

An Assessment of Propensity Score Matching as a Non Experimental Impact Estimator: Evidence from a Mexican Poverty Program

Juan Jose Diaz & Sudhanshu Handa¹

This version: April 2004

Abstract:

Social experiments have become the benchmark method for estimating the program impacts of policy interventions, but are not always available. Non experimental evaluation methods provide an alternative to the experimental design, but their results depend on stronger non-testable assumptions and therefore are less clear and more controversial. We present evidence on the reliability of a non experimental method, propensity score matching, in the estimation of program impacts. We apply propensity score matching to estimate the bias associated with the treatment on the treated effect of a Mexican anti-poverty program on outcomes such as food expenditure and its composition, and children's schooling and employment. We compare the results of the experimental impact evaluation of the program with those using matched samples drawn from a nationally representative household survey. Our results show that simple cross-sectional matching does well in replicating the benchmark for outcomes that are measured using similar survey instruments. However, for outcomes that are measured using different survey instruments we find significant differences between the benchmark and the results based on matching.

¹ Juan Jose Diaz is a graduate student in the Department of Economics, University of Maryland. Sudhanshu Handa is Associate Professor in the Department of Public Policy at the University of North Carolina at Chapel Hill. Address for correspondence: S. Handa, Department of Public Policy, CB#3435, University of North Carolina, Chapel Hill, NC 27599-3435. Tel: 919.843.0350; email: shanda@email.unc.edu. This paper could not have been written without the assistance of Monica Orozco of PROGRESA, who provided essential data and explained operational details of the program to us. We also thank Jeffrey Smith for detailed comments on earlier drafts, and seminar participants at the University of Maryland, the Latin American & Caribbean Economics Association (LACEA) Annual Meetings in Mexico, the Carolina Population Center and the Group for the Analysis of Development (GRADE-Lima) for constructive criticism.

1. Introduction

Social experiments are the benchmark method for estimating the impact of a program. However social experiments are usually not available for a variety of reasons such as high political and monetary costs, the inability to implement experiments for universal entitlements or on-going programs, and because the use of control groups can raise ethical concerns. Consequently, testing the reliability of non experimental methods is a central issue in the program evaluation literature. Non experimental methods typically identify program impacts by imposing non testable assumptions; randomized experiments, when available, can be used to assess the validity of those assumptions and the performance of alternative techniques of impact evaluation.

This study contributes to the small but growing literature on the performance of one particular type of non experimental technique, propensity score matching (PSM), as an impact estimator, using a unique data set from a Mexican social experiment designed to evaluate that country's new poverty program--PROGRESA. PROGRESA is a conditional cash transfer program, which was originally targeted to poor rural households. Eligible households receive benefits provided they enroll their children in school, send them for health check-ups and at least one adult attends a monthly health talk. The program has national coverage and is mandatory; all households in participant localities that satisfy program eligibility rules and comply with its requirements receive treatment.

PROGRESA expanded in phases beginning in late 1997, and by 2000 the program had incorporated 72,345 rural localities in all 31 states around the country covering approximately 2.6 million households. To evaluate the impacts of the program a randomized experiment was carried out during the second phase of incorporation. The evaluation data consist of four rounds of household surveys applied to residents in 506 program-eligible localities/villages across 6 Mexican states. Approximately one-third of these localities were randomly selected for delayed entry into the program, and thus served as the randomized-out ``control'' group for the impact evaluation.

We exploit the availability of experimental data from this social experiment to empirically assess the performance of several propensity score matching techniques in

producing a comparable sample of households with which to accurately evaluate the impacts of the program. We use a national household survey on income and expenditure carried out by the Mexican National Statistical Institute to identify poor households in localities similar to those in the experiment. We use a variety of matching methods to select a comparison group that resembles treatment units in the absence of program intervention. We then combine this non experimental comparison group with control the group from the experiment and compare the two groups across a range of outcomes related to the intervention in order to estimate the potential bias that arises when estimating program impacts using matching methods.

The results of this paper contribute to the existing literature in several ways. First, all the published research on the reliability of PSM as an impact estimator is based on employment and training programs inside the U.S.—ours is the first study to extend the evidence outside the U.S. and beyond employment programs. Our assessment of PSM based on a cash transfer poverty program is particularly valuable because at least five other countries in Latin America and the Caribbean have begun implementing programs similar to PROGRESA, and in several cases are in the process of setting up expensive social experiments to measure program impact without a clear idea of what the additional cost of these experiments buys in terms of accuracy. Second, we employ and compare a range of matching techniques including kernel and local-linear matching. Third, we are able to compare the bias in outcomes that are collected through different survey instruments, thus providing evidence on the importance of questionnaire versus other sources of bias.

Our main results are that PSM does not perform well in replicating the benchmark for outcomes that are collected using different survey instruments, but does perform well when outcomes are measured comparably. We find no systematic difference in results across matching techniques for outcomes measured differently, but some difference for outcomes that are measured the same way, such as child employment and schooling. The rest of the paper is organized as follows: section 2 provides a summary for the state of the literature in this field; section 3 describes the PROGRESA program; section 4 presents our methodology; sections 5 and 6 describe the data and main results, and section 7 summarizes the results and their implications.

2. Background and Selected Literature

A major challenge in estimating the impacts of policy interventions in the evaluation and treatment effects literature is to measure the outcome of interest in the counterfactual state. Given that potential outcomes cannot be observed for any single observational unit in all of the counterfactual states, the essence of an identification strategy is the estimation of missing counterfactual outcomes. Social experiments, where a group of program eligible units (individuals, households, localities, etc.) are randomly excluded from the treatment or intervention, provide the cleanest estimate of the counterfactual outcome, and have become the benchmark to evaluate policy interventions. Randomized social experiments are not a panacea however; they provide a consistent impact estimator when the experiment does not distort the environment in its absence (randomization bias), when there are no displacement effects, no substitution bias, and no drop-out bias. Additionally, a potential drawback of social experiments is that they may be too costly to implement in some contexts and may raise ethical concerns regarding the denial of treatment for randomized-out units. However, when applied correctly, the consensus among researchers is that this method produces the most accurate estimate of program impacts.

When experiments are not available, researchers have to rely on non experimental methods to overcome selection bias problems in the estimation of program impacts. Many statistical and econometric models have been developed to control for confounding variables and selectivity issues. These techniques require imposing assumptions which are non-testable, although many of their implications might be, and may or may not be tenable in actual data. Non experimental methods may produce substantial biases because of self-selection, environment differences stemming from local labor markets, and differences in data sources and quality. From this standpoint an important issue is to assess, when possible, whether non experimental methods are good substitutes for randomized experiments.

During the past three decades many federal and state sponsored programs in the U.S. have been evaluated using the experimental approach. These randomized evaluations had been used in several studies to assess the performance of non experimental methods, because they provide a suitable benchmark. Most of the

interventions have been employment and training programs, either voluntary programs such as the National Supported Work Demonstration (NSW), the AFDC Homemaker-House Health Aide Demonstration, and the National Job Training Partnership Act Study (JTPA) or mandatory programs such as the State Welfare-to-Work Demonstrations. Outside labor programs Tennessee's Student Teacher Achievement Ratio (Project STAR) was an experimental study on the impact of reduced class size on test scores. We present a brief review of the key findings from these studies.

2.1 Voluntary programs

Assessments based on the NSW Demonstration and the JTPA experiments have provided many insights on the reliability of non experimental methods applied to voluntary programs. For this type of intervention, with large eligible pools and relatively small numbers of participants, the problem is to find non-participants in the same labor market (or perhaps a very similar one) who look like the participants. In this context selection bias arises mainly due to individual self-selection. LaLonde (1986) raises concerns on the reliability of non experimental methods to estimate program impacts. His paper analyzes the NSW and demonstrates that common assumptions invoked by econometricians to justify traditional non experimental estimators such as cross-section, before-after, and difference-in-differences methods do not lead to reliable estimates of program impacts when compared to the experimental estimator. LaLonde's study uses experimental data from the NSW Demonstration to estimate the treatment on the treated effect of the program on earnings, and uses this estimate as a benchmark against which to assess the performance of non experimental methods to construct the counterfactual when applied to samples of non participants drawn from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). In response to Lalonde, Heckman and Hotz (1988) use additional samples from the NSW data to implement several specification tests that may help researchers in choosing among different non experimental estimators. Their tests perform well in rejecting models that give predictions considerably different from the experimental estimator. However, none of these studies analyze the performance of matching in constructing the desired counterfactuals, nor do they deal with the issue of data quality.

Using the NSW Demonstration data, Dehejia and Wahba (1999, 2002) suggest

that PSM performs well in constructing a comparison group that resembles the NSW participants in the counterfactual state. They use the same samples from the NSW, CPS and PSID previously analyzed by LaLonde to address the performance of several matching estimators. Specifically they combine treatment units drawn from the NSW experiment with non experimental comparison units drawn from the CPS and PSID samples as in LaLonde's study to generate nearest neighbor, radius (caliper) and stratified matched samples based on the propensity score. Their results show that nearest neighbor and radius matching do reasonably well in yielding accurate estimates of the treatment effect in non experimental settings, despite the fact that the comparison units come from different labor markets, the survey instruments are different and the set of conditioning variables is not particularly rich.

