , Y_{ij0} is the household income of the i th nonparticipant matched to the j th participant, P is the total number of participants, NP the total number of nonparticipants, and the W_{ij} 's are the weights applied in calculating the average income of the matched nonparticipants. There are several different types of parametric and nonparametric weights that one can use. In

this article, we use three different weights, and thereby report three different matching estimators. Our first matching estimator is the “nearest neighbor” estimator, where we find the closest nonparticipant match for each participant, and the impact estimator is a simple mean over the income difference between the participant and its matched nonparticipant. (In calculating our mean impact, if the income of the participant is less than the income of the matched nonparticipant, we treat the impact to be zero rather than the observed negative number.)

If the comparison sample is large enough, then “ m -to-1” matching with $m \geq 1$ can be used to reduce the standard errors of comparison. However, the gain in precision achieved by increasing the matched comparison sample size is typically modest (Rosenbaum and Rubin 1985; Rubin and Thomas 1996). So we construct a second estimator which takes the average income of the closest five matched nonparticipants, and compares this to the participant's income. Following Heckman et al. (1998), we also report a kernel-weighted estimator where the weights are given by

$$W_{ij} = K_{ij} / \sum_{j=1}^P K_{ij} \quad (3)$$

where

$$K_{ij} = \frac{K[(P(X_i) - P(X_j))/a_{N0}]}{\sum_{j=1}^P K[(P(X_i) - P(X_j))/a_{N0}]} \quad (4)$$

and where a_{N0} is the bandwidth parameter, and $K(\cdot)$ is the kernel as a function of the difference in the propensity scores of the participants and the nonparticipants. In our analysis, we have used Silverman's (1986) optimal bandwidth parameter and a biweight kernel function. (The results were very similar using either a rectangular or Parzen kernel function.)

Lastly, in each of these cases, the associated standard errors of the mean impact estimator are also calculated. We calculated both the parametric and bootstrapped standard errors for the impact estimators. The two were virtually identical. We report the parametric standard errors in the article. (The bootstrapped standard errors are available from the authors on request.)

2. THE PROGRAM AND DATA

Argentina experienced a sharp increase in unemployment in the mid-1990s, reaching 18% in 1996/1997. This was clearly hurting the poor; for example, the unemployment rate (on a comparable basis) was 39% among the poorest decile in terms of household income per capita in Greater Buenos Aires (Permanent Household Survey (EPH) for Greater Buenos Aires in May 1996). Unemployment rates fell steadily as the income per person increased.

2.1 The Trabajar Program

In response to this macroeconomic crisis, and with financial and technical support from the World Bank, the Government of Argentina introduced the Trabajar II Program in May 1997. This was a greatly expanded and reformed version of a previous program, Trabajar I. The program aimed to help

in two ways. Firstly, by providing short-term work at relatively low wages, the program aimed to self-select unemployed workers from poor families. Secondly, the scheme tried to locate socially useful projects in poor areas to help repair and develop the local infrastructure. This article only assesses progress against the first objective (on the second, see Ravallion 2000).

The subprojects are proposed by local governmental and nongovernmental organizations that must cover the nonwage costs. The proposals have to be viable with respect to criteria, and are given priority according to ex ante assessments of how well targeted they are to poor areas, what benefits they are likely to bring to the local community, and how much the area has already received from the program. Workers cannot join the program unless they are recruited to an accepted proposal. The process of proposing suitable subprojects is thus key to worker participation in the program. There are other factors. The workers cannot be receiving unemployment benefits or be participating in any other employment or training program. So our eligible nonparticipant pool excluded all those individuals currently receiving some form of unemployment insurance payments. It is unlikely that a temporary employment program such as this would affect residential location, although workers can commute.

The wage rate is set at a maximum of \$200 per month. This was chosen to be low enough to assure good targeting performance, and to help assure that workers would take up regular work when it became available. To help locate the Trabajar wage in the overall distribution of wages, we examined earnings of the poorest 10% of households (ranked by total income per person) in Greater Buenos Aires (GBA) in the May 1996 Permanent Household Survey. For this group, the average monthly earnings for the principal job (when this entails at least 35 hours of work per week) in May 1996 was \$263. This includes domestic servants. This is an unusual labor-market group, given that they often have extra income-in-kind. If one excludes them, the figure is \$334. (As expected, the poorest decile also received the lowest average wage, and average wages rose monotonically with household income per person.) So the Trabajar wage is clearly at the low end of the earnings distribution.

There are other questions that the evaluation can answer. Trabajar I had been targeted to middle-aged heads of households (typically male). However, under the modified program, it was decided not to impose this restriction since there was a risk of increasing the foregone income of participants by constraining their ability to adjust work allocation within the household in response to the program. In practice, however, the restrictions on participation may still have been imposed at the local level. If that were the case on average in all localities, then one might expect to find that there are unexploited income gains by increasing participation by the young and by women. We will test this.

2.2 Data

Two household surveys are used. One is of program participants, and the other is a national sample survey, used to obtain the comparison group. Both surveys were done by

the government's statistics office, the Instituto Nacional de Estadística y Censos (INDEC), using the same questionnaire, the same interviewing teams, and at approximately the same time.

