MEASURE Evaluation

Working Paper Series

Matching the gold standard: Evidence from a social experiment in Nicaragua

October 2007

WP-07-100

MEASURE Evaluation is funded by the U.S. Agency for International Development (USAID) through Cooperative Agreement No. GPO-A-00-03-00003-00 and is implemented by the Carolina Population Center at the University of North Carolina in partnership with Constella Futures, John Snow, Inc., Macro International Inc., and Tulane University.

Carolina Population Center
University of North Carolina at Chapel Hill
206 W. Franklin Street
Chapel Hill, NC 27516
Phone: 919-966-7482
Fax: 919-966-2391
measure@unc.edu
www.cpc.unc.edu/measure

Printed on recycled paper

This working paper series is made possible by support from the U.S. Agency for International Development (USAID) under Cooperative Agreement No. GPO-A-00-03-00003-00. The opinions expressed are those of the authors, and do not necessarily reflect the views of USAID or the U.S. government.

The working papers in this series are produced by MEASURE Evaluation in order to speed the dissemination of information from research studies. Most working papers currently are under review or are awaiting journal publication at a later date. Reprints of published papers are substituted for preliminary versions as they become available. The working papers are distributed as received from the authors. Adjustments are made to a standard format with little further editing.

This and previous working papers are available, free of charge, from the MEASURE Evaluation Web site, http://www.cpc.unc.edu/measure.

Matching the gold standard: Evidence from a social experiment in Nicaragua

Sudhanshu Handa
Department of Public policy, University of North Carolina, Chapel Hill, NC
shanda@email.unc.edu

John A. Maluccio
Department of Economics, Middlebury College, Middlebury, VT
john.maluccio@middlebury.edu

August 2007

Abstract: We compare non-experimental impact estimates using propensity score matching with those from a social experiment to determine whether this non-experimental approach can 'match' the gold standard. The social experiment we use was carried out to evaluate a conditional cash transfer program implemented in Nicaragua in 2000. The outcomes we assess include total and food expenditure and a variety of children's health outcomes including vaccinations, morbidity, and breast feeding. We find that PSM does better at replicating the benchmark for individual outcomes but does poorly for expenditure outcomes. Judicious choice of sample improves the performance of PSM for all outcomes. A more detailed analysis of the components of expenditures shows the degree of bias is related to the importance of the item in the household budget and persists even when differences in prices and consumption habits are controlled for by comparing households from the same geographic region. The PSM technique seems most promising for evaluating individual, and easily measured outcomes, such as those related to child schooling and health, but less so for more complex outcomes such as expenditures.

JEL Classification: I18, I38, C9
Key words: impact evaluation, propensity score matching, cash transfers, Nicaragua

---

We thank Tia Palermo for excellent research assistance, Dan Gilligan and David Guilkey for useful comments, and Juan Jose Diaz for the bootstrap programs. Funding for this research was provided by USAID's MEASURE Evaluation project at the Carolina Population Center, University of North Carolina at Chapel Hill.

1. Introduction

This article compares experimental and non-experimental estimates of the effectiveness of a Nicaraguan integrated antipoverty and human development program. We assess the non-experimental technique, propensity score matching (PSM), which has become increasingly popular as an approach to estimating program impacts in a variety of situations.[1] Assessing whether PSM-based approaches to evaluation lead to valid conclusions, and thus can substitute for social experiments, is important for several reasons. Social experiments are complex to implement, cannot be used to evaluate universal or on-going programs, and raise ethical concerns due to the denial of benefits to eligible groups.

Moreover, such experiments are costly. While full costs for evaluations are often not widely available, comparable estimates for three similar programs in Latin America that carried out rigorous experimental evaluations exist. The evidence shows that larger programs, while they spend more on evaluation, devote a smaller percentage of total costs to evaluation compared with smaller or pilot programs. For example, Mexico's Progresa program, with over 2.5 million household beneficiaries, spent 0.4% of its budget on evaluation from 1997 to 2000. Honduras's PRAF program, with approximately 50,000 households, spent about 11.7% from 1999 to 2002. The Nicaraguan pilot program we study, the *Red de Protección Social*, with approximately 10,000 households, spent just under $900,000 (8.5%) from 2000 to 2002, $600,000 (5.7%) of which was for the quantitative evaluation considered in this paper (Caldés, Coady, and Maluccio 2006). In all three cases, substantial resources were devoted to evaluation, and in the poorer, smaller countries, those resources represented a much larger proportion of the project budget.

[1] Some examples include Gilligan and Hoddinott (2006), Godtland et al. (2004), Jalan and Ravallion (2003), Levine and Painter (2003), Larsson (2003), Pradhan and Rawlings (2002), and Sianesi (2004).

Although the best-practice methodology used in this paper is not new, we provide new evidence on the feasibility of carrying out a non-experimental evaluation in a low income context, extending the literature in at least three ways. First, our study is only the second assessment of PSM in the context of a developing country and a conditional cash transfer program — the other study is for Mexico's Progresa program (Diaz and Handa 2006). Second, we study health outcomes, an area not previously assessed with these techniques but of significant policy interest. As part of their operations, both conditional cash transfer programs like the one considered here and many health programs (whether or not conditional) collect detailed administrative information on inputs and outcomes, making them naturally suited to this sort of evaluation strategy. Third, we provide a detailed analysis of the feasibility of evaluating expenditure outcomes using PSM. Outcomes, such as increased per capita expenditures or reductions in poverty, are fundamental objectives of antipoverty and human capital development programs such as conditional cash transfer programs, and their assessment requires detailed information on household consumption expenditures. Collecting expenditure information, however, is time-consuming, expensive, and the resulting data is subject to substantial measurement error (Deaton and Zaidi 2002). These issues raises concerns about the comparability of such information when collected from different surveys, as often is required for PSM. Diaz and Handa (2006) find that PSM does not work well for expenditure outcomes, but their results are based on data that is collected using different survey instruments in different populations. The analysis of expenditure outcomes in this paper, in contrast, is based on data collected using identical survey instruments but by different implementing agencies and in two different samples. Our analysis allows for a sharper test of the ability of PSM to accurately replicate experimental impact estimates.

We find that PSM accurately replicates experimental estimates of the program effects on individual health outcomes, but only when an appropriate comparison group sample is selected and

stringent common support criteria are applied to exclude outliers. After making those adjustments, however, PSM does not perform well for expenditure outcomes, despite the fact that survey instruments across the experimental and nonexperimental data sets are identical. The estimated bias for the impact on expenditures stems from the three largest components of total household expenditure: food, nonfood, and housing expenditures.

This paper is organized into sections. Section 2 briefly reviews the literature on assessments of PSM as a nonexperimental impact estimator. Section 3 describes the structure of the Nicaraguan conditional cash program we evaluate, the social experiment and associated data, and the nonexperimental data sources. Section 4 outlines the theoretical and empirical framework for the analysis. Section 5 presents the estimation of the balancing score and common support criteria, and Section 6 reports the main results. Lastly, Section 7 concludes our analysis, highlighting the policy implications of our findings.

2. Selected literature

Most of the existing literature on the performance of PSM is based on social experiments from United States employment and training programs, either voluntary programs, such as the National Supported Work Demonstration (NSW) and the National Job Training Partnership Act Study (JTPA), or mandatory programs, such as the State Welfare-to-Work Demonstrations. Voluntary interventions are those where eligible participants decide whether or not to enroll in the program. These programs are typically characterized by large pools of eligible candidates but a relatively small number of participants. The challenge of a non experimental evaluation strategy is to find non-participants in the same (or similar) labor market that look like participants. In this context, bias in the non-experimental estimation of impact arises mainly due to individual self-selection.

Using the JTPA experiment, Heckman, Ichimura, and Todd (1997; 1998) and Heckman et al. (1998) find that PSM performs well-provided researchers work with a rich set of control

variables, use the same survey instruments, and compare participants and non-participants from the same local labor market. Dehejia and Wahba (1999; 2002) use the NSW experiment combined with the CPS and PSID and show that PSM does well in replicating the experimental results. However Smith and Todd (2005) show that the results in Dehejia and Wahba are particularly sensitive to their sample restrictions, and that PSM actually exhibits considerable bias when applied to a less restrictive sample. This bias stems from differences in survey instruments as well as differences in local labor market conditions, although difference-in-difference matching is able to overcome the latter source of bias.

Mandatory programs are universal entitlements for all eligible individuals in an area. For these interventions (such as the U.S. welfare-to-work programs), the challenge for a non-experimental study is to find welfare recipients from non-participant locations similar enough to welfare recipients from participant locations. In this case, bias in non-experimental impact estimates arise mainly because of geographic differences in labor markets. This is the type of selection bias most relevant to the RPS evaluation. Assessments of PSM in this context are reported in Friedlander and Robins (1995) and Michalopoulos, Bloom, and Hill (2004), both of whom use experimental control units (or earlier cohorts) from one location as a non-experimental comparison group for treatment units in a different location. Both studies conclude that substantial biases arise when comparing recipients residing in different geographic areas, but that PSM helps in reducing differences on pre-treatment characteristics in out-of-state comparisons.

