FINAL REPORT (REVISED)

Impact Evaluation of the Farmer Training and Development Activity in Honduras

Millennium Challenge Corporation Contract

MCC-10-0133-CON-20 TO01

NOVEMBER 26, 2013

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

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

NORC at the UNIVERSITY of CHICAGO

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This document is the final report for the impact evaluation of the Farmer Training and Development Assistance (FTDA) project funded by the Millennium Challenge Corporation (MCC) in Honduras over the period 2007-2010. The project was implemented by the Millennium Challenge Account - Honduras (MCA-H) under a Compact between the governments of Honduras and the United States of America.

The Goal of the Compact in Honduras, which ended on September 30, 2010, was to stimulate economic growth and poverty reduction. To accomplish this goal, the MCA - Honduras Program aimed to achieve the following objectives:

- Increase the productivity and business skills of farmers who operate small and medium sized farms and their employees (the "Agricultural Objective"); and
- Reduce transportation costs between targeted production centers and national, regional, and global markets (the "Transportation Objective").

Over the course of the Compact, two projects were implemented by MCA - Honduras to achieve these Objectives:

- The Rural Development Project (RDP), which comprised of four activities: (i) farmer training and development, (ii) facilitation of access to credit by farmers, (iii) upgrading of farm to market roads, and (iv) provision of an agriculture public grants facility.
- The Transportation Project, which upgraded two major sections of the CA-5 Logistical Corridor, and pave approximately 65 km of secondary roads.

Between May 2007 and September 2012, NORC undertook rigorous impact evaluations of two MCA - Honduras Program activities: the Farmer Training and Development Activity (FTDA), and the Transportation project. This report discusses and presents the findings of the FTDA impact evaluation.

The MCA Honduras Rural Development Project and Farmer Training and Development Assistance Activity

The MCA - Honduras Rural Development Project sought to increase the productivity and improve competitiveness of smallholder farmers who 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. Towards this end, under the RDP, MCA-H implemented four activities, one of which was the FTDA.

The FTDA provided farmers with a comprehensive assistance package that focused on all stages of production from field preparation and planting, to the administration of fertilizers, herbicides, insecticides and improved varieties of seeds, to the negotiation with buyers and the marketing of their high-value horticultural crops. In addition to technical assistance, eligible farmers also received a limited amount of financial support to install better irrigation systems.

FTDA program participants (Program Farmers) were expected to significantly increase their agricultural productivity and income by improving yield through the use of improved technology, and changing their crops mix to emphasize horticultural over basic crops. There was also an expectation that this would lead to increased employment on farms.

Based on these hypotheses, we focused our evaluation on the following expected outcomes and associated impact indicators:

Expected outcomes	Indicators
The FTDA will lead to:	
Increased cultivation of horticultural crops (change in crop mix)	▪ Net income from horticultural crops ▪ Net income from basic grains ▪ Input expenditures on horticultural crops ▪ Input expenditures on basic grains
Increased household income	▪ Net household income ▪ Total household consumption
Increased employment on farms	▪ Labor expenses

Evaluation Design: From a Randomized Control Trial to an Econometric Model

The impact evaluation design for this activity changed over the course of the evaluation due to problems faced during its implementation. In its original conception, NORC and MCA-H planned to use a "group randomized" experimental design involving randomized assignment of communities (*aldeas*) to treatment. Following a series of implementation problems, the final approach adopted was a causal modeling approach that relied on econometric analysis to estimate impact.

The Original Evaluation Design: A Design-Based Approach, Using an Experimental Design (Randomized Controlled Trial)

The planned experimental design, or randomized control trial (RCT), consisted of an analytical survey design in which members of 200 matched pairs of aldeas were randomly allocated to treatment and control groups. In both treatment and control aldeas, "potential FTDA farmers" were selected using criteria provided by the implementing agency, replicating the program's selection process as closely as possible in the experimental sample. Baseline data were collected from these potential FTDA farmers and a probability sample of 20 additional households in each treatment and control aldea, and soon thereafter, the implementing agency entered the treatment aldeas to select and provide technical assistance to a final group of treatment farmers. Follow-on data collection was to occur either 18 or 24 months after the baseline.

Since the sample design was based on randomized selection of treatment and control communities, this design would have provided a sound basis for making causal inferences from the collected data.

Problems in Implementing the Experimental Design

NORC encountered significant problems in implementing the experimental design described above. The proposed group-randomized design, which randomly assigned aldeas to treatment and control groups, was designed such that the program implementer was free to select the final Program Farmers within randomly selected treatment aldeas; it was agreed, however, that NORC would select potential treatment farmers in these aldeas from which the Fintrac would make the final selection. For this purpose, we used Fintrac's selection criteria. A similar screening process for farmers would occur in control aldeas. It turned out that, despite concerted efforts, undertaken with the support and involvement of Fintrac, NORC could not replicate Fintrac's farmer selection criteria. This fact alone would not have compromised the reliability or validity of the experimental design, had the randomly selected sample of treatment aldeas been acceptable to Fintrac. NORC had selected these sample aldeas from a list of all aldeas that Fintrac would enter in the final three program years, randomly selecting over 100 for the treatment group. We selected a treatment aldea sample larger than was necessary for the evaluation design, which allowed for rejection of some aldeas as "out of scope," according to well-defined quantitative criteria. However, acceptance of selected treatment aldeas as program-eligible by the implementer was far smaller than expected, resulting in insufficient treatment aldeas and farmers to implement the experimental design[1].

The randomized selection and assignment to treatment of aldeas was essential to the experimental design. If some of the randomly selected treatment aldeas were identified as "out of scope," this would not have presented a problem. However, of the almost 1,000 potential treatment farmers in 113 treatment aldeas that were deemed eligible for the FTDA program, according to objective selection criteria provided by the implementer, Fintrac chose to provide technical assistance to only 28 farmers in 19 aldeas. This low rate of acceptance into treatment resulted in a small sample that was inadequate for use as a basis for evaluation, because the small farmer and *aldea* sample sizes would result in low precision (reliability) and, perhaps, low validity as well.

Revised Evaluation Design: A Causal Modeling Approach, Using Econometric Analysis

The original evaluation design concept of using an experimental design is a *design-based* approach to impact evaluation. Our resolution to the problem described above was to use a *causal-modeling* approach (revised approach) that would make use of almost all the data that were collected for the original experimental design, and complement these data with additional sample data from Fintrac's program clients[2]. Specifically, the additional sample comprised of new recruits selected by Fintrac from the sample of experimental treatment aldeas and a sample of Fintrac's clients, with no relation to the evaluation design, who were randomly selected from its program client list. Baseline data were collected from the additional sample, increasing the

[1] The primary reason for the loss of *aldeas* from the treatment sample was the following. The eligibility of *aldeas* "flows up" from the eligibility for farmers, and for an *aldea* to be eligible, it had to contain at least some program-eligible farmers. The rejection of large numbers of screened farmers effectively eliminated (as out-of-scope) larger than expected numbers of the randomly selected *aldeas* from our treatment sample.

[2] Both the design-based approach and the model-based approach use models – causal models that describe the relationship of outcomes of interest to explanatory variables, and statistical estimation models derived from the causal models, to estimate impact. The causal and statistical models involved in the *model-based* approach are generally more complex than those used for the design-based approach, hence the use of the term *model based*. With the revised approach there is much more focus on causal modeling, and the approach may also be characterized as a *causal modeling* approach.

and Erkki Pahkinen (Wiley, 2004). (The Lohr book is the most informative.)

sample size to achieve a satisfactory level of precision and power. Because the data no longer corresponded to a structured, properly randomized experimental design, the usual design-based estimates no longer applied; hence, we used special causal modeling procedures to construct good (unbiased, consistent) estimates of impact.

The validity of the results of an impact evaluation rests on the soundness of the causal model used to represent the system under study, and the soundness of the associated statistical model and estimators used to estimate impact. The causal model underlies both the experimental design and the revised design. For an experimental design, randomized assignment to treatment assures that the treatment and control aldea samples are distributionally equivalent with respect to all factors that might affect outcomes of interest, except for treatment. This condition greatly simplifies the procedures used to estimate impact, and eliminates the possibility of selection bias. For the revised approach, to reduce selection bias we used statistical procedures, such as propensity-score matching or regression analysis, to adjust for distributional differences between the treatment and control aldeas. The approach used in the FTDA evaluation involved the development of a "selection" model that estimates the probability of selection, or *propensity score*, of a sample unit for treatment, and the use of an impact estimator that is based on the estimated propensity score.

This evaluation uses data from a two-round panel survey. In this case, for both the experimental and revised designs, the usual measure of impact is the double-difference measure, or the difference, before and after the program intervention, of the difference in means of the treated and untreated units. For the experimental design, an unadjusted double difference in sample means of the four design groups (treatment before, treatment after, control before, control after) is a good estimate of program impact (i.e., a good estimate of the double-difference *measure*). For the revised design, the estimate of impact is more complicated, and usually obtained from regression analysis.

The Results

The table below presents the impact estimates for selected outcomes of interest based on econometric analysis of the data from the revised design:

Table ES1. Estimates of Average Treatment Effect (ATE), Using the Modified Regression-Adjusted Propensity-Score-Based Estimate of Impact

Outcome Variable	Estimate	Standard Error
Basic Grains (BG)		
Income, basic grains (IncBG)	-120	837
Total expenses, basic grains (ExpBG)	837*	393
Net income, basic grains (NetBG)	-957	750
Labor expense for basic grains (LabExpBG)	435	264
Other Crops (OC)		
Income, other crops (IncOC)	16773*	4298
Total expenses, other crops (ExpOC)	5413*	1078
Net income, other crops (NetOC)	11360*	4175
Labor expense for other crops (LabExpOC)	1911*	742
Labor Market Employment and Household Income and Expenditures		
Labor market income (IncEmp)	149	733
Total household expenditures (TotHHExp)	204	496
Net household income (NetHHInc)	18926*	13306

Table ES1. Estimates of Average Treatment Effect (ATE), Using the Modified Regression-Adjusted Propensity-Score-Based Estimate of Impact

Outcome Variable	Estimate	Standard Error
Production of Horticultural Crops		
Horticulture	-.0397	.0194

Note: Income and expense measured in Honduran lempiras.

These results show a positive effect of the FTDA program. Net income change from horticultural crops is on average 11,360 lempiras (USD 600) higher for program participants than for nonparticipants. Input expenditures on these crops increased far more than they did for basic crops, implying a higher level of activity in cultivation of high-value crops among program farmers[3]. The results suggest a corresponding decline among program farmers in income from basic crops, as might be expected with changing crop mix; however, this decline is not statistically significant. These results are consistent with the program logic and hypotheses for the FTDA.

Some of the results do not conform to expectations. For example, the program does not appear to have had a positive effect on the proportion of farmers growing horticultural crops. This could well be because the implementer primarily chose as program participants farmers who showed a proven ability to grow horticultural crops. This suggests that increments in income from horticultural crops came from increased production among farmers already growing horticultural crops and not from farmers who switched over for the first time.

Conclusion

The results of the impact evaluation show that the FTDA activity had a positive impact on its primary area of focus: activities related to horticultural crops. However, those effects were small in magnitude. Furthermore, we did not detect a broader positive impact on household income and expenditures.

The impact estimates were based on all of the data obtained from the original experimental design, augmented by data collected from a sample of program farmers recruited by Fintrac in the course of its normal project operations. Statistical/econometric analysis was used to adjust for differences between the treatment and control samples and to thereby reduce potential selection bias associated with a lack of proper randomization. The statistical/econometric analysis procedures used to estimate impact are based on sound causal models and causal modeling theory[4]. The impact estimates constructed in this evaluation are estimates of the causal effect of the FTDA program intervention. An *ex post* statistical power analysis (presented in Annex 1) shows that the study was not “underpowered.” We, therefore, consider the inferences made in this evaluation analysis to be sound (valid and of adequate precision and power), within the constraints of the assumptions and limitations described below.

[3] The impact results do not imply that the incomes or expenditures of program farmers did or did not increase by a substantial amount. They show changes in income (and other indicators) for program farmers relative to changes in the same indicators among control farmers.

[4] Neyman-Fisher-Cox-Rubin Causal Model, potential outcomes model, counterfactuals model

Summary of Assumptions and Limitations

The major assumptions associated with this analysis, which should be taken into consideration in reviewing the evaluation results, are the following:

1. The stable unit treatment value assumption (SUTVA, no macro effects assumption, partial equilibrium assumption) is made. This means that the effect (potential outcomes) on one individual are not affected by potential changes in the treatment exposure of other individuals. This implies, for example, that the program is not so large that the outcomes are correlated (e.g., that farmers would produce such a large amount of horticultural crops that the market would collapse).

2. The causal models are correct. The key assumption here is that all important unobserved variables affecting selection are time invariant (i.e., are constant between the two survey rounds).

3. The program intervention represents a "forced change" in (experimental control of) the agricultural system in Honduras.

4. The half of the country that Fintrac had treated before this evaluation began is similar to the half yet to be treated, with respect to relationships among the important causal variables represented in the causal model underlying the statistical analysis.

Other more specific assumptions are listed for particular estimation equations in the detailed analysis presented in Annex 1.

The limitations of the evaluation are:

1. The causal analysis used to estimate impact is based on assumptions about the selection process. The original evaluation design was based on randomized assignment (of *aldeas*) to treatment, and represented a firmer basis for making causal inferences. With the original approach, randomized assignment assures that the distributions of explanatory variables (other than treatment) are the same for the treatment and control samples. With the revised design, this assertion depends on the correctness of the causal model, and the assumption that unobserved variables affecting selection for treatment are time-invariant.

A. INTRODUCTION

This document is the final report for the impact evaluation of the Farmer Training and Development Assistance (FTDA) project funded by the Millennium Challenge Corporation (MCC) in Honduras over the period 2007-2010. The project was implemented by the Millennium Challenge Account - Honduras (MCA-H) under a Compact between the governments of Honduras and the United States of America.

The Goal of the Compact in Honduras, which ended on September 30, 2010, was to stimulate economic growth and poverty reduction. To accomplish this goal, the MCA - Honduras Program aimed to achieve the following objectives:

- Increase the productivity and business skills of farmers who operate small and medium sized farms and their employees (the “Agricultural Objective”); and
- Reduce transportation costs between targeted production centers and national, regional, and global markets (the “Transportation Objective”).

Over the course of the Compact, two projects were implemented by MCA - Honduras to achieve these Objectives:

(1) The Rural Development Project, which comprised of four activities: (i) farmer training and development; (ii) facilitation of access to credit by farmers; (iii) upgrading of farm to market roads; and (iv) provision of an agriculture public grants facility.

(2) The Transportation Project, which upgraded two major sections of the CA-5 Logistical Corridor, and paved approximately 65 km of secondary roads[5].

Under the NORC–MCA - Honduras contract (May 2007 to September 30, 2010) and the follow-on contract between NORC and MCC (September 30, 2010 to December 31, 2011), NORC undertook rigorous impact evaluations of two MCA - Honduras Program activities: the Farmer Training and Development Activity (FTDA), and the Transportation Project[6]. This report discusses and presents the findings of the FTDA impact evaluation. A separate report presents the findings of the Transportation Project impact evaluation.

The remainder of this report is organized as follows. Section B presents a brief description of the Farmer Training and Development Activity. Section C discusses in-depth the evaluation design and its implementation. This section discusses the original experimental design, as it was developed in 2007, as well modifications to that design necessitated by problems that were encountered during its implementation. These implementation problems had a major effect on

[5] The initial scope of the Transportation Project called for upgrading and paving two major sections of Highway CA-5, paving at least 70 km of secondary roads, and developing a vehicle weight control system. Due to increases in costs and a partial re-scoping of the road rehabilitation component, the project was scaled back and ultimately only about 65 km of secondary roads were rehabilitated. The vehicle weight control system was not implemented.

[6] MCA - Honduras rehabilitated 495 km of rural roads under the Rural Development Project. However, given that these rural roads form part of the national road network, for the purpose of the evaluation, NORC considered the evaluation of the rural roads improvement within the framework of the Transportation Project.

the analysis of the impact evaluation and are described in Section C. Section D describes the household survey conducted to collect the primary data on which this impact evaluation is based. Section E presents a summary of results of the impact evaluation. Annex 1 presents a detailed technical discussion of the impact analysis and results. Annex 2 presents a more detailed description of the evaluation design. Annex 3 describes the statistical power analysis used to determine the sample sizes for the evaluation. The survey questionnaire is separately bound.

B. THE FARMER TRAINING AND DEVELOPMENT ACTIVITY

The MCA - Honduras 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 agricultural inputs to farm-gate. The MCA - Honduras Program sought to alleviate these constraints and contribute to increased productivity among farmers through four activities:

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

The Farmer Training and Development Assistance (FTDA) Activity[7], implemented by Fintrac, provided direct technical assistance and training to more than 7,500 smallholder farmers in 16 departments of Honduras. The program, which emphasized high-value crops and crop and market diversification, used a market-driven production system approach to enable growers to implement technologies that increase yields, quality and competitiveness. The program worked closely with all members of the horticultural value chain and integrated growers with buyers, financial institutions, and equipment, input and services provides. Assistance and training was also provided to technicians from NGOs, agriculture schools, universities, associations, and the public sector, as well as to staff from private-sector allied agribusinesses (wholesalers, retailers, exporters, processors, other buyers, and input providers of both goods and services). This

[7] Program description information was obtained from the following sources:
(a) http://www.mcahonduras.hn/historico.php?o=17&i=2; (b) http://www.fintrac.com/past-projects.aspx; (c) Fintrac EDA Impact Report, 2006-2010.

integrated approach was intended to ensure that program farmers would continue to benefit from the assistance after the program ended.

More specifically, the program consisted of the following activities:

1. Identify existing market demand for commercial crops that Program Farmers can supply.
2. Identify Program Farmers who are willing and able to supply such demand. In its implementation of the FTDA, Fintrac used strict eligibility criteria for accepting farmers into the program. These criteria evolved over the life of the MCA - Honduras program, and included a host of objective and subjective criteria. Measurable criteria included volume of land under cultivation (no more than 50 hectares), access to water during at least six months per year, access to a paved road (i.e., less than 2 hours away), flooding situation, slope of land, depth of soil, and access to at least 70,000 lempiras/hectare for investments. Less quantifiable criteria included an interest and desire to cultivate horticultural crops and the motivation to follow Fintrac’s guidance and adopt new techniques. Farmers who were deemed eligible according to these criteria were accepted into the FTDA program and received weekly visits from Fintrac Field Technicians for a period of between 18 and 24 months.
3. Develop business plans that enable Program Farmers to meet market demand, and work with lenders, suppliers and buyers to ensure that these business plans are realistic.
4. Help Program Farmers obtain credit to finance their business plans. In addition, eligible farmers received a limited amount of financial support in the form of agricultural equipment used to install better irrigation systems.
5. Provide Program Farmers with technical assistance (TA) in production (including field preparation and planting, and administration of fertilizers, herbicides and insecticides, drip irrigation and hybrid seed varieties), business skills, marketing, postharvest handling and standards certification. Small farmers receive TA via project technicians and NGO partners’ technical staff, all highly trained in Fintrac’s market-led extension methodology. The program ensured that Program Farmers employed environmentally sustainable agricultural practices. It also developed instruments (e.g., purchase contracts) and market-based support services (e.g., farmer associations, processing arrangements) to help Program Farmers to successfully execute their business plans.
6. Certify that no crops supported by the Rural Development Project will substantially displace U.S. production.

Program Farmers were expected to significantly increase their agricultural productivity and income by increasing the number of hectares under cultivation, improving yield through the use of improved technology, changing their crops mix to emphasize horticultural over basic crops, and working with local buyers as well as exporters to select and produce those crops that are more marketable.

C. THE ORIGINAL FTDA EVALUATION DESIGN

The evaluation design for the FTDA underwent a significant change since the inception of the project in 2007. The original evaluation design concept was a “group-randomized” experimental design in which paired samples of program-eligible aldeas were selected and one member of each pair was assigned to treatment. In constructing the aldea sample frame from which the aldea sample was chosen, NORC relied on Fintrac’s assertions about the universe of aldeas that they would work in during the remainder of the FTDA activity, and used Fintrac’s eligibility criteria to select potential treatment farmers within the aldeas. Despite diligent efforts to comply with program eligibility criteria, many of the aldeas and farmers selected by NORC as FTDA-eligible were not acceptable to Fintrac. As a result, the experimental design approach ceased to be a viable option, and NORC opted to choose an alternative approach for the evaluation.

The revised evaluation design retained much of the structure and all the respondent data from the RCT design, which included data from all treatment farmers accepted by Fintrac under the original design and the complete sample of control aldeas. We supplemented this sample with new Fintrac recruits in the treatment aldeas and program farmers from Fintrac’s regular operational lists. Since the revised design was no longer based on proper randomization, the design-based estimates associated with the original experimental design no longer applied, and it was necessary to use an alternative approach (causal modeling and econometric analysis) to construct good (unbiased, consistent) estimates of impact.

The impact evaluation results presented in Section E of this report pertain to the revised model-based approach. This section of the report describes the original experimental design, the implementation problems encountered, and the alternative approach used for this evaluation.

C.1 THE ORIGINAL APPROACH: AN EXPERIMENTAL DESIGN

C.1.1 Evaluation Goals

During an initial trip to Honduras, NORC’s evaluation team met with the staff from the MCA - Honduras M&E and Rural Development Project teams, as well as MCC evaluation staff to discuss the evaluation goals and design options for the FTDA. Given that evaluation activities for the MCA - Honduras commenced close to two years after the start of the compact, the FTDA was already underway. Therefore, during the trip, the NORC team also had the opportunity to travel to the field and visit several Fintrac Program Farmers who were already receiving assistance. These visits and discussions with Fintrac provided the NORC team with an opportunity to understand the structure of the FTDA activity, and the proposed rollout of the project throughout Honduras over the next three years. Based on these discussions with various stakeholders and a thorough review of compact documents, including reports submitted by Fintrac, NORC developed several key hypotheses for the evaluation of the FTDA activity – namely that improved farmer training would:

- Increase cultivation of horticultural crops;
- Increase incomes of farm households; and

— Increase employment income on farms.

Based on these hypotheses, we proposed to focus the evaluation on the following impact variables: changes in household income (farm and off-farm) – net and gross; and changes in farm employment.

C.1.2 The Experimental Design Approach and Its Implementation Requirements

The experimental-design evaluation model developed in 2007 called for randomly allocating farming communities – in this case, aldeas (communities, villages) – into two groups: those that receive technical assistance now (the treatment communities) and those that receive it approximately 18 months later (control communities). Baseline and endline 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[8].

This experimental design is a pretest-posttest-randomized-control-group design, where the usual measure of program impact is the population double difference measure, which is the double difference of the four group true (population) means, $(\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})$, where μ_{11} = mean of treatment group at endline, μ_{10} = mean of treatment group at baseline, μ_{01} = mean of control group at endline, and μ_{00} = mean of control group at baseline. The double-difference *estimator* is the double difference of the four corresponding sample means.

For a pretest-posttest-comparison-group design, the (unadjusted, "raw") double-difference estimate is given by the following formula (see Box 1):

$$\text{Estimate of Impact} = \text{DD}_{\text{raw}} = (\bar{y}_{t1} - \bar{y}_{t0}) - (\bar{y}_{c1} - \bar{y}_{c0})$$

where

DD_{raw} = double-difference estimate (raw, unadjusted)
$\bar{y}_{t1}$ = mean outcome for treatment sample at time 1
$\bar{y}_{t0}$ = mean outcome for treatment sample at time 0
$\bar{y}_{c1}$ = mean outcome for control sample at time 1
$\bar{y}_{c0}$ = mean outcome for control sample at time 0.

In the preceding equation, the variable $\bar{y}_{ij}$ refers to any outcome variable of interest, such as income. For a pretest-posttest-randomized-control-group design, the double-difference *estimator* is a consistent estimate of the double-difference *measure*. For pretest-posttest designs that are not based on randomized assignment to treatment, the double-difference estimator is not necessarily an unbiased or consistent estimate of the double-difference measure, and more complicated

[8] NORC worked closely with the M&E Director for MCA - Honduras and the MCC Resident Country Director in Honduras in developing the evaluation design for the FTDA activity. NORC also drew on the expertise of its expert group, which was comprised of evaluation experts from the University of California at Berkeley and the National Institute of Public Health in Mexico, both parties that had been heavily involved in the design and implementation of evaluation of Mexico's conditional cash transfer program, *PROGRESA/OPORTUNIDADES*, one of the best known randomized control designs conducted in the development field. The formulation and finalization of the FTDA evaluation design required a second trip by the NORC team to Honduras for presentation and further discussion of the design and its implementation requirements.

estimators, such as matching estimators or regression estimators, must be used to obtain an unbiased or consistent estimate of the double difference measure.

If the sample design is based on randomized selection of treatment and control communities, it provides a sound basis for making causal inferences from the collected data.

It is important to note here that the original RCT design called for program treatment to be varied *among sample communities* rather than *among farmers within sample communities*[9]. This type of design is referred to as a "group-randomized" design, where randomization occurs at the community level because randomization at the farmer level is not feasible. As a randomized experimental design, this approach would have been a sound basis for making causal inferences about the effect of the program intervention. Note that the unit of treatment is the aldea, whereas the unit of analysis is the household.

Box 1: The Double-Difference Estimator

For a pretest-posttest-control-group experimental design, the sample double-difference estimator is a consistent estimate of the population double-difference measure. The standard approach of calculating double differences with respect to projects is based on the two situations faced by households or communities relative to the program intervention: those that receive an intervention – technical assistance in this case – and those that do not. The first differences (of which there are two) are the differences in average values for the outcome variables between treatment and non-treatment communities, before and after the program intervention. The second difference is between the two pre-treatment and post-treatment differences. The steps to be taken can be summarized as follows:

- Undertake a baseline survey before the intervention is started, covering the treatment and control communities.
- After the project is completed, undertake one or more follow-up surveys. These should be highly comparable to the baseline survey, both in terms of the questionnaire and the sampled observations (ideally the same sampled observations as the baseline survey).
- Calculate the mean difference between the pre- and post-treatment values of the outcome indicators for each of the treatment and comparison groups.
- Calculate the difference between these two mean differences to obtain the double-difference estimate of impact of the program.

The implementation of this rigorous randomized control design had operational implications for the implementation of the FTDA by Fintrac. In recognition of this burden, prior to receiving approval of the evaluation design, NORC presented it to Fintrac and reached consensus on the implementation requirements that were crucial to making the evaluation a success. Subsequently, MCA - Honduras, MCC, NORC, and Fintrac agreed to an implementation approach that consisted of the following steps:

- Identification of geographic areas that Fintrac would expand into during 2008 and 2009[10]; NORC reached agreement with Fintrac on aldeas, as a manageable community that would be used as the primary sampling unit (PSU) of a two-stage sample design.

[9] 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 might 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.

[10] The target population for this evaluation was the Fintrac expansion area for the three years following the start of this evaluation project. Since the FTDA project had already been operating in part of the country prior to the start of the evaluation, the expansion area was not the entire country and, it was understood that the inferential scope of the study would be restricted to this area. While this was a concern, it was not considered to be "debilitating," because the expansion area was a large portion of the country and because the study was concerned with estimation of an

- Use of a matching algorithm to group aldeas into pairs that have similar characteristics, based on available data on observable characteristics
- Selection of a probability sample of 100 pairs of matched aldeas, using an appropriate survey design[11]. The experimental design took into account aldea data that were available from Honduras Census and geographic information system (GIS) sources. These data included elevation, climatic zone, soil capacity, rainfall, temperature, vegetation cover, protected area status, distance to nearest major river, population, and a variety of travel times to point of interest. These data were used to match aldeas prior to randomized assignment of one member of each matched pair to treatment, and to stratify the sample according to variables that were believed to have an effect on outcomes of interest (to assure that the sample had adequate variation on these variables). This type of experimental design, in which units are paired and one member of each pair is randomly assigned to treatment, is called a "matched-pairs" design. Annex 2 presents a description of the sample design used.
- In each sample community, identification of a list of prospective program farmers that Fintrac would use later to make its final selection of lead farmers in both the treatment and control groups. Because Fintrac did not wish to prematurely enter control communities that it would not be working in till 18 months later, and because the design preferred that a similar farmer selection process be used in both treatment and control communities, NORC and MCA - Honduras decided to use independent "screeners" to select these potential program farmers by using the exact same procedures and criteria used by Fintrac. To this end, Fintrac participated in the training of these screeners, who also accompanied Fintrac technicians on site visits to further familiarize themselves with the project, and reviewed and approved screening forms. All measures were taken to ensure that identification of farmers for the sample aldeas mirrored Fintrac's selection process[12].

interaction effect (the "double difference"), rather than with estimating means or totals for a specific population or subpopulation. Although means and totals of socioeconomic variables usually vary substantially over regions, relationships such as effect interactions are usually less sensitive to regional variation. This situation was not perfect for an evaluation study, but it was the best solution, given the reality on the ground.

[11] At the request of MCC, the design was modified slightly, to include more treatment aldeas than control aldeas, thereby reducing the number of aldeas that Fintrac would be restricted from working in. A total of 113 matched pairs was selected (by marginally stratified probability sampling, 226 aldeas in all), and randomly divided into treatment and control groups. 23 of the control aldeas were dropped (randomly), resulting in a desired sample of 113 treatment aldeas and 90 control aldeas. This unbalancing of the design decreased its efficiency, but did not impair its ability to produce unbiased estimates of impact. To allow for replacement of out-of-scope aldeas, a total sample of 266 matched aldea pairs was selected (266 treatment aldeas and 266 control aldeas).

[12] In this group-randomized design, randomized assignment to treatment is done at the aldea level, not at the farmer level. The randomized design did not require that potential lead farmers all be selected for treatment by Fintrac. The validity of the original experimental design did not depend on the procedures used to select program farmers within aldeas; it rested on maintaining the classification (treatment or control) of the randomized selection of treatment and control aldeas. Within an aldea, farmers were stratified into two strata: a certainty stratum of program participants (or, in control aldeas, potential treatment farmers) and a non-certainty stratum of all others (from which a random sample of 20 farmers was selected). Since all farmers of an aldea are subject to (probability) sampling, it would have been possible to obtain an unbiased estimate of program impact, no matter how the treatment farmers were selected. We recommended a similar selection process for potential program farmers in treatment and control aldeas, because it imposed an additional level of control and would hence lead to higher precision and power. There were in effect three control groups: the untreated farmers in a treatment aldea; the potential program farmers in a control aldea; and the others in a control aldea.

- Collection of baseline data from treatment and control communities, from all potential program farmers who were identified (a certainty sample) and a probability sample of other farmers in both sets of sample aldeas.
- Provision of technical assistance by Fintrac to treatment aldeas soon after baseline data collection. The FTDA intervention consisted of three initial visits to test and identify Program Farmers from among the screened potential treatment farmers and the subsequent provision of technical assistance to Program Farmers by Field Technicians. Fintrac's stated rejection rate of farmers deemed to be eligible according to the eligibility criteria was 8-10 percent. Control aldeas would not receive Fintrac assistance until 18 months later.
- Collection of endline data from treatment and control communities, approximately 18 months later. In the treatment communities, follow-on data were to be collected from all program farmers and a probability sample of other farmers. In the control communities, data were to be collected from all potential lead farmers and a probability sample of other farmers.

The evaluation design, as it was developed in 2007, along with the agreed-upon implementation plan was intended to provide baseline and follow-on data from a sample of farmers that had been randomly assigned (at the aldea-level) to treatment and control groups. Because of random assignment of aldeas into treatment and control groups, we could assume that the farmer groups were the same distributionally in the treatment and control groups except for the intervention being measured. Thus, any significant differences observed between the two groups at the end of the treatment period could be attributed to the program intervention, allowing the calculation of unbiased estimation of the effect of treatment (i.e., of the program intervention).

Annex 2 presents additional details on the analytical survey design for the original experimental design. Annex 3 presents information on the statistical power analysis that was used to determine the *aldea*-level and farmer-level sample sizes. Below is a summary of the sample sizes planned for the experimental design:

Table 1: Sample sizes planned for Experimental Design

Aldeas	Farmers/households	Total farmers/households
113 (treatment)	9*	1,017 Program Farmers households (certainly sample)
	20	2,260 other households (probability sample)
90 (control)	9	810 potential Program Farmers households
	20	1,800 other households

* 9 was an estimate of the average number of farmers that would be accepted by Fintrac for its program in each treatment aldea

Although experimental evaluation designs such as the randomized control trial described above are complicated to implement, largely because they require some adaptation of program implementation processes to fit the needs of the evaluation design, we were confident at the outset that Fintrac, MCA - Honduras, and NORC were clear on these requirements and that numerous discussions and a firm agreement, in the form of a Memorandum of Understanding, would ensure the appropriate implementation of the evaluation design. Despite this expectation, NORC ran into serious obstacles in implementing this evaluation design.

C.2 PROBLEMS ENCOUNTERED IN IMPLEMENTING THE EXPERIMENTAL DESIGN

Despite multiple attempts to do so, NORC and MCA - Honduras were unable to replicate Fintrac's selection procedures and criteria for identifying eligible aldeas and program farmers, leading to significant reductions in sample size, and final treatment and control groups that were largely non-comparable due to differing criteria for aldea selection. Below, we describe the problems encountered, and the steps taken to address them.

NORC selected the original sample design consisting of 113 treatment and 90 control aldeas from a sample frame of aldeas obtained from official Honduras Census records from which. aldeas in areas already treated by Fintrac were removed. The remaining aldeas represented about one-third of the country. In addition, aldeas located in Gracias a Dios, national parks, and tourist areas (Islas de la Bahia) were removed (since Gracias a Dios is very remote, and the other areas have little to do with rural farming). The final list of aldeas from which the sample was selected included 1,822 aldeas out of a total of 3,675.

We anticipated that some of the remaining aldeas in the sample frame would not meet the FTDA eligibility criteria; hence, an "oversample" of 166 treatment aldeas and 166 control aldeas was selected, to allow for replacement of "out-of-scope" or ineligible aldeas (aldeas in national parks and non-farming communities), to maintain the sample size. Removing out-of-scope units from the sample frame, based on specified eligibility criteria does not introduce a selection bias. It is a standard procedure when a perfect frame is not available from available data sources. The scope of inference for the evaluation is the population of aldeas (not yet treated) that meet the eligibility criteria.

Initially, during the farmer screening conducted by MCA and NORC, very few of the sample aldeas were found to be out-of-scope and, hence, ineligible. Reasons for the ineligibility included being in a protected area or absence of agricultural production. Eligibility criteria used by NORC to identify potential farmers and, hence, potentially eligible aldeas (through the upward flow of eligibility criteria from farmer-level to aldea-level) included the following:

1. Access to water
2. Less than 50 hectares land
3. Interest in adopting FTDA
4. Not receiving other technical assistance.

It turned out, however, that basing the screening on the aforementioned criteria was insufficient. After visiting these "potential" program farmers in the treatment aldeas, only 15 percent were accepted into treatment by the implementer; as a result, entire treatment aldeas dropped out of the sample. After a thorough review of Fintrac's reasons for rejecting farmers and aldeas, we developed a new expanded set of aldea and farmer eligibility criteria (Box 2), and applied them to new group of randomly selected treatment and control aldeas. This second cohort (Cohort 2) of aldeas was selected from the remaining aldeas that Fintrac had not yet visited. NORC undertook the Cohort 2 screening, employing a well-established local survey firm, ESA Consultores, for this task.

