Accounting for selection bias in impact analysis of a rural development program:

An application using propensity score matching.

Maria Alexandra Peralta[a1], Scott M. Swinton[b] and Mywish Maredia[c]

a, b, c Department of Agricultural, Food and Resource Economics

Michigan State University

Selected Poster prepared for presentation at the International Association of Agricultural Economiss (IAAE) Trietnnial Conference, Foz do Iguaçu, Brazil, 18-24 August, 2012

The authors gratefully acknowledge funding by the Howard G. Buffett Foundation through the Catholic Relief Services Central America office. We thank Songqing Jin and Jeffrey Wooldridge for helpful comments.

[1] Email: peralta3@msu.edu

Abstract

When evaluating the impact of a program, the effects of interventions on program outcomes must be measured against a valid counterfactual case. Constructing a valid counterfactual is especially important when experimental data is not available. Building a baseline ensuring that treatment and comparison groups are similar as well as identifying potential sources of bias are essential first steps towards constructing a valid counterfactual. This paper assesses the comparability of groups of participants and non-participants for conducting an impact evaluation of the Agriculture for Basic Needs (A4N) program in Nicaragua. We examine the degree of similarity between A4N participant and non-participant comparison households using propensity scores (estimated probability of program placement). Propensity scores are matched for the two groups, comparing results from using caliper matching and nearest neighbor matching without and with replacement, and caliper matching. The analysis uses the pretreatment characteristics of households belonging to the treatment (participant) and comparison (non-participant) groups in order to verify whether the comparison group is statistically similar to the treatment group.

Introduction

When evaluating the impact of a program, challenges arise in estimating the benefits of interventions on program outcomes, particularly when evaluations are conducted in a non-experimental setting. In general, when conducting an impact assessment, we are able to observe realized outcomes for program beneficiaries when they participate in the program interventions, but we do not know what might have happened if the beneficiaries had not participated in the program. Therefore, program evaluation is a problem of missing data.

A key challenge is to construct a valid counterfactual or comparison group comprised of individuals who did not participate in the program but have characteristics similar to the program participants. Evaluating program performance in comparison to such a counterfactual group permits assessment of whether observed changes are due to the program, as opposed to changes that would have occurred anyway. A complicating factor is that many anti-poverty programs aim to target individuals with certain characteristics (like low incomes), so the targeted program population (i.e. people who meet the eligibility criteria) will often differ systematically from the non-targeted population. In the program evaluation literature this is known as selection bias on observable characteristics. Another source of bias may arise within the targeted eligible population when some individuals voluntarily select to participate in all the components of a program intervention offered to them, some participate in only some components of a program, and some elect not to participate at all. Such self-selection bias is known as bias on un-observables. In other words, such selection bias occurs when unobservable characteristics that affect the decision to participate or not to participate in a program are correlated with characteristics that affect the process that determines program outcomes of interest.

Conducting a program evaluation to isolate true program impacts requires determining an effective way to construct a valid counterfactual that is as similar as possible to the factual case in both observable and unobservable characteristics (i.e., is free from any bias). The literature on impact evaluation proposes different methods to correct selection bias and construct a valid counterfactual under different assumptions about the choice of program participants. This includes methods such as difference in difference (DID), propensity score matching (PSM) and regression discontinuity design (RDD). The latter two correct for bias on observable characteristics, usually when there is cross sectional data available. In PSM, the estimated probability of being selected for program participation, known as the propensity score, is used to match members of an untreated comparison group with members of the treatment group for the purpose of estimating impacts as the difference in outcomes between the treatment and comparison group(Becerril & Abdulai, 2009; Cox-Edwards & Rodríguez-Oreggia, 2009; Dillon, 2011; Jyotsna & Ravallion, 2003; Mendola, 2007; Cavatassi, Salazar, González-Flores, & Winters, 2011). RDD can be used when program participants are chosen according to a threshold value for a given characteristic that determines program eligibility. A comparison group with similar characteristics to the ones of the treated can be selected from nonparticipants who fall just beyond the eligibility threshold. In a sharp regression discontinuity design, the assignment is a deterministic function of the covariate used for selecting program participants. Alternatively, in a fuzzy regression discontinuity design, where probability of being eligible does not necessarily have to switch from zero to one at the threshold, producing a jump on the probability distribution between participants and nonparticipants(Carrella, Hoekstra, & West, 2011; Klinger & Schündeln, 2011; Lalive, 2008; Skoufias & Buddermeyer, 2004).

Among the methods that correct for selection bias on un-observables, the difference in difference (DID) estimator is effective and feasible with panel data regression (Centeno, Centeno, & Novo, 2009; Centeno, Centeno, & Novo, 2009; Schultz, 2004; Del Carpio, Loayza, & Datar, 2011; Bravo-Ureta, Nunes-Almeida, Solis, & Inestrosa, 2011). The DID estimator calculates an impact estimate by comparing the sample data between program participants and non-participants, calculating the difference between the mean outcomes of each group before and after the intervention and then the difference in means between these two differences. By taking differences, it removes any bias related to unobservable common time trends. Conducting a program evaluation to isolate true program impacts requires determining an effective way to construct a valid counterfactual.

Building a panel data set requires a baseline survey and one or more follow up surveys, which is what is planned for the impact evaluation of the Agriculture for Basic Needs (A4N) program. The economic impact evaluation of the A4N program implemented in Nicaragua to be based on two different methods to estimate program gains. One is comparing the post-treatment outcomes between the treatment and comparison households based on PSM using the pre-treatment (baseline) data, and the other is the DID analysis The estimates of program gains from these two methods will be checked for robustness. Since the project is still in progress and only the baseline data have been collected so far, this paper focuses on the comparison of different PSM methods to establish a valid counterfactual.

The estimation of the propensity score (PS) considers the eligibility criteria of the implementing agency, while attempting to control for household characteristics, asset endowments and village characteristics (i.e., observable characteristics). The A4N program eligibility criteria include households with limited resources, as measured by the area of cultivated land, access to water, sanitation and electricity, and good quality house construction materials. The program also targets households with children under five years old, those headed by women, and those facing food scarcity.

