"Matching the gold standard: Comparing experimental and non-experimental evaluation techniques for a geographically targeted program"

by

Sudhanshu Handa
And
John A. Maluccio

September 2008

MIDDLEBURY COLLEGE ECONOMICS DISCUSSION PAPER NO. 08-13

DEPARTMENT OF ECONOMICS
MIDDLEBURY COLLEGE
MIDDLEBURY, VERMONT 05753

http://www.middlebury.edu/~econ

Matching the gold standard: Comparing experimental and non-experimental evaluation techniques for a geographically targeted program

Sudhanshu Handa
Department of Public Policy, University of North Carolina
shanda@email.unc.edu
Phone: (919) 843-0350

John A. Maluccio
Department of Economics, Middlebury College, Middlebury, VT 05753
Phone: (802) 443-5941 Fax: (802) 443-2084
john.maluccio@middlebury.edu

September 2008

JEL Classification: I18, I38, C9
Key words: impact evaluation, non-experimental matching, conditional cash transfers, Nicaragua
Running title: Matching the gold standard

We thank Tia Palermo and Yanmin Choo for excellent research assistance, Alan de Brauw, Juan José Díaz, Dan Gilligan, David Guilkey, John Strauss, Harsha Thirumurthy, and three anonymous referees for many useful comments. Funding for this research was provided by USAID's MEASURE Evaluation project at the Carolina Population Center.

Abstract: We compare non-experimental impact estimates based on matching methods with those from a randomized evaluation to determine whether the non-experimental approach can "match" the so-called gold standard. The social experiment we use was carried out to evaluate a geographically targeted conditional cash transfer antipoverty program in Nicaragua. The outcomes we assess include several components of household expenditure and a variety of children's health outcomes including breast feeding, vaccinations, and morbidity. We find that using each of the following improves performance of matching for these outcomes: 1) geographically proximate comparison samples; 2) stringent common support requirements; and 3) both geographic- and household-level matching variables. Even for a geographically targeted program, in which the selection is at the geographic-, rather than at the individual- or household-level, and in which it is not possible to find comparison individuals or households in the program locales, matching can perform reasonably well. The results also suggest that the techniques may be more promising for evaluating the more easily measured individual-level binary outcomes, than for outcomes that are more difficult to measure, such as expenditure.

1. Introduction

In recent years, non-experimental matching methods have become an increasingly popular way to estimate program impacts for antipoverty or other social programs.[1] As a result, careful assessments of the extent to which these evaluation approaches are accurate—and thus adequate substitutes for social experiments—are needed. This is particularly so since large-scale field experiments are often not feasible, due to political or operational constraints. In this article, we provide evidence on the validity of such techniques by comparing experimental and non-experimental matching estimates of the effectiveness of a Nicaraguan antipoverty and human development program.

Our main contribution is to assess matching techniques for a geographically targeted program. Most previous assessments have considered individual- or household-level targeted programs, where individual- or household-level characteristics are considered most critical for matching treated observations with non-experimentally selected comparison observations. In contrast, in a geographically targeted program, the success of the technique may depend more on the extent to which one can identify good comparison observations from distinct, yet similar, locations, which may be more difficult if there are important observable or unobservable differences across those locations.

We also provide evidence on the reliability of matching methods in a low income context. To our knowledge, only two other articles carry out assessments of matching in the context of a developing country, Díaz and Handa (2006) and McKenzie et al. (2006). Finally, in

[1] Some examples include Pradhan and Rawlings (2002), Jalan and Ravallion (2003), Levine and Painter (2003), Gotland et al. (2004), Sianesi (2004), Gilligan and Hoddinott (2007), and DeBrauw and Hoddinott (2008). For other examples, see the reviews by Imbens and Wooldridge (2008), Todd (2008), and Ravallion (2008).

addition to expenditure, our analysis includes an assessment of the effects of the program on individual-level health indicators, which are outcomes not previously assessed with these techniques, but of particular policy importance.

We find that matching can replicate experimental estimates of the effects of a geographically targeted program reasonably well, particularly when it is done using: 1) geographically proximate comparison samples; 2) both geographic- and household-level matching variables; and 3) stringent common support. Even for a geographically targeted program, in which the selection is at the geographic-level and for which it is not possible to find comparison observations from the same geographic areas where the program is operating, matching can perform well. The results also suggest that the techniques may be more promising for evaluating more easily measured outcomes such as vaccination status, than for outcomes that are more difficult to measure, such as expenditure.

Section 2 briefly reviews the literature assessing non-experimental matching estimators. Section 3 describes the Nicaraguan antipoverty program we examine, the social experiment implemented to evaluate it, and the source of non-experimental data we use. Section 4 outlines the theoretical and empirical frameworks for the analysis and section 5 presents the estimation of the propensity score and common support criteria. Section 6 reports the matching results and section 7 concludes.

2. Selected literature

Over the past three decades, several federal and state sponsored programs in the U.S. have been evaluated experimentally, and some of these randomized evaluations have been used in studies assessing the performance of non-experimental evaluation methods. Most of the evaluations examined have been of employment and job training programs, which are either voluntary, as

with the National Supported Work Demonstration (NSW) and the National Job Training Partnership Act (JTPA), or mandatory, as with a number of state welfare-to-work programs.

Assessments based on the NSW and the JTPA experiments have provided substantial evidence on the reliability of using non-experimental methods for evaluating voluntary programs.[2] For such interventions, which typically have a large pool of eligible candidates but a relatively small number of participants, the central problem for a non-experimental study is finding non-participants in the same labor market (i.e., the intervention areas) who are similar to the actual participants. In this context, selection bias due to program participation arises mostly due to individual-level self-selection. In a series of studies, Heckman et al. (1997, 1998a) and Heckman et al. (1998b) use the JTPA experiment to assess the empirical performance of matching estimators. They find that propensity score matching provides reliable, low-bias estimates of program impact when the researcher is: 1) working with a rich set of control variables capable of predicting program participation; 2) using samples that have the same survey instrument; and 3) identifying participants and non-participants from the same local labor market.

Friedlander and Robins (1995) and Michalopoulos et al. (2004) carry out assessments of non-experimental methods applied to two welfare-to-work programs (examples of mandatory interventions). With this type of program, the central problem for a non-experimental study is to find welfare recipients from non-intervention locations similar to welfare recipients from intervention locations. In this context, selection bias due to program participation arises mainly due to geographic-level differences in labor markets—a selection problem similar to the one we

[2] Todd (2008) and Ravallion (2008) provide thorough reviews of matching and other non-experimental estimators and Cook et al. (2007) a general discussion on observational versus experimental evaluation results.

face. They conclude that substantial biases arise when comparisons are drawn from different geographic areas, consistent with the findings from the JTPA studies.

There is much less research that assesses the performance of non-experimental evaluation techniques on something *other* than employment programs[3] or in developing countries.[4] The analysis most closely related to ours is Díaz and Handa (2006), who provide an assessment of non-experimental evaluation methods for Mexico's *Progresa,* an antipoverty and human capital development program that provides conditional cash transfers. They use survey data from the randomized evaluation of the program and select comparison households from a Mexican national household survey using propensity score matching methods. They find significant bias in the non-experimentally estimated impacts of expenditure but not for schooling enrolment. Since the latter is measured with identical questions across surveys, but not the former, one of their conclusions is that small differences in questionnaires can lead to bias.

Our analysis is based on a Nicaraguan conditional cash transfer (CCT) program modeled after *Progresa*. In addition to examining another CCT program in a different (and poorer) country at a different time, our analysis extends that of Díaz and Handa (2006) in several ways. First, the Nicaraguan and Mexican CCT programs differed in their beneficiary selection, with Nicaragua using only geographic-level targeting and Mexico a combination of geographic- and household-level targeting. Consequently in Nicaragua, (nearly) all households in a chosen

[3] Examples for the U.S. include Hill et al. (2004) for the Infant Health and Development Program, and Agodini and Dynarski (2004) for a school dropout program, both in the U.S.

[4] McKenzie et al. (2006) assess matching methods in a developing country context, using experimental (a lottery) and non-experimental data to explore the effects of immigration from Tonga. They find that matching works well relative to a set of other non-experimental methods, but still overstates income gains by about 20%. Nevertheless, they cannot reject the hypothesis that the bias-adjusted impact estimate is the same as the experimental estimate.

locality were eligible for the program. If there are important differences across geographic areas that were unobservable, it may be difficult to find areas without the intervention that were similar to intervention areas and thus matching may not perform as well.

Second, we assess program effects on child health, a key component of both programs that Díaz and Handa (2006) were unable to assess because the relevant information was not available in their comparison sample. Finally, we explore whether results for expenditure outcomes are improved when identical survey instruments are used. Díaz and Handa (2006) found significant and substantial bias for program effects on expenditure and hypothesized that this was due in part to the feature that expenditure was measured differently across the survey instruments. To put that hypothesis to the test, we examine what happens to the estimated effect on expenditure when the survey instruments are the same.

