DESIGN REPORT

Design and Implementation of MCA Honduras Program Evaluation

Original Contract with MCA dated May 25, 2007

Current Contract #: MCC-10-0113-CON-20 TO01

REVISED
DECEMBER 17, 2010
AND
APRIL 10, 2011

PRESENTED TO:
Millennium Challenge Corporation
875 15th Street, NW
Washington, DC 20005

PRESENTED BY:
NORC at the University of Chicago
Shana Brown
Survey Director
55 East Monroe Street
30th Floor
Chicago, IL 60603
(312) 759-4000
(312) 759-4004

NORC at the UNIVERSITY of CHICAGO

Table of Contents

A. EXECUTIVE SUMMARY 1

A.1 Evaluation Designs 2

Transportation Project 2

Rural Development Project 3

A.2 Data Sources. 7

Transportation Project 7

Rural Development Project 8

Use of Geographic Information System as an input into the evaluation model 8

B. INTRODUCTION: MCA HONDURAS PROGRAM OVERVIEW 11

C. TRANSPORTATION PROJECT 13

C.1 Evaluation Goals 13

C.2 Key Features of the Evaluation Design 14

Use of existing external models (GIS and HDM) and derived variables 17

C.3 Data Collection Requirements of the Evaluation Design 19

Household Survey 19

Price-and-product survey 21

Traffic surveys and data on roads 21

GIS data.. 22

D. RURAL DEVELOPMENT PROJECT: FARMER TRAINING AND DEVELOPMENT ASSISTANCE 25

D.1 Evaluation Goals 25

D.2 Evaluation Design 25

D.2.1 Experimental-Design Evaluation Approach 26

D.2.2 Model-Based Evaluation Approach 27

D.3 Data Collection Requirements of Evaluation Design 31

E. QUALITATIVE ANALYSIS OF PROGRAM EFFECTS 34

Key informant interviews 34

Focus groups.. 35

F. IMPLEMENTATION SCHEDULE.. 36

ANNEX I: SAMPLE SIZE & ALLOCATION 39

A. EXECUTIVE SUMMARY

NORC is pleased to be working with MCC and MCA Honduras in designing and helping to implement the evaluation component of the MCA Honduras Monitoring & Evaluation Plan. The evaluation design will assess the impact of the two projects funded under the Honduras Compact: the Transportation Project and the Rural Development Project. The Compact has as its overall goal the stimulation of economic growth and poverty reduction through these two projects.

Specifically, the evaluation design will estimate the causal relationship between the Projects and the Compact Goal by means of a rigorous impact evaluation. The evaluation will also include some supplementary analyses to understand reasons why the Compact Goal was or was not achieved; identify positive and negative unintended results of the Program; highlight lessons that may be applied to similar projects; and assess the sustainability likelihood of its results over time. The quantitative data collection being undertaken for the evaluation will feed directly into the rigorous impact evaluation, which is the primary purpose of this contract. Qualitative data collection, which may be added to the scope of work for the current contract, would provide depth to supplementary analyses mentioned above.

In particular, the impact evaluation seeks to measure six benefit streams that feed into the income of program beneficiaries (with results to be disaggregated by socioeconomic characteristics such as income, gender, and age where appropriate):

- Transportation Project
  - Benefits of Highway CA-5 upgrading
  - Benefits of secondary road paving
  - Benefit of farm-to-market roads
- Rural Development Project
  - Increase in farm income
  - Increase in employment income
  - Benefits of Agricultural Public Goods Grant Facility[1]

This report presents methodologies for estimating these benefit streams and evaluating the impact of each project. For each project, the report presents key hypotheses to be tested; the evaluation methodology to

[1] The evaluation of the Agricultural Public Goods Grant Facility was completed under NORC's contract with MCA Honduras, and will not be further pursued under the current contract with MCC.

be used, along with an explanation of why it was selected; and quantitative and qualitative data collection requirements for the evaluation design. The report also contains a detailed work plan for implementing the evaluation methodology.

Summaries of these evaluation methods for each project are presented below.

A.1 Evaluation Designs

Transportation Project

The Transportation Project aims to reduce transportation costs between targeted production centers and key national markets and ports. The initial scope called for upgrading and paving two major sections of Highway CA-5, pave at least 70 km of secondary roads, and develop a vehicle weight control system. Under the Rural Development Project, MCA Honduras is upgrading and paving at least 600 km of rural roads (farm-to-market roads). Given that it is part of the national road network, for the purpose of this evaluation we are considering the evaluation of the rural roads improvement within the framework of the Transportation Project. Due to increases in costs and a partial re-scoping of the road rehabilitation component, about 65 km of secondary roads and 495 km of rural roads were rehabilitated.

The evaluation design presented in this report differs from that outlined in the original MCA Honduras M&E Plan, which proposed the estimation of a single before-after benefit stream that accrues in the form of decreased vehicle operating costs and decreased travel time for CA-5 and secondary roads, and a separate estimation of changes in income for those households within a specified zone of influence around the rural roads, as compared to a comparison group of households outside the zone of influence. The design that NORC is using for the evaluation of the Transport Project:

- Calls for integrating the separate evaluation models initially considered for the major road projects into a *single, integrated model* that better represents interrelationships and interdependence among different road segments. The integrated model represents the physical road network as a mathematical network (using a GIS system), allowing us to capture network interrelationships. It is a more accurate representation of the real world that yields evaluation results that are more valid and precise.
- Seeks to measure the impact of the Transportation Project on an expanded set of benefit stream variables such as household income and employment; business income and employment; prices of products; and accessibility to health facilities and schools. These benefit streams, as well as the original benefit streams in the M&E Plan will be calculated, with careful attention to avoiding

double counting. The intent is to capture the *multiple benefit streams* that accrue to the large-scale improvement of road networks that MCA Honduras is undertaking.

- Allows us to relate changes in these benefit stream variables to changes in travel-time/costs on a continuous scale, enabling us to estimate varying levels of *program impacts at different levels of treatment effect*. This is a considerable improvement over a dichotomous or before-and-after model[2].

The evaluation model will be national in scope, including all roads for which GIS data are available from existing sources. Data collection will also be national in scope, but with heavier sampling in MCA-H project areas. Extending the model from a local or regional scope to national scope makes very little difference in the level of effort required to collect the data. This requires no additional effort with respect to the GIS data, and no additional effort with respect to the sample sizes for the household and firm surveys.

Rural Development Project

The Rural Development Project seeks to increase the productivity and business skills of farmers who operate small- and medium-sized farms and their employees. Four activities will be implemented: (1) farmer training and development assistance (FTDA); (2) facilitation of access to credit by farmers; (3) upgrading of farm-to-market roads; and (4) provision of agriculture public goods grants. The ongoing evaluation of the FTDA activity is summarized below.

The evaluation design to be used for the FTDA/EDA was intended to be an experimental design that called for randomly allocating farming communities ("aldeas") into two groups: those that received technical assistance at the start of the program (the treatment communities) and those that received it as the program continued. Baseline and follow-up data collected from individual program farmers in these two groups would be used to assess – through "double difference" estimates – the impact of program interventions on changes in several variables including farm and off-farm income, farm employment, crop types (subsistence versus horticultural cash crops), and access to, sources and use of credit. Under this randomized control trial, NORC sought to replicate the EDA implementing agency, Fintrac's, farmer selection criteria and process to identify farmers in treatment and control aldeas. NORC worked closely with Fintrac to recreate the exact Fintrac screening process at the aldea and farmer level. Fintrac's

[2] The model initially proposed (in October 2007) for the evaluation of the Transportation Project also called for using randomization in the selection of rural roads to be upgraded. This feature would have provided us with a wide range of variation in the treatment effect, with some of that variation introduced through randomization, thereby allowing us to more confidently attribute *causality* of observed changes to the program. As the project progressed, however, it was not feasible to implement randomization in the selection of rural roads. As such this element of the design was dropped.

technicians were involved in training screeners and identifying the quantifiable selection criteria to be used in the screening process. However, despite Fintrac's participation in both the development of selection criteria, input on screener forms and their participation in the interviewer training NORC and MCA were unable to fully replicate Fintrac's farmer selection criteria, partly because some criteria were subjective and could not be articulated and measured quantitatively, and partly because the selection criteria kept evolving as the program implementation progressed. Because Fintrac was unable to articulate an objective and fixed set of selection criteria to inform the evaluation, the screening process yielded a large number of farmers that were then rejected by Fintrac from the experimental-design sample. A second attempt to expand the selection criteria and redo the sample selection (Cohort 2) had similar results, with Fintrac accepting relatively few of the selected aldeas and potential program farmers into the FTDA/EDA Program. As a result of these failed attempts it became increasingly clear that that the evaluation would have to proceed with significantly diminished sample sizes, which would result in low precision for estimates of program impact and low power for tests of hypothesis about program impact.

There are two impact measures of interest in this evaluation. The **average treatment effect** (the ATE or the intent-to-treat effect) serves as a measure of the impact of a treatment program or policy on a population of interest (whether they receive treatment or not, or drop out, or "cross over"). The **average treatment effect on the treated** (or average treatment effect conditional on treatment, or ATT) serves as a measure of the impact of the treatment on a treated individual (i.e., on an individual who actually receives the treatment).

It is important to recognize that the ATE always refers to a particular population, such as a population of eligible individuals, or a population that is believed to include a substantial proportion of eligible individuals, or perhaps even an entire country. The value of the ATE will differ for each of these populations of interest, even for the same program intervention. For the FTDA program evaluation, the experimental design was applied to all areas of Honduras that had not yet been treated by Fintrac; that is aldeas for the study were selected from the population of aldeas that Fintrac had not yet entered, but planned to enter in the future. In order for an experiment to be able to detect a program-caused effect of specified size with high power (probability), the population of interest (i.e., the population to which the program is to be applied, and from which samples will be selected) must be relatively "rich" or "dense" in individuals who actually receive treatment. Unfortunately, the "density" of eligible farmers falling in the areas sampled in this study turned out to be low. As a result, the magnitude of the ATE relative to this population is likely to be very small. The key issue here is not the magnitude of the ATT – it is that the ATE for the two-thirds of Honduras that Fintrac had not yet treated is small.

This situation begs two questions:

(1) How large an income effect (ATT) would have been necessary to detect an ATE of a specified size with high power with the current sample and design?

(2) How large a sample of treated farmers would have had to have fallen in the experimental-design sample to detect an ATE of a specified size with high power (i.e., what "density" of Fintrac-eligible and treated farmers would have been required for the experimental design, to detect an ATE of specified size with high power)?

From direct observation of the Fintrac treatment, it is apparent that the income change caused by the treatment can be substantial, and large income changes (25%, 50%, and 100%) were therefore used in the power calculations done at the beginning of the project. However, because of the small percentage of Fintrac-treated farmers that ultimately ended up in the experimental-design sample, it is quite likely that the magnitude of the change would have to be extremely high to detect an ATE of a modest size with high power. NORC will use the data collected to conduct analyses to identify the upper bound of this income change. We will also estimate the sample size that answers the second question, which asks how "rich" the experimental-design sample would have had to be in Fintrac-treated farmers, in order to detect an ATE of a modest size with high power.

In summary, for the FTDA evaluation, while experimental-design data can be used to construct an unbiased estimate of an intent-to-treat or average treatment effect estimate of program impact, the small number of farmers accepted by Fintrac for the EDA program assistance makes it likely that this effect would be too small to detect. Furthermore, the very small sample size of treatment farmers would makes it difficult to estimate the average treatment effect on the treated (ATT). In view of the likelihood that the outcome of an intent-to-treat analysis of the experimental-design sample is unlikely to show impact, NORC and MCC have decided to supplement the originally envisioned experimental-design with additional analyses of impact.

Towards this end, NORC will use additional data on treated Fintrac clients (from Fintrac's own client lists that are not linked to the evaluation), to enable the construction of an estimate of the average treatment effect on the treated (ATT). This revised approach would include:

(1) Analysis of the original experimental design data as an experimental design intended to produce a design-based intent-to-treat estimate of the average treatment effect. This intent-to-treat analysis of the experimental design data is referred to as a "design-based" approach.

As discussed above, we do not now expect to detect significant changes in income, as measured by the ATE, using this design-based approach, because of the small number of treated farmers in the population sampled for the experimental design. We will, however, estimate the number of Fintrac-treated farmers that would have had to have fallen in the experimental-design sample for us to have been able to detect a change of a specified size at high power, using the design-based approach (with the same design structure and sample size).

(2) Analysis of the supplemented data comprised of experimental design data plus the added sample of Fintrac clients, intended to produce a model-based estimate of the average treatment effect on the treated[3]. This analysis of the augmented sample is an "observational study" or a "model-based" approach.

