Office of Evaluation and Oversight

An Assessment of Propensity Score Matching as a Non Experimental Impact Estimator: Evidence from Mexico's PROGRESA Program

*Juan Jose Diaz and Sudhanshu Handa **

** Juan Jose Diaz is an Economist at the Group for the Analysis of Development (GRADE-Lima, Peru). Sudhanshu Handa is Associate Professor in the Department of Public Policy at the University of North Carolina at Chapel Hill.* *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 several earlier drafts, William Evans, Alex Whalley and seminar participants at the University of Maryland, the Johns Hopkins University, the Latin American & Caribbean Economics Association (LACEA) Annual Meetings in Mexico, the Carolina Population Center, GRADE and the Research Department at the Inter American Development Bank for constructive criticism. The findings and interpretations of the authors do not necessarily represent the views of the Inter-American Development Bank. The usual disclaimer applies.* *Corresponding author: Handa, Dept. of Public Policy, CB# 3435, University of North Carolina, Chapel Hill, NC 27599-3435; email: shanda@email.unc.edu*

Released by
Stephen A. Quick

Working Paper:
OVE/WP-04/05
July 22, 2005

TABLE OF CONTENTS

ABSTRACT

Not all policy questions can be addressed by social experiments. 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. In this paper we present evidence on the reliability of propensity score matching to estimate the bias associated with the effect of treatment on the treated, exploiting the availability of experimental data from a Mexican antipoverty program on several outcomes such as food expenditure and child schooling and labor. We compare the results of the experimental impact estimator with those using matched samples drawn from a (non-experimental) national survey carried out to measure household income and expenditures. Our results show that simple-cross sectional matching produces significant bias for outcomes measured in different ways. Results are more positive for outcomes measured similarly across survey instruments, but even in this case there are indications of bias depending on sample and matching method.

Keywords: Propensity score matching, treatment effects, program evaluation, PROGRESA

INTRODUCTION

Social experiments have become the standard method for estimating program impacts. By randomizing observational units into treatment and control groups, social experiments provide a clean estimate of program impacts because both observable and unobservable characteristics get uncorrelated with treatment assignment and thus no selection bias problem arises. However, social experiments are usually not available for several reasons such as high political, ethical and monetary costs, and the inability to implement experiments for universal entitlements or on-going programs, even more, randomization does not allow answering all policy relevant questions. Non-experimental methods, the alternative to randomized evaluations, identify program impacts by imposing stronger non-testable assumptions than randomization and researchers have to make the case for justifying them in particular applications. In this context, testing the reliability of non-experimental methods is a central issue in the evaluation literature and the availability of a randomized experiment provides the opportunity to assess the validity of non-experimental identification assumptions.

This study contributes to the small but growing literature on the performance of propensity score matching (PSM). PSM estimates treatment effects under the assumption of selection on observables and is appealing because it does not impose a functional form and highlights the support condition. Given that the popularity of PSM is spreading rapidly, it is important to assess whether it provides a suitable substitute for the experimental method. In this regard, we exploit the availability of experimental data to empirically assess the performance of several propensity score matching techniques as compared to the (unbiased) experimental estimator. These unique data come from a social experiment designed to evaluate the Education, Health and Nutrition Program (*Programa de Educación, Salud y Alimentación* – PROGRESA), Mexico's anti-poverty program.

PROGRESA is a conditional cash transfer program 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 family member 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 the incorporation process in late 1,997 and by 2,000, the program had incorporated 72,345 rural localities in all 31 states around the country covering approximately 2.6 million households. During the second phase of incorporation, a social experiment to evaluate the program was carried out in 506 program-eligible localities/villages across six Mexican states. One-third of these localities were randomly selected for delayed entry into the program, and thus served as the randomized-out control group. We combine this experimental control group from the PROGRESA social experiment with a non-experimental comparison group from a national household survey on income and expenditure carried out by the Mexican National Statistical Institute in order to estimate the potential bias that arises when estimating program impacts using the matching method.

The results of this paper contribute to the existing literature in several ways. First, given that social experiments are not abundant, our paper is among the few in the literature to exploit experimental data to "evaluate" non-experimental evaluation methods. Second, all the published research on the reliability of matching as an impact estimator is based on employment and

training programs in the U.S, our is the first paper that uses experimental data from a nation-wide and mandatory intervention in a developing country. It is also one of the first papers to study a program different to an employment or training program. Our assessment of matching based on a cash transfer poverty program is particularly valuable because at least five other countries in Latin America and the Caribbean region have begun implementing programs similar to PROGRESA and it is likely that they will not include randomized evaluations. Third, we employ and compare a range of matching techniques including kernel and local-linear matching. Finally, we are able to compare the bias arising for outcomes collected through both the same and different questionnaires, thus providing evidence on the importance of questionnaire versus other sources of bias.

Our main results show that PSM does not perform well in replicating the benchmark for outcomes measured using different questionnaire designs such as household expenditure, although there is some improvement in performance for outcomes that are measured comparably (such as child employment and schooling). In addition, we find that these results are robust to the specific method applied including caliper, kernel and local linear matching; sensitivity analysis suggests that having a rich set of control variables improves the results. The rest of the paper is organized as follows: section I provides a summary for the state of the literature in this field; section II describes the PROGRESA program; section III presents our methodology and data; section IV describes our main results, and section V concludes.

I. BACKGROUND AND SELECTED LITERATURE

The typical parameter of interest in a program evaluation 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 ($D=1$). Since these potential outcomes cannot be observed for any single observational unit in both counterfactual states at the same time - the evaluation problem[1] - the essence of an identification strategy is the estimation of the missing counterfactual outcome (the outcome for a treated unit had it not received treatment).

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 standard to evaluate policy interventions. Bear in mind, however, that randomized social experiments are not a panacea; 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 dropout bias.[2] Additionally, a potential drawback of social experiments is that they may be too costly to implement in some contexts and may raise ethical or political 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. (e.g. Burtless 1995, Fraker and Maynard 1987, Friedlander and Robbins 1995, LaLonde 1986, LaLonde and Maynard 1987, Michalopoulos et al. 2004).

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 methods have been developed to control for confounding variables and selectivity issues. To achieve identification, these methods impose non-testable assumptions (although many of their implications might be) that may or may not be tenable in actual data (Heckman, Hotz and Dabos 1987, Heckman and Hotz 1989, Heckman and Smith 1995). Non-experimental methods may produce substantial biases because of self-selection, environment differences such as differences in 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

[1] Heckman and Robb 1985, Holland 1986.
[2] See Heckman and Smith (1995) and Heckman, LaLonde and Smith (1999) for a discussion.

(Project STAR) was an experimental study on the impact of reduced class size on test scores. We present a brief review of these studies.

