Impact Assessment with Opt-in Treatments:

Evidence from a rural development project in Nicaragua

M. Alexandra Peralta and Scott M. Swinton

Department of Agricultural, Food, and Resource Economics

Michigan State University

June 3, 2013

Selected Paper prepared for presentation at the Agricultural & Applied Economics Association's 2013 AAEA & CAES Joint Annual Meeting, Washington, DC, August 4-6, 2013.

Acknowledgements

The authors gratefully acknowledge funding by the Howard G. Buffett Foundation through the Catholic Relief Services Central America office. We also would like to thank Andrew Dillon and Songqin Jin for comments on the methods used in this paper and to the collaboration from Nitlapan at the Universidad Centroamericana during the data collection and data cleaning conducted for this research. We also like to thank Catholic Relief Services (CRS) office in Nicaragua, Caritas and the Foundation for Research and Rural Development (FIDER) for their collaboration during the fieldwork stage of this research.

Abstract

In this study we conduct an impact evaluation of a complex rural development project in Central America with multiple treatments taking place simultaneously, purposive program placement and project participant freedom to opt in to project interventions. For this purpose we use propensity score matching difference-in-differences estimation, and compare results of this method with weighted propensity score regression and simple difference-in-differences estimation. We find short term project impacts in household savings, in participation in groups and associations, and in reduction of stored grain losses. However, we find no project impacts in long-term outcomes associated with increased agricultural income or household asset accumulation. These results are not surprising, since the project evaluation was conducted two years into a three-year project, before beneficiaries had realized its full benefits. Our study calls attention to the need of more research on linking short term to long-term impacts and on longer term strategies to evaluate impacts of agricultural technology.

JEL categories: O10, O13, Q1

Introduction

The needs of development project implementers and donors for rigorous measurement of impacts have triggered an explosion of recent studies. The most rigorous of these have measured impacts of well-defined treatments (Becerril & Abdulai, 2010; Buddelmeyer & Skoufias, 2004; Chowa & Elliott III, 2011; Dillon, 2011a, 2011b; Edmonds, Mammen, & Miller, 2005; Gilligan & Hoddinott, 2007; Imai & Azam, 2012; Islam, 2011; Mendola, 2007; Solís, Bravo-Ureta, & Quiroga, 2009; Towe & Tra, 2013), sometimes using experimental design (Ashraf, Giné, & Karlan, 2009; E. Duflo, Kremer, & Robinson, 2009; Duflo, 2004; Duflo, Kremer, & Robinson, 2008; Kremer, 2003; Lai, Sadoulet, & Janvry, 2011). The literature on empirical impact evaluation of agricultural development projects that involve multiple interventions without experimental design, is scant (Wanjala & Muradian, 2013).

Techniques for evaluating multiple treatments are available by means of using multiple dummy variables to determine the impact of each treatment(Wooldridge, 2010), or using matching for different types of interventions (Cuong, 2009; Plesca & Smith, 2007). Data collection requires a sample size that allows for meaningful inferences about the effects of individual interventions as well as some combinations. Yet when project participants self-select into different program interventions, it is difficult *ex ante* to forecast levels of participation

We investigate the overall impact on agricultural income and household wealth of a pro-poor rural development program in Nicaragua that involved many possible interventions over a two-year period. Project beneficiaries were not randomly assigned, so selection bias was a concern for impact evaluation.

Within designated villages, project treatments were made available to all households that met a set of need criteria, and the same treatment packages were not offered in all the treatment villages. Due to heterogeneous availability of treatments and the freedom of beneficiaries to opt-in to individual treatments, the application of multi-treatment impact evaluation techniques is inappropriate, because participation in different treatments is an intended outcome of this program.

The project did not lend itself to an experimental design for impact evaluation. When experimental designs are not feasible, program evaluation can be designed using quasi-experimental (QE) methods. Propensity score matching difference in differences (PSM-DID) is a QE method that can be used to correct for selection bias on observables and to control for time invariant unobservable characteristics (A. Smith & E. Todd, 2005). We use PSM-DID and consider as treatment exposure involvement in any of the opt-in treatments offered by the program. We compare the estimates for different matching methods with estimates of a propensity score weighted regression and the simple difference in difference estimator. The comparisons are first applied to the two major long-term goals of the project, increased agricultural income and increased household asset ownership. Next, we examine short-term outcomes that are expected to contribute to reaching the long-term goals.

Although we find no evidence if gains in agricultural incomes and household asset holdings during the first two years of the project, we do find short term outcomes that are likely to lead to these desired long-term impacts. The rest of this document provides details and is organized as follows: First we describe the Agriculture for Basic Needs project, then we introduce the problem of impact evaluation. After that we present the methods used to estimate program impacts, we present our results and finally we discuss our results and conclude.

The program to be evaluated

The Agriculture for Basic Needs (A4N) program was an integrated rural development project implemented in four Central American countries during 2009-2012. It was managed by Catholic Relief Services (CRS) and implemented in the field by its partners Caritas and the Foundation for Research and Rural Development (FIDER). The A4N program aimed to provide farmers with a set of five skills for achieving sustainable farm production and increased agricultural income: (1) group management; (2) saving and lending; (3) marketing; (4) basic experimentation and innovation skills for accessing new technology; (5) agricultural production and natural resource management skills.

To accomplish these objectives, the program promoted conservation agriculture and nutritious crops, improved crop varieties, micro-livestock management, integrated pest management and practices to reduce post-harvest crop loss. Other program interventions included saving and lending groups, post-harvest processing, expanded participation in markets, and promotion of farmer innovation groups. The program also provided beneficiaries with agricultural assets, such as metallic silos, material for building agricultural infrastructures, water storage infrastructure, and small animals, such as poultry, pigs and goats.