3. Study Setting and Data Sources

3.1. Nicaragua's Red de Protección Social and the social experiment

In 2000, the Nicaraguan Government piloted the *Red de Protección Social* (RPS), an antipoverty and human development program, in selected rural areas of the central region. RPS was a CCT program that provided transfers to eligible families that fulfilled specific co-responsibilities. The three components to the transfers were:

1) a food security transfer contingent on a designated household member (typically the mother) attending monthly educational workshops that covered *inter alia* nutrition, sanitation, breast feeding, and hygiene, and on all children below five attending regular preventive health checkups that included growth monitoring and vaccinations;

2) a school supplies transfer contingent on children ages 7 to 13 who had not completed fourth grade enrolling in school at the beginning of the school year; and

3) a school attendance transfer contingent on the enrolled children maintaining regular school attendance during the school year.

The combined potential transfer for a family with one child eligible for the school supplies and school attendance transfers was substantial—approximately 20% of preprogram average total household expenditure. Consequently, participation rates were well above 90%. While RPS was a voluntary program, in analyzing it we face the similar problems associated with mandatory programs discussed in Section 2, i.e., that the geographic-level targeting and nearly universal take-up make it nearly impossible to find good comparison households in the specific geographic areas where the program operated.

For the first phase of RPS, the government selected six (out of the 63) municipalities from the central region.[5] There were two main selection criteria. The first was the municipalities' capacity to implement the program. The six chosen municipalities were accessible (e.g., they were less than one day's drive from the capital Managua, where RPS was headquartered) and had reasonably good coverage of health posts and schools. The second criterion was their relatively high levels of poverty. In 1998, approximately 80% of the rural population in the selected areas was poor, and 40% extremely poor, according to Nicaragua-specific poverty lines (World Bank 2003). The central region experienced slightly increasing poverty between 1998 and 2001, in contrast to Nicaragua's other three regions (greater Managua, Pacific coast, and Atlantic coast) where there were declines in poverty (World Bank 2003).

[5] The municipalities were chosen from two of the seven departments in the region, and thus geographically clustered. See Maluccio (2008) for more details.

Within the six chosen municipalities, the 42 (of 59) worst off rural census *comarcas*[6] (hereafter, localities) were selected for the pilot program based on a locality-level marginality index associated with poverty comprised of the following four locality-level indicators: 1) the proportion of households without piped water; 2) the proportion of households without a latrine; 3) the proportion of adults who were illiterate; and 4) the average family size. An evaluation of RPS in the 42 selected localities took place from 2000–2002. The evaluation was based on a randomized, locality-based intervention in which 21 of the localities were randomly designated as treatment and 21 as control.[7] The RPS evaluation sample was a stratified (at the locality level) random sample of households from all 42 localities. A household panel survey was started before the program began in 2000, and implemented again in 2001 and 2002. The sample size of the October 2001 survey round was 1,453 households (53% from treatment localities); we use the 2001 survey round as our principal data source because its timing approximates that of the national survey described next.

3.2 Comparison group

We draw a non-experimental comparison group sample from the 2001 Nicaraguan Living Standards Measurement Survey (LSMS) (World Bank 2003), fielded from April through July 2001. This multipurpose, nationally representative, household survey included 4,191 households.

[6] Census *comarcas* are administrative areas that typically include between one and five small communities averaging 100 households each.

[7] Randomization began by ordering the 42 localities by their marginality index scores and stratifying them into seven groups of six each. Within each stratum, three localities were randomly selected as treatment and three as control. Maluccio and Flores (2005) demonstrate equality across the two groups for a number of household characteristics prior to the program.

Since RPS was targeted to areas with high poverty in one region of the country, a potential concern is that many households from the LSMS national rural sample, for example those from less impoverished or further away areas, may prove to be poor matches for those in the RPS evaluation sample. While this may not be the ideal situation for using matching methods, it is commonly confronted by researchers devising non-experimental evaluation strategies. From an applied perspective, then, the use of a national representative survey to assess the performance of non-experimental techniques for such a program is particularly informative. Because RPS was targeted to rural areas only, then, we exclude from the LSMS sample all households living in urban areas. Furthermore, to avoid potential contamination bias, we also exclude from the LSMS sample (but not from the RPS evaluation sample) nine localities from the same municipalities where RPS was operating. We refer to the resulting sample, with 1,718 households in 169 localities, as the LSMS national rural sample.

In the analyses, we also analyze two further refined sub-samples of the LSMS national rural sample. The first of these retains only those households that also were located in localities with marginality index scores above the cut-off point used for the RPS locality selection described in Section 3.1. We refer to this as the LSMS national high priority rural sample (1,316 households in 154 localities), since with their marginality index scores they were deemed "high priority" localities by RPS. Lastly, because Nicaragua has substantial regional variation (even beyond the differing trends in poverty incidence), for example areas in the Atlantic coast have weaker infrastructure, different prices, and relatively large indigenous populations, we consider a final refinement limiting further the sample to only those households in the central region; the LSMS central region high priority rural sample (638 households in 75 localities). While this third sample is necessarily the smallest, thus reducing the number of potential matches, it includes

areas closer to the RPS targeted areas and thus is likely to be the most geographically similar to the RPS evaluation sample.

3.3 Outcomes to be evaluated

From information available in *both* the RPS evaluation and LSMS samples, we calculate the following expenditure and health outcomes: household expenditure by category (measured in current Nicaraguan Córdobas); breast feeding practices; vaccination coverage; morbidity; and whether a child has had a preventive health checkup. Indicators are measured for appropriate age groups (shown in Table 1). For example, we consider breast feeding practices for infants 0–12 months of age.

With the exception of preventive health checkups,[8] these outcomes were measured using identical questions in the RPS evaluation and LSMS survey instruments. For preventive health checkups, the RPS evaluation survey questionnaire asked, "Did you bring [name] for a checkup in the last six months?", whereas the LSMS survey questionnaire asked, "In the last 12 months (since [month]), has [name] been given a growth checkup?" Thus there are two differences across the questions: 1) the reference period; and 2) the specific type of checkup. These differences provide an opportunity to explore how questionnaire inconsistency affects the performance of matching techniques for an individual-level indicator.

4. Estimation framework

4.1 Theoretical framework

[8] One other minor difference (in the expenditure module) was that the LSMS survey asked for the value of food received in school, while the RPS evaluation survey did not. LSMS expenditure was adjusted to exclude these amounts, which averaged less than ½% of food expenditure. We deflated all expenditure values to a common base using the department-level spatial price index constructed for the LSMS (World Bank 2003).

A key parameter of interest in program evaluation is the (average) treatment effect on the treated (*TT*), which compares the outcome of interest *Y* in the treated state (Y_1) with that in the counterfactual untreated state (Y_0), both conditional on receiving treatment (represented by the indicator $D = 1$).[9] Since both these potential outcomes cannot be observed for any single observational unit (e.g., individual or household), what is needed for the identification of *TT* is estimation of the missing counterfactual outcome, i.e., the outcome for a treated unit had it not received treatment $(Y_0 \mid D = 1)$. Matching techniques are non-parametric estimation methods that draw from a comparison sample to construct an estimate of this counterfactual. A sufficient identification assumption for matching is that conditional on a set of observable characteristics, outcomes in the untreated state are independent of treatment status, i.e., of program participation. This is known as the conditional independence assumption or the assumption of selection on observables (Rosenbaum and Rubin 1983).

Rosenbaum and Rubin (1983) show that if the conditional independence assumption holds for a set of covariates *X*, then it also holds for $P(X)$, a propensity score derived from a nonlinear combination of the components of *X*. Operationally, this is much more tractable since it reduces the dimensionality of the problem—treatment and comparison group units can be matched on one composite score instead of on a set of individual- or household-level characteristics. Denoting by *X* the set of observables, the conditional independence assumption

[9] While we frame our analysis as estimating *TT*, strictly speaking we estimate the average intent-to-treat (*ITT*) effect of the program because we do not condition on *actual* participation. Accordingly, we compare our non-experimentally estimated effects with experimental average *ITT* effects. Since over 90% of households living in the treatment localities participated, however, the *TT* and *ITT* are very similar. Estimating *ITT* has the advantage of allowing us to use the control localities, where we do not observe actual participation status.

becomes $Y_0 \perp D \mid P(X)$, where $\perp$ denotes independence. The slightly weaker assumption of conditional mean assumption is necessary to identify *TT*:

$$E(Y_0 \mid D = 1, P(X)) = E(Y_0 \mid D = 0, P(X)) \tag{1}$$

By conditioning on $P(X)$, we can estimate the unobserved component of *TT*. In particular, we identify the parameter as follows:

$$\begin{aligned} TT(X) &= E(Y_1 \mid D = 1, P(X)) - E(Y_0 \mid D = 1, P(X)) \\ &= E(Y_1 \mid D = 1, P(X)) - E(Y_0 \mid D = 0, P(X)). \end{aligned} \tag{2}$$

In addition to computing *TT*, an alternative approach to assessing how well matching performs is to estimate directly the bias associated with *TT* (Smith and Todd 2005; Díaz and Handa 2006). This is done by comparing control units (e.g., households) from the experimental data with non-experimental comparison units. Intuitively, the performance of matching hinges on the ability to select a comparison group that is similar to the experimental control group. A test of matching performance, then, amounts to a test of the differences in mean outcomes between these two groups. This difference is the expected bias in the matching estimator and can be given by:

$$B(X) = \underbrace{E(Y_0 \mid D = 1, P(X))}_{\textbf{Experimental Controls}} - \underbrace{E(Y_0 \mid D = 0, P(X))}_{\textbf{Matched Non-experimental Comparisons}} \tag{3}$$

An estimated bias of zero demonstrates that matching performs well (Smith and Todd 2005). While the above equations hold in expectation, empirically it is not necessarily the case that the bias plus the non-experimental estimate exactly equals the experimental estimate, due to sampling error. Therefore, in the analyses that follow we estimate and present both *TT* and $B(X)$.