A. 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 voluntary participation, 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. Fraker and Maynard (1987), LaLonde (1986) and LaLonde and Maynard (1987) raise serious concerns about the reliability of non-experimental evaluation methods. LaLonde's paper combines experimental data from the NSW Demonstration to non-experimental data on non-participants drawn from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). He shows 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. In response to LaLonde, Heckman, Hotz and Dabos (1989) and Heckman and Hotz (1989) 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 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. Their results show that nearest neighbor and radius matching do reasonably well in yielding accurate estimates of the treatment effect in a non-experimental setting, despite the fact that the comparison units come from different labor markets, the survey instruments between NSW, CPS and PSID are different and the set of covariates is not particularly rich. It is also important to bear in mind that their standard errors are too big relative to their point estimates to draw unambiguous conclusions.

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 PSM. Using experimental data from the JTPA experiment and three groups of non-experimental units,[3] they find that like for other non-experimental estimators, PSM performs best 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. They also 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.

[3] 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.

Smith and Todd (2003) reconcile the contradictory evidence on the performance of PSM as reported in Heckman et al. and Dehejia and Wahba studies. Using NSW, CPS and PSID samples from LaLonde's study they find that the results in the Dehejia and Wahba are particularly sensitive to the choice of their sample and conditioning variables. In particular, they find that sample restrictions imposed by Dehejia and 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).

B. Mandatory programs

Assessments based on mandatory interventions such as Welfare-to-Work provide evidence on a different type of selection bias. In this case, given that participation is mandatory the problem of a non-experimental study is to find welfare recipients from non-participant locations similar enough to welfare recipients from participant locations. In this context, selection bias arises mainly due to geographic differences in labor markets and not by self-selection. In our application, we are concerned with this type of selection bias.

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. They stress the importance of ensuring similarity of local conditions when using non experimental comparison groups.

Michalopoulos et al. (2004) 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.

Wilde and Hollister (2002) provide evidence on the performance of PSM for a different type of intervention. They apply this technique to estimate non-experimentally the treatment effect of reduced class size on achievement test scores using experimental data for kindergartners from schools in Tennessee's Project STAR. For each of their 11 schools with 100 or more kindergartners, they construct comparison groups using out-of-school units; that is, they combine treatment children from a given school with control children from all other schools. They conclude that propensity score matching estimates of the treatment effect differ substantially from the experimental estimate.

In summary, assessments of PSM as a reliable impact estimator are lukewarm at best. Surprisingly matching 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 covariates.

II. 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).[4] PROGRESA differed from previous national poverty programs in two major 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.

A. 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, starting at about USD12 per month for primary school and increasing to USD22 per month for middle school attendance, with girls receiving slightly higher subsidies (by about USD2 per month) than 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 (USD13 per school semester). The average transfer represents about one-third of total monthly household income.

B. 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.

C. 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.

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

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.[5]

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 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.

D. The social experiment

During the second phase of incorporation (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, Querétero, 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.[6]

The impact evaluation of PROGRESA was conducted independently by the International Food Policy Research Institute (IFPRI), 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. We consider these two outcomes in our assessment.[7]

[5] See Skoufias, Davis & de la Vega (2001) for an assessment of the Progresa targeting procedure.
[6] See Behrman & Todd (1999) for an assessment of the randomization process.
[7] 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.

III. METHODOLOGY AND DATA

A. Propensity score matching

PSM is a non-parametric estimation method that works by re-weighting the comparison sample to provide an estimate of the counterfactual of interest--what the outcome of a beneficiary household would have been had it not received program benefits. The identification assumption of PSM is that outcomes in the untreated state are independent of program participation conditional on a particular set of observable characteristics. This is the conditional independence assumption, the ignorable treatment assignment (Rosenbaum and Rubin 1983), and the assumption of selection on observables (Heckman and Robb 1985). 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. Actually, we require an even weaker condition to identify our treatment parameter, that of 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 get an estimate of the unobserved component in the TT parameter. In particular, we can identify the parameter as follows:

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

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 a national household survey (see below). The estimated bias can be expressed as:

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

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 selection bias.

B. The balancing score and the common support region

We implement the matching procedure using a balancing score estimated from a logit model. 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 $\log\big(p/(1-p)\big)$, for each observation in the control and comparison samples. All the experimental units we use are poor households but not every non-experimental unit is, 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 we use in X are precisely the variables that PROGRESA uses in calculating its point score to determine household eligibility.

1. 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.

2. 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. We define the region of common support by dropping observations below the maximum of the minimums and above the minimum of the maximums of the balancing score.[8]

C. 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:

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

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 (formal equations are presented in the appendix):

[8] Although this definition does not seem restrictive in our particular application, it might entail 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.

a) 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.

b) Caliper matching. This estimator chooses the nearest neighbor inside a caliper or tolerance width δ. This is an alternative way of imposing the common support condition.

c) 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.

d) 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.

To account for the fact that the balancing score is estimated, we use the bootstrap method to estimate standard errors for all matching estimators reported below. 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. We repeat this procedure 100 times to obtain the standard errors.

D. Data and 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 (poor) treatment and control households from 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 matching, which we discuss later.

The non-experimental sample comes from a biannual nationally representative household survey that collects information on income, expenditures, household demographic composition, and school enrollment (*Encuesta Nacional sobre Ingresos y Gastos de los Hogares* - ENIGH). We use rural households from the 1998 wave of ENIGH to construct the non-experimental comparison group. The overall sample size in ENIGH is approximately 13,000 households of which 4,000 reside in rural localities.

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 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 these latter localities (if there are any) 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, it will be interesting to find out whether the technique identifies enough good matches from ENIGH to allow for meaningful comparisons with the control group from ENCEL.

1. Differences in questionnaire design across survey instruments

Aside from differences in the sample frame, which may inhibit good matches, a critical issue is the different questionnaire design between the ENCEL and ENIGH surveys. The expenditure module in ENIGH is much more detailed than in ENCEL, and while the surveys were fielded at around the same time of year, many of the recall periods are also different. 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 will allow us to assess whether the results from PSM are sensitive to data quality and variations in questionnaire design, an important issue stressed by Heckman, Ichimura, Smith and Todd (1998) and Smith and Todd (2003).

E. Selection on observables

Matching will perform well if the assumption of selection on observables is valid. This is a reasonable assumption in the PROGRESA context because the incorporation of poor households into the program is based only on observable locality and household characteristics, the program is mandatory, enrollment is supply-driven, and there is little noncompliance with treatment assignment, that is, household self-selection is not a major concern. The real concern will be to find eligible households from non--experimental localities similar enough to those in the experiment. In these terms our problem is closely related to that addressed in Friedlander and Robins (1995) and Michalopoulos et al. (2004). Our goal is to construct a sample of potential beneficiary households in non-treated localities from the non experimental sample such that poverty levels and associated household characteristics in these localities are similar to those in experimental localities. Accordingly, we construct a non experimental comparison group that resembles poor households in the untreated state (experimental control units) using information from the nationwide household survey.

Our estimation procedure consists of two steps. First, we use the non-experimental national survey in order to select a sample of rural localities for which the marginality index makes them qualify to receive program benefits but which were not enrolled in the program during the first two phases of incorporation. At this stage, we use PROGRESA's administrative records to

identify the phase in which each rural locality from the non-experimental survey actually entered the program. We select two non-experimental samples: one that only excludes rural localities already incorporated by PROGRESA to avoid contamination bias, and another that further excludes localities that never enter the program in order to eliminate (or at least mitigate) the selection bias that might arise because of differences across localities in terms of community-level poverty.