Agodini and Dynarski (2004) is one of two published studies assessing PSM outside the context of an employment or labor market program. They use data from the U.S. Federal evaluation of the targeted program of the School Dropout Demonstration Assistance Program (SDDAP) and construct comparison groups from two sources: control schools in the evaluation of the restructuring program of SDDAP (who are all out-of-state), and the National Educational Longitudinal Study

(NELS). The authors find that PSM does not perform well in replicating the experimental results for outcomes such as school dropout, self esteem, expectation of completing high school, and absenteeism. They argue that this is due to the highly voluntary nature of program participation. Furthermore, difference-in-differences PSM does not perform any better than standard cross-section matching in the final period, implying that the differencing is not sufficient to remove the selectivity bias stemming from program participation.

Diaz and Handa (2006) present the first assessment of PSM from a developing country, where social experiments are less common, evaluating an integrated antipoverty and human capital development program — Mexico's Progresa program. They use data from the Progresa social experiment and select the comparison group from a Mexican national household survey (ENIGH) using PSM. They find significant bias in the impact estimates derived from PSM compared to the benchmark from the social experiment. The degree of bias is negatively related to the degree of consistency in questionnaire design across different outcomes, with smaller bias for child employment and schooling enrollment outcomes for which the questionnaires are similar (highly consistent), and with larger bias for food expenditures which are measured differently.

The analysis that follows is most closely related to Diaz and Handa (2006) in that it investigates PSM in the context of a conditional cash transfer program in Nicaragua that was modeled after Progresa. Although an important component of both programs is health, Diaz and Handa (2006) were unable to assess health outcomes because they are not available in the Mexican ENIGH. Moreover, they suggest that their finding of significant bias for expenditures outcomes is partly due to the fact that expenditures are measured differently across survey instruments, a likely explanation. Strictly speaking, however, their approach can only provide partial evidence for that hypothesis—to confirm it, the estimates for expenditures would need to be unbiased when the survey instruments are similar or identical. In the data we use in this paper, food and total

expenditures are captured with identical survey instruments, allowing that test. Finally, the RPS and Progresa programs differ in their selection of beneficiaries, with RPS using geographic-level targeting[2] and Progresa combining geographic- and household-level targeting, including only the poorest. Consequently, in RPS, nearly all households in a chosen community are eligible for the program. If heterogeneity within localities in the program is substantial, then PSM may perform better in the RPS evaluation because it may be easier to match program participants with households from a representative national household survey.

3. Estimation framework

3.1 The evaluation problem

The usual parameter of interest in program evaluation is the (average) effect of the treatment on the treated (TT). This parameter compares the outcome of interest in the treated state (Y_1) with the outcome in the untreated state (Y_0), conditional on receiving treatment ($D = 1$). The fundamental evaluation problem is that these potential outcomes cannot be observed for any single observational unit (e.g., individual or household) in both the treated and untreated states at the same time since the observation is either in the program or not. The challenge is thus the estimation of the missing counterfactual outcome (i.e., the outcome for a treated unit had it not received treatment). Social experiments solve the evaluation problem by randomly assigning otherwise eligible individuals to a control group that does not receive the treatment. This control group provides a measure of Y_0, the outcome of the eligible individual in the untreated state. Of course, social experiments provide accurate impact estimates only if the randomization is done well so that the control group is similar in terms of both observable and unobservable characteristics to the treated group.

[2] RPS did contain a small element of household targeting in the localities studied here related to possession of a vehicle or substantial landholdings, eliminating less than 3% of the households. They used a larger component in other localities not included in this evaluation (Maluccio 2005).

3.2 Non-experimental methods and the PSM approach

As mentioned in the introduction, social experiments may not be feasible for a variety of reasons so it is important to understand whether non-experimental methods can replicate experimental methods and thus provide accurate estimates of program impact. Probably the most popular non-experimental method is multivariate regression analysis. In this approach, a national household survey is typically used to measure the mean difference in outcomes between treated and untreated units. Observed differences in relevant characteristics between treated and untreated units are controlled for by their explicit inclusion in the regression equation. Additional refinements can be done by limiting the estimation sample to the rural sample only, if, for example, the program intervention only occurs in rural areas. The weakness of the regression method is that it does not account for unobservable differences between the treated and untreated units. This is especially a concern for the analysis of voluntary programs where individual self-selection into the program is important. Differences in unobserved characteristics (such as motivation or innate ability) often affect the probability of participation as well as the outcome of interest. The second weakness of the regression method is that it assigns equal weight to all untreated observations in the estimation sample, even those that are very different from treated observations in terms of observed characteristics. For example, if the program is poverty-targeted, the regression method would assign an equal weight to untreated households in the wealthiest deciles even though they are not eligible for the program. In this scenario, the estimate of Y_0 can be very different from its true value, leading to biased estimates of impact.

The PSM technique can be seen as a refinement of the regression approach in that it selects a comparison group that most 'looks' like the treated group in terms of observed characteristics. It essentially re-weights the comparison sample to provide a better estimate of the counterfactual—what the outcome for a beneficiary individual or household would have been had it not received

program benefits. In the example of a poverty-targeted program, untreated observations from the wealthiest deciles would receive very little weight (and possibly a weight of 0 depending on the exact technique used) in the estimation of the counterfactual outcome Y_0. Note however that PSM does not address the problem of unobserved differences between treated and untreated units. For this reason, the identification assumption of PSM is that conditional on a set of observable characteristics, outcomes in the untreated state are independent of program participation. In the evaluation literature, this is known as the conditional independence assumption or the assumption of selection on observables. Practically speaking, this means that PSM is better suited to mandatory programs rather than voluntary ones.

How does PSM select the comparison group? The technique, which is developed in the seminal work of Rosenbaum and Rubin (1983), is to match treated and untreated observations based on a propensity 'score' P(X), derived from a nonlinear combination of observed characteristics X. This is operationally much more tractable than traditional or case-control matching methods since it reduces the dimensionality problem to one—treatment and comparison group units are matched on one composite score instead of a set of individual characteristics (e.g., race, sex, and age) as is traditionally done. In addition, PSM opens up the possibility of using readily available national survey data to construct P(X), which can substantially reduce the cost of an evaluation.

3.2 Application of the PSM technique

The application of the PSM is done in three steps. In the first step, the propensity score P(X) is estimated for each observation (treated and untreated) using a set of characteristics (X) which are most likely to predict participation (and eligibility) in the program. The dependent variable is '1' for treated and '0' for untreated observations, and the model is estimated via a logit or probit. It can interpreted as the probability of program participation. The set of covariates (X) used in this estimation is crucial to the success of the technique—the variables must explain program

participation as well as the outcomes under study (Heckman and Navarro-Lozano 2004). Apart from that, however, there is little guidance in how to best estimate the propensity score equation (Smith and Todd 2005), in part because standard z- or t-tests on the significance of individual parameters and goodness-of-fit measures are incorrect (Heckman and Navarro-Lozano 2004). Following Diaz and Handa (2006) and Gilligan and Hoddinott (2006), we focus on finding a set of covariates that, on theoretical and practical grounds due to the design of the program, are likely to be highly associated with program participation, as well as with the outcomes of interest (though we do not estimate directly the latter relationships). The specific variables we choose are listed and justified in Section 5.1. Based on the coefficients from the logit model we predict the probability of participation—this is the propensity score P(X).[3] Only observations (both treated and untreated) within the region of common support are used in the subsequent steps (see section 5 for discussion of common support).

In the second step, once a propensity score has been calculated for each unit in the sample, it is used to match each treated unit to a comparison group unit (within the region of common support) using either nearest neighbor or kernel matching. Nearest neighbor (NN), the most common matching algorithm, matches the treated unit to the (one) comparison group unit with the closest propensity score, selecting randomly among them in the case of ties. Kernel matching, on the other hand, matches each treatment unit to a weighted average of the outcomes of all comparison group units (within the region of common support), with higher weights given to units with propensity scores close to the treatment unit and smaller weights given to observations with propensity scores that are farther away. NN-matching is done with replacement so a comparison group unit may be used more than once as a 'match' for a treated unit.

3 Standard balancing tests are performed to ensure that within small intervals of the propensity score (deciles) both the propensity scores and the mean values of the set of covariates are balanced between participant and comparison group households.

The third and final step is to take the difference in outcome values between the treated unit and the matched comparison group unit for each treated unit, that is (Y_1-Y_0) for each treated observation. Note that Y_0 is a weighted average of all comparison group units in the case of kernel matching—the technique used here. The mean difference in the outcome values is the PSM treatment effect—the impact estimate based on the PSM technique, or the ATT parameter mentioned above.

3.3 Presentation of results: direct versus indirect estimates of bias

Our objective is to assess whether PSM can replicate the benchmark ATT from the social experiment. Thus we will compare impact estimate using PSM described above with the actual impact estimate based on the experiment. If PSM does well in replicating the experiment, these two estimates will be the same. Significant differences between the two indicate bias in the PSM method. This approach is known as the indirect estimate of bias.