4.2 Econometric methodology

We first construct a propensity (or balancing) score for each household, by estimating a logit regression that predicts the probability of "participation," defined here as living in one of the 42 localities targeted by RPS. To increase precision, we use all households from the RPS evaluation sample (designated as participants), as well as those from the LSMS comparison sample being considered (designated as non-participants) (Smith and Todd 2005). Program participation takes on a value of one if the household was living in one of the 42 targeted RPS localities (i.e., is from the RPS evaluation sample) and zero otherwise (i.e., from the LSMS sample).

For the technique to be valid, the set of covariates used in the propensity score equation should help explain both selection into the program (or program participation) as well as the outcomes of interest (Heckman and Navarro-Lozano 2004; Smith and Todd 2005; Todd 2008). Moreover, it is important to choose covariates that are unlikely to have been influenced by the program. One way to ensure this is to use variables measured before the program. In our main analyses estimating the non-experimental impacts (though not the bias estimates), however, we use data collected in RPS treatment areas after the program began, to better mimic the type of data we expect would be available to most researchers carrying out a non-experimental evaluation. The specific variables we choose are described and justified in Section 5.1, and in Section 6.3 we consider an alternative to using data from the 2001 RPS evaluation sample measured after the intervention began. Lastly, we perform balancing tests to ensure that, within small intervals of the propensity score, both the mean propensity scores and the mean values of each of the covariates are "balanced" (i.e., not statistically different) between participant and comparison group households (Todd 2008). All results reported below are based on balanced propensity score models.

Once a propensity score has been calculated for each household, we carry out two types of matching. The first is nearest neighbor matching (Abadie and Imbens 2006) where, for each sample considered, the set of covariates we match on (using the algorithm described in Abadie et al. 2004) are those resulting from estimating and balancing the propensity score relation for that particular sample.[10] The second type of matching we use is Gaussian kernel matching, based directly on the estimated propensity scores (Todd 2008). The kernel method matches each treatment unit to a weighted average of the outcomes of comparison group units within the bandwidth, with weights inversely related to their distance from the propensity score of the treated household.

5. Propensity score model and common support

5.1 Propensity score equations

Results from logit estimates used to derive the propensity score for the LSMS national rural sample are presented in Table A1.We include as covariates locality- and household-level characteristics that are, as previously suggested, likely to affect both the probability of participating in the program as well as the various outcomes being evaluated, but that are at the same time unlikely to have been themselves influenced substantially by the program. They include each of the four locality-level indicators that were used to construct the marginality index used for selecting localities for the program (Section 3.1). These were taken from the 1995 National Population and Housing Census (NPHC), and include the proportion of households without piped water, the proportion of households without a latrine, the proportion of adults who were illiterate, and the average family size for each locality. Because of the important role of

[10] We do not use nearest neighbor matching based on the propensity scores themselves (used in much previous work), because calculation of the standard errors is not feasible, even with bootstrapping (Abadie and Imbens 2008).

these variables in locality selection, we also include a number of transformations (quadratics and interactions) of them in the model. Other locality-level variables (also based on the 1995 NPHC) in the model include the logarithms of the total population and total number of households. In addition, we constructed a number of locality-level measures based on the 2001 household data, including the distances from the locality to the nearest health clinic and primary school.

We also include a set of household-level variables (measured in 2001) commonly associated with poverty (e.g., household head's sex and schooling; household size and demographic composition; indicators of dwelling characteristics such as main material of walls, roof, and floor; toilet and kitchen facilities; access to piped water; availability of durable goods; and adult occupations).[11]

5.2 Common support

The region of overlap or common support of the propensity scores for households in the RPS and LSMS samples determines the extent to which one can find "good" matches. Since the control and treatment areas were randomly assigned and were similar before the program, in replicating the experimental results, what is most relevant is that the support for the LSMS households contains that of the RPS households. Figure 1 presents histograms of the propensity scores for each sample based on predictions using the estimated propensity score model shown in Table A1.[12]

[11] We do not ascribe causal interpretations, nor assess statistical significance of the propensity score models (Heckman and Navarro-Lozano 2004).

[12] The propensity score models based on the LSMS national high priority rural and LSMS central region high priority rural samples were balanced using slightly different sets of variables than those used for the LSMS national rural sample. Nevertheless, they yielded similar density patterns to those presented in Figure 1, though with a little less probability mass in the extremes.

Within the RPS evaluation sample, the distribution of propensity scores is similar between control and treatment households, though the latter has more probability mass at lower predicted propensity scores. In Section 5.1, we argued that the right-hand-side covariates in the propensity score model were unlikely to have been affected substantially by the program. The observed difference between the distributions of propensity scores between randomly determined control and treatment households, however, raises the possibility that some of the right-hand-side covariates were affected. Given that RPS began in late 2000, it is clear that it had no effect on the locality-level variables measured in the 1995 NPHC, or on parental characteristics such as age and education. There is some evidence that household demographics and work patterns were affected modestly by the intervention, though not durables (Winters et al. 2007; Maluccio 2005). We assess the robustness of our findings to potential biases from using the household variables measured in 2001 by using household variables measured in 2000 in Section 6.3, but continue with the 2001 sample for our main results to better synchronize the timing of observations with the comparison LSMS survey. Since the program areas (including treatment areas) were undergoing a general downturn from 2000 to 2001, changes over that period would not have been due solely to the program. Moreover, those data better mimic what would commonly be available for researchers considering non-experimental matching.

In both the RPS control and treatment sub-samples, there are observations with predicted propensity score over nearly the entire range from 0 to 1, though large proportions have predicted propensity scores of 0.95 or higher (Figure 1). Households from the LSMS national rural sample, on the other hand, have much lower predicted propensity scores, with the majority below 0.20 and a large proportion below 0.05. As with the RPS evaluation sample, however, there are households in the LSMS national rural sample with predicted propensity scores over

nearly the entire range from 0 to 1. As a result, it is feasible to find matches for most households in the RPS evaluation sample.

A key aspect of our examination of experimental versus non-experimental evaluation techniques is that the program being evaluated was geographically targeted. To the extent that unobservable factors were correlated with selection of program areas, or, worse still, played an important role in the selection of the program areas, it may be more difficult to find good matches. To address this potential difficulty, in the matching we directly use the key factors known to have been incorporated into locality selection decisions. We find that it is indeed possible to distinguish between participants and non-participants using the logit model, despite the large number of similar localities in the comparison sample. As such, there are only a small number of potentially good matches. This concern is mitigated, however, by our finding that the region of common support is substantial.

For our main results, we follow usual practice and only consider observations that lie on the common support. We us the standard approach to construct the common support, retaining all households in the RPS evaluation sample or the LSMS comparison sample being considered that have propensity scores above the larger of the minimum propensity scores for the two distributions and below the smaller of the maximum propensity scores. For the comparisons based on the LSMS national rural sample, this decision rule eliminates 31% of the LSMS sample, 39% of the RPS evaluation control sub-sample, and 26% of the RPS evaluation treatment sub-sample. These are substantial proportions, suggesting that the resulting sample may not be representative of all participants, and therefore may be altering the underlying parameters (e.g., *TT*) being estimated.

6. Propensity score matching results

6.1 Nearest neighbor

Table 1 presents nearest neighbor matching estimates of the bias, $B(X)$, and the non-experimental impact (*TT*), using observations on the common support for each of the three different comparison samples.[13,14] Column 7 reports the experimental impact estimated from all observations in the RPS evaluation sample and, to facilitate an assessment of the magnitudes of the effects, Column 8 presents the mean of the outcome variable in the RPS evaluation control sub-sample. Since our main objective is recovering *TT*, then, despite limiting the matching estimates to observations on the common support, the relevant comparison is with the experimentally estimated impact for *all* households in the RPS evaluation sample. Our discussion of Table 1 focuses on: 1) the statistical significance of the estimates of bias; 2) an assessment of how well the non-experimental estimated impacts approximate the experimental impacts; and 3) the differences in the findings across the three samples.

Columns 1, 3, and 5 show direct estimates of the bias (RPS control - LSMS) using the three different comparison samples. For all three, the estimated bias for the total expenditure per capita using nearest neighbor matching is statistically significant and negative, which suggests that the true program effects were underestimated (by about one-fourth to one-half). Díaz and Handa (2006) also find large bias estimates for total expenditure per capita for Mexico's *Progresa*. However, our assessment differs from theirs in that the survey instruments used here

[13] We also considered the bias corrected estimator described in Abadie and Imbens (2006) but this algorithm was unstable in our sample and led to results for many of the individual-level (binary) indicators substantially outside any feasible range (e.g., greater than one).

[14] When estimating program impacts and direct measures of bias for child outcomes, we assign each child his or her household's covariates (or propensity score, in the case of kernel matching) and then match. The restriction of the samples to households with children does not substantially change the common support patterns described above.

are identical in the evaluation and comparison surveys. As a result, the bias for Nicaragua is not due to the incomparability of survey instruments.