The A4N project first targeted villages considered poor. These villages tend to be located in areas of natural resource degradation with relatively high vulnerability to natural disasters. Within these villages, in order to be eligible to participate in the A4N program, households were expected to be characterized by 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.
- Household experiences hunger during some period of the year.
- Household head is female.
- Household includes children younger than five years old.

In spite of these formal eligibility criteria, the A4N's village-level managers found it difficult to exclude participation of village members. So the program allowed some technically ineligible individuals to participate, in the hope that they would help to spread A4N interventions during and after program implementation.

Within each A4N village, promoters trained by the A4N program formed groups of 15 to 20 eligible farmers and offered them the opportunity to participate in selected program interventions. Not all interventions were offered to every group of farmers. Due to limited project resources, interventions were prioritized and offered where the need was considered greatest. Hence, there was not a standard set of interventions offered in each village.

In the end, participation in specific A4N interventions was driven by two very different selection processes. Hence, the impact evaluation must account for potential selection bias of two distinct forms. First, official eligibility criteria that were not evenly enforced, so households permitted to participate in the A4N program vary on observable traits. Second, the self-selection of individuals into specific A4N activities means that unobservable traits may also affect participation assignments.

The problem of program evaluation

We approach program evaluation though Rubin's potential outcome framework (Rubin, 1974). The objective of program evaluation is to determine how the intervention or applied treatment (in the contest of A4N, treatment means opportunity to participate on a given intervention or set of interventions) has an effect on a given outcome, evaluating the treatment effect against a counterfactual. Participation of individual *i* in the project is referred to as a "treatment" given by w_i=1, so w_i=0 if the individual has not been exposed to treatment. The observed outcome for individual *i* is $w_i=w_iy_{1i}+(1-w_i)\ y_{0i}$, which means that the outcome for an individual who participates is y_{1i} and if she does not participate the outcome is y_{0i}. The treatment effect of the program intervention is $\tau_i=\Delta y_i= y_{1i} - y_{0i}$. But the resulting outcome attributable to a program cannot be observed in an individual participating and not participating in the program at the same time. Therefore, the problem of program evaluation is a problem of missing data, and the program effect cannot be calculated for the same individual, but instead requires constructing a counterfactual to calculate average treatment effects across individuals in (a sample from) the population.

The parameters of interest are the average treatment effect on the population, ATE, and the average treatment effect on the treated, ATT. The ATE is the difference between the expectation of the outcome with and without the program, for individuals given a vector of characteristics **x** is:

$$\mathrm{ATE} = E(\boldsymbol{\tau}(\mathbf{x})) = E[y_1|\mathbf{x}] - E[y_0|\mathbf{x}] \qquad (1)$$

ATE measures the effect of the treatment on both participants and non-participants. But for program evaluation purposes, the parameter of interest is the average treatment effect on the treated, ATT, which is the expected value of the outcome for those who participated in the program, conditional on the individual characteristics that determine program participation, **x**:

$$\mathrm{ATT} = E(\boldsymbol{\tau}(\mathbf{x}) \mid w=1) = E(y_1|\mathbf{x}, w=1) - E(y_0|\mathbf{x}, w=1) \qquad (2)$$

As already mentioned, $E(y_0|\mathbf{x}, w=1)$, the expected outcome of the treated if they were not exposed to the treatment, cannot be observed directly, whereas we can observe $E(y_0|\mathbf{x}, w=0)$, the expected outcome of the untreated, given that they were not exposed to the treatment. We can define:

$$E(y_1|\mathbf{x}, w=1) - E(y_0|\mathbf{x}, w=0) = \mathrm{ATT} - E(y_0|\mathbf{x}, w=1) + E(y_0|\mathbf{x}, w=0) \quad (3)$$

Therefore,

$$\mathrm{ATT}=E(y_1|\mathbf{x},w=1) - E(y_0|\mathbf{x},w=0) + E(y_0|\mathbf{x}, w=1) - E(y_0|\mathbf{x}, w=0) \qquad (4)$$

Subject to the assumption of no selection bias, in the absence of the program, those who participated in the program would have had equal outcomes to those who did not,

$$E(y_0|\mathbf{x}, w=1) - E(y_0|\mathbf{x}, w=0) = 0 \qquad (5)$$

However, if program selection has not been made randomly and conducted conditional on a given set of individual characteristics, then selection bias occurs, and individuals exposed to the treatment will systematically differ from those not exposed to the treatment. Hence, program impact appears as a consequence of these differences and program intervention, distorting the measure of the benefits from the program.

Selection bias is a consequence of the difference in the covariates **x**, between participants and non-participants. Some covariate differences can be observed by the researcher, such as housing characteristics, land allocated to agricultural production, and location on steep slopes. These characteristics are defined by the A4N program, and they determined eligibility for program participation. Other covariate differences are not observed by the researcher and can be assumed not to change over time, including such individual characteristics as motivation, cognitive learning ability, and attitudes towards innovation.

Two assumptions about program assignment mechanisms underlie the two major classes of quasi-experimental methods to correct for selection bias used when conducting program evaluation (Imbens and Wooldridge, 2008). The first is that expected values of outcomes, y conditional on covariates, $\boldsymbol{x}$, are independent of program assignment w. This is known as the conditional independence assumption (CIA), unconfoundedness or selection on observables. The second is that unobserved characteristics that affect selection are time invariant. This is referred to as the selection on un-observables. The challenge is to correct for these two sources of selection bias when conducting impact evaluation. In this paper we use propensity score matching difference in differences (PSM-DID) (A. Smith & E. Todd, 2005; J. J. Heckman, Ichimura, & Todd, 1997) to control for these two sources of bias. We compare PSM-DID estimates with simple propensity score weighting regression estimates. We also compare results with the simple DID estimator which provides the estimated mean of the difference between treatment and comparison groups.