In a series of important studies analyzing the JTPA Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998) assess the empirical performance of matching estimators. In addition to traditional nearest neighbor and caliper matching from the statistics literature, they present evidence on newly developed techniques such as kernel and local-linear matching (for which they work out distributional properties) and extend the method to a longitudinal setup. Using experimental data from the JTPA experiment and three groups of non experimental units,² they find that propensity score matching performs well only under certain conditions: when working with a rich set of control variables, using the same survey instruments and placing participant and non participant units in the same local labor market.

These two studies also estimate the relative importance of selection versus other sources of bias when using PSM. They argue that the selection bias problem can be broken down into 3 components: biases arising by comparing the wrong units (comparing units outside the common support region), biases arising by comparing the right units in the wrong proportion (differences in the densities of observable characteristics between treatment and comparison units) and bias arising due to unobservables, or self-selection bias rigorously defined. Since matching controls for differences in observable

² Eligible non-participants who were interviewed especially for study using the same survey instrument; a sample of eligible individuals drawn from the Survey of Income and Program participation; and no-shows from the JTPA experimental treatment group sample.

characteristics inside the common support region, thus making treatment recipients and comparison units as close as possible on pre-treatment characteristics, matching is suitable to eliminate the first two sources of bias but not the third. When they combine JTPA recipients with non experimental comparison units to obtain an estimate of the evaluation bias in their matching estimators, they find that it is observable rather than unobservable characteristics that are the main source of bias. Even when the self-selection bias is high compared to the program impact, it is not as important as the bias associated with differences in supports and distributions of observable characteristics. They believe this result can be generalized to other job training programs in the U.S. due to the similarities in program goals and design. The main conclusion from these papers is that an evaluation strategy that does well in controlling for observed characteristics (including local economic conditions), and which gathers information from program participants and non-participants in a similar way (i.e. through the same questionnaire) can yield estimates of program impact which are close to the actual impact.

Smith and Todd (2003) reconcile the contradictory evidence on the performance of PSM as reported in Heckman et.al and Dehejia & Wahba. They also use the NSW, CPS and PSID samples in the LaLonde study and find that the results in the Dehejia & Wahba are particularly sensitive to the choice of their sample and conditioning variables. In particular, they find that sample restrictions imposed by Dehejia & Wahba to LaLonde's samples in order to include an additional variable in the estimation of their propensity score model considerably reduce the selection bias problem by dropping high earners from their final samples. Smith and Todd also show that traditional econometric estimators (such as regression, before-after and difference-in-differences) also perform well when applied to the Dehejia and Wahba restricted samples. They find that several matching estimators (nearest neighbor, caliper, kernel and local linear) applied to the NSW often exhibit substantial biases because of differences in the way earnings data is recorded in the NSW compared to CPS and PSID data, and because treated and non-participants units are not taken from the same local labor markets. Furthermore, because of the differences in earnings recording and in local labor markets conditions between NSW, CPS and PSID data, they find that difference-in-differences matching estimators

perform better than cross-sectional matching estimators because the former removes time invariant differences between treatment and comparison units. These results support the findings in Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998).

2.2 Mandatory programs

Assessments of non-experimental methods applied to mandatory interventions such as the welfare-to-work programs are provided by the studies of Friedlander and Robins (1995) and Bloom et al. (2002). In this case, the problem of a non experimental study is to find welfare recipients from non-participant locations similar enough to welfare recipients from participant locations, thus, selection bias arises mainly due to geographic differences in labor markets. Friedlander and Robins (1995) present evidence on the performance of cross-sectional regression adjustment methods and Mahalanobis matching as estimators for treatment effects of the interventions on employment. Their assessment strategy consists of using experimental control units (or earlier cohorts) from one location as a non experimental comparison group for treatment units in a different location. They compare the impact estimates produced by these non experimental procedures to those provided by the actual experiments, which compare treatment and control groups in the same location at the same time, and conclude that substantial biases arise when comparing recipients residing in different geographic areas.

Bloom et al. (2002) assess several non experimental methods in estimating the treatment effect of welfare-to-work interventions on earnings in a six state random assignment experiment: Atlanta, California, Michigan, Ohio, Oklahoma and Oregon. Estimators considered in the study are cross-sectional regression, PSM, difference-in-differences, and random-growth models. They construct non experimental comparison groups using a similar procedure to that in Friedlander and Robins--their non experimental samples are classified as in-state, out-of-state and multi-state comparison groups. The evidence from this study shows that in-state comparison units perform better than other type of comparisons and that cross-sectional OLS outperforms other methods when applied to in-state comparisons, while propensity score matching helps in reducing differences on pre-treatment characteristics in out-of-state and multi-state comparisons. They conclude that even their best non experimental methods do not work well enough to

replace the experimental estimator.

In summary, the assessments to date on PSM as an impact estimator are lukewarm at best. Surprisingly the technique appears to perform better for voluntary programs relative to mandatory ones despite the fact that selection on unobservables should be higher in the former case and PSM only controls for selection on observables. However even the more optimistic results from voluntary programs indicate somewhat strict conditions for success—identical survey instruments, similar local labor market conditions, and a rich set of control variables.

3. The PROGRESA Program

In 1996, the Mexican government launched a new anti-poverty program in rural areas, the Education, Health and Nutrition Program (*Programa de Educación, Salud y Alimentación*—PROGRESA).³ PROGRESA differed from previous national poverty programs in two key respects. First, it provided benefits conditional on beneficiaries fulfilling certain human capital enhancing requirements: school enrolment of children age 6-16; attendance by an adult at a monthly health seminar and compliance by all family members to a schedule of preventive health check-ups. Second, the program was implemented based on a very detailed targeting process aimed at reaching the poorest population in rural areas and avoiding local political influence in designating program beneficiaries.

3.1 Benefits

Child labor is a common subsistence strategy for poor households in rural Mexico, delaying school entry, reducing attendance, and leading to eventual early dropout. PROGRESA explicitly attempted to stimulate human capital investment and break the inter-generational cycle of poverty by setting the level of cash transfers according to the opportunity cost of children's time. Thus benefits increase according to the age of the child, with a significant jump between primary and middle school and larger payments for girls relative to boys. In addition to the schooling benefits, each eligible household receives a fixed monthly payment of approximately USD12 for food, and a lump-sum for school uniforms and books. The average transfer represents about one-third of total monthly household income.

³ In 2000 the program expanded to cover poor urban communities and changed its name to *Oportunidades*.

3.2 Coverage

PROGRESA has expanded in phases. Phase one began in August 1997, when 3,369 localities covering 140,544 households; phase two began in November 1997, incorporating 2,988 additional localities and 160,161 households. By the end of phase 11 in 2000 PROGRESA had incorporated over 70,000 localities in all 31 states of the country, covering approximately 2.6 million rural households.

3.3 Beneficiary selection

Targeting of poor households is implemented centrally at the PROGRESA headquarters in Mexico City and entails three stages. First, all localities in the country are ranked using a “marginality index” constructed from 1990 National Census data; this index is stratified into five categories and localities in the bottom categories (high and very high levels of marginality) are pre-selected to be part of the program. Out of 200,151 localities in Mexico, 76,098 rural localities (14.8 million people) were identified as having high or very high marginality levels and thus pre-selected for the program.

In the second stage the program identifies poor households within targeted localities. A community census is administered to all households in the selected localities to retrieve information about household characteristics that determine poverty status, including household income, which is used to identify households below the official poverty line. Predicted poverty status is then computed using the results from a discriminant analysis of the poverty indicator that selects the household characteristics that best discriminate between poor and non-poor households. In general, the best predicting variables are a dependency index (number of children to number of working age adults), an overcrowding index (persons per bedroom), the sex, age and schooling of the household head, the number of children, dwelling characteristics such as dirt floor, bathroom with running water, and access to electricity; and possession of durable goods such as a gas stove, a refrigerator, a washing machine and a vehicle. These characteristics are used to compute the discriminant score that separates eligible and non-eligible households in the selected localities.⁴

In stage three, the list of potential beneficiaries of the program is presented to a community assembly where the composition of the list is reviewed; if the assembly

⁴ See Skoufias, Davis & de la Vega (2001) for an assessment of the ProgresA targeting procedure.

rejects a household in the list or an omitted household is alleged to be poor, an administrative process is implemented and the central office delivers a final decision.

3.4 The social experiment

During the second phase of implementation (November-December 1997) a social experiment was launched to evaluate the impacts of the program on outcomes such as health and schooling for children and household consumption. A total of 506 rural localities from 6 states--Guerrero, Hidalgo, Michoacán, Puebla, Querétaro, San Luis Potosi, and Veracruz--were selected randomly as the experimental evaluation sample: 320 localities were randomly assigned to the treatment group and incorporated into the program while the remaining 186 localities were assigned to the control group and were incorporated later during phases 10 (November-December 1999) and 11 (March-April 2000). All eligible households in treatment localities were offered program benefits and services; none of those in the control localities received any benefit or service from the program until phases 10 or 11 of incorporation, that is, for eligible households in the control group localities all program benefits were delayed for approximately 24 months.⁵

The impact evaluation of PROGRESA was conducted independently by the International Food Policy Research Institute (IFPRI), and an overview of the main results can be found in Skoufias (2000). The overall evaluation used both qualitative and quantitative techniques to explore a variety of outcomes such as parents' attitudes towards the education of girls, the use of time by household members including children, and women's empowerment, but two of the most important outcomes analyzed were those related to school enrollment and spending behavior, and these are the outcomes we consider in this paper.⁶

4 Methodology

4.1 The evaluation problem

The parameter of interest is the effect of 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. Since one cannot observe the outcome for any given household in both the treated and the untreated states--the

⁵ See Behrman & Todd (1999) for an assessment of the randomization process.