We note that the preceding lists of eligibility criteria are not the full list that is contained in Fintrac's contract with MCA - Honduras. Instead, it represents the specific criteria provided to us by Fintrac to use in aldea and farmer screening. It was not the goal of NORC to select Fintrac's program clients. Our goal was to identify a sufficiently large sample of randomly selected aldeas, which contained sufficient numbers of program-eligible farmers that Fintrac would accept into the FTDA. This component of aldea selection was fundamental to the experimental design.

Box 2: List of Eligibility Criteria used for Cohort 2 Screening

Eligibility criteria for aldeas:

1. Located less than 2 hours from a paved road
2. Accessible year round, including winter or rainy season (at least 10 months a year)
3. Access to water for irrigation during at least 6 months of the year
4. Not regularly flooded
5. Not covered by forests
6. A slope of less than 47%
7. At least 15 inches of topsoil (can be easily plowed)

Eligibility criteria for the farmer:

1. Owns or rents a plot with an area of at least 0.18 hectares
2. Plants or is willing to plant vegetables or other crops promoted by FTDA (not including crops grown exclusively for home consumption)
3. Willing to adopt, implement and follow recommendations of the FTDA technician
4. Possesses or can secure 70,000 lempiras per hectare to invest in crops
5. Not a current recipient of agricultural assistance from another institution.

The purpose of identifying potential treatment farmers (program-eligible farmers) in the sample aldeas was to stratify treatment and control aldeas into potential treatment farmers and others for the purpose of constructing a treatment stratum and three control strata and thereby increase the precision and power of the design. Since all farmers in all treatment and control strata were subject to sampling, this procedure would not introduce bias.

It turned out that most of the farmers and aldeas screened using the expanded set of criteria still were unacceptable to Fintrac for its program. Specifically, of the 343 farmers in 85 aldeas, who were screened and deemed eligible according to the expanded list of eligibility criteria, Fintrac Field Technicians ended up accepting only fewer than 30 farmers in 16 aldeas.

At this point, it was clear that the original design concept of stratifying farmers based on the set of quantitative criteria provided to us by Fintrac was counterproductive: it resulted in a large proportion of aldeas containing no or very few farmers that were accepted into treatment. Hence, we concluded that we could not replicate Fintrac's selection process for two reasons: (1) it contains elements/criteria that could not be quantified and depended on some element of subjective assessment by the Fintrac Field Technician of a farmer's motivation, ability to learn and grow (*potencial para crecer*), and willingness to follow program requirements; and (2) the selection criteria evolved over time, based on lessons learned during implementation.

In order to implement the group-randomized experimental design, it was necessary for NORC to identify by means of quantitative eligibility criteria a list of aldeas that would be acceptable to Fintrac for its program. Despite diligent efforts, it turned out that this could not be done.

To summarize, implementing the original experimental design required us to identify the target population for the program, i.e., the aldeas that met the eligibility criteria for FTDA and contained program-eligible farmers. To enable this approach, the sample included a number of replacement aldeas, to maintain the sample size if some of the aldeas selected from the sample frame were ineligible according to the screening criteria. However, despite multiple efforts, it was not possible to identify program-eligible aldeas, using quantitative criteria – the aldea sample that NORC constructed through screening was largely rejected by Fintrac. At that point, the experimental design became impractical because, at the end of the evaluation, it would not be possible to state in quantitative terms the target population or scope of inference for the program. An experimental design has two randomization aspects: randomized selection of units from an identified eligible population, and randomized assignment of units to treatment. It was the first of these randomization aspects that could not be implemented for this evaluation.

At this juncture, NORC and MCC decided to move away from using an experimental design and focus on identifying an alternative evaluation design. The new design we opted to use was a causal-modeling-based design. It incorporated much of the (responding) sample from the experimental design. In particular, in addition to containing a number of the treatment farmers and aldeas from the experimental design, it contained all of the originally selected control aldeas and sample of farmers selected from them. All of the control aldeas and farmers were selected using randomization and probability sampling.

The next section presents additional detail on the revised evaluation design.

C.3 ALTERNATIVE EVALUATION DESIGN CONCEPT: A CAUSAL MODELING APPROACH

When it became apparent that implementing the original experimental design was no longer a viable option, MCA - Honduras and MCC requested that NORC propose an alternative approach to completing the impact evaluation for the FTDA. To this end, NORC proposed a causal modeling approach that would make use of data that had been collected for the original design, supplemented with additional sample data.

C.3.1 Overview and Key Considerations of the Causal Modeling Approach

The original evaluation design concept, of using an experimental design, is a *design-based* approach to impact evaluation. Our resolution to the problem described above was to use a *causal modeling* approach that would make use of almost all the data that was collected for the original design, and complement it with additional sample data. The additional sample consisted of data collected from new recruits selected by Fintrac from the sample of experimental treatment *aldeas* and a sample of Fintrac's clients, with no relation to the evaluation design, who were randomly selected from Fintrac's client lists. Baseline data were collected from these additional samples, which served to increase the sample size to achieve a satisfactory level of precision and power.

The validity of the results of an impact evaluation rests on the soundness of the causal model used to represent the system under study and the selection of sample units from it, and the soundness of the associated statistical model and estimator used for estimation of impact. The causal model underlies both the experimental design and the revised design. For an experimental design, randomized assignment to treatment assures that the treatment and control *aldea* samples are distributionally equivalent with respect to all factors that might affect outcomes of interest, except for treatment. This condition greatly simplifies the procedures used to estimate impact, and eliminates the possibility of selection bias.

For the causal modeling approach that NORC used for the FTDA evaluation, in the absence of full randomization, selection bias may 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 be different for the treatment and control samples, even after *ex post* matching[13]. The approach that NORC used in this evaluation involved the development of a "selection" model that estimates the probability of selection, or *propensity score*, of a sample unit for treatment, and the use of an impact estimator that is based on the estimated propensity score.

In a design-based approach, such as a RCT, unbiased or consistent estimates of program impact are determined by taking into account the sample design structure and the probabilities of selection of the sample units. With this approach, the estimation formulas depend on the structure of the sample design, not on the relationship of outcome to explanatory variables other than treatment. With the causal modeling approach, the estimate of program impact is based on a causal model that describes the relationship of selection and/or outcome to treatment and to other explanatory variables, and on a statistical model that corresponds to the causal model.[14] For either approach, with data available from a two-round panel survey, the usual measure of impact is the double-difference *measure* (the double-difference *measure*, not the double-difference *estimator*!). Under the two approaches, however, the forms of the impact estimate and the procedures for constructing it are quite different. For the design-based approach using a highly-structured experimental design, the impact estimate can be represented as a simple double difference in sample means of the four design samples. For the model-based approach, the estimate is more complicated, and is usually constructed using multiple regression analysis.[15] In essence, the estimation process "adjusts" the double-difference estimate, to account for differences in the distributions of explanatory variables (other than treatment) between the treatment and control samples, to reduce the chance and the magnitude of selection bias.

In its simplest form, this estimator may be represented as a function of a variety of explanatory variables (treatment variables, design parameters and other explanatory variables (covariates)):

O*utcome measure = f(treatment indicator variable, other explanatory variables)*

[13] Both methods (matching and covariate adjustment) may be used together to reduce selection bias. This approach is referred to as being "doubly robust," in the sense that if either the matching model or the covariate-adjustment model is correct, the selection bias is removed.

[14] See the cited references, especially the Lohr book, for a detailed discussion of these two approaches.

[15] For both approaches, the conceptual framework is the Neyman-Fisher-Cox-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* 2nd ed. by Judea Pearl (Cambridge University Press, 2009 (1st ed. 2000)).

As mentioned previously, with a lack of proper randomization, non-treatment variables may have different distributions for the treatment and control samples, and this difference may bias the estimate of program impact if not properly taken into account. Since the differential effect of all of these variables between the treatment and control groups has not been removed by randomization, it is necessary to adjust for them in the analytical (estimation) model. The average measure of program impact is obtained by determining a regression-equation (or other) model showing the relationship of selection for treatment and outcome to explanatory variables or of program outcome to explanatory variables, and estimating the impact from the model. The estimators used with this approach are called "model-assisted," "model-based" or "model-dependent." The impact estimate may be a coefficient in a regression model (e.g., the coefficient of the interaction of treatment and time) or it may be obtained in a different way such a "matching" estimator.

Because of the focus on causality, the approach is also called a causal modeling approach. A randomized experimental design is also based on causal modeling, but the causal model is rather trivial, since randomized assignment to treatment assures that the distributions of explanatory variables (other than treatment) are the same for the treatment and control samples.

Absent the experimental design, the simple double-difference estimator is not an unbiased or consistent estimate of impact. However, with the full range of variables that we collected for this study, we were able to develop models of the relationship of program selection and outcome to explanatory variables and thereby obtain, under certain assumptions, a consistent estimate of impact from that model. 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. This is the reason why a sample of ordinary Fintrac clients may be combined with the probability sample of clients selected for the experimental design. 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.

In order to predict the effect of implementing the Fintrac program in other settings, it is also necessary that the program intervention represent a "forced change" to (experimental control of) the agricultural system in Honduras. (This requirement follows from the fact that estimation of the effects of changes to a system must be based on data in which forced changes were made to the system.) The Fintrac program does represent a forced change to the agricultural system, so estimation of overall program impact is appropriate. No effort was made, however, to exercise experimental control over the explanatory variables (other than program participation). For this reason, the evaluation project is in a good position to estimate the overall impact of the program, but not to make assertions about the effect of making changes in explanatory variables. To make valid inferences about the effect of specific variables on impact, it is necessary to make forced changes in those variables, as in an experiment.

C.3.2 Conceptual Framework for the Evaluation

This section describes the conceptual framework for the impact evaluation. It discusses the causal model; statistical model specification; the analytical survey design; the statistical power analysis used to determine sample size; identification of parameters and effects related to impact; analysis of selection factors and variables; estimation procedures; estimation of standard errors;

test-of-hypothesis procedures; *ex post* statistical power analysis; scope of inference, external validity; and a summary of assumptions and limitations

The three subsections dealing with estimation and *ex post* statistical power analysis are very brief. Additional details will be provided for these subsections later.

A. Causal Model

A causal model shows the causal relationship among factors or variables relevant to the evaluation[16]. Causal models are useful because the nature of the relationships among entities relevant to a process under study determines what quantities (e.g., impacts of interest) may be estimated from the available data, and which model specifications and estimation procedures are appropriate. Causal models are represented in various ways, such as structural equations and directed graphs. This evaluation will employ both of these methods.

Figure 1 is a high-level causal model (entity-relationship diagram) for the FTDA Project. Each arrow of the figure indicates that a causal relationship exists between the entity at the tail of the arrow and the entity at the head of the arrow. The causal model diagram shown in Figure 1 differs from the usual causal model diagrams, which show the relationship of variables in the survey questionnaire to each other and to certain exogenous variables, such as those used to select the project and the survey sample. These latter causal models show causal relationships among survey variables, such as income and program participation. The causal model diagram shown in Figure 1 addresses entities outside of the household survey questionnaire. (The figure depicts the final evaluation design, not the original randomized-groups design.) Causal relationships within the survey questionnaire will be addressed again, when specific analytical models are considered.

[16] The term "factors" refers to higher-level constructs, and "variables" to measurable lower-level quantities. For example, "wealth" and "ambition/motivation" are factors, while "value of home," "savings," "hectares owned," "highest grade achieved," and "number of civic offices and awards" are variables. Factors typically refer to constructs that are difficult to measure and cannot be directly observed or measured, and are reflected in (associated with) a number of (measurable) variables. Causal models are specified in terms of both factors and variables.

Figure 1. High-Level Causal Model for the FTDA Project

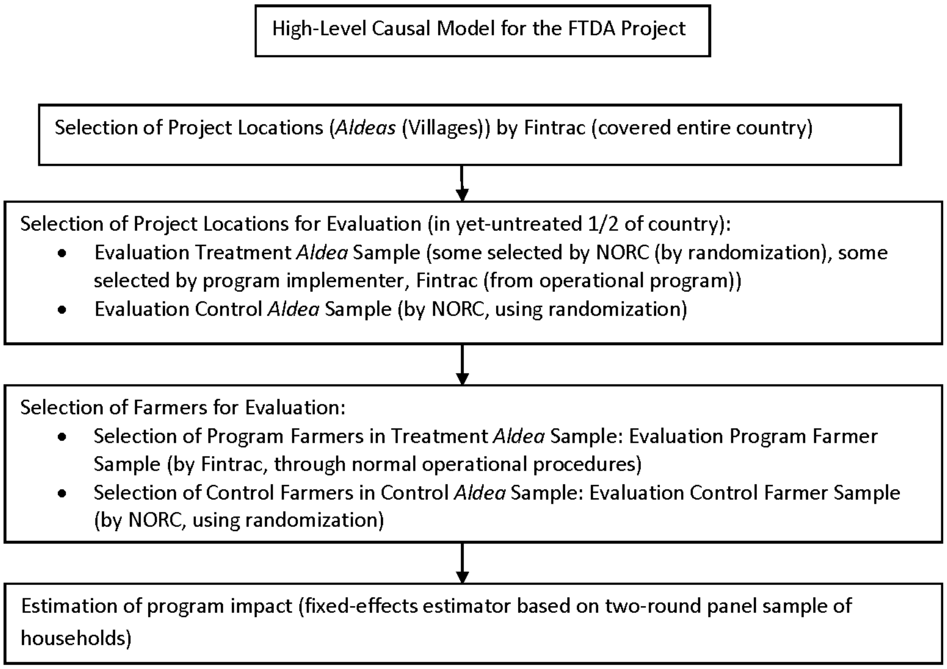

High-Level Causal Model for the FTDA Project
Selection of Project Locations (*Aldeas* (Villages)) by Fintrac (covered entire country)
Selection of Project Locations for Evaluation (in yet-untreated 1/2 of country):
• Evaluation Treatment *Aldea* Sample (some selected by NORC (by randomization), some selected by program implementer, Fintrac (from operational program))
• Evaluation Control *Aldea* Sample (by NORC, using randomization)
Selection of Farmers for Evaluation:
• Selection of Program Farmers in Treatment *Aldea* Sample: Evaluation Program Farmer Sample (by Fintrac, through normal operational procedures)
• Selection of Control Farmers in Control *Aldea* Sample: Evaluation Control Farmer Sample (by NORC, using randomization)
Estimation of program impact (fixed-effects estimator based on two-round panel sample of households)

Figure 1 is a directed acyclic graph (DAG). The directed arrows imply a time sequence. In this model, no simultaneous (nonrecursive) causal relationships are represented. The main points to note from the figure are that the evaluation is based on the half of the country that Fintrac had not yet treated when the evaluation project began, and that selection of the aldea sample is not based on randomization (as was planned in the original design). The first condition is not considered a serious limitation on the scope of inference of the evaluation, since the untreated half of the country was similar agriculturally to the treated half. The second condition can pose a serious threat to the validity of the impact estimates, because the lack of randomization may introduce selection bias into the impact estimates. The impact estimators must address this fact, to minimize selection bias.

The estimation of impact was based on fixed-effects estimators. Under this approach, the sample of aldeas and households of the sample survey are treated as nonstochastic variables. This assumption restricts the scope of inference a little, but not much (since the aldea sample covered much of the country). (The term "fixed" is somewhat ambiguous. It may indicate that a variable is nonstochastic, or is a conditioning variable, or that it is stochastic and not correlated with a model error term. While the latter is the usual interpretation in econometrics, the former is easier to understand and to justify. The term may also refer to making a forced change in a variable (as in Judea Pearl's "do" calculus), rather than statistical conditioning. The intended meaning of the term is important, and should be clear from context. Here, we use the term to indicate that the scope of inference is restricted to the sampling units (aldeas and households) of the household survey, and that the effects associated with aldeas and households are not stochastic, given the sample.)

Within a household, mutual (simultaneous, nonrecursive) causal relationships exist. For example, household income may affect ownership in farm equipment, and vice versa. A graphical representation that includes simultaneous causal relationships would be a directed cyclic graph (DCG), not a directed acyclic graphs (DAG). When particular estimating equations are considered later, mutual causal relationships will be described. Some models (e.g., a model containing only treatment and survey round) may contain only fixed effects, and are "fixed-effects" models. Other models, however, may contain household variables that may be considered to be random variables and may be correlated. For example, household income and size of land farmed may be considered to be endogenous (mutually causally related). Such models (containing both fixed effects and random effects) are "mixed-effects" models. For specific models involving questionnaire variables, explanation will be provided about model specification and identification, additional to the description provided by the preceding high-level causal model.[17] [18]

The project represents a "forced change," or intervention, to the agricultural system in Honduras. This intervention was not decided by randomization, but it is a forced change nonetheless, caused by MCA – Honduras and the program implementer. (Since the program intervention is a forced change, it is an experiment. It is not an experimental design based on randomized assignment to treatment, but it is an experiment, nonetheless. This is an important consideration. In order to estimate the effect of making a change to a system, it is necessary to base the estimate on data for which a forced change was made to the system. Were it not for the forced changes represented by the program intervention, the analysis would be based on "observational data.") Although randomization was not used to select the treatment aldeas, causal modeling may be used to estimate the effect of applying the program intervention to program-eligible farmers in a randomly selected aldea.

Causal Model Diagrams

Causal models may be specified in different ways, such as by structural equations or by directed graphs (as described in Pearl op. cit. and Morgan and Winship op. cit.). A comprehensive theory of causal modeling has been developed by Judea Pearl for directed acyclic graphs (DAGs),

[17] For more discussion of causal modeling, refer to *Causality: Models, Reasoning and Inference*, 2nd ed by Judea Pearl (Cambridge University Press, 2009, 2000). For a summary, see *Counterfactuals and Causal Inference: Methods and Principles for Social Research* by Stephen L. Morgan and Christopher Winship (Cambridge University Press, 2007). See also "Statistics and Causal Inference" by Paul W. Holland (*Journal of the American Statistical Association*, Vol. 81, No. 396 (Dec., 1986)). The theory of structural equation modeling is presented in *Linear Causal Modeling with Structural Equations* by Stanley A. Mulaik, (Chapman & Hall / CRC Press, 2009), *Nonrecursive Causal Models* by William D. Berry (Sage Publications, 1984); *Introduction to Structural Equation Models* by Otis Dudley Duncan (Academic Press, 1975); and *Structural Equation Models in the Social Sciences* by Arthur S. Goldberger and Otis Dudley Duncan (Seminar Press, 1973).

[18] The assumption of whether to use fixed-effects, random-effects, or mixed-effects models makes a substantial difference in the scope of inference and in testing of hypotheses. These notions are discussed at length from an econometric viewpoint in *Econometric Analysis of Cross Section and Panel Data* 2nd ed. by Jeffrey M. Wooldridge (MIT Press, 2010, 2002). From a statistical-analysis viewpoint, references on this topic include *Generalized, Linear and Mixed Models* 2nd ed. By Charles E. McCulloch, Shayle R. Searle and John M. Neuhaus (Wiley, 2008); *Variance Components* by Shayle R. Searle, George Casella and Charles E. McCulloch (Wiley, 1992, 2006); and *Linear Models* by S. R. Searle (Wiley, 1971).

which correspond to situations in which variables are not mutually causally related (i.e., to causal models that are recursive). Pearl describes criteria (such as the "back door" criterion) that a causal graph must satisfy for a causal effect to be identifiable (i.e., estimable – may be estimated from data on the variables represented in the graph).

The following figure (adapted from Morgan and Winship op. cit.) presents examples of DAGs for the statistical and econometric approaches. For these graphs, Y denotes an outcome of interest, W denotes treatment, X denotes the set of all variables other than W that affect outcome (i.e., are direct causes of Y), S denotes all variables that affect selection, Z denotes an observed subset of S, and U denotes unobserved variables. In the graphs, a solid directed arrow signifies a causal relationship, and a dashed arrow indicates that the endpoints are affected by common variables. (In the last figure, the term "fix" is used to mean physically (or in a "thought experiment," mentally) setting the value of a variable (as in Pearl's "do" calculus), rather than statistically conditioning on it.)

If selection is based on observed variables (which may or may not affect outcome), the effect of W on Y can be estimated (i.e., an unbiased or consistent estimate is available). If there exist unobserved variables that affect both W and Y, the effect of W on Y cannot be estimated (or, more accurately, a good (unbiased or consistent) estimate of the effect of W on Y is not available), without making certain assumptions about the distribution of the unobserved variables (such as time-invariance of unobserved variables that affect both selection and treatment in a two-round panel study). The purpose of constructing causal model diagrams is to assist determination of whether the effect of W on Y is estimable (i.e., is "identified").

Note that interest focuses on unobserved variables that affect both selection and outcome. If a variable affects selection but has no effect on outcome, then it is not relevant (e.g., if selection is based on eye color, and eye color has no effect on outcome, then it may be ignored). In later discussion, we will make reference to the penultimate panel of Figure 2, Figure 2d. This panel illustrates the fact that unobservable variables affect both selection for treatment (W) and outcome (Y). We shall seek conditions on Z, X, and U under which it is possible to obtain an unbiased (or consistent) estimate of impact (the effect of W on Y).

Figure 2. Examples of Causal Diagrams

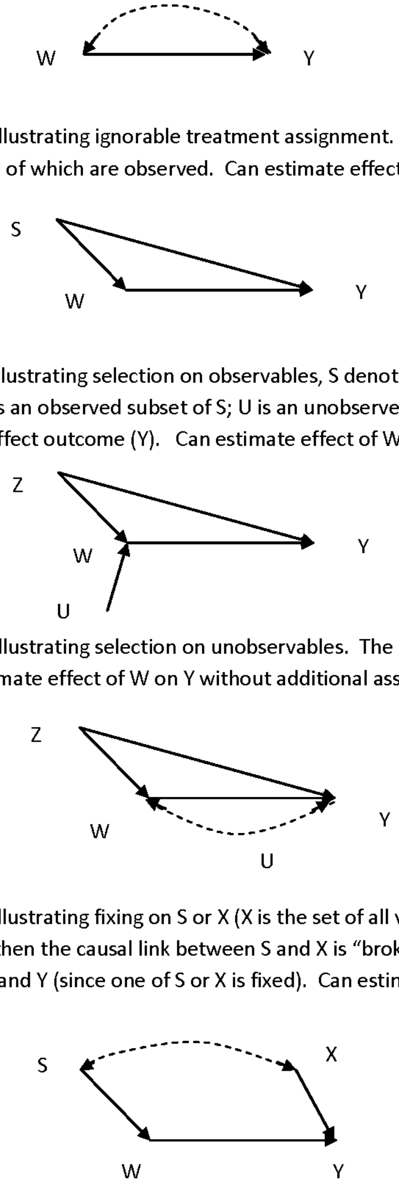


Figure 2a. Causal diagram illustrating nonignorable treatment assignment. W denotes selection for treatment. Y denotes an outcome of interest. Both W and Y are affected by unobserved variables (dashed line). Cannot estimate effect of W on Y (i.e., cannot construct a good (unbiased or consistent) estimate of the effect of W on Y).

Figure 2b. Causal diagram illustrating ignorable treatment assignment. S denotes all variables that affect selection for treatment, all of which are observed. Can estimate effect of W on Y.

Figure 2c. Causal diagram illustrating selection on observables, S denotes all variables that affect selection for treatment; Z is an observed subset of S; U is an unobserved subset of S. The unobservables, U, do not affect outcome (Y). Can estimate effect of W on Y.

Figure 2d. Causal diagram illustrating selection on unobservables. The unobserved variables may affect both W and Y. Cannot estimate effect of W on Y without additional assumptions about the distribution of U.

Figure 2e. Causal diagram illustrating fixing on S or X (X is the set of all variables other than W that affect outcome (Y)). If fix S or X, then the causal link between S and X is "broken" – there are no unobserved variables affecting both W and Y (since one of S or X is fixed). Can estimate effect of W on Y (fixing either S or X).

In order to obtain an unbiased estimate of a causal effect, it is necessary to establish the reasonableness of the assumption of conditional independence of the counterfactuals and treatment, given the covariates. Unfortunately, there is no statistical test for this. The standard approach is to specify a causal model in terms of general causal factors, and assess the extent to which those factors are represented by observable variables (i.e., variables from a survey questionnaire or other data sources). If one or more important factors are not represented by, or poorly represented by, observable variables, an assessment must be made about the effect that this situation will have on the impact estimate. In some cases, it is possible to construct a causal estimate even with selection on unobservables, under certain conditions (i.e., if certain assumptions are made). In some cases, it is possible to construct bounds on the estimate by making assumptions about the magnitude of the effect of the unobservables on selection.

Causal Modeling Theory Relevant to the FTDA Evaluation

The detailed causal model used in this evaluation is referred to as a "Neyman-Fisher-Cox-Rubin" causal model, or potential outcomes model, or counterfactuals model. Under this conceptual framework, each sample unit (household) is considered to possess two alternative possible outcomes, conditional on the project intervention. From the viewpoint of estimating impact, a difficulty associated with this approach is that for a particular sample unit, only one of the two outcomes can be directly observed – whichever is observed is a "counterfactual" for the other. It is therefore not possible to observe an estimate of impact for a single sample unit (household). The estimate of impact is obtained by comparing *groups* of similar individuals under the alternative treatment specification (project intervention or no intervention). With randomized assignment of treatment, it is straightforward to construct unbiased estimates of impact. In the absence of randomization, the properties of impact estimates depend on assumptions made about the model (such as conditional independence of the counterfactual responses and treatment, given the values of covariates), and the estimators are more complicated (e.g., matching estimators and regression-adjusted estimators).[19]

Alternative Approaches to Estimation of Causal Effects

As mentioned, estimation of impact is based on causal modeling, specifically, on the "potential outcomes" (or "counterfactuals") approach. Under this approach, causal relationships are specified in a causal model, and statistical estimations of impact are derived from a statistical model that corresponds to the causal model. The existence of causal relationships among variables is specified in the causal model. No inferences about the existence of causal relationships are derived from the statistical model – only from the causal model. The statistical model estimates the magnitude of causal effects that are specified in the causal model. By themselves, the statistical estimates simply assess the strength of associational relationships. They reflect causal relationships if the causal model is correct, the statistical model is also correct (i.e., correctly specified), the parameters / effects of interest are identified (estimable), and the estimation procedures are correct [20].

There are two slightly different approaches to causal inference (estimation of causal effects) based on counterfactuals, sometimes referred to as the "statistical" approach and the

[19] See Wooldridge, op. cit.

[20] A comprehensive description of causal modeling is presented in the book, Causality: Models, Reasoning, and Inference, 2nd edition, by Judea Pearl (Cambridge University Press, 2009. 1st ed. 2000). A summary of basic concepts that relate to estimation of impact is presented in the text, Counterfactuals and Causal Inference: Methods and Principles for Social Research, by Stephen L. Morgan and Christopher Winship (Cambridge University Press, 2007). References on structural equation modeling include Linear Causal Modeling with Structural Equations by Stanley A. Muliak (Chapman & Hall / CRC Press, 2009); and Introduction to Structural Models by Otis Dudley Duncan (Academic Press, 1975).

"econometric" approach[21]. The statistical approach is useful for assessing overall impact under a minimum of assumptions about outcome. The econometric approach is useful for estimating not only overall impact, but also for estimating the relationship of outcome to explanatory variables. Both of these approaches were used in the analysis of impact done in this project[22].

B. Analytical Survey Design

The survey design for this evaluation was an analytical survey design. This type of survey design differs substantially from the survey designs used for descriptive sample surveys. The purpose of descriptive sample surveys is to estimate overall characteristics of a population or subpopulations of interest, such as means, proportions and totals. The purpose of analytical sample surveys is to collect data to enable the construction of analytical models, such as a model that estimates the impact of a program intervention, or of the relationship of impact to explanatory variables.[23]

In a descriptive survey, it is generally attempted to keep the sample selection probabilities as uniform as possible, subject to achieving high precision for estimates for overall population characteristics. For an analytical survey, it is attempted to achieve adequate variation in explanatory variables that are considered to have an important relationship to outcomes of interest, so that those relationships may be estimated with high precision. These two designs are

[21] The statistical approach is described in the articles "The central role of the propensity score in observational studies for causal effects," by Paul R. Rosenbaum and Donald B. Rubin, *Biometrika*, vol. 70, no. 1, pp. 41-55 (1983); and "Statistics and Causal Inference," by Paul W. Holland, *Journal of the American Statistical Association*, Vo. 81, no. 396, pp. 945-960 (1986). The econometric approach is described in a number of books and articles by James J. Heckman and others, including *Causal Analysis after Haavelmo* by James J. Heckman and Rodrigo Pinto, Working Paper 19453, http://www.nber.org/papers/w19453 , National Bureau for Economic Research, September 2013; *Longitudinal Analysis of Labor Market Data*, edited by James J. Heckman and Burton Singer, Econometric Society Monographs, Cambridge University Press, 1985; "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," by James J Heckman and V. Joseph Hotz, *Journal of the American Statistical Association*, Vol. 84, No. 408, pp. 862-874 (1989); *Handbook of Econometrics, Volume 6B* (of Handbooks in Economics 2), Editors James J. Heckman and Edward E. Leamer, North-Holland / Elsevier, 2007; "Matching As an Econometric Evaluation Estimator," by James J. Heckman, Hidehiko Ichimura and Petra Todd, *Review of Economic Studies*, vol. 65, pp. 261-294 (1998); *Econometrics Counterfactuals and Causal Models*, Keynote Address, International Statistical Institute, Seoul, South Korea, August 27, 2001, by James J. Heckman; "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," by James Heckman and Salvador Navarro-Lozano, *The Review of Economics and Statistics*, February 2004, vol. 86, no. 1, pp. 30-57; "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," by James Heckman, Justin L. Tobias and Edward Vytlacil, *The Review of Economics and Statistics*, Vol. 85, no. 3, pp. 748-755 (2003); "The Scientific Model of Causality," by James J. Heckman, *Sociological Methodology*, vol. 35, pp. 1-97 (2005); and "Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs," by Arild Aakvik, James J. Heckman and Edward J. Vytlacil, *Journal of Econometrics*, vol. 125, pp. 15-51 (2005).

[22] Both of these approaches (statistical and econometric) are represented in the impact estimators presented in the book, *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, by Jeffrey M. Wooldridge (The MIT Press, 2010, 1st ed. 2002)

[23] 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.)

often very different. The selection probabilities for an analytical design usually vary substantially more than for a descriptive survey design.

Our sample design took into account *caserío* data that were available from Honduras Census and geographic information system sources[24]. These data included elevation, climatic zone, soil capacity, rainfall, temperature, vegetation cover, protected area status, distance to nearest major river, population, and a variety of travel times to point of interest. We used these data to construct an analytical survey design that had substantial variation on these variables. We used a two-stage sample design, with marginal stratification on the variables just listed. Marginal stratification was implemented using the method of variable probabilities of selection. Some additional information on the sample survey design is presented in the next section. A detailed description of the sample design is presented in Annex 3.

C. Statistical Power Analysis Used to Determine Sample Size

For a two-stage sample design, there are two sample sizes – the sample size for the first-stage sample units (primary sample units, PSUs; in this case, aldeas), and the sample size for the second-stage units, or households, within each first-stage unit. We determined the sample size for households by taking into account the relative cost of sampling aldeas vs. households, and the intra-unit (intra-aldea) correlation coefficient. We determined that an efficient household sample size was about 20 households per aldea. A detailed description of how this per-aldea household sample size was determined is presented in Annex 3.

We conducted a statistical power analysis to determine an aldea sample size that would achieve high power for detecting effects (change in the double-difference measure) of specified size. The minimum detectable effect sizes were varied over a range, from an effect equal to .25 times baseline income to 1.0 times baseline income. We determined a sample size, using statistical power analysis that would detect impacts in this range with high probability (power), for outcome variables having typical ranges of values for the coefficient of variation and intra-unit (*aldea*) correlation coefficient. Annex 3 presents a summary of the statistical power analysis.

D. Statistical Model Specification

Below we present a brief summary of essential aspects of the counterfactuals approach to estimation of impact. For ease of discussion, we first discuss the simpler case of a one-round cross-sectional survey, before discussing the two-round survey of the present application.

We denote the response (observation, measurement) for the i-th unit of an experiment or observational study as Y_i. Also, denote the two potential ("counterfactual") outcomes for unit i as Y_{1i} and Y_{0i}, and the event that the i-th unit receives treatment as W_i, where $W_i=1$ signifies that the i-th unit was treated and $W_i=0$ signifies that the i-th unit was not treated. Then, in terms of the observed outcome, we have:

$$Y_i = Y_{1i} \text{ if } W_i=1$$

[24] *Caseríos* are the smallest local-level governmental administrative unit in Honduras. The country is hierarchically divided into 18 departments, 298 municipalities, 3,721 *aldeas*, and 27,969 *caseríos*. The *caserío* data were aggregated to obtain *aldea* data.

and

$$Y_i = Y_{0i} \text{ if } W_i=0$$

or

$$Y_i = Y_{0i} + (Y_{1i} – Y_{0i})W_i.$$

Note that we may observe only one of the two counterfactual outcomes. In discussing concepts, for simplicity, we shall drop the index i that refers to the individual unit (subject), and write

$$Y = Y_0 + (Y_1 – Y_0)W.$$

The index i is used in formulas that refer to the individual units.

Two quantities of interest are the average treatment effect, or ATE = $E(Y_1 – Y_0)$ and the average treatment effect on the treated, or ATT = $E(Y_1 – Y_0|W=1)$. The ATE is the mean effect of treatment over all units in a population of interest (whether treated or untreated, e.g., all persons eligible for a program), and the ATT is the mean effect over all units conditional on selection for treatment. (In evaluation research, the populations of interest are often conceptually infinite.) If W is independent of (Y_0, Y_1) (slightly weaker conditions suffice) then both of these quantities (ATE and ATT) are equal, and may be estimated as the difference in means between the treated and untreated samples. ATE and ATT are referred to as “causal effect estimates” or “causal estimates” (although the term “causal” is sometimes reserved for use with (randomized-assignment-of-treatment) experimental designs, since causality can be established with a high degree of certainty only in this case (forced randomized assignment of treatment levels)).

In the case of a single time period (cross-sectional data), the ATE and ATT are “single difference” estimates of impact (where the difference is between treated and untreated groups, not a difference over time).

E. Identification of Parameters and Effects Related to Impact

Statistical Approach to Estimation of Causal Effects (Rosenbaum-Rubin Approach)

Using the statistical approach to causal analysis, conditions are sought under which the counterfactual responses Y_0 and Y_1 are independent of treatment, W. The conditions are specified in terms of the values of a random variable, X (referred to as a covariate). In general , X is a vector. The conditional independence of treatment and (counterfactual) response, given X, is symbolically denoted as $(Y_0, Y_1) \perp W |X$. (Conditional independence of treatment and the counterfactuals (given X) is also referred to as “ignorability of treatment,” or “selection on observables”.) If covariates X can be found under which the counterfactual responses are independent of treatment, then an unbiased estimate of impact may be obtained by taking the expectation of $Y_1 – Y_0 |X$ over the covariate, X. In the most basic application of the method, the sample is stratified into subsets for which the distribution of covariates is similar for the treatment and control samples. The problem that arises immediately in most applications is that X is a vector, and it is not practical to find sets of values of X for which conditional independence holds. Rosenbaum and Rubin showed that the counterfactual responses are conditionally independent, given the propensity score, or probability of selection for treatment (or, in the present application, probability of participation). This fact simplifies the problem tremendously, since the conditioning may now be done on a scalar rather than a vector.

