

# Estimating the Selection Bias of Matching Estimators using Experimental Data from PROGRESA\*

Juan-Jose Diaz<sup>†</sup>  
Sudhanshu Handa<sup>‡</sup>

This version: October 2004

## Abstract

Randomized experiments applied to the social sciences have become the benchmark method for estimating program impacts of policy interventions. However, not all policy questions have and 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 does well in replicating the benchmark for outcomes measured using similar survey questionnaires. However, for outcomes measured using different survey instruments, we find significant differences between the benchmark and the results based on matching.

---

\*Preliminary version, please do not quote. Comments welcome. This paper could not have been written without the assistance of Monica Orozco from OPORTUNIDADES, who provided essential data and explained operational details of the program to us. We are grateful to Jeffrey Smith, whose comments and suggestions let us improve earlier drafts of the paper. William Evans and Alex Whalley also provided useful suggestions. We also thank seminar participants at the University of Maryland, the Latin American and the Caribbean Economics Association (LACEA) 2003 Annual Meetings in Mexico, the Carolina Population Center, the Third Annual Johns Hopkins-Maryland Joint Workshop in Applied Macro- and Micro-econometrics, and the Group for the Analysis of Development (GRADE-Lima) for constructive criticism. All errors are ours.

<sup>†</sup>Graduate student, Department of Economics, University of Maryland at College Park; email: [diaz@econ.bsos.umd.edu](mailto:diaz@econ.bsos.umd.edu).

<sup>‡</sup>Associate Professor, Department of Public Policy, University of North Carolina at Chapel Hill; email: [shanda@email.unc.edu](mailto:shanda@email.unc.edu).

# 1 Introduction

After LaLonde's critique to non-experimental evaluation methods, social experiments have become the benchmark method for estimating program impacts. By randomizing observational units into treatment and control groups, social experiments provide a clean method to estimate program impacts because both observable and unobservable characteristics are uncorrelated with treatment assignment and thus, no selection bias problem arises. However, social experiments are usually not available in practice for several 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 may raise ethical concerns. Even more, randomization may also suffer from threats to external and internal validity. Consequently, testing the reliability of non experimental methods, the alternative to randomized evaluations, is a central issue in the evaluation literature. Non-experimental methods identify program impacts by imposing stronger non testable assumptions than randomization and the researcher has to make the case for justifying them in her particular application. In this context, the availability of a randomized experiment provides the opportunity to assess the validity of non-experimental assumptions and the performance of alternative impact evaluation techniques since one can compare these estimates to the experimental one.

In the U.S. literature most of the assessments of non experimental methods are based on employment and training programs for which randomized assignment is available. This literature analyzes two types of the selection bias problem in the estimation of program impacts. For voluntary interventions the problem is to find nonparticipants similar enough to program participants in the same labor market, in this case selection bias arises due mainly to self-selection. There is a recent debate in this branch of the literature about the reliability of matching methods as a non-experimental impact estimator. Dehejia and Wahba (1999 and 2002) contend that matching performs well in predicting the experimental treatment effect of the U.S. National Supported Work (NSW) Demonstration, a labor training program implemented in the 1970s. The studies by Dehejia and Wahba seems to contradict previous evidence on the reliability of matching provided by Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998). These studies find that matching performs well in replicating experimental estimator results only under certain conditions, conditions not met in the NSW data. However, Smith and Todd (2003) report that the matching estimates of Dehejia and Wahba are highly sensitive to their sample inclusion criteria and thus cannot be generalized.

On the other hand, for mandatory interventions the problem is to find eligible individuals in nonparticipating locations similar to eligible individuals in participant locations, in this case selection bias arises due mainly to differences in the economic environment (different labor markets). Friedlander and Robins (1995) and Bloom, Michalopoulos, Hill and Lei (2002) exploit the availability of Welfare-to-Work randomized evaluations to assess the performance of several non- experimental econometric evaluation methods, such as cross-section, before-after and difference-in-differences regressions and matching. The evidence

available, up to date, on the performance of non-experimental methods applied to this type of interventions bring negative conclusions.

In this paper we present evidence on the performance of propensity score matching as a non-experimental impact estimator using a unique data set from Mexico’s anti-poverty program, the Education, Health and Nutrition Program (PROGRESA). This is a conditional cash transfer program –based on children’s school attendance, health check-ups and mandatory health talks for at least one adult member of the household–targeted to rural households. The program has national coverage and is mandatory, all households in participant localities that satisfy program’s eligibility rules and comply with its requirements receive treatment. The incorporation process of localities and households began in August 1997 and by 2000 the program incorporated 72,345 localities in all 31 states around the country, including approximately 2.6 million households. To evaluate the impacts of the program a randomized experiment was carried out during the second phase of incorporation. These 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 one wave from a national survey on income and expenditure carried out biannually by the Mexican National Statistical Institute to identify poor households in localities similar to those in the experiment in order to construct a comparison group that resembles treatment units in the absence of program intervention. Specifically we combine randomized-out (control) units from the experiment with non-experimental households from the national survey to estimate the bias that arise when estimating program impacts using matching methods.

Our identification strategy is based on the assumption of “selection on observables”. We claim this is a valid assumption in the present context because the incorporation of poor households into PROGRESA 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 in our exercise is to find eligible households from non-experimental localities similar enough to those in the experiment. In these terms our problem is more closely related to that in Friedlander and Robins (1995) and Bloom et al. (2002). First, we select non-experimental localities similar enough to their experimental counterparts using administrative data from PROGRESA in order to eliminate or mitigate the selection bias that arises because of differences across localities. 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. This balancing score is used to match experimental to non-experimental units.

The results of this paper contribute to the existing literature in several ways. First, given that social experiments are not abundant, our paper contributes to the literature by exploiting experimental data from PROGRESA 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 inside the U.S, our paper extends the evidence outside the U.S. and beyond employment programs. 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 –social experiments. 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 that are collected through both the same and different questionnaires, thus providing evidence on the importance of questionnaire versus other sources of bias.

Our main results are that matching performs well when outcomes are measured in a comparable way (child schooling and labor outcomes), however, performs very poorly in replicating the experimental benchmark for outcomes that are collected using different survey instruments (household expenditure outcomes). This is of particular interest because in our application there is no problem of common support (so we compare “comparable units”) and matching re–weights the samples such that observational units are compared using the appropriate proportions, thus any remaining bias can be narrowed down to differences in the way outcomes are recorded. Further, we find no systematic differences in our results across matching techniques, although there are minor differences for child employment and schooling.

The rest of the paper proceeds as follows. Section 2 discusses the evaluation problem, the parameter of interest and presents a brief review of the literature on program evaluation, specifically on the performance of matching estimators and the types of bias that arise in observational studies. Section 3 describes PROGRESA, the targeting mechanism of the program and the social experiment. Section 4 discusses the non-experimental identification strategy pursued in our application. Section 5 reports the results of our assessment on matching methods and Section 6 concludes.

## 2 The Evaluation Problem 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. However, 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 drop-out bias.<sup>1</sup> 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 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.

## 2.1 The Evaluation Problem

In many applications as ours, the parameter of interest is the effect of **treatment on the treated** ( $TT$ ). The treatment on the treated parameter answers the question “how does the treatment of a program or policy intervention change the outcomes of participants relative to what they would have experienced had they not received the treatment?”. Lets denote outcomes by  $Y$  and program participation by  $D$ , and let  $D = 1$  for those who receive the treatment and  $D = 0$  for those who do not. Then, 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 ( $D = 1$ ).