Determining program impacts

The main assumptions for estimating the impact of the program are for constructing the counterfactual using propensity score matching are:

1) Unconfoundedness:

$$y_0, y_1 \perp w \mid \mathbf{x} \qquad (6)$$

where y_0 is the outcome for non-participants and y_1 is the outcome for participants, w is participation and **x** represents a set of variables that may influence participation. The sign $\perp$, denoting orthogonality, means that program outcomes are independent of program participation, conditional on **x**.

2) Mathematically, there is common support (overlap) between the probability distributions of program participants and non-participants (Caliendo & Kopeinig, 2008; Imbens & Wooldridge, 2008; Ravallion, 2008) (Eq. 7):

$$0 < \Pr(w=1 \mid \mathbf{x}) < 1 \qquad (7)$$

Propensity score matching (PSM) 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. To estimate the propensity score (PS), we use a logit model. The expected probability of program participation is

$$\Pr(w=1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta}) \tag{8}$$

Here, $0 < G(\mathbf{x}\boldsymbol{\beta}) < 1$, G refers to the logistic probability distribution function, where **x** represents a vector of explanatory variables and $\boldsymbol{\beta}$ is a parameter vector. In this case, the explanatory variables refer to program eligibility criteria, household characteristics, village characteristics, farm characteristics, and wealth. Including a rich set of variables that determine both participation in the project and pretreatment outcomes reduces bias in estimates (J. Heckman, Ichimura, Smith, & Todd, 1998) .

We apply Dehejia and Wahba's (2002) suggested algorithm for estimating the propensity score 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, by plotting the estimated probability distributions of the treated and comparison groups. Overlap is crucial to be able to implement propensity score based methods, the failure of this assumption is a major source of bias in impact evaluation estimates, basically because the counterfactual is not similar to the treatment group to conduct valid comparison. In addition we trim the observations with an estimated PS above 0.90 and below 0.10 to improve overlap. With this trimmed sample we re-estimate the PS and conduct matching.

We conduct balancing tests to check for the similarity of the marginal distributions of the covariates used to estimate the PS. The tests aim to determine whether the matching procedures have served the purpose of making participants and non-participant groups more similar. Covariates are compared via a measure of standardized bias or normalized differences in means defined as follows (Caliendo & Kopeinig, 2008; Wooldridge, Jeffrey, 2010):

$$\%bias = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_1^2 + s_0^2}} * 100, \qquad (9)$$

where $\bar{x}_1$ and $\bar{x}_0$ are the sample averages of variable j for the groups of participants (1) and non-participants (0), and s_1 and s_0 are estimated standard errors for variable j for participants and non-participants. An absolute value of percent bias above 25 is typically interpreted to mean that the two groups are not similar by those covariates (Wooldridge, Jeffrey, 2010). We also conducted two-sample t-tests for equal means. The advantage of the standardized difference of means with respect to the t-test, is that the former does not depend on the sample size. We compare these bias measures before and after matching.

To estimate the ATT we match participants to non-participants using the estimated propensity scores using four different matching methods. We use two kernel estimators (Epanechnikov and normal or Guassian with bandwith 0.06), local linear regression (tricube kernel and bandwith 0.8), and nearest neighbor (NN) with replacement (five nearest neighbors). Bootstrapped standard errors are calculated for all four matching estimates to compare the sensitivity of estimates to different matching methods (though it is disputed whether bootstrapping is valid with NN (Abadie & Imbens, 2008)).

Kernel and local linear regression (LLR) are non parametric matching methods. Kernel matching uses a weighted average of all the observations in the comparison group to construct the counterfactual outcome for each treated observation, whereas LLR estimates a nonparametric locally weighted regression using for comparison observations in the neighborhood of the treated ones (Smith & Todd, 2005; Khandker, Koolwan & Samad, 2010). The weights depend on the type of kernel function chosen. An advantage of kernel and LLR matching is that it reduces the variance of the estimates by using more information. However, a problem arises if there is insufficient overlap between the distributions of the treated and comparison groups, as poor matches may be used for comparison, resulting in biased estimates.

Nearest neighbor matching with replacement consists of matching each treated observation with one or more having the nearest value of estimated propensity score, so a control observation may be used more than once. When using more than one NN, the estimator constructs a counterfactual mean with the closest comparison observations. Matching with replacement using more than one NN reduces bias in the estimates but increases its variance (Smith & Todd, 2005; Caliendo & Kopeinig, 2008). Unlike kernel and LLR methods, NN matched observations all have the same weight. NN matching tends to works best with a large sample of comparison observations to match treated ones with.

Propensity score matching assumes that after controlling for observable characteristics, outcomes are mean independent of participation in the program. But it is likely that there are systematic differences in outcomes for participants and non-participants due to unobservable characteristics, known as bias on unobservables. Assuming that unobserved heterogeneity is time invariant and uncorrelated with treatment assignment, we can control for this source of bias using the difference in difference matching estimator, defined by Smith and Todd (2005) as follows:

$$\hat{\tau}_{ATTPSM-DID} = \frac{1}{N_1} \sum_{i \in I_1 \cap Sp} \left\{ \left(y_{1t'j} - y_{1tj}\right) - \sum_{j \in I_0 \cap S_p} \varphi(i,j)\left(y_{0t'j} - y_{0tj}\right) \right\} \qquad (10)$$

Where the subscripts *1* and *0* refer to treated and untreated respectively, the subscript *i* refers to the treated observations that are in the common support S_p, *j* refers to the untreated observation in the common support S_p, *t* refers to previous period and *t'* to current period, N_1 corresponds to the number of treated observations, and *φ(i,j)* refers to the weight, which depends on the matching method. By taking the difference between the pretreatment and after treatment outcomes we control for unobserved time invariant characteristics. Smith and Todd (2005) compared longitudinal methods with cross-sectional PSM methods and found that PSM-DID perform best in correcting for selection bias, when compared with experimental results. By using the PSM-DID estimator we control for both observable sources of bias by building our comparison groups using PSM and time invariant characteristics, by taking the difference of outcomes before and after treatment.