With randomized assignment to treatment, the distributions of all variables other than treatment are the same for the treatment and control samples. Conditioning on the propensity score achieves this same condition. For this approach to be useful, it is necessary to know the propensity score. In practice, the (true) propensity score is not known, but is estimated from the data. The validity of the method then depends on whether (or the extent to which) variables that affect treatment selection are known, and whether (or the extent to which) the model relating the propensity score to those variables is correct (i.e., correctly specified).

In practical applications, the propensity score is usually estimated using a logistic regression model (or other generalized linear statistical model, such as a probit model). This model is called the "selection" model. The validity of the approach rests on whether the accuracy (validity and reliability) of the estimated propensity score, i.e., of the selection model. If the model is correct, then the counterfactuals and treatment are conditionally independent, and a correct estimate of impact is obtained. The validity of the logistic regression model rests on statistical properties of the model error terms (residuals), such as whether they are independent and identically distributed and are uncorrelated with the explanatory variables of the model. A major concern is whether the model error term includes unobserved variables (hidden variables) that are correlated with the explanatory variables of the model. If it does, the method fails.

There are alternative approaches to estimating impact using the statistical approach, including stratification of the observations with respect to the estimated propensity score, or using the estimated propensity score in inverse-probability weighting.

The statistical approach is sometimes referred to as a "balancing" method (or "conditioning to balance"), since it identifies subsets of observations for which the distributions of covariates (variables other than treatment) are similar within each subset (i.e., are "balanced"). The propensity score is a "balancing score." The econometric approach is referred to as an "adjusting" method (or "conditioning to adjust"), since it provides methods to accommodate differences in the distributions of covariates in the treatment and control samples (i.e., it does not rely on stratification into subsets for which the distributions of covariates are similar for the treatment and control samples).

Econometric Approach to Estimation of Causal Effects (Heckman Approach)

The econometric approach differs from the statistical approach in that it includes consideration of an "outcome" model (that relates response to explanatory variables), as well as the selection model. The outcome model may differ for the two counterfactual responses. Since the same selection variables are available to both the statistical and econometric approaches, the selection model is the same in both approaches (since the causal model underlying it is the same). If the models are correctly specified, both approaches produce similar estimates of impact. A major difference is that, in addition to estimating overall impact (e.g., the average treatment effect), the econometric approach also estimates the relationship of outcome to the explanatory variables included in the outcome model. This information is useful for policy analysis since it may be useful to know the effect of making changes to policy variables. Developing such models from observational data is fraught with problems. As mentioned earlier, in order to estimate the effect of making changes to a system, it is necessary to develop the model from data in which forced changes were made to the system. An additional problem is associated with the fact that many of the explanatory variables are correlated, so the effect estimates (regression coefficients associated with the variables) are confounded. If it is not possible to make forced changes to important explanatory variables of interest, or if such variables are correlated, use of the econometric approach may be counterproductive, since it requires additional assumptions, which may not be justified.

Advocates of the econometric approach assert that it is easier to assess model validity if the relationship of the causal variables to potential outcomes is explicit. To this end, the following model structure is often assumed[25]. In the econometric approach, models are specified for selection and for outcome. In the case in which linear statistical models are used, these models may be represented as follows.

Outcome Model:

$$Y = X\beta + W\delta + U$$

where β denotes a vector of parameters and X denotes a vector of explanatory variables. (In this section, we drop the convention of using boldface font for vectors.) U denotes a model error term, for which $E(U) = 0$. It is assumed that X and U are uncorrelated, so that $E(U \mid X) = 0$. When selection is nonrandom, W and U may be correlated (given X), in which case $E(U \mid W, X) \neq E(U \mid X) = 0$. In this case, $E(Y \mid W, X) = X\beta + W\delta + E(U \mid W, X) \neq X\beta + W\delta$. In this case, the standard estimation procedures (ordinary least squares, OLS) fail to produce an unbiased or consistent estimate of δ (the impact).

Dependence may arise between W and U for a number of reasons. In the present application, the source is selection by Fintrac (in offering the program services to a farmer, conditional on his passing the screening tests) and selection by the individual farmer (i.e., his decision to participate). (Note that the various eligibility criteria, such as access to water, are not necessarily factors affecting selection for treatment. "Selection for treatment" refers to the eligible population defined by the eligibility criteria, not to construction of the eligible ("in-scope," target) population.)

The issue to be faced here is how to construct a good estimate of δ for the preceding model. This is done by taking into account information from the selection model.

Selection Model:

[25] There is a variety of estimators in the econometric approach. The method used here, which is based on propensity scores, is described on pp. 920-927 of Wooldridge op. cit.

The standard model of selection is a logistic regression model (or other generalized linear statistical model), described as follows. This model is the same as the selection model discussed for the statistical approach.

The model, a binary response model, is as follows:

$$P(y=1|\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta}) \equiv p(\mathbf{x})$$

where **x** denotes a (column) vector of explanatory variables, P(y=1|**x**) denotes the probability that y=1 (i.e., is treated) conditional on **x**, **β** is a vector of parameters and g(.) is a the logistic link function,

$$g(z) = \exp(z)/(1 + \exp(z)).$$

If we define z as

$$z = \mathbf{x}'\boldsymbol{\beta} + e,$$

where e denotes a random error term uncorrelated with **x** and with mean zero, then

$$y = 1 \text{ if } g(z)>.5 \text{ and 0 otherwise.}$$

The expression **x'β** is referred to as an index. The variable z is called a latent variable (since it is unobserved). The parameters **β** are estimated by the method of maximum likelihood. The expression **x'β** does not have any economic meaning (or units) – it is simply a modeling artifact.

The assumptions required to obtain an unbiased estimate of impact are as follows. Z is assumed to be independent of (U_W, Y_0, Y_1). (U_W, U_0) and (U_W, U_1) are assumed to be independent of (Z, X) (or of Z given X). It is assumed that ZβW is is nondegenerate, i.e., takes on more than a single value. For these (independence and nondegeneracy) conditions to hold, it is required that the exclusion restriction holds, that there is at least one variable that affects participation that does not directly affect outcome. What is required to identify (estimate) the treatment effect is that the distributions of the unobservables (that affect both selection and outcome) be the same for the treatment and control samples, given Z.

There are a number of ways of taking into account the results of the selection model to obtain an estimate of δ in the outcome model. One method is to note that, since given Z and X, W is uncorrelated with U, it follows that E(U | W, X, Z) = E(U | X, Z), so that E(Y|W, X, Z) = Xβ + Wδ + E(U | W, X, Z) = Xβ + Wδ + E(U | X, Z). A functional form for E(U | X, Z) may be specified, and OLS applied to estimate the regression equation Xβ + Wδ + E(U | W, X, Z) and hence obtain δ (the exclusion restriction stated earlier is necessary for δ to be estimable). This method is called the control function method (since E(U| X, Z) is called a control function).

The method we use to estimate δ is to note that, under the assumption of conditional independence (selection on observables or selection on unobservables under certain circumstances, such as time-invariance), selection for treatment and the counterfactual outcomes are independent, given the propensity score (which is based on the observables, Z). In this case, an estimate of impact is obtained from the regression equation Xβ + Wδ + $\beta_p\hat{p}(\boldsymbol{x})$, or a more general form, Xβ + Wδ + $\beta_p\hat{p}(\boldsymbol{x}) + \beta_{pw}(\hat{p}(\boldsymbol{x}) - \hat{p})$, where $\hat{p}$ is a consistent estimate of ρ = E[p(**x**)] = P(W=1).

Selection on Unobservable

If the dependence between U and W is not eliminated by controlling for Z, selection depends on unobservables. In this case, $E(U|W, X) \neq 0$ and $E(U|W, X, Z) \neq E(U|X, Z)$. To construct unbiased or consistent estimates of impact in this situation requires assumptions about the distributions or moments of U, U_W and Z. The approach that we shall use (see Heckman and Hotz op. cit.) is to use fixed-effects (or first-difference) estimators for the two-round panel data. In this case, all time-invariant unobservables drop out of the model, and an unbiased or consistent estimate of impact (δ) is obtained. We are assuming that $E(U_{t1} – U_{t0}|W, X) = 0$ (where t_0 and t_1 denote the times of the survey rounds). This holds if we assume the following model for U: $U_t = \varphi + V_t$ where φ is a zero-mean fixed effect (e.g., a household or aldea characteristic) and V_t is a zero-mean random component independent of φ and $V_{t'}$ for $t \neq t'$.

In the present application, the preceding approach to selection on unobservables applies, since the unobservables are time-invariant variables, which drop out of the two-round fixed-effects regression model. It is emphasized that the validity of the impact estimates depends on the assumption that the variables affecting selection for treatment and outcomes of interest are observable or, if not, are time-invariant. This issue will be addressed shortly, by constructing a list of factors and variables considered to affect selection for treatment, and indicating those that are unobserved. The validity of the impact estimates rests on an assessment of the degree to which unobserved variables affecting selection and outcome are time-invariant.

For both approaches (statistical and econometric), it is assumed that the program is sufficiently small that no “macro” economic effects (general equilibrium effects) manifest (i.e., that the partial equilibrium assumption, or stable unit treatment value assumption (SUTVA) is valid). The FTDA project treated about 6,000 farmers across the nation. It is arguable whether the program is likely to cause general equilibrium (macroeconomic) effects.

The ability to test the adequacy of the outcome (counterfactual-response) models is a major difference between the econometric approach and the statistical approach. For the statistical approach, it is necessary to make the assumption of conditional independence of the counterfactual responses and treatment. The validity of this assumption rests on the degree of belief about the adequacy of the selection model. This assessment rests on consideration of the causal variables that have been included in the selection model. This same assessment must be made for the econometric approach. The econometric approach, however, extends the statistical approach to include consideration of causal and statistical models for the counterfactual outcomes. In doing this, it is necessary to be explicit about the nature of the selection process, and how it relates to outcome. There is a tendency in applying the statistical approach to simply assume that conditional independence applies. With the econometric approach, the selection and outcome models are specified in detail. The econometric approach is a more general approach than the statistical approach for two reasons: (1) it provides models of the relationship of outcomes to explanatory variables, which may be useful for policy analysis (if forced variation is imposed on the policy-relevant variables); and (2) by considering specific details of the selection and outcome models, there is greater assurance of the validity of impact estimates based on them. A disadvantage of the econometric approach is that it requires more assumptions. If all that is wanted is an overall estimate of impact, it can be obtained more easily from the statistical approach. If data are not available for which forced changes were made to explanatory variables of interest, there may be little point to implementing the econometric approach.

In the estimation of impact that follows, we employed both the statistical and econometric approaches. Several estimators of the average treatment effect were considered in the analysis presented in Annex 1, but only one is presented in the main text. Three of these estimates are discussed in detail in Annex 1. These include (1) a basic propensity-score based estimator (i.e., a "statistical" approach); (2) a regression-adjusted propensity-score-based estimator; and (3) a "modified" regression-adjusted propensity-score based estimator. (The last two are examples of the "econometric" approach. They represent simple versions of the econometric approach, since the primary covariate included is the estimated propensity score. It is for this reason that the estimators are referred to as "propensity-score-based" estimators.) For the second estimator listed, the regression models used to describe the counterfactuals are similar, i.e., there is allowance for dependence of outcome on the propensity score (as a main effect), but no allowance for heterogeneity of response (differences in the dependence (model) of the counterfactual outcomes on explanatory variables other than treatment) for the two counterfactual responses. The third estimator allows for the inclusion of interaction terms (i.e., for a heterogeneous response to treatment).

F. Analysis of Selection Factors and Variables

For both approaches, the fundamental issue to be addressed is whether all of the factors S that affect selection (participation) are represented in the set of variables Z (the observables that are considered to affect selection), or whether some selection factors are not observed. In the first case (selection on observables), it is required that all variables affecting selection be observed. In the second case (selection on unobservables), it is required that the unobserved factors be time-invariant (so that they drop out of fixed-effects models for two-round panel data). In this application, the second estimator applies (selection on unobservables).

We shall now identify factors that are considered to be involved in selection, and observable and unobservable variables that affect those factors. Note that many of these factors are the same as the factors involved in determination of program eligibility (e.g., access to water, size of land holding, financial status). While many of these factors were not recorded in the questionnaire (i.e., are not observed), they are either *aldea* characteristics, household characteristics, or farmer characteristics which do not, or are unlikely to, change between survey rounds, in which case they drop out of the fixed-effects panel models.

The list that follows pertains to factors and variables relating to *selection*. A list of factors and observable variables related to *outcomes* of interest is not presented, since it is considered to be similar to the list for selection.

The DAGs that were presented earlier are simple because they involve only basic factors. Each factor is represented by a large number of subfactors and a large number of observable variables. It is cumbersome to represent large numbers of variables in DAGs, and so the factors represented in the DAGs will be discussed in terms of lists rather than graphs.

Table 2. Factors, Subfactors and Observable Variables Affecting Selection

Factor	Subfactor	Observable Variables (in questionnaire)
Land quality and quantity	Land size, value.	Hectares owned, hectares rented
	Soil depth	Observed in assessment of eligibility.
	Slope	Observed in assessment of eligibility.
Access to water	Distance from home	Observed in assessment of eligibility. On property, outside property <100 m, outside property >100 m.
Access to markets	Road type; travel times	Road type by seasonal availability. Travel times to places of interest.
Income	Monetary and in-kind.	Amounts by source (e.g., salary, labor, crops (by type), rent, remittances, pension). Employment by type and amounts earned.
Expenses	Expenses	Household (many categories, e.g., food (by type), clothing, transportation, medical); factor expenses (labor, fertilizer, seed, insecticide, herbicide, transport) by crop type.
Assets	Able to invest L70,000 per ha	Observed in assessment of eligibility.
	House	Numerous variables (e.g., construction type, number of rooms, type of roof).
	Household furnishings	By type (e.g., refrigerator, television, telephone, truck, motorcycle, bicycle)
	Equipment	By number and type (e.g., truck, tractor, draft animals, pups, harvesters)
	Livestock	Numbers, by type.
	Other wealth	Type of lighting, wells, water tanks, silos, drying patio.
Access to labor	Hired manual labor	By number and type.
Other resources	Family and social network	Family size, characteristics of family members (age, education, income, employment (by sector)); marital status.
	Access to credit	
	Technical assistance	By source and type.
	Remittances	
Personal attributes	Intelligence	Unobserved variable (may be reflected in assets)
	Education	Level of education for each household member.
	Health	Loss of work for health and disability reasons; number of visits and payments to medical facilities, by type.
	Enthusiasm	Unobserved variable (may be reflected in assets)
	Industry	Unobserved variable (may be reflected in assets)
	Discipline; ability to follow instructions	Eligibility test administered by Fintrac; unobserved variable.
	Acquisitiveness / motivation / skills for success	Unobserved variable (may be reflected in assets)
	Experience	Years of employment
	Willingness and ability to recruit beneficiary farmers	Unobserved variable.
	Motivation, ability to learn and grow (potencial para crecer), and willingness to follow program requirements	Unobserved variable
Awareness of program	Fintrac presence	Fintrac does not operate in a community unless a minimum number of beneficiary farmers is

Table 2. Factors, Subfactors and Observable Variables Affecting Selection

Factor	Subfactor	Observable Variables (in questionnaire)
		available.
Social pressure	Community size	Population.
Perception / anticipation of benefits	Fintrac outreach	Unobserved variable.

The preceding is a comprehensive list of factors considered in the course of the analysis. The variables listed are not unique. The questionnaire contains hundreds of related variables. The ones listed above are representative examples of the variables associated with the factors

A particular causal factor is in general reflected in a number of questionnaire variables. These variables are correlated, and there is no unique model representation. The approach to developing a selection model is to identify a small selection of observable variables that, as a whole, represent the factors that are considered important to selection. Most of the models finally considered include a small number of variables.

We emphasize that the validity of the impact estimates constructed in this project rests on the validity of the assumptions made about the selection and outcome models, viz., that (for the statistical approach) the counterfactual responses are independent of treatment, given the covariates, or that (for the econometric approach) the error terms of the counterfactual outcome models are independent of the error terms of the slection-for-treatment model. The reasonableness of the selection model is assessed by examining the explanatory variables included in the model, and judging whether they reflect all important factors that Fintrac may have taken into consideration is selecting farmers to participate in the program, and farmers may have taken into account in agreeing to participate. For variables that are unobserved, the key consideration is whether they are time-invariant.

The significant aspect to observe from the preceding table is that the unobserved variables listed in the table may reasonably be assumed to be time-invariant. Most of these are innate farmer characteristics or program characteristics that are unlikely to change or change very much over the term of the study. In any event, it is a key assumption of the analysis that unobserved variables that affect both selection and outcome are time-invariant over the course of the project. One selection variable for which this assumption may be arguable is “Fintrac outreach.” If the procedures used by Fintrac to select clients were to have changed substantially over the course of the project, and this variable has a substantial effect on outcomes of interest, then the impact estimates would be biased (relative to the goal of evaluating the original program) – in effect, the original program would have evolved to a different program.

In addition to the key model assumption that unobserved variables that affect both selection and outcome are time-invariant over the study term, the other major assumption affecting the validity of the results is the stable unit treatment value assumption (SUTVA) (which implies, for example, that the program is not so large that it could cause, for example, a market glut of vegetables, leading to depressed prices).

G. Estimation Procedures

Estimation of regression model parameters were done using procedures in the Stata statistical software package, such as *regress* (for single-round (cross-sectional) regressions), *logit* (for logistic regression models) and *xtreg* (for two-round regressions). These procedures estimate the parameters of a general linear model (e.g., regression model) or generalized linear model (e.g., a logistic regression model).

H. Estimation of Standard Errors

We estimated standard errors for all impact estimates in two ways – by using closed-form formulas or simulation (resampling, "bootstrapping"). All standard errors were estimated using Stata procedures.

I. Test-of-Hypothesis Procedures

The procedures for making tests of hypotheses are substantially simplified by the fixed-effects assumption. The reason for this is that the standard errors of the effects of interest are substantially simplified. Because of the large sample size of the household survey, use of the normal approximation is appropriate, and tests of the hypothesis of zero impact are based on a z statistic (the estimate divided by its estimated standard error).

J. Ex Post Statistical Power Analysis

It turned out that in this evaluation, the estimates of impact were very small, and in many cases they were not statistically significant. In such cases it is appropriate to ask whether the impact estimates are not statistically significant because they are small in magnitude, or whether the sample size for the evaluation was not sufficiently large to detect effects of the observed size with high power (probability). To address this question, we conducted an *ex post* statistical power analysis.[26] This analysis showed that the study was not underpowered.

K. Summary of Assumptions and Limitations

The major assumptions associated with this analysis are the following:

- The stable unit treatment value assumption (SUTVA, no macro effects assumption, partial equilibrium assumption) is made. This means that the effect (potential outcomes) on one individual are not affected by potential changes in the treatment exposure of other individuals. This implies, for example, that the program is not so large that the outcomes are correlated (e.g., that farmers would produce such a large amount of horticultural crops that the market would collapse).

- The causal models are correct. The key assumption here is that all important unobserved variables affecting both selection and outcomes of interest are time-invariant (i.e., are constant between the two survey rounds).

[26] For a detailed description of *ex post (post hoc)* statistical power analysis, refer to David M. Murray, *Design and Analysis of Group-Randomized Trials*, Oxford University Press, 1998.)

- The program intervention represents a “forced change” in (experimental control of) the agricultural system in Honduras.

- The half of the country that Fintrac had treated before this evaluation began is similar to the half yet to be treated, with respect to relationships among the important causal variables represented in the causal model underlying the statistical analysis.

Other more specific assumptions are listed for particular estimation equations in the detailed analysis presented in Annex 1.

The limitations of the evaluation:

- The causal analysis used to estimate impact is based on assumptions about the selection process. The original evaluation design was based on randomized assignment (of *aldeas*) to treatment, and represented a firmer basis for making causal inferences. With the original approach, randomized assignment assures that the distributions of explanatory variables (other than treatment) are the same for the treatment and control samples. With the revised design, this assertion depends on the correctness of the causal model, and the assumption that unobserved variables affecting both selection for treatment and outcomes of interest are time-invariant.

D. DEVELOPMENT AND IMPLEMENTATION OF THE HOUSEHOLD SURVEY

As has been noted in previous sections, the evaluation design underwent significant changes during the course of the impact evaluation. These multiple changes in the design had a significant impact on the implementation of the household survey. While the actual survey instrument underwent only minor modifications, these changes to later versions of the questionnaire called for a much greater reliance on recall data that spanned longer periods of time. The greatest impact of the design changes from an operational and cost perspective was on the implementation of the baseline data collection, which occurred in three distinct rounds that occurred between July 2008 and July 2010. Data from the first of these rounds (July 2008) is not used in this analysis, since Fintrac accepted only two potential program farmers from this cohort. Hence, while the original study design required only pre- and post-intervention data collections, the problems described in Section C.2 above meant that there were five rounds of data collection (four different baseline collections and one follow-on).

D.1 QUESTIONNAIRE DEVELOPMENT

In early 2008, INE (*Instituto Nacional de Estadísticas* de Honduras) and NORC initiated development of a household questionnaire that would provide the data to support the Impact Evaluation of the FTDA in Honduras. The questionnaire drew largely upon the key elements listed below, as well as input from the team of evaluation experts on the NORC team.

Table 3. Household Survey Elements

Key Elements	Item Description
1. Labor and Income	Detailed information on employment activity and household income and their sources
2. Consumption and expenditures	Retrospective family consumption and expenditure measures (week, month, quarter and year). Health and education measures
3. Travel information	Travel times, cost, access to major employment, highways, markets, school, clinics, etc.
4. Micro enterprises and agriculture	Involvement in micro enterprises including the informal sector. Agricultural practices and products, changes, and additional items for program farmers.
5. Housing costs and prices	Land value items including “How much did you pay for your home/land?” “If you were to sell this land today, how much do you think a buyer would be willing to pay you for it?”
6. Loans and credit	Sources and uses of credit, value of loans, etc.
7. SES/demographics	Basic HH demographic information. Relying upon many standard Census and national household survey items.
8. Perceptions of MCA- Program[27] elements	Qualitative questions on impact of program activities, negative consequences, etc.

[27] These questions on the FTDA program were only included in the second round of the survey and asked of those who had received the FTDA program. We also asked farmers about any other technical assistance they might have received.

In the first phase of questionnaire development, NORC conducted a systematic review of existing questionnaires that collected data on similar subject areas to those proposed for the FTDA survey. Preference was given to surveys that had been applied and field-tested in the region. The team determined that the ENCOVI (Encuesta de Condiciones de Vida), a national survey of the conditions of life of Honduran households, was the best source of existing items for the survey because it focused on many of the same content areas that we proposed to inform key indicators and elements of the evaluation. Furthermore, INE had experience with ENCOVI, having just fielded the survey nationally in Honduras in 2006.

INE delivered the first draft of the questionnaire in February 2008, and then, over the course of the next several months, the NORC and INE teams conducted multiple reviews and conference calls, engaging agricultural experts, to arrive at agreement on the best version of each question or series of questions needed to gather data on a particular impact indicator or series of indicators. In all, six different versions of the questionnaire were generated, reviewed, and revised before a final version was ready for pilot testing in May 2008. As the questionnaire evolved during those months, based on discussions of how to best inform the indicators, more emphasis was placed on developing and expanding the sections on economic and agricultural activities and household consumption, making those the core sections of the questionnaire.

While a significant percentage of the items incorporated into the final household questionnaire were taken directly from previous surveys, many items, particularly those on transportation, household consumption, and agricultural production, were modified or expanded to gather the more detailed data deemed necessary to inform particular impact indicators. Response categories were modified and adjustments were made to ensure adherence to local norms. INE also assisted with many adjustments to the "language" and "terminology" used in instructions, items, and response categories to ensure that we were using appropriate terms and a level of language that was accessible to respondents with lower levels of education.

For Cohorts 1 and 2, both of which were part of the experimental evaluation design, the questionnaire remained the same. However, for subsequent baseline data collections (of new Fintrac recruits in Cohort 2 *aldeas*, and the 545 supplemental sample of Fintrac program farmers), small but important modifications were made to the baseline survey instrument. These subsequent baseline data collections occurred among farmers who had entered the FTDA several months prior to the data collection. Some of these farmers, particularly those in the supplemental sample of 545, had entered the program as much as one year before data collection commenced. As a result, for the sake of comparability with other baseline data, we were compelled to re-word instructions and questions in the income and agriculture modules such that these farmers were obliged to recall the twelve month period prior to their entry into the FTDA when responding to questions on agricultural activity. This modification added a significant recall burden for respondents since the questions often asked them to recall specific agricultural income and input cost practices that occurred as much as 18 to 24 months prior to the date of the interview.

D.2 PILOT TEST OF THE QUESTIONNAIRE AND PROTOCOLS

NORC worked in close conjunction with INE during all phases of the data collection process. INE has more than 10 years' experience conducting national household surveys in Honduras, as well as particular experience in conducting agricultural and surveys of economic activity. With this experience, their suggestions on how to tailor and improve particular questions, as well as how best to organize the study nationally were indispensable.

The final questionnaire, “Encuesta de Hogares, Agricultores y de Precios y Productos,” was comprised of the following modules:

1. Housing Structure
2. Household Composition (Roster)
3. Migration (Internal and International)
4. Household Demographic Information (including education and health)
5. Employment and Economic activity
6. Other sources of income
7. Household consumption (both foodstuffs and other purchases)
8. Agricultural activities (information on all plots and crops, whether on owned or rented lands, production and commercialization, loans and credit obtained, equipment used, technical training received, farm animals)

In May 2008 INE trained 10 field staff to conduct a pilot test of the survey instrument. The questionnaire was revised and refined based on findings from the pilot test.

D.3 FIELD STAFF MATERIALS DEVELOPMENT AND TRAINING

NORC worked closely with INE staff to develop the project-specific training materials. The materials were developed to meet the specific evaluation requirements, but also incorporated NORC’s standard administrative protocols for surveys. We required that INE follow many standard elements of NORC trainings, including training sessions on good interviewing techniques, how to gain and maintain respondent cooperation, preventing interviewer bias, and protecting respondent confidentiality. INE was responsible for developing the interviewer manual that addressed each of these target areas. This manual also included sections that 1) provided a brief description of the study and its goals; 2) described protocols and procedures for survey administration; 3) overview of the study sample and field data collection protocols and procedures; 4) administrative responsibilities; and 5) quality control measures that staff was obliged to follow. INE, together with NORC, developed question by question explanations (QbyQs) for nearly every item in the questionnaire to ensure that field staff would have a source that provided a consistent and correct interpretation of each questionnaire item.[28]

For the first round of baseline data collection, INE trained 60 field staff, all of who had prior data collection experience and over 50% of who had experience administering agricultural surveys. Since the most significant portion of the survey, and perhaps the most difficult to master are those related to agricultural production, we stipulated that INE should make a concerted effort to recruit staff with experience using these types of tools. Additionally, most of these experienced staff had worked on the 2006 ENCOVI and were, therefore, familiar with most sections in the

[28] NORC worked closely with INE to ensure that the training was more interactive and included various techniques like “round robin” and “mock interviews” that engage interviewers more in administering and testing the instrument than a lecture style training that is often typical of many research organizations. NORC provided feedback to INE to ensure that interviewers were observed and received a “pass” on an exit interview if they were to be contracted as interviewers for the household data collection.

survey instrument. Interviewer training occurred in mid-June 2008 and lasted 10 days. Each of the data collection trainings included some classroom activities, but also included interactive modules that permitted the interviewers to practice the survey, section by section, both in groups and then in 2 by 2 interviews, until they demonstrated that they had mastered the instrument.

From this group of 60 interviewers, INE identified the “best” candidates to become supervisors and conducted a subsequent supervisor training. Most supervisors had both prior experience supervising projects and all had worked with INE as field interviewers and had experience administering agricultural surveys. They were also selected to be supervisors based upon their demonstration of a thorough understanding of the study, the survey instrument and all pertinent procedures and protocols.

As we discussed above, several rounds of baseline data collections were undertaken for this evaluation. For each of these rounds of baseline data collection, as well as the follow-on data collection, INE recruited and trained field staff following the protocols described in this section. To the extent possible, INE sought to recruit interviewers from the same pool of field staff for each survey round. During trainings and NORC field observation visits, we were satisfied to observe that the majority of field staff participated in multiple rounds of the baseline data collection and, as such, were very familiar with the instrument and addressed any problems that arose with expertise. The obvious unintended benefit of repeated data collection was that field staff became increasingly familiar with the field instruments, study protocols, and very specific aspects of agricultural practices and activities. Most supervisor and interviewer trainings for subsequent rounds were thorough, but for most field staff they served as more as refresher training.

D.4 FIELD DATA COLLECTIONS

Four rounds of baseline data collection (between July 2008 and July 2010), and one endline data collection in 2011, were conducted for the FTDA evaluation by INE and its staff. The first baseline data collection of Cohort 1 *aldeas* took place in July and August 2008; data were collected from nearly 900 potential program farmers as well as an average of 20 additional households in each of 203 control and treatment villages (n=4800). However, by late 2008, it became apparent that Fintrac had inducted only a handful of the potential program farmers identified into the FTDA. To try to retain the potential-farmer control-group stratification of the original experimental design, NORC identified a second cohort of treatment and control *aldeas* using a new, more detailed, list of criteria provided by Fintrac (this process is described in greater detail above in Section C.2). INE, working with NORC, collected data from what we now refer to as Cohort 2 *aldeas* (179) and farmers (658 potential program farmers plus other households in each *aldea*) in June 2009. This second effort also proved unsuccessful in replicating the Fintrac selection process and identifying farmers acceptable to Fintrac. Fintrac returned to many of these Cohort 2 *aldeas* in early 2010, to identify and recruit new farmers; they also provided NORC with lists of old recruits from Cohort 2 *aldeas* who had entered the FTDA as early as June 2009. Baseline data collection for these farmers (a total of approximately 200), as well as the random sample of 545 program farmers from Fintrac’s own client lists (from normal program operations) was conducted in two sub-rounds between April and July 2010. The follow-on data collection took place in Spring 2011.

The data collection for each round of the baseline, as well as for the endline, was completed by 5-person field teams during 30 day data collection periods.

Three senior technical supervisors oversaw each data collection effort and monitored progress on the ground during the entire data collection period. NORC provided the study sample for each round, along with any available geo-coding and contact information. INE used this information to organize the national data collection in the most cost-efficient manner possible, depending on the geographic dispersion of the cases.

Once data collection began, INE provided NORC with weekly production reports that included information on any anomalies that were occurring in the field, and when possible, potential solutions to resolve these difficulties. Based on weekly itineraries for the data collection provided by INE, NORC staff was able to conduct regular "unplanned" supervisory visits during each round of data collection. NORC's U.S.-based staff typically traveled to Honduras during data collection and observed field work in 2-3 sites during each mission. During these supervisory visits, NORC staff met with INE both in the field and in the central offices in Tegucigalpa to discuss any problems observed in the field, as well as to discuss solutions. NORC's local counterpart, ESA Consultores, also conducted independent supervision at 2-3 different intervals during each round of data collection. They provided timely feedback to INE, MCC and MCA regarding the progress of field work and any problems or issues that they encountered. Regular reports were submitted to MCA on the progress of each round of data collection.

To assure standards of quality in the field, INE used evaluation forms to assess the performance of supervisors, interviewers and team editors (*críticos*) during each round of data collection. These instruments, which were administered by direct the supervisor for each of the aforementioned groups, collected information on a range of tasks performed by each group. The data gathered using these forms was used to respond quickly and efficiently to any issue that was identified in the field.

During the course of NORC's field observations, our primary concern was that supervisors were not always as engaged as they should be in observing interviewers during the course of survey administration; especially during the first week of production. We found that supervisors, charged with other tasks such as gaining consent, were sometimes unavailable during the critical first few days when interviewers were still on unsteady footing or had questions or doubts. NORC brought this issue to INE on several occasions. We recommended that supervisory staff accompany interviewers for the entire length of several of their first few interviews to guarantee that staff fully understood all aspects of the instrument and had a resource if questions arose early in the data collection process. While we saw some progress in this area, it continued to be an issue during each round of data collection.

INE required that interviewers review and code any completed interviews and provide them to the editor by the end of each working day. The editor reviewed the completed questionnaire within one working day and, if necessary, discussed questions or problems with the interviewer and the supervisor. This rapid review permitted the interview staff to return to a household if data retrieval or verification were required. Since an average of just 2 to 3 days was spent in each zone, it was critical that these reviews be conducted promptly so updates could be made before the team left the zone. Completed questionnaires were reviewed by supervisors and if complete, returned in regular shipments to the Central Office in Tegucigalpa for receipting and processing.

D.5 DATA PROCESSING

Upon arrival at INE offices in Tegucigalpa, all surveys, delivered with corresponding control forms, were entered in the Receipt Control system before moving on to the data processing center on site. The Receipt Control system established by INE for the survey contained contact survey identification numbers as well information on the department, province, municipality, *aldea* (village), *caserío* and dwelling. Comparing this pre-loaded information data to the actual survey instruments permitted a strict control and tracking of the hard copy survey instruments.

For each round of data collection INE trained a team of 15 to 20 data entry clerks and two supervisors. INE would conduct 5 day-training of data entry staff prior to the start of data entry. Staff were expected to complete the data entry of 20 surveys per day during an 8 hour work day for the first week and then increase to as many as 25 per day as they became more familiar with the instrument.

First, the data processing staff conducted a review of all completed surveys to identify problems. If significant problems such as missing critical items or omitted sections, were found, staff would attempt, via telephone or through the field supervisor, to retrieve the missing information. If the case was deemed complete, it moved on to data entry.

The INE data entry team began data entry within two weeks of the start of each round of data collection. They performed data entry using an in-house program, which was developed and tested by INE programmers and approved by MCA and NORC prior to the start of data collection. INE protocols require 100% double data entry. To ensure quality and detect any data entry errors, we required that each questionnaire be data entered twice, using different clerks for each of the two entries. Then, supervisors performed a reconciliation of all data entries to identify and correct any errors that were identified. The data entry program was designed to conduct consistency checks and perform a series of validation measures automatically. The next step in processing was to conduct a number of additional consistency and error checks. INE then generated frequencies and crosstabs in SPSS for validation. The data were delivered to the client within 6 – 8 weeks of the end of data collection in the field.

Once the “raw” survey data were available from INE, they were prepared for analysis by the ESA Consultores, the Honduras subcontractor. This cleaning and aggregation process is documented in detail in a series of Stata command (.do) files, Do*FTDAImpact.do (where “*” represents digits 1-11).

D.6 THE SURVEY DATA: SOME KEY OBSERVATIONS

As discussed previously in this report, data for estimating impacts for the FTDA evaluation were obtained from a large-scale household survey administered in a two-round panel survey in which most households were interviewed in both survey rounds. The two rounds of surveys yielded 7,262 completed interview questionnaires, of which 4,526 were from the baseline surveys (Round 0) conducted in 2009 and 2010, and 2,736 were from the follow-on survey round (Round 1) conducted in 2011.

Due to the implementation problem described above, the final sample used for the FTDA evaluation included several farmer types that fell into different combinations of the following categories:

Farmers

- Potential Program Farmers – Cohort 2 farmers who were deemed eligible based on eligibility criteria as specified by MCC and MCA
- Program Farmers (FTDA Farmers) – Farmers who were selected by Fintrac to be part of the FTDA. A few of these came from the original Cohort 2 list selected for the experimental design; others were recruited directly by Fintrac in Cohort 2 *aldeas*; and a third group that was randomly selected from Fintrac's own lists, and had nothing to do with the Cohort 2 *aldeas* linked to the experimental design (i.e., it is a supplemental sample).
- Other Farmers – non-program farmer households who were randomly picked in each Cohort 2 *aldea* as part of a probability sample

Aldeas

- Treatment *Aldeas*
- Control *Aldeas*
- Other *Aldeas –aldeas* are associated with the group of farmers selected from Fintrac's program lists to supplement the diminished treatment sample of the original randomized experimental design

Design

- Original Experimental Design – all *aldeas* and farmers in Cohort 2 *aldeas*
- Not Original Experimental Design – farmers in the supplemental sample taken from Fintrac's program lists

Round

Baseline (Round0)
Endline (Round1)

Based on these various combinations of cohort, *aldea* and farmer, we classified the surveyed population into a number of categories. This made for a far more complex stratification than the original experimental design, which would have been comprised only of potential program farmers (the certainty sample) and the other households (probability sample).