Note that potential outcomes are not directly observed, what the researcher observes instead is the realization of the outcome ( $Y$ ) which depends on the particular state. This can be expressed as  $Y = DY_1 + (1 - D)Y_0$ , so we observe  $Y = Y_1$  only when  $D = 1$  and  $Y = Y_0$  only when  $D = 0$ . Thus, one cannot observe the outcome for any given observational unit in both the treated and the untreated states, this is **the evaluation problem** (Heckman and Robb 1985, Holland 1986). We can concentrate on the average effect of the program, but even so, we face the problem of constructing a suitable counterfactual outcome in the untreated state conditional on receiving treatment. To highlight this problem in a nonparametric setting, the treatment on the treated parameter can be expressed as:

$$\begin{aligned} TT &= E(Y_1 - Y_0|D = 1) \\ &= E(Y_1|D = 1) - E(Y_0|D = 1). \end{aligned}$$

The last term in this expression is the counterfactual of interest: what the outcome for treated units

---

<sup>1</sup>See Heckman and Smith (1995) and Heckman, Lalonde and Smith (1999) for a discussion.

would have been had they not received the treatment. The problem is that 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 in general we should expect that  $E(Y_0|D = 1) \neq E(Y_0|D = 0)$  because of selection bias.<sup>2</sup> Therefore, the central problem is to obtain a good estimate for the unobserved or “missing” component.

There are two methods to solve this problem. The first identification strategy is the experimental design (called social experiments when applied to the social sciences), which solves the evaluation problem by randomly denying treatment to analysis units (individuals, households, localities, etc.) which otherwise would receive it (Burtless 1995, Heckman, LaLonde and Smith 1999). 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 treatment and control units.<sup>3</sup> This is the evaluation method implemented by PROGRESA.

The second identification strategy is the non-experimental evaluation design. This strategy applies statistical and econometric techniques to identify non-treated households similar on pre-treatment characteristics to those in the pool of treatment 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 the 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).<sup>4</sup> The matching method assumes selection on observables.

---

<sup>2</sup>To see this add and subtract  $E(Y_0|D = 1)$  to the difference of means  $E(Y_1|D = 1) - E(Y_0|D = 0)$  and obtain:

$$\begin{aligned} E(Y_1|D = 1) - E(Y_0|D = 0) &= E(Y_1|D = 1) - E(Y_0|D = 0) + E(Y_0|D = 1) - E(Y_0|D = 1) \\ &= E(Y_1 - Y_0|D = 1) + \{E(Y_0|D = 1) - E(Y_0|D = 0)\} \\ &= TT + B. \end{aligned}$$

The term in curly brackets in the middle equation is the selection bias, which arises because of systematic differences between treatment and comparison units. Thus,  $E(Y_1|D = 1) - E(Y_0|D = 0)$  identifies  $TT$  when  $B = 0$ , that is when those differences can be eliminated so there is no selection bias.

<sup>3</sup>This strategy has become the benchmark for employment and training programs evaluations in North America because of its intuitive simplicity and expositional tractability. Additionally social experiments compare units in the same local areas and administer the same survey questionnaires thereby making the information fully comparable (Burtless 1995, Heckman and Smith 1995). However, social experiments may face some problems as described by Heckman and Smith (1995), Heckman, Lalonde and Smith (1999), and Heckman, Hohmann, Smith and Khoo (2000).

<sup>4</sup>See Heckman and Robb (1985) for more details.

## 2.2 Selected Literature

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 were employment and training programs, either voluntary programs such as the National Supported Work Demonstration, the AFDC Housemaker-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 next a brief review of findings from those studies.<sup>5</sup>

Assessments based on the NSW Demonstration and the JTPA experiment 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.<sup>6</sup> 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 nonparticipants drawn from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). As a response to the negative conclusions stressed by 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 propensity score matching methods perform well in constructing a comparison group that resembles the NSW participants

---

<sup>5</sup>A more complete description of each program, except for Project STAR, is provided by Bloom et al. (2002).

<sup>6</sup>LaLonde reports estimates for the normal selection model in his paper, but he didn't account for the choice-based sampling of the data and the specification used in the assignment equation was inappropriate because it included a dummy variable for residence in an urban area which is a one-way predictor of participation –in this case– and as such should not have been estimable. Therefore the results on selection models he presents do not provide any evidence on the performance of this estimator. This point was raised by Smith and Todd (2003).

in the counterfactual state. They also use data from the NSW, CPS and PSID 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. The results of these studies have contributed to the increasing popularity of matching although their standard errors are too big and, as we will discuss next, they obtain positive conclusions about the performance of matching on the propensity score even when the comparison units came from different labor markets, the survey instruments are different and the set of conditioning variables is not particularly rich.

In a series of studies analyzing the U.S. National Job Training Partnership Act (JTPA) Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998) also 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 and extend the method to a longitudinal setup. Using experimental data from the JTPA experiment and three groups of non-experimental units,<sup>7</sup> they find that propensity score matching performs well only under certain conditions: when working with a rich set of conditioning variables, using the same survey instruments and placing participant and nonparticipant units in the same local labor market.

These studies further point out that the selection bias problem can be decomposed in three components: (1) biases arising by comparing the wrong units (comparing units outside the common support region); (2) biases arising by comparing the right units in the wrong proportion (differences in the densities of observable characteristics between treatment and comparison units); and (3) bias arising due to unobservables, or self-selection bias rigorously defined. Since matching controls for differences in observable characteristics inside the common support region, thus making treatment recipients and comparison units as close as possible on pre-treatment characteristics, this method is suitable to eliminate biases (1) and (2). But it may happen that selection bias occurs on unobserved characteristics, a problem for which matching is not suitable. However, 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

---

<sup>7</sup>Eligible non-participants (ENP) which were interviewed specially for the JTPA using the same survey instruments; a sample of eligible individuals drawn from the Survey of Income and Program Participation (SIPP); and no-show units from the JTPA experimental treatment group sample.

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.

More recently, Smith and Todd (2003) reconcile this contradictory evidence on the performance of propensity score matching using the NSW, CPS and PSID samples in the LaLonde study to assess the performance of matching on the propensity score. Smith and Todd find that results in the Dehejia and Wahba studies 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. Further, Smith and Todd 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 nonparticipants units are not taken from the same local labor markets. Even more, 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 the cross-sectional version 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).

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 nonparticipant 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. They further stress the requirement that non-experimental studies demonstrate the similarity of local conditions between areas to establishing the validity of such comparisons.

Bloom et al. (2002) assess several non-experimental methods in estimating the treatment effect of

welfare-to-work interventions on earnings in six states random assignment experiments: Atlanta, California, Michigan, Ohio, Oklahoma and Oregon. Estimators considered in the study are cross-sectional regression, propensity score matching, 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.

Evidence on the performance of propensity score matching for non employment/training interventions is provided by Wilde and Hollister (2002). 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 kindergartner, they construct comparison groups using out-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.

We present further evidence on the performance of matching applied to a different type of policy intervention, a nationwide mandatory anti-poverty program that aims to reduce the intergenerational transmission of poverty among Mexican rural households. To evaluate the program a social experiment was carried out, we exploit the availability of these experimental data and combine them with non-experimental data from an outside source to estimate the bias that arises in the estimation of program impacts when matching estimators are used.