The national survey is the Encuesta de Desarrollo Social (EDS), a large socioeconomic survey done in mid-1997. The EDS sample covers the population residing in localities with 5,000 or more residents. The comparison group is constructed from the EDS. According to the 1991 census, such localities totaled to 420 in Argentina, and represented 96% of the urban population and 84% of the total population. 114 localities were sampled.

The second dataset is a special purpose sample of Trabajar participants done for the purpose of this evaluation. The sample design involved a number of steps. First, among all of the projects approved between April and June 1997, 300 projects in localities which were in the EDS sample frame were randomly selected, with an additional 50 projects chosen for replacement purposes. The administrative records on project participants did not include addresses, so the Ministry of Labor (MOL) had to obtain these by field work. From these 350 projects, the Labor Ministry could find the addresses of nearly 4,500 participants. However, for various reasons, about 1,000 of these were not interviewed. The reasons given by INDEC were that the addresses were found to be outside the sample frame, or they were incomplete, or even nonexistent, or that all household members were absent when the interviewer went to interview the household, or that they did not want to respond. In all, 3,500 participant households were surveyed. (The number of Trabajar participants during May 1997–January 1998 was 65,321.)

We restrict the analysis to households with complete income information, and those who completed all of the questions asked of them. Also, we only consider participants who were actually working in a Trabajar project at the time they were surveyed. Since the EDS questionnaire does not ask income-related questions to those below 15 or above 64 years of age, we also had to restrict our attention to the age group 15–64 years for our analysis. We focus on current Trabajar participants in the reference week, fixed at the first week of September 1997, who received wages from the Trabajar Program during August 1997. 80% of the Trabajar sample had current participants by this definition. The remaining 20% of participants are assumed to be beneficiaries who had left work by August 1, 1997 (i.e., at the start of the survey) or who had not yet started the Trabajar job. With these restrictions, the total number of active participants that we have used is 2,802.

3. RESULTS AND INTERPRETATION

3.1 Descriptive Statistics

In Table 1, we present selected descriptive statistics for the Trabajar and EDS samples. The Trabajar sample has a lower average income, higher average family size, is more likely to have borrowed to meet their basic needs, receives less from informal sources, is more likely to participate in some form of political organization, and less likely to own various consumer durables (with the exception of a color TV, which appears to be a necessity of life in Argentina).

Table 1. Some Descriptive Statistics of the Treatment and Control Groups

	Trabajar sample	National sample		
Per capita income (\$/person/month)	73.205 (101.843)	366.596 (792.033)		
Average household size	4.894 (2.509)	3.448 (1.981)		
Private pensions (\$/person/month)	10.821 (36.106)	18.927 (67.813)		
Social pensions (\$/person/month)	1.250 (6.719)	.749 (6.896)		
Help from friends and relatives (\$/person/month)	1.515 (16.013)	11.893 (71.977)		
% of households that need to borrow to meet basic needs	32.777 (.887)	18.820 (.263)		
% of population participating in some form of political organization	2.910 (.318)	1.450 (.009)		
% of households that own a telephone	22.660 (.791)	66.150 (.318)		
% of households that own a color TV	75.600 (.811)	77.040 (.283)		
% of households owning a refrigerator with built-in freezer	26.450 (.833)	48.280 (.336)		
% of households owning an automatic washing machine	11.660 (.606)	37.680 (.326)		
	Male	Female	Male	Female
Average age at which currently active household members started working (years)	15.945 (9.716)	17.809 (9.683)	15.658 (6.193)	17.689 (6.772)
Average age at which those household members who are no longer at school dropped out of school (years)	15.333 (8.137)	15.455 (8.813)	16.857 (8.649)	16.789 (7.1306)
% of people in household who were unwell (accident or sick) in the last month	19.030 (.742)	23.260 (.798)	22.130 (.279)	26.700 (.298)

NOTE: Above averages are population-weighted averages. Monetary units are in \$/month, 1997 prices. Standard deviations are reported in the parentheses.

Table 2 gives the percentage distribution of Trabajar participants' families across deciles formed from the EDS with households ranked by income per capita, excluding income from Trabajar. (The poorest decile is split in half.) This is the type of tabulation that is typically made in assessing such a program. It assumes zero foregone income, so each participating family's preintervention income is simply actual income minus wage earnings from the program.

Table 2 suggests that a high proportion of the families of participants come from poor families. In calculating the

Table 2. Location of Trabajar Participants in the National Distribution of Household Income per Capita

	Trabajar sample households	Persons	National sample households	Persons
Poorest 5%	40.2	38.8	5.0	5.6
Next 5%	18.0	21.3	5.0	7.8
Decile 2	17.5	18.5	10.0	13.1
Decile 3	9.9	9.5	10.0	11.7
Decile 4	6.8	5.8	10.0	10.9
Decile 5	2.2	1.9	10.0	9.7
Decile 6	2.5	1.6	10.0	9.1
Decile 7	1.7	1.6	10.0	9.2
Decile 8	.6	.5	10.0	8.2
Decile 9	.4	.3	10.0	7.9
Decile 10	.2	.1	10.1	6.7
Total	100.0	100.0	100.0	100.0

impact, we have only considered participants who have earned at least \$150 from a Trabajar job. The minimum wage offered under Trabajar is \$200. Those reporting less than \$150 as their Trabajar earnings must either be in the last phase of their Trabajar job and or have misreported their income. Since we are interested in the impact on currently active participants in the program, we excluded the observations to get a better, "albeit a more conservative" measure of the impact. We find that 40% of the program participants have a household income per capita that puts them in the poorest 5% of the national population; 60% of participants are drawn from the poorest 10% nationally. By most methods of measuring poverty in Argentina, the poverty rate is about 20%. So 75–85% of the participants are poor by this standard. Such targeting performance is very good by international standards.