The observational study will make use of the data that were collected for the original experimental design, and complement it with additional sample data. The additional sample data have been collected from a random sample of Fintrac clients that entered the program at the same time as the second set of Cohort 2 farmers. Sufficient additional sample units have been selected so that the total sample size should be adequate to achieve a satisfactory level of precision and power for inferences about the average treatment effect on the treated (ATT). The issue of selection (nonresponse) bias will be addressed by two means: ex-post matching of treatment and control units to reduce model dependency / selection bias; and covariate adjustment of estimates to account for the fact that the distribution of explanatory variables will be different for the treatment and control samples.

The total sample size for the alternative (model-based) approach will take into account the following:

- All responding sample units (households) from the experimental design. This will include all treatment and control group farmers with baseline data from Cohort 2 and Fintrac's new recruits from Cohort 2 aldeas, as well as "other" non-program farmer households from a probability sample.
- An additional simple random sample of 600 program farmers (who entered the Fintrac program at approximately the same time as the Cohort 2 farmers)

In the original experimental design, aldeas were randomly assigned to treatment. For the additional sample of 600, this is not the case – the aldeas were selected by the program implementer (FINTRAC).

[3] For additional discussion of treatment effects, including intent-to-treat estimators, see *Counterfactuals and Causal Inference: Methods and Principles for Social Research* by Stephen L. Morgan and Christopher Winship (Cambridge University Press, 2007.

For the experimental design, the measure of impact was to be a "double difference" estimate of change: the difference, before and after the program intervention, of the difference in means of the treatment and control groups. The double difference estimator will still be used to estimate the average treatment effect (or intent to treat effect) using the design-based approach. For the model-based design, however, it is not possible to use this simple estimator as a measure of program impact since, without randomized assignment of treatment, the treatment and control groups may differ significantly with respect to variables ("covariates") that have an effect on outcome. Therefore, while the impact estimates for the experimental-design data is the usual double-difference estimate (a "design-based" estimate), for the impact estimate for the revised design is a covariate-adjusted single-difference estimate (a "model-based" estimate). This change in the form of the estimator will be discussed further later in this report.

NORC plans to use the data collected from the supplementary sample of 600 Fintrac clients to conduct one additional analysis. We will estimate income growth (between baseline and endline data collection) for these 600 farmers, using our calculations of income, and the household data collected by INE for the evaluation. We will then compare these income estimates to those of Fintrac for the same 600 farmers and analyze any differences that we observe. There are two questions of interest here.

(1) How do the "raw" (unadjusted) estimates of income change compare with Fintrac's estimates; and
(2) After performing appropriate matching and covariate adjustments, how do the two estimates (Fintrac's and this evaluation's) compare.

A.2 Data Sources.

The impact evaluation of the MCA Honduras projects as described above requires a significant data collection effort. The following data collections are currently underway:

Transportation Project

- Household survey – national, with heavier sampling near MCA Honduras intervention sites; using Census as the sampling frame
- Traffic surveys
- Price and product survey – data will be collected by visiting key markets near each sample cluster of the household survey
- Geolocations of health centers, schools, and market centers obtained or purchased through relevant government ministries or other organizations

- Data on roads conditions, such as International Roughness Index (IRI) measures – from existing sources such as SOPTRAVI and supervisory firms for the MCA Honduras Transportation Project, and the World Bank and Inter-American Development Bank CA-5 projects[4].

Rural Development Project

- Survey of farm households – program farmers and a sample of non-program farmers in treatment and control communities

Table 1: Details of Major Surveys

	Final sample size	Estimated Timing (Multiple Rounds
Household survey for Transportation Project (inclusive of Price-and-Products Survey)	1600 hholds	Feb-Apr 2008 Mar 2011
Household survey for FTDA – includes program farmers and other HH	3063 hholds	Feb-Mar 2008 Feb-Mar 2011
Traffic Survey	2000 vehicles	Nov 2008-Jan 2009 Dec 2010-Jan 2011

Notes:
(1) The baseline and follow-on household surveys for the Transportation Project and EDA will be conducted jointly.
(2) Three rounds of the Traffic Survey are being undertaken

The evaluation design requires the collection of household data for both the Transportation and the Rural Development Projects. We will use the same base instrument for both household surveys and rely on additional modules to collect project-specific data from different types of households – for example, there will be a detailed agricultural module with specific questions relevant to the EDA, which will be applied only to program farmers. This combined approach will render significant cost-savings in data collection, and possibly provide opportunities for looking at interactive effects for the two projects.

Use of Geographic Information System as an input into the evaluation model

We will also implement the use of a geographic information system (GIS) as an important tool to support the evaluation of both MCA Honduras projects. The GIS will be used for specific purposes beyond that of a simple spatial display tool of Compact elements. These general purposes are outlined in Box 1 below.

[4] We will use existing IRI measures to estimate IRI for other road segments through a regression model that uses other variables – such approximate road surface type, number of lanes, elevation, etc. – for which we do have data. IRI measures will be used as inputs into the HDM to estimate road maintenance and, if calculated by Soptravi and available, vehicle operating costs (VOC).

Box 1: Use of GIS in MCA Honduras Impact Evaluation

- ***To Model the Entire Honduran Road Network for Road Improvement Evaluation.*** Existing GIS Honduran road network data, continually updated and improved through primary data collection, will be used to calculate changes in travel time (for households and firms) that result from MCA road improvements. Travel-time calculations will be approximated within the GIS as a function of road quality, topography, and other factors that can influence travel time, such as bridges, landcover, and road surface type. Travel-time calculations will be calibrated by actual measurements. Because the GIS will be able to model the entire road network as a single network, network effects can be considered, and the relative change in travel time due to road improvements compared with household and firm economic changes over time.
- ***To Control for the Effects of Other Projects.*** The GIS can be used to calculate measures of accessibility to the location of other projects (for example other on-going infrastructure and road improvements). This in turn provides a method for statistically controlling for the affects of these projects, to disentangle their effects from MCA road improvement and program effects
- ***To Qualify Observations.*** We will use the GIS to qualify Honduran observations (households, firms, aldeas) with physiographic and spatial access variables. This will increase the power of our sampling in certain cases, through delineation of control and treatment groups, and by bringing into our statistical analyses important variables that would otherwise not be considered.
- ***To Predict Future Hypothetical Impacts and Outcomes.*** In conjunction with previously measured outcomes and statistical impact models in Honduras, such as the estimated coefficients between travel time and economic impact in the Transportation Project GIS model, the GIS can then be used to predict the likely influence or outcomes of future projects, or future investment, including spatial display of those outcomes.
- ***In Conjunction with Statistical Analysis.*** GIS can be used in conjunction with numerous statistical techniques. Variables calculated by the GIS can be exported directly to statistical processing software, and then predicted or residual values mapped in the GIS to examine spatial patterns in model residuals. Spatial statistical techniques can be used to identify spatial clusters and trends, and to stratify observations by spatial proximity or access (to infrastructure, major towns or markets, etc.).
- ***As a Spatial Display Tool.*** The GIS can be used to provide spatial display of MCC or MCA Honduras programs, or the location and extent of the influence of these programs. Access to all spatial displays and data (including viewing spatial data) can be setup to be done remotely through a web page or internet connection. Such displays could be considered when designing future MCC programs.

In the remainder of the report we proceed as follows: After this introduction (Section A), in Section B we provide an overview of the MCA Honduras Program. Next, in Sections C and D, we present detailed descriptions of evaluation designs for the two MCA projects being evaluated. For each project, we present our understanding of MCC's evaluation goals and hypotheses; key features, benefits, and challenges of the evaluation designs; and data collection requirements. Section E presents an implementation schedule for the evaluation design.

Annex I presents a detailed statistical explanation of how sample sizes were calculated for the original evaluation designs. Other Annexes that were included in the October 2007 Design Report – a focused review of studies related to this evaluation that was conducted in 2007, descriptions of a the geographic information system (GIS) to be used in the evaluation, the World Bank Highway Development and Management Model to be used in conjunction with the GIS – have been omitted from this revised report, but are available upon request.

B. MCA HONDURAS PROGRAM OVERVIEW

The Honduras MCA Program comprises two projects: (1) the Transportation Project, and (2) the Rural Development Project.

The **Rural Development Project** sought to increase the productivity and improve competitiveness of owners, operators, and employees of small- and medium-sized farms. Although Honduras enjoys a comparative advantage in horticulture given its rich growing conditions, year-long growing season, and proximity to the U.S. market, most farmers predominantly grow basic grains. They are constrained by several barriers to cultivating horticultural crops: the requirement of sophisticated techniques and infrastructure for production and marketing; lack of credit necessary to meet the higher working capital requirements of horticultural crops; and poor transportation infrastructure that increases the cost of getting crops to market and inputs to farm-gate. The MCA Honduras Program sought to alleviate these constraints and contribute to increased productivity among farmers through four activities:

(a) Farmer Training and Development Assistance (FTDA) - provision of technical assistance in the production and marketing of high-value horticultural crops.
(b) Farmer Access to Credit - provision of technical assistance to financial institutions, loans to such institutions and support in expanding the national lien registry system.
(c) Farm-to-Market Roads - construction and improvement of feeder roads to connect farms to markets.
(d) Agricultural Public Goods Grant Facility - provision of grants to fund agricultural "public goods" projects that the private sector cannot provide on its own.

Poor road conditions on the CA-5 Norte Highway and key secondary roads in Honduras resulted in poor accessibility to key production centers and markets and high transportation costs, which undermine the competitiveness of producers. The goals of the **Transportation Project** are to reduce transportation costs between targeted production centers and national, regional, and global markets by addressing problems stemming from high congestion, travel time, and overweight vehicles. The key activities of the project included:

(a) Highway CA-5 – Improvement of a 50 km segment of CA-5 between Taulabe and Comayagua, and a 59 km segment between Villa de San Antonio and Tegucigalpa.
(b) Secondary roads – Upgrade of key secondary routes to improve access of rural communities to markets

(c) Weight Control System – Construction of a weight control system and issuance of contracts to operate it effectively.[5]

As mentioned, for the purpose of the evaluation, the Farm-to-Market Roads activity is integrated into the evaluation of the Transportation Project.

The overall goal of the MCA Honduras Program, as stated in the Compact, was to advance the goal of economic growth and poverty reduction in Honduras. This goal is reiterated in the MCA Honduras M&E Plan, which also goes on to say "The Goal for the Compact is to alleviate poverty by increasing the income of Hondurans affected by the program (the Beneficiaries)." Hence, the primary focus of this evaluation will be to provide an independent assessment of the impact of the MCA Honduras Program on economic growth and income change, and its effects on poverty *through* economic growth. Assessment of a program's effect on poverty is concerned with the distributional effects on income. Hence, we consider it of primary importance to assess both the income growth and, whenever feasible, distributional impacts of the program. The impact of the MCA Program on other goals of the larger Millennium Challenge Program such as the environment, access to education and health, and gender equality will be considered, but will not be the main focus of the research activities under this contract.

[5] The Weight Control System activity was eliminated from the Compact in 2009, when MCC terminated a portion of the Compact.

C. TRANSPORTATION PROJECT

C.1 Evaluation Goals

The Transportation Project sought to improve conditions of the national road network by constructing and upgrading a number of secondary and rural roads, two segments of the CA-5 Highway and implementing a national weight control system. Improved conditions throughout the road network are expected to:

- Lower transport costs and travel time for businesses, including farm households;
- Provide better access to a wider range of job opportunities for individuals (labor market effects);
- Lower price of consumables and inputs by increasing competition and reducing barriers to entry posed by poor transport infrastructure;
- Improve access to health establishments and schools
- Reduce vehicle operating costs and road maintenance costs

The overall expected result of these changes is higher incomes and employment at the business and household level, as well as an increase in total factor productivity. We also hypothesize a possible increase in use of health facilities (improved health-seeking behavior) and school attendance.

Based on these hypotheses, we will estimate the following **benefit streams**, taking care not to double count them:

- Changes in household income and/or consumption
- Changes in household employment
- Changes in use of health facilities (non-dollar values)
- Changes in school attendance (non-dollar value)

C.2 Key Features of the Evaluation Design

Box 2: Key Features of the Evaluation Design

- Estimation of multiple benefit streams – captures widespread benefits of the road projects
- A single integrated network model for all three roads projects – captures network interrelationships
- Measures of incremental impact on a continuous scale (instead of a simple before-and-after model)
- Use of derived variables from external models – improves precision

The evaluation design we present below expands the scope of that outlined in the M&E Plan. As described above it estimates multiple benefit streams accruing to households. The evaluation design that we will use to estimate impact has two important features. We describe them below, along with their benefits.

A single, integrated network model that has greater efficiency and validity. The evaluation design presented here aims to evaluate the impact of all MCA Honduras roads projects (CA-5, secondary and rural road improvements) through a **single, integrated road network model** that more accurately represents the real-world system under study. This new approach, which represents the physical road network as an integrated computer/mathematical network (through the GIS system), recognizes, for example, that in reality, rural households are likely to benefit not only from rural-road improvements, but also from improvements to secondary roads. For example, for a farmer who must travel to a distant location to obtain fertilizer, improvements to his local rural roads may not reduce his travel-time cost nearly as much as secondary-road improvements might. The integrated model captures synergies and interaction effects between improvements made to different parts of the total road network, thereby allowing us to “disentangle” impacts of different road improvements. Furthermore, this unified approach enables assessment of economy-wide impacts of road improvements.