An alternative way of presenting results is to compare the mean outcomes from the control group in the social experiment with the (untreated) comparison group (Diaz and Handa 2006; Smith and Todd 2005). The logic behind this approach is that the performance of PSM hinges on its ability to select a comparison group, which is similar to the experimental control group. A test of the performance of PSM thus amounts to testing for differences in mean outcomes between the experimental control group and the comparison group selected by PSM. If PSM works well, there should be no significant difference in mean outcomes between these two groups. This approach is known as a direct measure of bias. In the results below, we show both the direct and indirect estimates of bias.

4. Study Setting and Data Sources

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

In 2000, the Nicaraguan Government piloted an integrated anti-poverty and human development program, the *Red de Protección Social* (RPS), in the rural areas of the Central Region. Modeled after Mexico's Progresa program, RPS is a conditional cash transfer program that provides transfers to eligible families who fulfill specific 'co-responsibilities' or obligations. The transfers are divided into three parts: a food security transfer is provided contingent on a household member (typically the mother) attending monthly educational workshops (that cover, e.g., health, nutrition, sanitation, breastfeeding, hygiene, and diet) and children less than five attending regular preventive health checkups (that include growth monitoring and vaccinations)[4]; a school attendance transfer is given to the household provided all children ages seven to 13 who have not completed fourth grade are enrolled in school and maintain regular attendance; and a school supplies transfer is a lump sum payment for each child enrolled in school to assist in purchasing school supplies and uniforms (while this transfer varies with the number of eligible children, the school attendance transfer is a lump sum per household, regardless of the number of children). The combined transfer for a family with one child eligible for the school attendance and school supplies transfers represents approximately 21% of total average pre-program household expenditures among the target population, and is thus a significant proportion of the household's potential income. Consequently, compliance rates are well above 90%. While a voluntary program in that households are not obligated to participate formally, empirically the program more closely resembles the mandatory programs discussed in Section 2.2. It is not feasible to locate comparison households within the same localities in which the program is functioning because of geographic targeting (described below) and nearly universal take-up.

For the pilot phase of RPS, the Government of Nicaragua selected the departments of Madriz and Matagalpa from the northern part of the Central Region on the basis of poverty as well

[4] Children less than two years of age have appointments monthly and those between 2 and 5, every two months.

as their capacity to implement the program. This region was the only one that showed worsening poverty between 1998 and 2001, a period during which both urban and rural poverty rates were declining nationally (World Bank 2003). In 1998, approximately 80% of the rural population of Madriz and Matagalpa were poor, and half of those were extremely poor. In addition, these departments had easy physical access and communication (including being less than a one-day drive from the capital, Managua, where RPS is headquartered), relatively strong institutional capacity and local coordination, and reasonably good coverage of health posts and schools. By purposively targeting, RPS avoided devoting a disproportionate share of its resources to increasing the supply of educational and healthcare services to meet the increased demand for services that the program was expected to stimulate (Maluccio and Flores 2005). The program was later expanded to nearby municipalities in 2003.

Within these municipalities, 42 census *comarcas*[5] (hereafter, localities) were selected using a marginality index based on four indicators – average family size, percentage of households with piped water, percentage of households with a latrine, and average literacy rate – taken from the 1995 census. Over 95% of households in the selected localities were eligible to participate in the program and approximately 90% did so.

The evaluation design comprised a randomized, community-based intervention with measurements before and after in both treatment and control areas. The RPS evaluation sample includes 21 treatment and 21 control localities, and 42 households were selected from each locality using a stratified random sample. The data collected for the evaluation included a household panel survey (hereafter, the RPS evaluation survey) implemented in the early fall in both intervention and control areas of RPS before the start of the program in 2000, and again in 2001 and 2002 after the

[5] Census *comarcas* are administrative areas within municipalities that typically include between one and five small communities averaging 100 households each. They are comprised of census segments and determined by the National Institute of Statistics and Censuses and in some cases do not coincide with locally defined areas also referred to as *comarcas*.

program began operations. The sample size in 2001, the survey round we use in this analysis in order to match the timing of the national survey, is 1,453 households, spread evenly across treatment and control localities. The survey instrument was a comprehensive household questionnaire based on the 1998 Nicaraguan Living Standards Measurement Survey instrument (EMNV, for its acronym in Spanish, *Encuesta Nacional de Hogares sobre Medición de Nivel de Vida*). The instrument was expanded in some areas (e.g., maternal and child health and education) to ensure that all the necessary program indicators were captured, but cut in other areas (e.g., income from labor and other sources) to minimize respondent burden and ensure collection of high-quality data in a single interview. Maluccio and Flores (2005) present details on the sample design and results of the impact evaluation. We use households from both the treatment and control localities from the RPS evaluation survey.

4.2 Comparison group

The non-experimental comparison group is selected by PSM from the 2001 Nicaraguan EMNV (World Bank 2003), fielded in the summer and early fall of 2001. This multipurpose, nationally representative, household survey covered 4,191 households, approximately 42% of which exist in rural areas. One potential problem with using the 2001 EMNV is that some of the sample households may exist in areas where the RPS program is operating and therefore be RPS beneficiaries, thus leading to contamination bias when selecting a comparison group. As described above, the RPS evaluation sample was concentrated in 42 localities in six municipalities of the rural Central Region. We identify nine treatment and control localities in the evaluation sample that are also part of the EMNV survey. These are excluded from the sample used to construct a comparison group (but retained in the RPS evaluation survey data).

Another potential problem with using the 2001 EMNV is that it is a nationally representative survey, whereas RPS is a program geographically targeted to areas in extreme poverty. It may prove

difficult to find suitable matches to construct a comparison group, even after limiting only to rural households. While this may not represent the ideal circumstances for using PSM, researchers and policymakers typically confront these circumstances when devising a non-experimental evaluation strategy for such programs. Therefore, from an applied policy perspective, the use of a nationally representative survey to assess the performance of PSM for such a program is particularly informative.

4.3 Outcomes and samples

The health, nutrition, and expenditure outcomes of interest that can be calculated from information available in both the RPS evaluation and EMNV surveys include self reports on current vaccination coverage, diarrhea and (non-diarrheal) morbidity in the previous month, whether the child had had a preventive health checkup at a health clinic, breast feeding practices in early childhood, and food and total expenditures. For all of these outcomes, only growth monitoring is measured differently across the RPS evaluation and EMNV survey instruments, an important consideration as Heckman et al. (1997, 1998) and Smith and Todd (2005) conclude that the performance of PSM is sensitive to differences in questionnaire design. The reference period for a preventive health checkup is six months in the RPS evaluation survey and 12 months in the EMNV survey. This variation in questionnaire design provides an opportunity to test the relationship between questionnaire inconsistency and performance of PSM. Moreover, Diaz and Handa (2006) demonstrate very large biases for impacts on food and total expenditures, which they attribute to differences in questionnaire design and prices and regional consumption patterns. In our analysis, the expenditure modules are identical across survey instruments, which permits a cleaner test of whether expenditures can be accurately evaluated using PSM, and an indirect test of whether differences in questionnaire design led to Diaz and Handa's (2006) results.

In the analysis (e.g., estimating the balancing equations), we begin with all households from the RPS evaluation survey treatment and control localities, and consider separately three increasingly restricted sub-samples of the 2001 EMNV. Because RPS is targeted to rural areas only, the first sub-sample of the EMNV we use is the rural only sample. We refer to this as the EMNV-all rural sample, and it has 1,740 households. A second component of the geographic targeting of RPS (described in Section 3.1) was that only localities with high marginality index scores were selected for geographic-only targeting.[6] The second sub-sample of EMNV that we consider is the sample of all rural households living in localities with high marginality index scores (i.e., the poorest localities). These localities were considered the highest priority by RPS, and thus we refer to this sub-sample as the EMNV-high priority rural sample (1,316 households). Lastly, recognizing that Nicaragua has substantial regional variation (e.g., areas in the east of the country have weaker infrastructure, different prices, and large indigenous populations), we consider a third sub-sample of households only in the Central Region of the country where the program operate – EMNV-Central Region high priority rural sample (638 households). While this third sample is necessarily the smallest, it is also likely to be the most similar to the RPS evaluation survey sample.