Does relaxing the assumption of zero foregone income change the results in Table 1? Using the matching methods described above, we will now see whether that assumption is justified, and how much it matters to an assessment of average gains and their incidence.

3.2 Propensity-Score Matching Estimates

Estimating the propensity score is a crucial step in using matching as an evaluation strategy. Different practices have been adopted to choose a suitable specification of the participation equation (see, for example, Dehejia and Wahba 1998;

Heckman et al. 1998). The underlying principle is that preintervention variables—that are not influenced by participation in the program—should be included in the regression. Existing applications to the evaluation of job training programs in the United States have found employment histories of individuals to be very good predictors of participation. In our data, we do not have any information on their employment histories prior to the implementation of the program. However, unlike the job-training program studies, we have substantial information on community characteristics in which the individuals reside and on the household characteristics of the individuals, as well as data on their educational and demographic backgrounds. We determined that there were over 200 potential variables at our disposal that could reasonably be treated as exogenous to participation. Among these variables, many of the community characteristics were very important predictors of participation, as were a number of household characteristics. We estimated alternative logistic models to predict participation in the program, and selected a final model chosen on the basis of the likelihood function.

Table 3 presents the logit regression used to estimate the propensity scores on the basis of which the matching is subsequently done. The results accord well with expectations from the simple averages in Table 1. Trabajar participants are clearly poorer, as indicated by their housing, neighborhood, schooling, and their subjective perceptions of welfare and expected future prospects (relative to their parents). The participation regression suggests that program participants are more likely to be males who are head of households and married. Participants are likely to be longer term residents of the locality rather than migrants from other areas. The model also predicts that (controlling for other characteristics) Trabajar participants are more likely to be members of political parties and neighborhood associations. This is not surprising given the design of the program since social and political connections will no doubt influence the likelihood of being recruited into a successful subproject proposal. However, participation rates in political parties and local groups are still low, even for Trabajar participants (Table 1).

After estimating the propensity scores for the treated and the comparison group, we plotted them to check the common support condition. We found that there were regions of no overlapping support. We excluded nonparticipants in the nonoverlap region. Furthermore, since we had enough participants in each of the provinces, we tried to limit the bias due to location differences by matching within provinces only.

Based on Table 3, the mean propensity score for the national sample is .075 (with a standard deviation of .125). This is, of course, much lower than the mean score for the Trabajar sample, which is .405 (.266). However, there is considerable overlap in support, with only 3% of nonparticipants having a score less than the lowest value for participants (suggesting that there is considerable unmet demand for work on the scheme). Figure 1 gives the histograms of estimated propensity scores for participants and nonparticipants. After matching, the comparison group of nearest neighbors drawn from the national sample has a mean score of .394 (.253), very close to that of the Trabajar sample.

Tables 4 and 5 give our estimates of average income gains, and their incidence according to fractiles of households ranked by preintervention income per capita. The latter is not observed. To estimate it, we first estimate the gain for each participating household by each of the three methods described in Section 2.2. We then assign each household to a decile using the same decile bounds calibrated from the EDS, but this time, the participants are assigned to the decile implied by their estimated preintervention income as given by actual income minus the estimated net gain.

The nearest neighbor estimate of the average gain is \$157, about three-quarters of the Trabajar wage. The nearest five and nonparametric estimator give appreciably lower gains, of around \$100. Comparing the standard errors across the three estimators, we find that, as expected, there is an increase in precision of the estimators when we move from the nearest neighbor to the nonparametric estimator. The difference in the standard errors between the nearest neighbor and the nearest five estimator is 16%. However, the difference in the standard errors between the nearest five and the nonparametric estimator is only 7%. Thus, while there is an improvement in the precision of the estimates when we resort to “*m*-to-1” matching with *m* > 1, the improvement is only modest compared to the increase in the number of matches. For this reason, computational convenience and to circumvent the small sample problem in the subgroup cases, the rest of the article is based on the “nearest five” estimate.

The average gain using the “nearest five” estimator of \$103 is about half of the average Trabajar wage. Given that there is sizable foregone income, the crude incidence numbers in Table 1 overestimate how pro-poor the program is since preintervention income is lower than is implied by the net gains. Where this bias is most notable is among the poorest 5%; while the nonbehavioral incidence analysis suggests that 40% of participant households are in the poorest 5%, the estimate factoring in foregone incomes is much lower at 10%. Nonetheless, over half of the participant households are in the poorest decile nationally, even allowing for foregone incomes. Given that the poverty rate in Argentina is widely reckoned to be 20%, our results suggest that four out of five Trabajar participants are poor by Argentinean standards.