⁶ All the evaluation studies are available on IFPRI's website: www.ifpri.org/themes/progresa.htm. The results show positive and significant impacts for schooling outcomes and food expenditures.

evaluation problem—researchers have focused on estimating the average effect of the program. Even so, we face the problem of constructing a suitable counterfactual outcome in the untreated state conditional on receiving treatment. To highlight this problem denote program participation by D , and let $D = 1$ for those who receive treatment and $D = 0$ for those who do not. In a nonparametric setting, the treatment on the treated parameter can be expressed as:

$$TT = E(Y_1 - Y_0 | D = 1) = E(Y_1 | D = 1) - \underbrace{E(Y_0 | D = 1)}_{\text{unobserved}}.$$

The last term in this expression is the counterfactual of interest: what the outcome for treated units would have been had they not received treatment; however, this counterfactual is not observable in the data. What we can observe instead is the average outcome in the untreated state $E(Y_0 | D = 0)$, which could serve as an estimate for the counterfactual, but we should expect in general that $E(Y_0 | D = 1) \neq E(Y_0 | D = 0)$ because of selection bias. Therefore, the central problem is to obtain a good estimate of the unobserved component.

There are two methods to solve this problem: experimental and non-experimental identification strategies. The experimental design applied to the social sciences, a social experiment, solves the evaluation problem by randomly denying treatment to analysis units (individuals, households, localities, etc.) which otherwise should receive it. From the pool of potential participant units some are randomly selected to receive treatment and some are randomized out; the former is the "treatment" group and the latter the "control" group. The outcome for control or randomized-out units is the counterfactual of interest; this counterfactual directly identifies the outcome in the untreated state from those units who otherwise would receive treatment. The key element in this setting is the randomization of units into and out of treatment so that outcomes are independent of treatment assignment and the selection into treatment is uncorrelated with either observable or unobservable characteristics, thus ensuring that no bias arises in comparing the observed outcome for treated and control units. This is the evaluation method implemented by PROGRESA.

The second identification strategy relies on non-experimental evaluation methods. This strategy applies statistical and econometric techniques to identify non-treated

households similar on pre-treatment characteristics to those in the pool of treated households, and comparing differences in mean outcomes between these two groups to identify the impact of the program. In this context the researcher may assume either that treatment assignment is independent of outcomes conditional on all observable characteristics that determine treatment assignment and outcomes (selection on observables), or that treatment assignment depends on unobservables whose bias should be corrected for (selection on unobservables). The matching method assumes selection on observables.

4.2 The propensity score matching technique

Our matching application is done in two steps. In step one we select all rural households from a nationally representative household survey, the *Encuesta Nacional sobre Ingresos y Gastos de los Hogares* (ENIGH). In step two we apply PSM (Rosenbaum and Rubin 1983; Heckman, Ichimura and Todd 1998) to construct a comparison group of households from the ENIGH sample. We combine background covariates from the experimental sample with these same variables from ENIGH and estimate the probability of being selected to participate in the program (the probability of being poor) and then match experimental control units with non experimental comparison units. The identification assumption in PSM is that outcomes are independent of program participation conditional on a particular set of observable characteristics. This is the conditional independence assumption, the ignorable treatment assignment of Rosenbaum and Rubin, and the assumption of selection on observables in the definition of Heckman and Robb (1985). Denoting by X the set of observables, the identification assumption can be expressed as $Y_0 \perp D \mid X$ where the symbol \perp denotes independence. Actually we require a weaker condition to identify our treatment parameter, that of conditional mean independence: $E(Y_0 \mid D = 1, X) = E(Y_0 \mid D = 0, X)$. By conditioning on X we can get an estimate of the unobserved component in the TT parameter. In particular, we can identify the parameter as follows:

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

Note that in the general application of matching X can be one or more specific characteristics of the observations being matched. The specific technique of PSM is due to Rosenbaum and Rubin (1983), who show that if the conditional independence assumption holds, then it also holds if instead of conditioning on X itself one conditions on the propensity score $P(X) = \Pr(D = 1 | X)$, that is $Y_0 \perp D | P(X)$. The advantage of this approach is that the dimension of the propensity score is one. In this case, the conditional mean independence assumption can be expressed as

$E(Y_0 | D = 1, P(X)) = E(Y_0 | D = 0, P(X))$. Therefore, we can estimate the TT parameter as follows:

$$TT(X) = E(Y_1 - Y_0 | D = 1, P(X)) = E(Y_1 | D = 1, P(X)) - E(Y_0 | D = 0, P(X)).$$

In our application we compute a direct measure of the bias associated with the TT parameter instead of computing the parameter itself. We compare control units from the experimental data with non-experimental comparison units from the ENIGH survey. The estimated bias can be expressed as:

$$B(X) = \underbrace{E(Y_0 | D = 1)}_{\text{Control units}} - \underbrace{E_{P(X)|D=1}\{E(Y_0 | D = 0, P(X))\}}_{\text{Matched comparison units}}.$$

Since control units did not receive any treatment, the estimated bias should be equal to zero. In this setting, any deviation from zero can be interpreted as evaluation bias.

4.3 Balancing score

We implement the matching procedure using a balancing score computed from a logit model. In particular we use the log odds-ratio as our balancing score because we are dealing with choice-based samples where the proportion of the treatment group is over-sampled in the dataset. In practice we generate a dummy variable that takes a value of one when the observation comes from the experimental sample (either from the treatment or control groups) and zero when it comes from the non experimental sample. We estimate a logit model using all of the observations available (treatment, control and non experimental units) in order to gain efficiency and use the estimated coefficients to

obtain the predicted probability (p) and then compute the log odds-ratio $\log(p/(1-p))$, for each observation in the control and comparison samples. All the experimental units are poor households but not every non experimental unit is poor though for the most part they come from localities that are targeted to participate in PROGRESA later (see below). Thus we are estimating the probability of being in poverty conditional on a set X of observable characteristics. Note that the variables used in X are precisely the variables that PROGRESA uses in calculating its point score to determine household eligibility.

4.4 Balancing test

In the estimation of the propensity score we perform a balancing test similar to the one employed in Dehejia and Wahba (1999, 2002). We estimate the score using our base specification of the logit model and obtain the predicted scores for control and comparison units. We then stratify the sample of controls and comparison units according to the score beginning with an arbitrary number of strata. We then test whether the average score between control and comparison units within each strata are statistically the same. If this is not the case, then we partition the sample further and test again, repeating the process until the scores are balanced inside each strata. Once all the strata are balanced, we perform individual mean t-test between controls and comparisons for each of the variables used to predict the score. For the tests that are not accepted we go back to the first step and include higher order or interaction terms for those variables where statistical differences in means remain. We continue this procedure until all the tests are accepted.

4.4 Common support

The common support is the region (S) where the balancing score has positive density for both treatment and comparison units. No matches can be formed to estimate the TT parameter (or the bias) when there is no overlap between the treatment (control) and comparison groups. To define the region of common support by dropping observations below the maximum of the mins and above the minimum of the maxs of the balancing score. This procedure entails some potential problems: the support condition may fail in interior regions, good matches could be lost near the boundary of the support region, and -applicable to other procedures as well- excluding observations in either group changes

the parameter being estimated.

4.5 Matching estimators

We examine the performance of several different matching methods. Applied to estimate the bias using control and comparison units, all matching estimators have the general form:

$$B_m = \frac{1}{n_1} \sum_{i \in I_1 \cap S}^{n_i} \left[Y_{1i} - \sum_{j \in I_0 \cap S} W(i, j) Y_{0j} \right],$$

where B_m denotes the matching estimator for the bias, n_1 denotes the number of observations in the control sample, Y_{1i} represent the outcome for controls and Y_{0j} represent the outcome for comparison units, I_1 and I_0 denote the set of control and comparison units respectively, S represents the region of common support, and the term $W(i, j)$ represent a weighting function that depends on the specific matching estimator.

We present empirical evidence on the performance of the following estimators:

Nearest-neighbor matching. For each control unit this method assigns a weight equal to one for the nearest comparison unit in terms of the balancing score and zero to all the other comparison observations. We implement the method with replacement, so that a single comparison unit can be used as a match for more than one control unit.

Caliper matching. This estimator chooses the nearest neighbor inside a caliper of width δ , that is, the set of matched comparisons can be represented by $\{j : |p_i - p_j| < \delta\}$, where p is the propensity score. This is an alternative way of imposing the common support condition.

Kernel matching. The weighting function is a (Gaussian) kernel density. All the observations in the comparison group inside the common support region are used, the farther the comparison unit from the control unit the lower the weight.

Local-linear matching. This estimator is similar to the kernel estimator but includes a linear term of the balancing score which is helpful when the data are asymmetric.

We use the bootstrap method to estimate standard errors for all of the matching

estimators which accounts for the fact that the balancing score is also estimated. For each estimator we estimate a logit model using all the experimental units (treatment and controls) and the non experimental comparison units and then drop the treatment units and predict the score for control and comparison units. Finally for each matching estimator we match the control and comparison samples inside the common support region and compute the bias estimate on the matched sample. This process is repeated 100 times to obtain the standard errors.