Following Mu and van de Walle (2011) and Chen, Mu and Ravallion (2009), we compare matching estimates with propensity score weighted regression and the simple difference in difference estimates to test the robustness of our PSM-DID estimates. The PS weighted regression estimator is based on Hirano, Imbens and Ridder (2003) and Hirano and Imbens (2001):

$$\hat{\tau}_{ATTwgtreg} = \frac{1}{N}\sum_{i=1}^{N}\left\{(y_{it'} - y_{it})w_i - \left[(y_{it'} - y_{it})(1 - w_i)\left(\frac{\hat{\Pr}(x_i)}{1 - \hat{\Pr}(x_i)}\right)\right]\right\} \qquad (11)$$

where w_{i} takes the value of 0 if untreated or 1 if treated by program participation. The weights are 1 for participants (treated observations) and $\frac{\hat{\Pr}(x_i)}{[1 - \hat{\Pr}(x_i)]}$ for non-participants. Hirano and Imbens (2001) propose this modification of weights to estimate the ATT. These weights, estimated using flexible logit, can be used to construct an efficient PSM-DID estimator of treatment effects (Hirano et al., 2003; Imbens & Wooldridge, 2008). We implement this weighted estimator of ATT as $\Delta y_i = \alpha + \tau w_i + \varepsilon_i$. The regression without weights corresponds to the simple difference in difference estimator, which is included only for comparison purposes.

The data

The dataset was based on two-stage sampling of treatment and non-treatment villages, where “treatment” refers to participation in the Agriculture for Basic Needs (A4N) project. We randomly selected villages from the list of beneficiary villages, and chose similar non-participant villages using census data. The sampled villages were selected according to the population weights of each of the municipalities where the project intervened. 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, 2011). From each village we randomly selected 10 households in the participant villages and 10 to 15 households in the non-participant villages, depending on village size. 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 baseline survey measured livelihoods and income for the agricultural year 2008-09, before project implementation, and a follow up survey did the same for the agricultural year 2010-11, the second year after project implementation. The survey was conducted in the departments of Estelí, Jinotega and Matagalpa, located in the northeast of Nicaragua. The final balanced panel includes 578 households, 284 in participant villages and 294 in non-participant villages. 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

We report results from the series of analytical steps described above. The propensity scores were estimated first using a logit model with the data from 272 treated and 282 non-treated households, due to observations that were dropped because of outliers, missing data and the trimming of observations with PS above 0.90 and below 0.10 (Imbens & Wooldridge, 2008; Wooldridge, Jeffrey, 2010). Upon application of Dehejia and Wahba's (2002) algorithm for estimating the propensity scores, it was determined that no interaction terms and higher level terms were justified to improve the estimation, so the logit model was estimated with all covariates entering linearly.

The logit model estimates the probability of program participation (Table 1). Focusing on variables that are statistically significant, the A4N households were more likely to be female-headed and to have lower value of farm infrastructure but also less inadequate services as defined by the basic needs index (housing lacking piped water and where a toilet is missing). They were situated in villages closer to markets but with fewer large farms and less likely to have a health facility. These variables reflect differences between treated and comparison observations.

The predicted probability distributions of selection into the A4N participant and non-participant groups are presented in Figure 1. Clearly, the two distributions are not mirror images. The non-participant distribution contains more observations with propensity scores below 0.6, and a disproportionate number of observations with propensity scores below 0.4. In spite of this, overlap does not seem to be a problem, and we have comparison observations to match treatment ones.

Table 1 here

Figure 1 here

We conduct matching of participant and non-participant observations according to the values of the propensity score using STATA's psmatch2 procedure. The results for the balancing tests after matching are provided in Table 2. For all covariates in the PS logit model, the table reports the sample average for treated and comparison observations before and after matching, the standardized difference in means (percentage bias), and the p-values for two-sample t test of differences in means. Matching clearly improved overlap between the marginal distributions of the covariates. As evidence, the percentage bias decreases for all the covariates except the number of children under 5 years old, and the value of the percentage bias goes from a maximum absolute value of 385% before matching to a maximum after matching of 18%, well below the benchmark of 25% for covariates balance (Imbens & Wooldridge, 2008).

Table 2 here

To estimate program impacts using PSM-DID we conducted two different types of kernel matching, normal and Epanechnikov, local linear regression with the tricube kernel, and nearest neighbor with replacement, using five neighbors. We estimated program impact using the difference in the outcome variables before and after the project as dependent variable. The two targeted outcomes are gains in agricultural income and household wealth. Several proxy variables were available to measure these project impacts. For agricultural incomes, they include: farm gross margins, total value of agricultural sales, total value of production of main crops (maize, bean, sorghum and millet), bean production, and maize production. For household wealth, they include tropical livestock units and the contribution of agricultural assets to a) the value of main crops, b) the total value of sales and c) total gross margin.