### **3 The Education, Health and Nutrition Program (PROGRESA)**

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).<sup>8</sup> PROGRESA differed from previous national poverty programs in two key respects. First, it provided benefits conditional on beneficiaries fulfilling certain human capital enhancing requirements: (i) sending children to school, (ii) attendance by an adult at a monthly health seminar and (iii) 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.

---

<sup>8</sup>In 2000 the program expanded to cover urban localities in extreme poverty and change its name to OPORTUNIDADES.

Child labor is a common subsistence strategy for poor households in rural Mexico, delaying school entry, reducing attendance, and leading to eventual early school dropout. By setting the level of transfer according to the opportunity cost of children's time (e.g. increasing the benefit for older children and for girls), PROGRESA explicitly attempted to stimulate human capital investment and break the inter-generational transmission of poverty in rural Mexico.

The precise structure of entitlement for eligible families depends on demographic characteristics. Each household receives a fixed monthly payment of 115 pesos (approximately 12 USD in 1999) plus a monthly payment for every child enrolled at school who maintains an 85% attendance record. This payment varies with the school level and gender of the child. Benefits increase with school grade at both primary school and high school, with a large jump between primary and secondary levels; also at the secondary level girls receive a higher benefit than boys because of their higher propensity to drop out. Additionally the household receives a fixed yearly payment for school supplies of 135 pesos for each child in primary and 170 pesos for each child in secondary school. There is a total cap of 695 pesos per month per family. In 1999 the average monthly transfer represented approximately one-third of a rural household's monthly income. While there are no explicit restrictions on how the cash transfer can be spent, the mandatory health talks stress the importance of food and nutrition for family health, and the cash transfer is specifically targeted to the mother.

PROGRESA has expanded in phases, incorporating localities pre-selected using a marginality index. Phase one began in August 1997, when 3,369 localities covering 140,544 households were incorporated while transfers to those households began in September-October 1997. Phase two began in November 1997, incorporating 2,988 additional localities and 160,161 households receiving their first transfer in January 1998. By the end of phase 11 in 2000 PROGRESA incorporated 72,345 localities in all 31 states of the country, including approximately 2.6 million households.

### **3.1 PROGRESA's Targeting Mechanism**

The targeting of poor households is implemented in a centralized way at the PROGRESA headquarters at Mexico City in three stages. First, all localities in the country are ranked using a marginality index constructed from Census data. This index is stratified into categories and localities with high and very high levels of marginality are pre-selected to be part of the program. The assignment of localities also combines information on accessibility to public schools and public health centers because of the conditionality aspect of program entitlement. Given that households must comply with PROGRESA's requisites it is necessary that selected localities have access to these publicly provided services. This implies that some localities that qualified to be included in the program based on the marginality index could not be served because of lack of access to a public school or a health care center. In the second stage, households within selected localities identified as being in poverty are selected to participate in the program. The third stage requires

the community assembly to consider any possible error in the process of selection of households within the locality and submit the case for revision.<sup>9</sup>

The identification of localities with a high level of marginality was performed using information from the 1990 National Census. For each locality a marginality index was computed using the method of principal components based on seven indicators: (i) percentage of dwellings without running water, (ii) percentage of dwellings without sewer, (iii) percentage of dwellings without electricity, (iv) percentage of dwellings with soil floor, (v) percentage of illiterate population aged 15 or more in the locality, (vi) the average number of household members per room and (vii) percentage of employed population working in agriculture. Once the marginality index is computed, localities were stratified in five marginality categories: very high, high, moderate, low and very low. 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. In addition to the marginality index the selection of beneficiary localities depends on accessibility to public schools and health care centers because of the conditionality aspect of program entitlement. Localities with high and very high marginality, with at least 150 residents, and with access to publicly provided school and health services were selected to receive the program.

The second stage identifies poor households within targeted localities. Once a locality is selected into the program PROGRESA proceeds to select beneficiary households, i.e., to identify households in the locality that can be considered as being in poverty. At this stage a special community census was 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. The selection of households into the program was based on household characteristics other than observed per-capita income, that is, the selection criteria was based on “structural” rather than temporary conditions such as income. A poverty indicator was constructed by comparing the per-capita household income (excluding earnings from child work) to a poverty line, with households below the line considered poor. Then, predicted poverty status was computed using the results from a discriminant analysis of the poverty indicator that selected the household characteristics that best discriminate between poor and non-poor households. In general, the best predicting variables are a dependency index (the ratio of the number of children over the 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.

Finally, the third stage is performed at a decentralized level: the list of potential beneficiaries of the program is presented to the community assembly where the composition of the list is reviewed; at this level

---

<sup>9</sup>See Skoufias, Davis and de la Vega (2001) for an assessment of this procedure.

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

### 3.2 PROGRESA's Experimental Evaluation

During the second phase of incorporation (November-December 1997) a social experiment was launched to evaluate the impacts of the program on several outcomes such as health and schooling for children and household consumption and nutrition outcomes. A total of 506 rural localities from 6 states –including 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, the other 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.<sup>10</sup>

The impact evaluation of the program was conducted independently by the International Food Policy Research Institute (IFPRI). We present some of these results based on the overview provided by Skoufias (2000).<sup>11</sup> For schooling outcomes Schultz (2000) finds that enrollment increased both at the primary and secondary level and that girls experienced the greater increases at both levels. School enrollment rates for PROGRESA children at primary schools increased by 0.74% to 1.07% points for boys and by 0.96% to 1.45% points for girls relative to non-PROGRESA children. For children at secondary schools, enrollment rates increased by 3.5% to 5.8% points for boys and from 7.2% to 9.3% points for girls.

In terms of health outcomes Gertler (2000) finds that PROGRESA children ages 0-5 have a 12% lower incidence of illness than non-PROGRESA children do. For young adults PROGRESA beneficiaries experienced 19% fewer days of illness than non-beneficiaries, being able to walk about 7.5% more without getting tired. Similar results were found for adults over 50. In addition Behrman and Hoddinott (2000) also find that PROGRESA children ages 0-5 experienced higher growth than non-PROGRESA children.

Household consumption and food expenditures also increased as a results of the intervention. Hoddinott, Skoufias and Washburn (2000) show that the average level of consumption (including purchases and consumption out of own production) increased by approximately 10.5% for beneficiary households relative to control households. Even more, food expenditures were also significantly higher among beneficiary households, so that by 1999 median food expenditures were 13% higher in the treatment group.

---

<sup>10</sup>Behrman and Todd (1999) explore the testable implications of randomization applied to the PROGRESA experiment. They couldn't reject the hypothesis that observable characteristics have the same distribution between treatment and control localities. However they find many rejections for household level characteristics. This can be explained in part by the larger sample size at the household level, thus even minor differences may lead to rejections.

<sup>11</sup>All the evaluation studies are available on IFPRI's web-site: <http://www.ifpri.org/themes/progresas.htm>. We refer the reader to those studies.

While the overall evaluation, using either qualitative or quantitative techniques, explored a variety of other outcomes such as the parents’ attitudes towards the education of girls, the use of time by household members including children, or women’s empowerment, two of the most important outcomes were those related to school enrollment and spending behavior, which are the outcomes we consider in this paper.

## 4 Methodology and Data

Our assessment consists on estimating the bias that may arise when non-experimental methods are applied to estimate the impact of the program by comparing experimental control and treatment units to non-experimental comparison units that resemble what a treated household would have experienced had the household not received treatment. In this section we describe how a non-experimental design might solve the evaluation problem. We will present estimates of the potential evaluation bias that might arise by using propensity score matching methods to evaluate the impact of PROGRESA on a set of outcomes of interest to the program: composition of food expenditures, expenditure on children’s clothes and children’s school enrollment.

