

# GENERAL RURAL METHODOLOGY NOTE

Instituto Nacional de Salud Pública

Coordinación Nacional de Programa de Desarrollo Humano Oportunidades, SEDESOL

November 2005

## Contents

1. Introduction .....	3
2. Identification of the localities for the initial intervention by the Program .....	4
2.1 Geographic focalization .....	4
2.2. Socioeconomic data collection in households .....	6
2.3 Presentation of the findings in an assembly .....	6
3. Basis of the evaluation design .....	7
4. Sample design for impact evaluation .....	8
5. Evaluation Surveys .....	12
5.1 Household Socioeconomic Characteristics Survey (ENCASEH) .....	12
5.2 Household Evaluation Surveys (ENCEL) .....	13
6. Database Organization .....	17
6.1. Key Variables .....	20
7. Proposal for analysis strategies .....	22
8. References .....	24

## 1. Introduction

The Oportunidades Human Development program (formerly known as Education, Health and Nutrition Program PROGRESA ) is a Program of the Executive branch of the Federal Government. Its main purpose is to promote and strengthen the capacities and potential of families living in extreme poverty, so they may be able to achieve a higher standard of living and partake in national development. Oportunidades strives toward equity by means of the interaction among three fundamental factors that constitute a poverty-fighting tool: nutrition, health, and education. Oportunidades began to operate on August 8, 1997 in highly deprived rural areas. Since then, its coverage has steadily extended to five million families nationwide.

From the beginning stages of the Program's operations the need to assess its effects on the day-to-day life of participating families was established. Evaluation has always been considered an essential element of the Program, since evaluation not only measures results and impacts, but can also point to corrective or reorientation measures, improve the functionality of processes, and, in general, increase the effectiveness and efficiency of resource usage to achieve the established goals. A strict evaluation of the Program's impacts and results on the beneficiary population provides evidence of accomplishment of objectives and goals, and contributes to the design of social policy activities.

The evaluation plan to assess the results and impacts of the Program includes the analysis of short, medium, and long-term effects on the beneficiary population, using different methodological approaches –including both quantitative and qualitative methods. These methods allow the accurate identification of results and impacts attributable to the Program, sorting out those resulting from other factors, at the individual, family, and community levels. Also, evaluation of the Program has revealed the actual impacts of the Program in its service areas, the synergies among its components, and the consequences on family dynamics and status of women. The impact evaluation of the Program has been entrusted to renowned academic institutions such as the International Food Policy Research Institute (IFPRI), the National Institute of Public Health (INSP, after its initials in Spanish), and the Social Anthropology Research and Higher Studies Center, West campus (CIESAS-Occidente, after its initials in Spanish). These institutions have been in charge of coordinating a large group of national and international researchers from a variety of institutions, all of them experts in social program evaluation. The Advisory Groups for the quantitative and qualitative evaluation are responsible for the technical and methodological guidelines for impact evaluation in their respective fields.

The present Methodological Note describes the general evaluation design in rural areas, the sampling procedures, the data collection instruments, and the structure and management of the databases generated in this evaluation. Also, some strategies for data analysis according to the proposed designs are suggested. The purpose of this information is to provide a general framework for the quantitative analysis of Oportunidades evaluation databases in rural areas.

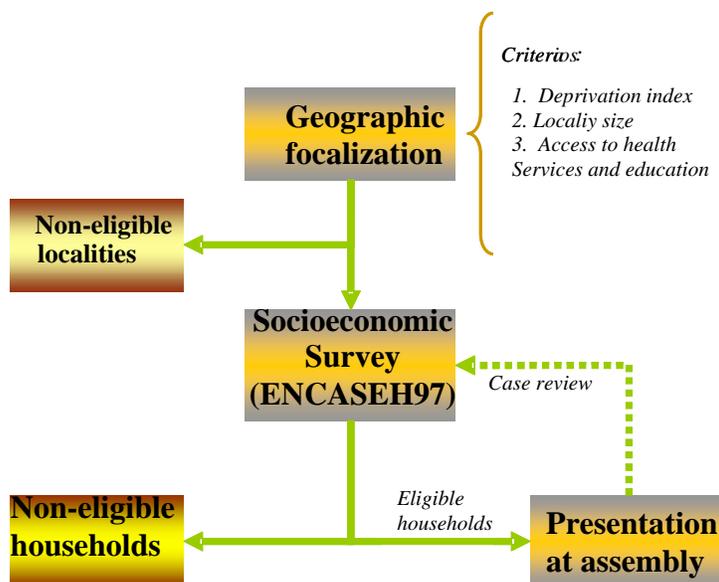
## **2. Identification of the localities for the initial intervention by the Program**

In rural areas the Program focalizes its assistance in three sequential phases: 1) Geographic focalization to find the localities with greatest deprivation and to verify access to basic school and health services; 2) Socioeconomic data collection for each household by taking census in the previously selected communities; and 3) presenting the findings to the community in an assembly to invite suggestions, corrections and additions.<sup>1</sup> Given the importance of these phases with regard to the impact evaluation project, they are described below (Figure 1).