The measure of total expenditure we use includes imputed values for both housing rent and the value of services rendered by the durable goods owned. Not only is information on these items collected in other parts of the survey outside the expenditure module, but also their calculation is subject to a number of assumptions about the comparability of different types of housing, and initial values and depreciation rates for durable goods. While the same methodology was used for both surveys we exclude the housing and durable goods components from total expenditures to assess whether these imputations could be responsible for bias. When we use instead this measure of "adjusted" total expenditure per capita (reported in the second row of Table 1) the estimates of bias are the same size (as a percent of the corresponding mean) and remain statistically significant. Further, when we consider only food expenditure, the estimated biases persist. Possible differences in the imputation of the monetary value of housing and durable goods, or of the assessment of expenditures on non-food items in general, do not appear to be driving the bias. An examination of the individual-level indicators in the bottom portion of Table 1, however, shows that the directly estimated biases are comparatively small. For example, when using the central region high priority rural sample as the comparison group (Column 5), only one of the seven indicators show statistically significant bias.

In Columns 2, 4, and 6 of Table 1, we report non-experimental impact estimates using nearest neighbor matching (RPS treatment - LSMS); these estimates can be compared to the experimental impact in Column 7. The latter are estimated as a first difference across treatment

and control groups using the entire RPS evaluation sample.[15] As we refine the comparison sample, that is, select comparison observations only from more similar and closer geographic areas, we improve the estimated impacts on expenditure and they approach the experimentally estimated impacts. Similarly, refining the comparison sample improves the estimated impacts on the individual-level indicators. The most improvement in accuracy comes from using comparison households from closer geographic areas, consistent with the results reported for the mandatory labor market programs described in Section 2. For the 12 indicators shown in Column 6, seven agree in sign and statistical significance with the experimentally derived results shown in Column 7.

6.2 Kernel matching

There are at least two potential problems with using the nearest neighbor techniques presented above. First, despite the fact that only about a third of the LSMS households lie outside the common support region, considerably fewer households from the LSMS samples are actually matched using nearest neighbor techniques; this is because the density of the predicted propensity scores for the LSMS is thin in regions where the density of the RPS predicted propensity scores is thick. The second problem pertains to child outcomes; since the propensity score is calculated at the household level and some households have multiple children, nearest neighbor (without further refinement) randomly picks one of possibly two or more children as the designated match for each treated (or control) child.

To address these potential problems, we first limit the age ranges under consideration to three years or less to avoid large numbers of households with multiple children; for example, the

[15] Experimental results reported in this article are comparable to the single-difference estimates presented in Barham and Maluccio (2008) for vaccination rates and Maluccio and Flores (2005) for most of the other outcomes.

problem of multiple children is unlikely to be of concern when considering breast feeding outcomes for children 0–12 months, since only 10% of households have two children in this age range. Next, we use kernel matching, whereby all children (and households) in the common support region and within a certain relevant bandwidth are used in the calculation of the counterfactual, with equal weight given to each child from the same household.

In Table 2, we present results for the same outcomes and comparison samples as presented in Table 1, but where the matches are constructed using a Gaussian kernel estimator with a bandwidth of 0.06, and standard errors are calculated via bootstrapping with 1,000 repetitions.[16, 17] In these analyses, bootstrapping leads to estimated standard errors that are in general larger than those calculated for the nearest neighbor matching shown in Table 1. These larger standard errors notwithstanding, the kernel estimation results are broadly similar to the nearest neighbor results, with only a few notable differences. First, the kernel results show an even clearer benefit from moving toward refined comparison samples that more closely approximate the geographical areas of the program. The central region results yield only two (out of 12) significant biases (Column 5), in contrast to seven for the national rural sample (Column 1). Comparing the non-experimental impact results from Column 6 with the experiment (again in terms of sign and statistical significance), only three of the indicators (never breast fed, BCG,

[16] As there is no formal guidance regarding bandwidth selection directly applicable to matching estimators, we follow other researchers and examine the sensitivity of our results to bandwidth selection in an ad hoc fashion (Imbens and Wooldridge 2008). In the tables, we report results using a bandwidth of 0.06, but the results are robust to both smaller (0.04) and larger (0.08) bandwidths.

[17] Unlike with propensity score matching, with kernel matching based on the propensity scores, bootstrapping is a valid method for calculating the standard errors (Todd 2008).

and DPT) disagree, so that these estimates also fare better than the nearest neighbor estimates presented in Table 1.[18]

While by the above metric, the non-experimental estimate for preventive health checkups is "correct" (as defined by having the same sign and significance as the experimentally estimated effect), the estimated effect is substantially, and significantly, biased for the central region high priority sample. This finding of bias is consistent with previous research on matching—it seems the technique cannot overcome differences in data collection methodology and performs poorly when survey questions are not identical.[19]

The overall results from Tables 1 and 2 confirm that when evaluating the impact of a geographically targeted program, non-experimental matching methods work best when the comparison group is taken from more similar, and in this case, closer, geographical areas. The following sub-sections further examine the robustness of the non-experimental results along three additional dimensions: 1) sensitivity to the propensity score model variables; 2) sensitivity to the common support restrictions; and 3) the components of household expenditure. We carry out these analyses using the method and sample that produced the "best" results (i.e., those which produced the greatest number of non-experimental impact estimates consistent with the

[18] We also examined results for nearest neighbor matching with the two nearest neighbors—the vast majority are in between those presented in Tables 1 and 2.

[19] For this indicator, two factors render an *a priori* assessment of the direction of potential bias difficult. First, the difference in reference periods likely exerts downward bias, because with the LSMS (past 12 months), more would have reported having taken a child to a checkup than with the RPS (past 6 months). Second, the difference in wording likely exerts upward bias, because with the LSMS which refers to a "growth checkup," fewer would have reported having taken a child to a checkup than with RPS, which refers to a "checkup" in general.

experimental ones in terms of sign and significance)—the kernel matching method applied to the central region high priority sample.

6.3 Consideration of different propensity score models

Smith and Todd (2005), and Todd (2008) provide evidence that matching results can be sensitive to the propensity score model employed. In this sub-section, we compare three different propensity score models (each estimated and balanced using the central region high priority sample). Also, we assess whether using only those variables directly used in geographic RPS targeting (i.e., the geographic-level indicators) is sufficient for successful matching.

The first model includes both geographic- and household-level variables; we have already presented the results for this model in Table 2 (Columns 5 and 6). The second model uses only geographic-level indicators; these include all those variables used to construct the marginality index (Section 5.1). If selection into the program was due mainly to geographic-level variables, and we included the relevant ones, then this second model should work just as well as the first. The third model we consider includes only the household-level variables. Household-level variables are likely to be correlated with the geographic-level variables, but if they are not sufficiently correlated with the variables driving geographic selection, then this model would not do as well as a model directly including those geographical-level variables.

In the first four columns of Table 3, we present results derived using the two additional propensity score models just described. Examining the expenditure results, it is clear that on the basis of bias, the model with both geographic- *and* household-level variables (Column 5 of Table 2) is preferred to the models based only on subsets of those variables. Indeed, when only geographic-level variables are used, estimated biases are large and significant (and estimated effects small or insignificant). This occurs even though it is the geographic-level variables only

model that retains the largest number of observations from the RPS evaluation sample in the common support (final row of Tables 2 and 3). To more closely approximate the experimental impacts on expenditure outcomes, it appears that *both* household- and geographic-level variables are necessary. This could reflect important household-level heterogeneity in the program areas. Regardless, the findings are consistent with Todd's (2008) general conclusion that a cruder set of matching controls leads to greater bias.

We reach similar conclusions when we examine the individual-level outcomes. For these outcomes, however, the combined geographic- and household-level variable model does not dominate as strongly. Six of the nine individual-level non-experimental estimates from the combined model (Column 6 of Table 2) have the same sign and significance as the experimental results versus five for the geographic-level only (Column 2 of Table 3).

The final variation in propensity score estimation we consider addresses the concern that by 2001 some of the household-level variables in RPS treatment areas may have been affected by the program (Section 5.2). For the results presented in Column 5, we drew all of the household-level variables for the RPS households from the year 2000 RPS evaluation sample (the baseline survey carried out before the program began), instead of the 2001 RPS sample. Consistent with the finding that some of those variables were affected, this estimated propensity score model yields histograms of the estimated propensity scores (not shown) for the treatment and control areas that are even more similar to each other than the those shown in Figure 1. For nearly every outcome, the point estimate of the bias or the impact is slightly smaller in magnitude (Column 5 of Table 3); nevertheless, the conclusions are qualitatively similar to those we make using the 2001 data.

6.4 Altering the common support regime

In Table 3, we also report (in Columns 6 and 7) findings using the geographic- and household-level variable propensity score model but where we have altered the set of observations retained for matching. It is not always preferable to drop observations based on the stringent common support rules employed above, because one loses potentially good matches just outside the cut-off points (Todd 2008). We first re-estimated the results including, in addition to all the observations on the common support, all observations within 0.02 of the common support cut-off points (Column 6 of Table 3). For household expenditure, the results improve somewhat in that the estimated impacts are now quite similar to the experimental impacts. Results for the individual measures, however, worsen. In particular, while the point estimates change little, now none of the results for vaccinations is statistically significant, as most are less precisely estimated with the addition of the "off" support observations.