### 4.1 How Does Matching Work?

Matching is a non-parametric estimation method that works by re-weighting the comparison sample so it provides an estimate of the counterfactual of interest, in this study, what the outcome of a beneficiary household would have been had it not received program benefits. After the re-weighting scheme treatment and comparison units look the same in terms of observables so any difference in outcomes between these two groups could be attached to the treatment effect if we can guarantee that there are no further systematic differences between the two groups besides those observables. Thus, in order to identify the treatment on the treated parameter using matching, the identification assumption is that outcomes in the untreated state ( $Y_0$ ) are independent of program participation ( $D$ ) conditional on living in marginal localities and on a particular set of observable characteristics that identifies households as being in poverty. This is the conditional independence assumption (Lechner 1999), the ignorable treatment assignment assumption (Rosenbaum and Rubin 1983), or the selection on observables assumption (Heckman and Robb 1985).

We need to make the assumption that treatment (control) and comparison units differ systematically only in terms of these observables, so once we condition on them matching will balance unobservables away. Denoting by  $X$  the relevant set of observable characteristics, the identification assumption can be expressed as  $(Y_0 \perp D)|X$  where the symbol  $\perp$  denotes independence. Actually we require a weaker condition to identify the treatment parameter, that of conditional mean independence:  $E(Y_0|D = 1, X) = E(Y_0|D = 0, X)$ .<sup>12</sup> By conditioning on  $X$  we can get an estimate of the unobserved component in the  $TT$  parameter by matching treatment and (non-experimental) comparison units; in particular, we can identify

---

<sup>12</sup>This weaker condition can be derived from the conditional independence assumption.

the parameter as follows:

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

However, a practical problem arises in the application of matching estimators. As the number of conditioning variables grows or continuous variables are present in the set  $X$ , the dimensionality of the matching procedure becomes cumbersome; this is the **curse of dimensionality**. We follow Rosenbaum and Rubin (1983) to reduce the curse of dimensionality in the application of our matching procedure. They have proved that if the conditional independence assumption holds by conditioning on  $X$ , then it also holds by conditioning on the propensity score, the conditional probability of participation:  $P(X) = Pr(D = 1|X)$ ; that is  $(Y_0 \perp D)|P(X)$ . The advantage is that the propensity score is of dimension one and can be estimated using parametric or semi-parametric methods, this reduces the dimension of the problem and matching becomes easy to implement.<sup>13</sup> Using the propensity score 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 treatment on the treated parameter as follows:

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

where the conditional mean independence assumption is already imposed in the right hand side of the expression.

## 4.2 Measures of Selection Bias

In our application, we compute two measures of the selection bias associated with the matching procedure. We compute a direct measure of the bias that compares control (non-treated) units from the experimental data –the randomized-out units– with non–experimental comparison units from the national survey. The estimated evaluation bias can be expressed as:

$$\hat{B}^{dir}(X) = E(Y_0|D = 1) - \hat{E}\{E(Y_0|D = 0, P(X))|D = 1\}.$$

The first term in the right side of the previous expression comes from the control sample, this is the experimental estimate of the counterfactual; the last term is the non–experimental estimator of the counterfactual, it is constructed from a comparison group sample by matching them to control units on the balancing score. Since randomized-out 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.

---

<sup>13</sup>Note that the Rosenbaum and Rubin (1983) result does not eliminate the curse of dimensionality, it just moves it to the level of the estimation of the propensity score. The problem “reappears” when the propensity score is estimated using fully nonparametric methods. See the discussion on this point in Smith and Todd (2003).

Several studies have addressed the performance of non-experimental methods by estimating an indirect measure of the bias associated with that particular method. The indirect way of measuring the bias consists on estimating the  $TT$  parameter using the treatment and comparison samples and then comparing this non-experimental estimate to the experimental figure, c.f. LaLonde (1986), Dehejia and Wahba (1999, 2002) and Friedlander and Robins (1995). Let  $TT_e$  and  $TT_n$  denote the experimental and non-experimental estimators respectively, then the indirect measure of the bias can be expressed as:

$$\begin{aligned}\hat{B}^{ind}(X) &= [E(Y_1 - |D = 1) - \hat{E}\{E(Y_0|D = 0, P(X))|D = 1\}] - [E(Y_1|D = 1) - E(Y_0|D = 1)] \\ &= TT_n - TT_e,\end{aligned}$$

which is the second measure of the bias we estimate in our assessment. Again, any deviation from zero can be interpreted as evaluation bias.

### 4.3 The balancing score and the common support region

**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 data.

In practice we generate a dummy variable  $D$  that takes a value of one when the observation comes from the experimental sample (either from the treatment or control group) and zero when it comes from the non-experimental sample. All the experimental units are poor households but not every non-experimental unit is poor, even though they come from localities targeted to participate later in PROGRESA. Thus we are estimating the probability of being eligible conditional on a set  $X$  of observable characteristics. We estimate a logit model using all of the observations available (experimental and non-experimental) in order to gain efficiency. Then we use the estimated coefficients from the logit to obtain the predicted probability ( $p$ ) and then compute the log odds-ratio,  $\log\left(\frac{p}{1-p}\right)$ , for each observation in the treatment, control and comparison samples.

**Balancing test.** In the estimation of the propensity score we perform a balancing test as described in Dehejia and Wahba (1999, 2002) to guide the specification of the logit model. Applied to our data, this test consists of:

1. Estimate the score using a parsimonious specification of the logit model and obtain the predicted scores for control and comparison units.
2. Stratify the sample of controls and comparison units according to the score beginning with an arbitrary number of strata. 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 partition the sample further and test again, repeating the process until the scores are balanced inside each strata.

3. Once all the strata are balanced, perform individual mean t-test between controls and comparisons for each of the variables used to predict the score.
4. If all the tests are accepted, then stop. If not, go back to the first step and include higher order or interaction terms for those variables with rejecting test.

**Common support.** In order to get a matching estimator we require to impose the common support condition. 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 (the bias) when there is no overlap between the treatment (control) and comparison groups. To define the region of common support we use a simple procedure which consists in 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.

#### 4.4 Matching estimators

We examine the performance of several different matching methods. Applied to estimate the direct measure of 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  $Y_{0i}$  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:

1. 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.
2. Caliper matching. This estimator chooses the nearest neighbor inside a caliper of width  $\delta$ , 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 to imposing the common support condition.
3. 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.

4. Local-linear matching. This estimator is similar to the kernel estimator but further 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, by doing this we can take into account 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, then we drop the treatment units and predict the score for control and comparison units; and finally, for each estimator we consider 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.

## 4.5 Data and Samples

We combine the experimental data used to evaluate PROGRESA with non-experimental data, drawn from a national household survey carried out for different purposes, and compare the outcomes for control households in PROGRESA to the outcomes for comparison matched pairs households from the non-experimental survey to estimate the bias that arises when non-experimental methods are applied to evaluate the program.

The experimental evaluation data we use come from PROGRESA's Household Evaluation Survey (*Encuesta de Evaluación de los Hogares* –ENCEL) and consist of four rounds of household surveys covering 506 localities and approximately 25,000 households from which about a half comprises the eligible sample. One third of the localities were 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 National Survey of Households' Income and Expenditures (*Encuesta Nacional sobre Ingresos y Gastos de los Hogares* –ENIGH). This is a biannual nation-wide 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 4,500 are rural households; we use the 1998 round of ENIGH to construct the non-experimental comparison group.