### **2.1 Geographic focalization**

Geographic selection of highly deprived localities is based on the Deprivation Index for each of the localities in the country, for which sociodemographic data from the census were available. These data were obtained from the XI General Population and Household Census 1990, the 1996 Population and Household Count, and afterwards from the XII General Population and Household Census 2000, all conducted by the National Institute for Statistics, Geography and Informatics (INEGI, for its initials in Spanish).<sup>1</sup>

Figure 1. Focalization process of Oportunidades benefits



This index is constructed by social indicators through the principal components method<sup>2</sup> to construct a summary variable that allows comparison of the information among localities (for further details on the construction of this index, see reference 2). This index serves to classify the localities into five deprivation categories: i) very high, ii) high, iii) medium, iv) low, and v) very low. To define these deprivation categories the Tore Dalenius optimum stratification method was applied.<sup>3,4</sup> This method allows grouping of localities based on the density function  $f(x)$  defined by the deprivation index. This statistical procedure can only be performed with localities having complete socioeconomic data.

Localities thus categorized as having a high or very high deprivation level were given priority for inclusion in the Program. Selected localities were also required to have access to elementary school, middle school, and a health clinic. Access to health services and education was defined, when these services were unavailable within localities, according to proximity to roads, considering the differential ground access between localities due to different distances from and to roads. It should be pointed out that almost 70% of all localities had access to health and education services (elementary and middle).<sup>1</sup> Starting in 2000, service capacity of schools and health clinics was verified, in addition to access to education and health services.

In the first stages of the program an additional criterion was used: localities had to have more than 50 but less than 2,500 inhabitants.<sup>1</sup> During the first year of operation the Oportunidades Program was running in eight Mexican states: Campeche, Coahuila, Guanajuato, Hidalgo, Puebla, Querétaro, San Luís Potosí, and

Veracruz. In 1998 the remaining states were also included, except Baja California, Baja California Sur, Aguascalientes, and the Federal District. By the end of 2004 the Program was operating in all Mexican states.

## **2.2. Socioeconomic data collection in households**

The second phase of the focalization process is centered on evaluating the poverty conditions of the households in the highly deprived localities, selected in the first phase. To this end, the Household Socioeconomic Characteristics Survey (ENCASEH) was administered as a census. Through this instrument the relevant socioeconomic and demographic data were obtained to identify households that will benefit from the Program.<sup>1</sup>

To identify the degree of poverty in households a multivariate statistical technique is used – discriminant linear analysis – which makes possible to assess the living conditions of the households and to construct the categories “extreme poverty” and “non-extreme poverty”, that would later be used for the subsequent classification of households.<sup>1</sup> This procedure includes household income, among other variables from the ENCASEH that were used to characterize the household members and their dwelling conditions. Upon comparison of the results of this identification procedure with the initial classification, based on the poverty line over the monthly per capita income, it is observed that they coincide in more than 80% of cases. This means that the discriminant analysis can pick up income as an important factor in determining the poverty condition of households and, moreover, that it is not the only factor determining living conditions. A more detailed explanation of these procedures can be found in another publication by the Mexican Social Development Ministry.<sup>1</sup>

## **2.3 Presentation of the findings in an assembly**

After identifying households considered in “extreme poverty,” thus eligible to benefit from the Program, and in keeping with the operation guidelines of the Program (especially item 5.2. (Initial Process for Inclusion in the Program) a community assembly was called in each of the intervention localities. Its purpose was to disclose to the members of the community the households selected to participate in the Program. Any comments were taken into account to review the cases according to inclusion criteria established for the identification of beneficiary families.<sup>6,7</sup>

### 3. Basis of the evaluation design

Given the uncertainty existing at the beginning of operations of the Program regarding the type and extent of the Program's effects,<sup>8</sup> it was decided from inception to evaluate its effectiveness.<sup>1</sup> To this end, two evaluation settings were ideal: one to measure the changes produced by participating in Oportunidades and another for counterfactual comparison,<sup>9-11</sup> to show what would have happened to the same population of beneficiaries, in an identical setting, but without the Program. These two settings would make it possible to evaluate the impact attributable to Oportunidades by direct comparison. However, this comparison is not feasible because it is not possible to simultaneously observe the same participants with and without the benefits of the Program.

The impossibility to observe the counterfactual scenario has been addressed in different ways, however all of them involve creating a hypothetical situation to estimate what would have happened in identical conditions, but without the Program. Evaluators often estimate the impact of programs by comparing the changes observed in the group of Program beneficiaries with those in another group. The group of non-beneficiaries is thought to provide empirical information on what would have happened in the absence of the Program. Therefore it is a simulation of the counterfactual setting, which can be valid if it meets the principle of exchangeability.<sup>9-11</sup> In other words, if the conditions of intervention or non-intervention were to be exchanged between the evaluation groups, they should show exactly the same expected result. That is to say that the populations are similar and that the observed differences can only be a result of the Program. This principle makes it necessary to assess the counterfactual scenario in a group that is as similar as possible to the group that will benefit from the Program; thus the definition of this control or counterfactual group is essential to obtain valid results.

Experimental evaluation designs randomly allocate individuals or other sampling units to the intervention or control groups. Therefore all participants have the same known probability of being allocated either to the intervention group or the control group. Random allocation assumes that there will be no differences between groups. Any difference will be due to random chance, not to bias.<sup>12</sup> This is why experimental designs are accepted to best meet the principle of exchangeability and are considered the most reliable and powerful study designs, from the statistical point of view, for the evaluation of social programs.