The rest of this paper presents an overview of the A4N program, introduces the PSM method, describes the survey data used for this exercise, presents the estimates of the probability of program placement and then tests for sufficient overlap in probability distributions between treatment and comparison groups in order to permit generating a valid counterfactual.

The Agriculture for Basic Needs (A4N) program

Catholic Relief Services (CRS) and its partners, Caritas and the Foundation for Research and Rural Development (FIDER), are implementing the A4N program in Nicaragua. Activities started in August of 2009 and are scheduled to run until August of 2012. This program aims to provide farmers with a set of skills for achieving sustainable farm production and increased agricultural income. To accomplish these objectives, the program promotes conservation agriculture and nutritious crops, improved crop varieties, micro-livestock (poultry and fish), integrated pest management and practices to diminish post-harvest crop loss. Other program interventions include saving and lending groups, post-harvest processing, expanded participation in markets, and promotion of farmer innovation groups.

The A4N project has first targeted villages that are considered poor. These villages are characterized by high levels of poverty, located in areas of natural resource degradation, and high vulnerability to natural disasters. Within these villages, eligible households are expected to meet most of the following official eligibility criteria:

- Cultivated land area less than two manzanas (1 Mz = 1.73 acres).
- Cultivated land on steep slopes.
- Lack of access to any of the following public services: piped water, sanitation, and electricity.
- Materials for house walls not brick or concrete; roof not concrete, zinc or brick; floor not concrete, ceramic or tile.
- The household experiences hunger during some period of the year.
- Household head is female.
- Household includes children younger than five years old.

Within the villages, promoters trained by the A4N form groups of 15 to 20 eligible farmers and offer them the opportunity to participate in selected program interventions. Not all interventions are offered to each group of farmers. Due to limited project resources, interventions are prioritized and offered where the need is considered greatest.

In the particular case of the A4N program, it is difficult to exclude the participation of village members who are not officially eligible, and the program allows for technically ineligible individuals to participate in the hope that they can facilitate spreading the benefits of the interventions during and after program implementation.

Since program participants select themselves to participate in the program interventions they are offered. In the case of A4N program, the impact evaluation must account for potential selection bias from two sources--selection into the A4N project via official eligibility criteria and self-selection into specific A4N activities by A4N participants.

Determining the comparison group by Propensity Score Matching.

The main conditions for using propensity score matching (PSM) for constructing the counterfactual are that 1) there is unconfoundedness in program assignment, given by Equation (1); and 2) there is a common support (overlap) between the probability distribution of program participants and non-participants, given by equation (2) (Caliendo & Kopeinig, 2007).

$$y_0, y_1 \perp \text{Partic} \mid \mathbf{x} \qquad (1)$$

$$0 < \text{prob}(\text{Partic}=1 \mid \mathbf{x}) < 1 \qquad (2)$$

where y_0 is the outcome for non-participants and y_1 is the outcome for participants, “Partic” is participation and **x** represents a set of variables that may influence participation. The sign ⊥ means that program outcomes are independent of program participation, conditional on x.

The PSM method consists of choosing the comparison group according to the probability of being selected for a treatment, given a set of observable pre-treatment characteristics and outcome values that do not change with program intervention but that affect program placement.

The characteristics of the treatment and the comparison groups (individuals participating and not participating in the program) prior to program interventions are used to determine their probability of participating in the program. The propensity score is the estimated baseline (i.e., pre-treatment) probability of being selected for program participation. It is used to test for the degree of overlap in the probability-of-selection distributions between the comparison and treatment groups to determine whether the two groups can be matched. Matching is important, because it means the two groups are sufficiently similar that the matched comparison group can serve as a valid counterfactual to estimate program gains after the treatment.

The empirical analysis is conducted in three steps. In the first step, sample means of individual covariate variables are compared in order to test for potential selection bias between non-participants and participants. Covariates are compared via a measure of bias defined as follows:

$$bias = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_1^2 + s_0^2}} * 100, \quad (3)$$

where $\bar{x}_1$ and $\bar{x}_0$ are the means of the sample averages of variable j for the groups of participants (1) and non-participants (0), and s_1 and s_0 estimated standard errors for variable j for participants and non-participants. An absolute value of bias above 25 is typically interpreted to mean that the two groups are not similar by those covariates.[2]

In the second step, the probability of program participation is estimated by the use of a binary response econometric model. The dependent variable is participation in the program, which takes values of one for farmers participating in the program and zero otherwise. Participation in A4N is defined as participating and benefiting from any one or more of the program interventions.

To assess the overall program participation, a logit model is estimated. The expected probability of program participation is

$$E(y|\boldsymbol{x}) = P(Partic{=}1|\boldsymbol{x}) = G(\boldsymbol{x\beta}) \quad (4)$$

Here, $0 < G(\mathbf{x}\boldsymbol{\beta}) < 1$, G refers to the logistic probability distribution function, and $\mathbf{x}\boldsymbol{\beta}$ is an index model, where **x** represents the set of explanatory variables (Heckman & Smith, 2004; Coady & Parker, 2009). In this case, the explanatory variables refer to program eligibility criteria (*elig*), household characteristics (*househ*), village characteristics (*village*), farm characteristics (*farm*), and wealth (*wealth*).

$$\mathbf{x}\boldsymbol{\beta} = \beta_1+\beta_2 eleg+\beta_3 househ+\beta_4 village+\beta_6 farm+\beta_7 wealth+\varepsilon \quad (5)$$

The partial effects of the beta coefficients can be used to measure the marginal effect of each of these characteristics on the probability of program participation. From the

[2] Personal communication, J. Wooldridge, Cross Sectional and Panel Data Econometrics II class (EC 821b), Michigan State University, spring, 2011.

estimation we obtain the estimated probability of participation for both participants and non-participants, given by equation (6),

$$\hat{P}(Partic) = G(\hat{\beta}_1 + \hat{\beta}_2 elig + \hat{\beta}_3 househ + \hat{\beta}_4 village + \hat{\beta}_5 farm + \hat{\beta}_6 wealth) \quad (6)$$

Dehejia and Wahba's (2002) suggested algorithm for estimating the propensity score is then applied to determine whether higher order terms and/or interaction terms need to be included in the model. With these estimated probabilities we check for the overlap of the probability distributions of selection into the two groups and conduct matching of participants to non-participants using the propensity scores (Wooldridge, 2010).