Alternatively, it is possible that using a stringent common support rule retains observations in regions of support where there is little density; as such, these observations may be better treated as outliers to be excluded (Todd 2008). We re-estimated the results but this time restricting the sample beginning with all observations on the common support by dropping observations just inside (in particular, within 0.02) of the common support boundaries (Column 7 of Table 3). The point estimates for expenditure measures are now slightly smaller; and, as with the alteration above, there are fewer significant impact estimates for the individual outcomes. Thus, small refinements to the common support regime can lead to changes in the estimated program impacts, supporting the view that good practice should include sensitivity analyses around the common support to provide additional evidence as to how reliably one should treat results based on the common support (e.g., Crump et al. 2006).

6.5 Exploring the bias in expenditure outcomes

Despite getting the signs and significance right, the results for total and food expenditure per capita show large biases, even though the components of the questionnaires related to expenditures were identical across both surveys. Given the importance of the various expenditure outcomes as welfare indicators in the assessment of antipoverty programs, we now explore these biases in more depth.

Collecting expenditure information is time consuming, complex, and expensive, and the resulting data is often subject to substantial measurement error (Deaton and Zaidi 2002). Even with identical questionnaires, then, there can be real concerns about the comparability of the information collected during different surveys. One possible source of differences for this study is interviewing techniques. Most, but not all, of the questions related to expenditures come from the questionnaire's expenditure module. This section tends to be quite long, in the Nicaraguan case involving 60 different food items and an additional 62 non-food items. Moreover, these questions occur near the end of the interview. For the LSMS, concerns about respondent fatigue in answering these questions were addressed by conducting the survey over the course of two visits to the household; this was not done for the RPS evaluation survey[20] due to budget constraints. Enumerator training and supervision were also less extensive in the RPS evaluation survey than in the LSMS.

We disaggregate total expenditure into seven components, shown in Table 4. Consistent with their high rates of poverty, households in the RPS evaluation sample spend nearly 70% of total household expenditure in 2001 on food. The second largest component of expenditures is the share on "other non-food" items, which includes *inter alia* personal and household necessities. As with food expenditures, these other expenditure items are asked about in the

[20] The RPS evaluation survey was about one-third shorter than the LSMS.

expenditure module of the questionnaire. In contrast, the remaining expenditures come mostly from other sections of the questionnaire. These include expenditures on health and education, which were relatively small on average, and taken largely from questions in the health and schooling modules. Household utilities and the value of housing and durables also are collected in other parts of the questionnaire, with the latter two imputed as described in Section 6.1.

In Table 5, we report the estimated bias and impact for each of the seven components of expenditure (per capita) reported in Table 4. The vast majority of the bias comes from the largest expenditure components, food and other non-food items, both measured in the expenditure module. The overall bias in adjusted per capita total expenditure is -392.5 (Column 5, row 2 of Table 2). The percent of this total accounted for by the bias on food expenditure is 49% and by the bias on other non-food items, is 33%. The next largest contribution to the overall bias is from household utilities expenditure, approximately 16%.

The allocation of the bias across components of expenditure, therefore, is roughly proportional to the budget share of each component in total spending. In addition, the largest bias is for those expenditure components taken from the relatively long and tedious expenditure module. This pattern of (negative) bias is consistent with the possibility that the LSMS data were collected by more experienced professional staff who were able to obtain more complete information over two visits instead of one. Nevertheless, we cannot rule out that it is also due to other possible differences (such as local prices, supply effects, or seasonality patterns across areas), though these have been controlled for to some extent by drawing all comparison observations from the high priority central region, the region most similar to the evaluation areas.

7. Conclusions

Evaluating the effectiveness of antipoverty programs is critical. Since social experiments are not always feasible, it is important to know whether, and to what extent, we can rely on non-experimental evaluation techniques. We assessed the performance of non-experimental matching techniques in the context of the *Red de Protección Social*, a conditional cash transfer antipoverty program in Nicaragua. The geographically targeted nature of RPS posed the challenge of finding good comparison households from geographical locations where the program was not operating.

How well could researchers have replicated the true experimental results for the expenditure and health outcomes of RPS had the program been evaluated using non-experimental matching techniques? Our preferred non-experimental program impact estimates were those using the combined geographic- and household-level propensity score model, with kernel matching, and based on the common support for the central region high priority rural sample (Column 6 of Table 2). The experimental results (Column 7 of Table 2) showed positive and significant effects of RPS on all three expenditure measures and six of the nine individual-level health indicators (namely, exclusive breast feeding, BCG, polio, MMR, full vaccination coverage, and preventive health checkups). For all but one of these outcomes (BCG), the non-experimental estimates were positive and statistically significant. The non-experimental estimates, however, also suggested an additional positive impact where one apparently did not exist, for DPT coverage.

With this sample and model specification, then, the non-experimental impact estimates would have generally "gotten it right," finding a positive program impact for the majority of the indicators. Of course, "getting it right" in terms of sign and significance depends in part on the magnitude of the actual treatment effect, the larger the true effect, the better the chance that matching techniques also will find an effect—and RPS had large effects. We found that the non-

experimental point estimates tended to substantially understate the magnitudes of program impacts for expenditure outcomes, but to overstate those for the individual-level health outcomes. These latter biases, however, were large relative to the mean for only a few of the individual-level indicators, one of which was preventive health checkups, which were measured differently across the two surveys.

Our findings, therefore, suggest caution when relying on non-experimental estimates. Three factors appear to be especially important for the performance of matching a geographically targeted evaluation: 1) the choice of comparison sample; 2) the choice of matching variables; and 3) the complexity of outcome variables, even when measured with identical survey questions. A fourth factor we would emphasize, though we could not directly test its importance, is that the researcher should have a clear understanding of the process by which individuals, households, and geographic areas were selected into the program. The less transparent this selection process is, at whatever level it occurs, the more difficult it is to replicate using observational data and statistical techniques.

Our results, therefore, confirm earlier research that matching performs better when the researcher more closely approximates the geographic-level characteristics of the program areas, by selecting households from nearer, and, more similar, locales. This was in spite of the fact that making such restrictions reduced the number of potential matches substantially. Overall results were also better when estimated on samples using stringent common support requirements.

We also explored whether, for a geographically targeted program in which the targeting process was transparent, matching can be done successfully using only geographic-level variables while ignoring household-level variables. For relatively simple to measure and easy-to-collect binary indicators such as the child health indicators we considered, geographic-level

variables alone performed nearly as well as when both geographic- and household-level variables were used to construct matches. For more difficult-to-measure continuous outcomes such as expenditure, however, matching using both geographic- and household-level variables performed better. Results for both sets of outcomes were poorer when geographic-level variables were not included, consistent with the importance of the geographic-level selection resulting from the targeting process.

Lastly, our results, in conjunction with Díaz and Handa (2006), suggest that matching techniques may be more promising for evaluating relatively easy to measure outcomes such as the individual-level binary health indicators we considered than for outcomes such as expenditure. As a consequence, the findings raise the possibility that household surveys may not be necessary to evaluate the effects of social programs. Instead, it may suffice to invest in monitoring systems that accurately track individual-level information and then compare that information to household survey data using matching techniques.

8. References

Abadie, Alberto, David Drukker, Jane Leber Herr, and G.W. Imbens. 2004. "Implementing Matching Estimators for Average Treatment Effects in Stata." *Stata Journal* 4(3): 290–311.

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1): 235–67.

Abadie, Alberto, and Guido W. Imbens. 2008 "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* (forthcoming).

Agodini, Roberto, and Mark Dynarski. 2004. "Are Experiments the Only Option? A Look at Dropout Prevention Programs." *Review of Economics and Statistics* 86(1): 180–94.

Barham, Tania , and John A. Maluccio. 2008. "Eradicating Diseases: The Effect of Conditional Cash Transfers on Vaccination Coverage in Rural Nicaragua." Working paper, Department of Economics, University of Colorado, Boulder, CO.

Cook, Thomas D., William R. Shadish, Jr., and Vivian C. Wong. 2007. "Within-Study Comparisons of Experiments and Non-Experiments: What the Findings Imply for the Validity of Different Kinds of Observational Studies." Working paper, Department of Economics, Northwestern University, Chicago, IL.

Crump, Richard K., V. Joseph Hotz, Guido Imbens, and Oscar Mitnik. 2006. "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand." NBER Working Paper No. T0330, Cambridge, MA.

Deaton, Angus, and Salman Zaidi. 2002. "Guidelines for Constructing Consumption Aggregates for Welfare Analysis." LSMS Working Paper Number 135, World Bank, Washington, DC.

de Brauw, Alan, and John Hoddinott. 2008. “Must Conditional Cash Transfer Programs Be Conditioned To Be Effective? The Impact of Conditioning Transfers on Schooling Enrollment in Mexico.” Food Consumption and Nutrition Division Discussion paper No. 757, International Food Policy Research Institute, Washington, DC.

Díaz, Juan José, and Sudhanshu Handa. 2006. “An Assessment of Propensity Score Matching as a Non-experimental Impact Estimator: Evidence from Mexico’s *PROGRESA* Program.” *Journal of Human Resources* 41(2): 319–45.

Friedlander, Daniel, and Phil Robbins. 1995. “Evaluating Program Evaluations: New Evidence on Commonly Used Non-experimental Methods.” *American Economic Review* 85(4): 923–37.

Gilligan, Daniel O., and John Hoddinott. 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): 225–42.