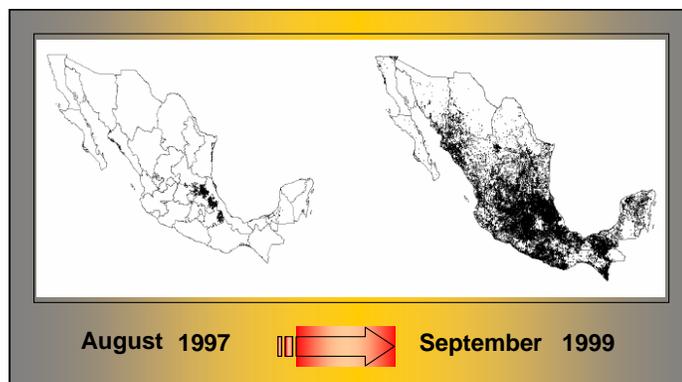
When random allocation is not possible due to ethical, financial, or practical reasons, a comparison group is selected; such a group should be as similar as possible to the group that is benefiting from the

Program in terms of the observable characteristics, thus maximizing the exchangeability principle. This can be done using matching or equalization of characteristics techniques, their purpose being to generate very similar groups. These groups are supposed to be similar to those obtained by random allocation. One drawback is that the degree of similarity can only be assessed with relation to the observed or the measurable variables. On the other hand, using random allocation, it is possible to suppose comparability even with relation to unobserved variables.

#### 4. Sample-based design for the impact evaluation

Randomization was done on the locality level specifically for the case of the Oportunidades Program in rural zones, taking advantage of the programmed expansion of the Program's coverage (see Figure 2). The process for randomly allocating localities to the control and intervention groups was as follows: i) a group of localities was selected that met inclusion criteria for the Program, ii) each locality was randomly allocated to the intervention or control groups, and iii) in each (intervention) locality identified to become a Program beneficiary, a poverty criterion (ENCASEH 97) was used to determine which households would be eligible to become beneficiaries of the Program. This was the selection process of the group of localities that would initially participate in the assessment of Oportunidades. It can also be denominated the original sample.

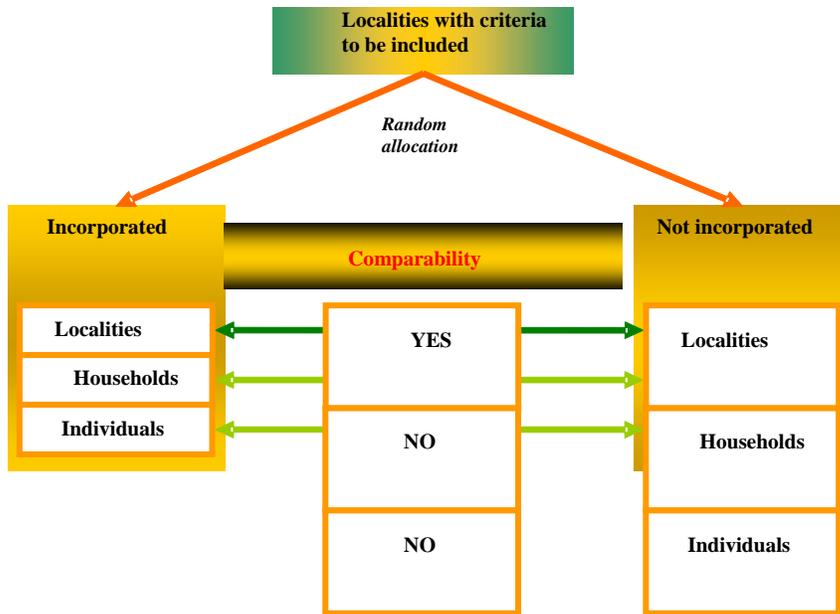
Figure 2. Expansion of Oportunidades coverage in rural zones between 1997 and 1999.



One of the costs of randomization on the locality level is that there is a greater likelihood of observing non-random patterns in terms of the differences between intervention populations and control populations, prior to receiving the benefits of the Program. A thorough assessment revealed that on the locality level randomization was satisfactory, but that on the households and individual levels there were significant

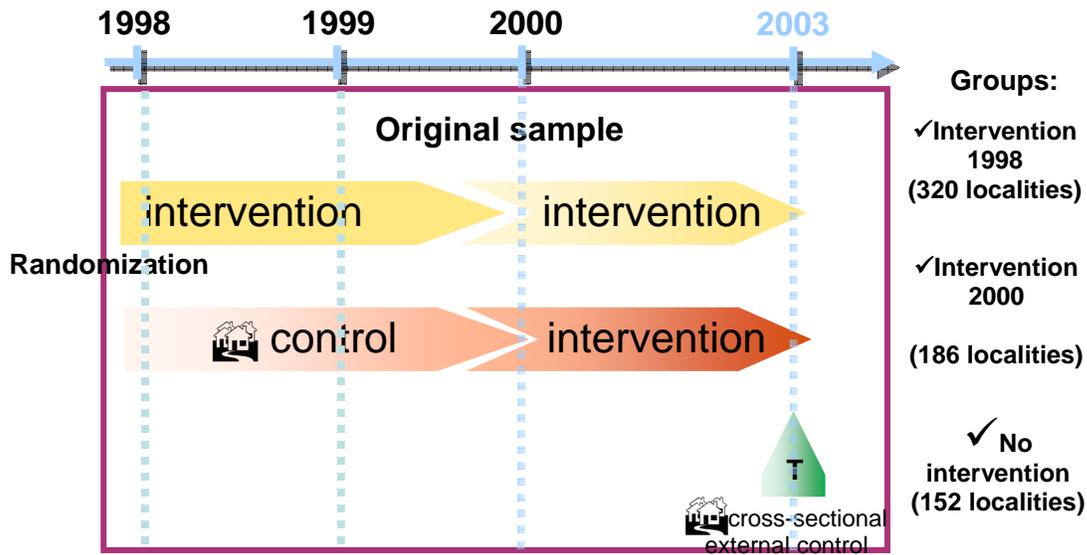
differences in some variables (see Figure 3).<sup>13</sup> This means any impact assessment of Oportunidades should include a method to correct for these initial differences, such as the difference in difference estimators (also called double-difference estimators).