1. Potential lead farmer in Cohort 2 treatment *aldeas* who were immediately accepted by Fintrac into the FTDA program
2. Other program farmers in treatment *aldeas*, who were not part of the original Cohort 2 list, but were recruited later by Fintrac in Cohort 2 *aldeas*
3. Potential Program Farmers in treatment *aldeas* (deemed eligible by screeners using Fintrac selection criteria) that Fintrac rejected (forever)
4. Other households (probability sample) in treatment *aldeas*
5. Potential Program Farmers in control *aldeas* (selected using Fintrac screening criteria)
6. Fintrac clients in control *aldeas* (there should not have been any of these)
7. Other households (probability sample) in control *aldeas*

8. Potential Program Farmers in treatment *aldeas*, initially rejected by Fintrac but then accepted
9. Fintrac clients in supplemental sample taken from Fintrac program lists (around 600)
10. Potential Program Farmers in treatment *aldeas* rejected by Fintrac (interviewed only in baseline)
11. Other households/farmers in treatment *aldeas* rejected by Fintrac (interviewed only in baseline)

All the categories listed above, with the sole exception of Category 9, formed part of the original experimental design. They entered the impact analysis as separate variables to account for their status in the evaluation, the survey design, and the FTDA program. The breakdown of baseline and follow-on survey respondents across these categories is presented in Annex 1. It is important to note here that the large drop in survey respondents from baseline to endline is largely due to the absence of producer categories 10 and 11 from the second-round survey.

All data analysis was conducted at the household level. Data below the household level, such as data at the level of individual household members or specific crops, were aggregated to the household level, such that the unit of analysis was the household. Since the goal of the MCC Compacts is to alleviate poverty among low-income households, analysis of the intervention's impact on income and expenditures at the household level is an appropriate level of analysis for the impact evaluation.

D.7 THE IMPACT INDICATORS

The primary objective of this evaluation is to assess the impact of the FTDA on household income (off-farm and on-farm) and employment, as well as its effect on the cultivation of horticultural crops. The expectation was that there would be a marked increase in net household income, due to increased income generated through the sale of horticultural crops. We might expect income from basic grains to decline as a result; however, that decline would be offset by the much greater gains in the area of horticultural crops. Since household expenditures are positively correlated with income, and because they are usually reported more accurately by respondents than income, expenditures are often a good proxy for income measures. Within this context, the evaluation analysis focused on the following household-level indicators:

For basic grains (BG) (annual amounts):

- Income from basic grains (including used for own consumption) (IncBG)
- Expenses for inputs for basic grains (FactorBG)
- Transportation expenses for basic grains (TranspBG)
- Other costs for basic grains (OthCostBG)
- Labor expense for basic grains (measure of employment associated with BG) (LabExpBG)
- Total expenses, basic grains (ExpBG) = FactorBG + TranspBG + OthCostBG + LabExpBG
- Net income from basic grains (NetBG) = IncBG – ExpBG

For other crops (OC) – horticultural crops (annual amounts):

- Income from other crops (including used for own consumption) (IncOC)
- Expenses for inputs for other crops (FactorOC)
- Transportation expense for other crops (TranspOC)

- Other costs for other crops (OthCostOC)
- Labor expense for other crops (measure of employment associated with OC) (LabExpOC)
- Total expenses, other crops (ExpOC) = FactorOC + TranspOC + OthCostOC + LabExpOC
- Net income from other crops (NetOC)= IncOC – ExpOC

For labor-market employment (monthly amount):

- Income from labor-market (“employee”) work (IncEmp)

For income and expenditures at the household level:

- Total household expenditures (TotHHExp) (monthly amount)
- Net household income (NetHHInc) = NetBG + NetOC + IncTotal*12 (annualized amount), where IncTotal = monthly household income from all sources (labor market, remittances, and other)

Additionally, the survey instrument included a question that recorded whether the household produced horticultural crops. The question asked respondents whether they had harvested horticultural crops (vegetables, fruits) in the last 12 months (not including home garden), with response categories of no = 1, yes = 2.

E. ESTIMATION OF PROGRAM IMPACT

The impact analysis conducted in support of evaluation of the FTDA was complicated and complex for several reasons. First, we examined a number of outcome indicators of interest, many of which are interrelated. Second, the problems encountered in implementing the experimental design (described above in Section C.2) meant that our approach, and analysis, switched from a straightforward pretest-posttest-randomized-control-group design for which the observed (sample) double-difference estimator would be an unbiased estimator of the double-difference measure, to a far more complex model-based approach in which an unbiased or consistent estimate of the double-difference measure is obtained from a statistical regression model and various assumptions. During the course of the analysis, we considered several different impact estimators, including propensity-score-based estimators and regression estimators based on selection and outcome models.[29] Information about these estimators is included in Annex 1; results are summarized for one of them in this section.

E.1 MODEL SPECIFICATION, IDENTIFICATION AND ESTIMATION

Section C3.2.E presents the selection model used in our analysis. It is based on the Round 0 (baseline) data. The outcome model is somewhat different from the single-round example presented earlier, since it involves a two-round panel sample. The following model (or variations of it) is used to represent outcome as a function of explanatory variables, and form the basis for estimation of program impact[30].

$$y_t = \mathbf{x}'_t\boldsymbol{\beta} + \theta d_t + \phi w_t + \delta d_t w_t + e_t,$$

where,

t = survey round index (0 for Round 0, which is the baseline, and 1 for Round 1, the follow-on, or endline)
y_t = explained variable (outcome variable, response variable, dependent variable)
$\mathbf{x}_t$ = vector of explanatory variables (the first component is one)
$\boldsymbol{\beta}$ = vector of parameters (the first parameter is a constant term)
d_t = indicator variable for survey round, = 0 for Round 0 and 1 for Round 1
θ = round effect
w_t = treatment variable
ϕ = treatment effect (not the impact, but the average difference in means between the treatment and control groups at baseline)
δ = impact (interaction effect of treatment and time)
e_t = model error term.

[29] Several estimators were considered because some estimators work better than others in different circumstances, and it is not generally known which estimators will perform best until after the analysis is completed.

[30] For discussion of this model, see *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, by Jeffrey M. Wooldridge (Massachusetts Institute of Technology Press, 2010, 2002).

The model error term is assumed to have mean zero, constant variance, and be uncorrelated with the explanatory variables. In this application, the treatment variable, w_t, is a binary variable having value one for sample units (households, farmers) who receive program services and zero otherwise. The coefficient δ is an estimate of the average treatment effect (ATE), which is the expected effect of the program intervention on a household in *aldeas* randomly selected from the program's target population. The preceding linear statistical model may be used directly to estimate impact, or in a two-step model that includes a "selection" model (which represents the probability of participation, or propensity score, as a logistic function of a linear form such as shown above) and an "outcome" model that includes the estimated propensity score as a covariate. For this evaluation, the two-step model turned out to be the best approach.

Below we present a summary description of the model, to facilitate understanding from a substantive (economic, econometric) viewpoint. Annex 1 presents a detailed description of the two-step estimation model.

The First-Step (Selection) Model

The first step model is a logistic regression model that estimates the probability of a farmer participating in the FTDA. This model, a binary response model, is as follows:

$$P(y=1|\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta}) \equiv p(\mathbf{x})$$

where **x** denotes a (column) vector of explanatory variables, $P(y=1|\mathbf{x})$ denotes the probability that y=1 (i.e., is treated) conditional on **x**, $\boldsymbol{\beta}$ is a vector of parameters and g(.) is a the logistic link function,

$$g(z) = \exp(z)/(1 + \exp(z)).$$

If we define z as

$$z = \mathbf{x}'\boldsymbol{\beta} + e,$$

where e denotes a random error term uncorrelated with **x** and with mean zero, then

$$y = 1 \text{ if } g(z)>.5 \text{ and } 0 \text{ otherwise.}$$

The expression $\mathbf{x}'\boldsymbol{\beta}$ is referred to as an index. The parameters $\boldsymbol{\beta}$ are estimated by the method of maximum likelihood. The expression $\mathbf{x}'\boldsymbol{\beta}$ does not have any meaning (or units) – it is simply a modeling artifact. The model is often referred to as a "latent variable" model, since the variable z is unobserved.

The identification of the explanatory variables to include in the selection model is guided by an underlying causal model. Variables that are considered likely to have an effect on selection are selected from the questionnaire. Since questionnaire variables are correlated, we attempt to make a selection that is not highly intercorrelated, yet reflects the underlying factors that may affect selection. The selection model uses data only from the first survey round (baseline).

For this application, the index is estimated to be

x'β = -10.87894 - .1250273*HouseholdSize + .1594562*FormalEducHead + .868967*AgEmployees - .0870384*TotHaOwnFarm + .0211418*TimeToSchool - .0103442*TimeToHosp + .9303271*LogTotHHExp + .2160815*LogIncBG - .2096164*LogLabExpBG + .2420389*LogIncOC - .2358687*LogLabExpOC

where

HouseholdSize = number of persons in the household
FormalEducHead = years of formal study of head of household
AgEmployees = number of household occupants in agricultural work
TotHaOwnFarm = total farm hectares owned
TimeToSchool = travel time in minutes to school
TimeToHospital = travel time in minutes to hospital.
LogTotHHExp = logarithm of total monthly household expenditures
LogIncBG = logarithm of value of production of basic grains
LogLabExpBG = logarithm of manual-labor expenditures for basic grains
LogIncOC = logarithm of value of production of other crops
LogLabExpOC = logarithm of manual-labor expenditures for other crops

The selection model presented above includes only variables that were highly statistically significant. The value of the "pseudo R^2" (a standard measure of model fit) for this model is .44, which is considered a relatively high value for this type of application.

There were a few missing values in some of the explanatory variables. In order to retain all of the observations for the regression analysis, these missing values were imputed as means of the non-missing values.

Some of the variables included in the model are logarithms of variables, and these are undefined for nonpositive values of the argument (of the logarithmic transformation). We replaced these undefined values by zeros, and included indicator ("dummy") variables in the model to account for the nonlinearity of this transformation. The inclusion of the dummy variables made little difference in the model fit (R^2 increased from .44 to .46), but the interpretation of the model parameters became more difficult. As a result, this alternative model was not considered further. (As noted, the coefficients in a logistic model have no economic meaning, and so inclusion of logarithmic terms without corresponding dummies does not present conceptual problems.)

Economic Interpretation of the Selection Model

The interpretation of each of the variables included in the first-step model follows:

- HouseholdSize (negative coefficient): larger households are less likely to participate
- FormalEducHead (positive coefficient): farmers with more formal education are more likely to participate
- AgEmployees (positive): households having more agricultural-sector employees are more likely to participate
- TotHaOwnFarm (negative): the larger the owned farm hectares, the less likely the farmer is to participate
- TimeToSchool (positive): the closer the school, the higher the likelihood of participation
- TimeToHospital (negative): the more remote the household, the lower the likelihood of participation

- LogTotHHExp (positive): households with larger total household expenses are more likely to participate
- LogIncBG (positive): the higher the basic-grains income, the higher the likelihood of participation
- LogLabExpBG (negative): the higher the basic-grains labor expense, the lower the likelihood of participation
- LogIncOC (positive): the higher the other-crops income, the higher the likelihood of participation
- LogLabExpOC (negative): the higher the other-crops labor expense, the lower the likelihood of participation.

All of the preceding variables are included in the list of causal factors and variables affecting selection, presented earlier. Note that the participation model reflects both the decision of Fintrac to accept a farmer into the program as well as the decision of the farmer to participate. The explanatory variables included in the model could reflect either type of decision, or both.

Remarks on Model Specification, Identification and Estimation

Estimation of program impact involves consideration of both the selection model and the models of outcomes of interest. The essential feature of these model pairs is that variables that affect both selection and outcomes of interest be observable (“selection on observables”), or, if not observable (“selection on unobservables”), be time-invariant. The following comments are made about the nature of the selection model and its relationship to the outcome models to be considered.

1. The household variables are correlated. There is not a unique model that describes the probability of selection, but an infinite variety of such models. The goal is to include a set of explanatory variables that reflects the important factors affecting selection, yet for which the intercorrelations are as low as possible. During the course of the analysis, a number of alternative selection model specifications were examined, including more, fewer and different variables than were listed above. For reasonable specifications, the results were similar (i.e., the value of R^2 was similar). The preceding model is one such model.

2. The selection model is determined solely from Round 0 (baseline) data, since selection into the program is made at Round 0.

3. It is not the goal to estimate individual parameters (coefficients) of the selection model. The goal is to estimate the propensity score (i.e., the explained variable), not individual coefficients. The individual coefficients of the selection model are not used in the analysis. For complex link functions, the parameters do not have a straightforward economic interpretation. Moreover, not only is the selection of explanatory variables not unique, but the explanatory variables are correlated, so that the estimates of individual coefficients are also correlated (confounded). The situation is similar to the problem of forecasting – the goal is to estimate, or forecast, the response variable, not to estimate the marginal effect of response to individual explanatory variables. It does not matter is the estimates of individual coefficients are biased or imprecise, because their magnitudes are of no interest – what matters is that they reflect factors affecting selection, so that the

value of R^2 is high. Care must be taken to avoid including too many explanatory variables in the selection model, to avoid overfitting the model.

4. There may be unobserved variables affecting selection for treatment, and these unobserved variables may be correlated with the explanatory variables included in the model (in which case ordinary least squares estimates of the model parameters are biased. As mentioned, the essential concern is whether the unobserved variables affecting selection and outcomes of interest are time-invariant, in which case they drop out of the (fixed-effects, two-round panel) outcome model, so that the assumption of conditional independence is justified. It is desirable to have a high value for R^2, since this promotes high precision of the impact estimates. From the viewpoint of bias, however, the value of R^2 is not important. What is important is that all variables affecting both selection for treatment and outcomes of interest be observable or, if not, be time-invariant.

5. Considered over time, a number of the explanatory variables in the selection model may be affected by the explained variable, i.e., they may be endogenous. Applying the ordinary least squares estimation procedure to cross-sectional data, estimates of the model parameters are biased. As discussed, estimation of the parameters is not the objective – the objective is estimation of the propensity score. Selection for the program is based on variables that are available at baseline (Round 0), and selection status does not vary over time. For the selection model, endogeneity is not an issue.

6. It is desired to include a set of observed variables that collectively do a good job of estimating the probability of selection (as reflected in the value of R^2). With respect to unobserved variables, the assumption is made that unobserved variables that have an important effect on both selection and on outcomes of interest are time-invariant. The issue of unobserved variables was discussed at length earlier. Unobserved variables that do not affect both selection and outcomes of interest are not a primary concern. They may reduce the value of R^2, which is certainly undesirable, and they may bias the estimates of the selection model parameters, but they do not corrupt (bias) the estimation of impact.

7. For the basic propensity-score method of estimation, it is required that the selection model include probabilities not equal to zero or one, since observations having such values are of little use to the estimation of impact for this method. For the econometric approach, it is required that the selection model contain at least one variable that is not included in the outcome model (this assumption will hold for all of the outcome models to be considered).

The Second-Step (Outcome) Model

Having estimated the propensity score, the second step in the two-step model is an outcome model that includes the estimated propensity score. This model is the “modified regression-adjusted propensity-score-based model,” which we obtained by regressing y (the outcome variable) on 1, Round, Treated, RoundTreated, $\hat{p}(\boldsymbol{x})$, Round($\hat{p}(\boldsymbol{x})$ - $\hat{\mu}_p$) and RoundTreated($\hat{p}(\boldsymbol{x})$ - $\hat{\mu}_p$), where $\hat{\mu}_p$ denotes the mean of the estimated propensity scores. The descriptor “modified” refers to the fact that this is the same model as a regression-adjusted propensity-score-based model, with the addition of a term representing the interaction of the demeaned propensity score with RoundTreated. The estimate of impact is the coefficient of the Round*Treatment interaction term or the interaction effect of treatment and time. This estimate is an estimate of the average treatment effect (ATE), or expected impact of the program intervention on a randomly selected program-eligible farmer (or, more accurately, of a program-selected program-eligible farmer in a randomly selected program eligible aldea).

This model allows for the impact effect to be directly related to the (demeaned) estimated propensity score. It may be used to estimate impact as a function of the covariates. This feature of the model is useful for estimation of the average treatment effect on the treated (ATT).

The modified regression-adjusted propensity-score-based model may be represented as:

$$y_t = \beta_0 + \beta_1 \text{ Round} + \beta_2 \text{ Treated} + \beta_3 \text{ RoundTreated} + \beta_4 \hat{p}(\mathbf{x}) + \beta_5 \text{ Round } \hat{p}(\mathbf{x}) + \beta_6 \text{ RoundTreated } (\hat{p}(\boldsymbol{x}) - \hat{\mu}_p) + e_t,$$

where

t = survey round index (0 for Round 0 and 1 for Round 1)
y_t = explained variable (outcome variable, response variable, dependent variable)
Round = survey round (0 or 1)
Treated = FTDA program client or not (0 or 1)
$\boldsymbol{\beta}$ = vector of parameters (the first parameter is a constant term)
$\hat{p}(\mathbf{x})$ = estimated propensity score
$\mathbf{x}$ = vector of covariates (the explanatory variable used in the propensity-score model).
e_t = model error term.

This estimator is a covariate-adjusted regression estimator, where the covariate is the estimated propensity score, and the interaction term with this demeaned covariate is included in the model.

The theory underlying the use of propensity scores in counterfactuals analysis is that the treatment variable and the counterfactual responses to treatment are independent, given the propensity score. That is, given the propensity score (i.e., a group of units having the same propensity score), an unbiased estimate of impact (conditional on the specified value of the propensity score) is the simple difference in means of the treated and untreated units. "(The average treatment effect is obtained by averaging over the propensity score distribution.) Under the assumption of conditional independence (of treatment and response), there is no need to include additional covariates in the outcome model, once the propensity score is included. For this reason, we do not include additional covariates are included in the preceding model, beyond the estimated propensity score.

It is important to allow for "flexible" specifications involving the propensity score. The relationship of outcomes of interest to the propensity score may be more complicated than a simple linear relationship. It was determined in the course of the analysis that a model that included a linear term and an interaction term with round worked well.

E.2 IMPACT ESTIMATES

Average Treatment Effect (ATE)

There is a separate response model for each outcome variable, each with its own set of β's. In each model, the estimate of impact is the RoundTreated effect (coefficient β_3 in the table to be presented). The full regression outputs for those models is not presented here, but are included in the Stata .log file that accompanies the project documentation. Here follows tables showing key model parameters (coefficients of treatment-related parameters), for both random-effects and fixed-effects models. The tables present the values of β_3, β_4, β_5 and β_6 for each outcome variable, along with their standard errors. The random-effects table is used to show the relationship of outcome to explanatory variables that are not estimable in the fixed-effects model, and the fixed-effects table is used to estimate impact. Note that the value of the propensity score is identical for the same household between survey rounds, and for this reason the parameter β_4 drops out of the fixed-effects model. It is retained in the model formula, to facilitate comparison between the fixed-effects and random-effects models.

In general, when assessing the economic meaning of a model, the random-effects model is preferred to the fixed-effects model. The reason for this is that the random-effects model is a structural representation with high face validity, whereas the fixed-effects model is in effect an "estimating equation," in which model parameters are dropped if they are have the same values in both survey rounds. As mentioned, since the value of the propensity score is the same in both rounds, the "P" term (corresponding to β_4) is dropped from all of the fixed-effects models. If it is of interest to see the relationship of the response to the propensity score, the random-effects model is used. In this application, the fixed-effects and random-effects models were generally similar (except for the fact that the propensity score drops out of the fixed-effects model). Because of the large sample size, the difference between the two models was usually statistically significant, but the difference is not large. Similarity of the fixed-effects and random-effects estimates is evidence that time-invariant unobserved variables are not correlated with the explanatory variables.

Table 4A presents results for the random-effects model. The coefficients β_4 and β_5 represent adjustments to the response (not to the impact), and may be of either sign. The interesting thing to observe here (in the case of NetHHInc) is that there is a very strong positive relationship of response to estimated propensity score (coefficient 202,469), and a modest negative relationship to the interaction of the estimated propensity score and RoundP (-12,088). This means that, in general, famers who had a high propensity for program participation tended to have high incomes in Round 0 and not quite so high incomes in Round 1. This situation is associated with a weak impact.

The fixed-effects table (Table 4B) shows that, in general, the direction of the impact (coefficient β_3) is as expected. For example, in the case of NetHHInc, the value of β_3 is positive. The coefficient β_5 represent adjustments to the *response*, not to the *impact*, and may be of either sign.

The coefficient β_6 represents an adjustment to impact. The coefficient β_6 (interaction RoundTreatedPstd) is not statistically significant (for any outcome variable). The interpretation of this is that there is not a strong relationship between impact (*impact*, not response) and the estimated propensity score, although the relationship of the response to the estimated propensity score is strong. We will revisit this fact later, in estimation of the average treatment effect on the treated (ATT).

Table 4A: Key Model Parameters for Modified Regression-Adjusted Propensity-Score-Based Random-Effects Estimator

Outcome Variable	β_3 (RoundTreated)		β_4 (P)		β_5 (RoundP)		β_6 (RoundTreatedPstd)	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
IncBG	-352	950	12635	1129	-5832	2091	2420	2491
ExpBG	815	360	2076	442	15.1	796	1111	953
NetBG	-1216	824	10559	888	-6182	1772	1889	2072
LabExpBG	396	234	-1121	248	1605	501	410	583
IncOC	10122	4531	80793	5223	-21002	9908	13265	11741
ExpOC	4833	1027	10553	1178	2422	2243	2615	2656
NetOC	4657	4086	70240	4504	-25729	8840	15206	10390
LabExpOC	2514	667	-1035	747	7297	1449	579	1708
IncEmp	-268	730	9953	817	1285	1585	-846	1868
TotHHExp	98	455	5740	460	-1856	959	-849	1102
NetHHInc	5414	11046	202469	12713	-12088	24030	-9250	28550
Horticulture	-.0572	.0237	-.0299	.0221	-.0443	.0514	.0931	.0566

Table 4B: Key Model Parameters for Modified Regression-Adjusted Propensity-Score-Based Fixed-Effects Estimator

Outcome Variable	β_3 (RoundTreated)		β_4 (P) (drops out)		β_5 (RoundP)		β_6 (RoundTreatedPstd)	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
IncBG	-120	979			-4428	2328	-278	2918
ExpBG	837	375			25.3	891	899	375
NetBG	-957	851			-4453	2022	-1177	2022
LabExpBG	351	248			1312	589	877	738
IncOC	26774	4665			-4752	11087	-37399	13900
ExpOC	5413	1075			6233	2556	-4622	3704
NetOC	11360	4206			-1481	9997	-32777	12534
LabExpOC	1911	707			6828	1681	1957	2107
IncEmp	149	755			2097	1794	-2020	1794
TotHHExp	204	465			-2865	1105	1466	1385
NetHHInc	18926	11411			21956	26892	-73123	33744
Horticulture	-.0397	.0258			-.0534	.0660	.0516	.0797

These results show a positive effect of the FTDA program. Net income change from horticultural crops (ATE estimate) is on average 11,360 lempiras (USD 600) higher for program participants than for nonparticipants. Input expenditures on these crops increased far more than they did for basic crops, implying a higher level of activity in cultivation of high value crops among program farmers[31]. The results suggest a corresponding decline among program farmers in income from basic crops, as might be expected with changing crop mix; however, this decline is not statistically significant. These results are consistent with the program logic and hypotheses for the FTDA.

Box 3: Two Key Findings

- The FTDA had a positive, but weak, impact on its primary area of focus: income, net income, expenditures, and labor associated with horticultural crops
- These positive impacts did not translate into increases in net household income and expenditures

Some of the results do not conform to expectations. For example, the program does not appear to have had a positive effect on the proportion of farmers growing horticultural crops. This could well be because the implementer primarily chose as program participants farmers who showed a proven ability to grow horticultural crops. This suggests that increments in income from horticultural crops came from increased production among farmers already growing horticultural crops and not from farmers who switched over for the first time.

These results of the impact analysis show strong statistical evidence that *the FTDA program had a positive, though weak, effect on income, net income, and expenditures and labor expenditures for horticultural crops. In other words, the interventions had a positive impact on its primary area of focus: activities related to horticultural crops. However, we did not detect a broader positive impact on household income and expenditures.*

Average Treatment Effect on the Treated (ATT)

We also estimated the average treatment effect on the treated. The average treatment effect (ATE, estimated above) is the expected impact of the program intervention on a randomly selected program-eligible farmer. The average treatment effect on the treated (ATT) is the expected impact of the program intervention on a treated farmer. The ATT was found to be similar to the ATE. The ATT is presented in Annex 1 for all of the outcome variables analyzed in detail.

For most outcome measures, the estimates of the ATE and ATT estimates of impact are generally similar in magnitude. This means that the effect on a randomly selected program-eligible individual in a randomly selected aldea is not much different from the effect on a Fintrac-selected program-eligible individual. This is consistent with our finding that the impact of the FTDA was not large. Had the ATE been small but the ATT large, we would conclude that, although the impact for a randomly selected program-eligible farmer may be small, the program had a substantial effect on Fintrac-selected clients, and that Fintrac knew how to select clients that would perform better-than-average in the program. This does not appear to be the case. We find, instead, that the program has a statistically significant, but weak, impact, which is about the same for Fintrac-selected clients as for randomly selected eligible farmers.

While the relationship of impact to treatment (program participation) is not strong, the relationship of income to the estimated propensity score is very strong. Farmers similar to those selected for treatment tend to do well, even though they do not participate in the program. This

[31] The impact results do not imply that the incomes or expenditures of program farmers did or did not increase by a substantial amount. They show changes in income (and other indicators) for program farmers relative to changes in the same indicators among control farmers.

indicates that Fintrac has the ability to select farmers who are likely to do well. This is not the same as a differential treatment effect (between treated and untreated farmers). Although Fintrac may have the ability to select farmers who are likely to do well, whether they participated in the FTDA or not, our results do not show that Fintrac has the ability to select farmers who are likely to perform noticeably better in the FTDA program than other program-eligible farmers.

E.3 SUMMARY OF RESULTS

The results of the impact evaluation show that the FTDA activity had a positive impact on its primary area of focus: activities related to horticultural crops. However, a broader positive impact on household income and expenditures was not detected.

The impact estimates were based on data that included all of the data obtained from the original experimental design, augmented by a sample of program farmers recruited by Fintrac in the course of its normal project operations. Statistical/econometric analysis was used to adjust for differences between the treatment and control samples and to thereby reduce potential selection bias associated with a lack of proper randomization. The statistical/econometric analysis procedures used to estimate impact are based on sound causal models and causal modeling theory[32]. The impact estimates constructed in this evaluation are estimates of the causal effect of the FTDA program intervention. An *ex post* statistical power analysis (presented in Annex 1) shows that the study was not "underpowered." We, therefore, consider the inferences made in this evaluation analysis to be sound (valid and of adequate precision and power), given the assumptions and limitation of the model described below. Based on our analysis, we conclude that the FTDA program produced positive results relative to horticulture production, but those results were small in magnitude.

Assumptions and Limitations

The major assumptions associated with this analysis are the following:

1. The stable unit treatment value assumption (SUTVA, no macro effects assumption, partial equilibrium assumption) is made. This means that the effect (potential outcomes) on one individual are not affected by potential changes in the treatment exposure of other individuals. This implies, for example, that the program is not so large that the outcomes are correlated (e.g., that farmers would produce such a large amount of horticultural crops that the market would collapse).

2. The causal models are correct. The key assumption here is that all important unobserved variables affecting selection are time invariant (i.e., are constant between the two survey rounds).

3. The program intervention represents a "forced change" in (experimental control of) the agricultural system in Honduras.

[32] Neyman-Fisher-Cox-Rubin Causal Model, potential outcomes model, counterfactuals model

4. The half of the country that Fintrac had treated before this evaluation began is similar to the half yet to be treated, with respect to relationships among the important causal variables represented in the causal model underlying the statistical analysis.

Other more specific assumptions are listed for particular estimation equations in the detailed analysis presented in Annex 1.

The main limitation of the evaluation is that the causal analysis used to estimate impact is based on assumptions about the selection process. The original evaluation design was based on randomized assignment (of aldeas) to treatment, and represented a firmer basis for making causal inferences. With the original approach, randomized assignment assures that the distributions of explanatory variables (other than treatment) are the same for the treatment and control samples. With the revised design, this assertion depends on the correctness of the causal model, and the assumption that unobserved variables affecting selection for treatment are time-invariant.

ANNEX 1: ESTIMATION OF IMPACT: A DETAILED DESCRIPTION OF THE ANALYSIS AND RESULTS

I. Introduction

This annex presents details on the analysis used to construct the impact estimates presented in the main text. It includes a description of the data processing that was done prior to the statistical analysis. It includes discussion of the impact estimates presented in the main text.

II. Preliminary Data Processing

Prior to conducting the data analysis (using the Stata statistical program package, version 10.0), NORC's data analysts conducted a rigorous quality review, cleaning and aggregation of the "raw" survey data. Since the primary unit of analysis for the survey data was the household, a major aspect of the initial data processing was aggregation of the detailed information included on the survey questionnaire into household-level data for analysis. This included aggregation of data for individual family members and items of income and expense, for various crops and for the household in general. The result of this initial data processing was a "flat file" (table) that included aggregated household-level data (one file record (row) per household)[33]. (It may be asked why the questionnaire collected disaggregated data, when the data were aggregated for the impact analysis. The primary reason for collecting disaggregated data (e.g., data for individual family members or for separate crops) is that collecting the detailed data and aggregating it is generally considered to produce more accurate aggregate measures than simply asking for aggregate amounts in the questionnaire. A secondary reason is that analysis of the detailed (disaggregated) data may provide additional insight into the mechanisms of impact, such as relationships to family-member characteristics or effects for individual crops. The scope of the evaluation contract was to estimate overall program impact, and it did not include time or resources to conduct extensive analysis of disaggregated data.)

In this section, we classify the household population (and survey population) into a number of different categories, according to their status in the evaluation design, the survey design, and the Fintrac program. We also present and discuss the impact indicators (outcome variables) of interest. This is a necessary first step, prior to discussing the impact analysis and its results.

II.A. Classification of the Survey Population

In the original evaluation experimental design, farmers were classified (stratified) into two categories: potential lead farmers and others. The intention was that only potential lead farmers in treatment *aldeas* would receive FTDA program services. A random sample of 20 farmers (a

[33] All of the initial data processing and analysis steps are documented in detail in Stata command files ("do" or ".do" files). The output from each .do file is a "log" file (or ".log" file). The processing and analysis may be replicated by executing the .do file, in which case the results will be presented in an associated .log file. For the FTDA evaluation project, .do files named Do1* - Do13* (where * denotes additional text) were used to clean and aggregate the questionnaire data to household level, and Do14FTDAImpactEstimation.do was used to construct the impact estimates.

probability sample) – “other” farmers - was selected from the treatment and control *aldeas*; they constituted a probability sample for the evaluation. Farmers selected for observation (survey) in the original experimental design are referred to as “Design” farmers. All others are “non-Design” farmers. The Design farmers fall into two categories: those selected for treatment (“DesSelForTrt”), and those selected for control (“DesSelForCtrl”). Farmers who received FTDA program services are referred to as “Treated” farmers.

As things turned out, the set of Design farmers who were treated differed somewhat from the set of farmers selected for treatment. This development had little effect on the results of the evaluation, since randomized assignment to treatment was conducted at the *aldea* level, not at the farmer level. However, Fintrac, the program implementer, also rejected entire *aldeas* that had been assigned to treatment; this had the effect of compromising the original experimental design. Faced with this reality, NORC and MCC decided to turn to an alternative evaluation design, which involved augmenting what remained of the original experimental sample (non-rejected farmers) with an additional 600 Fintrac clients (who entered the program around the same time as Cohort 2 farmers), and adopting a “model-based” evaluation approach, instead of the “design-based” approach that was originally planned. These 600 farmers are also referred to as “Treated” farmers, although they have no relationship to the original experimental design.

Because of the problem in implementing the evaluation and the complexities it introduced into the process, the surveys conducted for the FTDA evaluation included a number of different categories of farmers, *aldeas*, evaluation design and survey round. These categories are:

Farmers

- Potential Program Farmers – Cohort 2 farmers who were deemed eligible based on Fintrac’s stated selection/eligibility criteria
- Program Farmers (FTDA Farmers) – Farmers who were selected by Fintrac to be part of the FTDA. A few of these came from the original Cohort 2 list selected for the experimental design; others were recruited directly by Fintrac in Cohort 2 *aldeas*; and a third group that was randomly selected from Fintrac’s own lists, and had nothing to do with the Cohort 2 *aldeas* linked to the experimental design (supplemental sample).
- Other Farmers – non-program farmer households who were randomly picked in each Cohort 2 *aldea* as part of a probability sample

Aldeas

- Treatment *Aldeas*
- Control *Aldeas*
- Other *Aldeas* – these *aldeas* are associated with the group of farmers selected from Fintrac’s program lists to supplement the diminished treatment sample

Design

- Original Experimental Design – all *aldeas* and farmers in Cohort 2 *aldeas*
- Not Original Experimental Design – farmers in the supplemental sample taken from Fintrac’s program lists

Round

Baseline (Round 0, for the purpose of this report)

Endline (Round 1)

Not all 36 (3 x 3 x 2 x 2) different combinations of the preceding classification variables occurred in the survey population. For various reasons (discussed earlier), the FTDA baseline survey was conducted in several phases, or “cohorts.” The various combinations of cohort, *aldea* type and farmer type are as follows. Each of the preceding combinations is referred to as a “Producer Category,” or “PC”:

1. Potential program farmer in Cohort 2 treatment *aldeas* who were immediately accepted by Fintrac into the FTDA program
2. Other program farmers in treatment *aldeas*, who were not part of the original Cohort 2 list, but were recruited later by Fintrac in Cohort 2 *aldeas*
3. Potential program farmers in Cohort 2 treatment *aldeas* (deemed eligible by screeners using Fintrac selection criteria) that Fintrac rejected (forever)
4. Other households (probability sample) in treatment *aldeas*
5. Potential program farmers in control *aldeas* (selected using Fintrac screening criteria)
6. Fintrac clients in control *aldeas* (there should not have been any of these)
7. Other households (probability sample) in control *aldeas*
8. Potential Program Farmers in treatment *aldeas*, initially rejected by Fintrac but then accepted
9. Fintrac clients in supplemental sample taken from Fintrac program lists (around 600)
10. Potential Program Farmers in rejected Cohort 2 treatment *aldeas* (interviewed only in baseline)
11. Other households/farmers (probability sample) in Cohort 2 treatment *aldeas* rejected by Fintrac (interviewed only in baseline)

II.B. Distribution of the Survey Population across Key Sample Classifications

The household survey consisted of a total of 7,596 sample units (households) in both survey rounds, of which 4,533 are in Round 0 (baseline) and 3,063 in Round 1 (endline or follow-up). The number of nonrespondents (all table lines after the first) is 7 for Round 0 and 327 for Round 1. Only completed questionnaires (line 1 of the table) were retained for the analysis. Table A.1 shows the number of sample households by these response categories.