The 1998 wave of ENIGH was collected in the fall between September and early November, approximately 9 to 10 months after the start of PROGRESA in the evaluation sample, implying that some ENIGH households may actually have been participating in the program. Using PROGRESA retrospective admin-

istrative 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 in our matching estimates we exclude all localities from the ENIGH rural sample that had already entered PROGRESA at the time of the survey.<sup>14</sup> 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 will be to see whether the matching technique is able to identify enough good matches from ENIGH to allow for meaningful comparisons with the randomized-out group from ENCEL.

#### **Differences in survey instruments**

Aside from differences in the sample frame and survey purposes, a critical issue is the difference in survey questionnaires between ENIGH and ENCEL in terms of reference periods and report detail. 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 while 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, an important issue stressed out by Heckman, Ichimura, Smith and Todd (1998) and Smith and Todd (2003).

## **4.6 Making the case for “Selection on Observables”**

Matching will perform well if the assumption of selection on observables is valid. We claim this is a valid assumption in the present context because the incorporation of poor households into PROGRESA 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

---

<sup>14</sup>Contamination bias arises when the sample of comparisons used to compute the counterfactual of interest contains treatment units. This bias occurs when there is no way to identify whether units from the comparison sample have received or are receiving treatment.

in Friedlander and Robins (1995) and Bloom et al. (2002). 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 make them qualify to receive program benefits but which were not enrolled into the program during the first two phases of incorporation. At this stage, we use PROGRESA's administrative records to identify localities designated to be covered in future phases of implementation. Using these data we identify the phase in which each rural locality from the non-experimental survey was determined to enter the program according to the marginality index and the availability of public services near the community. We then select all households in those non-experimental localities that were assigned to be treated in future phases of implementation as our working sample. This procedure aims to guarantee that we are comparing "comparable" localities in terms of poverty in order to eliminate –or at least mitigate– the selection bias that arises because of differences across localities.

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 propensity score matching (Rosenbaum and Rubin 1983, Heckman, Ichimura and Todd 1998) 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.

## 5 Results

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 these data 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, this sample consist on 2,479 households; and Sample 2 refers to a further restricted sample which excludes all households in localities where PROGRESA was never implemented, there are 736 households in this sample.

Table 1 presents summary statistics on several variables including a dependency index (the ratio of the number of children over the number of working-age adults), the sex, age and schooling of the household head, the number of children in the household, an overcrowding index (number of persons per bedroom),

several dummies for dwelling characteristics such as soil floor, bathroom with running water, access to electricity; and dummy variables indicating possession of durable goods such as a gas stove, a refrigerator, a washing machine and a vehicle. 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. 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 13–16 years old and lower rates of child employment.

Results of the logit models to determine the probability of qualifying for the program are reported in Table 3. 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 one of the non-experimental samples. The independent variables include those used by PROGRESA in its score to determine program eligibility at the household level. 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. Columns 1 and 3 report estimated coefficients and marginal effects respectively when we combine ENCEL units with ENIGH-Sample 1 units, columns 5 and 7 report the same figures when we use ENIGH-Sample 2 instead.

There are a few differences worth noting between the estimates over the two combined 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 later case, while the bathroom indicators become smaller.

For each combined sample we perform balancing tests as 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 balancing score, the log 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 distribution of estimated odds-ratios are shown graphically in Figures 1 and 2. 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 0.851, still significantly lower than the mean for the ENCEL households, which is around 4.4. As in the previous combined sample, imposing the common support criteria results in the elimination of 1.62% of the control and 1.63% of the comparison groups.

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.

## 5.1 Bias Estimates for Matching Estimators

In this section we present estimates of the evaluation bias that arises from using non–experimental propensity score matching methods instead of randomization. We compare control units from ENCEL to matched comparison units drawn from ENIGH samples 1 and 2. We present evidence on the performance of the nearest–neighbor (with replacement), caliper, local–linear and kernel estimators.

### 5.1.1 Estimates of bias for household level outcomes

Table 5 presents estimates of the bias on 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 propensity score matching. 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 v.s. 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. Even more, the results for child labor outcomes are similar to those for schooling outcomes, this information is collected in a very 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 matching.

Comparing the different 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 kernel matching tend to produce larger point estimates of bias in 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. We find similar results when estimating the bias using the “indirect” measure, these results are reported in the appendix (available upon request).

### 5.1.2 Estimates of bias for child level 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 8-16 years old only for caliper matching ( $\delta = 0.01$ ), although this is primarily driven by the significant difference in outcomes for children 8-12 years old. The means for these outcomes (bottom of Table 2) reveal that enrollment rates for children ages 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 12-16 years old based on local linear and kernel matching. These results are somewhat surprising given the means for these outcomes in Table 2, especially for all kids 12-16 years old 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 8-16 years old 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 matching in measuring program impact as suggested by Heckman et al. (1998). The other noteworthy result is the difference across matching techniques. While nearest neighbor never identifies significant differences in the individual level outcomes, differences are detected using caliper and local linear matching.

## 5.2 Comparing Matching to Regression Estimates

While the main objective of this paper is to compare matching to the experimental estimates, it is of some interest to compare the matching 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 matching 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 matching 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 matching. 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 matching. Here the differences are quite stark; while none of the matching 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 single labor 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.

## 6 Conclusion

The validity of non-experimental evaluation estimators is an important issue in the evaluation literature given the potential difficulties in launching social experiments. All the published work that shed light on this question have used employment and training programs from the U.S. In particular the results from studies assessing the performance of matching seem to converge to the view that it 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.

In this paper 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. exploiting the availability of experimental data from Mexico’s anti-poverty program, the Education, Health and Nutrition Program (PROGRESA). 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.

There are two important implications that can be drawn from the results presented in the paper. 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 matching may be a valid impact estimator. These main conclusions are that matching requires appropriate and detailed covariates (in our case we have access to the same 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 non-experimental impact evaluations using matching which can provide reasonable estimates of program impacts. 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

- [1] Behrman, Jere, and John Hoddinott, “An Evaluation of the Impact of PROGRESA on Pre-School Child Height.” International Food Policy Research Institute, Washington, D.C. 2000.
- [2] Behrman, Jere and Petra Todd, “Randomness in the Experimental Samples of PROGRESA.” Research Report, International Food Policy Research Institute. Washington D.C. 1999.
- [3] Bloom, Howard, Charles Michalopoulos, Carolyn J. Hill and Ying Lei, “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. 2002.
- [4] Burtless, Gary, “The Case for Randomized Field Trials in Economic and Policy Research.” *Journal of Economic Perspectives*. 1995, 9(2), pp.63–84.
- [5] Dehejia, Rajeev and Sadek Wahba, “Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association*. 1999, 94, pp. 1053–1062.
- [6] Dehejia, Rajeev and Sadek Wahba, “Propensity Score Matching Methods for Non-Experimental Causal Studies.” *Review of Economics and Statistics*. 2002, 84, pp. 151-161.
- [7] Friendlander, Daniel and Phil Robbins, “Evaluating Program Evaluations: New Evidence on Commonly Used Non-experimental Methods.” *American Economic Review*. 1995, 85(4), pp. 923–937.
- [8] Gertler, Paul, “The Impact of PROGRESA on Health.” International Food Policy Research Institute, Washington, D.C. 2000.
- [9] Heckman, James, and Richard Robb, “Alternative Methods for Evaluating the Impact of Interventions.” In James Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data*. Cambridge, England: Cambridge University. 1985, pp. 156–246.