In the third empirical analysis step, non-participant observations are matched, one to one, with a participant observation by the value of its estimated propensity score, minimizing the absolute difference between the estimated propensity scores between observations. This procedure, known as matching by nearest neighbor (NN), was first conducted without replacement (a non-participant observation is used only once to match a participant observation), to compare participants and not participants on the estimated value of their propensity scores. Caliper matching was conducted as well, where individual observations for non-participants are paired with matching participant observations within a caliper or range of its propensity score. Then the NN matching procedure was conducted with replacement, which means that an observation of a non-participant can be used more than once as a match for a non-participant observation.

After matching, we conducted balancing tests, estimated the bias (as defined by equation (3)), and conducted t-tests for equal means before and after matching. These tests were conducted for each of the covariates included in the propensity score estimation for both participants and non-participants. The tests aim to determine whether the matching procedures have served the purpose of making participants and non-participant groups more similar, or if on the contrary, differences have increased after matching of observations in the two groups. For the purpose of constructing a valid counterfactual, it is desirable to identify nonparticipants whose estimated probabilities of selection are close to those of the control group.

The data.

The data used in this paper comes from a household and village baseline survey conducted in the departments of Estelí, Jinotega and Matagalpa, in northwestern Nicaragua between July and September of 2010.

The sampled villages were selected according to the population weights of each of the municipalities where the project intervened. The villages for the survey were randomly selected from the lists of participating villages provided by CRS. Non-participant villages were identified according to national census data on poverty levels, as measured by the index of unmet basic needs (Instituto Nacional de Información de Desarrollo, 2005), the importance of staple crops, small landholdings (Instituto Nacional de Información de Desarrollo, 2003) and location in the same agrarian zones (Nitlapan, 2001).

From each community, 10 - 15 households were selected randomly from the lists of all households. The aim was to collect information from 10 households in the small villages and 15 households in the larger villages. In A4N participant villages, CRS provided lists of participating households. In non-participant villages, sample lists were developed in consultation with village leaders, who were requested to identify households that would meet the eligibility criteria of the A4N program.

A total of 297 participant households in 31 participant villages, and 350 non-participant households in 32 non-participant villages were used for this exercise. More non-participant households were interviewed to take into account the trimming of observations to be done when applying the propensity score matching. A survey of village characteristics was conducted among village leaders in each of the 63 villages used here.

Results

Based on comparison of individual covariate sample means (Table 1), the A4N participant and the non-participant groups are similar by most of the covariates. All bias measurements for the eligibility criteria variables are less than 25 except for female household head, with a difference of 27, which is marginal. Household characteristics and asset endowments also do not differ significantly.

In terms of village characteristics, the bias for distance to nearest market is 32, indicating that A4N program participants are located in villages that are closer to markets than the non-participant villages. However, there is no significant bias in the distance to the nearest paved road.

Looking at the eligibility criteria, the average area of cultivated land among the participants is 3.2 Mz, which is greater than the 2.5 Mz maximum area of land proposed as a formal eligibility criterion for participation. In practice, this type of eligibility criterion is difficult to enforce.

The propensity score model was estimated first using a logit model with the data from 294 treated and 341 non-treated households. Fifteen observations across the two groups were dropped due to missing data. Dehejia and Wahba's (2002) algorithm for estimating the propensity score (see Appendix 1 for details) was used to determine whether interaction terms and higher level terms were required to improve the estimation. It was found that these terms were not required, so the model chosen was a parsimonious logit model with all covariates entering linearly.

The logit model estimates probability of program participation (Table 2). Results suggest that female-headed households are more likely to participate in the A4N program. Similarly, households in villages located farther away from markets, or in villages that lack access to a health facility or that include a high proportion of farms less than 10 Mz are more likely to participate in the A4N program. These coefficient estimates are significant at the 0.01 probability of a Type I Error, as measured by both the logit model in Table 2 and the tests for equal means in Table 1, implying that the participants and non-participants differ on these covariates.

The probability distributions of selection into the A4N participant and non-participant groups are presented in Graph 1. Only two A4N participant observations fall outside the common support. However, the two distributions are not mirror images, and the non-participant distribution contains more observations with propensity scores bellow 0.6, and a disproportionate number of observations with propensity scores below 0.2, compared with participants observations, for which there are not observations with propensity scores over this range.

Following standard procedures for propensity score matching, the next step was to conduct matching of participant and non-participant observations according to the values of the propensity score. First we started with matching on the nearest neighbor, or closest value of propensity score, without replacement, to test how close the estimated propensity scores is from the treated and untreated groups. The results in Graph 2, show that 294 pairs were formed, but after about 100 observations the quality of the matches decreases. The propensity scores for non-participants get farther and farther away from the estimated propensity score for the participants, which is a consequence of having fewer matches as the propensity scores get closer to one. Graph 3 displays the results for caliper matching with a range of 0.001. Within this narrow range, only 101 pairs could be formed. Upon expanding the caliper range to 0.01, 183 pairs could be formed (Graph 4). However there are still more than 100 participant observations lacking a match within that range.

The use of the nearest neighbor *with replacement* appears to be more attractive, as shown in Graph 5. NN with replacement allowed 294 pairs to be formed with 294 participants and 141 non-participants. As the propensity scores got closer to one, non-participant observations were used more than once to form pairs with corresponding participant observations. This method clearly improves the quality of the matches, and allows pairing each participant observation with a corresponding non-participant.

The results for the balancing tests for nearest neighbor matching with replacement are shown in Table 3. There are both reductions and increases in bias for different covariates included. This finding occurs because PSM only considers the estimated probability of program participation for pairing the data, and not the covariates directly. The most remarkable cases are for increases in bias for the village level variables - size of the village population and distance to closest paved road - with an increase in bias over 400%; however in absolute terms the value of the bias measure is still below 25.