Gotland, Erin M., Elisabeth Sadoulet, Alain De Janvry, Rinku Murgai, and Oscar Ortiz. 2004. “The Impact of Farmer Field Schools on Knowledge and Productivity: A study of Potato Farmers in the Peruvian Andes.” *Economic Development and Cultural Change* 53(1): 63–92.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1997. “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program.” *Review of Economic Studies* 64(4): 605–54.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1998a “Matching as an Econometric Evaluation Estimator.” *Review of Economic Studies* 65(2): 261–94.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998b. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017–89.

Heckman, James, and Salvador Navarro-Lozano. 2004 "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models." *Review of Economics and Statistics* 86(1): 30–57.

Hill, Jennifer L., Jerome P. Reiter, and Elaine L. Zanutto. 2004. "A Comparison of Experimental and Observational Data Analyses." In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. Andrew Gelman and Xiao-Li Meng. New York, NY: John Wiley & Sons, Ltd.

Imbens, Guido W. and Jeffrey M. Wooldridge. 2008. "Recent Developments in the Econometrics of Program Evaluation." IZA Discussion Paper No. 3640, Bonn, Germany.

Jalan, Jyotsna, and Martin Ravallion. 2003. "Estimating the Benefit Incidence of an Antipoverty Program Using Propensity Score Matching." *Journal of Business and Economic Statistics* 21(1): 19–35.

Levine, David, and Gary Painter. 2003. "The Schooling Costs of Teenage Out-of-Wedlock Childbearing: Analysis with a Within-School Propensity-Score-Matching Estimator." *Review of Economics and Statistics* 84(4): 884–900.

Maluccio, John A. 2005. "Coping with the Coffee Crisis in Central America: The Role of the Nicaraguan *Red de Protección Social*." Food Consumption and Nutrition Division discussion paper No. 188, International Food Policy Research Institute, Washington, DC.

Maluccio, John A. 2008. "Household Targeting in Practice: The Nicaraguan *Red de Protección Social*." *Journal of International Development* (forthcoming).

Maluccio, John A., and Rafael Flores. 2005. “Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan *Red de Protección Social*.” Research Report No. 141, International Food Policy Research Institute, Washington, DC.

McKenzie, David, John Gibson, and Steven Stillman. 2006. “How Important is Selection? Experimental versus Non-Experimental Measures of the Income Gains from Migration?” World Bank Policy Research Working Paper No. 3906, World Bank, Washington, DC.

Michalopoulos, Charles, Howard Bloom, and Carolyn Hill. 2004. “Can Propensity Score Methods Match the Findings From A Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?” *Review of Economics and Statistics* 86(1): 156–79.

Pradhan, Menno, and Laura B. Rawlings. 2002. “The Impact and Targeting of Social Infrastructure Investments: Lessons from the Nicaraguan Social Fund.” *World Bank Economic Review* 16(2): 275–95.

Ravallion, Martin. 2008. “Evaluation of Antipoverty Programs.” In *Handbook of Development Economics*, Volume 4, ed. T. Paul Schultz and John Strauss. New York, NY: North-Holland (pp. 3787–3846).

Rosenbaum, Paul, and Donald Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70(1): 41–50.

Sianesi, Barbara. 2004. “An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s.” *Review of Economics and Statistics* 86(1): 133–55.

Skoufias, Emmanuel. 2005. *PROGRESA* and Its Impacts on the Human Capital and Welfare of Households in Rural Mexico: A Synthesis of the Results of an Evaluation by IFPRI. IFPRI Research Report No. 139. International Food Policy Research Institute, Washington, DC.

Smith, Jeffrey, and Petra Todd. 2005 “Does Matching Overcome LaLonde’s Critique of Non-experimental Estimators?” *Journal of Econometrics* 125(1–2): 305–53.

StataCorp. 2007. *Stata statistical software: Release 10.0*. College Station, Texas: Stata Corporation.

Todd, Petra. 2008. “Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated.” In *Handbook of Development Economics*, Volume 4, ed. T. Paul Schultz and John Strauss. New York, NY: North-Holland (pp. 3847–3894).

Winters, Paul, Guy Stecklov, and Jessica Todd. 2007. “Household demography in the short-run: The response of household structure to economic change in Nicaragua.” Working paper, Department of Economics, American University.

World Bank. 2003. “Nicaragua Poverty Assessment: Raising Welfare and Reducing Vulnerability.” Report No. 26128-NI, World Bank, Washington, DC.

Table 1: Nearest neighbor matching using common support: By comparison sample

	Non-experimental						Experimental	
	National rural		National high priority rural		Central Region high priority rural			
	Bias	Impact	Bias	Impact	Bias	Impact	Impact	Mean
Household level	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Total expenditure per capita	**-391.5**	**486.7**	**-454.2**	**561.3**	**-516.5**	**834.1**	**1224.4**	3261.1
	(163.6)	(161.4)	(153.1)	(165.4)	(201.3)	(168.9)	(223.7)	[2598.4]
Adjusted total expend. p.c.[1]	**-362.9**	**468.6**	**-430.4**	**552.6**	**-494.2**	**773.5**	**1143.2**	2910.6
	(151.1)	(151.5)	(141.6)	(152.7)	(182.7)	(159.8)	(214.6)	[2400.6]
Food expenditure per capita	**-199.5**	**607.0**	**-256.6**	**659.4**	**-257.1**	**805.9**	**1008.9**	2166.1
	(107.6)	(116.4)	(102.4)	(116.1)	(118.1)	(121.8)	(169.8)	[1726.3]
Child level								
0–12 months								
Exclusive breast feeding	**0.240**	**0.256**	**0.202**	**0.233**	0.200	0.140	0.134+	0.569
first 3 months	(0.085)	(0.096)	(0.074)	(0.089)	(0.138)	(0.124)	(0.077)	
Never breast fed	**-0.078**	-0.064	-0.085+	-0.058	-0.140	-0.120	0.021	0.024
	(0.039)	(0.060)	(0.049)	(0.057)	(0.098)	(0.102)	(0.026)	
0–24 months								
BCG	0.016	0.064	0.029	0.051	0.017	0.008	0.035+	0.934
	(0.032)	(0.044)	(0.028)	(0.036)	(0.028)	(0.025)	(0.020)	
12–36 months								
Polio	0.034	**0.146**	0.040	**0.126**	-0.026	**0.060**	**0.041**	0.932
	(0.041)	(0.043)	(0.042)	(0.041)	(0.031)	(0.026)	(0.020)	
MMR	-0.015	**0.149**	-0.019	**0.086**	-0.011	0.087+	**0.063**	0.880
	(0.046)	(0.039)	(0.042)	(0.033)	(0.041)	(0.051)	(0.027)	
DPT/Pentavalent	0.040	**0.162**	0.056	**0.160**	0.004	0.073+	0.020	0.916
	(0.047)	(0.050)	(0.049)	(0.047)	(0.042)	(0.044)	(0.022)	
Up-to-date vaccination	-0.046	**0.202**	-0.024	**0.138**	-0.071	0.093+	**0.083**	0.801
	(0.057)	(0.055)	(0.056)	(0.040)	(0.047)	(0.056)	(0.040)	
0–36 months of age								
Illness in previous month	-0.076	-0.110+	-0.077	-0.087	**-0.108**	**-0.127**	0.003	0.294
	(0.052)	(0.059)	(0.047)	(0.055)	(0.064)	(0.059)	(0.040)	
Preventive health checkup	0.067	**0.346**	**0.114**	**0.359**	0.032	**0.214**	**0.169**	0.786
	(0.042)	(0.048)	(0.043)	(0.050)	(0.052)	(0.056)	(0.039)	
% RPS households dropped for common support	39	26	27	20	42	32	0	0

Notes: Nearest neighbor matching estimates of bias (RPS control – LSMS comparison) and impact (RPS treatment – LSMS comparison) for observations in the common support (Abadie et al. 2004). Columns 1–2 use the LSMS national rural, 3–4 the LSMS national high priority, and 5–6 the LSMS central region high priority samples. For each sample, a separate propensity score model was estimated to determine set of geographic- and household-level matching variables. Column 7 shows the experimental impact estimate based on the entire RPS evaluation sample, and Column 8 the average for the entire RPS evaluation control sub-sample. Expenditures measured in Nicaraguan Córdobas. Heteroskedasticity-robust standard errors (Columns 1–6) and standard errors accounting for clustering at the locality level (Column 7) shown in round parentheses (StataCorp 2007). Standard deviation shown in square brackets. + indicates significance at 10% and bold at 5%. Adjusted expenditure are expenditure less imputed housing and durable goods services.

[1] Adjusted total expenditure per capita = total expenditure per capita – housing and durable goods components.