Figure 3. Randomization of the participating localities in the assessment of Oportunidades in rural zones (original sample), and their impact on the comparability of the possible analysis units.



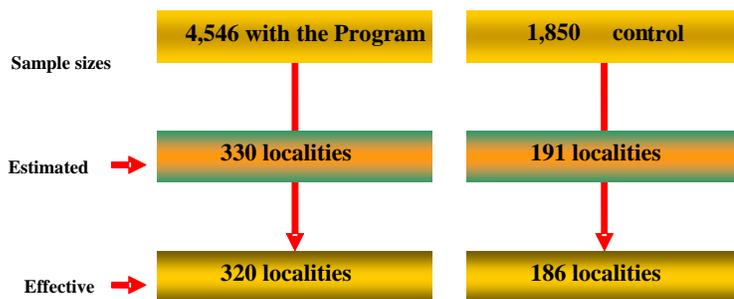
Since it was important, from the beginning of the Program, to estimate its short, medium, and long-term impacts, the design took into account data collection from a baseline (ENCASEH 97) and every six months thereafter, for an initial period of three years (1998 to 2000). This facilitated the short-term impact assessment. ENCASEH 97 is thus part of the baseline of the rural sample. These Household Evaluation Surveys (ENCEL) are identified by the year and month they were carried out, as ENCEL98M, ENCEL98O, etc. Later, another survey known as ENCEL 2003 was carried out to collect data to document the effects of the Program over a medium and short term (Figure 4). This is the latest survey carried out in rural zones.

Figure 4. Surveys that facilitate the impact evaluation of Oportunidades in rural areas.



Stratified sampling was used for the original evaluation sample. Stratification was conducted by locality size proportional to the number of localities, to generate a selection of localities with equal probability of selection for each locality size. Similarly, the sample size was estimated over a sampling universe of 4,546 localities to select 330 intervention localities, as well as over a universe of 1850 localities to select 191 control localities, using the distribution proportional to locality size. This estimate was done keeping in mind that there are no prominent differential costs of data collection across the different strata. The sample selection was done according to the sampling frame mentioned before, through systematic selection with a random start. After some adjustments for field work, the final sample was defined at 320 intervention localities, where the Program first operated, and 186 control localities, where Oportunidades began to operate after some time (Figure 5).<sup>14</sup>

Figure 5. Initial sample size of the localities included in the Oportunidades evaluation in rural areas.



The original impact evaluation simple in rural areas included localities in seven status of Mexico: Hidalgo, Puebla, Guerrero, Veracruz, Michoacán, San Luis Potosí and Querétaro, with a total of 506 localities (320 of intervention and 186 of control).

To evaluate the medium-term impact, ENCEL 2003 was carried out that year in all 506 localities that were part of the original evaluation sample. By that time all localities that had participated in the previous ENCEL surveys had been included in the Program (320 benefiting from the Program since 1998 and 186 starting in 2000). Therefore it was decided to have a comparison group without the benefits of the Program. In that case 151 additional localities where the Program had not yet been offered were included in the sample.<sup>15</sup> This strategy produced three comparison groups: two with different times of exposure to the Program (in 1998 and 2000) and one that was not benefiting from it until 2003 (new control group) ¶. The following is a more detailed description of this new control group.

The new control group was required to be as similar as possible to those that were already benefiting from Oportunidades, to quantify the effect of the Program over the results evaluated after six years of intervention. Therefore the selection procedure consisted of two steps: i) identifying some predictors measured in the inclusion of households that were already participating in ENCEL surveys and ii) using them to calculate the propensity or tendency to be included –this estimation being inclusive of the households already participating in the ENCEL survey and households included in the National Income and Expenditure of Households Survey (ENIGH for its initials in Spanish) found in rural areas.<sup>16</sup> This survey is statistically

¶ This new control group joined the Program in 2004.

representative of the Mexican Republic. To estimate the likelihood of incorporation to the Program, the multiple matching propensity score method was used, working with data added on the locality level as a function of several observable characteristics that permitted comparison with the localities of the original sample.<sup>17</sup> A more detailed and technical description of the variables included in the propensity score, the selection of the localities included as a new control group and first-round results of the comparison of some results, is found in other documents.<sup>15,17</sup> A technical note on the construction of the propensity score is available on our web page under the section Other Technical Notes.

## 5. Evaluation Surveys

### 5.1 Household Socioeconomic Characteristics Survey (ENCASEH)

ENCASEH is the main data collection instrument of the Program to identify beneficiary families. The data regarding the rural evaluation sample was obtained in 1997 by individual interview to each of the resident households in intervention and control localities. ENCASEH is a questionnaire applied prior to inclusion of eligible households to the Program. The data thus obtained served as the basis to determine the poverty condition of each household and to identify potential beneficiaries of the Program (eligible households). The following table summarizes the topics referred to in the ENCASEH Survey.

Table I. Main topics included in the ENCASEH Survey.