In summary, for this particular set of data, matching with replacement appears as a most suitable way to construct the comparison group. The need for replacement is driven by differences in certain household-level characteristics of non-participants, such as frequency of female household heads and by certain non-participant village level characteristics, such as distance to the nearest market, which reduce the observations that are suitable for matching with the participants.

Since the second round of panel data required for conducting impact evaluation of the A4N project is not available yet, the main goal of testing the level of overlap between participants and non-participants included in the sample is to determine whether there are enough observations to generate a subset of non-participant observations to perform propensity score matching of nearest neighbor with replacement.

Conclusions

When conducting baseline data collection for impact evaluation of a program that was not implemented based on the principles of randomized assignment, it is important to determine whether the non-project comparison group households are sufficiently similar to support subsequent comparisons. One method for checking if there is enough overlap between the two groups is to estimate the propensity scores and using the score for each observation as a matching variable.

Propensity score matching is frequently used to improve overlap between the probability distributions of program placement between participants and non-participants based on pre-treatment characteristics in order to estimate average treatment effects. In cases like the one presented in this study, overlap is not a problem. The predicted probabilities of program participation estimated in this paper allow us to conclude that there do exist some differences in the characteristics for participants and non-participants, such as gender of household head and distance between the village and nearest market. Nonetheless, there is sufficient overlap between the observable traits of participant and non-participant (comparison) households to enable valid comparisons. In this study, one to one matching without replacement or one to one matching based on caliper method, are not the best choices. Non-participants' propensity scores concentrate towards zero, and matches for participants' propensity scores decrease when we get closer to one. Therefore nearest neighbor matching with replacement emerges as the best alternative to construct a comparison group that is statistically similar to the treatment group for impact assessment.

References

Abebaw, D., Fentieb, Y., & Kassac, B. (2010). The impact of a food security programnext term on household food consumption in Northwestern Ethiopia: A matching estimator approach. *Food Policy , 35* (4), 286-293 .

Ashby, J., Heinriech, G., Burpee, G., Remington, T., Wilson, K., Quiroz, C. A., et al. (2009). What farmers want: collective capacity for sustainable entrepreneurship. *International Journal of Agricultural Sustainibility , 7* (2), 130-146.

Becerril, J., & Abdulai, A. (2009). The Impact of Improved Maize varieties on Poverty in Mexico: A propensity Score Matching Approach. *World Development , 38* (7).

Bezabih, M., Kohlin, G., & Mannberg, A. (2011). Trust, tenure insecurity, and land certification in rural Ethiopia. *Journal of Socio-Economics , 40* (6), 833-843 .

Bravo-Ureta, B. E., Nunes-Almeida, A., Solis, D., & Inestrosa, A. (2011). The Economic Impact of Marena's Investments on Sustainable Agricultural Systems in Honduras. *Journal of Agricultural Economics , 62* (2), 429-448.

Butcher, K. F., McEwan, P. J., & Taylor, C. H. (2010). The effects of quantitative skills training on college outcomes and peers. *Economics of Education Review , 29* (2), 187-199.

Caliendo, M., & Kopeinig, S. (2007). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys , 22*, 31-77.

Carrella, S. E., Hoekstra, M., & West, J. E. (2011). Does drinking impair college performance? Evidence from a regression discontinuity design approach. *Journal of Public Economics , 95* (1-2), 54-62 .

Cavatassi, R., Salazar, L., González-Flores, M., & Winters, P. (2011). You have free access to this content How do Agricultural Programmes Alter Crop Production? Evidence from Ecuador. *Journal of Agricultural Economics , 62* (2), 403–428.

Centeno, L., Centeno, M., & Novo, A. A. (2009). Evaluating job-search programsnext term for old and young individuals: Heterogeneous impact on unemployment duration. *Labour Economics , 16* (1), 12-25.

Chowa, G. A., & Elliott III, W. (2011). An asset approach to increasing perceived household economic stability among families in Uganda . *Journal of Socio-Economics , 40* (1), 81-87.

Coady, D. P., & Parker, S. W. (2009). Targeting Performance under Self-selecion and Administrative Targeting Methods. *Economic Development and Cultural Change , 57* (3), 559-587.

Cox-Edwards, A., & Rodríguez-Oreggia, E. (2009). Remmitances and Labor Participation in Mexico: An Analysis Using Propensity Score Matching. *World Development , 37* (5), 1004-1014.

Dehejia H., R., & Wahba, S. (2002). Propensity Score Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics , 84* (1), 151-161.

Del Carpio, X. V., Loayza, N., & Datar, G. (2011). Is Irrigation Rehabilitation Good for Poor Farmers? An Impact Evaluation of a Non-experimental Project in Peru. *Journal of Agricultural Economics , 62* (2), 449-473.

Dillon, A. (2011). Do Diffrerences in the Scale of Irrigation Projects Generate Different Impacts in Poverty and production? *Journal of Agricultural Economics , 62* (2).

Heckman, J. J., & Smith, J. A. (2004). The Determinants of Participation in a Social Program: Evidence from a Prototypical Job Training Program. *Journal of Labor Economics , 22* (2), 243-298.

Heckman, J. J., & Smith, J. A. (1999). The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme Implications for Simple Programme Evaluation Strategies. *The Economic Journal , 109*, 313-348.

Imbens, G., & Wooldridge, J. (2008). *Recent Developments in the Econometrics of Program Evaluation.* Cambridge, Massachusetts: National Bureau of Economic Research.

Instituto Nacional de Información de Desarrollo. (2003). Censo Nacional Agropecuario. Managua, Nicaragua.

Instituto Nacional de Información de Desarrollo. (2005). Censo Nacional de Poblacion. Managua, Nicaragua.

Jyotsna, J., & Ravallion, M. (2003). Estimating the Benefit Incidence of an Antipoverty Program by Propensity Score Matching . *Journal of Business & Economic Statistics , 21* (1), 19-30.