5. Data and Samples

5.1 Samples

The PROGRESA experimental evaluation data (*Encuesta de Evaluación de los Hogares -ENCEL*) consists of four rounds of household surveys covering 506 localities and approximately 25,000 households (poor and non-poor). One third of the sample was randomized out and serves as the control group to measure program impacts. Surveys were conducted in March and October 1998, and May and November 1999. We use the October 1998 round of ENCEL, which corresponds to approximately 8-10 months of program participation for treated households. PROGRESA expanded in phases, beginning its intervention in the poorest localities. Households in the evaluation sample were incorporated into the program during the second phase, and so are some of the poorest households in rural Mexico. This has important implications for the viability of the propensity score matching technique, which we discuss below.

The non experimental sample comes from the ENIGH, a biannual nationally representative household survey that collects information on income, expenditures, household demographic composition, and school enrollment. The sample size is approximately 13,000 households, of which approximately 4,000 are rural households; we use the 1998 round of ENIGH to construct the non experimental comparison group.

The 1998 wave of ENIGH was collected between September and early November, approximately 10-12 months after the start of PROGRESA, implying that some ENIGH households may actually have been participating in the program. Using PROGRESA retrospective administrative data, we are able to identify the date of entry (if entered) into the program for all rural localities that were sampled by ENIGH 1998. To avoid contamination bias we exclude all localities from the ENIGH rural sample that had

already entered PROGRESA at the time of the survey. The resulting sample of rural households is what we refer to as *Sample 1*. Additionally, since ENIGH is nationally representative and not poverty focused, there are many rural localities that never entered PROGRESA because they did not qualify. Since poor households in localities that did not qualify for PROGRESA may not provide good matches for poor households in localities that do qualify, we also present estimates based on a restricted sample that excludes all households in Sample 1, regardless of poverty status, from localities that do not qualify for PROGRESA. We refer to this more restricted group of households as *Sample 2*.

In general, because ENIGH is nationally representative while PROGRESA specifically targets the very poor, a big challenge is to see whether the matching technique is able to identify enough good matches from ENIGH to allow for meaningful comparisons with the control group from ENCEL.

5.2 Differences in survey instruments

Aside from differences in the sample frame which may inhibit good matches, a critical issue is the difference in survey instruments between ENIGH and ENCEL. The expenditure module in ENIGH is much more detailed than the ENCEL, and while the surveys were fielded at around the same time of year, many of the recall periods are also different, so that differences in expenditure outcomes may be entirely due to questionnaire design rather than evaluation technique. On the other hand, the questions on individual school enrollment are comparable across surveys. Finally, the questions on employment are slightly more detailed in the ENIGH survey, with a few additional questions included to probe for paid employment on the part of respondents. These differences allow us to assess whether the results from propensity score matching are sensitive to data quality and variations in survey instruments, a key issue pointed out by Heckman, Ichimura, Smith and Todd (1998) and Smith and Todd (2003).

6. Results

6.1 Mean characteristics of sub-sample

The experimental data from ENCEL 1998-October consist of 7,837 treatment household and 4,682 control households. The non experimental data drawn from ENIGH-1998 consist on 3,898 households from rural localities from which we extract two working

samples: Sample 1 refers to ENIGH households from rural localities not incorporated into PROGRESA prior to November 1998, i.e. excluding all localities already incorporated (2,479 households); Sample 2 refers to a further restricted sample which excludes all households in localities where PROGRESA was never implemented (736 households).

Table 1 presents summary statistics on the control variables used in the logit regression to estimate the balancing score—these are the exact variables used by PROGRESA in their targeting mechanism. Columns 1 and 2 provide means for the treatment and control units from ENCEL. These are virtually the same, indicating that control units in ENCEL are indeed a valid comparison group for the measurement of program impacts. The next three columns (columns 3, 4, and 5) present means for, respectively, the entire ENIGH rural sample and the two working samples. Rural ENIGH households are clearly better-off than their ENCEL counterparts as we would expect since ENIGH is nationally representative. For example, ENIGH household heads have significantly more schooling than ENCEL heads, significantly fewer children under age 13, and are more likely to have a refrigerator, a gas stove, a washing machine, and a vehicle. Note that the mean characteristics in the ENIGH-Sample 2 are closer to those of ENCEL, because we have excluded households from richer localities (those that never enter the program) in Sample 1.

Table 2 presents the means for the outcome variables we consider in our application. This table has the same structure as Table 1 and presents average outcome values for the treatment and control units from ENCEL and the comparison units drawn from ENIGH samples. This table again shows that rural ENIGH households are significantly better-off than the ENCEL households, with significantly higher per capita food expenditure and school enrollment rates for children ages 13-16 and lower rates of child employment.

6.2 Balancing score and common support

Results of the logit models to determine the probability of qualifying for the program are reported in Table 3. For efficiency reasons these estimates are based on all households in the evaluation sample (i.e. households from both the treatment and control localities) and all rural households (poor and non-poor) from either ENIGH-Sample 1 or ENIGH-Sample 2. The dependent variable is a dummy variable that takes a value of one

when a household comes from the experimental data and zero when it comes from the non experimental sample. Columns 1 and 3 report estimated coefficients and marginal effects respectively using ENIGH-Sample 1 while columns 5 and 7 report the same figures when we use ENIGH-Sample 2.

There are a few differences worth noting between the estimates over the different samples. Almost all variables are significant when we use ENIGH-Sample 1, which includes richer households in rural ENIGH, but several of these variables become insignificant when we use ENIGH-Sample 2, where households are more homogenous due to the exclusion of these richer households. Furthermore, the coefficients on heads' schooling become much larger in the latter case, while the bathroom indicators become smaller.

For each combined sample we perform the balancing tests described earlier to assess the specification of the logit model used to estimate our balancing score. Based on these results we included quadratic terms for the dependency and crowding variable, as well as an interaction between crowding and the number of kids under age 13. Table 4 reports summary statistics on the estimated propensity score, the odds-ratio, and the implied common support region - defined as the maximum of the mins and the minimum of the maxs of the balancing score between experimental and comparison units. The empirical distributions of the estimated odds-ratios are shown graphically in Figures 1a and 1b.

When we use households from ENIGH-Sample 1 as the comparison group, the mean odds-ratio is -0.709 for ENIGH households and around 3.2 for both control and treatment households from ENCEL; 1.62% of the control group and 1.65% of the non experimental comparison group do not satisfy the common support criteria and must be excluded from the subsequent analysis. In the case of ENIGH-Sample 2, the mean odds-ratio among the ENIGH sample is larger at 0.851 but still significantly lower than the mean for the ENCEL households, which is around 4.4. In this case as well, imposing the common support criteria results in the elimination of 1.62% of the control and 1.63% of the comparison groups.

6.3 Matched samples using nearest-neighbors

We now compare average characteristics from the experimental units to matched

comparison units from ENIGH samples 1 and 2. Columns 6 (sample 1) and 7 (sample 2) in Table 1 present average characteristics for the sample of households that have been matched on the balancing score using nearest-neighbor matching with replacement within the common support region. In both columns, mean characteristics are significantly different from the raw ENIGH samples before matching, and the matched households are clearly closer to ENCEL households in terms of those characteristics relative to the full rural ENIGH sample. For example, among the matched sample, the proportion of heads with secondary schooling is around 4-6%, compared to 5.5% in ENCEL and 11% in the overall rural sample from ENIGH. Similarly, the proportion of matched households without social security is 96% compared to 97% in ENCEL and 82% in overall rural ENIGH.

Average outcomes for the matched households drawn from samples 1 and 2 from ENIGH using nearest neighbor matching within the common support region are reported in columns 6 and 7 in Table 2. Average outcome values for these matched households are closer to the average outcomes for the experimental ENCEL households. In the case of school enrollment for older kids, the non experimental comparison group means are 0.51 and 0.47 (sample 2) compared to 0.48 in the control group and 0.58 in the full rural ENIGH sample; notice that the matched sample 2 mean is actually lower than the control group mean. For child labor the matched sample means are 0.17 (sample 1) and 0.11 (sample 2) compared to 0.18 in the control group and 0.14 in the overall rural ENIGH; here child labor is actually lower in matched sample 2 relative to the control group.

6.4 Bias estimates

6.4.1 Bias estimates for household outcomes

Table 5 presents estimates of the bias for household level outcomes using various matching estimators. These estimates of bias are calculated by taking the difference in means between the control group from ENCEL and the non experimental comparison group from ENIGH. If matching does well in replicating the experimental control group then this difference should be zero; thus, statistically significant deviations from zero indicate potential bias on impact estimates derived from the PSM technique. In Table 5 differences that are statistically different from zero (at 5%) are shown in bold. Virtually all expenditure composition outcomes are significantly different, whether measured in

levels or shares. Recall that there is significant variation in the data collection method for expenditures between the two surveys (ENCEL versus ENIGH) which may be driving these differences. This hypothesis is supported by the results of the schooling outcomes aggregated to the household level at the bottom of Table 5. None of these differences are statistically significant and this information is collected in a similar way in the two surveys. The child labor outcomes are also the same, and this information is collected in similar but not identical fashion across the two surveys.