Table A.1. Survey Responses

Response	Round		Total
	0	1	
Interviewed	4,526	2,736	**7,262**
Absent	0	71	**71**
Incomplete	0	17	**17**
Home Unoccupied	0	89	**89**
Home Destroyed	0	10	**10**
Two leaders in Same House	7	1	**8**
Refused	0	82	**82**
Deceased	0	2	**2**
Moved	0	3	**3**
Unknown/Not Located	0	51	**51**
Duplicate Farmer	0	1	**1**
Total	**4,533**	**3,063**	**7,596**

Table A.2 shows the number of respondents in each of the 11 producer categories listed in Section II.A, by survey round.

Table A.2. Respondents by Producer Category and Round

Response	Round		Total
	0	1	
1. Potential farmers Cohort 2 treatment *aldeas*, immediately accepted by Fintrac into FTDA program	20	18	**38**
2. Other program farmers in treatment *aldeas*, recruited later by Fintrac in Cohort 2 *aldeas*	8	8	**16**
3. Potential program farmers in treatment *aldeas* rejected by Fintrac	63	49	**112**
4. Other households (probability sample) in treatment *aldeas*	498	445	**943**
5. Potential program farmers in control *aldeas*	280	252	**532**
6. Fintrac clients in control *aldeas* (should not be any)	2	2	**4**
7. Other households (probability sample) in control *aldeas*	1,483	1,343	**2,826**
8. Potential program farmers in treatment *aldeas*, rejected and then accepted by Fintrac	157	140	**297**
9. Fintrac clients in supplemental sample taken from Fintrac program lists (around 600)	545	479	**1,024**
10. Potential program farmers in rejected treatment *aldea* (interviewed only in baseline)	224	0	**224**
11. Other households/farmers in treatment *aldeas* rejected by Fintrac (interviewed only in baseline)	1,246	0	**1,246**
Total	**4,526**	**2,736**	**7,262**

Table A.3 presents a breakdown of households by Producer Category. For the first two categories it is of interest to know how many *aldeas* were in the categories. The 18 households in Producer Category 1 were in 11 *aldeas* and the 8 households in Producer Category 2 were in 5 *aldeas*.

Table A.3 presents a breakdown of baseline (Round 0) respondents in each Producer Category by treatment status (Treated = 1 if client received program services, 0 otherwise). As the highlighted PCs in the table indicate, after multiple rounds of Fintrac-led recruitment, we ended up with 185 program farmers in Cohort 2 *aldeas*; of these 177 (20 + 157) were from the potential program farmers that were identified according to objective Fintrac eligibility criteria; 157 of these were initially rejected by Fintrac and then accepted later into the FTDA.

Table A.3. Respondents by Producer Category and Treatment Status ("Treated") in baseline (Round 0)

Response	Treated		Total
	No (0)	Yes (1)	
1. Potential farmers Cohort 2 treatment *aldeas*, immediately accepted by Fintrac into FTDA program	0	20	**20**
2. Other program farmers in treatment *aldeas*, recruited later by Fintrac in Cohort 2 *aldeas*	0	8	**8**
3. Potential program farmers in treatment *aldeas* rejected by Fintrac	63	0	**63**
4. Other households (probability sample) in treatment *aldeas*	498	0	**498**

5. Potential program farmers in control *aldeas*	280	0	**280**
6. Fintrac clients in control *aldeas* (should not be any)	0	2	**2**
7. Other households (probability sample) in control *aldeas*	1,483	0	**1,483**
8. Potential program farmers in treatment *aldeas*, rejected and then accepted by Fintrac	0	157	**157**
9. Fintrac clients in supplemental sample taken from Fintrac program lists (around 600)	0	545	**545**
10. Potential program farmers in rejected treatment *aldeas* (interviewed only in baseline)	224	0	**224**
11. Other households/farmers in treatment *aldeas* rejected by Fintrac (interviewed only in baseline)	1,246	0	**1,246**
Total	**3,794**	**732**	**4,526**

The penultimate category of the preceding table, Producer Category 10, includes potential program farmers in rejected treatment *aldeas*. Because these *aldeas* were rejected by Fintrac as a whole, a decision was made, for cost reasons, not to return to them to collect endline data. As such, they are of limited use to the impact analysis (which benefits much more from data from households that were interviewed in both survey rounds, than from households that were interviewed in only one round). In retrospect, this was probably a short-sighted decision. Had these potential program farmers been retained in the follow-up data collection they would have supported construction of an intention-to-treat-like (ITT-like) estimate of program impact. (It would not be a true ITT estimate, because NORC selected only *potential* treatment farmers; it was the program implementer, Fintrac, that selected the program farmers from the random sample of *aldeas*.)

For the purpose of the data analysis, we defined several additional indicator variables related to treatment status and whether or not the farmer was part of the original experimental design or part of the supplemental sample. (The following variables relate to farmers, not to *aldeas*. Hence a "nontreatment" farmer in a treatment *aldea* is classified as a control. The term "control" may refer either to a control *aldea* or a control farmer.) These variables are:

Design = 1 if PC = 1, 2, 3, 4, 5, 6, 7, 8, 10 or 11; 0 otherwise
Design, Selected for Treatment, DesSelForTrt = 1 if PC = 1, 2, 3, 8, 10 or 11; 0 otherwise
Design, Selected for Control, DesSelForCtrl = 1 if PC = 4, 5, 6, or 7; 0 otherwise
Treated = 1 if PC = 1, 2, 6, 8, or 9; 0 otherwise
TreatedAndDesSelForTrt = 1 if Treated=1 and DesSelForTrt=1, 0 otherwise
TreatedAndDesSelForCtrl = 1 if Treated=1 and DesSelForCtrl=1, 0 otherwise
*Aldea*Trt = 1 if PC = 1, 2, 3, 4, 8, 10 or 11; 0 otherwise
*Aldea*Ctrl = 1 if PC = 5, 6, or 7; 0 otherwise

A number of other indicator variables were defined and used during the course of the analysis (e.g., Married=1 if married or cohabitating, 0 otherwise), but these will not be discussed in this report unless they are worthy of note (i.e., of statistical and substantive significance).

Table A.4 shows the number of respondents in each of the preceding categories, by Round, for the binary categorical variables (0, 1) defined above for both the original experimental design and alternative design. The observations in the original experimental design correspond to Design = 1. The alternative design (i.e., the original design plus the sample of 600 Fintrac clients) corresponds to Design = 0 or 1.

Table A.4. Counts of Respondents by Categories of Interest in the Analysis

Classification	Level	Experimental + Alternative Design		Only Experimental Design	
		Round		Round	
		0	1	0	1
Design	0	545	479	0	0
	1	3981	2257	2735	2257
DesSelForTrt	0	2808	2521	2263	2042
	1	1718	215	472	215
DesSelForCtrl	0	2263	694	472	215
	1	2263	2042	2263	2042
Treated	0	3794	2089	2548	2089
	1	732	647	187	168
TreatedAndDesSelForTrt	0	4341	2570	2550	2091
	1	185	166	185	166
TreatedAndDesSelForCtrl	0	4524	2734	2733	2255
	1	2	2	2	2
*Aldea*Trt	0	2310	2076	1765	1597
	1	2216	660	970	660
*Aldea*Ctrl	0	2761	1139	970	660
	1	1765	1599	1765	1597

II.C. Impact Indicators of Interest

The FTDA program involves installation of high-productivity agricultural practices for horticultural crops (fruits and vegetables). The questionnaire collects data on household income and expenses in three categories: (labor market) employment, basic grains and other crops. The direct impact of the FTDA program is observed in the “other crops” category, which includes the crop types addressed by the program. Since households may substitute one form of income for another (e.g., plant less basic grains or engage in less employment while increasing other crops), we collected data on all sources of income and expense, to assess program impact.

The primary objective of this evaluation is to assess the impact of the FTDA on household income (off-farm and on-farm) and employment, as well as its effect on the cultivation of horticultural crops. The expectation was that there would be a marked increase in net household income, due to increased income generated through the sale of horticultural crops. We might expect income from basic grains to decline as a result; however, that decline would be offset by the much greater gains in the area of horticultural crops. Since household expenditures are positively correlated with income, and because they are usually reported more accurately by respondents than income, expenditures are often a good proxy for income measures. Within this context, the evaluation analysis focused on income and cost data for basic grains and other crops, employment income, as well as household net income and household expenditures. Income from crops is calculated as total crop value, not just the amount sold. That is, it includes the value of own consumption.

The key outcome variables (measures, indicators, response variables, explained variables, dependent variables) associated with income and expense are the following:

For basic grains (BG) (annual amounts):

- Income from basic grains (including used for own consumption) (IncBG)
- Expenses for inputs for basic grains (FactorBG)
- Transportation expenses for basic grains (TranspBG)
- Other costs for basic grains (OthCostBG)
- Labor expense for basic grains (measure of employment associated with BG) (LabExpBG)
- Total expenses, basic grains (ExpBG) = FactorBG + TranspBG + OthCostBG + LabExpBG
- Net income from basic grains (NetBG) = IncBG – ExpBG

For other crops (OC) – horticultural crops (annual amounts):

- Income from other crops (including used for own consumption) (IncOC)
- Expenses for inputs for other crops (FactorOC)
- Transportation expense for other crops (TranspOC)
- Other costs for other crops (OthCostOC)
- Labor expense for other crops (measure of employment associated with OC) (LabExpOC)
- Total expenses, other crops (ExpOC) = FactorOC + TranspOC + OthCostOC + LabExpOC
- Net income from other crops (NetOC)= IncOC – ExpOC

For labor-market employment (monthly amount):

- Income from labor-market ("employee") work (IncEmp)

For income and expenditures at the household level:

- Total household expenditures (TotHHExp) (monthly amount)
- Net household income (NetHHInc) = NetBG + NetOC + IncTotal*12 (annualized amount), where IncTotal = monthly household income from all sources (labor market, remittances, and other)

In addition to the preceding indicators of income, expense and employment, an indicator for harvesting of horticultural crops was available through the questionnaire:

> Production of horticultural crops: harvested horticultural crops (vegetables, fruits) in the last 12 months (not including home garden) (no = 1, yes = 2)

The indicators LabExpBG and LabExpOC are measures of employment. Since reported income is often not considered accurate, the expense measures (ExpBG and ExpOC) may constitute better measures of program impact than the reported income measures (IncBG and IncOC).

Table A.5A and A.5B presents basic characteristics of the distribution of income, expense and net income from basic grains (BG), other crops (OC), labor-market income (Emp) and total household (HH) for the baseline (Round 0) and endline (Round 1) data. The units for income and expense in the table (and most other tables that follow) are Honduran lempiras. The current exchange rate for the lempira is 18.9 lempiras to the US dollar for the period of the study.

Note, as discussed earlier, that income and expense amounts for crops are annual, household incomes and expenses (IncEmp, TotHHExp) are monthly, and NetHHInc is annualized. Since NetHHInc includes a component equal to 12 times the monthly total income (IncTotal), it has a very large variance. In the estimation of impact, this large variance will make it difficult to detect impacts relating toNetHHInc.

In all of the summary tables presented in this section, the means and standard deviations are simple unweighted values that do not take into account the design characteristics. The values presented in the tables are simply sample statistics, calculated using standard data-summary procedures such as Stata's *summarize* or *tabulate*. They should not be interpreted as estimates of population means or standard deviations – they are simply estimates of characteristics of the sample. In the impact estimation presented later, the design *is* taken into account, and the estimates and standard errors have desirable properties, such as unbiasedness or consistency.

Table A.5A. Basic Characteristics of the Distribution for Key Outcome Variables (Honduran Lempiras)
Baseline (Round =0), N=4,526

Indicator	Mean	Std. Dev	Min	Max
Income, basic grains (IncBG)	8976.86	39483.47	0	2166800
Expenses for inputs for basic grains (FactorBG)	2418.66	10568.58	0	507900
Transportation expenses for basic grains (TranspBG)	133.64	1719.45	0	112000
Other costs for basic grains (OthCostBG)	125.44	1120.71	0	30600
Labor expense for basic grains (LabExpBG)	1685.23	10032.47	0	324100
Total expenses, basic grains (ExpBG)	4362.98	17053.47	0	619900
Net income, basic grains (NetBG)	4613.87	29894.96	-287825	1546900
Income, other crops (IncOC)	24245.63	152281.1	0	7006750
Expenses for inputs for other crops (FactorOC)	3921.127	24822.28	0	939800
Transportation expenses for other crops (TranspOC)	335.633	3720.32	0	137500
Other costs for other crops (OthCostOC)	371.90	8885.19	0	557900
Labor expense for other crops (LabExpOC)	4482.36	46963.6	0	2052500
Total expenses, other crops (ExpOC)	9111.02	59633.56	0	2061450
Net income, other crops (NetOC)	15134.61	135858.1	-1267900	7005850
Labor market income (IncEmp)	6939.36	15994.66	0	460000
Total hhold expenditures (TotHHExp)	5375.21	4921.943	0	79644.13
Net household income (NetHHInc)	113914	263066	-1159058	8006891
Horticulture	1.9227	.2670	1	2
Note: All units of measure for the monetary indicators listed above are in Lempiras per year, with the exception of Labor Market Employment (IncEmp) and Total household expenditures (TotHHexp).				

Table A.5B. Basic Characteristics of the Distribution for Key Outcome Variables (Honduran Lempiras)
Endline (Round =1), N=2,736

Indicator	Mean	Std. Dev	Min	Max
Income, basic grains (IncBG)	9703.24	33967.62	0	995200
Expenses for inputs for basic grains (FactorBG)	2324.35	7519.96	0	201911
Transportation expenses for basic grains (TranspBG)	153.68	724.40	0	20000
Other costs for basic grains (OthCostBG)	147.16	1466.92	0	56000
Labor expense for basic grains (LabExpBG)	2433.22	11138.26	0	274000
Total expenses, basic grains (ExpBG)	5058.422	16016.66	0	277200
Net income, basic grains (NetBG)	4644.819	26480.76	-266294	727289
Income, other crops (IncOC)	34221.16	191685.90	0	6156000

Expenses for inputs for other crops (FactorOC)	4799.52	23963.5	0	456000
Transportation expenses for other crops (TranspOC)	449.73	3729.28	0	120000
Other costs for other crops (OthCostOC)	188.19	1816.42	0	50000
Labor expense for other crops (LabExpOC)	15106.66	229645.50	0	7776000
Total expenses, other crops (ExpOC)	20544.11	239307.2	0	7862500
Net income, other crops (NetOC)	13677.05	282061.6	-7784100	5895000
Labor market income (IncEmp)	9587.845	25095.99	0	900534
Total hhold expenditures (TotHHExp)	7760.885	11043.65	0	429396.70
Net household income (NetHHInc)	143183	465696	-5611403	16700000
Horticulture	1.9238	.2653	1	2

II.D. Treatment of Extreme Values

Virtually any large sample survey contains some extreme responses. Some of those extreme values may unduly influence the results, and decisions must be made on how to handle them. Standard alternatives for addressing this issue are imputation of missing values, censoring of extreme values and deletion (dropping) of observations containing missing or extreme values. We did not delete observations in this analysis, because it would have had an adverse effect on estimates of selection for treatment.

Casewise deletion of observations is routinely done by statistical software (such as Stata) during the course of model development (such as regression analysis), unless the missing values are imputed. Therefore, deletion of observations containing missing values or imputation of missing values is usually unavoidable at some point in the development of analytical models. The approach adopted here is to retain all observations in the model, and allow deletion of them only by the model-development software in cases in which missing values are not imputed. In cases in which there were few explanatory variables in a model and few missing values, casewise deletion was used. In cases for which casewise deletion would result in dropping many observations, such as a model containing many explanatory variables, missing values in regression models were imputed by substitution of the mean of the non-missing values. In general, the issue of missing values was not serious for regression models. (Before running regressions, tabulations were made to count missing values in each explanatory variable. The results of these tabulations are presented in the Stata .log file accompanying this document.)

Censoring of extreme values is problematic in the present application because the variables are interrelated (i.e., if a value is imputed for one variable, it must be consistent with the values of all related variables). In this analysis we examined the distribution of all components of income and expense for each of the two crop sources of income (BG and OC) and identified observations (households) for which any of the income or expense components exceeded the 99th percentile. For identified observations, we replaced the income value by the 99th percentile and the expense values by a value determined from a regression of the expense value on the income value. This procedure assures the consistency of all imputed income and expense components[34].

[34] The procedure used in the censoring is follows. If any of the components of income or expense exceeds the 99th percentile, then the values of all components were censored according to the following rules:

FactorBG = .22 IncBG
TranspBG = .036 IncBG
OthCostBG = .0063 IncBG
LabExpBG = .052 IncBG

The process of censoring is not without drawbacks. Some extreme observations are valid, and they will be censored along with erroneous ones. Although censoring will reduce bias by moderating the values of erroneous extreme values, it may introduce bias by altering values of legitimate extreme values. There is hence a trade-off between censoring at too high or too low a value. In the present study, all of the impact estimates involve the use of regression models, and it is considered that a somewhat stringent censoring is appropriate. Some legitimate large values of incomes and expense may be wrongly censored, but the nature of the relationships represented in the regression models will not be unduly affected. The observations that are censored in error will tend to be “well-off” households, and the focus of the program intervention is to reduce poverty, i.e., poorer households.

In addition to its role in reducing bias, censoring also has an effect on reducing variation, i.e., it is expected to reduce standard errors of estimates somewhat. Bias and precision (reliability) are two components of accuracy. Both are of concern, and it is viewed that the censoring contributed to improvements in both aspects in the present evaluation.

Note that in the analysis, a particular variable may appear in one instance as an explained variable (“dependent” variable) in a model and in another instance as an explanatory variable (“independent” variable) (and even sometimes as both, e.g., an endogenous variable). Once the decision was made to censor a variable, the censored values were used throughout the analysis, regardless of the role of the variable in a model (dependent or independent).

Table A.6 shows the same information as Table A.5, but for the censored data. It shows that the censoring caused a modest reduction in the means of the outcome variables, and a substantial reduction in the standard deviations. (While censoring of data may have some effect on estimation of means and totals, it usually has little effect on estimation of relationships, particularly when data are suitably transformed (e.g., logarithmic transformations of income and expense when used in linear regression models). In the present application, censoring was an effective means of removing erroneous data without unduly affecting the estimation of relationships and impact estimates based on them.)

FactorOC = .094 IncOC
TranspOC = .0065 IncOC
OthCostOC = .017 IncOC
LabExpOC = .067 IncOC.

Table A.6A. Basic Characteristics of the Distribution for Key Outcome Variables for Censored Data (Honduran Lempiras)
Baseline (Round =0), N=4,526

Indicator	Mean	Std. Dev	Min	Max
Income, basic grains (IncBG)	7682.06	14016.7	0	96680
Expenses for inputs for basic grains (FactorBG)	1926.24	3258.46	0	20940
Transportation expenses for basic grains (TranspBG)	92.53	263.78	0	2000
Other costs for basic grains (OthCostBG)	41.09	200.49	0	3000
Labor expense for basic grains (LabExpBG)	931.32	2923.82	0	31500
Total expenses, basic grains (ExpBG)	2991.39	5349.80	0	47900
Net income, basic grains (NetBG)	4690.67	10966.49	-33360	95330
Income, other crops (IncOC)	19102.70	65316.25	0	498825
Expenses for inputs for other crops (FactorOC)	2437.10	7737.55	0	80050
Transportation expenses for other crops (TranspOC)	120.2273	539.8585	0	7200
Other costs for other crops (OthCostOC)	96.57	588.36	0	6000
Labor expense for other crops (LabExpOC)	1877.78	8359.61	0	112500
Total expenses, other crops (ExpOC)	4531.68	13649.93	0	162400
Net income, other crops (NetOC)	14571.01	56736.83	-78880	497925
Labor market income (IncEmp)	6450.462	9851.20	0	70000
Total hhold expenditures (TotHHExp)	5375.22	4921.94	0	79644.13
Net household income (NetHHInc)	107379.1	157043.2	-16591	1395482
Horticulture	1.9227	.2670	1	2

Table A.6B. Basic Characteristics of the Distribution for Key Outcome Variables for Censored Data (Honduran Lempiras)
Endline (Round =1), N=2,736

Indicator	Mean	Std. Dev	Min	Max
Income, basic grains (IncBG)	8011.38	15467.11	0	96680
Expenses for inputs for basic grains (FactorBG)	1927.86	3252.45	0	20940
Transportation expenses for basic grains (TranspBG)	115.1102	303.746	0	2000
Other costs for basic grains (OthCostBG)	45.06	248.61	0	3000
Labor expense for basic grains (LabExpBG)	1350.26	3336.98	0	28800
Total expenses, basic grains (ExpBG)	3438.44	5790.30	0	40140
Net income, basic grains (NetBG)	4573.10	12197.07	-36700	95820
Income, other crops (IncOC)	25408.06	77543. 54	0	498825
Expenses for inputs for other crops (FactorOC)	2951.55	8514.69	0	82500
Transportation expenses for other crops (TranspOC)	176.99	661.21	0	7000
Other costs for other crops (OthCostOC)	86.46	570.84	0	6000
Labor expense for other crops (LabExpOC)	3229.49	10683.38	0	108000
Total expenses, other crops (ExpOC)	6444.51	17626.98	0	162000
Net income, other crops (NetOC)	18963.55	65655.95	-131550	495415
Labor market income (IncEmp)	8543.695	12257.64	0	70000
Total hhold expenditures (TotHHExp)	7626.59	7547.18	0	100000
Net household income (NetHHInc)	135910	195842.9	-26425	1330362
Horticulture	1.9238	.2653	1	2

Note that, after censoring, the maxima or minima may be exactly the same for both survey rounds.

Censoring was applied just to the income and expense variables specified above. There may be other extreme values in the data set. As models were developed, care was taken that the variables included in the model did not contain unreasonably extreme values, that might cause undue influence on the results.

To understand the reasons for differences in the various estimators discussed in this annex, it is helpful to compare the treated and untreated samples with respect to explanatory variables that are considered to have an effect on selection or on outcomes of interest, for Round 0 (for Round 1, the samples are likely to differ, because of the program intervention). Below is a table that compares selected variables that may affect selection or outcome, for the treatment and control samples, for Round 0 (a couple of outcome variables are also included in the table). Table A.7 compares the treatment and control populations with respect to variables that were statistically significant in the models used as bases for the estimators. The table shows the means of the variables. Ideally, it is desirable that the probability distribution of the explanatory variables be the same for the treatment and control populations. It is clear from this table that there are some substantial differences in these samples. This suggests that it may be necessary to include covariates in some models, either directly in an "outcome" model, or indirectly in a "selection" model (differencing removes differences in means, reducing the necessity to include covariates in models).

Table A.7. Basic Characteristics of the Distribution of Key Explanatory Variables for Treated and Untreated Samples (Round 0)

Indicator	Obs	Mean	Std. Dev	Min	Max
UNTREATED					
Household size	3974	4.95	2.32	1	17
Agricultural employees	3974	.576	.592	0	5
Total hectares of farm	3974	3.18	12.6	0	312.4
Mean education (years)	3974	3.55	2.24	0	20
Equipment value (lempiras, L)	2530	17775	83293	0	3007000
Rental value of installation (L/month)	2530	114	530	0	11000
Travel time to school (minutes)	3971	10.4	12.8	1	240
Travel time to hospital (minutes)	3738	122	64.6	1	690
Total household expenditure (L/month)	3974	4791	3876	0	67808.13
Income - basic grains (L/month)	3974	6623	12417	0	96680
Labor expenditure-basic grains (L/month)	3974	1028	3086	0	31500
Income – other crops (L/month)	3974	11501	46608	0	498825
Labor expense - other crops (L/month)	3974	2100	8953	0	112500
TREATED					
Household size	732	5.09	2.34	1	17
Agricultural employees	732	1.04	.849	0	7
Total hectares of farm	732	.929	5.23	0	74.55
Mean education (years)	732	5.11	2.92	0	17.5

Table A.7. Basic Characteristics of the Distribution of Key Explanatory Variables for Treated and Untreated Samples (Round 0)

Indicator	Obs	Mean	Std. Dev	Min	Max
Equipment value (lempiras, L)	732	56990	343805	0	8169000
Rental value of installations (L/month)	732	330	4567	0	120000
Travel time to school (minutes)	732	12.5	13.6	1	180
Travel time to hospital (minutes)	732	83.5	63.5	1	480
Total household expenditure (L/month)	732	8401	7815	0	79644.13
Income - basic grains (L/month)	732	13173	19498	0	96680
Labor expenditure-basic grains (L/month)	732	430	1789	0	25000
Income – other crops (L/month)	732	58501	115250	0	498825
Labor expense - other crops (L/month)	732	728	3891	0	45000

As a final summary of the “raw” data, we present tables of means for selected outcome variables and by treatment and round. (Standard deviations are included in parentheses, following each mean. As discussed earlier, these are not standard errors of the mean, considered as a population estimate. They are simply the standard deviation (ignoring the design) of the observations, within each table category.) This is done for all responding households (Table A.8A) and for households that were interviewed in both survey rounds (Table A.8B). (The reason for presenting the second table, for households interviewed in both survey rounds, is that difference estimates are more precise for such samples.) Also included, in the last column, is the double difference of the means of the four design groups (treatment before, treatment after, control before, control after). It is unadjusted for design features (household, selection probability), and is hence referred to as the unadjusted or “raw” double difference. These “raw” double differences are provided as a simple way of exhibiting the overall characteristics of the sample. No standard error is presented for the raw double differences – standard errors for design-adjusted or covariate-adjusted double-difference estimates will be presented later, in the analysis section, taking into full account the survey design and covariates. What is clear from the tables is that the raw double differences are small compared to the means. This preliminary review of the sample data suggests, before any analysis is done, that the estimated program impacts will likely be small. The reason for this is that for many programs, because of selection effects, the “raw” effects tend to be larger than the “adjusted” effects, taking into account design and covariates.

Table A.8A. Table of Means of Outcome Variables by Treatment and Round, for All Surveyed Households (standard deviations are included in parentheses)

Outcome Variable	Not Treated (Treatment = 0)		Treated (Treatment = 1)		Unadjusted (“Raw”) Double Difference
	Round 0 (n=3794)	Round 1 (n=2089)	Round 0 (n=732)	Round 1 (n=644)	
IncBG	6623(12417)	6966(13740)	13173(19498)	11394(19714)	-2122
ExpBG	2824(5398)	2889(5414)	3860(5002)	5214(7029)	1289
NetBG	3799(9420)	4077(10983)	9313(16081)	6180(15399)	-3411
LabExpBG	1027(3086)	1181(3159)	430(1789)	1901(3814)	1317
IncOC	11501(46608)	13029(51259)	58501(115250)	65520(112159)	5491
ExpOC	3578(13126)	3353(12547)	9472(15173)	16428(26028)	7181
NetOC	7923(38698)	9675(43050)	49029(103623)	49092(105469)	-1689
LabExpOC	2100(8953)	2059(8756)	728(3891)	6997(14731)	6310
IncEmp	5296(7935)	6908(9579)	12435(15208)	13781(17411)	-266
TotHHExp	4791(3876)	7043(6851)	8401(7815)	9387(8767)	-1266
NetHHInc	85625(124757)	106015(146916)	220079(238135)	232463(284073)	-8006
Horticulture	1.913(.281)	1.923(.266)	1.955(.208)	1.925(.264)	-.040

Table A.8B. Table of Means of Outcome Variables by Treatment and Round, for Households Interviewed in Both Survey Rounds (standard deviations are included in parentheses)

Outcome Variable	Not Treated (Treatment = 0)		Treated (Treatment = 1)		Unadjusted (“Raw”) Double Difference
	Round 0 (n=2089)	Round 1 (n=2089)	Round 0 (n=644)	Round 1 (n=644)	
IncBG	6738(12960)	6966(13740)	13173(19498)	11394(19714)	-2300
ExpBG	2764(5190)	2889(5214)	3860(5002)	5214(7029)	1203
NetBG	3974(9952)	4077(10983)	9313(16081)	6180(15399)	-3502
LabExpBG	1004(3010)	1181(3159)	430(1789)	1901(3814)	1309
IncOC	9623(42272)	13029(51259)	58501(115250)	65520(112159)	4302
ExpOC	2947(11407)	3353(12547)	9472(15173)	16428(26028)	6513
NetOC	6676(36102)	9675(43050)	49029(103623)	49092(105469)	-2211
LabExpOC	1699(7564)	2059(8756)	728(3891)	6997(14731)	5874
IncEmp	5266(7653)	6908(9579)	12435(15208)	13781(17411)	337
TotHHExp	5001(4126)	7043(6851)	8401(7815)	9387(8767)	-559
NetHHInc	85323(120939)	106015(146918)	220079(238135)	232463(284073)	37
Horticulture	1.902(.298)	1.923(.266)	1.955(.208)	1.925(.264)	-.048

It is of interest to compare the means of outcome variables for treated farmers of the original experimental design to those for the Fintrac-selected supplementary sample. The following table (Table A.9) is a table of means for selected outcome variables, by round, for the treated farmers in two cohorts of the original experimental design and the supplemental sample of 545 Fintrac-selected clients. The sample designated as “ED1” in the table is the sample from Producer

Categories 1 and 2. The sample designated “ED2” is the sample from Producer Categories 1, 2, and 8. The sample designated as “Fin” is the sample from Producer Category 9. (Student t tests may be applied (and were applied) to compare the means for ED1 to those for Fin, and to compare the means for ED2 to those for Fin. They are statistically significant in many cases.) The table shows (based on the t tests) that the baseline means are higher for a number of outcome variables for the supplementary sample of Fintrac clients, and that the increase from endline to baseline for other crops (which included horticultural crops) is substantially larger. (The values of zero for labor expenses in Round 0 for the Fintrac sample appear to be an error. If so, this error would tend to introduce a positive bias into the estimate of program impact. This bias would be small, since the mean for labor expenses is not large.)

Table A.9. Table of Means for Selected Outcome Variables, by Round, for Treated Famers in Two Cohorts of the Experimental Design (ED1 and ED2) and the Supplemental Sample of Fintrac-Selected Clients (Fin). (Standard deviations are in parentheses.)

Outcome Variable	Round 0			Round 1		
	ED1 (n=28)	ED2 (n=185)	Fin (n=545)	ED1 (n=26)	ED2 (n=163)	Fin (n=479)
IncBG	12077(13109)	12259(16713)	13511(20390)	8787(10706)	8499(15228)	12412(20976)
ExpBG	8060(8680)	4489(5235)	3656(4881)	3757(4936)	3913(4918)	5676(7637)
NetBG	4017(8686)	7771(15132)	9855(16396)	5030(7289)	4585(12681)	6736(16224)
LabExpBG	3271(6138)	1703(3426)	0(0)	1550(2317)	1547(2946)	2030(4068)
IncOC	57081(109018)	38399(86959)	65539(122846)	55170(104282)	34406(80661)	76381(131920)
ExpOC	14090(22276)	7020(14812)	10340(15232)	14765(25884)	11122(22690)	18302(26881)
NetOC	42991(294776)	31379(76474)	55199(110122)	40405(89604)	23284(68351)	58079(113144)
LabExpOC	5801(11690)	2881(7342)	0(0)	8278(20522)	5955(15452)	7381(14501)
IncEmp	11264(14583)	9547(11985)	13456(16058)	10524(14065)	10900(14025)	14800(18354)
TotHHExp	5837(3848)	8242(8243)	8480(7673)	8381(6284)	8480(6779)	9720(9341)
NetHHExp	190141(265538)	167052(202264)	238828(246829)	180530(181308)	169154(204283)	254707(303982)
Horticulture	1.857(.356)	1.935(.247)	1.961(.193)	2(0)	1.914(.281)	1.928(.259)

III. Estimation of Impact

We present impact estimates for all outcome variables listed in Section II.C of this Annex, with the exception of Factor Expense, Transport Expense and Other Cost. It is important to consider the estimates as a group, and not individually, since they are correlated. For example, if a farmer increased his production of other crops, he may have to reduce his production of basic grains (because of limitations on land or other resources).

For the sake of simplicity, when practical, models were developed in original (untransformed) variables. In some instances, however, to improve the quality of the model and of estimates based on it, we chose to transform incomes and expenses to logarithms.

Standard errors are presented for all statistical estimates of impact. To approximately assess the statistical significance of an estimate, divide the estimate by its standard error. Results exceeding two in magnitude are of moderate statistical significance (the likelihood that the estimated effect size exceeds its standard error in magnitude by a factor of two is about one in twenty, if the effect is in fact zero). An approximate 95 percent confidence interval for the estimate is defined by the estimate plus and minus two standard errors. (More precisely, an effect is considered statistically significantly different from zero if it differs from zero by more than 1.96 times its standard error, for two-sided tests of hypothesis (i.e., the effect may be either positive or negative), or by more than 1.645 times its standard error, for one-sided tests of hypothesis (i.e., the sign of the effect is specified). On the indication of statistical significance, the interpretation is that over many independent investigations, the probability that the confidence interval includes the true value of the parameter is approximately .95. Confidence intervals within the same investigation are correlated.)

III.A Impact Estimators of Interest[35]

For a highly structured experimental design, such as was originally planned for this evaluation, there is a single design-based formula that may be used to estimate impact. For the revised design that was ultimately used for the evaluation, there are a number of alternative approaches and formulas for estimation of impact. There are alternative models and estimators. We considered several of these alternative approaches, and identified one that appeared to produce the highest level of validity and power for the impact estimates. That estimator is a "modified regression-adjusted propensity-score-based estimator." In addition to estimating the average treatment effect (ATT), we also estimated the average treatment effect on the treated (ATT). The ATE is an estimate of the impact on a randomly selected program-eligible farmer. The ATT is an estimate of the impact on a randomly selected treated farmer.

Here follows a list of all of the impact estimators that were examined in detail in the course of the analysis:

"Statistical Estimator"

1. Basic propensity-score-based estimator of average treatment effect (ATE)

"Econometric Estimators"

2. Regression-adjusted propensity-score-based estimator of ATE
3. Modified regression-adjusted propensity-score-based estimator for ATE

Estimators Considered, but Not Presented

4. Regression estimator for ATE, not based on the estimated propensity score
5. Instrumental-variable (IV) regression estimator for ATE, based on the estimated propensity score.

In addition to the preceding estimators, consideration was given to estimating the Intention-to-Treat (ITT) estimator and the Local Average Treatment Effect (LATE) estimator, but the size of the treatment sample from the original experimental design was too small for these estimators to be of value.

Of the preceding estimators, results are presented in the main text only for the third estimator. This Annex presents discussion of several of the other estimators since it is of some interest in understanding why the third estimator was selected as the estimator of choice.

[35] The estimators used to assess program impact were discussed in detail in the *Analysis Plan*, and were summarized in the introduction to this chapter. The mathematical notation used for the estimation formulas follows *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, by Jeffrey M. Wooldridge (MIT Press, 2010, first edition 2002)). Many of the formulas presented in this reference pertain to the case of a single cross-section of data, and must be modified as appropriate for panel data. The estimators considered in the analysis does not include all estimators considered in Wooldridge, but it includes the major ones.