- [10] Heckman, James, and Joseph Hotz, "Choosing Among Alternative Non-experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association*. 1989, 84, pp. 862–880.
- [11] Heckman, James, Hidehiko Ichimura, and Petra Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies*. 1997, 64, pp. 605–654.
- [12] Heckman, James, Hidehiko Ichimura, and Petra Todd, "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies*. 1998, 65, pp. 261–294.
- [13] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd, "Characterizing Selection Bias Using Experimental Data." *Econometrica*. 1998, 66, pp 1017–1098.
- [14] Heckman, James, Neil Hohmann, Jeffrey Smith and Michael Khoo, "Substitution and Drop Out Bias in Social Experiments: A Study of an Influential Social Experiment." *Quarterly Journal of Economics*. 2000, 115, pp. 651–694.
- [15] Heckman, James, Robert LaLonde and Jeffrey Smith, "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Volume 3A. Amsterdam: North-Holland. 1999, pp. 1865–2097.
- [16] Heckman, James and Jeffrey Smith, "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*. 1995, 9(2), pp. 85–110.
- [17] Hoddinott, John, Emmanuel Skoufias, and Ryan Washburn, "The Impact of PROGRESA on Consumption: A Final Report." International Food Policy Research Institute, Washington, D.C.2000.
- [18] Holland, Paul, "Statistics and Causal Inference." *Journal of the American Statistical Association*. 1986, 81, pp. 945–970.
- [19] LaLonde, Robert, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*. 1986, 76(4), pp. 604–620.
- [20] PROGRESA, "Programa de Educación Salud y Alimentación." Mexico. 1997.
- [21] Rosenbaum, Paul and Donald Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*. 1983, 70, pp. 41–50.
- [22] Skoufias, Emmanuel, Benjamin Davis and Jere Behrman, "An Evaluation of the Selection of Beneficiary Households in the Education, Health and Nutrition Program (PROGRESA) of Mexico." Research Report, International Food Policy Research Institute. Washington D.C. 1999.

- [23] Skoufias, Emmanuel, Benjamin Davis and Sergio de la Vega, “Targeting the Poor in Mexico: An evaluation of the Selection of Households into PROGRESA.” *World Development*. 2001, 29, pp. 19769–1784
- [24] Schultz, T. P., “The Impact of PROGRESA on School Enrollments.” International Food Policy Research Institute, Washington, D.C. 2000.
- [25] Smith, Jeffrey and Petra Todd, “Does Matching Overcome Lalonde’s Critique of Non–experimental Estimators?.” Forthcoming, *Journal of Econometrics*. 2003.
- [26] Wilde, Elizabeth Ty and Robinson Hollister, “How Close Is Close Enough? Testing Non–experimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes.” Institute for Research on Poverty. Discussion Paper No. 1242-02.

Table 1: Observable Characteristics by Sample

Data set:	ENCEL			ENIGH				
	Sample:	Treatment	Control	All rural	Raw samples		Matched samples	
					Sample 1	Sample 2	Sample 1	Sample 2
Dependency ratio		1.461 (0.95)	1.487 (0.98)	1.140 (1.06)	1.031 (0.95)	1.120 (1.04)	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)	0.145 (0.35)	0.097 (0.30)	0.127 (0.33)
Head's schooling								
Complete Primary		0.446 (0.50)	0.458 (0.50)	0.399 (0.49)	0.402 (0.49)	0.395 (0.49)	0.458 (0.50)	0.503 (0.50)
Incomp. Secondary		0.247 (0.43)	0.237 (0.43)	0.204 (0.40)	0.223 (0.42)	0.193 (0.39)	0.270 (0.44)	0.224 (0.42)
Secondary/more		0.055 (0.23)	0.055 (0.23)	0.108 (0.31)	0.124 (0.33)	0.095 (0.29)	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)	47.120 (16.77)	41.361 (13.98)	41.540 (13.90)
Number kids ages $\leq 13$		2.457 (1.66)	2.483 (1.61)	1.487 (1.54)	1.339 (1.42)	1.439 (1.49)	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)	2.436 (1.69)	4.359 (2.20)	4.178 (2.01)
No social security		0.968 (0.18)	0.960 (0.20)	0.819 (0.39)	0.774 (0.42)	0.865 (0.34)	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)	0.405 (0.49)	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)	0.461 (0.50)	0.408 (0.49)	0.437 (0.50)
Soil floor		0.730 (0.44)	0.755 (0.43)	0.255 (0.44)	0.203 (0.40)	0.214 (0.41)	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)	0.288 (0.45)	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)	0.568 (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)	0.784 (0.41)	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)	0.773 (0.42)	0.944 (0.23)	0.943 (0.23)
Observations		7837	4682	3898	2479	736	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.

Matched samples are constructed using nearest neighbor with replacement and common support. Standard deviation in parenthesis.

Table 2: Summary Statistics for Outcome Variables

Data set:	ENCEL			ENIGH				
	Sample:	Treatment	Control	All rural	Raw samples		Matched samples	
					Sample1	Sample2	Sample1	Sample2
<b>Household outcomes</b>								
Food expenditure		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		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)
% of kids 8-16 in school		0.785 (0.3)	0.743 (0.3)	0.804 (0.3)	0.799 (0.3)	0.776 (0.4)	0.786 (0.3)	0.767 (0.4)
Observations		7837	4682	3898	2479	736	768	363
<b>Child outcomes</b>								
<b>School enrolment</b>								
Children 8-16		0.772 (0.42)	0.727 (0.45)	0.795 (0.40)	0.791 (0.41)	0.764 (0.43)	0.766 (0.42)	0.767 (0.42)
Observations		13589	8130	4407	2598	793	659	309
Children 8-12		0.921 (0.27)	0.891 (0.31)	0.948 (0.22)	0.947 (0.22)	0.948 (0.22)	0.93 (0.26)	0.938 (0.24)
Observations		8200	4877	2563	1493	464	531	243
Children 13-16		0.545 (0.50)	0.48 (0.50)	0.582 (0.49)	0.579 (0.49)	0.505 (0.50)	0.512 (0.50)	0.468 (0.50)
Observations		5389	3253	1844	1105	329	387	171
<b>Work for pay</b>								
All Children 12-16		0.111 (0.31)	0.116 (0.32)	0.107 (0.31)	0.126 (0.33)	0.121 (0.33)	0.118 (0.32)	0.112 (0.32)
Observations		7028	4250	2402	1259	422	458	197
Boys 12-16		0.164 (0.37)	0.18 (0.38)	0.141 (0.35)	0.164 (0.37)	0.132 (0.34)	0.167 (0.37)	0.112 (0.32)
Observations		3628	2146	1210	627	204	228	89
Girls 12-16		0.054 (0.23)	0.051 (0.22)	0.073 (0.26)	0.089 (0.28)	0.11 (0.31)	0.07 (0.26)	0.111 (0.32)
Observations		3376	2100	1192	632	218	230	108

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.

Matched samples are constructed using nearest neighbor with replacement and common support.

Standard deviation in parenthesis.