Estimates based on the more restrictive comparison group from ENIGH-Sample 2 are shown in Table 6. These results follow the same general pattern as those in Table 5, although fewer of the expenditure differences are statistically significant. However none of the schooling and child labor outcomes aggregated to the household are significant suggesting that differences in questionnaires may be an important determinant of the performance of PSM.

Looking across matching techniques and focusing on the results in Table 5, we find differences, but not major ones, in the point estimates of the bias across the various techniques. Local-linear and caliper matching tend to produce larger point estimates of bias in food and vegetable expenditure levels, but not in shares, while caliper matching is the closest to the nearest-neighbor in terms of point estimates. The patterns of statistical significance are also identical for all of our matching estimates. This pattern of results is the same when the bias is estimated on the restricted comparison group shown in Table 6.

6.4.2 Bias estimates for individual outcomes

Tables 7 (sample 1) and 8 (sample 2) present estimates of bias for children's schooling outcomes at the individual level. Here we first match households, then compare children in matched households using only households with children in the relevant age range. Results from Table 7 indicate significant bias in enrollment outcomes for children age 8-16 only for 0.01 caliper matching, although this is primarily driven by the significant difference in outcomes for children age 8-12. The means for these outcomes (bottom of Table 2) reveal that enrollment rates for all children age 8-12 years old among matched non experimental comparison units are the same as among the treated, and both are significantly higher (by at least 3 percentage points) than the rate among control children. The other statistically significant differences are for child labor among all kids and the

sub-sample of girls age 2-16 based on local linear and kernel (0.1) matching. These results are somewhat surprising given the means for these outcomes in Table 2, especially for all kids 12-16 where the mean in the nearest neighbor matched sample is the same as that of the control group.

Table 8 shows bias estimates based on the more restrictive sample 2 that excludes households from localities not targeted by PROGRESA in any phase. These results also indicate bias in the impact estimates for enrollment rates for children ages 8-16 when caliper matching is used, and again this result appears to be driven by differences among the 8-12 age group. In this sample none of the child labor estimates are significantly different from zero. Taken as a whole these results also suggest that differences in questionnaire design can have an important effect on the accuracy of PSM in measuring program impact as suggested by Heckman et al. (1998). The other noteworthy result is the differences across matching techniques. While nearest neighbor never identifies significant differences in the individual level outcomes, differences are detected using caliper and local linear matching.

6.4.3 Comparison with regression analysis

While the main objective of this paper is to compare PSM to the experimental estimates, it is of some interest to compare the PSM results with those from regression analysis since the latter is such a commonly used non experimental technique. Tables 9 and 10 compare the bias estimates derived from PSM with the mean difference in outcomes derived from a regression equation. These equations regress the outcome of interest on the same conditioning variables used in the logit regression reported in Table 3, and are estimated using the controls from ENCEL and the unmatched ENIGH sample 1 only. A dummy variable is included to indicate that the observation is from the control sample; for expenditure levels and shares we use OLS while for the individual schooling and child employment variables we use probit models.

Table 9 reports the estimates of the OLS coefficient of the dummy variable indicating that the household is from the control sample, and for ease of comparison we also report the estimated bias for each outcome using nearest neighbor PSM taken from Table 5. In every case except meat measured in levels, the regression estimates show a statistically significant difference in mean outcome between control and comparison

households. Moreover for the statistically significant outcomes measured in levels, the regression estimates are larger in absolute value than those using PSM. For the outcomes measured in shares however, the regression estimates are actually slightly smaller for food and cereal.

Table 10 presents the results of the same analysis using individual level schooling and employment outcomes along with the earlier results from Table 8 based on nearest neighbor PSM. Here the differences are quite stark; while none of the PSM estimates are different from zero, 6 out of the 9 regression based estimates are statistically significant. Notice that the regression coefficients for schooling outcomes are negative, indicating that control units have lower schooling outcomes than comparison group units, thus implying that regression based estimates of impact would lead to an under-estimate of program impact. On the other hand the lone employment outcome that is significant is also negative, which in this case indicates that the control group has better outcomes than the comparison group, implying that a regression based approach would lead to an over-estimate of program impact. Recall that while the survey instruments asked about school enrolment in the same way, the questions on paid employment are more detailed in the ENIGH survey and likely to lead to higher rates of reported child employment relative to ENCEL which may explain the negative coefficient in Table 8.

7. Conclusions

The validity of propensity score matching as an impact evaluation estimator is a key issue in the evaluation literature given the potential difficulties in launching social experiments. All the published work in that shed light on this question have used employment and training programs from the U.S; the results from these studies seem to converge to the view that matching can be a viable impact estimator under certain specific conditions, which include the availability of a rich set of control variables, the use of similar survey instruments, and control for local economic conditions. A variant of the method extended to a longitudinal setup (difference-in-differences matching) is thought to be superior to simple cross-sectional matching techniques because it eliminates important time invariant sources of bias such as local environment and systematic measurement error.

In this article we present further evidence on the performance of cross-sectional propensity score matching outside the scope of employment and training interventions

and from a country other than the U.S. We find significant bias in the matching estimates for outcomes that are measured differently due to different survey instruments-- household expenditure levels and composition. The results are more encouraging for children's schooling outcomes, which are measured in the same way across surveys. For these outcomes, we find bias in the matching estimates for school enrollment of children age 8-12, where the method significantly underestimates program impact. However there are no statistically significant biases for school enrollment among the 13-16 age group, where PROGRESA has the largest impact. For child employment which is measured in a similar but not identical way, we find some evidence of bias but only for caliper and local linear matching, and these imply over-estimates of true program impact. This is likely due to the extra effort in the ENIGH survey to capture paid employment.

7.1 Implications

There are several important implications of the results presented here. First, using a different type of program from another country we have been able to corroborate the main conclusions of the existing literature on the conditions under which PSM may be a valid impact estimator. These main conclusions are that PSM requires appropriate and detailed covariates (in our case we have the exact variables used by PROGRESA to select beneficiaries) and the outcomes of interest need to be measured in comparable fashion. Second, PROGRESA type programs are spreading rapidly around Latin America and the Caribbean (LAC) region, as is the interest to appropriately evaluate the impacts of these investments. Our analysis implies that there may be scope for the design of a non experimental impact evaluation using PSM which can provide reasonable estimates of program impact. Almost all countries in LAC have an annual or biannual national household survey with information on income or expenditures, schooling and in some cases health information. Under the right conditions (phased expansion of mandatory programs) and with advanced planning (coordinating survey instruments) these surveys can be combined with specific data on beneficiaries to produce credible estimates of program impact at considerably less political and financial cost to governments.

References

- Behrman, Jere and Petra Todd (1999) "Randomness in the Experimental Samples of PROGRESA." Research Report, International Food Policy Research Institute. Washington D.C.
- Bloom, Howard, Charles Michalopoulos, Carolyn Hill and Ying Lei, (2002), "Can Non Experimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" MDRC Working Papers on Research Methodology.
- Dehejia, Rajeev and Sadek Wahba, 1999), "Causal Effects in Non Experimental studies: Reevaluating the Evaluation of Training Programs." Journal of the American Statistical Association, Vol.94: 1053-1062.
- Dehejia, Rajeev & Sadek Wahba, (2002), "Propensity Score Matching Methods for Non Experimental Causal Studies." Review of Economics & Statistics, Vol. 84: 151-161.
- Friedlander, Daniel and Phil Robbins, (1995), "Evaluating Program Evaluations: New Evidence on Commonly Used Non Experimental Methods." American Economic Review, Vol.85: 923-937.
- Heckman, James and Joseph Hotz, (1989), "Choosing Among Alternative Non Experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," Journal of the American statistical Association, Vol. 84: 862-880.
- Heckman, James and Joseph Hotz, (1989), "Choosing Among Alternative Non Experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," Journal of the American statistical Association, Vol. 84: 862-880.
- 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: 605-654.
- Heckman, James, Hidehiko Ichimura, and Petra Todd (1998) "Matching as an Econometric Evaluation Estimator." Review of Economic Studies. 65: 261-294.
- Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998) "Characterizing Selection Bias Using Experimental Data," Econometrica. 66: 1017-1089.
- Lalonde, Robert, (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," American Economic Review, Vol.76: 604-620.
- PROGRESA (1997) PROGRESA: Programa de Educacion Salud y Alimentación. Mexico.
- Rosenbaum, Paul and Donald Rubin (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika. 70: 41-50.
- Skoufias, Emmanuel (2000) "Is PROGRESA Working? Summary of the Results of an Evaluation by IFPRI" International Food Policy Research Institute.
- Skoufias, Emmanuel, Benjamin Davis and Sergio de la Vega (2001) "Targeting the Poor in Mexico: An evaluation of the Selection of Households into PROGRESA." World Development. 29: 19769-1784

Smith, Jeffrey and Petra Todd (2003) “Does Matching Overcome LaLonde’s Critique of Non-experimental Estimators?” Forthcoming, Journal of Econometrics.