Second, based on PROGRESA's targeting mechanism at the household level, we construct a balancing score using the same observable characteristics used by the program to determine program eligibility. We apply PSM to construct a comparison group of households from non-experimental localities. At this stage, we combine experimental data from the program and non-experimental data from the national household survey to estimate households' eligibility (the probability of being poor) within targeted localities and then match control and treatment units to the non-experimental comparison units.

IV. RESULTS

A. Mean observable characteristics by 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 conditioning 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 are 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, although this sample is also clearly better off than the ENCEL controls. (We comment on columns 6 and 7 below.)

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. However, this table also illustrates the potential problem of different questionnaire designs between the two surveys. The ENCEL questionnaire breaks down children's clothing into different categories in order to obtain a more precise measure of this outcome while the ENIGH only offers one question on this category; Table 2 shows higher mean spending on children's clothing among the much poorer ENCEL control group relative to the richer ENIGH sample. On the other hand, the ENIGH is more detailed on the issue of employment status and paid employment, and Table 2 reports the same or slightly higher rates of child work for pay among the richer ENIGH sample compared to the poorer ENCEL control sample (column 2).

B. Balancing score and common support

Results of the logit models to determine the probability of qualifying for the program are reported in Table A1 in the Appendix. 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 using ENIGH-Sample 1 and ENIGH-Sample 2 respectively.

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 A2 in the Appendix 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 -0.710 for ENIGH households and around 3.2 for both control and treatment households from ENCEL; 0.1% of the control group and 12.6% of the non experimental comparison group do not satisfy the common support criteria and are 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 imposing the common support criteria results in the elimination of 2.6% of the control and only 1% of the comparison groups; the latter number is naturally due to the screening out of 'rich' households from this sample who would otherwise be excluded by the common support condition.

C. 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 incomplete secondary schooling is around 4-6%, compared to 5.5% in ENCEL and 12% in the overall rural sample from ENIGH. Similarly, the proportion of matched households without social security is 96% compared to 97% in ENCEL and 78% 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. Mean food expenditure is significantly lower in the matched samples ($696 and $647 respectively) relative to the full ENIGH samples in columns 3-5, but these means are still quite large relative to the control mean of $477 in column 2. In the case of school enrollment for older kids, the non experimental comparison group means are 0.48 and 0.40 (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 that the control group mean. For child labor the matched sample means are 0.11 (in both sample 1 and sample 2) compared to 0.12 in the control group and 0.11 in the overall rural ENIGH; here child labor is actually lower in the matched samples relative to the control group.

D. Bias estimates

1. Bias estimates for household outcomes

Table 3 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 3 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 3. 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 although not identical fashion across the two surveys.

While the main objective of this article 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. Using the sample of control and comparison units only, we regress each of the household expenditure outcomes on the same set of covariates used in the logit regression shown in Table A1, along with a 'control group' dummy variable; the coefficient estimate (and standard error) of this dummy variable is reported in column 9 of Table 3. 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 and about the same for the others. Also of note is that the unadjusted mean differences between control and comparison group reported in column 8 are not always larger than those from nearest neighbor matching or regression—see for example meat and children's clothing measured in levels.

Estimates based on the more restrictive comparison group from ENIGH-Sample 2 are shown in Table 4. These results follow the same general pattern as those in Table 3, 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 design may be an important determinant of the performance of PSM.

Comparing across matching techniques and focusing on the results in Table 3, we find very little difference in the point estimates of the bias. Caliper matching tends 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 4.

2. Bias estimates for individual outcomes

Tables 5 (sample 1) and 6 (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 5 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 higher than the rate among control children.

In the middle panel the results indicate significant differences for the 'never enrolled' outcome among all children which is also driven by differences among the younger age group. The other statistically significant differences are for child labor (bottom panel) among all kids and the sub-sample of girls age 12-16 based on local linear (0.2) matching. These latter 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.

Column 9 of Table 5 presents the regression adjusted estimates of bias derived from individual probit regressions using all the covariates of the logit regression in Table A1 plus a dummy indicator for a control group observation. 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 two questionnaires 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 coefficients for the estimated bias in the bottom panel of Table 5.

E. Sensitivity Analysis

In this section we report on two extensions to the analysis designed to assess the sensitivity of the results reported above. First, we repeat the analysis using a very basic set of conditioning covariates in the logit model to assess how sensitive the results are to alternative specifications of the balancing score and the availability (or absence) of a rich set of covariates. The covariates we use are what we consider the minimum information that might be available in "off-the-shelf" household surveys: age, sex and schooling of the household head, demographic composition and whether the household is covered by social security. We perform balancing tests on this model and the final specification therefore contains a number of higher order and interaction terms involving these core variables. Second, we impose a stringent criterion to set the common support region to analyze whether the (improved) composition of the underlying sample helps minimizing the bias in our PSM estimates. Figure 1a illustrates that the density of the comparison group is very thin at the upper tail of the distribution of the controls indicating that there may not be good matches in this region. We therefore eliminate the upper 25% of the distribution of controls (which occurs at an odds ratio value of 4.25) while maintaining the same lower support condition (eliminating all observations below the maximum value of the two minima).

1. Using a reduced set of covariates

The first 3 columns in Table 7 report a summary of results with a reduced set of covariates used to estimate the balancing score, along with the unadjusted mean differences in the first row for comparison. The food expenditures (column 1) point estimates of bias are significantly higher than those using a richer set of covariates (almost twice the size of those reported in Table 3); indeed, with this specification PSM gives point estimates that are closer to the unadjusted estimate in row 1. Note however that local linear and kernel matching do much better here, indicating that these techniques might mitigate some of the problems of poor data, although clearly not enough to eliminate the bias. The results for school enrolment in column 2 are generally the same; nearest-neighbor matching does very poorly once again, with PSM not providing any improvement over the straight unadjusted difference in means. The other techniques perform better than nearest neighbor, but the caliper and kernel bias estimates now become statistically significant. Finally, the results for child employment (column 3) are actually the only ones that are comparable to the initial results using the full set of controls, and here even nearest neighbor seems to give unbiased estimates of program impact. Taken as a whole however, these results show that a rich set of relevant covariates is an important determinant of the success of the matching technique.

2. Stringent common support criterion

Results using the more stringent common support regime are shown in the last 3 columns of Table 7. Even this criterion does not reduce the estimated bias related to food expenditure, with point estimates roughly the same as those reported in Table 3, nor is there any significant reduction in bias associated with the more complex matching techniques. The results are somewhat more encouraging for the individual outcomes reported in columns 5 and 6, where none of the point estimates of bias are significantly different from zero, in contrast to Table 5 where the employment outcomes are significant for kernel and local linear matching. Taken as a

whole these results suggest that the underlying composition of the comparison group is important regardless of matching technique, but that even more restrictive support conditions do not affect the results for outcomes that are measured in very different ways (i.e. food expenditures).