The contribution of agricultural assets to income offers a means of estimating how changes in specific assets will affect income, given that asset holdings are far easier to observe than asset values. Following Adato, Carter and May (2006) and Wanjala and Muradian (2013), we run regressions of the pretreatment agricultural income proxies on the pretreatment endowments of household head education, cultivated land and a set of agricultural equipment. We used the estimated coefficients to predict the contribution of assets farm gross margins, total value of agricultural sales, total value of production of main crops before and after treatment.

The estimates of PSM-DID, weighted PS regression and the simple DID estimator show that the project did not have an impact as measured by any of these outcomes (Table 3). This result is robust across the different methods, particularly among the different matching estimates. The magnitudes of the estimates are similar, although there are some differences among estimates using only weighted PS regression or simple DID. The sole inconsistent results came from use of the simple DID estimator for gross margins, value of agricultural sales and assets contribution to value of production of main crops. In the first two instances, the estimates changed signs, while the contribution of agricultural assets to gross margins appeared significant at 10% level. The results highlight the importance of control for selection bias on observable characteristics, which the simple DID estimator does not do.

Table 3. here.

To draw inferences about likely eventual project impacts, we estimated project effects on a set of intermediate implementation outcomes. Specifically, we measured the impact of the project on the change in the proportion of households that experienced food scarcity during a period of the year, that experienced stored grain losses, that are implementing agricultural conservation practices in at least one of their plots, the proportion of household participating in groups or associations, households with savings, households with credit, and households that use purchased inputs for agricultural production (Table 4).

Table 4. here

During its first two years, the A4N project increased the proportion of households participating in groups or associations, as well as those reporting savings, we also found weak evidence of a decrease in the losses of stored grains (Table 4). These outcomes are closely related to project interventions such as participation in saving groups, farmer field schools, local research committees and capacity building on postharvest management practices. According to our estimates the change in the percentage of participant households in a group or association after project implementation increased by over 13%, and this difference is significant at the 5% and 1% level (Table 4). The project also boosted the proportion of households with savings, also by approximately 13%, and the estimates are significant at the 1% level across the different estimators (Table 4). As mentioned already, we found some weak evidence of decreases in the percentage of households that experienced stored grain losses (the associated p-values are between 0.12 and 0.17). By its second year, the project had not had an impact in the proportion of households who experienced hunger, used agricultural conservation practices, used purchased inputs, or obtained credit.

Discussion and conclusion

Our results suggest that the PSM-DID and the weighted PS regression estimation, which controls for both selection on observable characteristics and time invariant characteristics, gave broadly similar results. The DID estimation alone gave quite different results, suggesting the importance of correcting for selection on observable characteristics by using PSM, combined with DID, when selection is nonrandom, as it is here.

Although we do not observe project impacts in agricultural income and household wealth, we do observe impacts on outcomes that measure shorter-term impacts of the project, closely related to participation in project interventions. Our analysis identifies intermediate project impacts that could serve as predictors of eventual impacts.

Particularly for agricultural development projects, it is common for impacts to result only after a series of project intervention stages (Gertler et al., 2011). Participants must first learn about and adopt the practices and technologies promoted. Only after that do they begin to realize benefits. With time, they master the use of the practices and technologies adopted, and only then do the benefits translate into income gains and asset accumulation. It is typical that rural development projects will only influence the first two or three stages of this results chain, and project implementation problems when starting many activities at the same time are normal. Hence, impacts do not necessarily take place during the implementation period of the project.

The timing of the impact assessment surveys gives particular cause to expect incomplete impacts. The final survey had to be conducted during the project funding period, so it measured outcomes a year prior to project completion, which in turn was well before full realization of likely impacts. Some short-term impacts of specific interventions suggested probable long-term impacts. Increases in household saving (linked to participation in saving and lending groups) can translate into investments in agricultural assets that are likely to increase household income and asset accumulation. Participation in groups and associations (linked with most project interventions), builds the capacity of participants in the different agricultural technologies promoted by the project, who are likely to move forward on the impact pathway toward the realization of the benefits of these practices that later translate into increases in agricultural incomes. Reduction in losses of stored grains is linked with improved postharvest management, which is likely to improve household food security and may also increase incomes by allowing delayed grain sales at better prices.

Complex rural development projects are challenging to evaluate. Two major areas deserve future attention by impact evaluation researchers. First, measuring long term impacts may require an evaluation strategy for measuring project impacts well after the project ends. More research is required understand how particular interventions are linked to intermediate impacts and how intermediate impacts (e.g., knowledge change and practices adopted) affect such long-term impacts as improved income.

Second, additional research is also needed to explore differential effects of specific interventions on various household types, which could enhance project targeting in future. Our study focuses on the ATT, the mean effect of a program among the treated. Yet as an overall average, the ATT can miss program impacts that vary among subsets of individuals or households (especially for groups that are small). For a program like A4N with a broad set of interventions, preliminary results suggest the existence of heterogeneous impacts across household types. Better understanding of what drives differential impacts could improve project targeting in future.