Figure 2 gives the mean income gain at each level of preintervention income, estimated by a locally weighted smoothed scatterplot of the data. The mean gain falls sharply (although not continuously) up to an income of about \$200 per person per month (which is about the median of the national distribution), and is roughly constant after that. The percentage net gain is highest for the poorest, reaching 74% for the poorest 5%.

To assess the impact on poverty incidence among participants, Figure 3 gives both the observed cumulative distribution of household per-capita income and the estimated counterfactual (preintervention) distribution. (There is automatically first-order dominance given that we have ruled out negative gains on a priori grounds.) We see a 15 percentage point drop in the incidence of poverty due to the program using a poverty line of \$100 per month (for which about 20% of the national population is deemed poor). The impact rises to about a 30 percentage point decline using poverty lines near the bottom of the distribution.

Table 3. Logit Regression of Participation in the Trabajar Program

	Coefficient	t ratio
Cordoba	3.5084	8.395
Chaco	1.0953	2.750
La Pampa	1.2023	3.053
La Rioja	3.1152	7.505
Misiones	1.4492	3.630
Neuquen	1.0367	2.597
Salta	1.3164	3.332
San Juan	1.4462	3.513
Santa Fe	1.5063	3.897
Santiago del Estero	1.4058	3.572
Whether household is located in an emergency town	-.5455	-3.284
... a settlement of 5+ years	-.9622	-3.998
... a social housing area	.3536	4.479
... an area in very damaged condition	-.3197	-2.747
Dwelling has one room (besides bathroom/kitchen)	.7733	7.654
... two rooms	.5247	6.805
... three rooms	.2734	3.902
Main material of interior floors is cement/bricks	.3028	2.579
Water is obtained from manual pumps	-.9468	-2.902
Water shortages in last 12 months	-.2707	-4.535
Portable gas is used for cooking	-.5661	-2.807
Household gets hot water through a central heating service	.6968	2.444
Located <3 blocks from a place where trash is placed habitually	-.3360	-5.015
... <3 blocks from a place which gets flooded	.2218	3.284
... in an area where there is daily collection of trash	.1795	2.016
... in an area with a water network	.7348	4.396
... in an area with sewer network	.2779	4.073
... <5 blocks from closest public transportation	-.2674	-2.202
... <5 blocks from closest public phone	-.3044	-3.109
... <5 blocks from closest public primary school	-.4211	-4.419
... 5-9 blocks from closest public primary school	-.3027	-3.180
... <5 blocks from closest neighborhood health center	.1675	2.309
... 5-9 blocks from closest neighborhood health center	.1678	2.315
... <5 blocks from closest pharmacy	-.4265	-5.129
... <5 blocks from closest mail	-.2709	-2.655
... <10 blocks from a secondary school	-1.0198	-4.231
... 10-30 blocks from a secondary school	-1.0127	-4.253
... 30-50 blocks from a secondary school	-.4955	-1.954
... <10 blocks from a public hospital	-.3943	-3.325
Safety is the major concern in the neighborhood	.2708	2.917
It is a dangerous street for pedestrians to cross	.1472	2.040
Shortages of electricity	.2925	3.084
Drug addiction problem in neighborhood	.3855	-3.786
Male	2.2307	13.961
Head of the household	.3169	2.735
Spouse of the household head	-.6185	-3.858
Legally married	.2211	2.343
Separated after being married	.4397	2.911
Divorced	.3769	2.202
During last 12 months has been absent from household for >1 month	-.4450	-3.182
Born in this locality	.8215	5.019
... in another locality of same province	.5672	3.373
... in another province	.6523	3.867
Lived habitually in this locality for last 5 years	.5326	4.876
Affiliated with a health system only through social work	-.6388	-7.750
... with a health system through unions and private hospital	-.4694	-3.839
... with a health system through social work & mutual benefit society	-1.0715	-3.291
... with a health system because he is a worker	-1.1213	-6.530
Currently attends an educational establishment for primary/secondary school	-.7551	-2.117
Currently a student at tertiary school	.8775	2.650
Dropped out of school because found syllabus uninteresting	-.5386	-3.656
... he/she was finding school difficult	.6700	3.048
... location of school was inconvenient	-.3996	-1.951
Dropped out of school for personal reasons	.3671	2.100
Taken a course in labor training in the last 3 years	.4252	5.244
Never a member of a sports association	.3444	2.826
Regular member of a neighborhood association with some administrative responsibilities	.9705	2.482
Regular member of a neighborhood association with no responsibilities	.8259	2.526
Never a member of union/student association	.5973	2.413
Member of a political party with some administrative responsibilities	.7523	1.900
Member of a political party	1.6387	6.020

(continued)

Table 3. (continued)

	Coefficient	t ratio
Occasional member of a political party	1.3609	5.041
Thinks that 20 years hence, economic situation will be the same as parents' now	.3981	5.401
Reason for above is lack of schooling	-.3705	-5.632
Reason for above is economic situation of country	-.7596	-7.291
Thinks that he and his family are very poor	.5976	6.078
Children born in the last 12 months	.2281	2.693
Pregnant currently	-.9295	-2.435
Constant	-5.6210	-4.390
Log likelihood	-5580	

NOTE: Only significant coefficients in the logit regression are reported in the above table. For omitted categories and for other variables included in the regression, see Addendum (available from the authors).