V. CONCLUSIONS

The reliability of non-experimental evaluation estimators, in general, is an important issue in the evaluation literature because of the potential difficulties in launching social experiments. The reliability of propensity score matching is of particular importance given the growing interest in this method as a substitute for randomized evaluations. Almost all the published studies on this issue have used experimental data from employment and training programs (voluntary or mandatory) from the U.S. The results from these studies are not encouraging. Indeed, the minimum conditions under which PSM may be viable also apply to other non-experimental methods - the availability of a rich set of conditioning variables, the use of similar survey instruments and control for local economic conditions.

In this paper, we present further evidence on the performance of cross-sectional PSM outside the scope of employment and training interventions and from a country other than the U.S. We use experimental control data from PROGRESA, a nationwide and mandatory intervention aimed to reduce poverty in Mexico, and compare them to non-experimental comparison units from a nationally representative household survey (ENIGH). In this context, selection bias might arise because of differences between participating and non-participating localities and not because of individual self-selection. Our results reveal that even when PSM narrows the unconditional differences between control and comparison units, significant bias remains in the matching estimates for outcomes that are measured differently (expenditures). We find slightly more encouraging results for children's schooling outcomes, which are measured in the same way across surveys. However, even for these outcomes we find bias in the matching estimates for school enrollment behavior (ever enrolled and current enrollment) of children 8-12 years old, where PSM significantly underestimates true (experimental) program impacts. On the other hand, there are no statistically significant biases for school enrollment behavior among 13-16 years old children, where PROGRESA has the largest impact. For child labor, measured in a similar but not identical way across survey instruments, we find some evidence of bias using kernel and local-linear matching, and these imply overestimation of true program impact. This may be related to the extra effort in the ENIGH survey to capture paid employment.

We pursue two types of sensitivity analysis. First, reducing the set of conditioning variables used to estimate the balancing reveals that having a rich set of covariates does matter, we find much larger point estimates for food expenditure and school enrolment when using the reduced set of covariates than we obtain using the richer one. Second, imposing a more restrictive common support criterion does not help eliminate the bias associated with food expenditure but does well for the individual outcomes, which tend to be measured in similar ways across questionnaires. Of course, it is important to keep in mind that the estimation of treatment effects on a restrictive support region fundamentally changes the parameter being estimated and thus the estimated effect may not be representative of all program participants.

There are several important implications of the results presented here. First, we have been able to corroborate the main - and rather pessimistic - conclusions of the existing literature on the performance of PSM as a non-experimental impact estimator. In our case, even though we have the exact same variables used by PROGRESA to select program beneficiaries, we are not able to replicate the experimental estimates for expenditure outcomes and some schooling and child

labor outcomes. Second, conditional cash transfers anti-poverty programs similar to PROGRESA are spreading rapidly around Latin America and other parts of the developing world, as is the interest to evaluate the impacts of this type of intervention. Our analysis implies that evaluating these interventions using PSM must be done with caution, with particular care taken to ensure the consistency of survey instruments and to justify the assumption of selection on observables.[9]

[9] Note that difference-in-difference matching, which Smith and Todd argue can eliminate time invariant sources of bias, such as differences in questionnaire design or local labor markets, will also not be available for these universal programs unless the phasing up period is very lengthy.

REFERENCES

Behrman, Jere and Petra Todd (1999) "Randomness in the Experimental Samples of PROGRESA." Research Report, International Food Policy Research Institute. Washington D.C.

Burtless, Gary (1995) "The Case for Randomized Field Trials in Economic and Policy Research." Journal of Economic Perspectives Vol.9(2): 63-84.

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.

Fraker, Thomas, and Rebecca Maynard (1987), "The Adequacy of Comparison Group Design for Evaluations of Employment-Related Programs," Journal of Human Resources Vol. 22(2): 194-227.

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 & Richard Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions," in James Heckman & Burton Singer (eds) Longitudinal Analysis of Labor Market Data. Cambridge, England: Cambridge University Press, pp. 156-246.

Heckman, James, Joseph Hotz and Marcelo Dabos (1987), "Do We Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?" Evaluation Review Vol. 11(4): 395-427.

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) "Charactrizing Selection Bias Using Experimental Data," Econometrica. 66: 1017-1089.

Heckman, James, Robert LaLonde and Jeffrey Smith (1999), "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card, eds., Handbook of Labor Economics, Vol.3A: 1865--2097.

Heckman, James and Jeffrey Smith (1995) "Assessing the Case for Social Experiments," Journal of Economic Perspectives Vol.9(2): 85-110.

Holland, Paul (1986), "Statistics and Causal Inference." Journal of the American Statistical Association Vol.81: 945-970.

LaLonde, Robert, (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," American Economic Review, Vol.76: 604-620.

LaLonde, Robert, and Rebecca Maynard (1987), "How Precise are Evaluations of Employment and Training Programs," Evaluation Review Vol. 11(4): 428-451.

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 & Statistics Vol.86(1):156-179.

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." Biometrica. Vol. 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, Banjamin 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.

Wilde, Elizabeth & Robinson Hollister (2002), "How Close is Close Enough? Testing Non Experimental Estimates of Impact against Experimental Estimates of Impact with Educational Test Scores as Outcomes," Institute for Research on Poverty, DP # 1242-02.

TABLES AND FIGURES

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	1.487	1.144	1.036	1.119	1.519	1.609
	(0.95)	(0.98)	(1.05)	(0.95)	(1.04)	(1.00)	(1.15)
Head's sex (female)	0.083	0.085	0.130	0.134	0.148	0.096	0.099
	(0.28)	(0.28)	(0.34)	(0.34)	(0.36)	(0.29)	(0.30)
Head's schooling							
None	0.448	0.460	0.402	0.404	0.399	0.455	0.515
	(0.50)	(0.50)	(0.49)	(0.49)	(0.49)	(0.50)	(0.50)
Incomplete primary	0.248	0.238	0.203	0.223	0.191	0.262	0.211
	(0.43)	(0.43)	(0.40)	(0.42)	(0.39)	(0.44)	(0.41)
Incomplete secondary	0.055	0.055	0.104	0.120	0.091	0.040	0.054
	(0.23)	(0.23)	(0.31)	(0.32)	(0.29)	(0.20)	(0.23)
Head'sage	42.204	42.529	47.141	47.210	47.421	41.509	41.699
	(14.58)	(14.88)	(16.26)	(16.37)	(16.72)	(14.11)	(14.08)
Number of kids ages 13 or below	2.457	2.489	1.488	1.341	1.431	2.567	2.541
	(1.66)	(1.61)	(1.54)	(1.43)	(1.49)	(1.61)	(1.55)
Crowding index	4.399	4.455	2.616	2.364	2.453	4.375	4.164
	(2.26)	(2.26)	(1.86)	(1.71)	(1.69)	(2.16)	(1.99)
Do not have social security	0.969	0.960	0.821	0.776	0.867	0.955	0.962
	(0.17)	(0.20)	(0.38)	(0.42)	(0.34)	(0.21)	(0.19)
No bathroom	0.482	0.489	0.346	0.289	0.412	0.568	0.525
	(0.50)	(0.50)	(0.48)	(0.45)	(0.49)	(0.50)	(0.50)
Bathroom no water	0.500	0.493	0.482	0.478	0.452	0.417	0.452
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)
Dirt floor	0.729	0.754	0.255	0.203	0.214	0.751	0.741
	(0.44)	(0.43)	(0.44)	(0.40)	(0.41)	(0.43)	(0.44)
Without gas stove	0.847	0.834	0.362	0.260	0.285	0.842	0.837
	(0.36)	(0.37)	(0.48)	(0.44)	(0.45)	(0.37)	(0.37)
Without refrigerator	0.959	0.962	0.559	0.474	0.566	0.960	0.971
	(0.20)	(0.19)	(0.50)	(0.50)	(0.50)	(0.20)	(0.17)
Without washer	0.986	0.988	0.762	0.687	0.783	0.983	0.985
	(0.12)	(0.11)	(0.43)	(0.46)	(0.41)	(0.13)	(0.12)
Without vehicle	0.979	0.980	0.789	0.737	0.773	0.941	0.942
	(0.14)	(0.14)	(0.41)	(0.44)	(0.42)	(0.24)	(0.23)
Observations	7703	4604	3837	2438	724	765	371