Table 3: Logit Estimates

	Sample 1		Sample 2	
	Coeff.	std.err.	Coeff.	std.err.
Dependency ratio	0.297	0.088	0.367	0.129
Head's sex	-0.150	0.101	-0.189	0.146
Head's schooling				
Complete Primary	0.466	0.084	0.717	0.127
Incomplete Secondary	0.887	0.110	1.146	0.168
Complete Secondary or more	0.661	0.153	0.869	0.229
Head's age	-0.084	0.041	-0.058	0.062
Head's age squared	0.002	0.001	0.002	0.001
Head's age cube	0.000	0.000	0.000	0.000
Number kids ages $\leq 13$	0.620	0.062	0.570	0.091
Crowding index	0.456	0.053	0.552	0.076
Without social security	1.522	0.107	1.193	0.167
No bathroom	0.517	0.144	0.136	0.205
Bathroom no water	0.629	0.138	0.456	0.198
Soil floor	1.257	0.073	1.487	0.116
Without gas stove	1.425	0.077	1.650	0.119
Without refrigerator	1.332	0.092	1.277	0.131
Without washer	1.076	0.129	0.729	0.178
Without vehicle	0.530	0.122	0.326	0.167
Crowding index squared	-0.011	0.008	-0.017	0.012
Crowding index $\times$ number of kids	-0.070	0.014	-0.063	0.021
Dependency ratio cube	-0.063	0.016	-0.077	0.022
Constant	-5.949	0.681	-4.873	1.033
Number of observations	14745		13031	
Likelihood ratio test	6292		2252	
Prob.	(0.00)		(0.00)	

The dependent takes a value of one if the observation comes from the experimental sample and zero if from the non-experimental sample.

Table 4: Estimated Propensity (Balancing) Score

	Statistics				Obs. inside common support	Obs. in each sample	Percentage excluded
	Mean	Std.Dev.	Min	Max			
A. Matched Sample 1							
Treatment	3.183	1.35	-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	7.62	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: Estimated Bias for household level outcomes – sample 1

Matching method:	Nearest Neighbor	Caliper		Local Linear		Kernel	
		d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2
<b>Expenditures</b>							
Food	<b>-219.114</b> (33.772)	<b>-223.855</b> (29.089)	<b>-238.126</b> (44.568)	<b>-226.493</b> (31.438)	<b>-216.869</b> (31.252)	<b>-215.466</b> (29.607)	<b>-214.611</b> (29.361)
Vegetables	<b>70.698</b> (2.132)	<b>71.22</b> (2.186)	<b>72.578</b> (5.305)	<b>69.177</b> (1.968)	<b>69.382</b> (1.862)	<b>69.484</b> (1.950)	<b>69.601</b> (1.882)
Cereals	21.038 (11.872)	<b>22.267</b> (9.895)	20.517 (16.995)	<b>22.045</b> (9.769)	<b>26.901</b> (10.106)	<b>26.52</b> (8.988)	<b>27.595</b> (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	<b>11.158</b> (0.798)	<b>11.265</b> (0.789)	<b>11.505</b> (1.571)	<b>10.916</b> (0.688)	<b>11.218</b> (0.643)	<b>11.116</b> (0.658)	<b>11.133</b> (0.657)
Expenditure shares							
Food	<b>0.267</b> (0.015)	<b>0.263</b> (0.012)	<b>0.248</b> (0.017)	<b>0.262</b> (0.013)	<b>0.268</b> (0.013)	<b>0.263</b> (0.013)	<b>0.263</b> (0.013)
Vegetables	<b>0.120</b> (0.001)	<b>0.121</b> (0.002)	<b>0.122</b> (0.005)	<b>0.119</b> (0.001)	<b>0.119</b> (0.001)	<b>0.119</b> (0.001)	<b>0.119</b> (0.001)
Cereals	<b>0.195</b> (0.007)	<b>0.195</b> (0.007)	<b>0.191</b> (0.011)	<b>0.194</b> (0.006)	<b>0.197</b> (0.006)	<b>0.195</b> (0.005)	<b>0.196</b> (0.005)
Meat	<b>0.069</b> (0.006)	<b>0.070</b> (0.005)	<b>0.065</b> (0.009)	<b>0.068</b> (0.006)	<b>0.07</b> (0.005)	<b>0.069</b> (0.006)	<b>0.069</b> (0.005)
Kids clothes	<b>0.018</b> (0.001)	<b>0.018</b> (0.001)	<b>0.018</b> (0.002)	<b>0.018</b> (0.001)	<b>0.018</b> (0.001)	<b>0.018</b> (0.001)	<b>0.018</b> (0.001)
<b>Schooling - labor</b>							
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)	-0.03 (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)
Child labor (% 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)
Match summary for non experimental controls							
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 localities 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: Estimated Bias for household level outcomes – sample 2

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
<b>Expenditure</b>							
Food	<b>-169.252</b> (67.164)	<b>-255.987</b> (54.835)	<b>-290.038</b> (83.193)	<b>-263.924</b> (59.692)	<b>-239.388</b> (52.183)	<b>-258.173</b> (59.396)	<b>-264.56</b> (58.733)
Vegetables	<b>71.777</b> (2.703)	<b>72.124</b> (3.672)	<b>76.468</b> (9.756)	<b>68.382</b> (2.401)	<b>68.745</b> (2.377)	<b>69.451</b> (2.368)	<b>68.917</b> (2.356)
Cereals	<b>56.921</b> (16.552)	<b>31.938</b> (13.513)	40.891 (25.565)	27.555 (16.233)	<b>35.115</b> (15.057)	28.486 (16.519)	27.671 (15.736)
Meat	19.13 (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	<b>10.566</b> (1.163)	<b>9.678</b> (1.289)	<b>10.448</b> (3.245)	<b>10.713</b> (1.025)	<b>10.828</b> (0.997)	<b>10.895</b> (1.000)	<b>10.732</b> (1.005)
Expenditure shares							
Food	<b>0.242</b> (0.021)	<b>0.232</b> (0.014)	<b>0.225</b> (0.024)	<b>0.237</b> (0.019)	<b>0.244</b> (0.020)	<b>0.228</b> (0.019)	<b>0.233</b> (0.018)
Vegetables	<b>0.12</b> (0.002)	<b>0.12</b> (0.003)	<b>0.121</b> (0.007)	<b>0.118</b> (0.002)	<b>0.119</b> (0.002)	<b>0.119</b> (0.002)	<b>0.119</b> (0.002)
Cereals	<b>0.209</b> (0.012)	<b>0.196</b> (0.010)	<b>0.199</b> (0.018)	<b>0.196</b> (0.011)	<b>0.202</b> (0.010)	<b>0.195</b> (0.012)	<b>0.196</b> (0.011)
Meat	<b>0.078</b> (0.009)	<b>0.078</b> (0.008)	<b>0.077</b> (0.014)	<b>0.073</b> (0.009)	<b>0.076</b> (0.008)	<b>0.076</b> (0.008)	<b>0.075</b> (0.008)
Kids clothes	<b>0.018</b> (0.001)	<b>0.016</b> (0.001)	<b>0.018</b> (0.003)	<b>0.018</b> (0.001)	<b>0.018</b> (0.001)	<b>0.018</b> (0.001)	<b>0.018</b> (0.001)
<b>Schooling - labor</b>							
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	-0.02 (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)
Child labor (% 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)
Match summary for non experimental controls							
Number of hhlds used	363	356	207				
Average times used	12.36	6.56	1.97				
Maximum use	195	37	12				

Sample 2 excludes ENIGH rural localities that were already in PROGRESA at the time of the survey, and those never scheduled to enter the program.