The integrated network model has the additional benefit of optimizing project resources: all resources that were planned for the three separate roads-project evaluation efforts can now be placed into the development of a single model. This will not only lead to a more valid representation of road-related phenomena (e.g., the interaction of road segments), but a substantial increase in the precision of model estimates and the statistical power of tests of hypotheses.

On the cost side, the survey sample size required to develop this integrated model for the three roads projects is about the same as that for a sample survey to develop a model for one sub-component project, such as rural roads. The reason for this counterintuitive fact is that in an evaluation study, sample size is not affected by the population size and, unlike simple “descriptive” surveys in which comparable sample sizes are required for each subpopulation of interest, all the data from the various projects will go into developing a single, integrated model.

Measuring of incremental impact The design will comprehensively evaluate the impact of road improvements by measuring before-and-after changes in the benefit stream variables (such as income and employment) for samples of households, businesses, and road users, relative to changes in travel time or travel cost (accessibility).

The approach will measure the impact of the roads projects on a continuous scale, enabling us to estimate varying levels of program impact at different treatment intensities through a mathematical (statistical) model that relates the incremental change in program impact (benefit-stream) variables to incremental changes of road improvements as reflected in travel time and/or travel cost. In technical terminology, the model is a path-analysis or structural-equation model, in which the relationship of the dependent variable (change in impact associated with program intervention) is represented in terms of explanatory variables derived from the GIS or other external models, such as travel time, travel cost, and accessibility.

The exact form of the analytical relationship between the outcome (response) variable and the explanatory variables – linear or nonlinear, including threshold levels, diminishing-returns phenomena, or a two-way table[6] – will be determined during the course of the data analysis. As a simple conceptual representation, the relationship may be presented as:

Change in outcome measure = f(change in travel time, covariates)

where f(.) denotes the functional relationship between change in the outcome measure and the change in travel time.

We consider this incremental-impact approach to be considerably better than a before-and-after or dichotomous (zero-one, treatment/no-treatment) model because it explicitly addresses the fact that the impact of road improvements varies as a continuum over users across space and geography (as a function of variation in travel-time accessibility to roads and markets) and by type of user (for example, commuter versus farmer versus business).

[6] Both linear and non-linear relationships will be statistically assessed for explanatory power.

Box 3: Accessibility Indicators and GIS

For estimating indicators of accessibility, such as travel time, and other spatial analysis relating to the impact of the road, the GIS should contain data such as:

- Existing road network spatial configurations, with approximate road speeds and conditions (determined from road maps, satellite imagery, and/or survey/census information)
- Elevation data
- Data on land quality (such as soil type, rainfall)
- Stream, river, lake and reservoirs
- Other infrastructure (power networks, railways, water access systems)
- Location of other features that are important for the calculation of welfare gains to the communities (such as hospitals, major market centers, government centers, health clinics)
- Data on the location and timing of construction of other road improvement projects

Once the GIS is assembled, access indicators can be constructed; example approaches include:

- *For households that are directly connected to the upgraded road segments by feeder roads, or that directly border the upgraded roads:* Travel time from each household to the nearest point on the infrastructure/road improvement section, or to other points of interest (markets, ports, etc.) calculating travel along connecting feeder roads, considering approximate road speed/road quality and other factors such as topography or traffic congestion.
- *For households that are not connected directly to the upgraded road segments by feeder roads:* A cost surface would be constructed for the relevant area, based on a combination of all GIS layers that provide data on factors that increase the cost/difficulty of travel (for example, steep slopes or unimproved roads). The GIS would then calculate the pathway from each unconnected community to the upgraded road, following the path of lowest travel cost, to determine the accessibility index value for that community.

As explained below, we will use a geographic information system (GIS) to estimate changes in accessibility (travel time, travel cost and access to services) and traffic flow along the entire road network as a function of measured road improvements resulting from the Transport Project. Change in travel time will be computed for every household using the GIS.

In its initial conception, in 2007, the evaluation design also called for using randomization in the selection of rural roads to be upgraded. By introducing this element of randomization we hoped to attribute causality of observed changes in outcome variables to the program interventions. However, in the implementation of the Transport Project, MCA Honduras determined that it was not feasible to employ randomization in the selection of the rural roads to be improved (treated) for two reasons: (1) one criterion that MCA added to the selection of eligible roads was that municipalities provide matching funding or in-kind contributions towards the road improvements; this adds a level of self-selection to the process that is not compatible with randomization; and (2) the number of kilometers rural roads to be improved by MCA declined dramatically from 1500 to 600 initially, and then to 495 towards the end of the Compact, resulting in insufficient sample for randomizing. Therefore, in place of randomization to identify comparison (control) roads, we used a matching procedure to identify a selection of untreated primary, secondary, and rural road segments that are similar to the segments being improved (treated segments). The matching process reduces the correlation of the program intervention (treatment) variable with other variables (known prior to conducting the survey) -- either variables that may have been related to the selection of project roads or variables that may be related to program impact. The matching process causes the distribution of these (explanatory) variables to be similar for both the treatment roads and the non-treatment roads. Reducing the correlation between the treatment and these other variables substantially improves the quality of the survey estimates (i.e., promotes increased precision and decreased bias). To preserve knowledge of the sample selection probabilities, the matching procedure will be implemented using stratified sampling (i.e., the stratification variables are the matching variables).

Use of existing external models (GIS and HDM) and derived variables

The evaluation model will use information and important variables derived from existing models such as the Geographic Information System (GIS) and the World Bank's Highway Development and Management (HDM) model. More detailed descriptions of the GIS (including methodologies for GIS travel-time calculations) and the HDM (including detailed description of selected engineering models) were included in the October 2007 Design Report and are available upon request.

The GIS will be used both as the repository of a detailed digital, geo-spatial road network database (which will include detailed and up-to-date data on primary, secondary and rural road networks, road surface types, and extensive physiographic data (elevation, land cover), and as an analysis/methodological tool to calculate detailed travel-time and accessibility measures for households, firms and road users across Honduras, allowing for the measurement of traffic-flow costs between any two locations. Travel-time measures will be calibrated and weighted by a sample of actual measurements in Honduras. Using this approach in conjunction with the GIS road network model will allow accurate derivation (with error estimations) of travel time for all road users in Honduras at far lower cost than on-the-ground measurement of this data. These travel-time estimates will be used in the Transportation Program evaluation mathematical model described in the previous section.

The Highway Development and Management (HDM) model, developed by the World Bank, has been used for over two decades to combine technical and economic appraisals of road projects for the purposes of estimating road maintenance costs, the application of rigorous engineering models quantifying the relationship between key road cost variables, such as Vehicle Operation Costs (VOC) and International Roughness Indices (IRI)), calculation of Net Present Value (NPV) and ERR for alternative road improvement options, and road department budget forecasting and planning. The HDM model includes a library of engineering models hand-picked from more than 40 years of road maintenance and traffic studies conducted around the world. They include models that:

- Calculate road deterioration as a function of vehicle weight, axle width and spacing, which in turn feed into estimations of road maintenance costs.
- Model pavement deterioration as a function of road roughness measures and pavement type and strength; and
- Traffic congestion models, which allow modification of travel times (and subsequent impacts on VOC and road deterioration) as a function of congestion cycles.

We will use HDM model libraries (including estimated model coefficients) to improve the GIS modeling and estimates of travel times (by vehicle type), traffic congestion, road roughness indices and vehicle weights.

We also foresee opportunities to use the GIS database (and models) in conjunction with the HDM model to generate network-wide outputs that will support the program evaluation. For example, we anticipate that we will be able to estimate HDM model outputs - such as IRI - for road segments that do not have complete on-the-ground measurements of these indicators, by using the GIS and other readily available road data such as approximate road surface type, number of lanes, elevation, and land cover. In these cases, IRI could be estimated for all primary and secondary road segments, by comparing observable characteristics of road segments for which IRI has been measured (elevation variation, road surface type, date of last repair, number of lanes, etc.) with the measured IRI to establish a correlation, perhaps using a regression model. This model could then be used to estimate IRI for other road segments in the GIS that have no IRI direct measures. This innovative integration of the HDM models in the GIS could substantially improve the evaluation of MCA Honduras' road projects.

C.3 Data Collection Requirements of the Evaluation Design

The evaluation model will be national in scope. Hence, all data collection will also be national in scope, with heavier sampling in the MCA-H project areas, since those areas will exhibit a wide range of variation in the quantities of interest (program-related changes in travel time, travel cost, accessibility to services).

Household Survey

The impact of the road improvements on household users will be estimated from data collected in a national household survey. The design for the household survey will be an "analytical" survey design, where the goal is to obtain a sample that is a good basis for estimating parameters of the analytical model representing the process under study. Briefly, this means that there should be substantial variation in the explanatory variables of the analytical model, and low correlation among them[7].

- ***Target population.*** The target population for the household survey is the population of all households in Honduras at the beginning and end of the project. We will use the sample frame constructed for the most recent national census. Since that census was conducted a number of years ago (2001), the survey field procedures will include processes to ensure that all current households are subject to sampling (e.g., use of systematic sampling over all of the current households of the entire census segment). The unit of analysis is the household.
- ***Variables of interest***. The variables of primary interest in the survey are household income and employment, but other information will be collected, including: access to and use of educational and health facilities; key socioeconomic and demographic indicators; consumption and expenditures; value of housing and land; and travel time and travel costs to points of interest will also be gathered. We will also include some questions on agricultural practices and products, similar to those that will be collected for the Rural Development Project, to enable an integrated analysis of both projects.
- ***Sample design and stratification.*** We are using a panel survey with a stratified two-stage sample design in which the first-stage sample units or "clusters" are caserios. Caserios were selected for use as the first-stage sample unit not only because they (like Census segments) are an efficient size for sampling, but also because a substantial amount of GIS data is available for them. This type of sample design (two-stage sampling) is widely used in surveys of households, because it affords a high return of precision for sampling effort expended.

[7] This type of survey differs from the usual sample survey designs, which are "descriptive" designs intended to estimate means of the target population or subpopulations of interest, or to compare means among subpopulations.

The design of an analytical survey must have variation in variables that have a substantial effect on impact (in this case, travel time). This variation is achieved through stratification of the sample. The variables of stratification (of the clusters) were estimated change in travel time to various points of interest (to be estimated using the GIS), urban/rural status and a number of geophysical variables (from GIS data sources). A GIS model was used to facilitate stratification by anticipated change in travel time.

Since we are using a panel survey design, the second survey of households will attempt to reinterview the same households that were interviewed in the initial survey. This promotes local control – the precision of estimates of change will be substantially higher for this longitudinal (time-series; panel) approach than for a “cross-sectional” approach of interviewing an independent sample of households the second time. Based on previous experience in Honduras, we anticipate that about 10-20 percent of the households will be different from those interviewed at the beginning of this project. In these few cases, the household currently occupying the dwelling will be interviewed on the rationale that its road-related behavior is probably similar to those of the previous occupants, and that the precision obtained by including it is probably greater than if it were dropped from the survey. Additionally, we will include several questions that attempt to identify why the previous occupants moved, and obtain location information for their new residence. If they have moved within the same census segment or *caserío*, interviewers will attempt to locate and reinterview them. To keep survey costs down and stay within time constraints, no attempt will be made to locate and interview displaced persons or households who no longer live in a primary sampling unit (cluster) of the sample.

- ***Sample size.*** In the initial data collection, a sample of 2,000 households located in 100 caserios was selected for the panel survey. Because some caserios did not contain 20 households, data was collected on a total of 1600 households. The final round of the household survey will be collected from these 1600 households.
- ***Survey instrument.*** The questionnaire is approximately 90 minutes in length. The head of the household, if available, will be the survey respondent. After two attempts have been made to locate and interview the head of the household, a proxy adult within the household may be selected to complete the survey. If another knowledgeable adult is not present and/or available, a replacement household will be interviewed.

We gathered and reviewed established household surveys – ENCOVI (Encuesta Nacional de Condiciones de Vida), the Honduran National Household Survey, and the Honduran National Census, among others – to use as models in developing our household survey instrument.

The instrument includes the following modules: labor and income; consumption and expenditures; travel times, cost, and access to services or destinations; housing and/or land costs/prices; socioeconomic scores (SES)/demographics. Also included are a set of questions on agriculture that overlap with the Farmer Survey for the Rural Development Project. Where appropriate, the instrument also contains a limited number of questions on respondents' views and perceptions of the effectiveness, sustainability, and unintended consequences of program activities (see Section E for further discussion of this).