5. Propensity score and common support

5.1 Propensity score or balancing equation

Results from the logit estimates used to derive the propensity score are presented in Table 1. All RPS households (treatment and control) are included in the estimation as well as the EMNV-all rural sample, with the former coded as '1' and the latter as '0.' We include in the X vector of household and community characteristics likely to affect both the probability of participating in the program and the various outcomes we study, but that are at the same time themselves not influenced

[6] Localities with lower scores were also enrolled in the program, but had household level targeting and were not part of the experimental evaluation.

by the program. These include a set of commonly used household level variables that describe the poverty status of the family (e.g., size and demographic composition, head's sex and schooling, dwelling characteristics, such as main material of walls, roof, and floor, toilet and kitchen facilities, access to piped water; and availability of durable goods). We also include the four locality-level indicators that were used to construct the marginality index which ultimately determined locality selection. These indicate for each locality the proportion of households without piped water or toilet, the proportion of adults in the locality that are illiterate, and average family size. These four variables were used in ordering and selecting the localities to be included in the program with geographic targeting, and are thus likely to be highly associated with program participation. Localities with lower scores (i.e., less poor) are not included in the experimental evaluation component of the research but did receive the program, though it was targeted at the household level. These census level variables proved difficult to balance, which explains why a number of transformations (quadratics and interactions) of them are included in the regression.

Although we neither formally ascribe a causal interpretation to the model nor assess significance (Heckman and Navarro-Lozano 2004), most of the coefficients shown in Table 1 have signs that are intuitive based on our understanding of the correlates of poverty and are consistent with results from the literature. Coefficients without intuitive signs, however, include the coefficients on piped water, electricity, and block wall construction, which all appear to be positively associated with program participation although they would typically be negatively associated with poverty. Since the purpose of the estimation is for matching, any such anomalies do not present any problem for our analysis.[7] Finally, the overall prediction appears to be very good, with a pseudo-R^2 of nearly 50%, though again we are unable to interpret this measure reliably.

[7] Moreover each of these characteristics is individually negatively associated with program participation in the RPS sample.

5.2 Common support

The region of overlap or common support of the balancing scores for RPS and EMNV households determines the extent to which PSM can find 'good' matches to estimate the counterfactual. Figures 1-3 present the distribution of scores by household type and sub-sample of the EMNV to assess the degree of overlap. Results for the EMNV-all rural sample are shown in Figure 1 and are based on predicted scores using the logit results shown in Table 1. The distribution of propensity scores is similar between treatment and control households within the RPS sample, but the control sample has a longer right tail (higher predicted propensity of being in the program) and a small number of outliers in the left tail (lower predicted propensity). The standard approach to constructing the common support is to retain all households in either sample with scores that are above the maximum value of the minimums of the two distributions and below the minimum value of the two maximums. This decision rule eliminates only 10% of the EMNV sample, even though the graph clearly shows that only a small proportion of EMNV households are likely to be good matches for control households. Under this regime, 2% of control households, those in the upper tail, are also eliminated. Effectively, the outliers (to the left) in the control sample allow many more EMNV households in the region of common support, even though these EMNV households are unlikely to provide good matches with households from the RPS evaluation survey. In contrast, when the same common support decision rule is imposed using treatment households from RPS, we eliminate 30% of EMNV households, the majority from the extended left tail.

Figures 2 and 3 depict similar graphs for the two restricted samples described earlier, where we have re-estimated the balancing equation in a fashion similar to that described above (but not shown) for each figure. Differences across figures arises because changing the EMNV portion of the estimating sample not only changes the figure for EMNV households but also changes the figures for treatment and control households since the estimated coefficients can differ. Limiting to

the EMNV-high priority rural localities does not change the story (in particular, it does not eliminate the outliers in the left tail of the control distribution) so that the resulting common support sample is similar to that in Figure 1. Further restricting to the EMNV-Central Region high priority rural localities does succeed in eliminating some outliers in the left tail of the control sample, though again this does not substantially affect the composition of EMNV households in the common support regime.

Table 2 provides percentiles of the distribution of balancing scores for each sample to further illustrate the divergence in distributions and the effect of the extended tails among control households. The RPS treatment and control groups have balancing scores that match quite well except for the right tail, where the 95th percentile value is much higher for control relative to treatment households. On the other hand, the scores from the three EMNV sub-samples are centered at much lower values (between -1.04 and -1.60 compared to about +0.96 for the RPS evaluation survey). The 25th percentile value for the EMNV-all rural sample is -3.82, well below the 1st percentile in either the RPS samples. Restricting to the EMNV samples to high priority rural localities within the Central Region (with better potential matches) shifts the distribution to the right. However, in these two samples, the 25th percentile is still below the 1st percentile in the two RPS samples. The medians of the RPS samples occur at about the 90th percentiles of the EMNV samples, highlighting again that the distribution of scores among EMNV households is significantly below that of the RPS households. We would expect this since the RPS targets impoverished areas and the EMNV is a nationally representative sample which is, on average, wealthier. Based on Figures 1-3 and Table 2, we are most likely to find multiple good matches in the EMNV samples for RPS households with a balancing score between -2.5 and 2.5. These cut-offs might provide a better alternative to the typical maximin and minimax common support regime. However such a

strategy has costs, since limiting the treatment sample means we are no longer estimating the average *ATT*.[8]

When estimating program impacts and direct measures of bias for child outcomes, we assign each child his or her household's propensity score and then match on these scores. Of course, not all households in the survey have children. In RPS, for example, 809 households have children age 12 or younger; in the EMNV-all rural sample, 941 have children age 12 or younger. It is possible that the distribution of scores for households with children may be different from those shown in Figures 1-3. Figure 4 presents the distribution of the balancing score for the sample of children using the EMNV-all rural sample. The RPS treatment and control samples have distributions that are similarly centered, but once again there is an extended right tail in the control sample and a few outliers (below -7) in the left tail. The region of 'thickest' overlap for households with children between the EMNV-all rural and control samples remains between -2.5 and 2.5. It turns out, then, that the restriction to households with children does not engender substantial changes in the patterns described above.

We close our discussion of common support by examining the means of selected household and community characteristics for each sample. Columns 1 and 2 of Table 3 show means for treatment and control households in the RPS evaluation survey. These should be similar if randomization was successful. At the household level, there is some slight indication that control households are worse off than treatment households, with a lower literacy rate, higher proportion of children, larger families, and more adults that are not economically active. This is consistent with the more pronounced right tail of the propensity score distribution observed in Figures 1-4. Randomization in the RPS experiment was done at the locality level after stratifying on the marginality index composed of the four, census-based variables shown at the bottom of Table 3

(described above). Like the household characteristics, these indicators are also broadly similar across the sample but suggestive that control localities were slightly less well off. The observed differences for both household and locality characteristics, however, are small, indicating that randomization was effective even if not perfect.

Columns 3-5 of Table 3 present the means for the three different EMNV sub-samples, beginning with the EMNV-all rural sample. On average, these households are markedly better off than those from the RPS evaluation sample, with higher rates of literacy, smaller family sizes, and lower proportions of children. Moreover, even households in the two more restricted samples (columns 4 and 5) do not provide average characteristics much more similar to the RPS ones. The localities that these latter households from the restricted samples come from, however, are much more similar to the RPS households. This occurs in part because the restricted EMNV samples are ones in which we have selected localities with similar marginality index scores.

The final column of Table 3 shows means for a matched sample of households based on nearest neighbor PSM between the EMNV-all rural sample and RPS control households. Using this procedure, only 342 EMNV households are matched to 666 control households, despite the fact that 90% of the 1,740 EMNV-all rural households are included in the area of common support. As expected, average household characteristics from the matched sample are much more similar to the RPS sample than the unmatched EMNV-all rural sample. For example, head's literacy is now 0.386 compared to 0.541 in the EMNV-all rural sample and 0.362 in RPS control households. Similarly, the proportion of households with a dirt floor is 0.772 in the matched sample compared to 0.684 in EMNV-all rural sample and 0.825 in the RPS control households. It is these more encouraging latter results that confirm the effectiveness of PSM in selecting households with similar characteristics and underlie the analysis below.