Tables 7–9 report the net wage gains by fractiles of pre-intervention incomes for three different demographic groups: female participants, participants between the ages of 15–24

years (typically identified as those who are new entrants into the job market), and workers in the age group 25–64 years.

The estimates in Table 7 are not consistent with the existence of income losses due to low female participation in the program. The net wage gains from the program accruing to female participants are virtually identical to the gains for male participants. However, the distribution of female participation is less pro-poor, as indicated by household income per capita; while over half of the members of participating families are in the poorest decile nationally, this is true of less than 40% of the members of female participants' families. This probably reflects lower wages for women in other work, making the Trabajar wage more attractive to the nonpoor.

For the younger cohort, however, the net gains are significantly higher (comparing Tables 8 and 9). Foregone incomes are lower for the young, probably reflecting their lack of experience in the labor market. Because of this, there would be income gains from higher participation by the young. (To the extent that any young participants leave school to join the program, future incomes may suffer.) This suggests that the older workers may well be favored in rationing Trabajar jobs. However, the distribution of gains is more pro-poor for

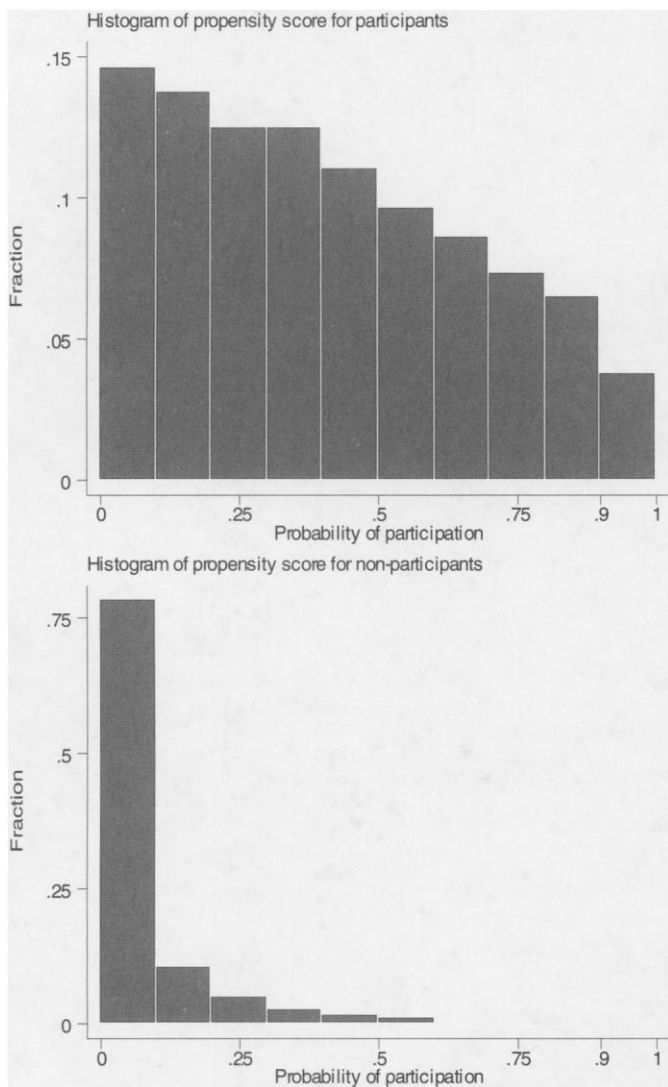


Figure 1. Histograms of Propensity Scores.

Table 4. Net Income Gains From the Program Using Different Estimators

Groups	Nearest neighbor	Nearest five estimator	Nonparametric estimator
Full sample	156.770 (296.083)	102.627 (247.433)	91.678 (230.327)
Ventile 1	372.010 (409.053)	108.543 (210.543)	107.862 (222.831)
Ventile 2	132.662 (260.851)	83.351 (200.379)	63.331 (161.769)
Decile 2	112.166 (230.161)	119.044 (285.357)	93.506 (197.679)
Decile 3	102.058 (176.515)	136.349 (263.939)	120.430 (240.703)
Decile 4	78.740 (248.272)	82.386 (281.863)	89.295 (277.294)
Decile 5	148.711 (434.210)	107.125 (208.313)	205.050 (597.605)
Deciles 6–9	80.965 (191.337)	111.229 (278.584)	114.913 (196.906)
Decile 10	No participants in this decile		

NOTE: Standard errors in parentheses.

Table 5. Persons of Participant Households Using Different Estimators

Groups	Nearest neighbor estimator	Nearest five estimator	Nonparametric estimator
Full sample	100.000	100.000	100.000
Ventile 1	21.525	10.207	8.671
Ventile 2	41.278	42.284	39.460
Decile 2	20.732	26.908	27.734
Decile 3	8.084	10.892	13.460
Decile 4	5.403	6.307	7.302
Decile 5	1.842	2.069	1.652
Deciles 6–9	1.135	1.334	1.722
Decile 10	No participants in this decile		

the older workers, with almost 60% coming from the poorest decile. Pushing for higher participation by the young entails a short-term tradeoff between average gains and a better distribution. It may also entail a longer term tradeoff with future incomes of the young by reducing schooling.