Price-and-product survey

This survey, intended to shed light on the consumption effects of road improvements, will be conducted in tandem with the household survey. While field interviewers administer the household survey in sampled dwellings, supervisors will visit three local markets in or near each census segment of the household survey and complete a price-and-product data for a market-basket of goods. The survey instrument will take the form of a data collection form that lists products in a basic market basket of goods, as defined by the government of Honduras. The sample size is 100 markets per round.

Traffic surveys and data on roads

Traffic data –volumetric counts, origin-destination data, and speed measures - are an important input to any rigorous evaluation of road network usage, or for estimation and projection of future road maintenance or improvement costs. The traffic data will provide a "picture" of daily, weekly and annual congestion and usage flows that are essential to evaluate who is using the road, when and where are they traveling, and which sections of the road network will be receiving high usage.

The traffic data will also be used as an important calibration on the GIS, as GIS and HDM estimations of travel time (see descriptions above) will vary depending on whether or not the user (and vehicle type) is traveling during a "congested" or "uncongested" time. Thus, there will be at least two travel-time estimates produced from the GIS/HDM model calculations for any areas of the road network where travel could possibly face traffic congestion: an "uncongested" travel-time estimate, and a "congested" travel-time estimate. These values will not be solely a function of road surface type, number of lanes, International Road Roughness Index (IRI), vehicle type, or road condition, but will also reflect the relative congestion amounts per road segment. The traffic data are considered essential to accurate travel-time estimates.

There will be three data collection points for all roads – one baseline and two follow-ups. Trained data collectors will collect data – traffic counts, origin-destination, and speed data - during a complete week cycle from Sunday to Saturday. They will collect data at twelve-hour intervals each day and, for volumetric data, several twelve hour intervals at night (from 6 am to 6pm or 6pm to 6am in each case) to account for variation in traffic during peak and non-peak periods as well as possible variation in the types of vehicles traveling during different periods (heavy vehicles often travel more during the evening hours). To conduct origin-destination surveys, which include passenger counts, the data collector, working with the MCA-H and local authorities, will randomly stop vehicles and administer surveys. Data collectors will also collect information on vehicle type. Traffic count and other road-related data will be collected on all road projects (CA-5, secondary and tertiary roads), and a matched sample of comparison roads.

Since we are evaluating the road projects as part of an overall integrated national road system, it is crucial that we monitor efforts and collect data on all other road improvement projects in Honduras. NORC is making every effort to access such data from SOPTRAVI and other funding and implementing agencies, since these changes could well affect network-wide travel-time and accessibility estimates and improve the precision of the model parameter estimates.

GIS data

Honduras possesses exceptionally comprehensive and high-quality GIS data compared to most developing countries. GIS data generated in multiple Honduran government agencies – and also collected by private firms and in non-governmental research studies – have already been pooled centrally into a separate, stand-alone entity backed by a 12-year World Bank financial support program. This entity – *Honduran Programa de Administracion de Tierras de Honduras (“PATH”)* Digital Land Information System – has pooled into GIS databases, stored on powerful servers extensive and comprehensive Honduran GIS datasets including:

- National and local political boundaries (departmentos, muncipios, aldeas and caseríos);
- National census and survey data already linked to caserío, aldea and municipio locations (including aggregated data by political district);
- Extensive primary, secondary and rural Honduran road network data. Table 2 below presents details about the availability and sources, both PATH and other, of GIS data on the Honduran road network.
- Elevation;
- Hydrology (rivers, streams, lakes, watersheds, etc.);
- Extensive collections of high-resolution digital orthophotos and geo-rectified satellite imagery;
- Derived land cover;
- Extensive GIS datasets on climate, temperature, rainfall variation, humidity, and other factors. (PATH climate data includes sub-annual (e.g. monthly and daily) data). ;

- For a subset of major cities, extremely high-resolution GIS data providing the digital polygons of buildings, detailed street networks, electric and sewer networks, and household locations (linked to census data);

GIS specialists from the NORC team toured and evaluated the PATH GIS archive, and secured an agreement for full GIS data sharing and usage at no cost. Crucial to our evaluation of the extent and quality of these data was the evaluation of the GIS national road network, including primary, secondary and (most importantly and typically difficult to obtain) rural road networks, since our Transportation Program evaluation framework is premised on a reasonably complete GIS Honduran road network. We concluded that the GIS road network for Honduras is almost complete, including more than 70 percent of the rural road network. This is considered sufficient for the analysis we are carrying out[8]. Hence, extending the model from a local or regional scope to national scope requires no additional effort or cost with respect to the GIS data.

[8] For the relatively few communities or households in our analysis for which we do not have direct road network in the GIS, we can proxy travel times using a GIS travel-cost surface as a function of physiographic and elevation data – more details on these can be provided on request.

Table 2: Honduran GIS Road Network Data: Anticipated Sources and Extent

Road Data Description	Data Format and Extent	Anticipated and Current Data Sources
Primary Roads	Complete Honduran primary road network coverage available in GIS format	PATH (Programa de Administración de Tierras de Honduras); SOPTRAVI (Ministry of Transportation)
Secondary Roads	Complete Honduran secondary road network coverage available in GIS format	PATH (Programa de Administración de Tierras de Honduras); SOPTRAVI (Ministry of Transportation)
Rural/Tertiary Roads	Approximately 70 percent of rural road GIS coverage (obtained through cars with GPS receivers) available	SOPTRAVI (Ministry of Transportation)
Rural/Tertiary Roads	Approximately 70-80 percent of rural road GIS coverage available, obtained from GIS data, maps, private sector projects, etc.	CIES (Centro de Investigaciones Económicas y Sociales de la Universidad Jose Cecilio del Valle) and COHEP (Consejo Hondureño de la Empresa Privada) CIAT (Centro Internacional de Agricultura Tropical)
Location of Improved Primary, Secondary and Rural road segments	Improved road segments can be labeled in the GIS database, for display or analysis	MCA, WORLD BANK, Road Coordination Committee, SOPTRAVI
Periodic updates over time for changes to Primary, Secondary and Rural road locations or improvements, and to locations of improved road segment	Data can be entered periodically into the GIS database, and then display or analysis modified or re-run	PATH, SOPTRAVI, CIES, COHEP, Road Coordination Committee, SOPTRAVI, WORLD BANK, MCA, etc.

NOTE: NORC personnel in coordination with ESA Consultores conducted a field visit to Honduras in August, 2007, and identified the current extent and availability of Honduran GIS road network data.

Anticipated Sources and Status of Honduran GIS Road Network Data, as of September 2007. These data will be entered into the NORC Honduras GIS database, and will be updated periodically as additional data becomes available. Note that these data and updates will include not only the geo-location of Honduran road segments, but also additional data per road segment on approximate road surface/type, number of lanes, variation in elevation, and (where available) roughness index data. In addition, the GIS database will include information on locations of improved road segments, both MCC/MCA improvements and also others (World Bank, Honduran government, etc.).

D. RURAL DEVELOPMENT PROJECT: FARMER TRAINING AND DEVELOPMENT ASSISTANCE

Under this program component, FINTRAC provided farmers (Program Farmers) with small- and medium-sized farms with market-oriented technical skills, namely agronomy and small-business skills.

D.1 Evaluation Goals

Improved farmer training and access to credit is expected to:

- Increase cultivation of horticultural crops;
- Increase incomes of farm households; and
- Increase employment income on farms

Based on these hypotheses, we will estimate the following benefit streams[9]:

- Changes in household income (farm and off-farm) – net and gross
- Changes in farm income employment

D.2 Evaluation Design

The evaluation design for this activity has undergone a significant change since the inception of the project in 2007. As discussed above, the original experimental design data was supplemented with collection of data on an additional sample of Fintrac clients, and we now plan to conduct an analysis of both the original experimental design sample and the augmented sample. The statistical models used in these two approaches are substantially different. The original experimental design approach is referred to as a design-based approach, and the alternative approach is referred to as a model-based approach[10].

[9] These two benefit streams represent the two primary impact variables that the evaluation design will measure. Impact on contributing factors to these benefit streams, such as change in crop type and yield, may also be measured as part of the impact evaluation; however, in the interest of avoiding double-counting, we will not consider these as primary impact variables.

[10] For background information on these two approaches to evaluation and survey design, see the following references: (1) "History and Development of the Theoretical Foundation of Survey Based Estimation and Analysis," by J. N. K. Rao and D. R. Bellhouse, *Survey Methodology*, June 1990, Vol. 16, No. 1, pp. 3-29 Statistics Canada; (2) *Sampling: Design and Analysis* by Sharon L. Lohr (Duxbury Press, 1999); (3) *Sampling*, 2nd edition by Steven K. Thompson (Wiley, 2002); (4) *Practical Methods for Design and Analysis of Complex Surveys*, 2nd edition by Risto Lehtonen and Erkki Pahkinen (Wiley, 2004). (The Lohr book is the most informative.)

D.2.1 Experimental-Design Evaluation Approach

The **experimental-design** model developed in 2007 called for randomly allocating farming communities – in this case, aldeas – into two groups: those that receive technical assistance and credit now (the treatment communities) and those that receive it approximately 18 months later (control communities). Baseline and follow-up data collected from individual program farmers in these two groups would be used to assess the impact of program interventions on changes in several variables including income and farm employment.

Under this experimental design, the measure of impact would be the interaction effect of treatment and time, or the double-difference (or "difference-in-difference") estimate:

Change in benefit stream variable or estimate of impact
$= (Y_{T,t2} – Y_{T,t1}) - (Y_{C,t2} – Y_{C,t1}),$

where,
Y = benefit stream or impact variable
T = treatment group
C = control group
t1 = baseline or beginning of study
t2 = end of study

Since the experimental design is based on randomized selection of treatment and control communities, it is an accepted basis for making causal inferences from the collected data.

It is important to note here that the evaluation design calls for program treatment to be varied *among sample communities* rather than *among farmers within sample communities*[11]. (This type of design is sometimes referred to as a "cluster-randomized" design.) This design would have allowed us to make comparisons at two levels – the community level and the farmer level. In other words, the design would have enabled estimation of both the average treatment effect (ATE) and the average treatment effect on the treated (ATT). At the community level, we would have estimated aldea-level program impact by calculating the "double difference" estimate described above. At the farmer level, we would have calculated a similar estimate that compares program farmers of the treatment group to (eventual) program farmers in the control group.

[11] This approach substantially simplifies the operational demands of the evaluation on the program implementer, since each community is processed in its normal fashion (with no need to treat farmers or aldeas differently in the evaluation from normal program operation.). The decrease in "local control" that would have been afforded by varying treatments across farmers within the same community is compensated by stratifying communities that are similar with respect to characteristics considered important with respect to program outcome, and randomly assigning half of each stratum to the treatment and control groups.

However, NORC ran into a serious obstacle in implementing this evaluation design. Namely, despite repeated attempts, we could not replicate the Fintrac procedure for selecting farmers to receive treatment.

NORC used selection criteria and processes employed by Fintrac to select potential program farmers in treatment and control aldeas. However, over 90 percent of these farmers screened and selected for the evaluation were rejected by Fintrac as "ineligible." As such, after two separate screening efforts, each supported by Fintrac, and two rounds of full-blown data collection, NORC concluded that the Fintrac selection process could not be replicated, because it contains elements/criteria that cannot be quantified and depend on a subjective assessment by the Fintrac field technician of the farmer's motivation and willingness to follow program requirements, and also because the selection criteria kept changing and evolving over time, based on lessons learned during implementation. The result was a very small sample of treatment aldeas and farmers for the evaluation.

The experimental data, despite the small sample size, can still be used to construct an unbiased "intent-to-treat" estimate of the overall impact of the Fintrac program; however, with so few treated farmers present in the sample (only 28 from the original group of 300), the magnitude of this estimate will not be very large. The original statistical power analysis assumed that NORC would be able to replicate Fintrac's selection process and that the resultant sample size of farmers receiving program services would be substantial. With a large number of treated farmers in the sample, the probability (power) of detecting the specified minimum detectable impact would have been high; with a small number of treated farmers in the sample, the overall program effect in the sampled areas is small, and the probability of detecting it is also small. In summary, while the experimental-design sample can and will be used to construct an unbiased estimate of overall program effect (average treatment effect, or ATE), this effect is likely to be small and undetectable at a reasonable level of power (because of the low "density" of Fintrac-eligible farmers in the population sampled for the experimental design).

As a corollary to this analysis, we will also use the sample data to estimate what "density" (or number) of Fintrac-treated farmers would have been required for the average treatment effect of a specified size to have been detected with high probability (power), given the size and structure of the experimental design. We will also estimate how large the income change for treated farmers would have had to be to detect an ATE of specified size with high power.

D.2.2 Model-Based Evaluation Approach

When it became apparent that completion of the original experimental design would produce results of little value, MCA Honduras and MCC requested that NORC propose an alternative approach to completing the impact evaluation for the EDA activity. To this end, NORC proposed a model-based approach that will make use of whatever data that was collected for the original design, and complement it with additional sample data.