Most of these estimators can be obtained by linear regression. For some of the models, the explanatory variables of the regression model are simply the design parameters, such as household, Round and Treated. For the more complex estimates, the regression models include both design parameters and other explanatory variable (covariates such as family size, education of head of household, assets, or an estimated propensity score). In the following, we will generally use the term "regression estimate" to refer to the case in which explanatory variables other than design parameters are included, and the term "propensity-score-based estimate" to refer to the case in which the major explanatory variable other than the design parameters is the estimated propensity score.

(The term "propensity score" arises frequently in evaluation, usually in the context of matching a non-randomly-selected control group to a treatment sample. For clarification, it is pointed out that the matching(of *aldea* pairs) prior to randomized assignment to treatment that was done in constructing the sample survey design for this evaluation project had nothing to do with propensity scores. The use of propensity scores in this evaluation is restricted to the analysis.

(In the "matched (*aldea*) pairs" design used in this study, propensity score matching was not used since, although it may reduce bias, it may also decrease precision, since units that match on the propensity score may be very dissimilar with respect to explanatory variables that have an important effect on outcomes of interest. Instead of propensity score matching, an "importance score" method was used to match units on variables considered to have an important effect on outcomes of interest. The purpose of the *aldea* matching prior to randomization was to increase precision and power. It was not at all to reduce selection bias – that was taken care of through the use of an experimental design that include randomized assignment to treatment.

(Although propensity score matching was not used to form the matched *aldea* pairs of the survey design, the analysis used propensity-score-based estimators (both nonparametric "matching" estimators based on the propensity score and (parametric) regression models that included the propensity score as a covariate.)

The analysis that follows discusses and presents results for several of the estimators listed above. It may be asked why this analysis examined several estimators of impact. The general approach to modeling used in this analysis is causal modeling and counterfactuals. (For a description of this approach, see *Causality: Models, Reasoning, and Inference 2nd ed.* by Judea Pearl, Cambridge University Press, 2009.) Within this general approach, a wide variety of estimators may be used. Which estimator is selected as a preferred one depends on a number of considerations, including: (1) the reasonableness (face validity) of the causal model with which the statistical model specification and estimator are associated; (2) statistical tests of the validity and reliability of the statistical models (e.g., goodness-of-fit tests; specification tests (such as a Hausman test of the equivalence of fixed-effects and random-effects models)); and (3) an *ex post* statistical power analysis, which showed that the tests of hypothesis based on the third estimator were substantially more powerful than those based on the other estimators. Based on these considerations, most confidence was placed in the third estimator, and estimates of impact are presented in the main text of this report only for this estimator. Pearl op. cit. presents a high-level discussion of causal models, discussing model specification and identification (estimability) in very general terms. For discussion of specific estimation techniques, see Wooldridge, op. cit.

Under certain conditions (such as conditional independence), all of the impact estimates presented above are consistent estimators of impact (i.e., of the average treatment effect). Under reasonable assumptions that apply to this project, the preceding estimators are consistent estimates of impact (i.e., the expected value of the sample estimate converges to the desired population value as the sample size becomes large).

Regardless of the estimator used, it is necessary that it take into account the design features (such as two-stage sampling of *aldeas* and panel sampling of households). In many cases the estimator has the same value whether the design is correctly accounted for, but to obtain correct estimates of the standard errors of the estimates, and to make correct statistical tests of hypotheses, the design characteristics must be correctly represented in the data model. Other than selection for treatment, the principal design feature for this evaluation is the fact that (in most instances) the same households are interviewed in both survey rounds (i.e., the design is a "strongly balanced" panel design). Once this (longitudinally matched pairs) feature has been taken into account, most other design features (e.g., *aldea*) and covariates (e.g., whether a farmer owns his own land) are of secondary importance.

The probabilities of selection (of the households) are variable, and are a prominent design feature. They are determined by the stratification of households according to variables believed to have an effect on outcomes of interest. The selection probabilities are used in two ways. First, models are constructed with and without consideration of the selection probabilities (i.e., with and without "weights," where the weight for a household is the reciprocal of its probability of selection), and compared. If the two models are similar, this is taken as evidence that the model specification is correct.[36] If the two models differ substantially, this is taken as evidence that the model specification is not correct, and a better model specification is sought. If a better specification cannot be found (either in terms of the same variables, or by adding other variables), then (the second way in which weights are used) consideration is given to use of weighted estimates. The Stata *xtreg* procedure used for much of the analysis does not accommodate weights, and so using weights is done only in particular circumstances (e.g., in analysis of a single survey panel, or by transforming the data using *xtdata*, and using non-panel procedures that allow weights).

Regression models were developed with and without sample "weights" (reciprocals of the probabilities of selection). Little difference was observed between the weighted and unweighted estimates. This is a strong indication that the model is correctly specified.

All of the impact estimators considered here take into account that the evaluation design is a pretest-posttest-comparison-group design. In all cases, the estimators are similar to double-difference estimators of impact (or the interaction effect of treatment and time). This type of estimator is not sensitive to the actual level of a variable of interest (e.g., income, or even net income), but to differences in the relative change in the variable over time, between the treatment and control groups.

The process of double differencing removes the mean levels of variables in the four design groups (treatment before, etc.). For this reason, the fact that income may be underreported in

[36] Note that there are other tests for specification, such as the "principle of conditional error" or the so-called "Hausman" test, which compares model parameters for fixed-effects and random-effects specifications.

some cases is not a concern, as long as the underreporting is not related to response (outcome). It is important to realize that the impact estimators measure the interaction effect of treatment and time, which is similar to a double difference. This effect is not a *level*, and it is not an *increase or decrease*. If it is reported that the income effect of treatment is 10,000 lempiras, this does not mean that income increases on average by this amount if the program services are received. Rather, it means that the incomes of the treated farmers after four years in the program will be about 10,000 lempiras more than the incomes of untreated farmers, for program farmers in *aldeas* randomly selected from an eligible population. It is important to keep this distinction in mind when reviewing the impact tables presented in this report.

Any impact indicator variable that is strongly correlated with another may be used as a surrogate, or alternative, estimate for it. For example, since income from basic grains is about three times total expense for basic grains, the income effect for basic grains is about three times as large as the expense effect. Since reported income may not be as accurate as reported expense, the expense impact times three may be a better estimate of the income effect than the income effect estimated from reported incomes.

III.B Assumptions about the Stochastic Nature of Explanatory Variables

As discussed in the main text, estimates of impact are based on fixed-effects estimators. Random-effects estimators are used to assess the relationship of outcome to time-invariant variables (such as propensity score). For all panel regression models considered in the analysis (i.e., all models estimated using Stata procedure *xtreg*), a Hausman test was applied to test the equivalence of fixed-effects and random-effects models. The Hausman test is a test of whether the unobserved variables of the model are correlated with the explanatory variables. The results of these test showed that the two model specifications were similar in most instances.

III.C Observed Treatment Effect

To facilitate understanding of the methodology, detailed description of the procedure (formula or regression-analysis procedure) for obtaining the impact estimates will be presented in the case of a single selected outcome measure (ExpOC). Readers familiar with Stata may execute the Do14FTDAImpactEstimation.do file (or examing the Do14FTDAImpactEstimation.log file) to obtain detailed information for the other outcome measures.

This Annex will now present a detailed description for three impact estimators. In the course of the data analysis other estimators were examined, but they did not show positive program effects. To better understand why some estimators fail to show positive program effects, it is helpful to examine the raw double-difference estimator, or Observed Treatment Effect (OTE). The OTE is the estimated double difference taking into account the sample design features but no other covariates (such as household characteristics or an estimated propensity score). The OTE is of interest as a basis to which the other estimates (to be presented) may be compared, to assess the effect of taking into account model features (such as a first-step selection model) and covariates. (The OTE differs from the “raw” double difference presented earlier in that the raw double difference does not take into account any design features – it is based solely on the unadjusted sample means of the four design groups (treatment before, treatment after, control before, control after).)

The following table presents the OTE for the outcome measures considered here, taking into account the design feature that the same households are interviewed in both survey rounds, but no other explanatory variables. Standard errors are included since these estimates take into account the design features.

Table A.10. Estimates of Observed Treatment Effect (OTE)

Outcome Variable	Estimate	Standard Error
Basic Grains (BG)		
IncBG	-2300*	672
ExpBG	1202*	257
NetBG	-3502*	585
LabExpBG	1310*	171
Other Crops (OC)		
IncOC	4301	3205
ExpOC	6513*	738
NetOC	-2211	2893
LabExpOC	5873*	490
Labor Market Employment and Household Income and Expenditures		
IncEmp	336	517
TotHHExp	-560	319
NetHHInc	237	7792
Production of Horticultural Crops		
Horticulture	-.0427*	.0183

Income and expense measured in Honduran lempiras.

The salient feature of the preceding list is that, at first look, it appears that the program has no positive effects. Even worse, although the effect for NetOC is not statistically significant, it is of unexpected sign (negative). The implication of this situation is that, unless selection (for participation) is an important factor affecting outcome, the program appears to be of no value. The propensity-score-based estimators are based on a strong logistic-regression model of selection, and they estimate positive results for the program. The other two estimators examined did not represent the selection process well (they did not explicitly represent the selection process), and they failed to show positive results. Because these models are weak, they reflect the OTE.

III.D. Calculation of Estimates

1. Basic Propensity-Score-Based Estimates of the Average Treatment Effect (ATE) ("Statistical Approach")

This section presents estimates of the average treatment effect (ATE) using a basic propensity-score-based estimator.

In order to obtain a good (unbiased or consistent) estimate of impact, it is necessary to take into account the procedure used to select farmers for treatment (receipt of FTDA services). This may be done in either of two ways. The first approach is to develop a linear statistical model that specifies the relationship of an outcome variable of interest (y) to explanatory variables, including all variables believed to affect outcome and selection, and obtain an estimate of impact from this model. The second approach is to develop a separate logistic-regression selection model that estimates the probability of selection, or "propensity score," and construct an estimate of impact based on the estimated propensity score. Both approaches are based on the same data (explanatory variables), but the model specifications differ. Both approaches were used in this analysis, but it was concluded that the second approach (based on propensity scores) was better suited to the present application (i.e., the face validity of the model relative to the causal model was better, and the model goodness-of-fit (based on statistical measures, such as power) was better). For this reason, more confidence is placed in the logistic-regression-model propensity-score-based estimators (the first three estimators listed earlier) than in the linear regression estimators (the last two estimators).

For both approaches, it is necessary to condition on variables (other than treatment) that can affect both selection for treatment and outcome. As discussed in the main text, this may be done by conditioning on variables that affect selection for treatment "conditioning to balance" or by conditioning on variables that affect outcome ("conditioning to adjust").

The following model was developed to estimate the probability of selection of a household for treatment (provision of program services by Fintrac). This model used all of the survey data, including not only the sample of Fintrac clients but also all of the households from the original experimental design and all of the potential lead farmers in the treatment *aldeas* rejected by Fintrac. Ordinarily, it is not appropriate (useful) to include data from a randomized experimental design in a selection model (since, because of randomized assignment to treatment, the distributions of all explanatory variables are the same for the treated and untreated populations). In the present application, however, the data from the original experimental design *were* included in the selection model, since randomization was applied at the level of the *aldea*, not the individual farmer.

The selection model was a binary selection model developed using a logistic regression model. The term "selection" here refers to selection of a farmer by Fintrac for provision of program services. It refers to program participation, not to selection in the original experimental design (i.e., by randomized assignment of *aldeas* to treatment, and selection of potential lead farmers by NORC). There is a potential for confusion here, since in many studies, in which selection for treatment implies treatment, this is referred to as selection for treatment. To avoid confusion, we shall generally use the terms "Participated" or (participation indicator) or "Treated" or "treatment indicator" rather than the more customary terms "selected" or "selection indicator."

In addition to participation, there are two other selection effects that could be taken into account in the present analysis. These effects are attrition (leaving the program prior to the second-round survey) and nonresponse in the second round (for a variety of reasons, such as not-at-home, death and relocation). An analysis of nonresponse in the second round failed to show a strong relationship to explanatory variables, and so no selection model was developed for second-round nonresponse (i.e., the "model" is "missing at random"). (It is noted that the 224 potential lead farmers in treatment *aldeas* that were rejected by Fintrac were not interviewed in Round 1, so these are not included in any selection model of nonresponse.) Unfortunately, in Round 0 there was no variable in the questionnaire that indicates whether a household was participating in the FTDA program, so it is not possible to accurately measure attrition. A rough measure could be obtained by counting households that grew other crops in Round 0 but not in Round 1, but this is considered a poor measure, and there is little point to considering an estimator based on a poor instrumental variable. For these reasons, the selection model is based solely on participation, as measured by Treated.

The selection model to be developed is used in all of the impact estimators to be analyzed in detail. The utility of the propensity score is that the counterfactual outcomes are independent of treatment, conditional on the propensity score. This condition is referred to as ignorability of treatment. The basic Rosenbaum-Rubin method (the "statistical approach") compares observations that have similar propensity scores. For this method, observations having propensity scores of 0 or 1 are not useful (since they are all treated or all untreated). The econometric approach can accommodate these cases. In the econometric approach, attention is focused on variables on which the propensity score is based, and the outcome models are conditioned on these variables plus other variables that affect outcome. In view of the fact that, with a large number of explanatory variables available, there may be many different choices of variables on which to base the selection model, this set of variables is not unique.

Let y denote the participation indicator random variable, which has the value 1 if a household is provided services and 0 otherwise. We define a binary response model:

$$P(y=1|\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta}) \equiv p(\mathbf{x})$$

where **x** denotes a (column) vector of explanatory variables, P(y=1|**x**) denotes the probability that y=1 (i.e., is treated) conditional on **x**, **β** is a vector of parameters and g(.) is a the logistic link function,

$$g(z) = \exp(z)/(1 + \exp(z)).$$

If we define z as

$$z = \mathbf{x}'\boldsymbol{\beta} + e,$$

where e denotes a random error term uncorrelated with **x** and with mean zero, then

$$y = 1 \text{ if } g(z)>.5 \text{ and } 0 \text{ otherwise.}$$

The expression **x'β** is referred to as an index. The parameters **β** are estimated by the method of maximum likelihood. The expression **x'β** does not have any meaning (or units) – it is simply a modeling artifact. The model is often referred to as a "latent variable" model, since the variable z is unobserved.

The identification of the explanatory variables to include in the selection model is guided by an underlying causal model. Variables are selected from the questionnaire that are considered likely to have an effect on selection. The questionnaire variables are correlated, and it is attempted to make a selection that is not highly intercorrelated, yet reflects the underlying factors that may affect selection. The selection model uses data only from the first survey round (baseline).

Using the Stata *logistic* procedure, the index was estimated to be

x'β = -10.87894 - .1250273*HouseholdSize + .1594562*FormalEducHead + .868967*AgEmployees - .0870384*TotHaOwnFarm + .0211418*TimeToSchool - .0103442*TimeToHosp + .9303271*LogTotHHExp + .2160815*LogIncBG - .2096164*LogLabExpBG + .2420389*LogIncOC - .2358687*LogLabExpOC where

HouseholdSize = number of persons in the household
FormalEducHead = years of formal study of head of household
AgEmployees = number of household occupants in agricultural work
TotHaOwnFarm = total farm hectares owned
TimeToSchool = travel time in minutes to school
TimeToHospital = travel time in minutes to hospital.
LogTotHHExp = logarithm of total monthly household expenditures
LogIncBG = logarithm of value of production of basic grains
LogLabExpBG = logarithm of manual-labor expenditures for basic grains
LogIncOC = logarithm of value of production of other crops
LogLabExpOC = logarithm of manual-labor expenditures for other crops.

The Stata program package includes two "panel" logistic regression procedures, *xtlogit*, for fixed and random effects, and *xtmelogit*, for mixed models. Neither of these was considered appropriate for this application. The selection indicator variable is determined by observables at Round 0. The programs *xtlogit* and *xtmelogit* are intended for use in applications in which the binary selection variable is changing in each round, such as membership in a union. The ordinary *logit* procedure was considered the appropriate procedure for this application.

The selection model presented above includes only variables that were highly statistically significant. The value of the "pseudo R^{2}" (a standard measure of model fit) for this model is .44. (In general, R^2, called the "coefficient of determination," is the square of the multiple correlation coefficient, R. R^2 indicates the proportion of the variation (variance) in the dependent variable that is explained by the model.) For this type of application, the value R^2 = .44 is considered a relatively high value.

There were a few missing values in some of the explanatory variables. In order to retain all of the observations for the regression analysis, these missing values were imputed as means of the non-missing values.

Some of the variables included in the model are logarithms of variables, and these are undefined for nonpositive values of the argument (of the logarithmic transformation). These undefined values were replaced by zeros, and indicator ("dummy") variables included in the model to account for the nonlinearity of this transformation. The inclusion of the dummy variables made little difference in the model fit (R^2 increased from .44 to .46), but the interpretation of the model parameters became more difficult. As a result, this alternative model was not considered further. (As noted, the coefficients in a logistic model have no meaning, and so inclusion of logarithmic terms without corresponding dummies does not present conceptual problems.)

The interpretation of each of the included variables is as follows:

HouseholdSize (negative coefficient): larger households are less likely to participate
FormalEducHead (positive coefficient): farmers with more formal education are more likely to participate
AgEmployees (positive): households having more agricultural-sector employees are more likely to participate

TotHaOwnFarm (negative): the larger the owned farm hectares, the less likely the farmer is to participate
TimeToSchool (positive): the closer the school, the higher the likelihood of participation
TimeToHospital (negative): the more remote the household, the lower the likelihood of participation
LogTotHHExp (positive): households with larger total household expenses are more likely to participate
LogIncBG (positive): the higher the basic-grains income, the higher the likelihood of participation
LogLabExpBG (negative): the higher the basic-grains labor expense, the lower the likelihood of participation
LogIncOC (positive): the higher the other-crops income, the higher the likelihood of participation
LogLabExpOC (negative): the higher the other-crops labor expense, the lower the likelihood of participation.

All of the preceding variables are listed in the list of causal factors and variables affecting selection, in the main text. Note that the participation model reflects both the decision of Fintrac to accept a farmer into the program as well as the decision of the farmer to participate. The explanatory variables included in the model could reflect either type of decision, or both.

Remarks on Model Specification, Identification and Estimation

Estimation of program impact involves consideration of both the selection model and the models of outcomes of interest. The essential feature of these model pairs is that variables that affect both selection and outcomes of interest be observable ("selection on observables"), or, if not unobserved ("selection on unobservables"), be time-invariant. The following comments are made about the nature of the selection model and its relationship to the outcome models to be considered.

1. The household variables are correlated. There is not a unique model that describes the probability of selection, but an infinite variety of such models. The goal is to include a set of explanatory variables that reflects the important factors affecting selection, yet for which the intercorrelations are as low as possible. During the course of the analysis, a number of alternative selection model specifications were examined, including more, fewer and different variables than were listed above. For reasonable specifications, the results were similar (i.e., the value of R^2 was similar). The preceding model is one such model.

2. The selection model is determined solely from Round 0 (baseline) data, since selection into the program is made at Round 0.

3. It is not the goal to estimate individual parameters (coefficients) of the selection model. The goal is to estimate the propensity score (i.e., the explained variable), not individual coefficients. The individual coefficients of the selection model are not used in the analysis. For complex link functions, the parameters do not have a straightforward economic interpretation. Moreover, not only is the selection of explanatory variables not unique, but the explanatory variables are correlated, so that the estimates of individual

coefficients are also correlated (confounded). The situation is similar to the problem of forecasting – the goal is to estimate, or forecast, the response variable, not to estimate the marginal effect of response to individual explanatory variables. It does not matter is the estimates of individual coefficients are biased or imprecise, because their magnitudes are of no interest – what matters is that they reflect factors affecting selection, so that the value of R^2 is high. Care must be taken to avoid including too many explanatory variables in the selection model, to avoid overfitting the model.

4. There may be unobserved variables affecting selection for treatment, and these unobserved variables may be correlated with the explanatory variables included in the model (in which case ordinary least squares estimates of the model parameters are biased. As mentioned, the essential concern is whether the unobserved variables affecting selection and outcomes of interest are time-invariant, in which case they drop out of the (fixed-effects, two-round panel) outcome model, so that the assumption of conditional independence is justified. It is desirable to have a high value for R^2, since this promotes high precision of the impact estimates. From the viewpoint of bias, however, the value of R^2 is not important. What is important is that all variables affecting both selection for treatment and outcomes of interest be observable or, if not, be time-invariant.

5. Considered over time, a number of the explanatory variables in the selection model may be affected by the explained variable, i.e., they may be endogenous. Applying the ordinary least squares estimation procedure to cross-sectional data, estimates of the model parameters are biased. As discussed, estimation of the parameters is not the objective – the objective is estimation of the propensity score. Selection for the program is based on variables that are available at baseline (Round 0), and selection status does not vary over time. For the selection model, endogeneity is not an issue.

6. It is desired to include a set of observed variables that collectively do a good job of estimating the probability of selection (as reflected in the value of R^2). With respect to unobserved variables, the assumption is made that unobserved variables that have an important effect on both selection and on outcomes of interest are time-invariant. The issue of unobserved variables was discussed at length earlier. Unobserved variables that do not affect both selection and outcomes of interest are not a primary concern. They may reduce the value of R^2, which is certainly undesirable, and they may bias the estimates of the selection model parameters, but they do not corrupt (bias) the estimation of impact.

7. For the basic propensity-score method of estimation, it is required that the selection model include probabilities not equal to zero or one, since observations having such values are of little use to the estimation of impact for this method. For the econometric approach, it is required that the selection model contain at least one variable that is not included in the outcome model (this assumption will hold for all of the outcome models to be considered).

The output of the Stata procedure for determining the logistic selection model is shown in Figure A.1.

Figure A.1. Logistic Regression Estimation of Participation Model

```
Logistic regression                                 Number of obs   =       4302
                                                    LR chi2(11)     =    1739.87
                                                    Prob > chi2=     0.0000
Log likelihood = -1092.3351                         Pseudo R2       =     0.4433

------------------------------------------------------------------------------
     Treated |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
HouseholdS~e |  -.1250273   .0264687    -4.72   0.000    -.1769051   -.0731495
FormalEduc~d |   .1594562   .0162117     9.84   0.000     .1276818    .1912306
 AgEmployees |    .868967   .0986173     8.81   0.000     .6756806    1.062253
TotHaOwnFarm |  -.0870384   .0162219    -5.37   0.000    -.1188328   -.0552439
TimeToSchool |   .0211418   .0040684     5.20   0.000     .0131678    .0291158
TimeToHosp~l |  -.0103442   .0010603    -9.76   0.000    -.0124223    -.008266
 LogTotHHExp |   .9303271   .0925013    10.06   0.000     .7490279    1.111626
    LogIncBG |   .2160815   .0171065    12.63   0.000     .1825534    .2496096
 LogLabExpBG |  -.2096164   .0208682   -10.04   0.000    -.2505173   -.1687156
    LogIncOC |   .2420389   .0138435    17.48   0.000     .2149061    .2691717
 LogLabExpOC |  -.2358687   .0223509   -10.55   0.000    -.2796757   -.1920617
       _cons |  -10.87894   .8008922   -13.58   0.000    -12.44866   -9.309223
------------------------------------------------------------------------------
Note: 2 failures and 0 successes completely determined.

.
. *Postestimation analysis.
.
. estat clas if Round==0 & !(idhh>4000 & idhh<5000)

Logistic model for Treated

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       424            99  |        523
     -     |       308          3471  |       3779
-----------+--------------------------+-----------
   Total   |       732          3570  |       4302

Classified + if predicted Pr(D) >= .5
True D defined as Treated != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   57.92%
Specificity                     Pr( -|~D)   97.23%
Positive predictive value       Pr( D| +)   81.07%
Negative predictive value       Pr(~D| -)   91.85%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    2.77%
False - rate for true D         Pr( -| D)   42.08%
False + rate for classified +   Pr(~D| +)   18.93%
False - rate for classified -   Pr( D| -)    8.15%
--------------------------------------------------
Correctly classified                        90.54%
--------------------------------------------------
```

```
. estat gof if Round==0 & !(idhh>4000 & idhh<5000)

Logistic model for Treated, goodness-of-fit test

       number of observations =      4302
 number of covariate patterns =      4302
           Pearson chi2(4290) =     25293.85
                  Prob > chi2 =         0.0000
```

The correlation between Treated and the propensity score estimated from the selection model is .70. (The square of this correlation ($.70^2 = .49$) is approximately the value of the pseudo R^2 (.44).) The preceding model does a relatively good job of predicting the households selected by Fintrac for provision of services.

Aside: Comparison of the Distribution of Estimated Propensity Scores for the Baseline Treatment and Control Samples

A useful way to compare the treatment and control samples is to compare the distributions of their estimated propensity scores (i.e., of the estimated probability of participation in the program). This comparison is much simpler than comparing the samples with respect to a large number of variables (as was done earlier). This comparison shows how similar or different the treatment and comparison samples are with respect to estimated probability of selection (for participation). For an experimental design based on randomized assignment to treatment, these distributions will be two “spikes” located symmetrically about the value .5. (If the proportions of treatment and control units in the sample are the same, then there is a single spike at .5. Otherwise, there are two spikes, representing the different proportions of the treatment and control units in the sample.) If the distributions of all variables on which the propensity score is based are the same for the treatment and control samples, the distributions will be rotations (of each other) about .5. (For example, in a design having the treatment and control samples of the same size, roughly speaking, for every treatment unit having a value p for the propensity score, there will be a control unit having (approximately) a value 1-p.) The following two graphs show the distribution of estimated propensity scores for the treated and non-treated household samples[37].

[37] It should be recognized that even if the treatment and control samples had the same distributions for estimated propensity score, they would not necessarily have the same distribution for all variables that affect outcome. Similarity of the distributions of the estimated propensity score is *necessary* to avoid selection bias, but it is by no means *sufficient*, since it is based simply on observables. Randomized assignment is the only sure method of assuring comparability of the treatment and control groups

Figure A.2a. Distribution of Estimated Propensity Score (P) for Households Treated by Fintrac, Full Data Set (Original Experimental Design and Additional Sample of "600" Fintrac Clients)

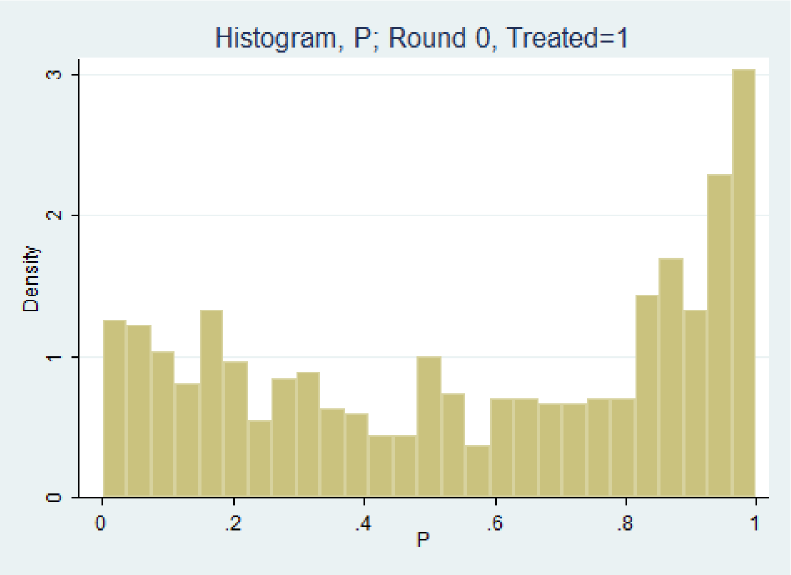
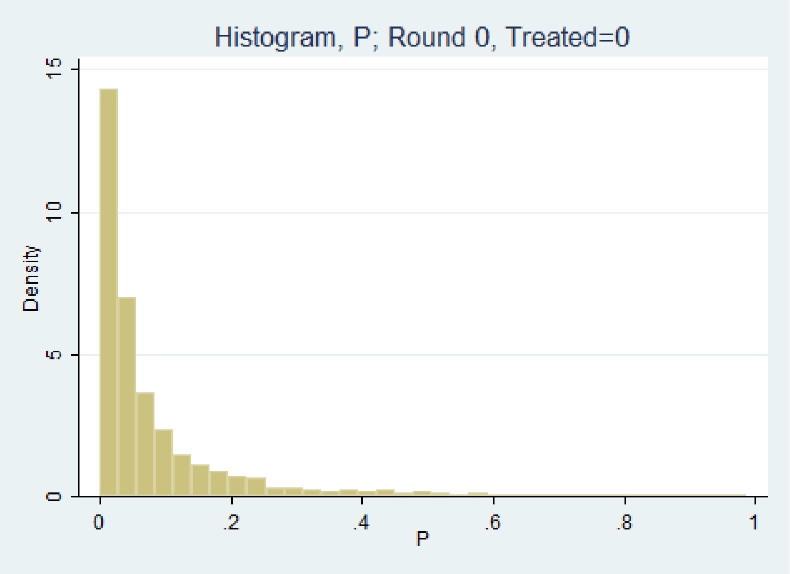


Figure A2b. Distribution of Estimated Propensity Score (P) for Households Not Treated by Fintrac, Full Data Set (Original Experimental Design and Additional Sample of "600" Fintrac Clients.



The preceding figures show that the distributions of estimated propensity score for the treatment and control samples are not rotations (of each other) about .5. This implies that the distributions of variables related to selection are different for the treatment and control samples. The basic (stratified) propensity-score estimator compares means for units having similar propensity scores, thereby accounting for this difference (in a way that reduces selection bias). The regression estimators take the propensity score into account as a covariate. The various estimators differ in magnitude since they adjust for the differences in the treatment and control groups in different ways. The two-step estimation model (first-step selection model (using a logistic regression model to estimate the propensity score) and second-step outcome model (using a regression model based on the estimated propensity score) turned out to be the best representation of the process under study (highest face validity, precision and power).

At first glance, it may appear that the supports (domains) of the two propensity score distributions are not similar. This is not at all the case. Because of the skewness of the distribution of controls, it appears to be flat over much of the unit interval, but it is not zero. Furthermore (as is also not clear from the histograms), the propensity score model developed above in fact does not have values of 1 or 0 for any units. Here follow the counts of units for various values and intervals of the propensity score.

```
p=0, treated=0, round=0: 0
p=1, treated=0, round=0: 0
p=0, treated=1, round=0: 0
p=1, treated=1, round=0: 0
p>0, p<1, treated=0, round=0: 3794
p>0, p<1, treated=1, round=0: 732

p=<.01, treated=0, round=0: 704
p=>.99, treated=0, round=0: 0
p=<.01, treated=1, round=0: 11
p=>.99, treated=1, round=0: 26
p>.01, p<.99, treated=0, round=0: 3090
p>.01, p<.99, treated=1, round=0: 695

p=<.05, treated=0, round=0: 2161
p=>.95, treated=0, round=0: 3
p=<.05, treated=1, round=0: 46
p=>.95, treated=1, round=0: 106
p>.05, p<.95, treated=0, round=0: 1630
p>.05, p<.95, treated=1, round=0: 580
```

The preceding statistics show that the supports for the propensity-score distributions for the treated and untreated households are completely overlapping. The common support is the entire open interval (0,1). The distributions are highly skewed because the logistic regression model did a very good job of predicting participation.

Statistical analysis was conducted to assess the degree to which the distributional characteristics of explanatory variables (listed in Table A.7) were similar for observations having similar propensity scores. To this end, the variable means were estimated, by treatment status, for the five quintile categories of the estimated propensity score (such a comparison is called a "balancing test"). While the means were substantially more similar within quintile categories, substantial differences remained.

Construction of the Basic Propensity-Score-Based Estimate of Impact

Once the estimated propensity score is available, it may be used to construct estimates of impact. These estimates include the average treatment effect (ATE) and the average treatment effect on the treated (ATT). In the original article by Rosenbaum and Rubin, a simple nonparametric estimate of ATE was proposed, in which the sample is stratified by the estimated propensity score, the difference in means (between treatments and controls) is calculated for each stratum, and a stratified estimate of the impact obtained from the differences in stratum means. In this evaluation, we will use more complex propensity-score-based estimators. The first one (nonparametrically identified) is similar to a Horvitz-Thompson estimator, and the next two are regression models.

If we define $\hat{p}(\boldsymbol{x})$ as the value of p(.) estimated for the value **x**, then the ATE is estimated by the following formula:

$$\widehat{ATE} = N^{-1} \sum_{i=1}^{N} (w_i - \hat{p}(\boldsymbol{x}_i)) y_i / (\hat{p}\,(\boldsymbol{x}_i)(1 - \hat{p}(\boldsymbol{x}_i)).$$

This formula is intuitively reasonable, since it is analogous to the usual formula for the estimate of the slope coefficient, β, in a regression model involving a single explanatory variable,

$$\hat{\beta} = cov(w, y)/var(w)$$

where the variance of the binary variate (w) in the denominator is given by p(1 – p), where p is the mean of the variate.

The preceding estimators are similar to Horvitz-Thompson estimators, and their precision is low if the values of $\hat{p}(\boldsymbol{x})$ are close to zero or one. Observations for which $\hat{p}(\boldsymbol{x})$ is zero or one are not included in the estimate, since for such values the term included in the sum would not be defined. To improve the precision of the preceding estimator, the estimated propensity score was censored at .1 and .9 (i.e., all values below .1 were set equal to .1 and all values above .9 were set equal to .9). (This censoring of the propensity score was done only for the basic propensity-score-based estimate discussed in this subsection; it was not done for the regression-adjusted propensity-score-based estimates discussed in the following two subsections.)

The assumptions under which the preceding propensity-score estimate provides a consistent estimate of ATE is that (1) conditional on **x**, w and (y_0, y_1) are independent; and (2) 0 < p(**x**) < 1 for all **x**. This condition is called “strong ignorability of treatment” (conditional on **x**). This assumption addresses the problem of p values of 0 and 1. Note that the requirement that p not be equal to zero or one applies to the basic propensity score estimate, not to the regression estimates.

In the main text, a list of variables that affect selection for treatment was presented. The important consideration is that variables affecting both selection and outcomes of interest are observable or, if not, that they be time-invariant.

Note that the participation model is based on household characteristics, not *aldea* characteristics, and so the model relates to participation at the farmer level, not at the *aldea* level. Data are not available to develop a participation model at the *aldea* level. It may be that Fintrac has taken into account variables not reflected in the farmer. On the other hand, Fintrac represented that *aldea* eligibility "flows up" from the farmer to the *aldea*, so this is considered unlikely. It is considered that an *aldea*-level selection model would provide little additional information, conditional on farmer-level selection.

A "naïve" estimate of the standard error of the preceding estimate may be obtained by calculating the standard deviation of the terms comprising the estimate and dividing by the square root of the number of terms. This estimator is "conservative," i.e., it converges to a value somewhat higher than the correct value, as the sample size increases. A correct estimate of the standard error is described on pp. 920 – 927 of *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, by Jeffrey M. Wooldridge (Wiley, 2010, 2002). (It is counterintuitive that the simple procedure is conservative. In general (e.g., in ordinary-least-squares regression models), when errors in a variable are ignored the standard error of the estimate is underestimated, not overestimated, i.e., the approach is not conservative. This fact is discussed on pp. 500 – 502 of this reference, in Section 13.10.2. "Surprising Efficiency Result When the First-Step Estimator Is Conditional Maximum Likelihood Estimator." The preceding estimator is a two-step M estimator in which the first step (estimation of the propensity score) is done by the method of maximum likelihood and the second step (estimation of impact, given the estimated propensity score) is done by the method of least squares. The result presented on pp. 500-502 applies to this case, under the conditional independence assumption (that conditional on **x**, the response (i.e., the counterfactual responses) (y_0, y_1) is independent of treatment, w). This assumption is called "ignorability (or unconfoundedness) of treatment" given **x**. If it is also assumed that 0 < P(w=1|**x**) < 1, the combined assumption is called "strong ignorability of treatment" given **x**. For the basic propensity-score-based estimate presented above, it is necessary to assume strong ignorability, since the estimator is undefined for values of $P=\hat{p}(\boldsymbol{x})$ equal to 0 or 1. Wooldridge observes on page 923 of op. cit. that "The naïve standard error that we obtain is [formula for the naïve estimate of the standard error] and this is at least as large as the expression (21.45) [the correct estimate of the standard error], and sometimes much larger.")