Klinger, B., & Schündeln, M. (2011). Can Entrepreneurial Activity be Taught? Quasi-Experimental Evidence from Central America. *World Development , 39* (9), 1592-1610 .

Lai, F., Sadoulet, E., & de Janvry, A. (2009). The contributions to School quality and teacher qualifications to school performance: Evidence from a natural experiment in Beijing Middle schools. *Journal of Human Resources, forthcoming* .

Lalive, R. (2008). How do extended benefits affect unemployment duration? A discontinuity design approach. *Journal of Econometrics , 142* (2), 785-806.

Mendola, M. (2007). Agricultural Technology adoption and poverty reduction: A propensity-score matching analysis for rural Bangladesh. *Food Policy , 32*, 372-393.

Nitlapan. (2001). Tipologia Nacional de Productores y Zonificacion Economica 2001. Managua, Nicaragua: Universidad Centroamericana.

Ravallion, M. (2005). *Evaluating Anti-Poverty Programs.* World Bank Policy Research Working Paper.

Rubin, D. (1974). Estimated Causal Effects of Treatments in Randomized and Non-Randomized Studies. *Journal of Educational Psichology* (66), 688-701.

Schultz, T. P. (2004). School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of Development Economics , 74* (1), 199-250.

Skoufias, E., & Buddermeyer, H. (2004). *An evaluation of the performance of discontinuity design on PROGRESA.* World Bank Policy Reseach Working Paper.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, Massachussets: The MIT Press.

Appendix 1.

From Dehejia and Wahba, 2002, p. 161:

A simple Algorithm for Estimating the Propensity Score:

1. Start with a parsimonious logit specification to estimate the score.
2. Sort the data according to estimated propensity score (ranking from lowest to highest).
3. Stratify all observations such that estimated propensity scores within a stratum for treated and comparison units are close (no significant differences); for example, start by dividing observations intro strata if equal range (0-0.2, 0.2-0.4, …, 0.8-1)
4. Statistical test: for all covariates, differences in means across treated and comparison units within each stratum are not statistically different from zero.
    a. If covariates are balanced between treated and comparison observations for all strata, stop.
    b. If covariates are not balance for some stratum, divide the stratum into finer strata and reevaluate.
    c. If a covariate is not balanced for many strata, modify the logit by adding interaction terms and/or higher-order terms of the covariates and reevaluate.

Table 1. Summary statistics for the variables included in the propensity score regression, for A4N participants and non-participants, Nicaragua, 2009

	Treatment N=297		Comparison N=350			p value t-test for equal means
Variable	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Bias**	
Cultivated land in Mz	3.20	3.35	3.44	3.32	-5.03	0.37
Steep slope(0,1)	0.34	0.47	0.31	0.46	3.81	0.49
Inadequate services(0,1)	0.68	0.47	0.77	0.42	-14.55	0.01
Inadequate housing(0,1)	0.88	0.33	0.83	0.38	9.33	0.09
Electricity(0,1)	0.61	0.49	0.67	0.47	-9.63	0.09
The household experienced food scarcity(0,1)	0.40	0.49	0.31	0.46	13.27	0.02
Female head(0,1)	*0.20*	*0.40*	*0.07*	*0.26*	*26.75*	*0.00*
Number of children less than 5 years old	0.53	0.73	0.51	0.71	1.93	0.73
Age of household head	49.06	15.43	48.04	14.68	4.76	0.39
Head works in own farm(0,1)	0.77	0.42	0.84	0.37	-12.36	0.03
Head years of education	2.88	2.70	3.19	2.72	-7.92	0.16
Number of people in the household	5.16	2.28	5.39	2.28	-7.32	0.19
Number of people per room	3.81	1.99	3.85	2.09	-1.52	0.79
Value of livestock in 1000s Cordobas*	7.59	11.27	9.89	15.21	-12.14	0.03
Value of farm infrastructure in 1000s Cordobas*	0.59	1.70	1.13	3.19	-15.06	0.01
Value of farm equipment 1000s Cordobas*	2.68	6.49	3.67	7.62	-9.89	0.07
Television(0,1)	0.42	0.49	0.54	0.50	-16.83	0.00
Cellphone(0,1)	0.37	0.48	0.43	0.50	-8.40	0.13
Population	632.79	479.96	637.04	593.62	-0.56	0.92
Distance to closest market Km/10	*1.40*	*0.68*	*1.95*	*1.59*	*-31.78*	*0.00*
Distance to closest road Km/10	0.95	0.94	0.94	0.90	1.23	0.83
Health facility(0,1)	0.20	0.40	0.32	0.47	-19.10	0.00
Proportion farms which produce basic grains(2003)**	0.84	0.24	0.87	0.10	-10.87	0.06
*Proportion farms size less than 10 Mz (2003)** *	*0.59*	*0.23*	*0.51*	*0.20*	*25.49*	*0.00*