Table 2: Gaussian kernel matching using common support: By comparison sample

	Non-experimental						Experimental	
	National rural		National high priority rural		Central Region high priority rural			
	Bias	Impact	Bias	Impact	Bias	Impact	Impact	Mean
Household level	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Total expenditure per capita	**-1027.3**	143.3	**-911.5**	262.9	-387.9	**723.4**	**1224.4**	3261.1
	(246.6)	(217.7)	(237.6)	(216.4)	(276.9)	(233.7)	(223.7)	[2598.4]
Adjusted total expend. p.c.[1]	**-931.0**	183.1	**-827.8**	272.3	-392.5	**662.2**	**1143.2**	2910.6
	(228.6)	(202.6)	(221.4)	(202.5)	(264.1)	(220.2)	(214.6)	[2400.6]
Food expenditure per capita	**-517.6**	**456.4**	**-516.2**	**483.8**	-193.2	**726.5**	**1008.9**	2166.1
	(151.8)	(137.3)	(170.1)	(159.7)	(188.8)	(160.0)	(169.8)	[1726.3]
Child level								
0–12 months								
Exclusive breast feeding	**0.307**	**0.476**	**0.231**	**0.451**	0.177	0.248+	0.134+	0.569
first 3 months	(0.090)	(0.099)	(0.113)	(0.117)	(0.118)	(0.138)	(0.077)	
Never breast fed	-0.191+	**-0.202**	-0.209+	-0.221+	**-0.207**	-0.163	0.021	0.024
	(0.099)	(0.110)	(0.115)	(0.124)	(0.090)	(0.119)	(0.026)	
0–24 months								
BCG	0.042	0.081	0.027	0.074	-0.001	0.018	0.035+	0.934
	(0.071)	(0.059)	(0.073)	(0.062)	(0.027)	(0.032)	(0.020)	
12–36 months								
Polio	0.082	**0.141**	0.062	**0.109**	0.096	0.133+	**0.041**	0.932
	(0.069)	(0.054)	(0.051)	(0.049)	(0.095)	(0.074)	(0.020)	
MMR	0.030	**0.131**	0.006	**0.093**	0.095	0.140+	**0.063**	0.880
	(0.063)	(0.050)	(0.049)	(0.047)	(0.096)	(0.079)	(0.027)	
DPT/Pentavalent	0.142+	**0.191**	0.152+	**0.178**	0.120	0.145+	0.020	0.916
	(0.077)	(0.063)	(0.082)	(0.074)	(0.099)	(0.080)	(0.022)	
Up-to-date vaccination	0.068	**0.214**	0.067	**0.172**	0.069	**0.167**	**0.083**	0.801
	(0.080)	(0.067)	(0.067)	(0.076)	(0.102)	(0.086)	(0.040)	
0–36 months of age								
Illness in previous month	0.045	0.059	0.026	0.033	0.066	0.039	0.003	0.294
	(0.062)	(0.060)	(0.068)	(0.063)	(0.058)	(0.058)	(0.040)	
Preventive health checkup	**0.164**	**0.353**	0.105	**0.319**	**0.251**	**0.388**	**0.169**	0.786
	(0.071)	(0.060)	(0.068)	(0.065)	(0.089)	(0.080)	(0.039)	
% RPS households dropped for common support	39	26	27	20	42	32	0	0

Notes: Gaussian kernel matching (with bandwidth 0.06) estimates of bias (RPS control – LSMS comparison) and impact (RPS treatment – LSMS comparison) for observations in the common support (Todd 2008). Columns 1–2 use the LSMS national rural, 3–4 the LSMS national high priority, and 5–6 the LSMS central region high priority samples. For each sample, a separate propensity score model was estimated to determine set of geographic- and household-level matching variables. Column 7 shows the experimental impact estimate based on the entire RPS evaluation sample, and Column 8 the average for the entire RPS evaluation control sub-sample. Expenditures measured in Nicaraguan Córdobas. Bootstrapped standard errors in Columns 1–6 (1000 repetitions) and standard errors accounting for clustering at the locality level in Column 7 shown in round parentheses (StataCorp 2007). Standard deviation shown in square brackets. + indicates coefficients is significant at 10% and bold indicates significant at 5%.

[1] Adjusted total expenditure per capita = total expenditure per capita – housing and durable goods components.

Table 3: Gaussian kernel matching: By propensity score model and varying common support

	Non-experimental						
	Geographic variables only		Household variables only		Geographic & household		
	Bias	Impact	Bias	Impact	Year 2000 household variables: Impact	Common Support Plus 0.02: Impact	Common Support Minus 0.02: Impact
Household level	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Total expenditure per capita	**-1058.1**	226.6	**-424.3**	**781.7**	**658.9**	**1145.0**	**531.0**
	(199.7)	(201.5)	(178.6)	(180.1)	(191.4)	(263.2)	(255.0)
Adjusted total expend. p.c.[1]	**-896.8**	313.0+	**-389.3**	**742.6**	**615.9**	**1084.49**	**469.9**
	(180.7)	(182.6)	(166.8)	(168.3)	(178.7)	(256.2)	(235.0)
Food expenditure per capita	**-387.2**	**668.2**	-100.6	**886.4**	**698.8**	**1056.5**	**583.8**
	(123.3)	(128.2)	(122.1)	(129.4)	(137.2)	(186.0)	(173.1)
Child level							
0–12 months							
Exclusive breast feeding	0.088	**0.235**	0.033	0.137	0.197	**0.330**	0.227
first 3 months	(0.091)	(0.106)	(0.098)	(0.099)	(0.123)	(0.139)	(0.157)
never breast fed	**-0.217**	**-0.236**	-0.077+	-0.044	**-0.197**	-0.168	-0.214+
	(0.076)	(0.089)	(0.043)	(0.045)	(0.099)	(0.119)	(0.123)
0–24 months							
BCG	-0.039	0.009	-0.035	-0.005	0.023	0.001	0.022
	(0.028)	(0.027)	(0.025)	(0.021)	(0.033)	(0.026)	(0.039)
12–36 months							
Polio	0.074	**0.115**	-0.002	0.040	**0.135**	0.158	0.096
	(0.054)	(0.051)	(0.031)	(0.034)	(0.061)	(0.114)	(0.061)
MMR	0.027	0.082	-0.015	0.047	0.084+	0.153	0.117
	(0.061)	(0.053)	(0.039)	(0.046)	(0.053)	(0.117)	(0.068)
DPT/Pentavalent	0.112+	**0.125**	0.041	0.062	**0.152**	0.156	0.126+
	(0.060)	(0.056)	(0.040)	(0.042)	(0.068)	(0.115)	(0.067)
Up-to-date vaccination	0.089	**0.151**	0.006	0.099+	**0.139**	0.148	**0.155**
	(0.066)	(0.062)	(0.050)	(0.055)	(0.071)	(0.118)	(0.077)
0–36 months of age							
Illness in previous month	0.030	0.045	0.001	-0.008	0.083	**0.124**	-0.002
	(0.048)	(0.049)	(0.046)	(0.050)	(0.059)	(0.049)	(0.072)
Preventive health checkup	0.058	**0.239**	0.037	**0.193**	**0.353**	**0.418**	**0.357**
	(0.046)	(0.045)	(0.043)	(0.043)	(0.065)	(0.116)	(0.076)
% RPS households dropped for common support	6	4	10	12	30	1	45

Notes: Gaussian kernel matching (with bandwidth 0.06) estimates of bias (RPS control – LSMS comparison) and impact (RPS treatment – LSMS comparison) for observations in the common support using the LSMS central region high priority sample (Todd 2008). Columns 1–2 use propensity score model with geographic variables only, 3–4 with household variables only, and 5–7 with both (showing impact estimates only). Expenditures measured in Nicaraguan Córdobas. Bootstrapped standard errors in Columns 1–7 (1000 repetitions). + indicates coefficients is significant at 10% and bold indicates significant at 5%.

[1] Adjusted total expenditure per capita = total expenditure per capita – housing and durable goods components.