The standard error was estimated in two ways: by defining a Stata "ado" file that calculates the naïve estimate, and by applying the bootstrap procedure to an ado file that calculates the estimate. (It is noted here that it is somewhat presumptuous to use the term "correct" in referring to the recommended (bootstrap) procedure for estimating the standard error, although this (or similar adjectives, such as "proper" or "valid" or "correct") is standard usage. The estimator (i.e., the bootstrapping algorithm used to implement it) is *consistent,* which means that for large sample sizes it converges to the true ("correct") value, if the various assumptions (e.g., conditional independence; first-order approximations; variable selection probabilities; fixed or random effects) hold. Furthermore, the bootstrap estimate is conditional on the particular sample from which the bootstrap samples are selected. For finite sample sizes, it is not "correct" in an absolute sense, and there is never an assurance that the assumptions hold. A more "correct" term for this estimator of the standard error would be "improved" or "reduced-bias.")

The estimated propensity score is used in most of the estimators that follow. The estimation of the standard errors of those estimates was undertaken using the procedures described in the preceding Wooldridge reference (implemented in Stata using the *bootstrap:* procedure and suitable ado files).

Below, we present estimates of ATE and its estimated standard error (two estimates, calculated as described) for all of the outcome measures listed earlier, using the basic propensity-score-based method of estimating impact. The units for income and expense are Honduran lempiras.

Table A.11. Estimates of Average Treatment Effect (ATE), Using the Basic Propensity-Score-Based Estimate of Impact

Outcome Variable	Estimate	Standard Error (naïve estimate)	Standard Error (bootsrap estimate)
Basic Grains (BG)			
IncBG	-758	606	655
ExpBG	636*	278	291
NetBG	-1394*	530	539
LabExpBG	522*	180	163
Other Crops (OC)			
IncOC	7745*	3373	3728
ExpOC	4401*	911	1208
NetOC	3344	3018	3064
LabExpOC	2791*	601	792
Labor Market Employment and Household Income and Expenditures			
IncEmp	-157	682	716
TotHHExp	-193	363	439
NetHHInc	3376	8472	7538
Production of Horticultural Crops			
Horticulture	-.0400	.0221	.0207

Income and expense measured in Honduran lempiras.

Recall that incomes and expenses for basic grains (BG) and other crops (OC) are annual amounts; IncEmp and TotHHExp are monthly; and NetHHInc is annualized.

These indicators provide evidence that the FTDA program has had a positive effect on income for other crops (i.e., the income increased more, or decreased less, for program farmers than it did for non-program farmers). The effect on total income for other crops (IncOC) was 7,745, the effect on total expense for other crops (ExpOC) was 4,401, and the effect on estimated net income for other crops (NetOC) was 3,344. The first two of these are statistically significant, but the third is not. The effect on income from basic grains is nil, and the effect on net income for basic grains is negative. The effect on labor expenditures (LabExpBG and LabExpOC) is positive.[38] The estimated effect on NetHHInc is positive, but not statistically significant. (As mentioned, relative to the indication of statistical significance, the interpretation is that over many independent investigations, the probability that the confidence interval includes the true value of the parameter is approximately .95. Confidence intervals within the same investigation are correlated.)

Note that the estimate of NetHHInc has a large standard error. This is true not only for this (basic propensity-score-based) estimator but for all the estimators that follow. The reason for this is that the expression for NetHHInc contains the term IncTotal multiplied by 12 (to convert it from a

[38] Most tests of hypothesis considered in this report are “one-sided” tests, of whether the program increased income for other crops (OC), not “two-sided” tests of whether the program increased or decreased income. With respect to income for basic grains (BG), two-sided tests are used.)

monthly amount to an annual amount). This term causes the standard error of NetHHInc to be large. It is noted that the effect is of the expected sign (positive).

2. Regression-Adjusted Propensity-Score-Based Estimates of ATE

The basic regression model on which impact estimates are based is the following:

$$y_t = \mathbf{x}'_t\boldsymbol{\beta} + \theta d_t + \phi w_t + \delta d_t w_t + e_t,$$

where

t = survey round index (0 for Round 0 and 1 for Round 1)
y_t = explained variable (outcome variable, response variable, dependent variable)
$\mathbf{x}_t$ = vector of explanatory variables (the first component is one)
$\boldsymbol{\beta}$ = vector of parameters (the first parameter is a constant term)
d_t = indicator variable for survey round, = 0 for Round 0 and 1 for Round 1
θ = round effect
w_t = treatment variable
ϕ = treatment effect (not the impact, but the average difference in means between the treatment and control groups at baseline)
δ = impact (interaction effect of treatment and time)
e_t = model error term.

(The usual convention of representing row vectors in boldface and denoting a column vector or matrix transpose with a prime is adhered to.) The model error term is assumed to have mean zero, constant variance, and be uncorrelated with the explanatory variables. In this application, the treatment variable, w_t, is a binary variable having value one for sample units (households, farmers) who receive program services and zero otherwise. This is the same variable as has been called "Treated." In this model formulation, the value of the treatment indicator variable, w_0, varies over the Round 0 sample units depending on whether the household receives program services, and the Round 1 value (w_1) is identical to the Round 0 value for a particular household. In this model formulation, the estimate of impact (the average treatment effect, ATE) is the coefficient, δ, of the interaction of treatment and round. (In some model formulations, such as a randomized experimental design, it is customary for w_t to have the same value of w_t for all Round 0 units, and for the value in Round 1 to reflect receipt of treatment. The impact is then simply the coefficient of w_t, not the coefficient of the interaction of w_t with round. The disadvantage of that specification for this application is that w_t cannot be used to represent selection effects in Round 0 (since it is identical for all Round 0 units).)

The preceding linear statistical model may be used directly to estimate impact, or in a two-step model that includes a "selection" model (which represents the probability of participation, or propensity score, as a logistic function of a linear form such as shown above) and an "outcome" model that includes the estimated propensity score as a covariate. For this evaluation, the two-step model turned out to be best.

The preceding model formulation is appropriate if there is no interaction between the treatment variable and the explanatory variables. If this assumption is not valid, then it is necessary to include interaction terms between treatment and the covariates. If this is done, the covariate factor of the interaction term must be the deviation from the mean. *This (use of deviations from the mean) is very important.* If the covariate factor is not demeaned, then the coefficient of the interaction of treatment and round will not be an unbiased estimate of impact. If $\boldsymbol{\varphi}$ denotes the mean of the covariate, $\mathbf{x}$ (i.e., $E(\mathbf{x}) = \boldsymbol{\varphi}$), then the additional term is $d_t w_t(\mathbf{x}_t - \boldsymbol{\varphi})$.

In the present application, models were examined with and without the interaction terms between treatment and demeaned covariates. The interaction terms were determined to be necessary, and the results presented below are for models including the interaction terms.

Some additional comments about the preceding models are the following. The (vector) parameter $\boldsymbol{\beta}$ contains not only substantively (economically) meaningful explanatory variables, such as farm size, educational level of the head of household, or estimated propensity score, but also design parameters, such as *aldea* and household. The sample consists of about 3,000 households, and there are hence about 3,000 household parameters (coefficients). The particular values of these parameters are of no interest, but they are essential to include in the model in order to obtain a correct estimate of the standard error of the parameter of interest, viz., δ. There is one household indicator variable for each household. These parameters are "nuisance" parameters. They are explicitly represented in the undifferenced model described above, but not in a first-difference model (any variable that has the same value in both rounds falls out of the differenced model). Once household has been included as a variable in a fixed-effects model, the effect of *aldea* (the design variable associated with two-stage sampling) becomes negligible.

For the propensity-score-based estimates, there is a single covariate, viz., the propensity score. It is considered to combine the influence of all other covariates. Since none of the covariates were subject to experimental control (forced variation), there is little point to representing them explicitly in the model, once the propensity score is included. Assertions about the causal effect of the program as a whole are justified, since the program represents a forced intervention. Since forced variation was not implemented at the explanatory variable level (as in a designed experiment), similar assertions are not justified for individual explanatory variables.

As was discussed earlier, it is important, when constructing estimates and making tests of hypotheses about model parameters, to specify the stochastic nature of the explanatory variables (fixed or random).

We conducted a survey to assess measurement errors ("errors in variables"). However, based on that survey, it was not possible to estimate the reliabilities of the variables with sufficient precision to be used to reduce possible attenuation bias that may be caused by errors in variables.

Under the usual conditional independence assumption (i.e., that conditional on $\mathbf{x}$, w and (y_0, y_1) are independent), it may be shown, for non-panel data, that the regression of y on 1, Treated and $\hat{p}(\mathbf{x})$, the coefficient on Treated is a consistent estimate of ATE. For panel data, the regression is on 1, Round, Treated, RoundTreated, $\hat{p}(\mathbf{x})$ and Round$\hat{p}(\mathbf{x})$, and the coefficient on RoundTreated is the estimate of ATE. This estimator is called a "regression-adjusted" propensity-score-based estimator.

Instead of the general-linear-model formula presented above, the regression-adjusted propensity-score-based model may be represented as:

$$y_t = \beta_0 + \beta_1 \text{ Round} + \beta_2 \text{ Treated} + \beta_3 \text{ RoundTreated} + \beta_4\, \hat{p}(\mathbf{x}) + \beta_5 \text{ Round } \hat{p}(\mathbf{x}) + e_t.$$

The estimate of impact is the coefficient of the RoundTreated interaction term, β_3 (i.e., the interaction effect of treatment and time). All explanatory variables except for the estimated propensity score are fixed effects, and hence uncorrelated with the model error term. Under the assumption that unobserved variables affecting outcome are time-invariant (over the two survey rounds), the error term of this outcome model is uncorrelated with the error term of the selection (propensity score) model.

This model allows for the response level (y) to differ according to the level of the estimated propensity score (probability of selection for participation). It was mentioned earlier that it is considered important to allow for "flexible" model specifications, relative to the propensity score. The preceding model is the simplest one – the propensity score is included as a linear regressor. The next impact estimator to be considered will be a more complex representation.

Economic Interpretation of the Outcome Model

There is a separate response model for each outcome variable, each with its own set of β's. In each model, the estimate of impact is the RoundTreated effect (coefficient β_3). The full regression outputs for those models is not presented here, but are included in the Stata .log file that accompanies the project documentation. Here follows tables showing key model parameters (coefficients of treatment-related parameters), for both random-effects and fixed-effects models. The tables present the values of β_3, β_4 and β_5 for each outcome variable, along with their standard errors. The random-effects table is used to show the relationship of outcome to explanatory variables, and the fixed-effects table is used to estimate impact. Note that the value of the propensity score is identical for the same household between survey rounds, and for this reason the parameter β_4 drops out of the model. It is retained in the model formula, to facilitate comparison between the fixed-effects and random-effects models.

In general, when assessing the economic meaning of a model, the random-effects model is preferred to the fixed-effects model. The reason for this is that the random-effects model is a structural representation with high face validity, whereas the fixed-effects model is in effect an "estimating equation," in which model parameters are dropped if they are have the same values in both survey rounds. Since the value of the propensity score is the same in both rounds, the "P" term (corresponding to β_4) is dropped from all of the fixed-effects models. If it is of interest to see the relationship of the response to the propensity score, the random-effects model is used. In this application, the fixed-effects and random-effects models were generally similar (except for the fact that the propensity score drops out of the fixed-effects model). Because of the large sample size, the difference between the two models was usually statistically significant, but the difference is not large. Similarity of the fixed-effects and random-effects estimates is evidence that time-invariant unobserved variables are not correlated with the explanatory variables.

Table A.12A presents results for the random-effects model. The coefficients β_4 and β_5 represent adjustments to the response (not to the impact), and may be of either sign. The interesting thing to observe here (in the case of NetHHInc) is that there is a very strong positive relationship of response to estimated propensity score (coefficient 202,471), and a modest negative relationship to the interaction of the estimated propensity score and RoundP (-17,957). This means that, in general, famers who had a high propensity for program participation tended to have high incomes in Round 0 and not quite so high incomes in Round 1. This situation is associated with a weak impact.

The fixed-effects table (Table A.12B) shows that, in general, the direction of the impact (coefficient β_3) is as expected. For example, in the case of NetHHInc, the value of β_3 is positive. The coefficient β_5 represents adjustments to the *response*, not to the *impact*, and may be of either sign.

Table A.12A. Key Model Parameters for Regression-Adjusted Propensity-Score-Based Random-Effects Estimator

Outcome Variable	β_3 (RoundTreated)		β_4 (P)		β_5 (RoundP)	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
IncBG	-121	920	12635	1129	-4292	1364
ExpBG	921	348	2075	442	722	516
NetBG	-1036	799	10559	888	-4980	483
LabExpBG	435	227	-1121	248	1866	336
IncOC	11394	4385	80793	5223	-12587	6501
ExpOC	5082	994	10553	1178	4084	1474
NetOC	6116	3958	70240	4506	-16078	5862
LabExpOC	2569	647	-1034	747	7666	958
IncEmp	-349	708	9953	817	747	1048
TotHHExp	17.2	443	5740	460	-2396	654
NetHHInc	4509	10688	202471	12712	-17957	15782
Horticulture	-.0485	.0231	-.0299	.0221	.0183	.0344

Table A.12B. Key Model Parameters for Regression-Adjusted Propensity-Score-Based Fixed-Effects Estimator

Outcome Variable	β_3 (RoundTreated)		β_4 (P) (drops out of fixed-effects model)		β_5 (RoundP)	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
IncBG	-147	939			-4605*	1404
ExpBG	923*	359			597	537
NetBG	-1070	815			-5202*	1219
LabExpBG	435	238			1870*	355
IncOC	13205*	4478			-19039*	6695
ExpOC	4972*	1031			3293*	1542
NetOC	8233*	4037			-22332*	6037
LabExpOC	2098*	678			8073*	1014
IncEmp	-44	729			812	1082
TotHHExp	344	446			-1933*	666
NetHHInc	11796	10933			-24487	16255
Horticulture	-.0348	.0246			-.0180	.0370

The big advantage of this estimator (and the one to be considered in the following subsection) over the basic propensity-score-based estimator just discussed is that it is not unduly affected by values of $\hat{p}(\mathbf{x})$ close to 0 and 1. All observations, even those for which the values of $\hat{p}(\mathbf{x})$ are zero or one, may be included in the analysis. That is, it does not require the assumption of *strong* ignorability of treatment, just ignorability.

The regression analysis of the outcome variable ExpOC (for example) is shown in Figure A.3. Figure A.3 presents the estimate and estimated standard error of the estimate, using the ordinary least squares (OLS) estimation procedure and ignoring the fact that the propensity score (regressor P in the model) is an estimate. The impact estimate is the coefficient ("Coef." in the printout) of RoundTreated, and the estimated standard error of this estimate is the standard error of this coefficient ("Std. Err." in the printout).

Obtaining an improved estimate of the standard error, taking into account the fact that the propensity score is an estimate, is problematic. If all that is desired is an estimate of the standard error, it suffices to draw on the order of 50 - 200 samples in the bootstrap procedure. If it is desired to use the bootstrap to estimate both the impact and the standard error of this estimate, then much larger samples are required, e.g., on the order of 1,000. The problem is that this procedure must be applied to a substantial number of impact estimates (IncBG, ExpBG, NetBG, IncOC, etc.). Even with a powerful recent-model microcomputer, the computer running times become prohibitive for large bootstrap sample sizes (since the complete model must be re-estimated for every bootstrap sample). The approach we adopted here is to present the estimate and its standard error using the standard OLS estimation procedure (i.e., ignoring the fact that the propensity score is an estimate), and also the bootstrap estimate of the standard error (but not the bootstrap estimate of the impact estimate) using a bootstrap sample of 50. This procedure corresponds to computer runs on the order of one-half hour for a full set of estimates. The results that follow show that there is not much difference between the estimated standard error produced by the OLS regression procedure and that produced by the bootstrap procedure.

Note that the value of R^2 (.0425 for the fixed-effects model) in the regression output is of little interest. The fact that it is low is not important. The explanatory power of the model comes from the "first-step" selection model of the propensity score (i.e., the logistic model described earlier), not from this "second-step" model. What is of interest in the second-step model is the statistical significance of the impact estimate (coefficient of RoundTreated). In this example (for ExpOC) the estimate is 4,972 and its estimated standard error is 1,031, a highly statistically significant result.

Figure A.3. Regression-Adjusted Propensity-Score-Based Estimate of ATE, for ExpOC

```
. xtreg ExpOC Round Treated RoundTreated P RoundP, re

Random-effects GLS regression                   Number of obs      =      7259
Group variable: idhh                            Number of groups   =      4526

R-sq:  within  = 0.0419                         Obs per group: min =         1
       between = 0.0826                                        avg =       1.6
       overall = 0.0909                                        max =         2

Random effects u_i ~ Gaussian                   Wald chi2(5)       =    594.89
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
```

```
------------------------------------------------------------------------------
       ExpOC |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Round |   -389.388   359.0729    -1.08   0.278    -1093.158    314.3819
     Treated |   858.8321   814.8137     1.05   0.292    -738.1735    2455.838
RoundTreated |   5081.853   994.3819     5.11   0.000       3132.9    7030.806
           P |   10553.48   1177.792     8.96   0.000     8245.053    12861.91
      RoundP |   4084.198   1474.061     2.77   0.006     1195.091    6973.305
       _cons |   2608.404   260.8199    10.00   0.000     2097.207    3119.602
-------------+----------------------------------------------------------------
     sigma_u |  8962.4599
     sigma_e |  11564.174
         rho |  .37525584   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store random_effects

. xtreg ExpOC Round Treated RoundTreated P RoundP, fe

Fixed-effects (within) regression               Number of obs      =      7259
Group variable: idhh                            Number of groups   =      4526

R-sq:  within  = 0.0425                         Obs per group: min =         1
       between = 0.0653                                        avg =       1.6
       overall = 0.0598                                        max =         2

                                                F(3,2730)          =     40.37
corr(u_i, Xb)  = 0.1338                         Prob > F           =    0.0000

------------------------------------------------------------------------------
       ExpOC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Round |   95.66886   386.2497     0.25   0.804    -661.7023    853.0401
     Treated |  (dropped)
RoundTreated |   4972.161    1031.14     4.82   0.000     2950.268    6994.054
           P |  (dropped)
      RoundP |    3292.88   1541.843     2.14   0.033     269.5828    6316.177
       _cons |   4517.251   179.7077    25.14   0.000     4164.875    4869.628
-------------+----------------------------------------------------------------
     sigma_u |  13512.174
     sigma_e |  11564.174
         rho |  .57721679   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4525, 2730) =     1.95         Prob > F = 0.0000

. hausman . random_effects

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       .        random_eff~s    Difference          S.E.
-------------+----------------------------------------------------------------
       Round |    95.66886     -389.388        485.0569        142.3217
RoundTreated |    4972.161     5081.853       -109.6922        272.8622
      RoundP |     3292.88     4084.198       -791.3187        452.1314
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg
```

```
Test:  Ho:  difference in coefficients not systematic

              chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                      =       12.29
            Prob>chi2 =      0.0065
```

Below, we present the estimate of ATE and its standard error for all of the outcome measures listed earlier, using the regression-adjusted propensity-score-based method of estimating impact. (The full regression output is presented, above, just for ExpOC.) The units for income and expense are Honduran lempiras. The table includes the standard error (se) as calculated by the regression program (using standard formulas) and also using the bootstrap method.

Table A.13. Estimates of Average Treatment Effect (ATE), Using the Regression-Adjusted Propensity-Score-Based Estimate of Impact

Outcome Variable	Estimate	Standard Error (naïve estimate)	Standard Error (bootsrap estimate)
Basic Grains (BG)			
IncBG	-147	939	834
ExpBG	923*	359	391
NetBG	-1070	815	735
LabExpBG	435	238	260
Other Crops (OC)			
IncOC	13205*	4498	4096
ExpOC	4972*	1031	1097
NetOC	8233	4037	3950
LabExpOC	2098*	678	745
Labor Market Employment and Household Income and Expenditures			
IncEmp	-44	724	750
TotHHExp	344	448	464
NetHHInc	11796	10934	13254
Production of Horticultural Crops			
Horticulture	-.0348	.0245	.0193

Income and expense measured in Honduran lempiras.

These results are similar to those for the basic propensity-score estimators presented earlier – the program intervention is associated with positive increase in IncOC, ExpOC, NetOC and LabExpOC. The NetHHInc effect is positive, but not statistically significant. (As mentioned, relative to the indication of statistical significance, the interpretation is that over many independent investigations, the probability that the confidence interval includes the true value of the parameter is approximately .95. Confidence intervals within the same investigation are correlated.)

Note that the standard errors calculated using the improved procedure (the “bootstrap se”) differ little from the standard errors produced by the regression model. Theoretically, as discussed earlier, the improved estimates should not be any larger than the estimates from the regression model based on the estimated propensity score. The fact that some of them are, is because of sampling variation (i.e., the bootstrap estimates of the standard errors are based on relatively small bootstrap sample of size 50). (In the bootstrap procedure, samples of the same size as the full data set were selected by replacement from the full data set, and the regression estimate was calculated for each sample. This was done 50 times, and the standard error of the estimate was calculated directly from these 50 replications.) In this study, both the naïve and bootstrap estimates of the standard error have been presented. In future studies that use similar models, it is recommended that only the naïve estimator of the standard error be used, since calculation of the bootstrap estimator consumes a substantial amount of computer running time (if a large number of estimates is involved, as is the case in the present study).

3. Modified Regression-Adjusted Propensity-Score-Based Estimates of ATE

A modified version of the preceding estimator is obtained by regressing y on 1, Round, Treated, RoundTreated, $\hat{p}(\boldsymbol{x})$, Round($\hat{p}(\boldsymbol{x})$ - $\hat{\mu}_p$) and RoundTreated($\hat{p}(\boldsymbol{x})$ - $\hat{\mu}_p$), where $\hat{\mu}_p$ denotes the mean of the estimated propensity scores. The descriptor "modified" refers to the fact that this is the same model as the regression-adjusted propensity-score-based model, with the addition of a term representing the interaction of the demeaned propensity score with RoundTreated. The additional assumption required for use of this estimator is that $E(y_0|p(\mathbf{x}))$ and $E(y_1|p(\mathbf{x}))$ are linear in $p(\mathbf{x})$. The estimate of impact is the Round*Treatment effect (i.e., the coefficient of the Round*Treatment interaction term (i.e., the interaction effect of treatment and time)). . This estimate is an estimate of the average treatment effect (ATE), or expected impact of the program intervention on a randomly selected program-eligible farmer.

This model allows for the impact effect (interaction of treatment and time) to be directly related to the (demeaned) estimated propensity score. It may be used to estimate impact as a function of the covariates. This feature of the model will be used in the next section, which is concerned with estimation of the average treatment effect on the treated (ATT).

The modified regression-adjusted propensity-score-based model may be represented as:

$$y_t = \beta_0 + \beta_1 \text{ Round} + \beta_2 \text{ Treated} + \beta_3 \text{ RoundTreated} + \beta_4 \hat{p}(\mathbf{x}) + \beta_5 \text{ Round } \hat{p}(\mathbf{x}) + \beta_6 \text{ RoundTreated } (\hat{p}(\boldsymbol{x}) - \hat{\mu}_p) + e_t.$$

This is the same model as used for the preceding estimator, with the addition of the three-component interaction term, RoundTreated($\hat{p}(\boldsymbol{x})$ - $\hat{\mu}_p$), the interaction of RoundTreated .with the demeaned estimated propensity score. (The mean and standard deviation of the estimated propensity score are .169 and .00379.) All explanatory variables except for those involving the estimated propensity score are fixed effects, and hence uncorrelated with the model error term. Under the assumption that unobserved variables affecting outcome are time-invariant (over the two survey rounds), the error term of this outcome model is uncorrelated with the error term of the selection (propensity score) model.

This estimator is a covariate-adjusted regression estimator, where the covariate is the (estimated) propensity score, and the interaction term with this demeaned covariate is included in the model. This model is a more "flexible" specification, with respect to how the propensity score is included in the model.

The theory underlying the use of propensity scores in counterfactuals analysis is that the treatment variable and the counterfactual responses to treatment are independent, given the propensity score. That is, given the propensity score (i.e., a group of units having the same propensity score), an unbiased estimate of impact is the simple difference in means of the treated and untreated units. Under the assumption of conditional independence (of treatment and response), there is no need to include additional covariates in the outcome model, once the propensity score is included (in flexible specifications). This is the reason why no additional covariates are included in the preceding model, beyond the estimated propensity score.

Economic Interpretation of the Outcome Model

[The analysis here is very similar to that presented for the preceding model. Some paragraphs are repeated, so that each section may be read independently.]

There is a separate response model for each outcome variable, each with its own set of β's. In each model, the estimate of impact is the RoundTreated effect (coefficient β_3). The full regression outputs for those models is not presented here, but are included in the Stata .log file that accompanies the project documentation. Here follows tables showing key model parameters (coefficients of treatment-related parameters), for both random-effects and fixed-effects models. The tables present the values of β_3, β_4, β_5 and β_6 for each outcome variable, along with their standard errors. The random-effects table is used to show the relationship of outcome to explanatory variables, and the fixed-effects table is used to estimate impact. Note that the value of the propensity score is identical for the same household between survey rounds, and for this reason the parameter β_4 drops out of the fixed-effects model. It is retained in the model formula, to facilitate comparison between the fixed-effects and random-effects models.

In general, when assessing the economic meaning of a model, the random-effects model is preferred to the fixed-effects model. The reason for this is that the random-effects model is a structural representation with high face validity, whereas the fixed-effects model is in effect an "estimating equation," in which model parameters are dropped if they are have the same values in both survey rounds. Since the value of the propensity score is the same in both rounds, the "P" term (corresponding to β_4) is dropped from all of the fixed-effects models. If it is of interest to see the relationship of the response to the propensity score, the random-effects model is used. In this application, the fixed-effects and random-effects models were generally similar (except for the fact that the propensity score drops out of the fixed-effects model). Because of the large sample size, the difference between the two models was usually statistically significant, but the difference is not large. Similarity of the fixed-effects and random-effects estimates is evidence that time-invariant unobserved variables are not correlated with the explanatory variables.

Table A.14A presents results for the random-effects model. The coefficients β_4 and β_5 represent adjustments to the response (not to the impact), and may be of either sign. The interesting thing to observe here (in the case of NetHHInc) is that there is a very strong positive relationship of response to estimated propensity score (coefficient 202,469), and a modest negative relationship to the interaction of the estimated propensity score and RoundP (-12,088). This means that, in general, famers who had a high propensity for program participation tended to have high incomes in Round 0 and not quite so high incomes in Round 1. This situation is associated with a weak impact.

The fixed-effects table (Table A.14B) shows that, in general, the direction of the impact (coefficient β_3) is as expected. For example, in the case of NetHHInc, the value of β_3 is positive. The coefficient β_5 represent adjustments to the *response*, not to the *impact*, and may be of either sign.

The coefficient β_6 represents an adjustment to impact. The coefficient β_6 (interaction RoundTreatedPstd) is not statistically significant (for any outcome variable). The interpretation of this is that there is not a strong relationship between impact (*impact*, not response) and the estimated propensity score, although the relationship of the response to the estimated propensity score is strong. This fact will be revisited later, in estimation of the average treatment effect on the treated (ATT).

Table A.14A: Key Model Parameters for Modified Regression-Adjusted Propensity-Score-Based Random-Effects Estimator

Outcome Variable	β_3 (RoundTreated)		β_4 (P)		β_5 (RoundP)		β_6 (RoundTreatedPstd)	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
IncBG	-352	950	12635	1129	-5832	2091	2420	2491
ExpBG	815	360	2076	442	15.1	796	1111	953
NetBG	-1216	824	10559	888	-6182	1772	1889	2072
LabExpBG	396	234	-1121	248	1605	501	410	583
IncOC	10122	4531	80793	5223	-21002	9908	13265	11741
ExpOC	4833	1027	10553	1178	2422	2243	2615	2656
NetOC	4657	4086	70240	4504	-25729	8840	15206	10390
LabExpOC	2514	667	-1035	747	7297	1449	579	1708
IncEmp	-268	730	9953	817	1285	1585	-846	1868
TotHHExp	98	455	5740	460	-1856	959	-849	1102
NetHHInc	5414	11046	202469	12713	-12088	24030	-9250	28550
Horticulture	-.0572	.0237	-.0299	.0221	-.0443	.0514	.0931	.0566

Table A.14B: Key Model Parameters for Modified Regression-Adjusted Propensity-Score-Based Fixed-Effects Estimator

Outcome Variable	β_3 (RoundTreated)		β_4 (P) (drops out)		β_5 (RoundP)		β_6 (RoundTreatedPstd)	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
IncBG	-120	979			-4428	2328	-278	2918
ExpBG	837	375			25.3	891	899	375
NetBG	-957	851			-4453	2022	-1177	2022
LabExpBG	351	248			1312	589	877	738
IncOC	26774	4665			-4752	11087	-37399	13900
ExpOC	5413	1075			6233	2556	-4622	3704
NetOC	11360	4206			-1481	9997	-32777	12534
LabExpOC	1911	707			6828	1681	1957	2107
IncEmp	149	755			2097	1794	-2020	1794
TotHHExp	204	465			-2865	1105	1466	1385
NetHHInc	18926	11411			21956	26892	-73123	33744
Horticulture	-.0397	.0258			-.0534	.0660	.0516	.0797

The regression analysis for the modified regression-adjusted propensity-score-based estimate of

ExpOC is shown in Figure A.4. Both the random-effects and fixed-effects models are shown, but the impact is taken from the fixed-effects model.

Figure A.4. Modified Regression-Adjusted Propensity-Score-Based Estimate of ATE, for ExpOC

```
. xtreg ExpOC Round Treated RoundTreated P RoundP RoundTreatedPstd, re

Random-effects GLS regression                   Number of obs      =      7259
Group variable: idhh                            Number of groups   =      4526

R-sq:  within  = 0.0409                         Obs per group: min =         1
       between = 0.0833                                        avg =       1.6
       overall = 0.0914                                        max =         2

Random effects u_i ~ Gaussian                   Wald chi2(6)       =    596.24
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       ExpOC |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Round |   -233.364   393.0394    -0.59   0.553    -1003.707     536.979
     Treated |   858.8321   814.6609     1.05   0.292     -737.874    2455.538
RoundTreated |    4832.58   1026.786     4.71   0.000     2820.117    6845.043
           P |   10553.48   1177.571     8.96   0.000     8245.486    12861.48
      RoundP |    2422.42   2242.667     1.08   0.280    -1973.126    6817.966
RoundTrea~td |   2614.542   2655.657     0.98   0.325     -2590.45    7819.535
       _cons |   2608.404   260.7709    10.00   0.000     2097.303    3119.506
-------------+----------------------------------------------------------------
     sigma_u |  8939.1204
     sigma_e |  11561.885
         rho |  .37412669   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store random_effects

. xtreg ExpOC Round Treated RoundTreated P RoundP RoundTreatedPstd, fe

Fixed-effects (within) regression               Number of obs      =      7259
Group variable: idhh                            Number of groups   =      4526

R-sq:  within  = 0.0432                         Obs per group: min =         1
       between = 0.0603                                        avg =       1.6
       overall = 0.0568                                        max =         2

                                                F(4,2729)          =     30.81
corr(u_i, Xb)  = 0.1251                         Prob > F           =    0.0000
```

```
------------------------------------------------------------------------------
       ExpOC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Round |  -181.6896   431.3957    -0.42   0.674    -1027.585    664.2056
     Treated |  (dropped)
RoundTreated |   5413.159   1075.313     5.03   0.000      3304.65    7521.669
           P |  (dropped)
      RoundP |   6232.988   2555.557     2.44   0.015     1221.966    11244.01
RoundTrea~td |  -4621.811   3204.129    -1.44   0.149    -10904.58    1660.953
       _cons |   4517.251   179.6721    25.14   0.000     4164.944    4869.559
-------------+----------------------------------------------------------------
     sigma_u |  13522.412
     sigma_e |  11561.885
         rho |   .577683   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4525, 2729) =     1.95          Prob > F = 0.0000

. hausman . random_effects

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       .       random_eff~s    Difference          S.E.
-------------+----------------------------------------------------------------
       Round |   -181.6896     -233.364        51.67444        177.8265
RoundTreated |    5413.159      4832.58        580.5795        319.3888
      RoundP |    6232.988      2422.42        3810.568        1225.282
RoundTrea~td |   -4621.811     2614.542       -7236.354        1792.744
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       28.77
                Prob>chi2 =      0.0000
```

Below, we present the estimate of ATE and its standard error for all of the outcome measures listed earlier, using the modified regression-adjusted propensity-score-based method of estimating impact. (The full regression output was shown just for ExpOC, above.) The standard errors of the estimated impact are estimated two ways: from the regression model using the full sample, and by a bootstrap sample of 50 (i.e., by calculating the regression estimate for 50 samples (the same size as the full sample but selected with replacement from the full sample), and calculating the standard error of this estimate).

Table A.15. Estimates of Average Treatment Effect (ATE), Using the Modified Regression-Adjusted Propensity-Score-Based Estimate of Impact

Outcome Variable	**Estimate**	**Standard Error (naïve estimate)**	**Standard Error (bootsrap estimate)**
Basic Grains (BG)			
IncBG	-120	979	837
ExpBG	837*	375	393
NetBG	-957	851	750
LabExpBG	435	248	264
Other Crops (OC)			
IncOC	16773*	4665	4298
ExpOC	5413*	1075	1078
NetOC	11360*	4206	4175
LabExpOC	1911*	707	742
Labor Market Employment and Household Income and Expenditures			
IncEmp	149	755	733
TotHHExp	204	465	496
NetHHInc	18926*	11411	13306
Production of Horticultural Crops			
Horticulture	-.0397	.0258	.0194

Income and expense measured in Honduran lempiras.

The results are similar to the earlier estimates, and the conclusions drawn are the same. The impacts for this estimator are somewhat stronger than for the previous estimator. For example, all income and expense components for “other crops” are statistically significant, as is net household income (NetHHInc).

As discussed earlier, it is these estimates, for the modified regression-adjusted propensity-score-based estimate of impact, that are presented in the main text.

As before (for the regression-adjusted estimator of the preceding subsection), the bootstrap estimates of the standard errors differ little from the estimate produced by the regression model based on the full sample.

These results show that the effect of the program is positive. For example, the table shows that, over the population of eligible *aldeas*, net income change from other crops is on average 11,360 lempiras (USD 601) higher for program participants than for nonparticipants. All of the income/expense components for other (horticultural) crops have positive effects.

While a number of the impact estimates are statistically significant, they are not as large as was anticipated for the program (represented as perhaps doubling a farmer’s income, or by an estimated economic rate of return of 36% (in the *M&E Plan*)). The relationship of income to the estimated propensity score is very strong.

While the relationship of impact to treatment (program participation) is not strong, it is noted that the relationship of income to the estimated propensity score is very strong. Farmers similar to those selected for treatment tend to do well, even though they do not participate. Another way of looking at this is that Fintrac has an ability for selecting farmers who are likely to do well.