Figure 1. Density of log odds ratio: EMNV-all rural

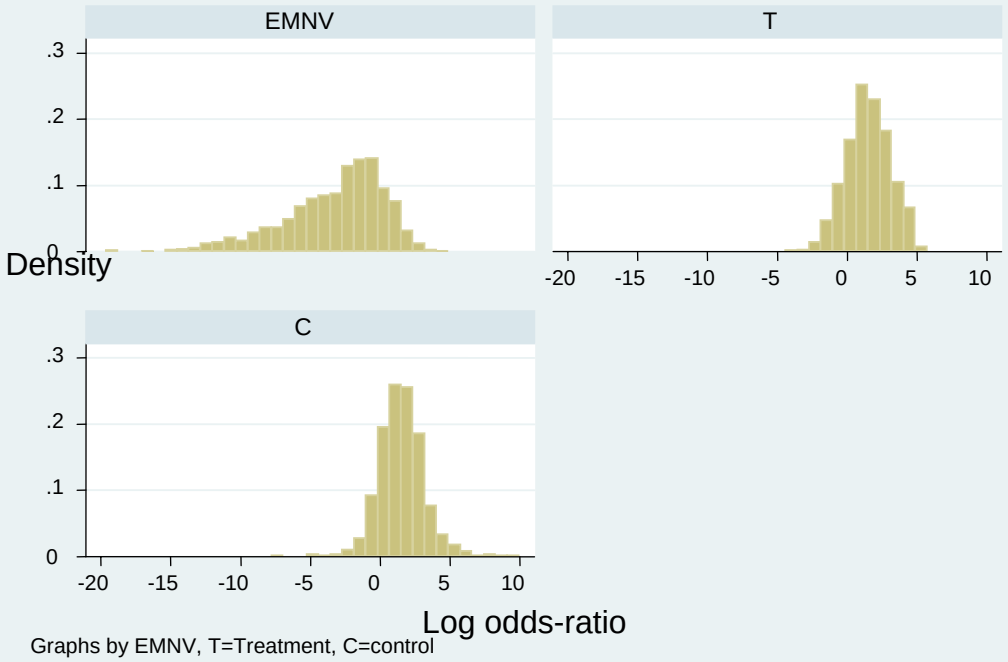


Figure 2. Density of log odds ratio: EMNV high priority rural

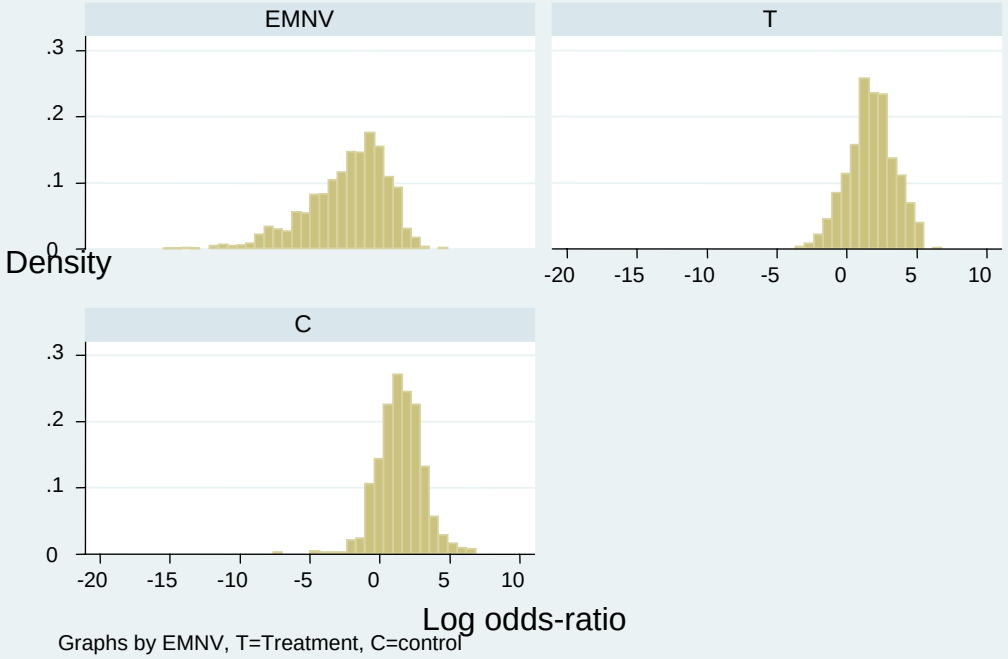


Figure 3. Density of log odds ratio: EMNV-Central Region

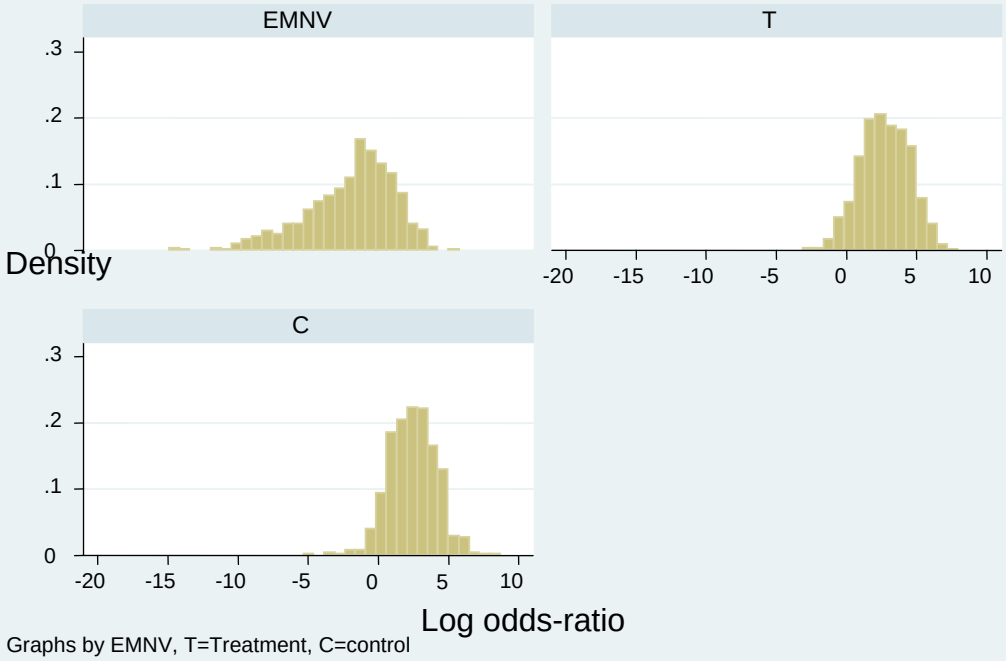
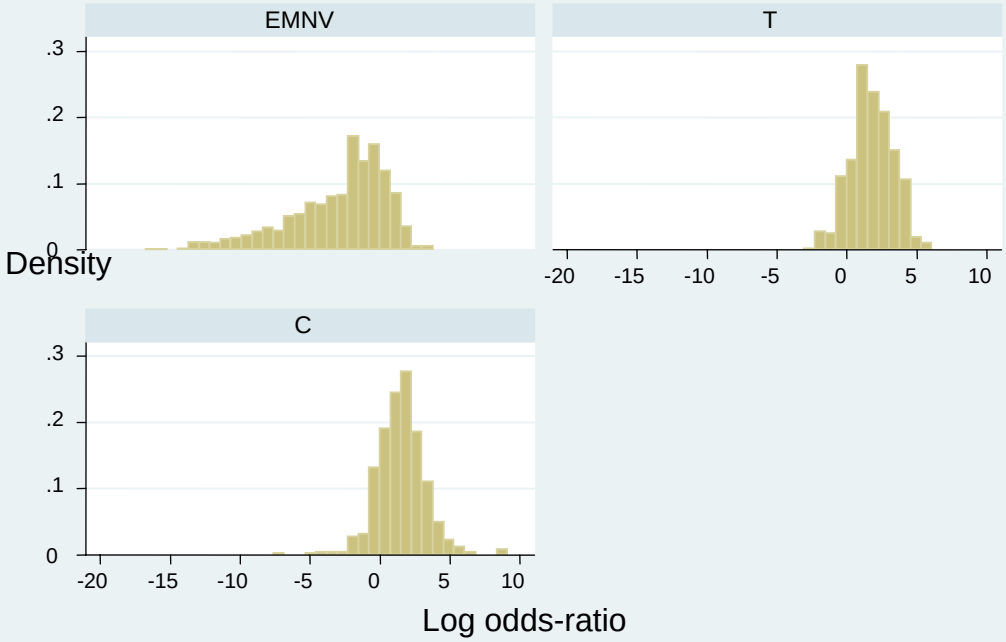


Figure 4. Density of log odds ratio: All children



6. Propensity score matching results

6.1 Base results

Table 4 presents estimates of the direct bias and the non-experimental impact using kernel matching (bandwidth=0.06) for the three different EMNV sub-samples, with common support defined by the minimax and maximin criteria.[9] Standard errors are bootstrapped,[10] and point estimates that are statistically significant from 0 (at a 5% significance level) are indicated in bold. The last two columns of the table provide the experimental impact, as well as the non-experimental impact estimated using OLS with the same set of controls that are used in the logit equation to estimate the balancing score. Our discussion of Table 4 considers for each sample the statistical significance of the direct estimates of bias, a comparison of matching versus regression adjusted non-experimental impact estimates, the contrast between the outcome measured differently (preventive health checkups) and those measured identically.

Columns 1, 3, and 5 show estimates of the direct bias (RPS control less EMNV) for the three sub-samples. PSM performs poorly when the EMNV-all rural sample is used — eight out of the 10 estimates of direct bias are statistically different from zero.[11] As we limit the EMNV to comparison groups that are on a priori grounds more similar to the RPS sample, however, performance improves. In column 5, only five-point estimates are significantly different, and one of these is for preventive health checkups which differs across the survey instruments.

Non-experimental impact estimates using PSM are shown in columns 2, 4, and 6 and can be compared to the experimental impact in column 7. The latter is estimated as a first difference across

[9] We have replicated the analysis using NN matching as well as using the more stringent common support regime of -2.5 to 2.5; these results are available from the authors upon request.

[10] Specifically, we select a sample of size (with replacement) equal to the original sample size, estimate the logit, compute the propensity score, impose common support (if applicable), and calculate the direct (or indirect) measure of bias. We repeat this one hundred times and report the standard error of the distribution of the estimate.