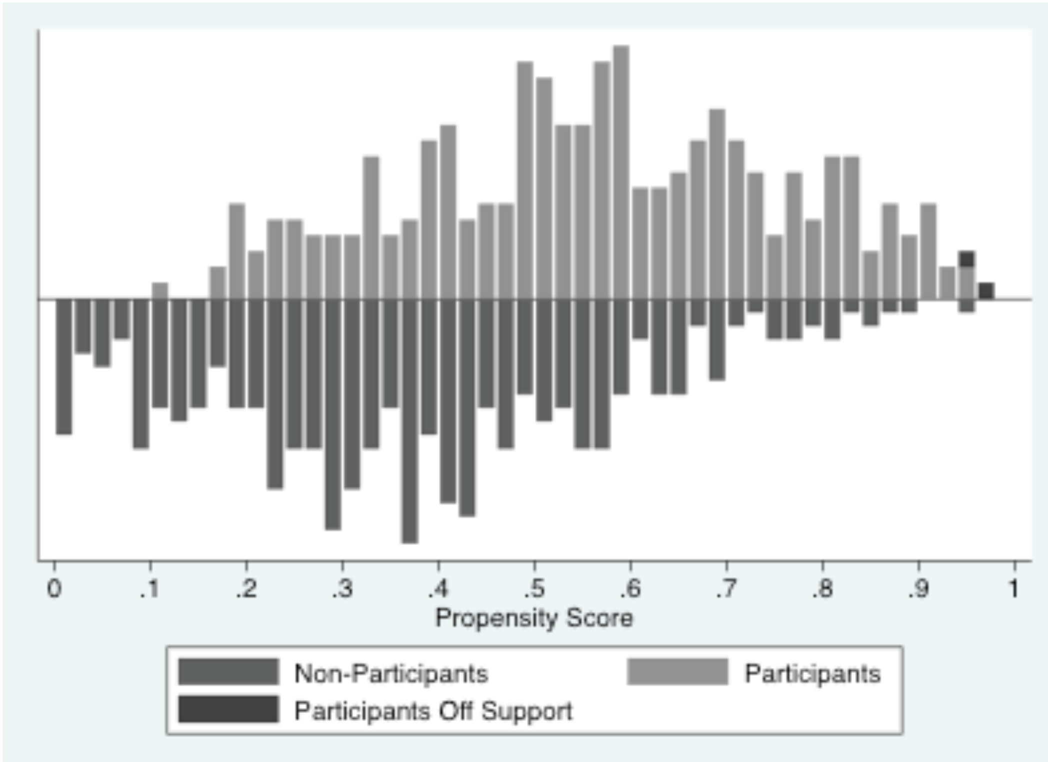
*Exchange rate in July 2009 was US$1.00 = C21.35

**Source: Instituto Nacional de Información de Desarrollo. (2003). Censo Nacional Agropecuario. Managua, Nicaragua.

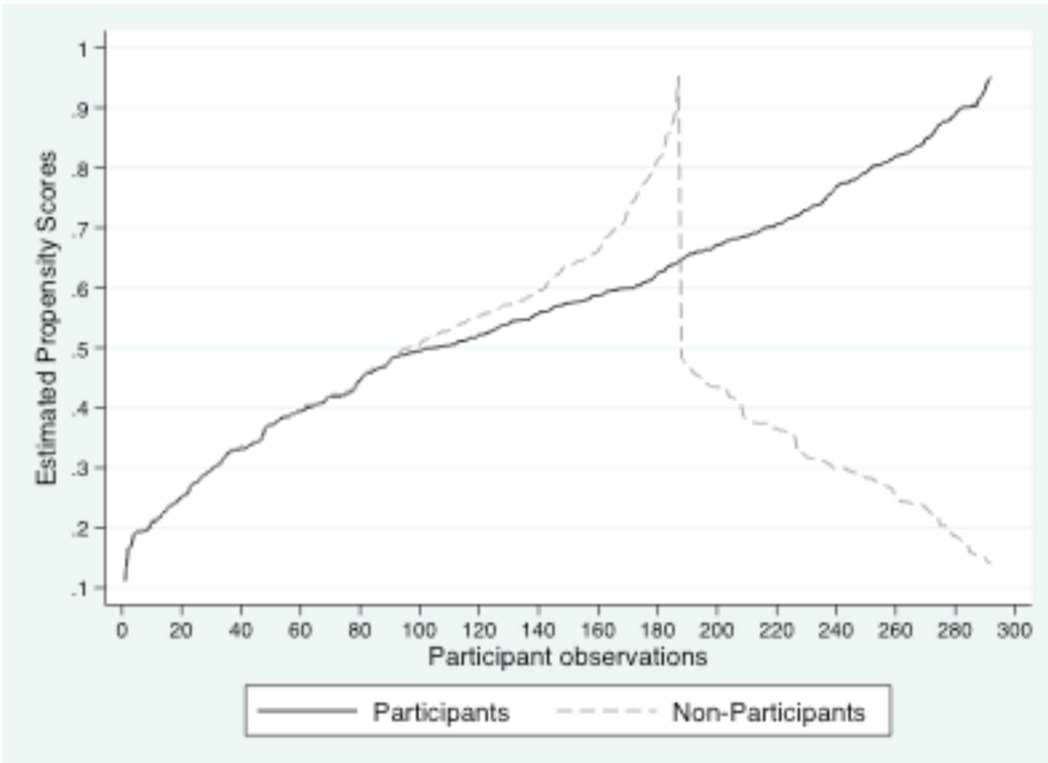
Table 2. A4N program participation propensity score logit model Nicaragua, 2009.

Dependent Variable: Participation in A4N (0, 1)		All households N=635		
	Covariates	**Coefficients**	**Margeff**	**p-value**
program eligibility criteria	Cultivated land in Mz	0.023	0.005	0.468
	Steep slope(0,1)	0.257	0.051	0.192
	Inadequate services(0,1)	-0.412	-0.083	0.052
	Inadequate housing(0,1)	0.048	0.010	0.864
	Electricity(0,1)	0.188	0.037	0.438
	The household experienced food scarcity(0,1)	0.347	0.070	0.075
	Female head(0,1)	1.563	0.303	0.000
	Number of children less than 5 years old	0.059	0.012	0.681
household characteristics	Age of household head	-0.001	0.000	0.913
	Head works in own farm(0,1)	0.485	0.094	0.087
	Head years of education	-0.003	-0.001	0.929
	Number of people in the household	-0.014	-0.003	0.802
	Number of people per room	-0.041	-0.008	0.498
agricultural assets	Value of livestock in 1000s Cordobas	-0.013	-0.003	0.084
	Value of farm infrastructure 1000s Cordobas	-0.082	-0.016	0.060
	Value of farm equipment 1000s Cordobas	-0.005	-0.001	0.753
other assets	Television(0,1)	-0.392	-0.079	0.091
	Cellphone(0,1)	-0.016	-0.003	0.934
village effects	Population	0.000	0.000	0.105
	Distance to closest market Km/10	-0.645	-0.129	0.000
	Log distance to closest road Km/10	0.183	0.037	0.092
	Health facility(0,1)	-0.999	-0.197	0.000
	Proportion farms which produce basic grains 03	-0.220	-0.044	0.711
	Proportion farms size less than 10 Mz 03	2.423	0.485	0.000
	Constant	-0.445		
	Log likelihood	-369.351		