3.3 Economic Interpretation

Although we find that program participation falls off sharply as household income rises, the net gains conditional on participation do not fall among the upper half of the income distribution (Fig. 2). Since the program wage rate is about the same for all participants, foregone income among participants appears to be independent of family income above about \$200 per person per month. This may be surprising at first sight. The standard model of self-targeting through work requirements postulates that foregone income tends to be higher for higher income groups (Section 1.1).

We can offer the following explanation. The Trabajar wage is almost certainly too low to attract a worker out of a regular job. For a worker with such a job, let the foregone income from joining the program be $f_e(Y) > W$, where (as in Section 1.1) Y is the preintervention income of the worker's household, W is the wage rate offered in the Trabajar Program, and the function f_e is strictly increasing.

For an unemployed worker, however, only miscellaneous odd jobs are available. Anyone can get this work, and it does

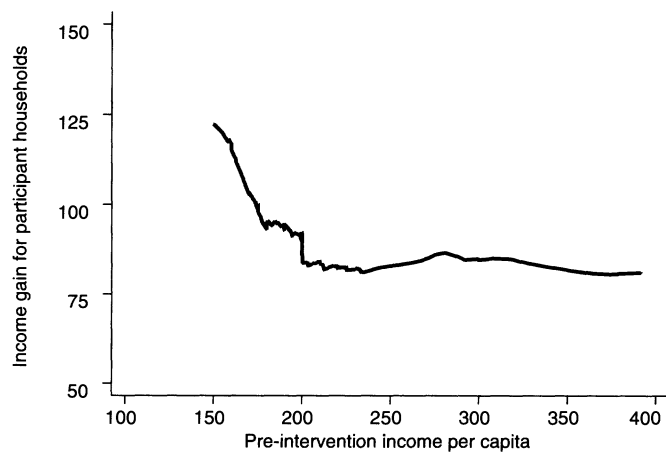


Figure 2. Mean Income Gain Plotted Against Preintervention Income.

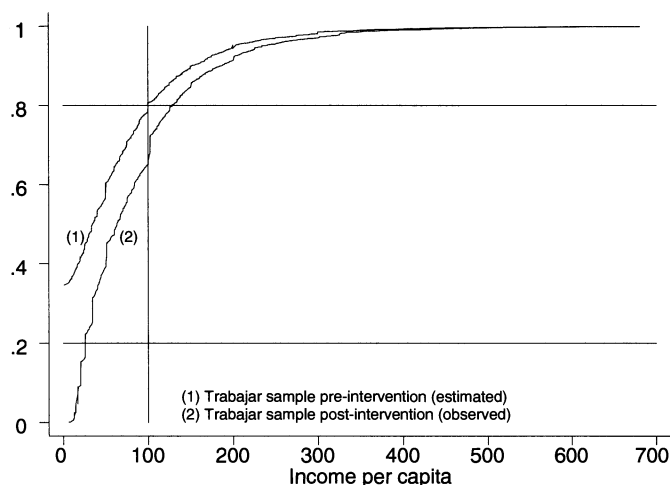


Figure 3. Impacts on Poverty Incidence.

not earn any more for someone from a well-off family than a poor one. Let this “odd-job” foregone income be $f_u < W$, and assume that f_u is independent of Y . Let the rate of unemployment be U , and assume that this is a decreasing function of Y ; that is also consistent with the evidence for this setting (Section 2). The average foregone income if one joins the Trabajar Program is then

$$F(Y) = U(Y)f_u + [1 - U(Y)]f_e(Y). \tag{5}$$

This is strictly increasing in Y , as in the standard model of self-targeting (Section 1.1).

In this model, unemployed workers will want to participate in the Trabajar program, while the employed will not be interested in participating (assuming that the alternative work is judged equal in other respects, although this can be relaxed without altering the main point of this model). The program will successfully screen the two groups. We will see a fall in Trabajar participation as income rises, as in Table 4. However, when we calculate the foregone income of actual participants, we will get f_u , not $F(Y)$. The measured net gains among actual participants will not vary systematically with preintervention income, even though self-targeting of the poor is excellent. Our finding that foregone income conditional on participation does not fall as income rises among the upper half of the distribution is still consistent with good overall targeting through self-selection.

4. CONCLUSIONS

A counterfactual income in the absence of the program is missing data, and assumptions will have to be made to make up for these missing data. The assumptions made in program evaluations are often dictated by data availability. In assessing the gains from antipoverty programs—programs that are often set up rapidly in response to a crisis—it is common to only have access to a single cross-sectional survey done after the program is introduced. Propensity-score matching methods of evaluation combine a single cross-sectional survey of program participants with a comparable larger cross-sectional survey from which a comparison group is chosen. With sufficiently