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 7: Estimated Bias for Child level Outcomes – sample 1

Matching method:	Nearest Neighbor	Caliper		Local Linear		Kernel	
		d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2
<b>Currently enrolled</b>							
All kids 8-16	-0.038 (0.030)	<b>-0.064</b> (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)	<b>-0.030</b> (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)
<b>Never enrolled</b>							
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)
<b>Work for pay</b>							
All kids 12-16	-0.029 (0.029)	0.021 (0.023)	-0.004 (0.039)	<b>-0.062</b> (0.023)	<b>-0.054</b> (0.023)	<b>-0.051</b> (0.021)	-0.041 (0.021)
Boys 12-16	-0.086 (0.048)	-0.009 (0.036)	-0.021 (0.077)	-0.054 (0.039)	-0.050 (0.039)	-0.041 (0.039)	-0.029 (0.038)
Girls 12-16	-0.004 (0.022)	0.001 (0.023)	0.000 (0.048)	-0.033 (0.019)	<b>-0.044</b> (0.021)	-0.019 (0.017)	-0.026 (0.018)
Match summary for nonexperimental controls							
Number of hhlds used	659	652	356				
Average times used	12.12	10.13	3.66				
Maximum use	127	63	15				

Sample 1 excludes ENIGH rural localities 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.

Match summary is for 8-16 year old schooling sample only.

Table 8: Estimated Bias for Child level Outcomes – sample 2

Matching method:	Nearest Neighbor	Caliper		Local Linear		Kernel	
		d=0.01	d=0.001	bw=0.1	bw=0.2	bw=0.1	bw=0.2
<b>Currently enrolled</b>							
All kids 8-16	-0.054 (0.057)	<b>-0.085</b> (0.035)	-0.053 (0.065)	-0.01 (0.036)	-0.016 (0.035)	-0.012 (0.041)	-0.015 (0.034)
Kids 8-12	-0.023 (0.042)	<b>-0.071</b> (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)	0.03 (0.055)	0.05 (0.063)	0.041 (0.056)
<b>Never enrolled</b>							
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)
<b>Work for pay</b>							
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)	0.02 (0.147)	0.04 (0.043)	0.06 (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)
Match summary for non experimental controls							
Number of hhlds used	309	299	117				
Average times used	24.78	9.72	3.56				
Maximum use	445	67	14				

Sample 2 excludes ENIGH rural localities that were already in PROGRESA at the time of the survey, and those never scheduled to enter the program.

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.

Match summary is for 8-16 year old schooling sample only.

Table 9: Regression vs Matching Estimates, Expenditure Outcomes

	Regression	Nearest Neighbor	Kernel
<b>Expenditure</b>			
Food	<b>-270.819</b> (18.342)	<b>-219.114</b> (33.772)	<b>-215.466</b> (29.607)
Vegetables	<b>79.202</b> (2.966)	<b>70.698</b> (2.132)	<b>69.484</b> (1.950)
Cereals	<b>29.901</b> (7.121)	21.038 (11.872)	<b>26.52</b> (8.988)
Meat	-10.7 (5.739)	-3.960 (8.382)	-4.597 (7.669)
Kid clothes	<b>12.086</b> (0.982)	<b>11.158</b> (0.798)	<b>11.116</b> (0.658)
<b>Expenditure shares</b>			
Food	<b>0.233</b> (0.006)	<b>0.267</b> (0.015)	<b>0.263</b> (0.013)
Vegetables	<b>0.124</b> (0.002)	<b>0.120</b> (0.001)	<b>0.119</b> (0.001)
Cereals	<b>0.183</b> (0.005)	<b>0.195</b> (0.007)	<b>0.195</b> (0.005)
Meat	<b>0.069</b> (0.004)	<b>0.069</b> (0.006)	<b>0.069</b> (0.006)
Kids clothes	<b>0.018</b> (0.001)	<b>0.018</b> (0.001)	<b>0.018</b> (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 regression estimates of the bias. Columns (2) and (3) report the bias estimates using nearest neighbor and kernel matching taken from Table 5 column (1) and (6). Standard errors in parentheses below marginal probabilities. Bold indicates significant at 5%.

Table 10: Regression vs Matching Estimates, Child Outcomes

	Regression	Nearest Neighbor	Kernel
<b>Currently enrolled</b>			
All kids 8-16	<b>-0.052</b> (0.012)	-0.038 (0.030)	-0.028 (0.015)
Kids 8-12	<b>-0.034</b> (0.009)	-0.045 (0.024)	-0.024 (0.017)
Kids 13-16	<b>-0.047</b> (0.024)	0.01 (0.041)	0.014 (0.027)
<b>Never enrolled</b>			
All kids 8-16	<b>-0.011</b> (0.005)	-0.029 (0.020)	-0.026 (0.014)
Kids 8-12	<b>-0.022</b> (0.007)	-0.024 (0.019)	-0.042 (0.016)
Kids 13-16	0.005 (0.007)	0.009 (0.024)	-0.002 (0.014)
<b>Work for pay</b>			
All kids 12-16	-0.014 (0.013)	-0.028 (0.029)	<b>-0.051</b> (0.021)
Boys 12-16	0.006 (0.022)	-0.075 (0.048)	-0.043 (0.039)
Girls 12-16	<b>-0.028</b> (0.014)	-0.01 (0.022)	-0.02 (0.017)

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 estimates of the bias. Columns (2) and (3) report the bias estimates using nearest neighbor and kernel matching taken from Table 7 column (1) and (6). Standard errors in parentheses below marginal probabilities. Bold indicates significant at 5%.

Figure 1: Empirical Density for Estimated Log Odds-Ratio – Sample 1

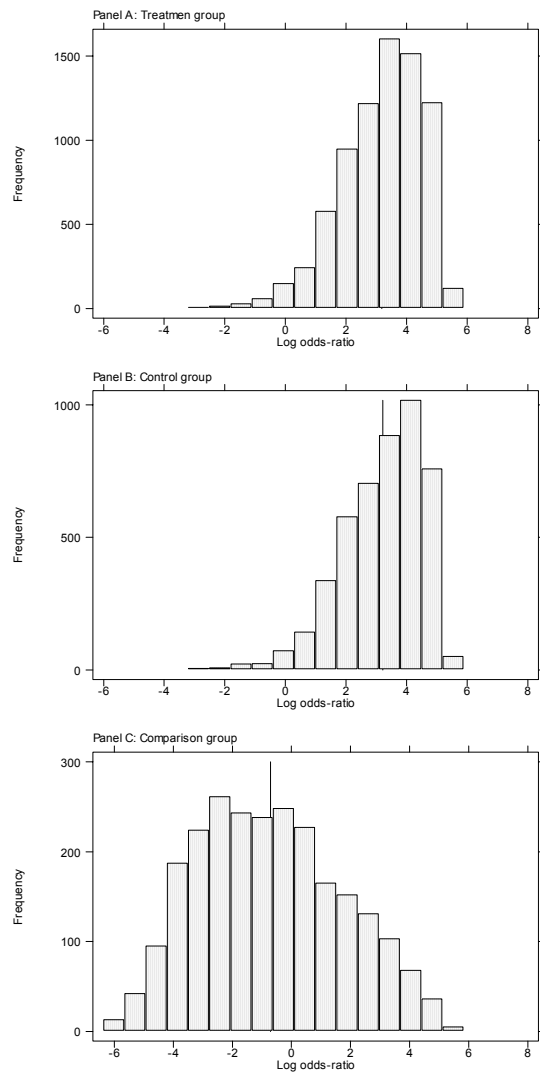


Figure 2: Empirical Density for Estimated Log Odds-Ratio – Sample 2