[11] The bias equation presented in Section 4.1 holds only on average (in expectation) so that in the results it is not always the case that the experimental estimate equals the nonexperimental estimate plus the bias.

treatment and control groups using the RPS evaluation survey data. The regression adjusted non-experimental impact presented in column 8 is the coefficient of the dummy variable for treatment estimated by OLS for the combined RPS treatment and EMNV-all rural samples. Therefore, it is most directly comparable to the non-experimental matching estimates in column 2.[12] For this sample, PSM offers only limited improvement over regression in replicating the experimental impact. Thus, neither PSM nor regression adjustment performs well in replicating the experiment in the EMNV-all rural sample.

In the case of preventive health checkups, the survey instruments differ across RPS and EMNV. The literature on PSM suggests the technique cannot overcome differences in data collection methodology and performs poorly when survey instruments are not identical. Direct estimates of bias for preventive health checkups are statistically significant in the first two samples but not in the preferred sample (high priority rural-central region in column 5).

The bias estimates for total expenditures are large even though the instruments are identical. The measure of total consumption expenditures, however, includes the imputed value of rent and services from durable goods. Not only is information on these items collected in another part of the survey outside the expenditure module, but their calculation is also subject to a number of assumptions about initial values for durable goods, depreciation rates, and the comparability of different types of housing. These calculations were done in similar fashion across the surveys, but their computation adds a significant element of noise to the expenditure calculation that comes from outside the survey instrument (use value of housing represent about 10% of total expenditure, for example). Even minor differences in how they were calculated could lead to differences across the surveys. We exclude these components from total expenditures to explore whether this improves the

[12] As discussed in section 3, regression is one of the more common non-experimental approaches used. Typically regression analysis is performed on observations from one (often nationally representative) survey; in column 8 we have combined observations from RPS treated sample with the EMNV.

performance of PSM. Results from this analysis, reported in the row entitled adjusted expenditures, show even larger estimates of bias than the original total expenditures outcome, indicating that possible differences in the calculation of the monetary value of housing and durable goods is not driving the poor performance of PSM for the total expenditures outcome. The insignificant PSM estimate of program impact for total expenditure in column 6 is worrying, because this is the theoretically 'best' sample, and actual program impacts reported in column 7 (1271.701) are quite large and statistically significant. PSM would totally miss this impact on overall household welfare.

634 Exploring the bias in expenditure outcomes

Our results show large biases for food and total consumption expenditures using the PSM technique despite the fact that the survey instrument used to collect expenditures is identical in the surveys. Household consumption expenditures is an important outcome for poverty alleviation programs such as the RPS, because total expenditures are the preferred welfare indicator for measuring poverty. Food expenditures are also an important outcome because conditional cash transfer programs like RPS emphasize the use of program money to purchase food, and supporters of such programs claim that cash transfers coupled with behavioral change are a more effective approach to poverty alleviation relative to traditional methods such as food subsidies.

Expenditure data is notoriously difficult to collect, and Living Standards Measurement Surveys such as the EMNV rely on professional interviewers who undergo extensive training and practice before entering the field. The actual expenditure module itself tends to be long in the Nicaraguan case involving 60 different food items and an additional 62 nonfood items. Respondent fatigue is a potentially serious problem, particularly as the core of the module occurs in the final section of the questionnaire. For the EMNV, this was somewhat mitigated by dividing the survey across two visits to the household, something not possible in the RPS evaluation survey due to budgetary constraints. The recall period for expenditures varies with the item, with high frequency

purchases involving a recall period of seven to 14 days and recall periods of six months or one year for low frequency and bulky purchases items. Despite these efforts, actual recall is fraught with difficulty, and measurement error can be significant. Enumerator training and supervision were also less extensive in the RPS evaluation survey than in the EMNV, though this was partly offset by using in the RPS evaluation survey interviewers who had worked previously on EMNV.

In this section, we explore in greater detail the potential sources of differences in the expenditure estimates. We disaggregate total expenditures into seven components and report budget shares for each of these components in Table 5. Among the RPS households, food and nonfood comprise nearly 80% of total spending, with an additional 10% for housing. Three components of total expenditures – housing, utilities, and use value of durables – are wholly collected from other parts of the questionnaire and not within the expenditure module itself. Two other components, health and education, are derived largely from questions in the health and schooling sections of the questionnaires, and are also primarily derived from information reported in other parts of the survey. These categories combine for about 20% of total consumption expenditures in the RPS and 25% in the rural EMNV.

The expenditure questions on housing, health, and education that occur outside the expenditure modules are the same across survey instruments except for one difference. The EMNV asks for the value of food received in school and includes this value in total food expenditure, while the RPS does not capture this one expenditure item. The use value of durable goods is also estimated identically in both data sets. In Table 6, we report direct and indirect estimates of bias for each of the seven components of expenditure reported in Table 5, including food expenditures that have been adjusted in the EMNV to exclude the value of food from school to make it exactly comparable to the RPS.

Table 6 shows that the estimate of the program effect on adjusted food expenditures (third row) performs slightly better than the original variable which contains the additional item in the EMNV not found in the RPS (value of food consumed in school — first row). In each of the three samples, the bias estimate for this outcome is smaller, and the resulting non-experimental impact estimate closer to the experimental estimate in column 7.

Examining each of the columns reporting the direct bias estimates (columns 1, 3, and 5) allows us to assess how the bias in total expenditures is distributed among the seven expenditure components. Most of the bias comes from the food and nonfood components of total expenditures, all of which are collected directly from the expenditure model and make up over 60% of total expenditures (see Table 5). The percent of bias accounted for by these two components ranges from 58% in column 1 to 69% in column 3. The remaining bias stems from the housing components (about 30%) and health expenditures (10% of the bias).

The allocation of the bias across components of expenditures is proportional to the budget share of that component in all cases except for health spending. This component is only 4% of the budget but it accounts for 10% of the bias in the impact estimate. This section of the questionnaire is particularly complex. First, the respondent is asked if s/he was sick. If the answer is yes, then a series of questions are asked about illness and treatment received, and an assessment of costs paid, usually broken into components associated with transportation, medical fees, and medicine costs, is conducted. Clearly there is substantial room for error here. In all cases where the bias in a sub-component of total expenditures is statistically significant, the point estimate is negative, implying that reported expenditures are higher in the matched EMNV sample relative to the RPS. This is consistent with the idea that the EMNV data were collected using more experienced professional staff who are better able to collect information over two visits instead of one, especially in a time-consuming and difficult area such as expenditures.

7. Discussion and conclusion

How badly would we be mistaken if the RPS was evaluated using PSM? Our best non-experimental program estimates are in column 6 of Table 4. The experimental results (column 7) show positive impacts in four areas – food and total expenditures, preventive health checkups, and MMR vaccinations – and negative impacts in one case (reported diarrhea). The non-experimental results in Table 5 column 6 are 'correct' for food expenditures, preventive health checkups and MMR, but show no impact for the remaining two indicators. The non-experimental results would therefore slightly overstate the positive impact of the program, mostly by not identifying the higher incidence of reported diarrhea among program beneficiaries. This finding, however, is likely related to increased awareness of mothers in the program who bring their children to the regular checkups, receive counseling during those sessions, and attend the healthcare workshops which emphasize themes related to childcare. Even though the sign and significance are correct, however, the magnitude of the positive impacts on food and preventive health checkup rates are substantially smaller than the experimental estimates, while for MMR they are actually substantially higher.

Taken as a whole, our results indicate that PSM must be implemented with extreme care to interpret the results confidently, as indicative of true program impact. This is especially true since it is rare that one has an experimental evaluation with which to benchmark the results. The key feature that seems to be especially important in ensuring results that are consistent with the true impact estimates is the choice of the comparison sample.

An important precondition highlighted in the previous literature, the alignment of survey instruments, appears not to affect the qualitative results in one special case. The recall period for preventive health checkups is longer in the EMNV relative to the RPS, yet PSM delivers qualitatively similar results to the social experiment for this outcome.

The performance of PSM for evaluating expenditure outcomes, however, remains a major concern. Handa and Diaz (2006), in a similar assessment to this one based on the Mexican Progresa experiment, showed that PSM was not capable of replicating the experimental results for food and total expenditures. In that case, the survey instruments for both food and nonfood differed substantially. In the RPS case, the food and total expenditure modules in the EMNV and the social experiment are identical, after one minor adjustment. The PSM non-experimental impact estimates for expenditures are positive and statistically significant, as are results from the experiment. However, the point estimates for the non-experimental estimators are about three-quarters those from the experiment for food, and less than half (and insignificant) for total expenditures. It is unlikely that variations in prices and consumption patterns explain the entire difference, because these have been partially controlled for by limiting the comparison group to households in the same region as those from the experiment. Consequently, an important policy implication of this work is that PSM techniques that seek to evaluate expenditures may still be substantially 'wrong' in a quantitative sense, even after controlling for differences in prices and consumption patterns.