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
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Household outcomes							
Food expenditure per capita	511.8	476.7	909.7	970.7	880.8	695.9	646.6
	(399.1)	(405.2)	(697.8)	(731.7)	(664.7)	(671.2)	(619.4)
Children's clothing per capita	20.4	14.8	5.3	6.2	6.0	3.7	4.2
	(35.5)	(28.6)	(23.0)	(26.3)	(22.9)	(12.4)	(12.5)
Percentage of kids 8-16 in school	0.785	0.744	0.806	0.800	0.778	0.781	0.785
	(0.31)	(0.33)	(0.32)	(0.33)	(0.35)	(0.32)	(0.32)
Observations	{7703}	{4604}	{3837}	{2438}	{724}	{765}	{371}
B. Children outcomes							
School enrolment							
Children 8-16	0.772	0.727	0.795	0.792	0.766	0.754	0.745
	(0.42)	(0.45)	(0.40)	(0.41)	(0.42)	(0.43)	(0.44)
	{13553}	{8126}	{4407}	{2577}	{786}	{1058}	{506}
Children 8-12	0.921	0.891	0.948	0.947	0.948	0.903	0.921
	(0.27)	(0.31)	(0.22)	(0.22)	(0.22)	(0.30)	(0.27)
	{8183}	{4876}	{2563}	{1481}	{460}	{659}	{313}
Children 13-16	0.545	0.480	0.582	0.581	0.509	0.475	0.396
	(0.50)	(0.50)	(0.49)	(0.49)	(0.50)	(0.50)	(0.49)
	{5370}	{3250}	{1844}	{1096}	{326}	{399}	{193}
Work for pay							
All Children 12-16	0.111	0.116	0.107	0.121	0.124	0.114	0.111
	(0.31)	(0.32)	(0.31)	(0.33)	(0.33)	(0.31)	(0.32)
	{7004}	{4246}	{2402}	{1423}	{428}	{458}	{198}
Boys 12-16	0.164	0.180	0.141	0.155	0.135	0.176	0.147
	(0.37)	(0.38)	(0.35)	(0.36)	(0.34)	(0.38)	(0.36)
	{3628}	{2146}	{1210}	{708}	{207}	{267}	{116}
Girls 12-16	0.054	0.051	0.073	0.090	0.113	0.082	0.085
	(0.23)	(0.22)	(0.26)	(0.28)	(0.32)	(0.27)	(0.28)
	{3376}	{2100}	{1192}	{715}	{221}	{281}	{117}

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

Table 3: Direct estimates of the bias for household level outcomes -- sample 1

Matching method:	Nearest Neighbor	Caliper		Local Linear		Kernel		Unadjusted	Regression
		d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2	Difference	Adjusted
Expenditure	(1)	(2	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Food	**-219.114**	**-223.855**	**-238.126**	**-226.493**	**-216.869**	**-215.466**	**-214.611**	**-494.168**	**-270.819**
	(33.772)	(29.089)	(44.568)	(31.438)	(31.252)	(29.607)	(29.361)	(13.419)	(18.342)
Vegetables	**70.698**	**71.220**	**72.578**	**69.177**	**69.382**	**69.484**	**69.601**	**54.009**	**79.202**
	(2.132)	(2.186)	(5.305)	(1.968)	(1.862)	(1.950)	(1.882)	(2.082)	(2.966)
Cereals	21.038	**22.267**	20.517	**22.045**	**26.901**	**26.520**	**27.595**	6.073	**29.901**
	(11.872)	(9.895)	(16.995)	(9.769)	(10.106)	(8.988)	(8.907)	(4.981)	(7.121)
Meat	-9.088	-7.745	-3.251	-5.391	-3.957	-4.597	-3.495	**-85.976**	-10.700
	(8.382)	(7.001)	(11.464)	(7.526)	(7.627)	(7.669)	(7.487)	(4.11)	(5.739)
Kid clothes	**11.158**	**11.265**	**11.505**	**10.916**	**11.218**	**11.116**	**11.133**	**8.560**	**12.086**
	(0.798)	(0.789)	(1.571)	(0.688)	(0.643)	(0.658)	(0.657)	(0.690)	(0.982)
Expenditure shares									
Food	**0.267**	**0.263**	**0.248**	**0.262**	**0.268**	**0.263**	**0.263**	**0.312**	**0.233**
	(0.015)	(0.012)	(0.017)	(0.013)	(0.013)	(0.013)	(0.013)	(0.005)	(0.006)
Vegetables	**0.120**	**0.121**	**0.122**	**0.119**	**0.119**	**0.119**	**0.119**	**0.115**	**0.124**
	(0.001)	(0.002)	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Cereals	**0.195**	**0.195**	**0.191**	**0.194**	**0.197**	**0.195**	**0.196**	**0.223**	**0.183**
	(0.007)	(0.007)	(0.011)	(0.006)	(0.006)	(0.005)	(0.005)	(0.004)	(0.005)
Meat	**0.069**	**0.070**	**0.065**	**0.068**	**0.070**	**0.069**	**0.069**	**0.058**	**0.069**
	(0.006)	(0.005)	(0.009)	(0.006)	(0.005)	(0.006)	(0.005)	(0.003)	(0.004)
Kids clothes	**0.018**	**0.018**	**0.018**	**0.018**	**0.018**	**0.018**	**0.018**	**0.019**	**0.018**
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Kids' schooling									
Percent enrolled	-0.037	-0.045	-0.051	-0.023	-0.026	-0.026	-0.030		
	(0.023)	(0.022)	(0.033)	(0.022)	(0.021)	(0.020)	(0.020)		
Percent never enrolled	-0.024	-0.017	-0.013	-0.019	-0.018	-0.016	-0.014		
	(0.016)	(0.013)	(0.016)	(0.013)	(0.012)	(0.012)	(0.011)		
Child labor	-0.007	0.004	-0.013	-0.026	-0.017	-0.028	-0.019		
(% working for pay)	(0.026)	(0.021)	(0.036)	(0.021)	(0.020)	(0.021)	(0.020)		

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. In column 1 768 non-experimental controls were matched an average of 6 times and the maximum times a unit was matched was 57. In column 3 564 non experimental control units were matched an average of 2.18 times, and the maximum match was 12 times. Column 8 is the mean difference computed from columns 2 and 4 in Table 2; column 9 is the regression adjusted mean difference using ENIGH sample 1—see text for details.