Figure 1. Probability of participation in the A4N program

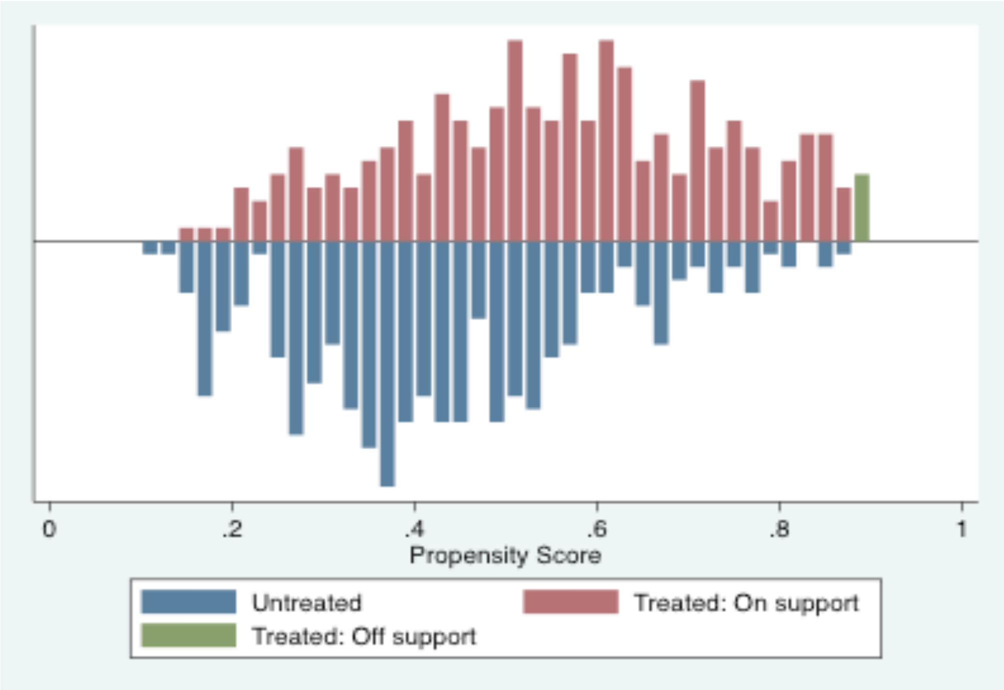

Table 1. Logit estimation of the probability of participating in the A4N program.

Dependent variable: Participation in A4N, n=554 **Explanatory variables**	**Coefficient**
Cultivated land Mz	0.03
	(0.03)
Steep slope=1	0.18
	(0.20)
Inadequate services=1	-0.51**
	(0.22)
Inadequate housing=1	0.11
	(0.29)
Electricity=1	-0.05
	(0.22)
Hunger=1	0.34*
	(0.20)
head female=1	1.19***
	(0.31)
#children<5	0.06
	(0.15)
head age	0.00
	(0.01)
head education	-0.01
	(0.04)
household size	-0.05
	(0.06)
people per room	-0.02
	(0.06)
Infrastructure C$/1000	-0.09*
	(0.06)
Livestock C$/1000	-0.02*
	(0.01)
Equipment C$/1000	0.00
	(0.02)
Population 2009	0.00
	(0.00)
Dist. Market Km/10	-0.05***
	(0.01)
Dist. Paved road Km/10	0.02
	(0.01)
Health facility=1	-0.82***
	(0.26)
Proportion basic grains 2003	-0.18
	(0.63)
Proportion landholdings<10Mz 2003	2.25***
	(0.50)
Constant	-0.20
	(0.84)
Log likelihood	-345

Standard errors in parenthesis. Levels of significance ***1%, **5%, *10%

Table 2. Balancing tests of covariates, before and after matching (after trimming)

	Before Matching				After Matching			
	Mean			t-test	Mean			t-test
Variable	Treated	Control	%bias	p-value	Treated	Control	%bias	p-value
Cultivated land Mz	3.30	3.48	-62.04	0.54	3.32	3.37	-1.6	0.86
Steep slope=1	0.32	0.31	21.54	0.83	0.32	0.37	-9.7	0.27
Inadequate services=1	0.68	0.79	-300.68	0.00	0.67	0.66	3.6	0.69
Inadequate housing=1	0.88	0.85	93.52	0.35	0.88	0.86	4.4	0.60
Electricity=1	0.61	0.63	-56.19	0.57	0.60	0.58	4.9	0.58
Hunger=1	0.38	0.32	158.29	0.11	0.38	0.40	-4.2	0.64
Head female=1	0.18	0.07	384.70	0.00	0.18	0.22	-12.8	0.23
#children<5	0.51	0.51	3.87	0.97	0.51	0.41	14.1	0.09
Head age	49	48	119	0.24	49	49	-1.2	0.89
Head education	2.84	3.04	-90.26	0.37	2.84	2.79	1.7	0.84
Household size	5.20	5.36	-79.82	0.43	5.20	4.99	9.3	0.23
People per room	3.84	3.87	-19.32	0.85	3.84	3.85	-0.6	0.95
Infraestructure C$/1000	0.53	0.80	-170.11	0.09	0.53	0.47	3.3	0.62
Livestock C$/1000	6.78	9.01	-197.43	0.05	6.80	6.08	5.7	0.40
Equipment C$/1000	1.79	2.03	-61.41	0.54	1.80	2.09	-6.2	0.46
Population 2009	642	635	14.68	0.88	645	678	-5.9	0.50
Dist. Market Km/10	14.26	16.31	-293.23	0.00	14.34	14.46	-1.5	0.86
Dist. Paved road Km/10	9.61	8.97	82.41	0.41	9.56	8.63	10	0.25
Health facility=1	0.21	0.28	-199.07	0.05	0.21	0.21	0.7	0.93
% basic grains 2003	0.86	0.88	-110.30	0.27	0.86	0.87	-4.9	0.58
% landholdings<10Mz 2003	0.58	0.52	350.25	0.00	0.58	0.54	18	0.03
n	274	291			272			

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

Table 3. Project impacts on major outcome variables using six methods to correct for selection bias.