TOPICS OF THE SURVEY
1. Demographic characteristics of the household
2. Human capital of the household members
3. Economic activities of the household members
4. Income of the household members
5. Condition of the dwelling
6. Availability of services
7. Ownership of goods

The following table presents the population of the original evaluation sample for which information is available in this survey.

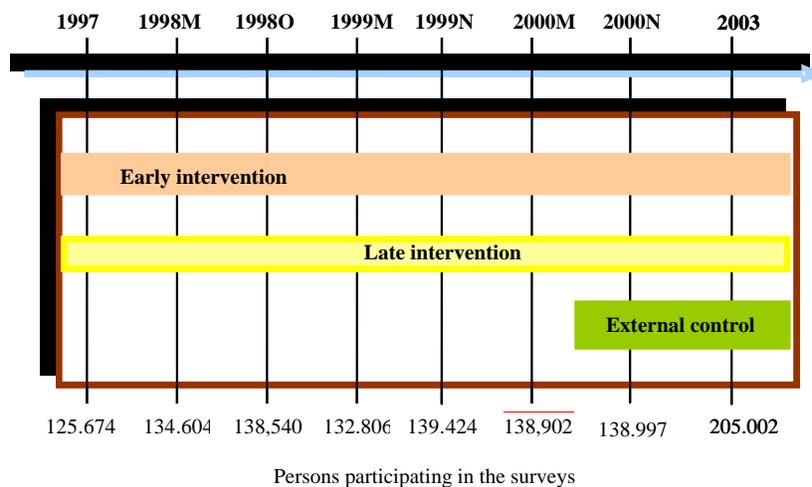
Table II. Population participating in the ENCASEH 1997, by sex and eligibility status

Type of Locality	Males	Females	No data	Total
Eligible	60,585	59,507	49	120,141
Non-eligible	2,748	2,785	0	5,533
Total	63,333	62,292	49	125,674

## 5.2 Household Evaluation Surveys (ENCEL)

The objective of ENCEL surveys is to collect data on different topics regarding the household as a whole as well as household members. ENCEL surveys are made up of different questionnaires addressing particular topics. Although the questions may vary from one survey to another, depending on the time period needed to observe, through certain indicators, the results attributable to intervention by the Program, preserving the comparability of the questions throughout time is always sought. The section describing the databases briefly explains the topics addressed in each of these surveys. Figure 6 summarizes the population (number of individuals) participating in the ENCEL and ENCASEH surveys, showing the changes between each survey.

Figure 6. Samples available in the databases of the Oportunidades evaluation in rural areas.



The first round of evaluation surveys, ENCEL 98M, sought to complement and describe in detail the data collected by the ENCASEH 97. The data from both these sources provided a baseline measurement of the situation of the households, before the families' inclusion in the Program. ENCEL 98M was performed in March 1998 on the totality of resident households of all 506 localities, intervention and control together, bringing in data from a total of 24,077 households.

ENCEL 98O, the second evaluation survey, comprised the first measurement panel of the Program impact. This survey was performed in November 1998 on the same 506 localities visited at the baseline data collection and on all households there (which were interviewed in the ENCASEH and ENCEL 98M). Using this same strategy the data collection of ENCEL 99M was done. The same applied to ENCEL 2000M, performed in May 2000, and ENCEL 2000N, performed in November 2000.

The main data collection instrument in the ENCEL 1998-2003 surveys was the questionnaire on socioeconomic characteristics of the households. In 2003, in addition to applying this questionnaire, and in order to evaluate the impacts of the Oportunidades Program on the beneficiary population's reproductive health, a fertility questionnaire was applied to a sample of 16,622 women selected between the ages of 15 and 49 years. For this data collection a representative selection of the three types of localities was done: localities participating in the Program since 1998, those participating since 2000, and those in the new control group. The three groups were divided into seven strata, each corresponding to one of the Mexican States where the survey took place. Afterward a selection of localities and households was performed taking into account the percent distribution of women of childbearing age by state and by locality. The selection procedure ended on the household level. All women of childbearing age in each household were surveyed. The selection of women was carried out by the National Population Council (CONAPO for its initials in Spanish).<sup>18</sup>

It was considered important from the beginning to have access not only to the data obtained on the household level, but also data that could account for the economic, social, infrastructure, and demographic characteristics that could affect the results of the indicators of interest. Therefore a questionnaire was applied as a part of ENCEL 2003. This questionnaire was applied in all localities of the evaluation sample, one per locality. The interview was carried out with the community authorities, leaders, or public officials who could provide relevant information. The objective of this questionnaire was to gather information on the local infrastructure, availability of services, main economic activities, and organization of groups or association that congregate to perform some community activities. In addition, a questionnaire regarding prices was applied to

obtain the prices of the main foodstuffs, farming products, and pharmaceutical products of each locality included in the sample. The data obtained through this questionnaire are important because they are useful in calculating price indexes that will to a great extent reflect the wellbeing of the population, thus partly representing the general state of the locality and in which way the products are being distributed and sold.