The additional sample data was collected from a random sample of 600 FINTRAC clients who entered the program around May/June 2009, which coincides with data collection for the second cohort of treatment and control farmers (Cohort 2). These additional sample units are intended to increase the sample size such that it is adequate to achieve a satisfactory level of precision and power. Since this sample was not determined by randomized allocation to treatment, estimates based on it may have selection bias. The magnitude of the selection bias will be reduced by two means: ex-post matching of treatment and control units, to reduce model dependency; and covariate adjustment of estimates to account for the fact that the distribution of explanatory variables may still be different for the treatment and control samples, even after ex-post matching. This approach maintains a considerable amount of the structure of the original experimental designs.

This alternative approach is called a "model-based" approach, as contrasted with the original "design-based" approach. With the design-based approach, unbiased estimates of program impact are determined by using the probabilities of selection of the sample units (e.g., in a Horvitz-Thompson estimate). With the model-based approach, the estimate of program impact is based on a statistical model that describes the relationship of treatment outcome to explanatory variables. (See the cited references, especially the Lohr book, for a detailed discussion of these two approaches.) Under these two approaches, the form of the impact estimate and the procedures for constructing it are quite different. For the design-based approach using an experimental design, the impact estimate can be represented (as discussed) as a simple double difference in means of the four design groups (treatment before, treatment after, control before, and control after). For the model-based approach, the estimate is more complicated, and is constructed using multiple regression analysis[12].

With the model-based approach, impact is represented as a single-difference covariate-adjusted estimator (as opposed to a double-difference estimator corresponding to the pretest-posttest experimental design). In its simplest form, this estimator may be represented as a function of a variety of explanatory variables.

Change in outcome measure = f(treatment indicator variable, covariates)

[12] For both approaches, the conceptual framework is the Neyman-Rubin causal model (the "potential outcomes framework" or "counterfactuals" model). For information on this approach, see *Mostly Harmless Econometrics* by Joshua D. Angrist and Jörn-Steffen Pischke (Princeton University Press, 2009); *Micro-Econometrics for Policy, Program, and Treatment Effects* by Myoung-Jae Lee (Oxford University Press, 2005); *Counterfactuals and Causal Inference: Methods and Principles for Social Research* by Stephen L. Morgan and Christopher Winship (Cambridge University Press, 2007); and *Causality: Models, Reasoning and Inference* by Judea Pearl (Cambridge University Press, 2000).

With the alternative approach, non-treatment variables may have different distributions for the treatment and control samples, and this difference may bias the estimate of program impact. Since the influence of all of these variables has not been removed by randomization, it will be necessary to adjust for them in the analytical model. As mentioned, this is done in a two-step process of ex-post matching (trimming, culling, pruning) and covariate adjustment[13]. The outcome variables of interest are referred to as response (dependent or explained) variables. Variables that have an effect on the response variables are referred to as explanatory (or independent) variables. The explanatory variables include both treatment variables and non-treatment variables, called covariates. The average measure of program impact is obtained by determining a regression-equation (or other) model showing the relationship of program outcome to explanatory variables, and substituting the mean values of these variables into the equation. The estimators that we will be using with this approach are called "model-assisted," "model-based" or "model-dependent."

When we depart from the experimental design, the simple double-difference estimator is not an adequate representation of the process under study, but with the full range of variables that we have collected for this study, we have strong reason to believe that we will be able (using ex-post matching and covariate adjustment) to develop good models of the relationship of program impact to explanatory variables. With this approach, it is not necessary to have a probability sample of the population under study – the model is assumed to apply to each unit of the population. What is important for estimation of the model is to have a sample in which there is a full range of variation in the explanatory variables of the model, and that the correlation among them is low. This is exactly what was done in the original sample design (see the project memo, *Sample Selection for MCA-Honduras Program Evaluation*, 30 January 2008).

The double-difference impact-estimation formula presented in section D1.2.1 is appropriate for the experimental design, but not for the modified design. In the experimental-design case, no covariate adjustment is necessary since, because of randomization the distributions of the treatment and control samples are the same. For the alternative FTDA evaluation design, it is necessary to modify the formula to reflect the covariate adjustment. In this case, the basic impact estimate (corresponding to the conceptual model presented earlier) takes the following form[14]:

$$\Delta y_t = \alpha + \beta\, x + \mathbf{z}_t'\, \boldsymbol{\gamma} + e_t$$

[13] (For information on matching, refer to Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (2007): 199-236 (also available from Internet website http://gking.harvard.edu/files/matchp.pdf .)

[14] For discussion of this model, see *Econometric Analysis of Cross Section and Panel Data* by Jeffrey M. Wooldridge (Massachusetts Institute of Technology Press, 2002).

where:

y_t = outcome variable (e.g., income) at time t
Δ = backward time difference operator (i.e., $\Delta y_t = y_t - y_{t-1}$)
α = mean effect
β = treatment effect
x = treatment indicator variable (= 1 for treatment households, 0 for control households)
$z' = (z_1, z_2, \ldots, z_m)$ = vector of covariates (explanatory variables other than the treatment indicator variable); some of these may be time differences (representing change in an explanatory between baseline and endline), while others (e.g., aldea, gender) may not be
$\gamma' = (\gamma_1, \gamma_2, \ldots, \gamma_m)$ = vector of covariate effects (regression coefficients)
e = model error term (model residual).

The difference operator is applied to the same household, at the two different times of the panel survey. The model parameters (α, β, and the γ's) are estimated using the general linear statistical model (regression analysis or similar approaches (such as Wald estimation)). The parameter β is the measure of program impact. (Note that this estimate is an estimate of the average treatment effect on the treated (ATT), whereas the double-difference estimate based on the experimental design is an estimate of the average treatement effect (ATE).) It is assumed that the model error terms are stochastically independent of the explanatory variables (x and the z's), and that the explanatory variables are measured without error. If these assumptions do not hold for a particular model (i.e., "endogeneity" is present), the model is respecified (e.g., using instrumental variables) so that they do hold, or alternative estimation procedures (e.g., Wald estimation) are used.

The model is referred to as a "single-difference" model or a "first-difference" model. The double-difference model (used for the experimental design) is applied when there were four distinct groups formed by randomization (treatment before, treatment after, control before, control after), in which case there is no need to include the covariates in the model, and the estimate of the β parameter turns out to be the difference in means between the Δy's for the treatment group and the Δy's for the control group, i.e., it is the double-difference estimator. (The regression coefficient, β, of the single-difference model takes the place of the second difference.)

In summary, the proposed alternative design is considered a valid approach that is feasible to implement for estimating the average treatment effect on the treated (ATT). It will make full use of the data already collected (i.e., both the data from the experimental design and the supplementary sample of Fintrac clients). It will make full use of the data on the large number of variables contained in the survey questionnaire.

The experimental-design data will be used to construct an estimate of the average treatment effect (ATE) for the population from which the experimental-design data were sampled. This estimate is expected to be small, because of the small number of Fintrac clients in the experimental-design sample. As discussed earlier, the ATE is always relative to a particular population of interest. An alternative approach to estimating the average treatment effect from the experimental-design data is to use the method of generalized regression (GREG) estimation or the method of synthetic (SYN) estimation (from the field of demography) to estimate the ATE for other populations of interest. In the synthetic-estimation approach, the ATE is estimated by multiplying the ATT (expressed as a rate of increase) for various demographic subgroups (estimated from the alternative model) times the population in the subgroup, and summing over all subgroups in the population of interest. The generalized-regression approach is similar, but a little more complicated. These are a model-based approaches to estimating the ATE, based on the model-based estimate of the ATT (conditional on demographic parameters) derived from the sample data. The generalized-regression and synthetic-estimation approaches overcome the shortcoming of the intent-to-treat estimate based on the experimental design, which (we now know), was not sufficiently "rich" in Fintrac-eligible farmers (i.e., it included a large proportion of farmers that Fintrac did not choose to treat). They may be applied to any populations of interest, such as all agricultural areas of Honduras (including those areas that were excluded from the experimental design sample because Fintrac had already worked there)[15].

D.3 Data Collection Requirements of Evaluation Design

- ***Target population.*** The target population is all farmers in the geographic areas into which FINTRAC expanded its program from 2007-2010. We used an area sampling plan, in which the primary sampling units are aldeas (these correspond to the communities used by FINTRAC in implementing its program). The sample frame was the same one used for the household survey - the sample frame from the 2001 Census (with field procedures to assure that new households are included in the sampling process). The unit of analysis is the household.
- ***Variables of interest.*** The primary variables of interest are farm household income (both farm and off-farm income), farm employment and wages, production types and yields, sales revenues, costs incurred (input costs, interest on loans, investments, depreciation, among others). In

[15] For more information on GREG and SYN estimates of ATE, see Practical Methods for Design and Analysis of Complex Surveys, 2nd edition by Risto Lehtonen and Erkki Pahkinen (Wiley, 2004)

addition, we will also collect data on demographics/SES such as information and consumption data. The questionnaire includes a component on sources, amounts, and uses of credit, which was intended to support the evaluation of the Farmer Access to Credit Activities. The instrument also contains a limited number of questions on respondents' views and perceptions of the effectiveness, sustainability, and unintended consequences of program activities.

- ***Sample design.*** We employ a panel survey (re-interview of the same households), with two waves (baseline data collection in 2009-2010, with follow up data collection in the spring of 2011). The sample design took the form of a two-stage design, with aldeas as the first-stage sample unit (cluster), and stratification of farmers within the cluster. The stratification is by program farmer and others in the treatment clusters. Program farmers are sampled with certainty, and the "other" stratum is sampled. In the control clusters, prospective program farmers are sampled with certainty, and the others are sampled with probabilities less than one.

- ***Sample Size***. Initially, we recommended that a panel survey of approximately 24 households per aldea (including program farmer households) in 200 aldeas be conducted – 4,800 for the baseline and 4,800 for the follow-up. Due to previously discussed problems with high rejections rates by Fintrac of farmers in the evaluation sample, and the fact that many aldeas contained fewer than 24 households, resulted in a final sample size that was much smaller than originally anticipated. Round 2 data collection occurred in a total of 3063 households from 344 aldeas. This includes 1765 households from 83 control aldeas and 1298 households from 261 treatment aldeas. Of these 261 aldeas, 212 are from the random sample of program farmers from FINTRAC's database. There are 545 program farmers in the augmented sample. The remaining 49 aldeas contain the treatment program farmers that were accepted by Fintrac as part of the evaluation. Of these 49 aldeas, only 28 are from the random selection of Cohort 2 aldeas; the other 21 comprise a subset of aldeas from which Fintrac selected program farmers, but were not part of the Cohort 2 list. There were a total of 124 program farmers in the 28 Cohort 2 aldeas.

- ***Survey Instrument.*** The household survey instrument used for FTDA evaluation is the same one used for the Transportation Project, with the addition of an agricultural module. The agricultural module includes questions on household income (both farm and non-farm), farm labor and wage bills, crop types and yields, and other variables necessary to calculate net incomes. We also include a series of questions on access to credit. The household surveys for the two projects will be combined and fielded as one large survey with the same timing for baseline and follow-up data collection.

E. QUALITATIVE ANALYSIS OF PROGRAM EFFECTS

Sections C and D of the Design Report focuses primarily on the approach and methodologies we will use to estimate the causal relationships between the Project and the Compact Goal. These methodologies will address the hypotheses for each program component and assess whether incomes of households, businesses, and farmers increased because of the MCA Honduras Program.

The original MCA Honduras also required assessing each program component – Transportation Project, Farmer Access to Credit activities, FTDA, and Public Goods Grants Facility to:

- Evaluate the efficiency and effectiveness of the Project Activities,
- Analyze the reasons why the Compact Goal was or was not achieved16,
- Identify positive and negative unintended results of the Program,
- Provide lessons learned that may be applied to similar projects, and
- Assess the likelihood that results will be sustained over time.

In the original design report, NORC propose to use qualitative data collection techniques such informant interviews, focus groups, and case studies to gather in-depth information on these issues. However, they were not mentioned in the most recent RFP, and at present we have no plans to evaluate the program using additional quantitative methods. NORC is nevertheless happy to discuss with the MCC the possibility of including them. Each qualitative technique and its potential use in evaluating the program are explained further below:

Key informant interviews

Key informant interviews provide in-depth information on most qualitative evaluation issues. We would rely on such in-depth interviews to gather information on how program goals and objectives were achieved or were not achieved to strengthen our understanding of the causal relationships behind the measured outcomes. If targets were not achieved, key informant interviews may help us to understand reasons for this; challenges faced and surmounted; unintended consequences (both positive and negative) of the program and how the negative effects might be avoided in the future; and lessons learned. If targets were met or exceeded, we would investigate how and why the program outperformed

[16] In the case of the Public Goods grants, we will attempt to explain any gaps between ex-ante and ex-post Economic Rates of Returns.

expectations. We would also use these interviews to get informants' detailed views on the perceived sustainability of the different program activities. A list of possible informants includes:

- Staff from MCA Honduras
- Implementing organizations (FINTRAC, ACDI-VOCA, Louis Berger, grantee organizations)
- FINTRAC field technicians
- Government officials from pertinent agencies such as SOPTRAVI, Ministry of Agriculture
- Banks (General Credit managers)
- Associations by sector (small farmers, transport users, banks, credit users)

Focus groups

We would rely on focus groups to gather information from program beneficiaries in greater detail than can be achieved in the surveys; focus groups allow us to interactively follow up with responses in a way that is not practical in a survey. We would conduct focus groups with community leaders (mayors, heads of aldeas, leaders of village organizations), household members (drawn from households that form part of the transport survey), program and beneficiary farmers, and beneficiaries of the public goods grants. Many of these focus groups could be conducted in conjunction with the second round of the household and farmer surveys, so that we could investigate more fully the reasons behind the outcomes observed.