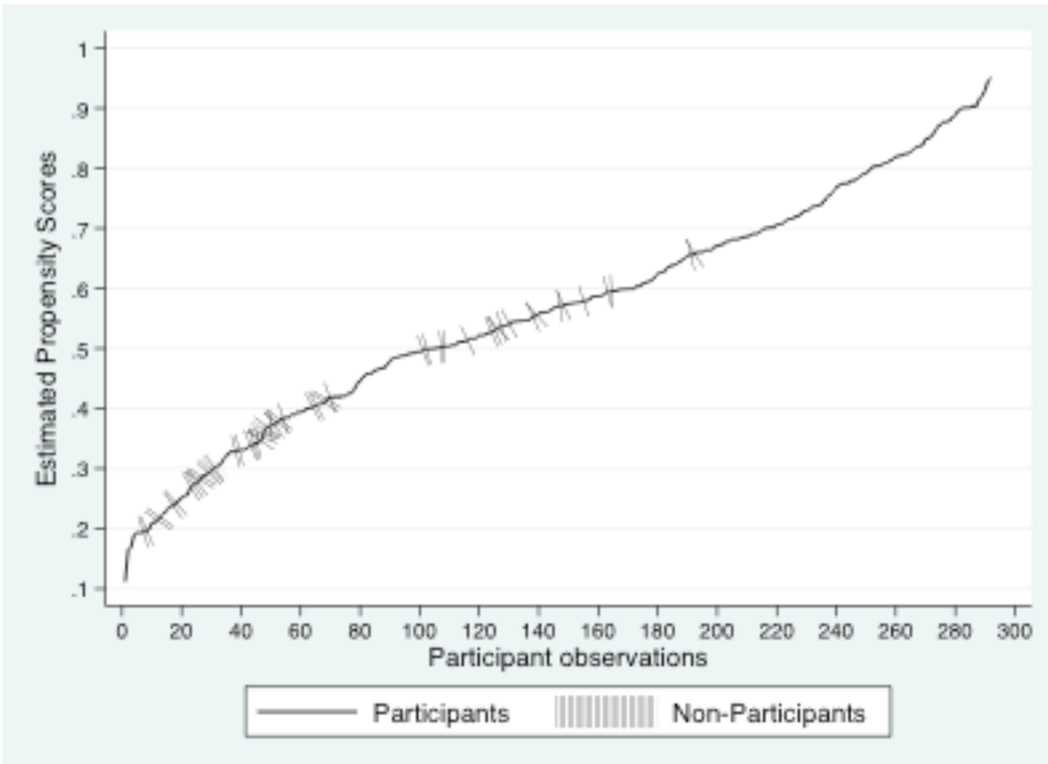
Graph 1. Estimated Propensity Scores, logit model results



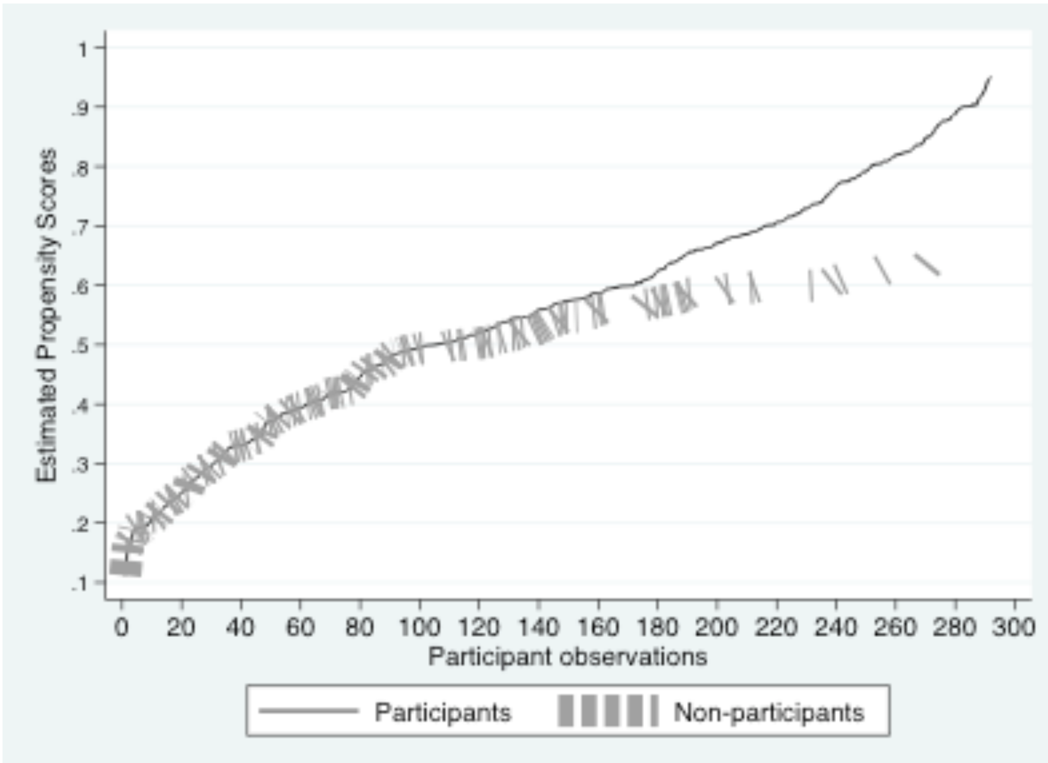
Graph 2. Propensity scores for participants and non-participants after matching without replacement.



Graph 3. Propensity score for participants and non-participants after caliper=0.001 matching.



Graph4. Propensity score for participants and non-participants after caliper=0.01 matching.



Graph 5. Propensity score for participants and non-participants after nearest neighbor matching with replacement.