The Program is concerned with evaluating the impact of Oportunidades on indicators of demand of education services such as school assistance, dropout rate, failing, reenrollment, etc. In addition, it also sought to assess the prevailing conditions of middle and high education provision. To this end, ENCEL 2003 included a questionnaire for schools that was applied to a sample of 987 schools, the questionnaire being answered by the Principal and by two teachers of each school. This information is to facilitate documentation of the characteristics of education services used by household members. To select the sample of schools that would be surveyed, priority was given to schools where the majority of Oportunidades beneficiaries (or potential beneficiaries, as in the case of control localities) attended. Schools where such subjects attended and which were within a 10 km radius of any of the 506 localities were selected. Weighting was done considering the number of beneficiaries per school so that every beneficiary would have the same probability of having his or school selected. Unlike intervention localities, here it was impossible to take into account the beneficiary attendance criterion as there were no beneficiary households. Thus the only relevant weighting was distance, which made nearer schools more likely to be selected.<sup>19</sup>

As a part of the 2003 rural data collection the National Institute of Public Health was in charge of planning and executing the biologic component of this survey. The data collection was carried out in all 506 intervention localities and all 151 non-intervention localities. It consisted of the collection of biologic samples, anthropometric measurements, cognitive development tests, and questionnaires from clinics. The target population for this data collection was 1) children 2 years to 5 years 11 months of age and household characteristics, 2) Adults and mothers of children 2 years to 5 years 11 months of age, and 3) adolescents aged 15 years to 20 years 11 months. The sample size that was originally proposed was 9,000 children in the 2 year to 5 year 11 month group (3000 from each type of locality). The same sample size was proposed for the adolescent group with the same distribution among the three population groups. The proposed sample size for the adults group was 6000.

The biologic data collection included several tests and questionnaires applied to the three population groups mentioned above. Most of these tests and questionnaires have been tested in many countries,

including Mexico. Therefore it was necessary to design only some sections of the questionnaires in which questions were included regarding the health status of the subjects and their education and work history. Table III shows the tests and measurements that were performed with each of the target groups.

Table III. Biologic tests applied to the study population

Tests	Children	Adolescents	Mothers	Adults
Saliva	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	
Anemia	<input checked="" type="checkbox"/> (H)	<input checked="" type="checkbox"/> (P/A)	<input checked="" type="checkbox"/> (H)	
Height and weight	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Glucose				<input checked="" type="checkbox"/>
Pregnancy		<input checked="" type="checkbox"/>		
General urine test		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Pulse	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Blood pressure			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Exertion test				<input checked="" type="checkbox"/>
Specimens		Urine Filter paper	Venous blood	Venous blood

\* Children 24 to 36 months of age; H: Hemocue; P/A: Presence/Absence

The additional data included in the questionnaire were

- Children's questionnaire:** sections regarding the health status and behavior of children
- Adolescents questionnaire:** sections related to risk behaviors, stress, depression, health status, eating patterns, and education and work history.
- Adults questionnaire:** sections related to eating patterns, risk behavior associated with chronic diseases, health status, stress, and morbidity.

The general data collection of ENCEL 2003 was carried out from September 22 to December 4, 2003. The biologic data collection was carried out from July to October 2003. A data-entry company was hired and directly supervised by the National Coordination of the Oportunidades Program and by INSP to ensure that data-entry programs were duly validated and that the quality of the process was satisfactory.

## 6. Database Organization

In this website all data from the rural evaluation were organized by survey (ENCASEH 97, ENCEL 98M, ENCEL 98O, ENCEL 99M, ENCEL 99N, ENCEL 2000M, ENCEL 2000N, and ENCEL 2003). For each survey folders were created containing the tables that make up the survey, in SPSS format, as well as the questionnaire in PDF format with the label for each question exactly as it appears on the database. All data concerning the household and its members was put together in a single table (or database). It was done for each survey except for ENCEL 2003.

The data from ENCEL 2003 are available according to the questionnaire type used in this survey (socioeconomic, biologic, fertility, locality, clinic, and school). A reference folder was created containing the different tables that comprise the total data of that questionnaire, as well as the questionnaire in PDF format with the label for each question exactly as it appears on the database. The data obtained by the socioeconomic data were divided into two tables: household-level data and household member data.

The folders were named in such a way that one could easily identify what survey it belongs to, the questionnaire or data it contains, and the date of the version it corresponds to. An example is `bd_rur_2003_socioeconomico_2004-11-29`. This folder contains the **socioeconomic** database from the **rural survey ENCEL 2003**, and it corresponds to the version from **November 29, 2004**. Although database cleanliness guarantees a consistency error lower than 10%, the cleanliness process is still current, and thus certain databases may be updated at any given time. It is advisable to regularly check for the latest version.

Other databases are also available on this website. These are closely related amongst themselves. The following tables summarize the total folders in each rural evaluation survey and briefly describe the data and the number of tables that they contain.

Database corresponding to ENCASEH 1997			
Folder	Format	Number of tables	Content
bd_rur_1997_2004-07-08	SPSS	1	Includes the information collected by the questionnaire on socioeconomic characteristics of households. (ENCASEH 1997) for the original evaluation sample

Database corresponding to ENCEL surveys 1998 through 2000			
Folder name	Format	Number de tables	Content
bd_encil98m_2005-02-28	SPSS	1	All of them contain information collected by the questionnaire on socioeconomic characteristics of households ) for the original evaluation sample
bd_encil98o_2005-02-28	SPSS	1	
bd_encil99m_2005-02-28	SPSS	1	
bd_encil99n_2005-02-28	SPSS	1	
bd_encil00m_2005-02-28	SPSS	1	
bd_encil00m_2005-02-28	SPSS	1	