The more detailed analysis of the bias in expenditures reveals that the bias is roughly proportional to the overall importance of the component in total expenditures and not concentrated in any single components. For example, food and nonfood account for about 80% of total expenditures and over 60% of the bias in total expenditure. In general, we found substantial bias on the impact on expenditure items reported within the lengthy expenditures module as well as among components of expenditures that are reported in other sections of the survey (housing and health). Our supposition is that different field methodologies can lead to different measurement errors, with the possibility that expenditures are systematically underreported in the RPS evaluation survey. These differences in measurement errors lead to the failure of PSM to estimate program effects accurately. Having identical survey instruments is not enough, at least in the case of the

measurement of complex items such as expenditures. Identical or very similar implementation of the field work also may be necessary.

Based on accumulated information from this assessment and the one from Mexico's Progresa program, the PSM technique seems most promising for evaluating individual and easily measured outcomes, such as those related to child schooling and health. These outcomes are less subject to the potential problems associated with collecting expenditure information (recall error, respondent fatigue) making field implementation a less important concern. Nor are they affected by relative price differences or consumption patterns in the same way that food expenditures are, and thus, are more likely to deliver reliable estimates of true program impact. Moreover, the interviewer teams used in social experiments in developing countries are possibly not as well trained as their full-time professional counterparts in the institute's of statistics who are responsible for collecting national expenditure surveys, which may also affect the quality of the expenditure data obtained from these experiments. Basic information on health and schooling are much easier to collect, and thus, more likely to be of comparable quality to similar information from a national household survey. This raises the question of whether a household survey is even necessary to evaluate accurately the impact of a conditional cash transfer or other human capital oriented social program on individual outcomes, such as vaccinations, preventive health checkup rates, and school enrollment. It may be cheaper to invest in stronger monitoring staff and systems that can accurately track information on school enrollment, health visits, and immunizations and compare these to household survey data using PSM. However, to the extent that poverty and food expenditures are important outcomes that require evaluation, the evidence to date suggests that a non-experimental approach using PSM is likely to significantly underestimate program impacts.

8. References

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.

Caldés, Natàlia, David Coady, and John A. Maluccio. 2006. “The cost of poverty alleviation transfer programs: A comparative analysis of three programs in Latin America.” *World Development* 34(5): 818–37.

Deaton A., and S. Zaidi. 2002. Guidelines for constructing consumption aggregates for welfare analysis. LSMS Working Paper Number 135, The World Bank, Washington, D.C.

Dehejia, Rajeev and Sadek Wahba. 1999. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association* 94: 1053–62.

Dehejia, Rajeev and Sadek Wahba. 2002. “Propensity Score Matching Methods for Nonexperimental Causal Studies.” *Review of Economics and Statistics* 84: 151–61.

Diaz, Juan José and Sudhanshu Handa. 2006. “An Assessment of Propensity Score Matching as a Nonexperimental 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 Nonexperimental Methods.” *American Economic Review* 85: 923–37.

Gilligan, Dan and John Hoddinott. 2006. 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* (forthcoming).

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. 1998 “Matching as an Econometric Evaluation Estimator.” *Review of Economic Studies* 65(2): 261–94.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. “Characterizing Selection Bias Using Experimental Data.” *Econometrica* 66: 1017–89.

Heckman, James, Salvador Novarro-Lozano. 2004 “Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models.” *Review of Economics and Statistics* 86: 30–57.

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.

Larsson, Laura. 2003. “Evaluation of Swedish Youth Labor Market Programs.” *Journal of Human Resources* 38(4): 891–927.

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, J.A. 2005. Household targeting in practice: The Nicaraguan *Red de Protección Social*, Food Consumption and Nutrition Division, IFPRI, Washington D.C., Photocopy.

Maluccio, John A. and Rafael Flores. 2005. “Impact evaluation of the pilot phase of the Nicaraguan *Red de Protección Social*.” Research Report No. 141, International Food Policy Research Institute, Washington D.C.

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): 275–295.

Rosenbaum, Paul and Donald Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrica* 70: 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.

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

World Bank. 2003. Nicaragua poverty assessment: Raising welfare and reducing vulnerability, Report No. 26128-NI. Washington DC: The World Bank.

Table 1: Logit coefficients for propensity (balancing) score estimation

	Coefficient	Std. Err	z-statistic
Stove	-1.476	0.496	-2.98
Vehicle	-1.708	1.341	-1.27
Household head is literate	-0.415	0.114	-3.65
fracchild0_5	0.703	0.356	1.97
fracchild~12	0.610	0.351	1.74
fracchild~17	1.619	0.436	3.71
Log(household size)	0.471	0.200	2.36
Female head	-0.314	0.150	-2.09
Ln(distance to health center)	1.265	0.072	17.62
Ln(distance to primary school)	0.080	0.051	1.57
Census mean household size	5.110	1.109	4.61
Census proportion w/o piped water	-0.108	0.050	-2.15
Census proportion w/o toilet	-0.063	0.027	-2.36
Census proportion adults illiterate	0.837	0.083	10.11
Census: hhld size*water	0.047	0.008	5.58
Census: hhld size*illiteracy	-0.028	0.009	-3.01
Census: hhld size*toilet	0.023	0.005	4.96
Census: water*toilet	0.000	0.000	-1.09
Census: water*illiteracy	-0.003	0.001	-4.43
Census: toilet*illiteracy	0.000	0.000	0.29
Census: hhld size squared	-0.800	0.100	-8.03
Census: water squared	0.000	0.000	0.22
Census: toilet squared	0.000	0.000	-4.72
Census: illiteracy squared	-0.003	0.000	-7.37
Number of rooms	-0.967	0.171	-5.66
Rooms squared	0.092	0.028	3.24
Bedrooms per capita	2.787	0.794	3.51
Bedrooms per capita squared	-0.873	0.383	-2.28
Block wall	0.766	0.162	4.74
Dirt floor	0.580	0.144	4.02
Own house	-0.781	0.120	-6.53
Electricity	0.875	0.156	5.61
Pipe in yard	0.940	0.178	5.28
No. Adults agricultural obreros	-1.610	0.364	-4.42
No. adults agricultural peones	0.185	0.087	2.13
No. adults agricultura self-employed	0.786	0.108	7.31
No adults nonagric. Obreros	0.415	0.284	1.46
No adults patrones	-2.992	0.474	-6.31
No of adults not economically active	-0.479	0.086	-5.57
Constant	-41.949	4.561	-9.20

Logit estimated over the EMNV-all rural sample and RPS evaluation sample.
Households in the RPS evaluation sample are given a 1 and EMNV households 0.
Total sample is 3143; Psuedo R-squared is 0.48 and the log likelihood is -1122.72.

Table 2: Distribution of balancing score by sample

	RPS			EMNV	
Percentile	Treated	Control	All rural	High priority rural	High priority rural in Central Region
1	-2.48	-2.22	-13.45	-7.80	-6.93
5	-1.18	-1.00	-9.38	-5.65	-5.62
10	-0.63	-0.52	-7.08	-4.29	-4.15
25	0.15	0.15	-3.82	-2.65	-2.68
50	0.97	0.95	-1.60	-1.04	-1.04
75	1.73	1.71	-0.13	0.10	0.16
90	2.39	2.47	0.70	0.82	0.91
95	2.85	3.38	1.20	1.31	1.39
99	3.40	5.76	2.14	1.31	2.73

Table 3: Means of selected household characteristics by sample

	RPS			EMNV		
	T	C	All rural	High priority rural	High priority in Central Region	Matched Sample
Head is literate	0.374	0.362	0.541	0.518	0.525	0.386
Head is female	0.132	0.148	0.176	0.168	0.172	0.170
proportion of children 0-5	0.194	0.212	0.162	0.170	0.165	0.176
proportion of children 6-12	0.206	0.215	0.191	0.195	0.192	0.207
proportion of children 13-17	0.126	0.127	0.110	0.109	0.107	0.119
family size	5.950	6.078	5.868	5.999	5.939	5.904
Cocina	0.004	0.009	0.083	0.046	0.038	0.009
Dirt floor	0.825	0.825	0.684	0.691	0.754	0.772
Crowding: bedrooms per person	0.280	0.267	0.304	0.298	0.304	0.301
# residents age>14 economically inactive	0.254	0.288	0.440	0.495	0.513	0.295
Community means from census						
family size	5.766	5.885	5.619	5.695	5.758	5.789
Households without piped water (%)	92.127	91.761	84.838	96.595	94.741	92.809
Households without toilet (%)	54.915	59.731	46.091	52.809	53.351	55.071
Percentage of adults illiterate	55.253	56.413	46.110	50.079	50.784	53.953
N	727	676	1740	1316	638	342