	PSM-DID					
Outcome variables (difference)	**kernel (Epan)**	**kernel (normal)**	**LLR (tricube)**	**NN(5)**	**weighted PS reg**	**simple DID**
Agricultural Income						
Dif gross margins C$ 2011	2,154	2,035	2,199	2,569	2,135	-675
	(3,116)	(2,857)	(3,185)	(3,447)	(2,831)	(3,311)
Dif total agricultural sales C$2011	2,305	2,208	2,163	2,760	2,351	-1,726
	(3,306)	(3,182)	(3,600)	(3,652)	(3,170)	(3,708)
Dif value main crops C$ 2011	-760	-823	-780	-601	-1,460	-2,388
	(1,704)	(1,601)	(1,962)	(1,966)	(1,700)	(1,600)
Dif bean production (qq)	-2.01	-1.75	-2.80	-2.38	-2.23	-2.69
	(1.62)	(1.66)	(1.89)	(1.89)	(1.64)	(1.58)
Dif maize production (qq)	-0.74	-0.63	-0.67	-0.07	-0.98	-1.24
	(1.88)	(1.86)	(2.19)	(2.32)	(1.79)	(1.81)
Household Assets						
Dif TLU	-0.03	-0.04	-0.05	-0.02	-0.04	0.04
	(0.20)	(0.20)	(0.22)	(0.24)	(0.19)	(0.18)
Dif ag asset value main crops	1,164	1,142	1,091	1,176	1172	610
	(857)	(753)	(940)	(880)	(787)	(556)
Dif ag asset Ag sales	-2,014	-1,934	-2,199	-1,411	-2,224	-2,835
	(1,980)	(1,783)	(2168)	(2,175)	(1,946)	(1825)
Dif ag asset gross margins	-1,732	-1,710	-1,920	-1,248	-2,058	-2,579*
	(1,645)	(1,511)	(1,831)	(1,846)	(1,617)	(1,491)

Standard errors (se) in parenthesis, PSM-DID se bootstrap with 1000 repetitions, weighted regression and simple DID robust se

Levels of significance ***1%, **5%, *10%

NN refers to nearest neighbor, LLR to local linear regression

n=554, with 272 participant observations and 284 non participant observations

Monetary variables in real cordobas of 2011.

The exchange rate for 2011 was U$1=C$22.4243

qq refers to bags of 100 Kg.

TLU refers to tropical livestock units, conversion factors are: horses 0.8; cattle and mule 0.7; asses 0.5; pigs 0.2; goat, sheep 0.1; poultry 0.01
source: http://www.ilri.cgiar.org/InfoServ/Webpub/fulldocs/X5443E/X5443E04.HTM

Table 4. Project impacts on intermediate outcomes using six methods to correct for selection bias.

	PSM-DID					
Outcome variables (difference take values -1, 0, 1)	**kernel (Epan)**	**kernel (normal)**	**LLR(tricube)**	**NN(5)**	**weighted PS reg**	**simple DID**
Dif experience food	0.04	0.03	0.05	0.05	0.04	-0.05
scarcity[1]	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)	(0.05)
Dif experienced stored	-0.11~	-0.09~	-0.13~	-0.07	-0.11~	-0.16***
grain losses	(0.08)	(0.07)	(0.08)	(0.08)	(0.07)	(0.06)
Dif conservation	-0.02	-0.01	-0.02	-0.03	0.00	0.04
agriculture practices	(0.05)	(0.05)	(0.06)	(0.07)	(0.05)	(0.05)
Dif groups or	0.15***	0.17***	0.14**	0.13**	0.16***	0.19***
associations	(0.06)	(0.05)	(0.06)	(0.06)	(0.05)	(0.04)
Dif use of purchased	0.03	0.04	0.01	0.01	0.04	0.02
inputs	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)	(0.04)
Dif savings	0.13***	0.13***	0.12***	0.13***	0.13***	0.14***
	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)	(0.04)
Dif credit	-0.01	-0.01	-0.03	-0.03	-0.01	-0.02
	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)	(0.04)

[1] Food scarcity refers to households that experience one period of time during the year when they lack enough food to cook one of the daily meals.

Standard errors (se) in parenthesis, PSM-DID se bootstrap with 1000 repetitions, weighted regression and simple DID robust se

Levels of significance ***1%, **5%, *10%

~p values between 0.12 and 0.17

NN refers to nearest neighbor, LLR to local linear regression

n=554, with 272 participant observations and 284 non participant observations

References

Abadie, A., & Imbens, G. W. (2008). On the Failure of the Bootstrap for Matching Estimators. *Econometrica*, *76*(6), 1537–1557. doi:10.3982/ECTA6474

Adato, M., Carter, M. R., & May, J. (2006). Exploring poverty traps and social exclusion in South Africa using qualitative and quantitative data. *Journal of Development Studies*, *42*(2), 226–247. doi:10.1080/00220380500405345

Ashraf, N., Giné, X., & Karlan, D. (2009). Finding Missing Markets (and a Disturbing Epilogue): Evidence from an Export Crop Adoption and Marketing Intervention in Kenya. *American Journal of Agricultural Economics*, *91*(4), 973–990. doi:10.1111/j.1467-8276.2009.01319.x

Becerril, J., & Abdulai, A. (2010). The impact of improved maize varieties on poverty in Mexico: a propensity score-matching approach. *World development*, *38*(7), 1024–1035.