Base de datos correspondiente a la ENCEL 2003			
Nombre de la carpeta	Formato	Número de tablas	Contenido
bd_rur_2003_localidad_2004-11-26	SPSS	1	Economic, social, and infrastructure data of the locality
bd_rur_2003_socioeconomico_2004-11-29	SPSS	2	Data from the socioeconomic questionnaire; two tables: one with household-level data and the other with household member data
bd_rur_2003_fecundidad_2004-11-26	SPSS	1	Data from the fertility questionnaire
bd_rur_2003_hogares_1997_2004-11-26	SPSS	1	Retrospective data to 1997 of resident households in the localities of the new control group. That is, households that were not beneficiaries of Oportunidades before 2003
bd_rur_2003_precios_2004-11-26	SPSS	1	Data on prices in the locality
bd_rur_2003_director_2004-07-14	SPSS	1	Data from the questionnaire applied to the School Principal
bd_rur_2003_maestro_2004-07-14	SPSS	1	Data from the questionnaire applied to the School Teacher
bd_rur_2003_transf_rurales_2004-11-26	DBF	1	Contains the households that were identified as included in the Program. This identification was done by means of money transfer administrative records. Also contains the record of money transfers received by the household between 1998 and 2004
bd_rur_2003_biologico_2004-11-26	SPSS	1	Data on the biologic data collection

## 6.1. Key Variables

To link the different analysis units the following will be considered: i) for the analysis on the household level the “key” variable, the variable that links the data of each household to the individuals members of that household is **folio**; ii) for the analysis of individuals the key variables are **folio** and **renglón**. **Folio** is the household identifier and **renglón** is an identifier for individuals, which indicates the position assigned to each person in the household. This means that to link the individual data through the panel it will be necessary to do it by combining the variables **folio** and **renglón**.

In addition to key variables, ENCEL databases contain other variables that could be useful for the analysis, as the ones shown in the following table.

Variable	Type	Content
<b>yycali</b>	continuous	Score according to the Program criteria used in 1997
<b>pobre</b>	{0, 1}	Classification of poverty level according to the Program criteria used in 1997 1: Poor 0: Not poor.
<b>contba</b>	{1, 2}	Type of sample: 1: Intervention 2: Control
<b>indice</b>	continuous	Marginality index of the locality
<b>grado</b>	{4, 5}	Degree of marginality of the locality: 4: High 5: Very High
<b>hog_nue</b>	{1, 2, 3, 4}	Present only in ENCEL 2003, and it indicates 1: Households that belong to the original sample; that is, those that were interviewed at the time of the first round of evaluation 2: Household in an intervention locality (98 or 00) that was recruited to the sample in ENCEL surveys posterior to the first data collection. 3: Household in an intervention locality (98 o 00) that was not found until the 2003 data collection 4: Household new to the sample due to belonging to a control locality in 2003
<b>tipo</b>	{1, 2, 3}	Present only in ENCEL 2003, and it indicates 1: Household in an intervention locality in 1998 2: Household in an intervention locality in 2000 3: Household in an intervention locality in 2003
<b>mpcalif</b>	continuous	Score according to the Program criteria used in 2003
<b>mppob</b>	{1, 2, 3}	Classification of poverty level according to the Program criteria used in 2003: 1: Poor 2: Almost poor 3: Not poor

## 7. Proposal for analysis strategies

Oportunidades databases can be of use in assessing the impact of the Program and for other research purposes as well. In this section some comments on methodology are given and some analysis strategies proposed for the impact evaluation. If the objective is to study the impact of the Program it is important to take advantage of the longitudinal nature of the study so as to obtain double-difference or Difference in Difference (DID) estimators that facilitate the control of baseline differences in the desired indicator that are common to both comparison groups. Also, they would allow the net effect of the Program to be isolated from other, changing factors throughout time that might affect that indicator. Generally speaking, the intention is to compare the situation the beneficiary families had previous to the Program (termed their base conditioned at an initial moment  $t_0$ ) and the changes these families show and which can be attributed to the benefits of the Program (measured at a later moment  $t_1$ ).

Conversely, to optimize specificity in this comparison, the above results are compared with other similar measurements (also taken at times  $t_0$  y  $t_1$ ) but in a population of families that are similar to beneficiary families except in that they do not receive the benefits of the Program. This second, comparison population is the control population. Thus the impact of the Program for a given impact B in the beneficiary population is the difference between measurements  $B_{t_1}-B_{t_0}$ . On the other hand, if C is assigned as the same indicator but this time measured in the control population, the difference  $C_{t_1}-C_{t_0}$  denotes the changes that may have occurred in said indicator for reasons unrelated to the Program. So to obtain the net effect of the program this difference is usually subtracted from the first difference,  $(B_{t_1}-B_{t_0}) - (C_{t_1}-C_{t_0})$ , which is an index known as the double difference. After subtracting the changes not attributable to the Program from those observed in the beneficiary population, the result is a reasonable measurement of its specific impact. Subtracting the difference  $(C_{t_1}-C_{t_0})$  adjusts for the observed  $B_{t_1}-B_{t_0}$  effect of the Program, removing possible changes that other factors may have introduced in indicator B and which cannot be attributed to the Program.

In general this analysis strategy was the most frequently used by researchers of the International Food Policy Research Institute (IFPRI) who documented the effects of the Program up until 2000, in a series of health, education, and other sociodemographic indicators. The technical details of those analyses can be viewed directly in the documents available at the website of the external evaluation of Oportunidades (<http://evaloportunidades.insp.mx>).