Table 1: Summary statistics for conditioning variables by sample

	Data set: ENCEL		ENIGH				
	Sample: Treatment	Control	All rural	Raw samples		Matched samples	
				Sample1	Sample2	Sample1	Sample2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Demographic dependency	1.461 (0.95)	1.487 (0.98)	1.140 (1.06)	1.031 (0.95)	0.997 (0.94)	1.537 (1.03)	1.645 (1.13)
Head's sex	0.085 (0.29)	0.090 (0.34)	0.129 (0.34)	0.132 (0.34)	0.133 (0.34)	0.097 (0.30)	0.127 (0.33)
Head's schooling							
None	0.446 (0.50)	0.458 (0.50)	0.399 (0.49)	0.402 (0.49)	0.402 (0.49)	0.458 (0.50)	0.503 (0.50)
Incomplete primary	0.247 (0.43)	0.237 (0.43)	0.204 (0.40)	0.223 (0.42)	0.226 (0.42)	0.270 (0.44)	0.224 (0.42)
Incomplete secondary	0.055 (0.23)	0.055 (0.23)	0.108 (0.31)	0.124 (0.33)	0.129 (0.34)	0.038 (0.19)	0.055 (0.23)
Head'a age	42.444 (14.95)	42.737 (15.14)	46.879 (16.33)	46.936 (16.43)	46.969 (16.31)	41.361 (13.98)	41.540 (13.90)
Number of kids ages 13 or below	2.457 (1.66)	2.483 (1.61)	1.487 (1.54)	1.339 (1.42)	1.290 (1.40)	2.576 (1.62)	2.534 (1.53)
Crowding index	4.399 (2.26)	4.460 (2.26)	2.601 (1.86)	2.348 (1.70)	2.289 (1.67)	4.359 (2.20)	4.178 (2.01)
Do not have social security	0.968 (0.18)	0.960 (0.20)	0.819 (0.39)	0.774 (0.42)	0.767 (0.42)	0.962 (0.19)	0.957 (0.20)
No bathroom	0.484 (0.50)	0.493 (0.50)	0.341 (0.47)	0.284 (0.45)	0.270 (0.44)	0.579 (0.49)	0.536 (0.50)
Bathroom no water	0.497 (0.50)	0.489 (0.50)	0.491 (0.50)	0.486 (0.50)	0.494 (0.50)	0.408 (0.49)	0.437 (0.50)
Dirt floor	0.730 (0.44)	0.755 (0.43)	0.255 (0.44)	0.203 (0.40)	0.190 (0.39)	0.736 (0.44)	0.748 (0.43)
Without gas stove	0.847 (0.36)	0.835 (0.37)	0.365 (0.48)	0.263 (0.44)	0.244 (0.43)	0.846 (0.36)	0.837 (0.37)
Without refrigerator	0.959 (0.20)	0.963 (0.19)	0.562 (0.50)	0.478 (0.50)	0.450 (0.50)	0.962 (0.19)	0.967 (0.18)
Without washer	0.986 (0.12)	0.988 (0.11)	0.764 (0.42)	0.691 (0.46)	0.683 (0.47)	0.985 (0.12)	0.986 (0.12)
Without vehicle	0.979 (0.14)	0.981 (0.14)	0.791 (0.41)	0.739 (0.44)	0.734 (0.44)	0.944 (0.23)	0.943 (0.23)
Observations	7837	4682	3898	2133	2291	768	363

Treatment and Control units are from PROGRESA's experimental sample. ENIGH sample 1 excludes PROGRESA localities; ENIGH sample 2 excludes 'rich' localities from sample 1—see text for details. Matched samples are constructed using nearest neighbor with replacement and common support. Standard deviation in parenthesis.

Table 2: Summary statistics for outcome variables by sample

	Data set: ENCEL		ENIGH					
	Sample:	Treatment	Control	All rural	Raw samples		Matched samples	
					Sample1	Sample2	Sample1	Sample2
A. Household outcomes								
Food expenditure per capita		511.5 (400.5)	476.5 (403.7)	905.9 (696.1)	970.7 (731.7)	878.5 (662.2)	687.4 (626.8)	645.1 (603.2)
Children's clothing per capita		20.3 (35.5)	14.7 (28.5)	40.4 (50.9)	45.3 (57.4)	38.6 (45.3)	25.4 (23.2)	23.5 (19.9)
Percentage of kids 8-16 in school		0.785 (0.310)	0.743 (0.330)	0.804 (0.320)	0.799 (0.330)	0.776 (0.350)	0.786 (0.319)	0.767 (0.350)
Observations		{7837}	{4682}	{3898}	{2479}	{736}	{768}	{363}
B. Children outcomes								
<u>School enrolment</u>								
Children 8-16		0.772 (0.420)	0.727 (0.446)	0.795 (0.404)	0.791 (0.407)	0.764 (0.425)	0.766 (0.423)	0.767 (0.423)
		{13589}	{8130}	{4407}	{2598}	{793}	{659}	{309}
Children 8-12		0.921 (0.270)	0.891 (0.311)	0.948 (0.221)	0.947 (0.224)	0.948 (0.222)	0.930 (0.255)	0.938 (0.241)
		{8200}	{4877}	{2563}	{1493}	{464}	{531}	{243}
Children 13-16		0.545 (0.498)	0.480 (0.500)	0.582 (0.493)	0.579 (0.494)	0.505 (0.501)	0.512 (0.501)	0.468 (0.500)
		{5389}	{3253}	{1844}	{1105}	{329}	{387}	{171}
<u>Work for pay</u>								
All Children 12-16		0.111 (0.314)	0.116 (0.321)	0.107 (0.310)	0.126 (0.332)	0.121 (0.326)	0.118 (0.323)	0.112 (0.316)
		{7028}	{4250}	{2402}	{1259}	{422}	{458}	{197}
Boys 12-16		0.164 (0.370)	0.180 (0.384)	0.141 (0.348)	0.164 (0.371)	0.132 (0.340)	0.167 (0.373)	0.112 (0.318)
		{3628}	{2146}	{1210}	{627}	{204}	{228}	{89}
Girls 12-16		0.054 (0.226)	0.051 (0.220)	0.073 (0.260)	0.089 (0.284)	0.110 (0.314)	0.070 (0.255)	0.111 (0.316)
		{3376}	{2100}	{1192}	{632}	{218}	{230}	{108}

See notes to Table 1 for explanation. Numbers in curly brackets indicate sample size.

Table 3: Logit estimates for program eligibility

	ENCEL + ENIGH (Sample 1)				ENCEL + ENIGH (Sample 2)			
	Coeff.	std.err.	Marginal	std.err.	Coeff.	std.err.	Marginal	std.err.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependency ratio	0.2972	0.0882	0.0200	0.0059	0.3671	0.1291	0.0052	0.0019
Head's sex	-0.1500	0.1013	-0.0106	0.0076	-0.1886	0.1457	-0.0029	0.0024
Head's schooling								
Complete Primary	0.4656	0.0840	0.0308	0.0055	0.7168	0.1270	0.0100	0.0018
Incomplete Secondary	0.8866	0.1104	0.0499	0.0053	1.1463	0.1683	0.0128	0.0017
Complete Secondary or more	0.6609	0.1529	0.0350	0.0063	0.8687	0.2286	0.0087	0.0017
Head's age	-0.0842	0.0408	-0.0057	0.0027	-0.0581	0.0619	-0.0008	0.0009
Head's age squared	0.0020	0.0008	0.0001	0.0001	0.0016	0.0012	0.0000	0.0000
Head's age cube	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Number of kids ages 13 or below	0.6195	0.0615	0.0416	0.0043	0.5700	0.0910	0.0081	0.0014
Crowding index	0.4556	0.0529	0.0306	0.0036	0.5517	0.0755	0.0078	0.0012
Do not have social security	1.5215	0.1069	0.1785	0.0190	1.1933	0.1674	0.0302	0.0069
No bathroom	0.5166	0.1437	0.0342	0.0096	0.1358	0.2047	0.0019	0.0029
Bathroom no water	0.6291	0.1384	0.0426	0.0097	0.4555	0.1982	0.0065	0.0029
Dirt floor	1.2571	0.0725	0.1024	0.0070	1.4868	0.1161	0.0309	0.0035
Without gas stove	1.4247	0.0767	0.1328	0.0093	1.6497	0.1189	0.0422	0.0050
Without refrigerator	1.3320	0.0917	0.1390	0.0141	1.2770	0.1310	0.0329	0.0059
Without washer	1.0763	0.1286	0.1084	0.0185	0.7289	0.1783	0.0147	0.0051
Without vehicle	0.5302	0.1217	0.0436	0.0121	0.3255	0.1670	0.0054	0.0032
Crowding index squared	-0.0113	0.0078	-0.0008	0.0005	-0.0172	0.0119	-0.0002	0.0002
Crowding index * number of kids	-0.0696	0.0135	-0.0047	0.0009	-0.0631	0.0214	-0.0009	0.0003
Dependency ratio cube	-0.0629	0.0158	-0.0042	0.0011	-0.0767	0.0222	-0.0011	0.0003
Constant	-5.9488	0.6809			-4.8726	1.0328		
Number of observations	14745				13031			
Likelihood ratio test	6292				2252			
Prob.	(0.000)				(0.000)			

See notes to Table 1 for definition of samples. The dependent takes a value of one if the unit comes from the ENCEL experimental sample, and zero if from the ENIGH non experimental sample.