Table 4: Mean budget shares by sample

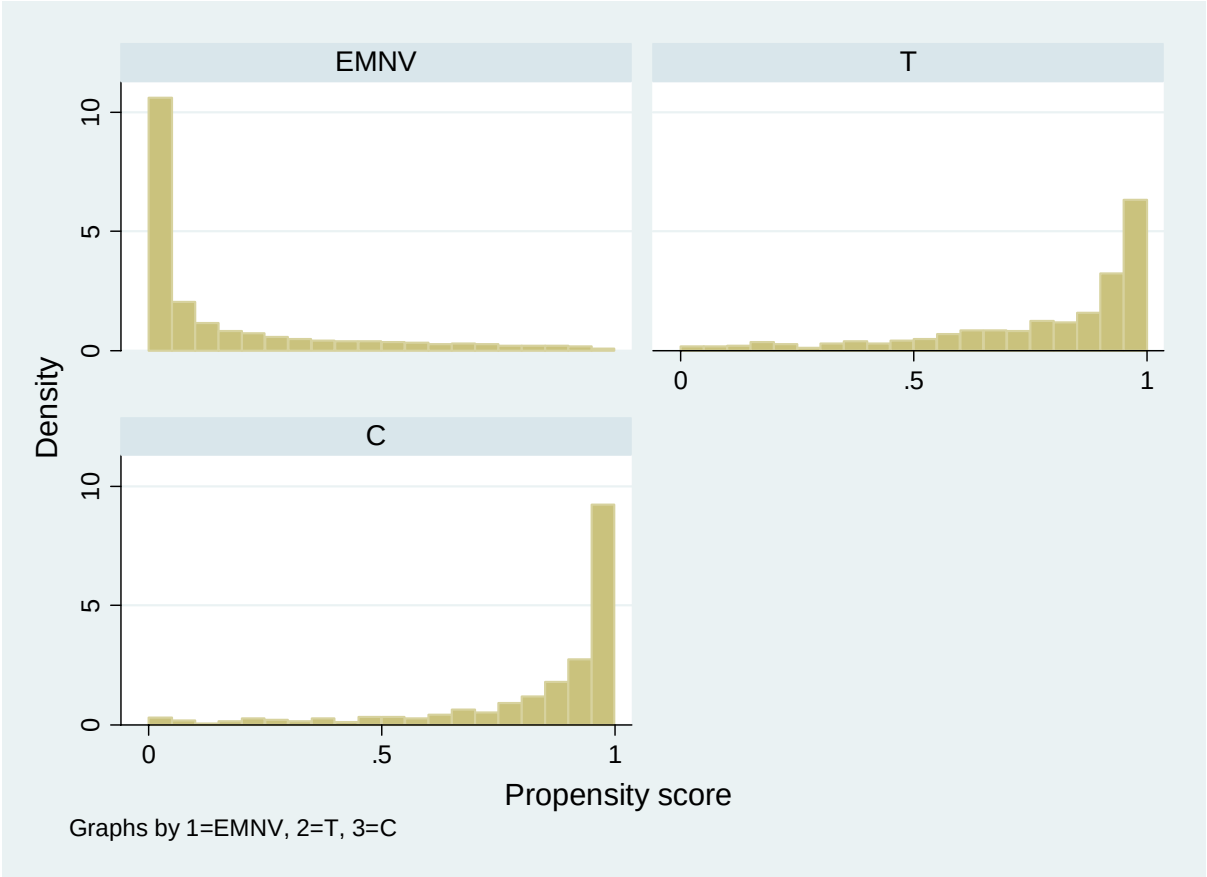
	RPS evaluation		LSMS national rural	
Expenditure item	Share	[SD]	Share	[SD]
Food	0.685	[0.13]	0.605	[0.14]
Other non-food items	0.110	[0.07]	0.143	[0.09]
Health	0.038	[0.06]	0.058	[0.09]
Education	0.025	[0.04]	0.025	[0.04]
Household utilities	0.027	[0.03]	0.053	[0.04]
Use value of housing	0.103	[0.09]	0.106	[0.08]
Use value of durables	0.006	[0.01]	0.010	[0.02]

Table 5: Gaussian kernel matching using common support: Expenditure categories

	Non-experimental		Experimental	
	Geographic & household			
	Bias	Impact	Impact	Mean
Household level per capita expenditure				
Food	-193.2	**726.5**	**1008.9**	2166.1
	(188.8)	(160.0)	(169.8)	[1726.3]
Other Non-food items	-131.3^{+}	-55.6	67.5^{+}	417.2
	(77.5)	(69.9)	(36.8)	[567.1]
Health	6.3	43.1	29.3	159.6
	(36.0)	(30.4)	(25.8)	[472.7]
Education	15.7	**43.7**	**28.0**	72.8
	(12.4)	(10.2)	(12.0)	[176.2]
Use value of housing	-4.5	54.3	**83.7**	324.3
	(37.6)	(37.8)	(25.4)	[350.3]
Household utilities	**-64.1**	**-67.1**	9.5	94.9
	(11.6)	(13.0)	(10.6)	[112.3]
Use value of durables	**9.1**	**6.9**	-2.6	26.2
	(4.6)	(3.3)	(4.4)	[70.0]
Other Non-food items	-131.3^{+}	-55.6	67.5^{+}	417.2
	(77.5)	(69.9)	(36.8)	[567.1]
% RPS households dropped for common support	42	32	0	0

Notes: Gaussian kernel matching (with bandwidth 0.06) estimates of bias (RPS control – LSMS comparison) and impact (RPS treatment – LSMS comparison) for observations in the common support using the LSMS central region high priority sample (Todd 2008). Columns 1–2 use both household- and geographic-level variables in the propensity score model. Column 3 shows the experimental impact estimate based on the entire RPS evaluation sample, and Column 4 the average for the entire RPS evaluation control sub-sample. Bootstrapped standard errors in Columns 1–2 (1000 repetitions) and standard errors accounting for clustering at the locality level in Column 3 shown in round parentheses (StataCorp 2007). Expenditures measured in Nicaraguan Córdobas. Standard deviation shown in square brackets. + indicates coefficients is significant at 10% and bold indicates significant at 5%.

Figure 1 – Histogram of propensity scores from LSMS national rural and RPS evaluation samples



Notes: Based on estimated propensity score relation shown in Table A1.

Table A1: Propensity score model estimation

Variable	Coeff.	SE
Geographic indicators (from 1995 National Population and Housing Census)		
% households in locality without piped water	0.057	(0.065)
% households in locality without latrine	-0.041	(0.030)
% over 15-year olds who are illiterate	0.779	(0.091)
Average household size in locality^2	-0.689	(0.089)
% households in locality without piped water^2	-0.001	(0.000)
% households in locality without latrine^2	-0.001	(0.000)
% over 15-year olds who are illiterate^2	-0.004	(0.001)
Average household size in locality × % households in locality without water	0.045	(0.010)
Average household size in locality × % over 15-year olds who are illiterate	-0.003	(0.011)
Average household size in locality × % households in locality without latrine	0.016	(0.005)
% households in locality without latrine × % households in locality without water	0.000	(0.000)
% over 15-year olds who are illiterate × % households in locality without water	-0.003	(0.001)
% over 15-year olds who are illiterate × % households in locality without latrine	0.000	(0.000)
Logarithm of number of households in locality	-16.227	(3.942)
Logarithm of number of population in locality	18.220	(3.753)
Fraction of household heads in locality who can read and write	-4.780	(0.403)
Logarithm of distance (km) from locality center to health center	2.149	(0.130)
Logarithm of distance (km) from locality center to primary school	-0.196	(0.050)
Logarithm of distance (time) from locality center to health center	-0.953	(0.125)
Logarithm of number of households in locality × Logarithm of number of population in locality	-0.285	(0.247)
Logarithm of household size × Number adult household members with post-secondary education	0.355	(1.140)
Logarithm of household size^2	-0.434	(0.195)
Logarithm of distance (km) from locality center to primary school × Number of bedrooms/number of persons × (1) if house walls made of brick or concrete block	-0.099	(0.110)
Demographics		
Logarithm of household size	1.391	(0.732)
Fraction of household members 0–5 years old	-1.174	(1.032)
Fraction of household members 6–17 years old	-0.584	(0.931)
Fraction of female household members 18–35 years old	0.732	(0.922)
Fraction of female household members 35–60 years old	-0.808	(0.926)
Fraction of female household members over 60 years old	-0.912	(0.843)
Fraction of male household members 18–35 years old	-0.854	(0.853)
Fraction of male household members 35–60 years old	-1.702	(0.746)
(1) if household head is female	-0.094	(0.201)
Age in years of household head	0.066	(0.030)
(Age in years of household head)^2	-0.001	(0.000)

Educational measures		
(1) if household head can read and write	-0.860	(0.229)
Completed grades of education of household head	0.245	(0.095)
(Completed grades of education of household head)^2	-0.005	(0.009)
Average completed grades of education of household members over 15 years of age	-0.124	(0.078)
Number adult household members with no education	0.023	(0.117)
Number adult household members with less than primary education	0.506	(0.114)
Number adult household members with completed primary education	0.307	(0.141)
Number adult household members some secondary education	0.705	(0.252)
Number adult household members some post-secondary education	-1.127	(2.182)
Household characteristics and assets		
(1) if dwelling is house	-0.153	(0.339)
(1) if dwelling is ranch (open walls)	0.659	(0.376)
(1) if own house	-0.519	(0.140)
Number of rooms pertaining to the household	-1.331	(0.235)
(Number of rooms pertaining to the household)^2	0.151	(0.048)
Number of bedrooms/number of persons	1.216	(0.778)
(1) if house walls made of brick or concrete block	1.051	(0.296)
(1) if house has dirt floor	0.324	(0.170)
(1) if use firewood for cooking	-0.261	(0.728)
(1) if house roof made of zinc	0.361	(0.204)
(1) if house roof made of tile	1.555	(0.230)
(1) if house has latrine	0.077	(0.144)
(1) if household drinking water source was well	-0.390	(0.168)
(1) if household drinking water source was river	-1.242	(0.179)
(1) if house has electricity	0.720	(0.181)
(1) if owned a gas/propane stove	-1.622	(0.734)
(1) if owned a fan	-1.000	(0.622)
(1) if owned a radio/tape cassette player	-0.594	(0.178)
(1) if owned a vehicle	-2.193	(1.617)
Working members in household		
Number of skilled agricultural workers	-1.405	(0.386)
Number of unskilled agricultural workers	0.165	(0.122)
Number of self-employed agricultural workers	0.781	(0.145)
Number of skilled non-agricultural workers	-0.301	(0.146)
Number of unskilled non-agricultural workers	0.687	(0.316)
Number of self-employed non-agricultural workers	-0.282	(0.199)
Number of bosses (“patron”)	-2.339	(0.508)
Number of members of cooperatives	-1.443	(1.197)
Number of unpaid family workers	-0.681	(0.122)
Constant	-56.327	(8.963)

Notes: Logit estimated on the combined RPS evaluation and LSMS national rural samples. Households in the RPS evaluation sample are given a value of one and those in the LSMS national rural sample, zero. N=3,171. Pseudo R^2=0.59. Standard errors allowing for clustering at the locality level shown in round parentheses (StataCorp 2007).