As has been mentioned before, the evaluation sample for the ENCEL 2003 data collection was modified in the design in that a new control group was included. This group was selected by matching based on a series of observable characteristics of the locations and households, in trying to maintain comparability in these characteristics with the early intervention (households benefiting since 1998) and late intervention (households included in the Program in 2000) groups. The proposal is that the impact evaluation between these groups can be based on the matching method with propensity score or with a discontinuous regression design.<sup>15</sup>

There are many possibilities for minimizing or controlling the problem of lack of comparability. However, the most commonly used by the Consulting Group of the quantitative assessment of Oportunidades in the recent analyses was the one based on the propensity score, to refine the comparability between the household groups according to their likelihood of being included in the Program (propensity score matching).<sup>20-</sup>  
<sup>23</sup> This methodology is the most recommendable for comparing the intervention groups and the group of control households, when the intervention is not randomly allocated. Matching by propensity scores is a way of “correcting” the estimate of the effects by controlling for possible confounding factors and selection bias, based on the fact that bias and the possibility of confusion are minimized when comparing the subjects that received the intervention with those who did not, all things being equal. Theoretical assumptions are that the likelihood of inclusion in the Program depends to a great extent on the observable characteristics, and that, on average, compared groups only differ with respect to participation in the program – they are exactly the same as far as other aspects. A condition for constructing the propensity score is that it must be a function of variables that are not modified by the Program. That means it depends of characteristics that are prior to participation in the Program. For this analysis in particular it is thus recommended to use the sociodemographic data retrospective to 1997 of the households that made up the new control group. Once the propensity score is constructed there are several methods available to perform matching. Of these, the most commonly used are the nearest neighbor method, the Kernel method, the Mahalanobis matching method, and the matching by stratification method.<sup>15,20</sup>

Moreover, the data from the evaluation surveys in rural areas can be used to perform an *Intention to treat analysis* considering only household allocation to intervention or control areas, regardless of whether the households in intervention areas are beneficiaries or not.

## 8. References

1. Secretaría de Desarrollo Social. Más oportunidades para las familias pobres. Evaluación de Resultados del Programa de Educación, Salud y Alimentación. Primeros Avances, 1999. México DF: Programa de Educación, Salud y Alimentación (Progresas); 1999.
2. Widaman KF. Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multiv Behav Res* 1993; 28:263-311.
3. Buhler A, Deutler M. Optimal stratification and grouping by dynamic programming. *Metrika* 1975; 22:161-175.
4. Consejo Nacional de Población. Índice de marginación a nivel localidad, 2000. México DF: CONAPO; 2002.
5. Huberty CJ. Applied discriminant analysis. New York: Wiley-Interscience; 1994.
6. Secretaría de Desarrollo Social. Acuerdo que establece los Lineamientos Generales para la Operación del Programa de Educación, Salud y Alimentación (PROGRESA). *Diario Oficial de la Federación*, Friday February 26, 1999.
7. Secretaría de Desarrollo Social. Acuerdo que establece las Reglas de Operación del Programa de Educación, Salud y Alimentación (PROGRESA). *Diario Oficial de la Federación*, Thursday March 15, 2001.
8. Petticrew M. Presumed innocent. Why we need systematic reviews of social policies. *Am J Prev Med* 2003; 24(3S):2-3.
9. Winship C, Morgan SL. The estimation of causal effects from observational data. *Annu Rev Sociol* 1999; 25: 659-706.
10. Pearl J. Causal inference in the health sciences: A conceptual introduction. *Health Serv Outcomes Res Method* 2001; 2:189-220.
11. Parascandola M, Weed DL. Causation in epidemiology. *J Epidemiol Community Health* 2001; 55:905-912.
12. Schulz KF, Grimes DA. Allocation concealment in randomised trials: Defending against deciphering. *Lancet* 2002;359:614-618.
13. Behrman JR, Todd PE. Evaluación preliminar de los tamaños de muestra para la evaluación del Programa de Educación, Salud y Alimentación (Progresas). Technical Report IFPRI; s.f.
14. Orozco M, Parker S, Hernández D. El modelo de evaluación de Progresas. En: Secretaría de Desarrollo Social. Más oportunidades para las familias pobres: evaluación de resultados del programa de educación, salud y alimentación. Metodología de la evaluación de Progresas 2000 (1a reimp). México: 2003:1-29.

15. Todd P. Technical note on using matching estimators to evaluate the Oportunidades program for six year follow-up Evaluation of Oportunidades in rural areas. Technical Report; 2004.
16. Instituto Nacional de Estadística, Geografía e Informática. Encuesta nacional de ingresos y gastos de los hogares. Available at: <http://www.inegi.gob.mx>
17. Díaz JJ, Sudhanshu Handa S. An assessment of propensity score matching as a non experimental impact estimator: evidence from Mexico's PROGRESA program. *J Hum Res* (Forthcoming).
18. Consejo Nacional de Población. Encuesta para medir el impacto en salud reproductiva del Programa Oportunidades 2003. Diseño muestral (technical document). México DF: CONAPO; 2004: 1-14.
19. Programa De Desarrollo Humano Oportunidades. Diseño de la muestra de escuelas para el levantamiento de los Cuestionarios de Director y Maestro, que forman parte de la ENCEL 2003 (Documento técnico). Oportunidades; s.f.
20. Becker SO, Ichino A. Estimation of average treatment effects based on propensity scores. *Stata J* 2002;2:358-377.
21. Rosenbaum, PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41-55.
22. Dehejia RH, Wahba S. Propensity score matching methods for non-experimental causal studies. *Rev Econ Stat* 2002; 84(1):151-161.
23. Heckman JJ, Ichimura H, Todd P. Matching as an econometric evaluation estimator. *Rev Econ Stud* 1998;65:261-294.