Table 4: Direct estimates of the bias for household level outcomes -- sample 2

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

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. In column, 363 non experimental comparison units were matched an average of 12 times an the maximum times a unit was matched was 195. In column 3 207 comparison units were matched an average of 2 times and the maximum match was 12.

Table 5: Estimates of the bias for individual schooling and work outcomes -- sample 1

Matching method:	Nearest Neighbor	Caliper		Local Linear		Kernel			Regression
		d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2	Unadjusted	Adjusted
Currently enrolled	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
All kids 8-16	-0.035	**-0.078**	-0.074	-0.024	-0.024	-0.028	-0.031	**-0.063**	**-0.052**
	(0.030)	(0.025)	(0.042)	(0.015)	(0.015)	(0.015)	(0.015)	(0.010)	(0.012)
Kids 8-12	-0.041	**-0.036**	-0.035	**-0.034**	-0.017	-0.024	-0.027	**-0.056**	**-0.034**
	(0.024)	(0.015)	(0.030)	(0.016)	(0.017)	(0.017)	(0.016)	(0.009)	(0.009)
Kids 13-16	0.022	-0.018	-0.003	0.026	0.013	0.014	0.011	**-0.098**	**-0.047**
	(0.041)	(0.039)	(0.073)	(0.026)	(0.027)	(0.027)	(0.027)	(0.017)	(0.024)
Never enrolled									
All kids 8-16	**-0.041**	-0.009	-0.007	-0.026	**-0.029**	-0.026	**-0.024**	0.005	**-0.011**
	(0.020)	(0.015)	(0.017)	(0.014)	(0.014)	(0.014)	(0.012)	(0.004)	(0.005)
Kids 8-12	-0.024	**-0.033**	-0.020	**-0.042**	**-0.044**	**-0.042**	**-0.038**	**-0.011**	**-0.022**
	(0.019)	(0.013)	(0.022)	(0.016)	(0.016)	(0.016)	(0.014)	(0.004)	(0.007)
Kids 13-16	0.011	0.008	0.012	-0.002	-0.007	-0.002	-0.003	**0.027**	0.005
	(0.024)	(0.020)	(0.033)	(0.014)	(0.015)	(0.014)	(0.014)	(0.007)	(0.007)
Work for pay									
All kids 12-16	-0.028	0.011	-0.030	**-0.062**	**-0.054**	**-0.051**	-0.041	-0.006	-0.014
	(0.029)	(0.023)	(0.039)	(0.023)	(0.023)	(0.021)	(0.021)	(0.010)	(0.013)
Boys 12-16	-0.075	-0.033	0.006	-0.057	-0.052	-0.043	-0.030	0.025	0.006
	(0.048)	(0.036)	(0.077)	(0.039)	(0.039)	(0.039)	(0.039)	(0.016)	(0.022)
Girls 12-16	-0.010	-0.005	-0.018	-0.034	**-0.045**	-0.020	-0.026	**-0.039**	**-0.028**
	(0.022)	(0.023)	(0.048)	(0.019)	(0.021)	(0.017)	(0.017)	(0.010)	(0.014)

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. In column 1, 659 comparison group units were matched an average of 12 times and the maximum times a unit was matched was 127. In column 3, 356 comparison group units were matched an average of 4 times and the maximum match was 15. Column 8 is the mean difference computed from columns 2 and 4 in Table 2; column 9 is the regression adjusted mean difference using ENIGH sample 1—see text for details.

Table 6: Estimates of the bias for individual schooling and work outcomes – sample 2

Matching method:	Nearest Neighbor	Caliper d=0.01	Caliper d=0.001	Local Linear bw=0.1	Local Linear bw=0.2	Kernel bw=0.1	Kernel bw=0.2
Currently enrolled							
All kids 8-16	-0.051	**-0.085**	-0.074	-0.010	-0.016	-0.012	-0.015
	(0.057)	(0.035)	(0.065)	(0.036)	(0.035)	(0.041)	(0.034)
Kids 8-12	-0.023	**-0.071**	**-0.106**	-0.016	-0.016	-0.018	-0.023
	(0.042)	(0.026)	(0.048)	(0.034)	(0.032)	(0.037)	(0.031)
Kids 13-16	0.037	0.086	0.092	0.058	0.030	0.057	0.047
	(0.077)	(0.064)	(0.142)	(0.056)	(0.055)	(0.063)	(0.056)
Never enrolled							
All kids 8-16	-0.017	-0.000	-0.012	-0.016	-0.016	-0.016	-0.014
	(0.021)	(0.020)	(0.034)	(0.017)	(0.017)	(0.016)	(0.016)
Kids 8-12	**-0.046**	-0.013	-0.013	-0.024	-0.024	-0.022	-0.021
	(0.020)	(0.018)	(0.030)	(0.018)	(0.018)	(0.016)	(0.017)
Kids 13-16	-0.009	0.006	-0.041	-0.009	-0.006	-0.013	-0.007
	(0.028)	(0.033)	(0.065)	(0.024)	(0.023)	(0.024)	(0.023)
Work for pay							
All kids 12-16	0.012	0.031	0.021	0.017	0.046	0.013	0.021
	(0.036)	(0.030)	(0.067)	(0.028)	(0.024)	(0.031)	(0.028)
Boys 12-16	0.033	0.071	0.020	0.040	0.060	0.036	0.039
	(0.051)	(0.049)	(0.147)	(0.043)	(0.042)	(0.049)	(0.046)
Girls 12-16	-0.019	0.011	-0.065	-0.022	-0.018	-0.029	-0.022
	(0.044)	(0.038)	(0.086)	(0.036)	(0.031)	(0.041)	(0.037)

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. In column 1, 309 comparison units were matched an average of 25 times; in column 3, 117 comparison units were matched an average of 14 times.