Table 4: Matching estimates by sample: Gaussian kernel with common support

	All rural		High priority rural		High priority rural in Central Region			
Outcomes	Direct bias	Nonexperimental impact	Direct bias	Nonexperimental impact	Direct bias	Nonexperimental impact	Experimental impact	Regression adjusted impact
Household level	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
food expenditure	**-370.462**	**681.327**	**-408.854**	**614.33**	**-330.267**	**677.658**	**1050.055**	**607.692**
	(87.96)	(115.60)	(101.46)	(149.89)	(115.69)	(192.26)	(108.25)	(90.77)
total expenditure	**-868.462**	**375.32**	**-827.808**	**389.238**	**-764.541**	457.221	**1271.701**	**379.568**
	(125.90)	(180.84)	(140.26)	(194.11)	(177.78)	(248.66)	(148.57)	(136.26)
adjusted total expenditure	**-1084.19**	-19.204	**-1112.258**	-91.824	**-1052.62**	-38.317	**1067.478**	-102.114
	(122.63)	(155.16)	(131.39)	(185.72)	(169.74)	(234.57)	(129.13)	(117.83)
Child level								
diarrhea last month	**-0.084**	0.001	-0.053	0.026	**-0.113**	-0.036	**0.092**	0.018
	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.03)	(0.03)
sickness last month	-0.006	-0.006	-0.005	-0.031	-0.05	-0.089	-0.008	-0.031
	(0.04)	(0.05)	(0.04)	(0.05)	(0.05)	(0.06)	(0.03)	(0.03)
0-36 months								
check up	**0.137**	**0.279**	0.091	**0.252**	0.001	**0.159**	**0.170**	**-0.277**
	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.08)	(0.02)	(0.03)
0-12 months								
Exclusive breast feeding	**0.171**	**0.277**	0.05	0.1	-0.131	-0.084	0.088	**0.197**
first 3 months	(0.05)	(0.10)	(0.07)	(0.11)	(0.07)	(0.08)	(0.06)	(0.07)
never breast fed	**-0.211**	**-0.159**	**-0.195**	-0.104	**-0.131**	-0.067	0.021	**-0.101**
	(0.04)	(0.08)	(0.04)	(0.09)	(0.04)	(0.04)	(0.03)	(0.04)
12-36 months								
DPT/Pentavalent	**0.111**	**0.17**	**0.113**	**0.131**	0.057	0.061	0.017	**0.178**
	(0.05)	(0.06)	(0.04)	(0.05)	(0.06)	(0.07)	(0.02)	(0.03)
MMR	0.069	**0.25**	0.059	**0.179**	0.11	**0.206**	**0.154**	**0.227**
	(0.06)	(0.07)	(0.05)	(0.07)	(0.08)	(0.09)	(0.03)	(0.04)

Column 7 shows the experimental impact estimate. Column 8 shows the nonexperimental impact estimate using OLS regression and is the coefficient of the treatment dummy. Standard errors in parentheses below point estimates.

Table 5: Mean budget shares by sample

	RPS		Rural EMNV	
Item	Share	St. Dev.	Share	St. Dev.
food	0.685	(0.13)	0.605	(0.14)
nonfood	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)
use value of housing	0.103	(0.09)	0.106	(0.08)
household utilities	0.027	(0.03)	0.053	(0.04)
use value of durables	0.006	(0.01)	0.010	(0.02)

Table 6: Matching estimates for expenditure components by sample: Gaussian kernel with common support

	All rural		High priority rural		High priority rural in Central Region			
		Nonexperimental		Nonexperimental		Nonexperimental	Experimental	Regression
Outcomes	Direct bias	Impact	Direct bias	impact	Direct bias	impact	impact	adjusted impact
Household level	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
food expenditure	**-370.462**	**681.327**	**-408.854**	**614.33**	**-330.267**	**677.658**	**1050.055**	**607.692**
	(87.96)	(115.60)	(101.46)	(149.89)	(115.69)	(192.26)	(108.25)	(90.77)
total expenditure	**-868.462**	**375.32**	**-827.808**	**389.238**	**-764.541**	457.221	**1271.701**	**379.568**
	(125.90)	(180.84)	(140.26)	(194.11)	(177.78)	(248.66)	(148.57)	(136.26)
adjusted food expenditure	**-347.691**	**704.403**	**-402.697**	**621.781**	**-302.062**	**700.35**	**1016.259**	**669.695**
	(98.47)	(111.93)	(94.66)	(150.80)	(125.40)	(175.11)	(109.00)	(119.44)
non food expenditures	**-134.305**	-61.205	**-166.261**	**-85.714**	**-179.489**	**-102.861**	**74.72**	**-131.882**
	(36.15)	(45.39)	(35.45)	(37.61)	(51.15)	(52.18)	(30.77)	(44.25)
Health	**-96.894**	**-70.808**	**-78.132**	**-70.754**	**-73.961**	-44.787	22.944	-9.349
	(27.69)	(35.19)	(33.90)	(33.25)	(35.42)	(37.59)	(24.58)	(41.10)
Education	6.388	**33.686**	8.011	**33.961**	7.235	**34.361**	**25.02**	15.992
	(8.02)	(7.55)	(8.15)	(8.89)	(11.71)	(12.46)	(8.60)	(11.45)
Use value of housing	**-167.024**	-105.112	**-106.571**	-20.39	**-105.119**	-9.758	**84.372**	-15.117
	(39.28)	(61.88)	(23.57)	(27.55)	(25.21)	(27.53)	(22.85)	(30.05)
Household utilities	**-111.962**	**-103.951**	**-106.654**	**-106.031**	**-113.719**	**-118.805**	5.4	**-74.94**
	(10.08)	(12.97)	(10.31)	(13.69)	(13.31)	(13.81)	(7.35)	(11.22)
Use value of durables	3.53	0.476	2.751	-0.092	**6.139**	2.449	-3.247	10.853
	(3.56)	(2.51)	(3.08)	(2.92)	(2.67)	(3.41)	(3.45)	(7.49)

First 4 rows are taken from Table 5. Column 7 shows the experimental impact estimate. Column 8 shows the nonexperimental impact estimate using OLS regression and is the coefficient of the treatment dummy. Adjusted food expenditure subtracts the value of food received at school from the EMNV. Standard errors in parentheses below point estimates.

Appendix 1: The Evaluation Problem and the PSM Technique

The usual parameter of interest in program evaluation is the (average) effect of the treatment on the treated (TT). This parameter compares the outcome of interest in the treated state (Y_1) with the outcome in the untreated state (Y_0), conditional on receiving treatment ($D = 1$). Since these potential outcomes cannot be observed for any single observational unit (e.g., individual or household) in both the treated and untreated states at the same time — the evaluation problem —the estimation of the missing counterfactual outcome (i.e., the outcome for a treated unit had it not received treatment) is needed for identification of TT. PSM is a non-parametric estimation method that re-weights the comparison sample to provide an estimate of the counterfactual of interest — what the outcome for a beneficiary individual or household would have been had they not received program benefits. The identification assumption of PSM is that, conditional on a set of observable characteristics, outcomes in the untreated state are independent of program participation. In the evaluation literature, this is known as the conditional independence assumption or the assumption of selection on observables.

The attractiveness of PSM is in large part due to the seminal result of Rosenbaum and Rubin (1983) who show that if the conditional independence assumption holds for a set of covariates X, then it also holds for P(X), the propensity score derived from a nonlinear combination of the components of X. This is operationally much more tractable than traditional or case-control matching methods since it reduces the dimensionality problem to one — treatment and comparison group units are matched on one composite score instead of a set of individual characteristics (e.g., race, sex, and age) as is traditionally done. In addition, PSM opens up the possibility of using readily available national survey data to construct P(X), which can substantially reduce the cost of an evaluation.

Denoting by X the set of observables, the identification assumption can be expressed as $Y_0 \perp D \mid P(X)$ where the symbol $\perp$ denotes independence and $P(X)$ is the propensity score. To identify the treatment parameter, we require a slightly weaker assumption known as conditional mean independence: $E(Y_0 \mid D = 1, P(X)) = E(Y_0 \mid D = 0, P(X))$. By conditioning on $P(X)$, we can estimate the unobserved component of TT. In particular, we identify the parameter as follows:

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

A common approach used to assess PSM is to use the above to compute a (direct) measure of the bias associated with TT, instead of computing the parameter itself (Diaz and Handa 2006; Smith and Todd 2005). This is done by comparing control units from the experimental data (in our case, these will be from the control group of the RPS evaluation survey data) with non-experimental comparison units (from the EMNV sub-samples). The logic behind this approach is that the performance of PSM hinges on its ability to select a comparison group which is similar to the experimental control group. A test of the performance of PSM thus amounts to testing for differences in mean outcomes between the experimental control group and the comparison group selected by PSM. This difference, the expected bias in the PSM estimator, can be written as:

$$B(X) = \underbrace{E(Y_0 \mid D = 1, P(X))}_{\text{Experimental Controls}} - \underbrace{E(Y_0 \mid D = 0, P(X))}_{\text{Matched Nonexperimental Comparisons}} .$$

Since experimental control units did not receive any treatment, when the estimated bias is zero, it demonstrates that PSM will perform well. Any statistically significant difference from zero, however, can be interpreted as statistically significant bias (Smith and Todd 2006).