Table 4: Propensity (Balancing) Scores Estimates

	Statistics				Obs. inside common support	Obs. in each sample	Percentage excluded
	Mean	Std.Dev.	Min	Max			
A. Matched Sample 1							
Treatment	3.183	1.350	-3.769	6.173	7704	7837	1.70%
Control	3.216	1.338	-3.592	5.876	4606	4682	1.62%
Comparison (ENIGH 1998)	-0.709	2.454	-6.359	5.611	2438	2479	1.65%
B. Matched Sample 2							
Treatment	4.411	1.502	-2.305	7.620	7704	7837	1.70%
Control	4.449	1.492	-3.233	7.358	4606	4682	1.62%
Comparison (ENIGH 1998)	0.851	2.197	-4.557	6.576	724	736	1.63%

The last column refers to observations outside the region of common support, defined as the maximum of the mins and the minimum of the maxs. Treatment and control units are from ENCEL. See notes to Table 1 for explanation of samples.

Table 5: Direct estimates of the bias for household level outcomes -- sample 1

Matching method:	Nearest	Caliper		Local Linear		Kernel	
	Neighbor	d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2
<u>Expenditure</u>							
Food	-219.114 (33.772)	-223.855 (29.089)	-238.126 (44.568)	-226.493 (31.438)	-216.869 (31.252)	-215.466 (29.607)	-214.611 (29.361)
Vegetables	70.698 (2.132)	71.220 (2.186)	72.578 (5.305)	69.177 (1.968)	69.382 (1.862)	69.484 (1.950)	69.601 (1.882)
Cereals	21.038 (11.872)	22.267 (9.895)	20.517 (16.995)	22.045 (9.769)	26.901 (10.106)	26.520 (8.988)	27.595 (8.907)
Meat	-3.960 (8.382)	-1.639 (6.992)	-4.027 (11.464)	-5.338 (7.526)	-3.937 (7.627)	-4.597 (7.669)	-3.495 (7.487)
Kid clothes	11.158 (0.798)	11.265 (0.789)	11.505 (1.571)	10.916 (0.688)	11.218 (0.643)	11.116 (0.658)	11.133 (0.657)
<u>Expenditure shares</u>							
Food	0.267 (0.015)	0.263 (0.012)	0.248 (0.017)	0.262 (0.013)	0.268 (0.013)	0.263 (0.013)	0.263 (0.013)
Vegetables	0.120 (0.001)	0.121 (0.002)	0.122 (0.005)	0.119 (0.001)	0.119 (0.001)	0.119 (0.001)	0.119 (0.001)
Cereals	0.195 (0.007)	0.195 (0.007)	0.191 (0.011)	0.194 (0.006)	0.197 (0.006)	0.195 (0.005)	0.196 (0.005)
Meat	0.069 (0.006)	0.070 (0.005)	0.065 (0.009)	0.068 (0.006)	0.070 (0.005)	0.069 (0.006)	0.069 (0.005)
Kids clothes	0.018 (0.001)	0.018 (0.001)	0.018 (0.002)	0.018 (0.001)	0.018 (0.001)	0.018 (0.001)	0.018 (0.001)
<u>Kids' schooling</u>							
Percent enrolled	-0.037 (0.023)	-0.045 (0.022)	-0.051 (0.033)	-0.023 (0.022)	-0.026 (0.021)	-0.026 (0.020)	-0.030 (0.020)
Percent never enrolled	-0.024 (0.016)	-0.017 (0.013)	-0.013 (0.016)	-0.019 (0.013)	-0.018 (0.012)	-0.016 (0.012)	-0.014 (0.011)
<u>Child labor</u>							
(% working for pay)	-0.007 (0.026)	0.004 (0.021)	-0.013 (0.036)	-0.026 (0.021)	-0.017 (0.020)	-0.028 (0.021)	-0.019 (0.020)
<u>Match summary for non experimental controls</u>							
Number of hhlds used	768	767	564				
Average times used	5.99	5.45	2.18				
Maximum use	57	50	12				

Sample 1 excludes ENIGH rural households that were already in PROGRESA at the time of the survey. Bootstrapped standard errors in parenthesis below the estimates account for the estimation of the propensity score. Significant estimates at 5% shown in bold. The nearest-neighbor estimator was computed with replacement. The kernel estimator uses the normal density.

Table 6: Direct estimates of the bias for household level outcomes -- sample 2

Matching method:	Nearest Neighbor	<u>Caliper</u>		<u>Local Linear</u>		<u>Kernel</u>	
		d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2
<u>Expenditure</u>							
Food	-169.252 (67.164)	-255.987 (54.835)	-290.038 (83.193)	-263.924 (59.692)	-239.388 (52.183)	-258.173 (59.396)	-264.560 (58.733)
Vegetables	71.777 (2.703)	72.124 (3.672)	76.468 (9.756)	68.382 (2.401)	68.745 (2.377)	69.451 (2.368)	68.917 (2.356)
Cereals	56.921 (16.552)	31.938 (13.513)	40.891 (25.565)	27.555 (16.233)	35.115 (15.057)	28.486 (16.519)	27.671 (15.736)
Meat	19.130 (13.931)	11.496 (12.150)	15.432 (19.465)	5.912 (12.068)	8.758 (12.240)	9.008 (12.611)	7.182 (12.340)
Kid clothes	10.566 (1.163)	9.678 (1.289)	10.448 (3.245)	10.713 (1.025)	10.828 (0.997)	10.895 (1.000)	10.732 (1.005)
<u>Expenditure shares</u>							
Food	0.242 (0.021)	0.232 (0.014)	0.225 (0.024)	0.237 (0.019)	0.244 (0.020)	0.228 (0.019)	0.233 (0.018)
Vegetables	0.120 (0.002)	0.120 (0.003)	0.121 (0.007)	0.118 (0.002)	0.119 (0.002)	0.119 (0.002)	0.119 (0.002)
Cereals	0.209 (0.012)	0.196 (0.010)	0.199 (0.018)	0.196 (0.011)	0.202 (0.010)	0.195 (0.012)	0.196 (0.011)
Meat	0.078 (0.009)	0.078 (0.008)	0.077 (0.014)	0.073 (0.009)	0.076 (0.008)	0.076 (0.008)	0.075 (0.008)
Kids clothes	0.018 (0.001)	0.016 (0.001)	0.018 (0.003)	0.018 (0.001)	0.018 (0.001)	0.018 (0.001)	0.018 (0.001)
<u>Kids' schooling</u>							
Percent enrolled	-0.045 (0.046)	-0.055 (0.035)	-0.018 (0.059)	-0.015 (0.040)	-0.024 (0.040)	-0.016 (0.040)	-0.018 (0.039)
Percent never enrolled	-0.020 (0.022)	-0.013 (0.018)	-0.023 (0.031)	-0.024 (0.018)	-0.018 (0.017)	-0.026 (0.019)	-0.024 (0.018)
<u>Child labor</u>							
(% working for pay)	0.041 (0.034)	0.014 (0.037)	-0.037 (0.052)	0.005 (0.036)	0.026 (0.033)	-0.002 (0.036)	0.006 (0.037)
<u>Match summary for non experimental controls</u>							
Number of hhlds used	363	356	207				
Average times used	12.36	6.56	1.97				
Maximum use	195	37	12				

Sample 2 excludes from the ENIGH rural sample localities already in PROGRESA and those never scheduled to enter the program. Bootstrapped standard errors in parenthesis below estimates account for estimation of the propensity score. Significant estimates at 5% shown in bold. Nearest neighbor done with replacement; kernel uses normal density.

Table 7: Estimates of the bias for individual schooling and work outcomes -- sample 1

Matching method:	Nearest	Caliper		Local Linear		Kernel	
	Neighbor	d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2
<u>Currently enrolled</u>							
All kids 8-16	-0.038 (0.030)	-0.064 (0.025)	-0.053 (0.042)	-0.024 (0.015)	-0.024 (0.015)	-0.028 (0.015)	-0.031 (0.015)
Kids 8-12	-0.045 (0.024)	-0.030 (0.015)	-0.047 (0.030)	-0.021 (0.016)	-0.017 (0.017)	-0.024 (0.017)	-0.027 (0.016)
Kids 13-16	0.010 (0.041)	0.002 (0.039)	-0.003 (0.073)	0.025 (0.026)	0.013 (0.027)	0.014 (0.027)	0.011 (0.027)
<u>Never enrolled</u>							
All kids 8-16	-0.029 (0.020)	-0.018 (0.015)	-0.003 (0.017)	-0.026 (0.014)	-0.029 (0.014)	-0.026 (0.014)	-0.024 (0.012)
Kids 8-12	-0.024 (0.019)	-0.033 (0.013)	-0.020 (0.022)	-0.042 (0.016)	-0.044 (0.016)	-0.042 (0.016)	-0.038 (0.014)
Kids 13-16	0.009 (0.024)	0.000 (0.020)	0.015 (0.033)	-0.002 (0.014)	-0.007 (0.015)	-0.002 (0.014)	-0.003 (0.014)
<u>Work for pay</u>							
All kids 12-16	-0.028 (0.029)	0.011 (0.023)	-0.030 (0.039)	-0.062 (0.023)	-0.054 (0.023)	-0.051 (0.021)	-0.041 (0.021)
Boys 12-16	-0.075 (0.048)	-0.033 (0.036)	0.006 (0.077)	-0.057 (0.039)	-0.052 (0.039)	-0.043 (0.039)	-0.030 (0.039)
Girls 12-16	-0.010 (0.022)	-0.005 (0.023)	-0.018 (0.048)	-0.034 (0.019)	-0.045 (0.021)	-0.020 (0.017)	-0.026 (0.017)
<u>Match summary for non experimental controls</u>							
Number of hhlds used	659	652	356				
Average times used	12.12	10.13	3.66				
Maximum use	127	63	15				

Bootstrapped standard error in parenthesis below estimates account for estimation of propensity score. Estimates in bold are significant at 5%. Nearest neighbor is done with replacement; kernel uses the normal density. Sample 1 excludes PROGRESA localities identified in the ENIGH sample. Match summary is for 8-16 year old schooling sample only.