Table 7: Summary of bias estimates with alternative specification and common support regime

	Reduced set of covariates[1]			Alternative common support[2]		
	Food expenditure	School Enrolment (13-16)	Work for pay (12-16)	Food expenditure	School Enrolment (13-16)	Work for pay (12-16)
	(1)	(2)	(3)	(4)	(5)	(6)
Unadjusted – sample 1	**-494.17**	**-0.098**	-0.006	**-494.17**	**-0.098**	-0.006
	(13.42)	(0.017)	(0.010)	(13.42)	(0.017)	(0.010)
Nearest neighbor	**-466.96**	**-0.105**	-0.004	**-228.95**	-0.010	-0.004
	(26.49)	(0.036)	(0.023)	(38.99)	(0.048)	(0.030)
Nearest neighbor plus	**-424.53**	-0.019	0.003	**-222.31**	0.108	-0.011
Screening (sample 2)	(42.41)	(0.049)	(0.028)	(80.37)	(0.074)	(0.051)
Caliper matching	**-465.29**	-0.020	-0.017	**-256.61**	-0.008	-0.030
(d=0.001)	(24.99)	(0.038)	(0.021)	(45.45)	(0.082)	(0.050)
Local linear (bw=0.2)	**-416.99**	**-0.074**	-0.001	**-228.84**	-0.002	0.011
	(21.42)	(0.026)	(0.014)	(30.75)	(0.035)	(0.018)
Kernel matching	**-416.14**	**-0.070**	0.003	**-236.70**	0.002	-0.016
(bw=0.2)	(19.92)	(0.025)	(0.013)	(31.29)	(0.033)	(0.019)

1/ Balancing score logit estimated with reduced set of covariates. 2/ Observations above the 75th percentile of the distribution of controls excluded from the sample in addition to observations below the maximum of the minima among the two distributions. All estimates are based on sample 1 except for screening, which uses sample 2. Unadjusted differences in the first row are taken from column 8 in Tables 3 and 5. Statistically significant (5%) estimates in bold.

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

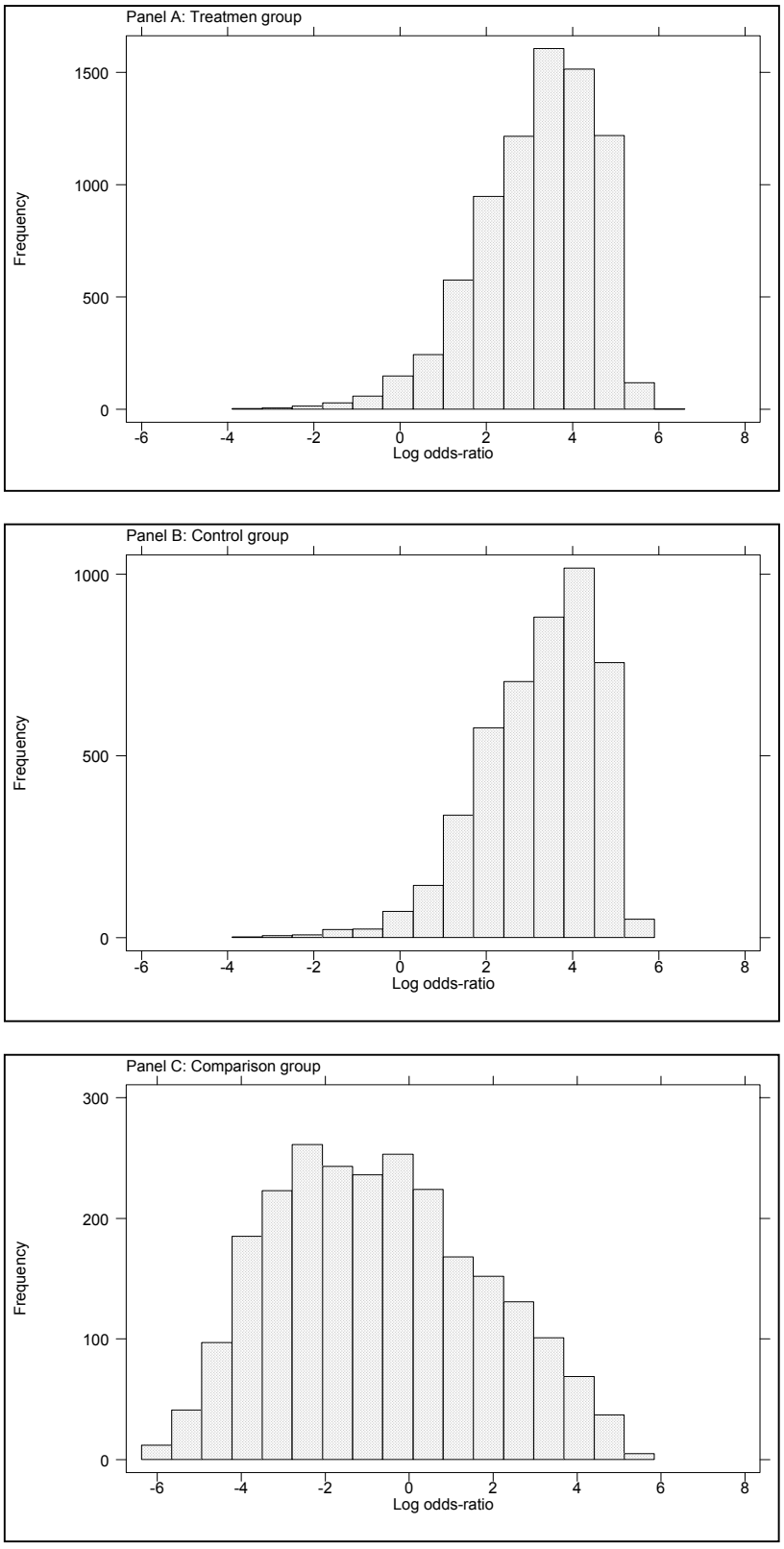
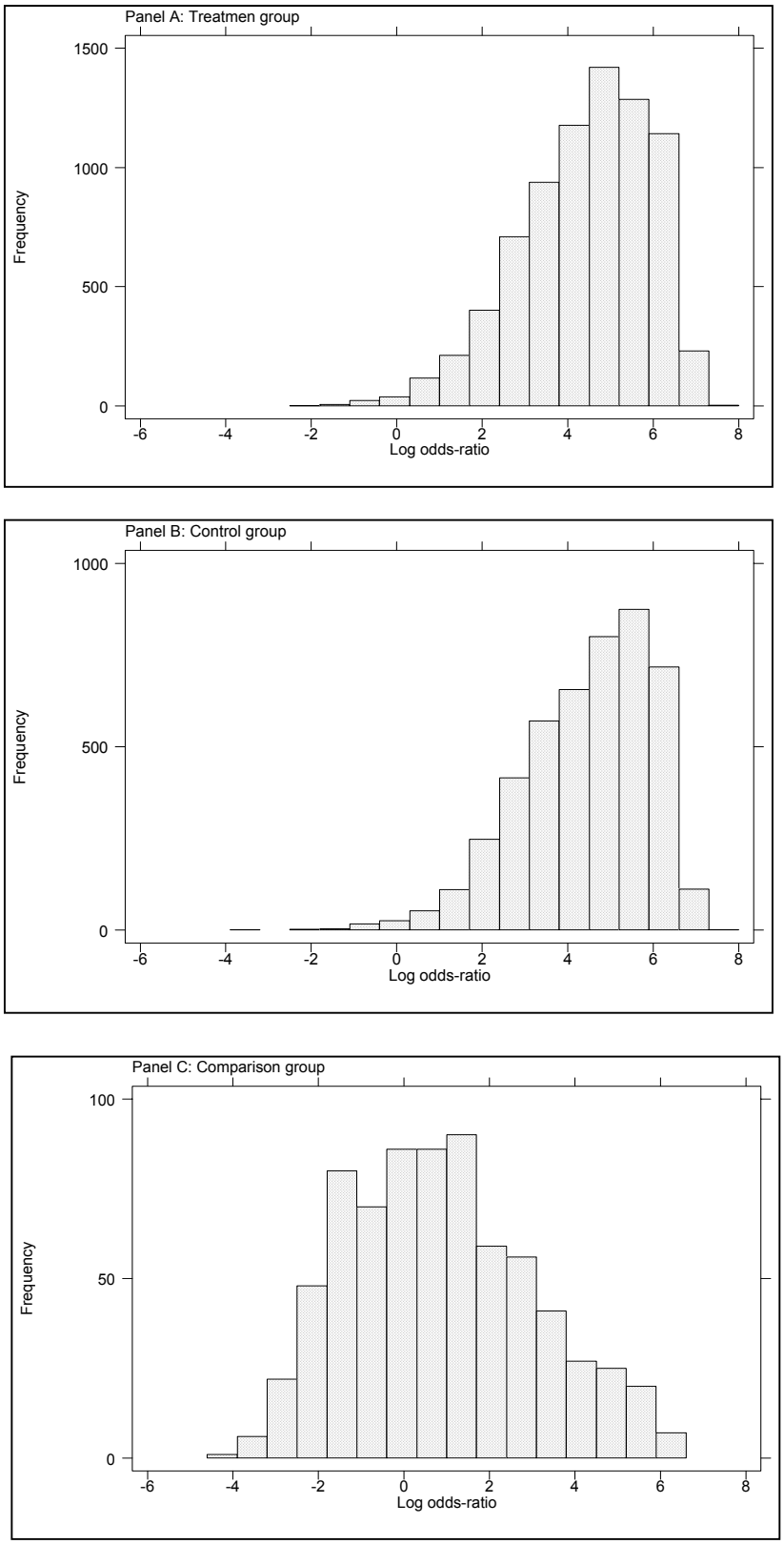