Table 6. Net Income Gains From the Program

Groups	% of participants in ventile/decile	Persons of participant households	Household income of Trabajar participants	Net income gain due to the program	Net gain as % of preintervention income
Full sample	100.000	100.000	501.181 (364.632)	102.627 (247.433)	25.926
Ventile 1	6.070	10.207	299.102 (221.119)	108.543 (210.543)	74.830
Ventile 2	36.535	42.284	369.194 (265.054)	83.351 (200.379)	24.746
Decile 2	26.700	26.908	548.789 (353.237)	119.044 (285.357)	26.566
Decile 3	12.601	10.892	685.413 (358.139)	136.349 (263.939)	23.056
Decile 4	11.833	6.307	543.680 (441.794)	82.386 (281.863)	13.483
Decile 5	3.496	2.069	749.443 (384.025)	107.125 (208.313)	14.975
Deciles 6–9	2.766	1.334	879.382 (496.091)	111.229 (278.584)	11.469
Decile 10	No participants in this decile				

NOTE: These numbers correspond to the nearest five estimator reported in Table 4. Standard errors in parentheses.

detailed cross-sectional data on both participants and nonparticipants, these methods can allow an assessment of behavioral responses without preintervention baseline data or randomization. The accuracy of this method will depend on how well one can assure that treatment and comparison groups come from the same economic environment and were given the same survey instrument. The method cannot rule out the possibility of selection bias due to unobserved differences between participants and even a well-matched comparison group, although there is evidence this may well be an overrated problem (Heckman et al. 1998; Dehejia and Wahba 1998, 1999).

We have applied recent advances in matching methods to Argentina's Trabajar Program. While neither a baseline survey nor randomization were feasible options in this case, the program is well suited to matching methods.

We find that program participants are more likely to be poor than nonparticipants by a variety of both objective and subjective indicators. The participants tend to be less well educated, they tend to live in poorer neighborhoods, and they tend to be members of neighborhood associations and political parties. The relatively low wage rate clearly makes the program unattractive to the nonpoor.

Using our model of program participation to find the best matches from the national sample for each Trabajar worker, we have estimated the net income gain from the program. We find that ignoring foregone incomes greatly overstates the average gains from the program, although sizable gains of about half the gross wage are still found. Even allowing for foregone incomes, the program's benefit incidence is decidedly pro-poor, reflecting the self-targeting feature of the

Table 7. Net Income Gains for Female Participants

Groups	% of participants in ventile/decile	Persons of participant households	Household income of Trabajar participants	Net income gain due to the program	Net gain as % of preintervention income
Full sample	100.000	100.000	571.890 (382.580)	103.904 (277.340)	22.818
Ventile 1	3.289	5.645	351.300 (428.177)	158.240 (409.963)	82.298
Ventile 2	25.000	31.948	424.370 (320.742)	101.360 (281.681)	30.767
Decile 2	32.895	34.000	520.800 (286.501)	87.490 (202.641)	18.400
Decile 3	16.447	15.261	718.660 (493.045)	136.284 (420.507)	21.166
Decile 4	12.500	8.251	655.579 (322.183)	92.353 (196.851)	14.123
Decile 5	4.605	2.605	696.143 (224.638)	79.000 (126.926)	12.558
Deciles 6–9	5.263	2.295	963.663 (473.150)	132.663 (248.887)	14.006
Decile 10	No participants in this decile				

NOTE: These numbers correspond to the nearest five estimator for the subgroup of female participants. Standard errors in parentheses.

Table 8. Income Gains for Those 15–24 Years of Age

Groups	% of participants in decile	Persons of participant households	Household income of <i>Trabajar</i> participants	Net income gain due to the program	Net gain as % of preintervention income
Full sample	100.000	100.000	618.789 (401.990)	125.241 (255.903)	25.592
Decile 1	30.214	37.012	434.619 (332.660)	121.500 (261.500)	35.287
Decile 2	31.567	34.431	636.060 (353.555)	143.657 (272.418)	28.629
Decile 3	16.234	14.776	738.666 (383.006)	133.560 (275.162)	19.921
Decile 4	11.838	8.313	620.135 (378.544)	73.146 (169.706)	10.559
Decile 5	10.034	3.618	886.735 (422.0520)	152.898 (262.636)	17.400
Deciles 6–9	3.495	1.850	1,069.600 (608.221)	102.142 (176.652)	9.550
Decile 10	No participants in this decile				

NOTE: These numbers correspond to the nearest five estimator for the subgroup of 15–24 year participants. Standard errors in parentheses.

programs' design. Average gains are very similar between men and women, but are higher for younger workers. Higher female participation would not enhance average income gains, and the distribution of the gains would worsen. Higher participation by the young would raise average gains, but would also worsen the distribution.

We do not have to drop any participants from our sample in computing the impact estimator. This is probably the result of having a large comparison group sample from which we could draw our matches. Finally, as discussed in the articles by Rosenbaum and Rubin (1985) and Rubin and Thomas (1996), we find that the precision of the matching impact estimators does improve when we use “*m*-to-1” matching with $m > 1$. However, the increase in the precision is moderate. Thus, matching each participant to five nonparticipants does not reduce the standard errors fivefold. The reduction in the

standard errors is even smaller when we compare the nearest five estimator to the nonparametric estimator which uses all of the information in the nonparticipant sample.