F. IMPLEMENTATION SCHEDULE

Below we present the revised timeline for final data collection – please note that this timeline has not been approved by MCC as yet. It differs significantly from the workplan schedule in the proposal due to changes in the timing of the traffic and household data collections, and the expressed need for significantly more time for the impact analysis.

Work Plan Schedule

Dates	Task
Oct 1 – 31	Meet internally to take stock of developments and progress to date, discuss analysis plan and framework for the two evaluations, and agree on minor changes that are needed to the evaluation design and implementation plans. Deliverables: - Summary of changes/updates to evaluation designs (Oct 22) - Detailed workplan and timeline for completing both evaluations and data collections (Nov 5)
Nov 1 – 30	Travel to Honduras to meet with stakeholders from MCC and MCA Honduras to discuss the evaluation design, data collection activities, and analysis plan. Meet with INE and ESA Consultores to discuss detailed action plans for the follow-on household and traffic surveys. Prepare terms of reference for data collection activities. Deliverables: - Trip Report (Nov 19)
Nov 30 – Feb 23	Pre-fieldwork preparatory tasks for the household surveys: finalize TOR; finalize samples; review and improve, as needed, survey protocols and procedures for data collection; modify survey instruments, as needed; develop data quality control procedures for each data collection; recruit and train field supervisors and interviewers. Deliverables: - Terms of reference (Nov 15, Nov 22) - Survey planning documentation, FTDA HH Survey (Jan 14) - Survey planning documentation, Transport HH Survey (Jan 14)
Feb 16 – Mar 21	Field work — Field work for EDA HH Survey (Feb 16-Mar 9, 2011 — Field work for Transport Project HH Survey (Mar 2-21, 2011)

Dates	Task
Feb 24 – May 15	Data cleaning, processing and entry; prepare and submit clean datasets and final survey documentation — For EDA HH Survey (Feb 24- Apr 25) — For Transport Project HH Survey (Mar 9 – May 2) Deliverables: - Clean datasets with codebooks, FTDA (Apr 25, 2011) - Clean data set with codebook, Transport (May 2, 2011) - Final survey documentation, FTDA (May 5, 2011) - Final survey documentation, Transport (May 15, 2011)
Nov 15 – Dec 14	Pre-fieldwork preparatory tasks for the survey: finalize the sample; review and improve, as needed, survey protocols and procedures for data collection; modify survey instruments, as needed; develop data quality control procedures for data collection; recruit and train field supervisors and interviewers. Deliverables: - Terms of reference (Nov 22) - Survey planning documentation, Traffic Survey (Nov 30)
Dec 16 – Jan 28	Field work
Jan 20 – March 25	Data cleaning, processing and entry; prepare and submit clean datasets and final survey documentation Deliverables: - Clean data set with codebook, Traffic Survey (Mar 31) - Final survey documentation, Traffic (Apr 20)
Weekly	Weekly calls with COTR, accompanied by minutes Deliverables: - Meeting minutes (weekly)
Quarterly	Prepare and submit Quarterly Progress Reports Deliverables - Quarterly Progress Reports (Dec, Mar, June, Sep)

Dates	Task
Apr 1 – July 8	Data analysis - full statistical analysis of the impact-evaluation models, using the multiple rounds of data from each survey. Preparation of statistical tables, syntax files, and updated ERR calculations. Prepare draft evaluation reports. Deliverables: - Statistical tables from data analysis, FTDA (July 8) - Statistical tables, Transport HH & Traffic (July 8) - Syntax files, FTDA (July 8) - Syntax, Transport HH & Traffic (July 8) - Updated ERR calculations, FTDA (July 8) - Updated ERR calculations, Transport HH & Traffic (July 8)
Aug 11 – Sep 12	Finalize and submit two draft evaluation reports, engage review process, and prepare & submit Final Evaluation Report. Prepare PowerPoint slides for dissemination activities. Deliverables: - Detailed outline of Final evaluation Report (Mar 1) - Draft final evaluation report, FTDA (Aug 30) - Draft final evaluation report, Transport HH & Traffic (Aug 30) - Final evaluation Report, FTDA (Sep 12) Final evaluation Report (Sep 12)
Sep 15 – 30	Conduct 2-3 outreach sessions in Washington DC for MCC staff and other interested parties. Deliverables: - PowerPoint slides (Sep 30)

ANNEX I: SAMPLE SIZE & ALLOCATION

Transportation Project

Household Survey

The standard approach to determining sample size (for first-stage sample units ("clusters") and households within clusters) and allocation (to strata) are: (1) to specify a total budget for the survey and then configure the design to maximize precision of certain estimates or power of certain tests of hypothesis; or (2) to specify desired or required levels of precision (of certain estimates) or power (of certain tests of hypothesis), and configure the design to minimize cost. Since a survey budget has not been specified, but it is anticipated that it would be sufficient to fund a survey of several hundred clusters and several thousand households, the approach used in this case will be mainly the second one, with some iterations expected if the total cost becomes "too large."

To make sample-size estimates, information is needed about the relative cost of sampling clusters (caserios or aldeas) and elements (households) within clusters; about the variances of estimates of interest; about the intracluster correlation coefficient for estimates of interest; and about the intraclass correlation coefficient of strata for estimates of interest. Information is known from previous similar surveys about sampling costs but, as noted earlier, this evaluation design is a new one, and little information is known about estimate variances or the intracluster or intrastratum correlation coefficients for the variables of primary interest (e.g., estimates of change in economic impact as a function of travel time). There is certainly *some* prior information available on variability, however, since the proposed survey will in fact include many of the same variables that have been included in previous surveys – just not on the primary phenomenon of interest (the relationship of change in impact (income, employment, access) to program interventions or its surrogate (latent / endogenous variable), change in travel time.

To assist the survey design, a statistical analysis was conducted to estimate the value of the intracluster correlation coefficient (icc) for a selection of about a dozen variables of the 2001 Honduran Census, and for a general measure of socioeconomic well-being (households lacking three or more basic necessities ("Necesidades Básicas Insatisfechas" (NBI)). The analysis was conducted using formulas presented by Kish (*Survey Sampling*, Wiley, 1965) and using the Stata statistical analysis program. The intracluster correlation coefficient was estimated for census segments,aldeas and municipios. The results are as follows:

Conglomerado	Rho (formula Kish)	Rho (Stata)
Segmento censal	0.1949	0.1941
Aldea	0.1184	0.1471
Municipio	0.0599	0.0804

It is noted that the icc's presented in the table are for a single composite variable, NBI, and that the icc varies depending on the variable and cluster definition. For the selection of other (raw) variables taken from the Census (e.g., presence or absence of a refrigerator, presence of a specified type of water, attainment of 4th grade education), the icc varied for census segments varied from close to zero to as high as .8. The value for NBI, about .2, is fairly typical, and, in the absence of icc estimates for the variables of primary interest in the present survey (change in impact measures associated with program interventions (as reflected in change in travel time) over two years), it will be used to suggest reasonable sample sizes. (The results presented in the table vary a little by estimation type (Kish, Stata) because of different estimation approaches. The Kish formulas assume a fixed cluster size, which is true for segments but not so for aldeas and municipios, and so they were calculated for a subsample of clusters of similar size.)

Because substantial GIS data were available for caserios, and they were a convenient size for use as first-stage sample units, it was decided to use them as the first-stage sample unit for the transportation project survey. Caserios are somewhat larger (in terms of population) than Census segments – slightly over 200 people each vs. less than 100 for Census segments – and smaller than aldeas. It will be conservatively estimated that the icc for caserios is the same as for Census segments.

It is expected that an interviewer could conduct two to four household interviews (lasting about an hour) per day, once present in the caserio. In this case, the ratio of cluster sampling cost to household-within-cluster sampling cost varies from approximately 10:1 to 100:1, depending on how long the questionnaire is. If travel costs between clusters are not very large, then the following formula specifies the "optimal" within-cluster sample size, as a function of the sampling cost ratio and the icc:

$$m_{opt} = \frac{S_2}{\sqrt{S_1^2 - S_2^2 / M}} \sqrt{c_1 / c_2} \approx \sqrt{\frac{c_1(1-icc)}{c_2 icc}}$$

where

S_1^2 = variance among primary (cluster) means
S_2^2 = variance among subunits (households) within primary units
M = cluster size (number of households per cluster)
n = number of clusters in sample

m = number of households sampled per cluster
c_1 = variable cost of sampling per cluster
c_2 = variable cost of sampling per household
C = total variable sampling cost = $c_1 n + c_2 nm$
icc = intracluster correlation coefficient.

Substituting $c_1/c_2 = 10$ and icc = .2, we obtain $m_{opt} = 6$. If $c_1/c_2 = 100$, then $m_{opt} = 20$. The preceding estimates are based on a number of assumptions, and the results vary according to the variable (since the icc varies according to the variable). Prior survey experience in Honduras suggests that the value of m = 20 is a reasonable one for the within-cluster sample size, and that is what is proposed for the present survey.

It remains to specify the number of clusters to select. As in the case of determining a reasonable intracluster sample size, prior information can, along with a number of assumptions, suggest a reasonable value or range of values. Since this evaluation design is unlike any other, however, it should be recognized that the sample size estimates that follow are simply rough guidelines, making use of best available prior information.

The objective of the evaluation research design is to provide estimates of adequate precision for the relationship of change in impact (income, employment, access) to program interventions, as reflected in change in travel time. There are two standard approaches to sample-size estimation, specifying either the precision of an estimate (e.g., by specifying the width of a confidence interval) or the statistical power of a test of hypothesis. The present study is more concerned with estimation rather than tests of hypotheses (e.g., determining whether results are different for different subpopulations, or at different times (e.g., before-and-after an intervention)), and so the former method will be emphasized. Sample sizes will be estimated, however, using both methods (since the survey data will be used both to make estimates of means and to conduct tests of hypotheses (e.g., about differences among subpopulations, such as comparisons by gender, level of education, urban/rural status, sector, and region).

Since we have no prior information about the variance of the estimates of primary interest (change in impact as a function of change in travel time), we shall limit consideration to estimation of proportions, for which the variance is a function of the mean. It is recognized that the estimates of primary interest in this project are *not* proportions, but if the survey is designed to efficiently produce estimates of adequate precision for proportions for a variety of socioeconomic variables in the population of interest, it is reasonable to expect that it would to provide adequate precision for a the socioeconomic variables of interest in this evaluation. This cannot be affirmed with certainty, but it is the best that we can do with the information that is already available, without undertaking a costly and time-consuming preliminary ("first-phase") survey to collect preliminary data which would enable a better full-scale survey design to be constructed. In any event, the planned survey will collect data on many socioeconomic variables that are similar to those collected in the Census, and the results presented here will certainly pertain quite well to those variables.

The following are the standard formulas used to make sample-size estimates.

Estimation of Sample Size Based on Specification of Precision

The formula for the half-width of a confidence-interval estimate of a population mean (assuming sample sizes sufficiently large that the central limit theorem applies, which is the case in the present application) is:

$$E = z_{1-\alpha/2}\sigma\sqrt{\frac{(1-n/N)deff}{n}}$$

where

E = half-width of confidence interval
1-α = confidence coefficient (e.g., .95)
$z_{1-\alpha/2}$ = 1-α/2 percentile point of normal distribution (e.g., 1.96 for α=.05)
N = population size
deff = Kish's design effect
n = sample size.

The confidence coefficient is often characterized in terms of its ones-complement, denoted by α. The parameter α corresponds to the probability of a Type I error in a test of hypothesis (i.e., the probability of rejecting the null hypothesis when it is true). Here, it is the probability that the confidence interval does *not* contain the true value (of the parameter being estimated).

In the present application, N may be assumed very large (finite-population corrections are not relevant in evaluation studies, since the goal is to make inferences about a process, not a particular finite population). The design effect, "deff," is the ratio of the variance of the mean using the specified complex design (involving stratified cluster sampling) and the variance of the mean using simple random sampling with the same sample size. The value of deff in the case of cluster sampling (ignoring stratification) is:

$$deff = 1 + (m-1)\rho,$$

where ρ denotes the intracluster correlation coefficient. Substituting ρ = .2 and m = 20 we obtain deff = 5, and sqrt(deff) = 2.2. This means that, because of the loss of precision due to clustering, the sample size for the proposed survey is (for variables for which ρ = .2) about twice as large as for a simple random sample of the same size. (This loss in precision is compensated for by the cost saving afforded by cluster sampling.)