Buddelmeyer, H., & Skoufias, E. (2004). *An evaluation of the performance of regression discontinuity design on PROGRESA* (Vol. 3386). World Bank Publications. Retrieved from http://books.google.com/books?hl=en&lr=&id=5tuCP0PMNvMC&oi=fnd&pg=PA32&dq=%22set+from+rural+Mexico+collected+for+the+purposes+of+evaluating+the+impact+of%22+%22poverty+alleviation+program+to+examine+the+performance+of+a%22+%22the+Regression+Discontinuity+Design+(RDD).+Using+as+a+benchmark+the%22+&ots=_M7LHAYqMJ&sig=Cz8PG5zMCke-ko7lULFNgBhKYug

Caliendo, M., & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, *22*(1), 31–72. doi:10.1111/j.1467-6419.2007.00527.x

Chen, S., Mu, R., & Ravallion, M. (2009). Are there lasting impacts of aid to poor areas? *Journal of Public Economics*, *93*(3–4), 512–528. doi:10.1016/j.jpubeco.2008.10.010

Chowa, G. A. N., & Elliott III, W. (2011). An asset approach to increasing perceived household economic stability among families in Uganda. *The Journal of Socio-Economics*, *40*(1), 81–87. doi:10.1016/j.socec.2010.02.008

Cuong, N. V. (2009). Impact evaluation of multiple overlapping programs under a conditional independence assumption. *Research in Economics*, *63*(1), 27–54. doi:10.1016/j.rie.2008.10.001

Dehejia, R. H., & Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics*, *84*(1), 151–161. doi:10.1162/003465302317331982

Dillon, A. (2011a). The Effect of Irrigation on Poverty Reduction, Asset Accumulation, and Informal Insurance: Evidence from Northern Mali. *World Development*, *39*(12), 2165–2175. doi:10.1016/j.worlddev.2011.04.006

Dillon, A. (2011b). Do Differences in the Scale of Irrigation Projects Generate Different Impacts on Poverty and Production? *Journal of Agricultural Economics*, *62*(2), 474–492. doi:10.1111/j.1477-9552.2010.00276.x

Duflo, E. (2004). The medium run effects of educational expansion: Evidence from a large school construction program in Indonesia. *Journal of Development Economics*, *74*(1), 163–197.

Duflo, E., Kremer, M., & Robinson, J. (2009). *Nudging farmers to use fertilizer: theory and experimental evidence from Kenya*. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w15131

Duflo, Esther, Kremer, M., & Robinson, J. (2008). How High Are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya. *The American Economic Review*, *98*(2), 482–488. doi:10.2307/29730068

Edmonds, E. V., Mammen, K., & Miller, D. L. (2005). Rearranging the Family? Income Support and Elderly Living Arrangements in a Low-Income Country. *The Journal of Human Resources*, *40*(1), 186–207. doi:10.2307/4129570

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2011). *Impact evaluation in practice*. World Bank Publications.

Gilligan, D. O., & Hoddinott, J. (2007). Is There Persistence in the Impact of Emergency Food Aid? Evidence on Consumption, Food Security, and Assets in Rural Ethiopia. *American Journal of Agricultural Economics*, *89*(2), 225–242. doi:10.1111/j.1467-8276.2007.00992.x

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, *66*(5), 1017–1098. doi:10.2307/2999630

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, *64*(4), 605–654. doi:10.2307/2971733

Hirano, K., & Imbens, G. W. (2001). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, *2*(3-4), 259–278. doi:10.1023/A:1020371312283

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, *71*(4), 1161–1189. doi:10.1111/1468-0262.00442

Imai, K. S., & Azam, M. S. (2012). Does Microfinance Reduce Poverty in Bangladesh? New Evidence from Household Panel Data. *Journal of Development Studies*, *48*(5), 633–653. doi:10.1080/00220388.2012.661853

Imbens, G. M., & Wooldridge, J. M. (2008). *Recent developments in the econometrics of program evaluation*. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14251

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.

Islam, A. (2011). Medium- and Long-Term Participation in Microcredit: An Evaluation Using a New Panel Dataset from Bangladesh. *American Journal of Agricultural Economics*, *93*(3), 847–866. doi:10.1093/ajae/aar012

Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2010). Handbook on impact evaluation: quantitative methods and practices. World Bank Publications.

Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: Some lessons. *American Economic Review*, *93*(2), 102–106.

Lai, F., Sadoulet, E., & Janvry, A. de. (2011). The Contributions of School Quality and Teacher Qualifications to Student Performance Evidence from a Natural Experiment in Beijing Middle Schools. *Journal of Human Resources*, *46*(1), 123–153.

Mendola, M. (2007). Agricultural technology adoption and poverty reduction: A propensity-score matching analysis for rural Bangladesh. *Food Policy*, *32*(3), 372–393. doi:10.1016/j.foodpol.2006.07.003

Mu, R., & Van de Walle, D. (2011). Rural Roads and Local Market Development in Vietnam. *Journal of Development Studies*, *47*(5), 709–734. doi:10.1080/00220381003599436

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

Plesca, M., & Smith, J. (2007). Evaluating multi-treatment programs: theory and evidence from the U.S. Job Training Partnership Act experiment., *32*, 491–528. doi:10.100/s00181-006-0095-0

Ravallion, M. (2008). *Evaluation in the Practice of Development* (SSRN Scholarly Paper No. ID 1103727). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1103727

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1–2), 305–353. doi:10.1016/j.jeconom.2004.04.011

Solís, D., Bravo-Ureta, B. E., & Quiroga, R. E. (2009). Technical Efficiency among Peasant Farmers Participating in Natural Resource Management Programmes in Central America. *Journal of Agricultural Economics*, *60*(1), 202–219. doi:10.1111/j.1477-9552.2008.00173.x

Towe, C., & Tra, C. I. (2013). Vegetable Spirits and Energy Policy. *American Journal of Agricultural Economics*, *95*(1), 1–16. doi:10.1093/ajae/aas079

Wanjala, B. M., & Muradian, R. (2013). Can Big Push Interventions Take Small-Scale Farmers out of Poverty? Insights from the Sauri Millennium Village in Kenya. *World Development*, *45*, 147–160. doi:10.1016/j.worlddev.2012.12.014

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