ACKNOWLEDGMENTS

The work reported in this article is one element of the ex-post evaluation of the World Bank's Social Protection II Project in Argentina. The support of the Bank's Research Committee (under RPO 681-39) is gratefully acknowledged. The article draws on data provided by the SIEMPRO unit of the Ministry of Social Development, Government of Argentina. The authors are especially grateful to Joon Hee Bang and Liliana Danilovich of SIEMPRO for their help with the data. The authors' thanks also go to staff of the *Trabajar* Project Office in the Ministry of Labor, Government

Table 9. Income Gains for Those 25–64 Years of Age

Groups	% of participants in decile	Persons of participant households	Household income of <i>Trabajar</i> participants	Net income gain due to the program	Net gain as % of preintervention income
Full sample	100.000	100.000	443.443 (328.253)	85.820 (231.032)	22.241
Bottom 5%	7.423	13.062	307.386 (251.260)	97.474 (221.489)	38.564
Next 5%	39.451	45.767	342.499 (252.305)	71.809 (205.962)	22.207
Decile 2	26.651	24.938	487.939 (251.477)	86.833 (180.047)	21.204
Decile 3	11.046	8.851	625.097 (395.1020)	122.505 (334.238)	25.578
Decile 4	10.812	5.046	476.941 (410.221)	74.724 (271.968)	13.594
Decile 5	2.221	1.343	755.921 (561.663)	123.176 (331.995)	13.996
Deciles 6–9	2.396	.993	753.736 (437.869)	115.478 (224.021)	15.834
Decile 10	No participants in this decile				

NOTE: These numbers correspond to the nearest five estimator for the subgroup of 25–64 year participants. Standard errors in parentheses.

of Argentina, who provided the necessary data on their program, and gave this evaluation their full support. They also thank the journal's associate editor and an anonymous referee for useful comments. Petra Todd kindly advised the authors on matching methods. Useful comments were received from Polly Jones, Dominique van de Walle, and seminar participants at the World Bank, the Indian Statistical Institute, Delhi, and the Institute of Fiscal Studies, London.

[Received October 1999. Revised June 2001.]

REFERENCES

- Atkinson, A. (1987), "On the Measurement of Poverty," *Econometrica*, 55, 749–764.
- Besley, T., and Coate, S. (1992), "Workfare vs. Welfare: Incentive Arguments for Work Requirements in Poverty Alleviation Programs," *American Economic Review*, 82, 249–261.
- Datt, G., and Ravallion, M. (1994), "Transfer Benefits From Public Works Employment: Evidence From Rural India," *Economic Journal*, 104, 1346–1369.
- Dehejia, R. H., and Wahba, S. (1998), "Propensity Score Matching Methods for Non-Experimental Causal Studies," NBER Working Paper 6829, Cambridge, MA.
- (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.
- Friedlander, D., Greenberg, D., and Robins, P. (1997), "Evaluating Government Training Programs for the Economically Disadvantaged," *Journal of Economic Literature*, 35, 1809–1855.
- Heckman, J., Ichimura, H., and Todd, P. (1997), "Matching as an Econometric Evaluation Estimator: Evidence From Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.
- Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1099.
- (1996), "Nonparametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA. Part II, Theory and Methods and Monte-Carlo Evidence," mimeo, University of Chicago.
- Heckman, J., and Robb, R. (1985), "Alternative Methods of Evaluating the Impact of Interventions: An Overview," *Journal of Econometrics*, 30, 239–267.
- Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.
- Hotz, J., Imbens, G., and Mortimer, J. (1999), "Predicting the Efficacy of Future Training Programs Using Past Experiences," NBER Technical Working Paper T0238, Cambridge, MA.
- Lechner, M. (1999), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany," *Journal of Business and Economic Statistics*, 17, 74–90.
- Lipton, M., and Ravallion, M. (1995), "Poverty and Policy," in *Handbook of Development Economics, Volume 3*, eds. J. Behrman and T. N. Srinivasan, Amsterdam: North-Holland.
- Lopez, M. (1999), "Does Bilingual Education Affect Educational Attainment and Labor Market Outcomes?," Working Paper, University of Maryland School of Public Affairs.
- Manski, C., and Lerman, S. (1977), "The Estimation of Choice Probabilities From Choice-Based Samples," *Econometrica*, 45, 1977–1988.
- Ravallion, M. (1991), "Reaching the Rural Poor Through Public Employment: Arguments, Evidence and Lessons From South Asia," *World Bank Research Observer*, 6, 153–175.
- (1999), "Appraising Workfare," *World Bank Research Observer*, 14, 31–48.
- (2000), "Monitoring Targeting Performance when Decentralized Allocations to the Poor Area are Unobserved," *World Bank Economic Review*, 14, 331–345.
- Rosenbaum, P., and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.
- (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *American Statistician*, 39, 35–39.
- Rubin, D., and Thomas, N. (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 249–264.
- Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.
- Subbarao, K., Bonnerjee, A., Braithwaite, J., Carvalho, S., Ezemenari, K., Graham, C., and Thompson, A. (1997), *Safety Net Programs and Poverty Reduction: Lessons From Cross-Country Experience*, World Bank, Washington, DC.
- Todd, P. (1995), "Matching and Local Linear Regression Approaches to Solving the Evaluation Problem With a Semiparametric Propensity Score," mimeo, University of Chicago.