Figure 1b: Empirical density of estimated log odds-ratio: sample 2



APPENDIXES

APPENDIX 1: Alternative Matching Estimators

Nearest-Neighbor estimator

Let $C(P_i)$ denote the neighborhood for each treatment unit i, which consists of comparison units $j \in I_0$ for whom $P_j \in C(P_i)$. The nearest-neighbor matching estimator sets

$$C(P_i) = \{j : \left\|P_i - P_j\right\| = \min_{k \in I_0} \left\|P_i - P_k\right\|\},$$

in this case there is only one comparison unit matched to each treatment unit. Weights are given by $W(i,j) = 1[j = k]$, where $1[\cdot]$ is the indicator function.

Caliper estimator

The caliper matching estimator is a variation of nearest-neighbor that allows matches only under a tolerance δ on the distance $\left\|P_i - P_j\right\|$ in the attempt to avoid matches where the comparisons are "too far away" from the treatment unit (and thus, this is an alternative way of setting the support region). For caliper matching the neighborhood for treatment i is:

$$C(P_i) = \{j : \delta > \left\|P_i - P_j\right\| = \min_{k \in I_0} \left\|P_i - P_k\right\|\},$$

and a single comparison unit is matched when its distance to the treatment unit is the lowest and below the tolerance δ. If none of the comparison units is within the tolerance criterion, then treatment unit i is left unmatched. Weights are given by $W(i,j) = 1[j = k]$.

Kernel estimator

The kernel estimator matches treatment units to a kernel a weighted average of comparison units. This can be thought as a non-parametric regression of the outcome on a constant term. For the treatment units, weights are given by:

$$W(i,j) = \frac{G\left(\frac{P_j - P_i}{h_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{h_n}\right)},$$

where $G(\cdot)$ is a kernel function and h_n is a bandwidth parameter.

Local-Linear estimator

Local-Linear matching is a generalized version of Kernel matching that adds a linear term in P_i to the constant on a non-parametric regression of the outcome variable. The weights are given by:

$$W(i,j) = \frac{G_{ij}\left(\sum_{k\in I_0} G_{ik}(P_k - P_i)^2\right) - \left(G_{ij}(P_j - P_i)\right)\left(\sum_{k\in I_0} G_{ik}(P_k - P_i)\right)}{\sum_{j\in I_0} G_{ij} \sum_{k\in I_0} G_{ik}(P_k - P_i)^2 - \left(\sum_{k\in I_0} G_{ik}(P_k - P_i)\right)^2},$$

where $G_{ij} = G\left(\frac{P_j - P_i}{h_n}\right)$.

APPENDIX 2: Tables

Table A1: Logit Estimates for Balancing Score

	ENCEL + ENIGH Sample 1		ENCEL + ENIGH Sample 2	
	Coeff.	z-statistic	Coeff.	z-statistic
Dependency ratio	0.296	(3.36)	0.366	(2.83)
Head's sex	-0.130	(1.38)	-0.169	(1.25)
Head's schooling				
Complete Primary	0.468	(5.57)	0.719	(5.67)
Incomplete Secondary	0.889	(8.05)	1.149	(6.83)
Complete Secondary or more	0.662	(4.34)	0.870	(3.81)
Head's age	-0.084	(2.07)	-0.058	(0.94)
Head's age squared	0.002	(2.53)	0.002	(1.34)
Head's age cube	0.000	(2.79)	0.000	(1.58)
Number of kids ages \le 13	0.620	(10.09)	0.571	(6.27)
Crowding index	0.456	(8.63)	0.553	(7.32)
Without social security	1.521	(14.23)	1.193	(7.13)
No bathroom	0.516	(3.59)	0.135	(0.66)
Bathroom no water	0.629	(4.55)	0.455	(2.30)
Soil floor	1.257	(17.34)	1.487	(12.81)
Without gas stove	1.425	(18.58)	1.650	(13.88)
Without refrigerator	1.332	(14.52)	1.277	(9.75)
Without washer	1.076	(8.37)	0.729	(4.09)
Without vehicle	0.529	(4.35)	0.324	(1.94)
Crowding index squared	-0.011	(1.45)	-0.017	(1.45)
Crowding index \times number of kids	-0.070	(5.15)	-0.063	(2.95)
Dependency ratio cube	-0.063	(3.98)	-0.077	(3.46)
Constant	-5.947	(8.73)	-4.869	(4.71)
Number of observations	14748		13034	
Likelihood ratio test	6292		2253	
Prob.	(0.00)		(0.00)	

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

Table A2: Propensity (Balancing) Score Estimates

	Statistic				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	6.170	7690	7703	0.2%
Control	3.216	1.338	-3.590	5.874	4600	4604	0.1%
Comparison (ENIGH 1998)	-0.710	2.454	-6.375	5.613	2132	2438	12.6%
B. Matched Sample 2							
Treatment	4.410	1.502	-2.302	7.615	7448	7703	3.3%
Control	4.449	1.492	-3.207	7.355	4484	4604	2.6%
Comparison (ENIGH 1998)	0.851	2.196	-4.553	6.575	717	724	1.0%

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.

Inter-American Development Bank
Washington, D.C.