The corresponding formula for the sample size is:

$$n = \frac{z_{1-\alpha/2}{}^2 \sigma^2 deff}{E^2 + z_{1-\alpha/2}{}^2 \sigma^2 deff / N}$$

As it stands, this formula applies to the case in which we are estimating the mean of a population. This applies to some estimates of interest in this study, but the primary estimates of interest – estimates of the change in economic impact as a function of change in travel time – are similar in precision to differences (actually, to double differences) between subpopulations, not to means. In simple random sampling the variance of an estimated difference in means is four times the variance of the mean (using the same total sample size), and the variance of an estimated double difference is 16 times the variance of the estimated mean.

To enhance precision, the proposed survey design is a panel survey. A panel survey is proposed since it generally increases the precision of estimates of differences (but at the cost of decreasing precision for estimates of means or totals). If r denotes the correlation of a variable for the same household at the two survey times, then the variance of an estimated difference of means between the two survey times is reduced by the factor 1-r (whereas the variance of the mean is increased by the factor 1+r).

As a guide to estimating sample size, we shall consider the case of estimating double differences, and apply the formula given above with σ^2 replaced by $16(1\text{-}r)\sigma^2$ (or, equivalently, with deff replaced by 16(1-r)deff). Since we have no prior information about the value of r, we can only guess at it – we shall assume the value of .5, in which case the value of $16(1\text{-}r)\sigma^2$ is $8\sigma^2$ (or 16(1-r)deff is 8deff).

In sampling for proportions, the variance σ^2 is equal to p(1-p), where p denotes the true value of the proportion. This quantity assumes a maximum value, .25, for p=.5. If we set α=.05 (a standard value), then $z_{1-\alpha/2}$=1.96. From above, we determined that a reasonable estimate of deff is 5. For these values, let us consider three cases, in which E (the half-width of a confidence interval) is set equal to .05, .1, and .2 (i.e., to ten percent of the mean (p=.5), twenty percent of the mean, and 40 percent of the mean). The value of N is not relevant here (i.e., may be assumed very large). In summary, the following values are assumed for the parameters of the sample-size formula:

p = .5 (corresponding to $\sigma^2 = .25$)
deff = 8(5) = 40
$\alpha = .05$ (corresponding to $z_{1-\alpha/2} = 1.96$).

In this case, the formula for the estimated sample size is:

$$n = (1.96)^2(.25)(8)(5)/E^2 = 38.416/E^2$$

For the three values of E, the corresponding values of n are 15,366, 3,842, and 960. These are the sample sizes of households. If 20 households are selected from each cluster, then the number of clusters to be sampled is 1/20 of these amounts, or 768, 192 and 48. (Note that these sample sizes are the total for both waves of the panel survey.)

The following table summarizes the preceding discussion.

Estimate of Household-Survey Sample Size (Roads Evaluation) Based on Specification of Precision

Half-Width of 95% Confidence Interval, as a Fraction of the Value of the Quantity Being Estimated	Estimated Sample Size (based on 20 households per cluster)	
	Households	Clusters
.1	15,366	768
.2	3,842	192
.4	960	48

In the preceding discussion, the value of the confidence coefficient (CC, or 1 – α) was set equal to .95 (α = .05). The sample size could be calculated for other confidence coefficients, such as .99, .90 and .80. The value .95, which corresponds to a confidence-interval half-width of approximately two standard errors, is widely used, and so sample-size estimates will not be presented here for other values.

Estimation of Sample Size Based on Specification of Statistical Power

If we consider the power as the basis for estimating sample size, the results are as follows.

The formula for the power of a statistical test of an hypothesis (i.e., the probability of rejecting a null hypothesis) is:

$$\Pr\left(\frac{\hat{\mu}_1 - \hat{\mu}_2}{[deff\,(\sigma_1{}^2/n_1 + \sigma_2{}^2/n_2 - 2\rho(\sigma_1/\sqrt{n_1})(\sigma_2/\sqrt{n_2}))]^{1/2}}\right) > z_{1-\alpha} \middle| \mu_1 - \mu_2 = D) = 1 - \beta$$

where

μ_1 = mean for group 1
μ_2 = mean for group 2
n_1 = sample size for group 1
n_2 = sample size for group 2
σ_1 = standard deviation for group 1
σ_2 = standard deviation for group 2
ρ = correlation between the group estimated means
α = significance level of *one-sided* test of hypothesis of equality of group means (the probability of Type I error, i.e., the probability of rejecting the hypothesis of equality of group means, when it is in fact true) (e.g., .05)
β = the probability of making a Type II error, i.e., the probability of accepting the hypothesis of equality of the group means, when it is in fact false) (e.g., .1)
$1 - \beta$ = power of the test (e.g., .9)
$z_{1-\alpha}$ = 1-α percentile point of normal distribution (e.g., 1.6449 for α=.05, or 1.2816 for α=.1)
deff = design effect (assumed the same for both groups)
D = (true) size of difference between group means
and a caret (ˆ) over a parameter (symbol) denotes a sample estimate.

(Note that α refers to the significance level of a *one-sided* test of hypothesis. This corresponds to the case in which there is little doubt about which of the two group means will be larger. For situations in which it is not known which of the two group means might be larger, a *two-sided* test would be appropriate. In that case, the value $z_{1-\alpha}$ in the preceding formula (and the one given below) should be replaced by $z_{1-\alpha/2}$. Whether to use a one-sided test or a two-sided test depends on the circumstances. Assuming a one-sided test corresponds to a smaller value of z (i.e., $z_{1-\alpha}$ is less than $z_{1-\alpha/2}$), and to a smaller sample size estimate (a smaller sample is required since there is some *a priori* information about which group mean might be larger).)

The corresponding formula for the sample size is:

$$n_1 = \frac{deff\,(z_{1-\alpha} + z_{1-\beta})^2 (\sigma_1^{\,2} + (1/ratio)\sigma_2^{\,2} - 2\rho\sigma_1\sigma_2/\sqrt{ratio})}{D^2}$$

where
ratio = ratio of the group sizes = n_2/n_1 (i.e., n_2 = ratio x n_1).

As was discussed above (in the case of estimating sample size based on specification of precision), this formula applies to estimation of means (or proportions). While some of the quantities of interest in the survey will be means or proportions, the estimates of primary interest (change in impact as a function of change in travel time) are similar to differences or double differences. Applying the same reasoning as before, the formula for estimating sample size in this case is the above formula, but with deff replaced by 8deff.

In estimating sample size based on power calculations, it is important to realize that the power refers to the probability of detecting differences in the double-difference estimator (the estimate of program impact, such as the change in income as a function of change in travel time). The power for detecting single differences, such as a comparison of income for two groups or comparing income for the same group before and after the program, would be much higher. (For a comparable level of power, the sample size for comparing single differences is about four times that required to compare group means, and the sample size for comparing double differences is about 16 times that required to compare group means (ignoring correlations) – precise measurement of double differences, or high power for comparing double-differences, requires large sample sizes!)

In the present application, we shall assume the following values for the parameters of the above formula:

> deff = 8(5) = 40
> $p_1 = .45$ (so $\sigma_1^2 = .2475$ and $\sigma = .4975$)
> $p_2 = .55$ (so $\sigma_2^2 = .2475$ and $\sigma = .4975$)
> $\alpha = .05$ (so $z_{1-\alpha} = 1.6449$)
> $\beta = .1$ (so power = .9, and $z_{1-\beta} = 1.2816$)
> $\rho = .5$
> ratio = 1.

With these values, the formula for n_1, as a function of D, is:

$$n_1 = 84.78758228/D^2 .$$

Let us denote the total sample size (of both groups) as $n = n_1 + n_2$.

If we set D=.1, .2 and .5 (corresponding to differences equal to 20%, 40% and 100% of the mean (p=.5), respectively), we obtain the following values for n: 16,958, 4,239, and 678. These are household sample sizes, so that, with 20 households selected from each cluster, the corresponding cluster sample sizes are 851, 212, and 34. (Note that these are the total number of households and clusters for both waves of the panel survey.)

The following table summarizes the preceding discussion.

Estimate of Household-Survey Sample Size (Roads Evaluation) Based on Specification of Power (.9)

Size of Difference to Detect, as a Fraction of the Value of the Quantity Being Estimated	Estimated Sample Size (based on 20 households per cluster)	
	Households	Clusters
.2	16,958	848
.4	4,240	212
1.0	678	34

In the preceding discussion, a single value was assumed for the power, i.e., 1 – β = .9. The preceding table shows sample size estimates for detecting differences of varying size, corresponding to this value of the power. Here follows a table that presents estimated sample size for varying values of the power, as well as varying values of the specified size of the difference to be detected. This table is based on a single value for the probability of Type 1 error (significance level of the test), i.e., α = .05.

Estimate of Household-Survey Sample Size (Roads Evaluation) Based on Specification of Power

Power (1-β)	Size of Difference to Detect, as a Fraction of the Value of the Quantity Being Estimated	Estimated Sample Size (based on 20 households per cluster)	
		Households	Clusters
.95	.2	21,430	1,072
	.4	5,358	268
	1.0	858	44
.90	.2	16,958	848
	.4	4,240	212
	1.0	678	34
.85	.2	14,236	712
	.4	3,560	178
	1.0	570	28

Based on the preceding considerations (estimating sample size by specifying precision and power), the size of the NORC evaluation contract, and the perceived level of funds available for data collection, it is viewed that a **minimum acceptable sample size** for the household survey is about 4,000 households, selected from 200 clusters. Since this is a panel survey, the same households are interviewed in both survey waves, each wave involving 2,000 interviews in 100 clusters. (For costing purposes, it does not matter that the same households and clusters are sampled in the second wave – the cost is the same as if they were sampled independently.)

Based on the variety of cases examined, this is judged to be about as small a sample size possible to produce adequate precision for the policy objectives of the study. This survey will also include visits to 400 markets to collect price and product data (200 at the beginning of the study and the same 200 at the end). (The precision of these price-and-product estimates, based on a cluster sample with two elements (markets) sampled per cluster, have not been estimated.) If it is desired to track households that were relocated because of the roads projects, then it is recommended that all of them be included (there are several hundred of them). (It has not been decided whether they should be interviewed once or twice, but once seems adequate.) In addition, the second-round survey will include interviews of a certain number of households that have moved from their locations at the time of the first survey.

It is recognized that the preceding sample-size estimates are based on a number of assumptions, made in the face of limited prior information. The problem that we are facing here is that the proposed design (i.e., the "path analysis" model based on primary (sample-survey) and secondary (GIS, HDM) data sources) is rather different from most approaches (based on analysis of primary data from sample surveys). After this study has been completed, the next one to use this approach will be on much firmer ground with respect to estimation of sample sizes.

Sample Allocation

"Sample allocation" refers to allocation of the cluster sample to the proposed strata of the survey design. It is recommended that a "balanced" two-way stratification be employed, in which equal numbers of sample units (clusters) all allocated to each stratum. If, for example, it is decided to use five categories of anticipated change in travel time and two socioeconomic categories (urban / rural), then the sample allocation would be as follows:

		Rural/urban status	
		Rural	**Urban**
Travel-time change category	0-20%	20	20
	20-40%	20	20
	40-60%	20	20
	60-80%	20	20
	80-100%	20	20

Rural Development Project

Household and Program Farmer Survey

In the case of the rural development project, the experimental model is much simpler than in the case of the transportation project. The basic model involves comparison of income changes before and after the program, between the program farmers and a control group. The basic estimator is the "double difference" estimate: the difference, between the program and control groups, in the change in income before and after the program. The main variable of interest is income, and some information about income is known from the annual national household survey. A statistical analysis was conducted of rural incomes and determined that, for rural households, the mean income was 4,617 (lempiras) and the standard deviation was 10,857. The intracluster correlation coefficient was not estimated in that analysis, and so the value observed in the last census for aldeas will be used (i.e., the value of .15, reported above). If we select a sample of 20 households from each cluster, then the design effect is deff = 1 + (m-1)ρ = 1 + (20-1)(.15) = 3.85. (The within-cluster sample size of 20 is the same as for the household survey, since the sampling costs and the intracluster correlation coefficient are similar.)

In these calculations, we shall assume that the standard deviation is proportional to the mean.

From the point of view of estimating sample size from a precision requirement, we shall determine the sample size required to estimate mean income within one-hundred percent (i.e., specify the half-width of a 95% confidence interval to be equal to the mean). (It has been represented that the FINTRAC program causes very large increases in income, e.g., a doubling.) From the point of view of estimating sample size from a power specification, we shall determine the sample size required to detect a doubling of income. We shall then estimate sample size based on specification of precision and power for the basic quantity of interest in the rural development study, viz., the "double difference" estimator.