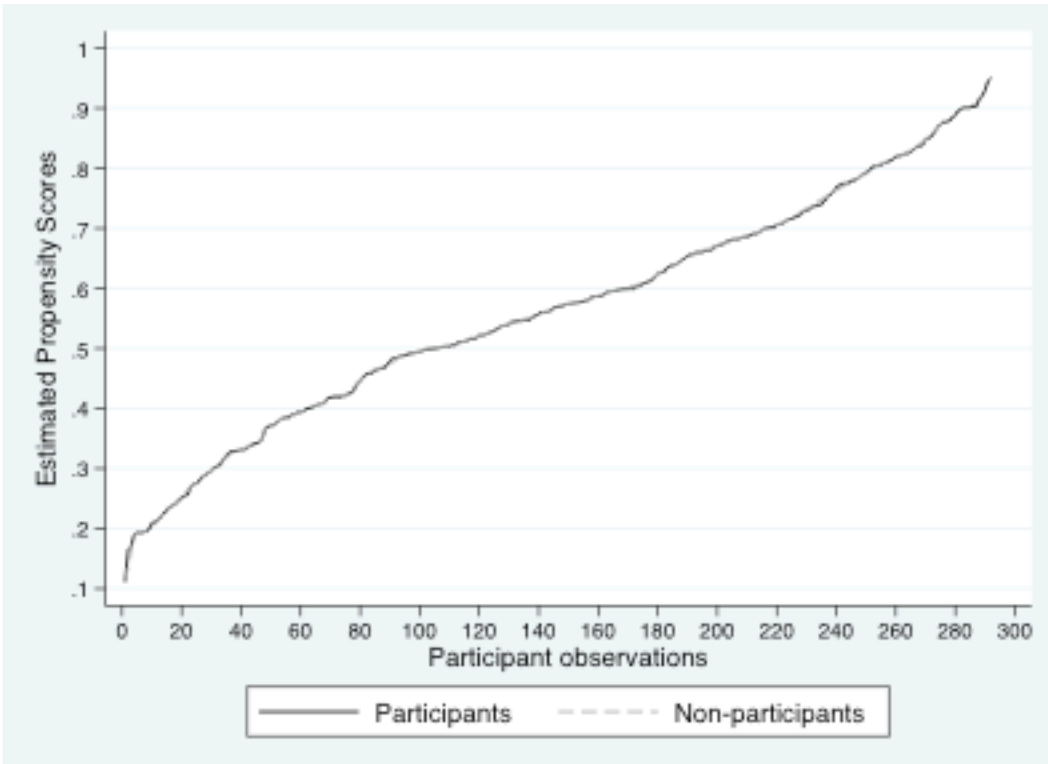


Table 3. Balancing tests after conducting Nearest Neighbor Matching

N=294		Sample average			% reduc bias
Variable	**Sample**	**Treated**	**Non Treat**	**Bias**	
Cultivated land in Mz	Unmatched	3.21	3.41	-6.20	
	Matched	3.21	3.20	0.20	97.50
Steep slope(0,1)	Unmatched	0.34	0.32	4.30	
	Matched	0.33	0.30	5.80	-36.90
Inadequate services(0,1)	Unmatched	0.67	0.76	-19.80	
	Matched	0.67	0.74	-14.50	26.90
Inadequate housing(0,1)	Unmatched	0.87	0.82	14.00	
	Matched	0.87	0.89	-3.80	72.70
Electricity(0,1)	Unmatched	0.60	0.69	-17.60	
	Matched	0.60	0.49	23.70	-34.30
The household experienced food scarcity(0,1)	Unmatched	0.40	0.31	18.30	
	Matched	0.40	0.45	-10.70	41.30
Female head(0,1)	Unmatched	0.20	0.07	37.90	
	Matched	0.19	0.21	-4.10	89.20
Number of children less than 5 years old	Unmatched	0.52	0.50	3.10	
	Matched	0.52	0.49	5.30	-68.60
Age of household head	Unmatched	49.12	47.84	8.50	
	Matched	49.15	47.55	10.60	-24.30
Head works in own farm(0,1)	Unmatched	0.77	0.84	-16.90	
	Matched	0.77	0.82	-11.20	33.60
Head years of education	Unmatched	2.88	3.18	-11.10	
	Matched	2.88	2.95	-2.70	75.90
Number of people in the household	*Unmatched*	*5.15*	*5.38*	*-10.20*	
	Matched	*5.17*	*4.73*	*19.20*	*-88.20*
Number of people per room	Unmatched	3.81	3.87	-2.50	
	Matched	3.83	3.77	2.90	-17.10
Value of livestock in 1000s Cordobas	Unmatched	7.63	9.89	-17.00	
	Matched	7.68	6.35	10.00	41.10
Value of farm infrastructure 1000s Cordobas	Unmatched	0.59	1.15	-21.90	
	Matched	0.59	0.50	3.40	84.30
Value of farm equipment 1000s Cordobas	Unmatched	2.70	3.74	-14.60	
	Matched	2.72	2.43	4.00	72.30
Television(0,1)	Unmatched	0.42	0.55	-25.50	
	Matched	0.42	0.36	13.10	48.60
Cellphone(0,1)	Unmatched	0.37	0.43	-13.00	
	Matched	0.36	0.32	8.40	35.50
Population	*Unmatched*	*635.89*	*650.29*	*-2.70*	
	Matched	*637.08*	*562.24*	*13.80*	*-419.60*
Distance to closest market Km/10	Unmatched	1.41	1.99	-47.00	
	Matched	1.42	1.30	9.20	80.40
Distance to closest road Km/10	*Unmatched*	*0.95*	*0.94*	*0.90*	
	Matched	*0.95*	*0.88*	*7.30*	*-672.10*
Health facility(0,1)	Unmatched	0.20	0.32	-28.80	
	Matched	0.20	0.15	11.80	59.00
Proportion farms which produce basic grains	Unmatched	0.85	0.87	-13.20	
	Matched	0.85	0.88	-15.90	-20.50
Proportion farms size less than 10 Mz	Unmatched	0.59	0.52	33.50	
	Matched	0.59	0.52	29.10	13.10