Table 8: Estimates of the bias for individual schooling and work outcomes – sample2

Matching method:	Nearest	Caliper		Local Linear		Kernel	
	Neighbor	d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2
<u>Currently enrolled</u>							
All kids 8-16	-0.054 (0.057)	-0.085 (0.035)	-0.053 (0.065)	-0.010 (0.036)	-0.016 (0.035)	-0.012 (0.041)	-0.015 (0.034)
Kids 8-12	-0.023 (0.042)	-0.071 (0.026)	-0.084 (0.048)	-0.016 (0.034)	-0.016 (0.032)	-0.018 (0.037)	-0.023 (0.031)
Kids 13-16	-0.002 (0.077)	0.081 (0.064)	0.052 (0.142)	0.058 (0.056)	0.030 (0.055)	0.050 (0.063)	0.041 (0.056)
<u>Never enrolled</u>							
All kids 8-16	-0.017 (0.021)	-0.004 (0.020)	-0.012 (0.034)	-0.016 (0.017)	-0.016 (0.017)	-0.016 (0.016)	-0.014 (0.016)
Kids 8-12	-0.045 (0.020)	-0.012 (0.018)	-0.013 (0.030)	-0.024 (0.018)	-0.024 (0.018)	-0.022 (0.016)	-0.021 (0.017)
Kids 13-16	-0.009 (0.028)	-0.013 (0.033)	0.000 (0.065)	-0.009 (0.024)	-0.006 (0.023)	-0.013 (0.024)	-0.007 (0.023)
<u>Work for pay</u>							
All kids 12-16	0.012 (0.036)	0.031 (0.030)	0.021 (0.067)	0.017 (0.028)	0.046 (0.024)	0.013 (0.031)	0.021 (0.028)
Boys 12-16	0.033 (0.051)	0.071 (0.049)	0.020 (0.147)	0.040 (0.043)	0.060 (0.042)	0.036 (0.049)	0.039 (0.046)
Girls 12-16	-0.019 (0.044)	0.011 (0.038)	-0.065 (0.086)	-0.022 (0.036)	-0.018 (0.031)	-0.029 (0.041)	-0.022 (0.037)
<u>Match summary for non experimental controls</u>							
Number of hhlds used	309	299	117				
Average times used	24.78	9.72	3.56				
Maximum use	445	67	14				

Bootstrapped standard error in parenthesis below estimates account for estimation of propensity score. Estimates in bold are significant at 5%. Nearest neighbor is done with replacement; kernel uses the normal density. Sample 2 excludes PROGRESA localities identified in the ENIGH sample as well as localities that never entered the program. Match summary is for 8-16 year old schooling sample only.

Table 9: Regression estimates of the bias for expenditure outcomes

	<u>Regression</u>	<u>Nearest Neighbor</u>
	(1)	(2)
Expenditure		
Food	-270.819 (18.342)	-219.114 (33.772)
Vegetables	79.202 (2.966)	70.698 (2.132)
Cereals	29.901 (7.121)	21.038 (11.872)
Meat	-10.700 (5.739)	-3.960 (8.382)
Kid clothes	12.086 (0.982)	11.158 (0.798)
Expenditure shares		
Food	0.233 (0.006)	0.267 (0.015)
Vegetables	0.124 (0.002)	0.120 (0.001)
Cereals	0.183 (0.005)	0.195 (0.007)
Meat	0.069 (0.004)	0.069 (0.006)
Kids clothes	0.018 (0.001)	0.018 (0.001)

Column (1) reports the estimated OLS coefficients on a control-group dummy in regressions of the outcomes on this dummy and the other observable characteristics used to estimate the balancing score; these coefficients are the OLS estimate of the bias. Column (2) reports the bias estimates using nearest neighbor matching taken from Table 5 column (1). Standard errors in parentheses below marginal probabilities. Bold indicates significance at 5%.

Table 10: Regression estimates of the bias for individual level outcomes

	<u>Regression</u>	<u>Nearest neighbor</u>
	(1)	(2)
Currently enrolled		
All kids 8-16	-0.052 (0.012)	-0.038 (0.030)
Kids 8-12	-0.034 (0.009)	-0.045 (0.024)
Kids 13-16	-0.047 (0.024)	0.010 (0.041)
Never enrolled		
All kids 8-16	-0.011 (0.005)	-0.029 (0.020)
Kids 8-12	-0.022 (0.007)	-0.024 (0.019)
Kids 13-16	0.005 (0.007)	0.009 (0.024)
Work for pay		
All kids 12-16	-0.014 (0.013)	-0.028 (0.029)
Boys 12-16	0.006 (0.022)	-0.075 (0.048)
Girls 12-16	-0.028 (0.014)	-0.010 (0.022)

Column (1) reports the estimated probit coefficients on a control-group dummy in regressions of the outcomes on this dummy and the other observable characteristics used to estimate the balancing score; these coefficients are the regression estimate of the bias. Column (2) reports the bias estimates using nearest neighbor matching taken from Table 7 column (1). Standard errors in parentheses below marginal probabilities. Bold indicates significant at 5%.

Figure 1a: Empirical density of estimated log odds-ratio: sample 1

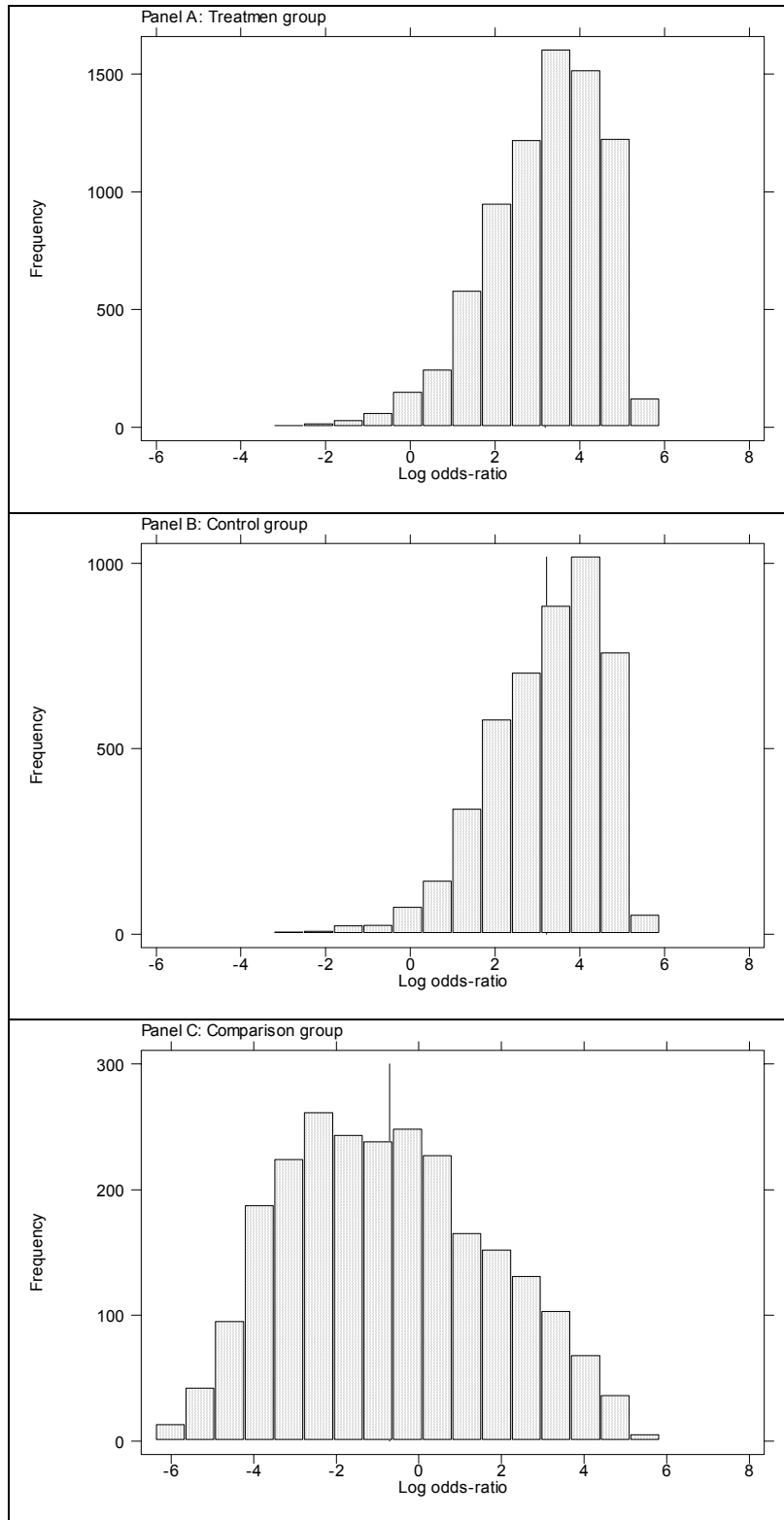


Figure 1b: Empirical density of estimated log odds-ratio: sample 2