From the point of view of specifying precision as a basis for estimating sample size, we have the following values for the parameters of the sample-size estimation equation (presented earlier, in the section on roads):

$\mu = 4{,}617$
$\sigma = 10{,}857$
deff = 3.85
$\alpha = .05$ (corresponding to $z_{1-\alpha/2} = 1.96$).

As discussed earlier, the sample-size formula presented is for estimation of means, and for estimation of a double difference using a panel survey, this sample size must be multiplied by 16(1-ρ), where ρ denotes the correlation coefficient of reinterviewed households. If we assume that ρ=.5, then this factor is 16(1-.5) = 8.

Substituting these in the formula (ignoring N), we obtain:

$$n = (1.96^2)(10{,}857^2)(8)(3.85)/E^2 = 13947055690/E^2 .$$

We shall consider three values for E, corresponding to 100 percent of the mean, 50 percent of the mean, and 20 percent of the mean. These values (4,617, 2308.5, and 923.4) produce the following values for n: 654, 2,617, and 16,357. Since 20 households are selected from each cluster (PSU), the corresponding cluster sizes are 33, 131, and 818.

The following table summarizes the preceding discussion.

Estimate of Household-Survey Sample Size (FTDA Evaluation) Based on Specification of Precision

Half-Width of 95% Confidence Interval, as a Fraction of the Value of the Quantity Being Estimated	Estimated Sample Size (based on 20 households per cluster)	
	Households	Clusters
.2	16,358	818
.5	2,618	132
1.0	654	34

For estimating sample size based on specification of power of a test of hypothesis, we use the following parameter values:

$\mu_1 = 4{,}617$
$\mu_2 = 9{,}234$
$\sigma_1 = 10{,}857$
$\sigma_2 = 21{,}714$
ratio = 1
deff = 3.85
$\rho = .5$
$\alpha = .05$ (corresponding to $z_{1-\alpha} = 1.6449$)
$\beta = .1$ (power = .9, corresponding to $z_{1-\beta} = 1.2816$).

For specifying power of a test of the double-difference estimate (the interaction of treatment and time), the required sample size is eight times the result given by the formula for testing of means (i.e., replace deff by 8deff).

With these values, the formula for n_1, as a function of D, is:

$$n_1 = (9.328 \times 10^{10})/D^2 .$$

If we set D = 2,308.5, 4,617 and 9,234 (corresponding to differences equal to .5, 1.0, and 2.0 times the mean, respectively), we obtain the following values for n (the total sample size, = $n_1 + n_2$): 35,007, 8,752, and 2,188. These are household sample sizes, so that, with 20 households selected from each cluster, the corresponding cluster sample sizes are 1,760, 219 and 109. (Note that these are the total number of sample households and clusters for both waves of the panel survey combined.)

The following table summarizes the preceding discussion.

Estimate of Household-Survey Sample Size (FTDA Evaluation) Based on Specification of Power (.9)

Size of Difference to Detect, as a Fraction of the Value of the Quantity Being Estimated	Estimated Sample Size (based on 20 households per cluster)	
	Households	Clusters
.5	35,008	1,750
1.0	8,752	438
2.0	2,188	110

From this table we see that very large samples are required to be able to detect modest differences in income impact (as measured by the double difference estimator) with high power. This is because the standard deviation of income is very large. It is understood that the FINTRAC program results in very large income changes (e.g., doubling or more), and hence very large sample sizes are not required to detect the changes that are likely to occur. Note that the power calculations are for the double difference, not for a single difference (such as a comparison of income of two groups, or a comparison of one group at two different times).

In the preceding discussion, a single value was assumed for the power, i.e., $1 - \beta = .9$. The preceding table shows sample size estimates for detecting differences of varying size, corresponding to this value of the power. Here follows a table that presents estimated sample size for varying values of the power, as well as varying values of the specified size of the difference to be detected. This table is based on a single value for the probability of Type 1 error (significance level of the test), i.e., $\alpha = .05$.

Estimate of Household-Survey Sample Size (FTDA Evaluation) Based on Specification of Power

Power (1-β)	Size of Difference to Detect, as a Fraction of the Value of the Quantity Being Estimated	Estimated Sample Size (based on 20 households per cluster)	
		Households	Clusters
.95	.5	44,238	2,212
	1.0	11,060	554
	2.0	2,764	138
.90	.5	35,008	1,750
	1.0	8,752	438
	2.0	2,188	110
.85	.5	29,386	1,470
	1.0	7,346	368
	2.0	1,836	92

Based on the preceding considerations, we recommend the following, **at a minimum**: a sample of 200 aldeas, with selection of a random sample of 20 households selected from each aldea (in addition to certainty sampling of lead and beneficiary farmers in the treatment aldeas and certainty sampling of prospective lead farmers in control aldeas). (The indicated sample size is for both waves of the panel survey combined.) It is noted that with this sample size, it is possible to detect substantial differences in impact (as measured by the double difference), but not small differences. If it is desired to have increased power for comparing small differences, either the sample size should be increased or a stratification should be employed such that the within-stratum variance of income is substantially less than for the general (rural) population.

The household sample size ought to be somewhat over 4,000, because of the certainty sampling of prospective leads and beneficiaries.

As in the transportation household survey, it is expected that the panel survey may lose about 10 percent of the households each year. If a family that was interviewed in the baseline has moved to a new location in the aldea, it will be reinterviewed at its new home.

Aldeas vary greatly in size, and so, to increase precision, the sample will be stratified by aldea size. In addition, it is expected that program effects may vary substantially by agricultural region, and so the sample will also be stratified by agricultural region.

Sample Size Modifications Resulting from Adoption of Model-Based Approach

In proposing the additional model-based approach, NORC recommended that a random sample be selected of at least 500 program farmers who entered the FTDA program in the spring of 2009. The MCC ultimately agreed to a random sample of 600 program farmers.

This sample size is recommended even to detect an impact of a doubling of income caused by the program intervention. At first glance, it would appear that this sample size (when combined with available data from the experimental design) is greater than that required by an experimental design (e.g., such as the 30-aldea example cited earlier). It is for two main reasons.

First, in model based designs, explanatory variables are necessary to determine an overall estimate of program impact. In experimental designs, randomization "washes out" the effects of all variables other than treatment. Thus, it is not necessary to include any explanatory variables in the model, and the plain, unconditional double-difference estimate is all that is needed to assess program impact, and the variation in the explanatory variables is irrelevant to the estimate. It is not necessary to do ex-post matching, and it is not necessary to do covariate adjustment. (The impact estimate is the average value of the double difference over the four groups of the sample design (treatment before, treatment after, control before, and control after.) To be sure, it is also of interest to estimate the relationship of impact to various explanatory variables (such as the variables observed in the questionnaire), but this is not necessary.

With the model-based, observational-data approach being suggested here, the role of the explanatory variables is quite different from their role in an experimental design. Since randomization is not involved either in the selection of experimental units or the assignment of treatment, the model must describe the relationship of impact to all observable variables that may affect outcome (more exactly, are causally prior to treatment, empirically related to treatment, and affect outcome conditional on treatment). The explanatory variables are no longer just "desirable" (to describe the relationship of impact to them), but essential to determine a good estimate of overall program impact (via covariate adjustment). Estimates of program impact are obtained by substituting the values of explanatory variables into the estimated relationship.

The model must be a correct specification of this relationship. In designing a survey to be used as the basis for developing an analytical model, it is important to assure that there is substantial variation in all of the important explanatory variables, and low correlation (low collinearity; high orthogonality) among them. Prior to the survey, controlling variation is usually possible only at the level of the primary sampling unit, since information about the ultimate (lowest-level) sampling unit, the household, is not known until the survey is completed. With respect to variables measured at the household level, there is little or no control over the variation (spread, balance) or the degree of orthogonality (correlation) among them. S*ince there is no control over the variation in the household-level explanatory variables, and since adequate variation in the explanatory variables is necessary to develop a covariate-adjustment model, the sample size must be larger than in a designed experiment.*

There is a second (but related) reason why sample sizes are usually substantially larger in observational studies than in designed experiments or quasi-experimental designs (structured observational studies), and that has to do with the data analysis. It was mentioned that the data analysis involves ex-post matching, or "trimming" the sample to cause the distributions of the treatment and control samples to be more similar, and hence reduce model dependency (dependence of the outcome estimate on model specification). The process of trimming is also used to increase orthogonality among explanatory variables, to lower the correlation among regression-coefficient estimates. This process loses data. For this reason, it is desirable to start the data analysis with a larger sample than is the minimal needed for the statistical estimation procedures. In performing ex-post matching, it is generally the case that more control units are dropped than treatment units. (In the present application, we in fact have more control units available from the experimental design than treatment units.)

To emphasize: For this approach, in which we have no opportunity to control the values of the explanatory variables, *the sample size must be quite large*. Experience has shown that, using simple random sampling, sample sizes of *at least 400* are generally required to produce useful results with the model-based approach, if the observations can be included in a parametric (analytical; multiple regression) model. If it is desired to do nonparametric analysis, such as displaying results for domains (subpopulations) of interest (such as tables of means), a reasonable sample size must be available for each domain of interest. It is our recommendation that if the analytical-model approach is adopted, then a sample of at least 400 treatment units (program farmes) and 400 control units (potential program farmers) be available, after ex-post matching (trimming). This is considered a minimal sample size, and it was on this basis that we conducted a supplementary sample of 600 program farmers. In the interest of efficiency (high precision and power), it is generally desirable to have "balance" in the sizes of the treatment and control samples. This is the reason for the statement that the treatment sample and the control sample be comparable in size (numbers of farmers, not numbers of aldeas), after ex-post matching. The control-sample baseline data have already been collected and are useful for the new approach.

A final point must be made with respect to sample size. The survey questionnaire is large, i.e., contains many variables. Under the original design concept (a pretest-posttest-with-randomized-control-group experimental design), it did not have to be large – all that was required was to measure the outcome variables of interest. It was made large in order to enable estimation of the relationship of impact to explanatory variables (i.e., of dependent variables to independent variables). Now that the original experimental design concept cannot be implemented, the variables of the questionnaire are much more important than before, since they are necessary for use in covariate adjustment. In the data analysis, we will be "searching for structure." On the one hand, we will use "causal analysis" to suggest relationships (linkages of outcome variables to input or other explanatory variables), but we will be using the sample data to identify and suggest relationships, i.e., we will be conducting what is known as "exploratory data analysis." (This is often done in data-mining contexts using an automatic interaction detection or classification-tree program, such as Salford Systems' Classification and Regression Technique (CART) or SPSS Chi-Square Automatic Interaction Detection (CHAID) program.) In doing this sort of (nonparametric) analysis, it is important to have a lot of data.

In doing multiple-regression analysis, the ability to estimate the model is highly dependent on the number of degrees of freedom available for estimation of the model error term. This is approximately the number of observations less the number of parameters being estimated in the model. (The number of parameters can be much larger than the number of variables, e.g., when a categorical variable is replaced by a set of indicator variables, one for each category (or group of related categories) of interest.) In many situations, it is not possible to combine all of the data into a single model, and separate regression models are necessary or desirable. A further loss in precision is introduced by the fact that two-stage sampling was used for the selection of the control units in the original design (i.e., a sample of aldeas was selected, and than a sample of farmers was selected from each sample aldea). For all of these reasons, it is highly recommended that if this alternative design approach is adopted, the program-farmer sample size (including both the original design and the additional sample of 500 program farmers) be on the order of 500-1,000, preferably on the "high side." As things appear at present, we will have the simple random sample of 500 program farmers plus about 200 program farmers from the experimental design (Cohort 2 aldeas), for a total of about 700 program farmers in the treatment group.

We mentioned earlier that the model must be "correctly specified." This means that its functional form is correct, relative to the variables included in it. The model does not have to be an absolutely correct representation of the outcome with respect to all variables, seen and unseen, but simply with respect to the observable variables (i.e., those measured in the survey and available from other sources). The ex-post matching and the covariate adjustments are done relative to the *observables*. While the experimental design approach removes the influence of all non-treatment variables from the impact estimate, the analysis of observational data accounts only for the effects of observed variables. Whether there are important unknown ("hidden") variables that may bias the results is considered through causal modeling.

Thus, a supplementary simple random sample of 600 program farmers was selected from the population of 1,797 program farmers who entered the FINTRAC FTDA program between March 1 and August 31 of 2009. We decided to keep the sample design for the supplemental sample simple, if the budget for data collection would permit. The principal reason for this decision was that additional stratifications were not expected to improve the quality of the household-level analytical model for which the survey data would be used, and might even impair it. It is expected in this application that the household-level variables will have a much stronger effect on the outcome than the aldea-level characteristics, and the stratification of aldeas would have introduced substantial variation into the household selection probabilities. While the selection probabilities are not very important to the model development (since it is expected that we will use mainly unweighted regression estimates rather than weighted regression estimates), they are of some interest, and of greater value if comparable in size, as from a simple random sample.