An interesting result is that the program does not appear to have a positive effect on the proportion of farmers growing horticultural crops, as measured by the question asking respondents whether they had harvested horticultural crops in the last 12 months (not including home garden), with response categories of no = 1, yes = 2. The impact estimate for this indicator is not statistically significantly different from zero. This could be because Fintrac chose only farmers who showed a proven ability to grow horticultural crops to be part of their program. This suggests that increments in income from other crops came from increased production among farmers already growing horticultural crops and not from farmers who switched over for the first time.

The data analysis provides strong statistical evidence that the FTDA program had a positive effect on income, net income, expenditures and labor expenditures for other crops (the category that includes those crops addressed by the FTDA program). The results of the impact evaluation show that the FTDA activity had a positive impact on its primary area of focus: activities related to horticultural crops. However, a broader positive impact on household income and expenditures was not detected.

The impact estimates were based on data that included all of the data obtained from the original experimental design, augmented by a sample of program farmers recruited by Fintrac in the course of its normal project operations. Statistical analysis was used to adjust for differences between the treatment and control samples, i.e., to reduce potential selection bias. The statistical analysis procedures used to estimate impact are based on sound causal models and causal-modeling theory (Neyman-Fisher-Cox-Rubin Causal Model, potential outcomes model, counterfactual model). The impact estimates constructed in this evaluation are estimates of the causal effect of the FTDA program intervention. An *ex post* statistical power analysis was conducted (to be discussed below), that showed that the study was not "underpowered." It is considered that the inferences made in this evaluation project are sound – valid and of adequate precision and power (an *ex post* statistical power analysis will be presented, below). It is the conclusion of this evaluation study that the FTDA program produces positive results relative to horticulture production, but those results are small in magnitude.

Ex Post Statistical Power Analysis

One of the issues to address with respect to the estimated impacts is whether the small number of statistically significant results is an indication of low power. That is, to address whether the sample size may not be sufficiently large to detect effects of anticipated or realized size. Statistical power analysis was done at the beginning of the project to estimate sample size (see Annex 3). That *ex ante* power analysis was complicated by the fact that the standard error of the impact estimates was not known (at that time, prior to the survey). For that reason, the power analysis was based on a model that involved a number of parameters about the test, the population under study, and the sample design. Now that the data analysis has been completed, estimates are available for the standard errors of the impact estimates, and an *ex post* (or *post hoc*) power analysis may be conducted much more easily than the *ex-ante* power analysis. It depends on just the test parameters (significance level; test direction (one-sided or two-sided)) and the standard error of the impact estimate. The significance level, α, of the test is the probability of a Type I error of making a decision that the effect (impact) is present (different from zero) when it is not. The probability of a Type II error of making a decision that the effect is not present when it is, is β. The power is 1 – β.

There are a number of indicators that may be examined in an *ex post* power analysis. Two standard indicators are the power of the test to detect a true effect equal in magnitude to the observed effect, and the minimum detectable effect (MDE) that can be detected for a specified level of power, which we shall set at 90%. The formula for the first indicator is:

$$Power = prob\left(\hat{t} \geq \hat{t}_{1-\beta}\right), where\ \beta\ is\ defined\ by\ \hat{t}_{1-\beta} = t_{critical:\alpha/2} - \frac{\hat{\Delta}}{\hat{\sigma}_{\Delta}}$$

for a two-sided test and

$$Power = prob\left(\hat{t} \geq \hat{t}_{1-\beta}\right), where\ \beta\ is\ defined\ by\ \hat{t}_{1-\beta} = t_{critical:\alpha} - \frac{\hat{\Delta}}{\hat{\sigma}_{\Delta}}$$

for a one-sided test, where $\hat{\Delta}$ denotes the impact estimator and $\hat{\sigma}_{\Delta}$ denotes the standard error of this estimate. (The power formulas and notation presented here are from David M. Murray, *Design and Analysis of Group-Randomized Trials*, Oxford University Press, 1998.)

The formula for the second indicator is:

$$\hat{\Delta} = \hat{\sigma}_{\Delta}(t_{critical:\,\alpha/2} + t_{critical:\,\beta})$$

for a two-sided test, where α = .05 and β = 1 – power = .1, and

$$\hat{\Delta} = \hat{\sigma}_{\Delta}(t_{critical:\,\alpha} + t_{critical:\,\beta})$$

for a one-sided test.

For α = .05 and β = .1, the critical t values are

$$t_{critical:\,\alpha/2} = 1.96;\ t_{critical:\,\alpha/2} = 1.645;\ and\ t_{critical:\,\beta} = .84.$$

Two other indicators of interest in an *ex post* power analysis are the power to detect an effect equal to 10 percent of the mean of an outcome variable of interest and the power to detect an effect equal to 10 percent of the standard deviation of an outcome variable of interest. These are standard cases often considered in *ex ante* power analysis, and it is of interest to estimate the power for these two cases after the data have been analyzed and values are known for the various parameters that were unknown at the beginning of the study. The power is calculated for these two indicators from the same formula given above (for the first indicator), simply by substituting the effect size (ten percent of the mean or standard deviation) in place of $\hat{\Delta}$.

It is also of interest to calculate the ratio of the standard error of the estimate to the mean and to the standard deviation. These indicators are related to the two just described. The latter one is of interest for estimating the (Kish) design effect of the study.

The following table presents the indicators just described, for a selection of the outcome variables, for primary roads. The table is constructed using one-sided tests, in which case, for α = .05 and β= .1 the value of $t_{critical:\alpha} + t_{critical:\beta}$ = 1.6449 + 1.2816 = 2.9265.

The power to detect an effect equal in magnitude to the observed effect is shown in column 4 of the table. This indicator is of interest only for the larger effects, since if a true effect is small, the power to detect it will be, too. The minimum detectable effect for a test of power 90% is shown in column 5. The most interesting indicators are shown in columns 9-14 – the power to detect effects equal in magnitude to various percentages of the variable (base year) mean and standard deviation.

The sample size for the evaluation was determined by statistical power analysis, i.e., by determining the sample size required to achieve a specified level of power for detecting an effect (impact, measured by the double difference measure) of specified size. When this evaluation project began, it was represented that the program intervention could easily double the income of a rural farmer, from that provided by raising traditional crops (basic grains). The initial power calculations were based on this assumption. As time passed, the assessment of program impact grew more conservative, and the sample size was estimated to achieve high power for detecting impacts equal to .5, .33, and .25 of the baseline income. As mentioned, the *M&E Plan* estimated an economic rate of return of 36% for the FTDA project.

The table that follows shows the power of the sample size used in the evaluation to detect impacts equal to 1.0, .5 and .25 of the base year mean, for selected outcome variables. For most outcome variables, the power to detect impacts of these magnitudes is very high. *The evaluation was not "underpowered." The power to detect impacts of the size anticipated was very high.*

A revealing indicator of the power of the design is the ratio of the standard error of the estimated impact to the variable mean. This is shown in the penultimate column of the table. For an effect to be statistically significant, it has to be about twice as large as the entry in this column, as a fraction of the mean. This means that for some of the outcome variables (the smaller components of income), impacts would have to be a substantial proportion of the mean, in order to detect them with high power. (For example, the relative standard error of the estimate of impact for NetHHInc is .124. Twice this is .248. This means that for the impact of NetHHInc to be statistically significant, the effect would have to be about 25 per cent of the mean NetHHInc. This is in line with the minimum detectable effects specified at the beginning of the project (e.g., an ERR of 36%). This magnitude change may be expected for some of the indicators, but this magnitude change would not be expected for all indicators. As another example, the relative standard error of IncOC is .225. Twice this is .45. In the planning phase of the study, it was represented that the program could produce changes of this magnitude.

The last column is useful for estimating the design effect of the study. For estimation of double differences, the standard deviation of the double difference estimator if simple random sampling is used for all four design groups (if of equal size) is 4σ/sqrt(n), where σ denotes the standard deviation and n denotes the total sample size for all four groups. The value of the design effect is deff = (standard error of estimate) / (4σ/sqrt(n)) = (sqrt(n)/4) (standard error of estimate) / (Round 0 sd). From the last column, it is seen that the average value (over the outcome variables) of the ratio of the standard error of the estimate to the Round 0 standard deviation is about .08. The value of n is 7262, so deff is approximately equal to .08 sqrt(7262)/4 = 1.70. This value of deff is in line with what was expected for the design (e.g., for an intra-unit (*aldea*) correlation coefficient of icc = .1 and a within-unit household sample size of m = 20, deff = 1 + (m-1)icc = 2.9; for icc= .03 and m=20, we obtain deff = 1 + (20-1).03 = 1.57).

The table includes a column that specifies the coefficient of variation (CV) of the outcome variables for Round 0. The coefficient of variation is the standard deviation divided by the mean. It is presented in the column headed “CV (sd/mean)” in the table. In the power calculations done at the beginning of the project, not much was known about the statistical properties of the population with respect to the variables of interest. Data were available from which the CV for income could be estimated, and it was seen to be in the range 1-2. The sample data show that the CV is often much larger than this.

Table A.16. *Ex Post* Statistical Power Analysis for Selected Impact Estimates																
Outcome Variable	**Estimate of Impact (observed effect)**	**Standard error (se) of estimate**	**Power of test to detect a true effect equal in magnitude to the estimate of impact**	**Minimum detectable effect (MDE) for a test of power 90%**	**Round 0 mean**	**Round 0 standard deviation (sd)**	**CV (sd/mean)**	**Power to detect a true effect equal to 1.0 of Round 0 mean**	**Power to detect a true effect equal to 1.0 of Round 0 sd**	**Power to detect a true effect equal to .5 of Round 0 mean**	**Power to detect a true effect equal to .5 of Round 0 sd**	**Power to detect a true effect equal .25 of Round 0 mean**	**Power to detect a true effect equal to .25 of Round 0 sd**	**Std error of estimate relative to Round 0 mean**	**Std error of estimate relative to Round 0 std dev**
IncBG	-120	837	.070	2450	7682	14017	1.82	1.0	1.0	.997	1.0	.742	.993	.109	.0600
ExpBG	837	393	.686	1150	2991	5350	1.79	1.0	1.0	.983	1.0	.602	.959	.131	.0735
NetBG	-957	750	.356	2195	4691	10966	2.34	1.0	1.0	.928	1.0	.468	.977	.160	.0684
LabExpBG	351	264	.376	773	931	2924	3.14	.969	1.0	.547	1.0	.223	.868	.284	.0903
IncOC	16773	4298	.987	12578	19103	65316	3.42	.996	1.0	.718	1.0	.297	.983	.225	.0658
ExpOC	5413	1078	.999	3155	4532	13650	3.01	.993	1.0	.676	1.0	.277	.933	.238	,0788
NetOC	11360	4175	.858	12218	14571	56737	3.89	.996	1.0	.540	1.0	.221	.959	.287	.0736
LabExpOC	1911	742	.823	2171	1878	8360	4.45	.811	1.0	.350	1.0	.156	.878	.395	.0888
IncEmp	149	733	.078	2145	6450	9851	1.53	1.0	1.0	.995	1.0	.710	.956	.114	.0744
TotHHExp	204	496	.110	1452	5375	4922	.916	1.0	1.0	1.0	1.0	.856	.798	.092	.1008
NetHHInc	18926	13306	.412	38940	107379	157043	1.46	1.0	1.0	.991	.903	.645	.903	.124	.0847
Horticulture	-.0397	.0194	.656	.0568	1.9227	.2670	.139	1.0	1.0	1.0	1.0	1.0	.962	.010	.0727

Intra-unit correlation coefficients

In the *ex-ante* statistical power analysis that was done at the beginning of this project to estimate sample size, one of the key parameters involved in the calculations was the intra-unit correlation coefficient for outcome variables of interest, at two levels of sampling (aldea and household). That parameter was not known for any specific outcome variable, and “nominal” values of .1 and .5 were assumed, for aldeas and households, respectively. Once the survey data are available, the intra-unit correlation coefficient can be calculated for various levels of aggregation. These values are not of direct interest to the analysis presented in this report, but they would be of interest to assist power analysis and sample size estimation for future studies. Here follows a table of the intra-unit correlation coefficients for the outcome variables of this study, for various levels of aggregation (household, aldea, municipality, and department). The lower levels of aggregation (household, aldea) are the ones of interest for use as sampling units in multistage sampling. (The intra-unit correlation was not calculated for all levels of sampling for all variables).

The intra-unit correlation coefficients are estimated by using the Stata procedure *loneway*. Here follows a sample output (for variable ExpOC at the household level). (Both rounds of survey data were used to calculate the intra-unit correlations for household, and the Roung 0 (baseline) data were used to calculate the intra-unit correlations for the higher levels.) The program output included the estimated intra-unit correlation and its standard error. The standard errors are not included in the table shown below, but are include in the .log file. As a general rule, intra-unit correlations are positive and increase as the size of the sample unit increases. The intra-unit correlations are estimated from an analysis of variance procedure, and are restricted to be positive (the zero entries in the table represent truncated estimates).

In the ex-ante power analysis done at the beginning of the project (to estimate aldea and household sample sizes), “nominal” values were assumed for the intra-unit correlations. The intra-unit correlation associated with households was assumed to be .7. The correlation associated with matched pairs of aldeas was assumed to be .5. The intra-unit correlation associated with aldeas was assumed to be .15. It is seen from the table that the intra-unit correlations at a particular level of sampling vary substantially over the various outcome variables, and that these assumed values were conservative for aldeas (e.g., about .03 vs. .15) and not conservative for households (about .4 vs. .7). The design effect for the experimental design was assumed to be deff = 1 + (m-1)icc = 1 + (20-1) .15 = 3.85. For the revised design, the design effect, taking into account loss of precision for multistage sampling and increase in precision from regression, was taken to be 1.0.

The size of the intra-unit correlation coefficient does not affect the *ex post* power analysis presented earlier. It is presented here (along with the values of the coefficients of variation) to assist the design of future similar evaluations.

```
. loneway ExpOC Aldea if Round==0

                    One-way Analysis of Variance for ExpOC:

                                              Number of obs =      4526
                                                  R-squared =    0.0368

    Source                SS         df      MS            F     Prob > F
-------------------------------------------------------------------------
Between Aldea          3.101e+10     45    6.891e+08      3.80     0.0000
Within Aldea           8.121e+11   4480    1.813e+08
-------------------------------------------------------------------------
Total                  8.431e+11   4525    1.863e+08

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ------------------------------------------------
            0.02876     0.01112       0.00696     0.05055

         Estimated SD of Aldea effect            2316.719
         Estimated SD within Aldea               13463.66
         Est. reliability of a Aldea mean         0.73695
              (evaluated at n=94.62)
```

Table A.17. Intra-unit correlation coefficients for sampling units of various sizes

Outcome Variable	Sampling unit (level of sampling in multistage sampling)			
	Household	***Aldea***	**Municipality**	**Department**
IncBG	.475	.033	.016	.044
ExpBG	.461	.027	.016	.051
NetBG	.348	.031	.020	.029
LabExpBG	.223	.013	.017	.027
IncOC	.486	.042	.026	.054
ExpOC	.404	.029	.028	.059
NetOC	.433	.045	.022	.046
LabExpOC	.267	.022	.022	.030
IncEmp	.423	.033	.026	.024
TotHHExp	.272	.040	.066	.040
NetHHInc	.504	.040	.025	.028
Horticulture	.051	.007	.029	.032

4. Estimate of the Average Treatment Effect on the Treated (ATT)

Using the modified regression-adjusted propensity-score-based approach, we shall also estimate the average treatment effect on the treated. The average treatment effect (ATE) is the expected impact of the program intervention on a randomly selected program-eligible farmer. The average treatment effect on the treated (ATT) is the expected impact of the program intervention on a treated farmer. (The ATT may be estimated for the other estimators considered earlier. In fact,, they were, and they are included in the Do14* Stata .log file.).

The ATT estimator is obtained from the model used for ATE, simply by substituting Treated=1 in the formula, calculating the formula for every treated unit in the sample, and averaging (over the Treated=1 sample). The response (outcome) model is:

$$y_t = \beta_0 + \beta_1 \text{ Round} + \beta_2 \text{ Treated} + \beta_3 \text{ RoundTreated} + \beta_4 \hat{p}(\mathbf{x}) + \beta_5 \text{ Round } \hat{p}(\mathbf{x}) + \beta_6 \text{ RoundTreated } (\hat{p}(\boldsymbol{x}) - \hat{\mu}_p) + e_t.$$

Substituting Treated=1 we obtain:

$$y_t = \beta_0 + \beta_1 \text{ Round} + \beta_2 + \beta_3 \text{ Round} + \beta_4 \hat{p}(\mathbf{x}) + \beta_5 \text{ Round } \hat{p}(\mathbf{x}) + \beta_6 \text{ Round}(\hat{p}(\boldsymbol{x}) - \hat{\mu}_p) + e_t$$
$$= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{ Round} + \beta_4 \hat{p}(\mathbf{x}) + \beta_5 \text{ Round } \hat{p}(\mathbf{x}) + \beta_6 \text{ Round } (\hat{p}(\boldsymbol{x}) - \hat{\mu}_p) + e_t.$$

This formula is evaluated for every treatment unit in the sample, and averaged, to obtain the estimate of the ATT. An approximate estimate of the standard error of the estimated ATT may be obtained by using the formulas for the standard error of a linear function of the parameters for a general linear model, i.e., se(**c**'**b**) = sqrt (**c**'V**c**) where c denotes the vector of coefficients of the linear function and V denotes the covariance matrix of the estimate **b** of **β**, where **β** denotes the vector of parameters (regression coefficients). (This estimate is approximate since it ignores the fact that the propensity score is an estimate. Bootstrapping could be used to obtain a sample-unbiased estimate of the standard error, but, in view of the similarity of the bootrstrap and naïve estimates of the standard error for ATE, this is not worth the effort.)

Here follow the estimated ATT for selected outcome variables:

Table A.18. Estimates of Average Treatment Effect on the Treated (ATT), Using the Modified Regression-Adjusted Propensity-Score-Based Estimate of Impact

Outcome Variable	Estimate	Standard Error
Basic Grains (BG)		
IncBG	-172	974
ExpBG	1002*	373
NetBG	-1174	846
LabExpBG	513	246
Other Crops (OC)		
IncOC	9896*	4638
ExpOC	4563*	1069
NetOC	5333	4183
LabExpOC	2271*	703
Labor-Market Employment and Household Income and Expenditures		
IncEmp	-222	750
TotHHExp	474	462
NetHHInc	5479	11309
Production of Hortucultural Crops		
Horticulture	-.0302	.0256

Income and expense measured in Honduran lempiras.

In many cases, the estimates of the ATT do not differ much from the estimates of the ATE. In fact, for those that are markedly different (NetOC and NetHHInc), the random-effects estimates are similar (4657 and 5414, respectively, not shown here for other estimates, but included in the .log file). This is not surprising, since the ATT simply adjusts the ATE by the demeaned-covariate interaction term, which in this case is the demeaned propensity-score interaction term (with RoundTreated), and this effect is not large (also not shown here, but included in the .log file).

The significant observation to make here is that the ATE and the ATT are generally similar in magnitude. This means that the effect on a randomly selected program-eligible farmer (or, more specifically, from a randomly selected program-eligible farmer in a randomly selected *aldea*) is not much different from the effect on a Fintrac-selected program-eligible individual. This is additional evidence that the program impact is not large. Had the ATE been small but the ATT large, it would have been concluded that, although the impact for a randomly selected program-eligible farmer may be small, the program had a substantial effect on Fintrac-selected clients, and that Fintrac knew how to select clients that would perform better-than-average in the program. This does not appear to be the case. The program has a statistically significant, but weak, impact, and it is about the same for Fintrac-selected clients as for randomly selected eligible farmers.

It was observed earlier that farmers similar to those selected for treatment (i.e., having similar propensity scores) tend to do well, even though they do not participate. This is not the same as a differential treatment effect (between treated and untreated farmers). Although Fintrac may have an ability to select farmers who are likely to do well (whether they participate or not), it does not appear that Fintrac has an ability to select those who are likely to perform noticeably better in the FTDA program (than other program-eligible farmers). (A differential treatment effect is one of the two components of selection bias, the other component being baseline bias.)

Economic Interpretation

The economic interpretation for the ATT is similar to that for the ATT, and is not repeated. The reason for this is that the model coefficients are very similar (this is seen from the model equation for ATT, given above – the coefficients on P (β_4) and RoundP (β_5) are identical to those for ATE, and the coefficient β_6 is not statistically significant).

5. Regression Estimates of ATE, Not Based on the Estimated Propensity Score

As discussed earlier in this section, an unbiased estimate of the ATE may be obtained from a regression model that expresses outcome as a function of explanatory variables. The explanatory variables include those that affect selection for treatment or outcome.This approach was examined, but not found to be not as useful (precise) as the propensity-score-based approaches. This estimator did not show statistically significant positive results for the program. The results are included in the .log file, but not presented here.

The situation is that while a good model could be developed to describe the probability of participation (“Treated”) as a function of explanatory variables, it was not possible to develop a very good regression model to describe outcomes of interest as a linear-model function of explanatory variables, without explicit (structural) representation of the probability of selection (through the highly nonlinear logistic regression model). In other words, the propensity-score-based models are considered to be better representations (specifications) of the causal model.

It is noted that the lack of statistically significant results (failure to reject the null hypothesis of zero impact) does not imply that the impact results produced by this model are inconsistent with the statistically significant results produced by the modified regression-adjusted propensity-score-based model. In statistical theory, a lack of evidence to reject a null hypothesis is not equivalent to acceptance of an alternative hypothesis (as in criminal justice, a verdict of not guilty is not equivalent to a declaration of innocence).

6. Instrumental-Variable Regression Estimates of ATE, Based on Estimated Propensity Score

The final set of impact estimators we examined is based on instrumental variables that reflect program participation. These models account for selection effects in a different way than the propensity-score approach, but they rely on a similar conditional-independence assumption (viz., that the instrument p is independent of (y_0, y_1, w_0, w_1)).

With the propensity-score method, a selection model is developed such that, conditional on **x**, w is uncorrelated with the response (y_0, y_1). With the instrumental-variables approach, the basic regression model that is specified is for outcome, rather than for selection. The situation that motivates the use of instrumental variables is that participation (Treated) may be correlated with the model error term, leading to biased and inconsistent estimates of the model parameters (regression coefficients) if they are estimated using the ordinary-least-squares (OLS) procedure. To obtain improved estimates, the model is supplemented with variables that are correlated with Treated but uncorrelated with (or have less correlation with) the model error term, given the covariates. The supplementary variables are called “instruments” or “instrumental variables.” The standard procedure for constructing estimates in this situation is the method of two-stage-least-squares (2SLS).

Using this approach, the full data set is used, including both the data from the original experimental design and the additional sample of Fintrac clients. In this approach, the estimated propensity score, $\hat{p}(\boldsymbol{x})$, is used as an instrumental variable for Treated.

The mathematical form of the instrumental-variable model is the same as described in the preceding subsection on regression estimates. What is different is the procedure for estimating the model parameters. In this application, the variables Treated and Round*Treated are considered endogenous, and the variables P and Round*P are used as instruments for them. The Stata procedure *xtivreg* is used to perform the estimation calculations.

Note that there is a fundamental difference between the regression-adjusted propensity-score-based estimator discussed earlier and the instrumental-variable regression estimator. Both involve the same variables (i.e., the estimated propensity score and explanatory variables from the questionnaire), but in the regression-adjusted propensity-score-based estimator the propensity score is a *regressor*, whereas in the instrumental-variable regression estimator the estimated propensity score is an *instrumental variable*. If the relationship of the instrumental variable (P) to the endogenous variable it represents (Treated) is weak, the instrumental-variable regression estimator is not very useful.

The instrumental variable (IV) models and associated estimates were constructed, but they were weak. This estimator did not show statistically significant positive results for the program. They are not presented here (they are included in the .log file).

As we discussed in the preceding section, the lack of statistically significant results (failure to reject the null hypothesis of zero impact) does not imply that the impact results produced by this model are inconsistent with the statistically significant results produced by the modified regression-adjusted propensity-score-based model.

The results for this IV regression estimator are weak because, as shown in the preceding subsection, the relationship of the outcome variables to the explanatory variables (using a simple linear statistical model) is weak. The weakness of the relationship is not because of weak instruments – the first-stage regression showed that the relationship of the endogenous variates to the instruments was in fact rather high. Consideration of more complex linear-regression models would improve the model fit, but it is viewed that the logistic-regression propensity-score-based models are a better structural representation of the causal model, and that this better representation is the main reason why those models provide much more precise estimates of impact. Consideration of precision (standard errors) is just one aspect of assessing the adequacy of a model. The greater face validity of the logistic regression selection model would suggest that the bias of estimators based on that model would be less than the bias of models that do not represent the selection process as well.

III.E Summary of Program Impact

The results of the impact evaluation show that the FTDA activity had a positive impact on its primary area of focus: income and activities related to horticultural crops. However, a broader positive impact on household income and expenditures was not detected.

The impact estimates were based on data that included all of the data obtained from the original experimental design, augmented by a sample of program farmers recruited by Fintrac in the course of its normal project operations. Statistical analysis was used to adjust for differences between the treatment and control samples, i.e., to reduce potential selection bias. The statistical analysis procedures used to estimate impact are based on sound causal models and causal-modeling theory (Neyman-Fisher-Cox-Rubin Causal Model, potential outcomes model, counterfactuals model). An *ex post* statistical power analysis was conducted, that showed that the study was not "underpowered." It is considered that the inferences made in this evaluation project are sound – valid and of adequate precision and power. It is the conclusion of this evaluation study that the FTDA program produces positive results relative to horticulture production, but those results are small in magnitude.

While the relationship of impact to treatment (program participation) is not strong, it is noted that the relationship of income to the estimated propensity score is very strong. Farmers similar to those selected for treatment tend to do well, even though they do not participate. Another way of looking at this is that Fintrac has an ability for selecting farmers who are likely to do well. This is not the same as a differential treatment effect (between treated and untreated farmers). Although Fintrac may have an ability to select farmers who are likely to do well (whether they participate or not), it does not appear that Fintrac has an ability to select those who are likely to perform noticeably better in the FTDA program than other program-eligible farmers. (A differential treatment effect is one of the two components of selection bias, the other component being baseline bias.)

The Impact Estimates Are Estimates of Causal Relationships

The impact estimates presented in this report are causal estimates of the effect of the program intervention. The ATE estimates are estimates of the effect on a randomly selected program-eligible farmer. The ATT estimates are estimates of the effect on a randomly selected treated farmer. The validity of these results depends on the validity of the household survey data and the supplementary sample of Fintrac-selected clients. The statistical models developed in this analysis were based on underlying causal models, and the effect estimates are estimates of the *causal relationship* of the outcomes to the program intervention.

We recognize that not all of the data used in the analysis is from a fully randomized experimental design (i.e., an experimental design in which randomization is used to select farmers from a population of program-eligible farmers, and to randomly assign the selected farmers to treatment or control). To overcome the lack of randomization, causal models were developed to describe the relationship of selection and outcome to explanatory variables, and impact estimates were obtained from these models that are unbiased or consistent, under the stated assumptions (about conditional independence (of treatement and the potential outcomes), given the explanatory variables in the model). The estimates of program impact are based on consistent estimators derived from statistical models that are in turn derived from causal models.

It is important to keep in mind that the term “statistical significance” refers to associational relationships, quite independently of whether the relationships are causal or simply associational (“correlational”). Whether those relationships represent causal relationships is not determined by the statistical model, but by a causal model and the relationship of the statistical model to it. *The estimates resulting from this analysis are estimates of the causal effect of the program intervention on the outcome variables.*

We recognize that some researchers refuse to attribute causality to program interventions unless randomization has been used in every instance to select experimental units and randomization has been used to assign treatment levels to the selected experimental units. There is a large body of scientific opinion that does not support this point of view. Much scientific progress has been made in recent centuries in settings in which randomization was not used at all. This progress belies the assertion that causal inferences can be made only if treatment levels are assigned using randomization for all experimental units. A very important consideration is the presence of forced changes in treatment variables (whether caused by randomization or other means). (The forced change supports compliance with Judea Pearl’s “back door” criterion of valid statistical estimation (identification) of causal relationships (causal effects).) The lack of randomization for some sample units makes the analysis and interpretation of results more difficult and more subject to threats to validity, but it does not alter the fact that the analysis presented here is based on causal modeling, and the estimates are estimates of the causal impact of the program intervention.

Summary of Assumptions and Limitations

The major assumptions associated with this analysis are the following:

1. The stable unit treatment value assumption (SUTVA, no macro effects assumption, partial equilibrium assumption) is made. This means that the effect (potential outcomes)

on one individual are not affected by potential changes in the treatment exposure of other individuals. This implies, for example, that the program is not so large that the outcomes are correlated (e.g., that farmers would produce such a large amount of horticultural crops that the market would collapse).

2. The causal models are correct. The key assumption here is that all important unobserved variables affecting selection are time invariant (i.e., are constant between the two survey rounds).

3. The program intervention represents a "forced change" in (experimental control of) the agricultural system in Honduras.

4. The half of the country that Fintrac had treated before this evaluation began is similar to the half yet to be treated, with respect to relationships among the important causal variables represented in the causal model underlying the statistical analysis.

Other more specific assumptions are listed for particular estimation equations in the detailed analysis presented in Annex 1.

The limitations of the evaluation are:

1. The causal analysis used to estimate impact is based on assumptions about the selection process. The original evaluation design was based on randomized assignment (of *aldeas*) to treatment, and represented a firmer basis for making causal inferences. With the original approach, randomized assignment assures that the distributions of explanatory variables (other than treatment) are the same for the treatment and control samples. With the revised design, this assertion depends on the correctness of the causal model, and the assumption that unobserved variables affecting selection for treatment are time-invariant.

ANNEX 2. HOUSEHOLD SAMPLE SURVEY DESIGN

I. Introduction

This annex is extracted from the project report, *Sample Selection for MCA - Honduras Program Evaluation*, 30 January 2008. The extract presented here is a verbal description of the sample design and sample selection process. That report also contained the sample design for the evaluation of the Transportation project, which is not relevant to this report. It also contained a list of all of the selected sample units and a number of tables describing the population (frame) and selected sample. Those tables and the sample list are not reproduced here. (This extract is an exact copy of selected material from the cited report, and it contains references to the omitted tables.) The Microsoft Excel file, *SurvDes4FTDAAldeas.mdb*, contains a complete listing of the sampling frame and selected sample, along with sample selection data such as the selection probabilities (used to generate the weights used in the analysis presented in this report).

II. Household Survey design for FTDA Program Evaluation

The sample selection for the household survey for the FTDA evaluation is similar in many respects to that for the household survey for the transportation-program evaluation. The major difference is that it uses *aldeas* instead of *caseríos* for the primary sampling unit, and it includes matching to identify a comparison sample.

The total number of *aldeas* in the sample frame (from the GIS, also from Census) was 3,675. After deleting *aldeas* in Islas de la Bahia and Gracias a Dios departments, those having 100% of *caseríos* in protected status, and those not to be processed by Fintrac over the next two years (mainly those already processed), the sample frame was reduced to 1,822 *aldeas*. These are the primary sampling units for the survey.

The design variables used for this sample selection are the same as those used in the household survey for the transportation evaluation. In addition to deleting all *aldeas* located in Islas de la Bahia and Gracias a Dios departments, only *aldeas* to be processed by Fintrac (the FTDA contractor) over the next two years are included in the sample. The *aldea* values for the design variables were obtained by aggregating the *caserío* values, using a suitable aggregation function. For variables on an interval or ordinal measurement scale, the mean or total was used as the aggregation function, as appropriate (e.g., total for population, mean for temperature). The statistical program package being used for the aggregation of the GIS-derived data (Stata) did not allow for aggregation using the mode, which is appropriate for non-ordinal variables (such as climate zone, vegetation cover, or protected status); those variables were hence dropped from the list of *aldea* design variables.

The procedure for selecting a matched sample of treatment and comparison *aldeas* is described in the “Sample Survey Design Details” memorandum. This procedure involves matching all population units with other units that are similar with respect to known variables that are considered to be related to program impact. These are the design variables listed earlier.

Table 6, entitled "*Aldea* Population Frequencies," shows the number of population units in each cell of the stratification. Table 7, entitled, "*Aldea* Desired Sample Allocation," shows the allocation of the sample to the stratum cells.

The matching process involves the use of weights to combine the distance between two units with respect to each matching variable into a single number. These weights (called "importance weights") reflect the relative importance of the variable in affecting the impact variable. They were specified so that the sum of the weights is approximately equal for the demographic / administrative variables, the GIS-derived travel-time variables, and the GIS-derived physiographic variables.

The process for selecting a matched sample is described in the referenced memo. At the end of the process, there are an equal number of matched items in the sample. For this application, it is desired that the sample contain fewer comparison units than treatment units. To accomplish this, a number of comparison units are randomly dropped from the sample.

The sample size desired for this survey is 113 treatment *aldeas* and 90 comparison *aldeas*. This sample is constructed by selecting a sample of 113 matched pairs (226 units in all), randomly dividing them into treatment and comparison *aldeas*, and dropping 23 of the comparison *aldeas* (resulting in the desired sample size of 90 comparison *aldeas*).

The desired stratum allocation for this survey differs significantly from the desired allocation for the transportation survey. For this survey, travel-time variables are of less interest for stratification.

Table 8, entitled, "*Aldea* Actual Sample Frequencies," shows the distribution of the sample units over the strata. Table 9, entitled, "*Aldea* Sample for the FTDA Program Evaluation," contains a list of the *aldeas* selected for the sample. The column "InSample" specifies the nature of the sample item: 3 indicates a comparison *aldea* that has been dropped from the original sample of 226 matched pairs (23 of them); 2 indicates a comparison *aldea* that is retained in the sample (90 of them); and 1 indicates a treatment *aldea* (113 of them).

A map was prepared (in the referenced file Sample5_maps1.pdf) to show the geographic distribution of the treatment *aldeas*, the comparison *aldeas*, and the dropped comparison *aldeas*. The map displays reasonable geographic distribution.

Supplementary Material from Transportation Sample Survey Design

The following supplementary material about the sample design is taken from the description of the sample survey design for the Transportation Project. It is relevant since the sample frame for the FTDA study was obtained from the sample frame for the Transportation study, by aggregating the *caseríos* to the *aldea* level. Some of the design variables used in both surveys were obtained from 2001 Honduras Census of Households data, and many were obtained from geographic information system data obtained from the government of Honduras.

The sample frame was a list of 22,816 *caseríos* stored in the GIS (and the same as those available from the Census). *Caserío*s in the Islas de la Bahia and Gracias a Dios departments, and *caseríos* in protected status, were excluded as out of scope, reducing the sample frame size to 20,467. These are the primary sampling units (PSUs) for the study.

The design process began by identifying all known variables that might have had an effect on selection of roads for the program, and that may have a significant relationship on impact. (The word "known" means that data on these variables are available to assist survey design, i.e., are available prior to implementation of the survey data collection.) Three sets of variables were identified: (1) basic demographic and administrative variables, such as population size, agricultural region and urban/rural status; (2) data from the 2001 Census of households; and (3) GIS data.

The purpose of the planned survey is to collect data in support of the development of an analytical model, and the approach to designing the survey followed established procedures for constructing an analytical survey design (see the memorandum, "Sample Survey Design Details" for some background on this methodology).

The first step in the process was to examine the Cramer coefficient of correlation (a nonparametric measure of correlation) among all variables. The variables fall into two categories – those that are known to be closely related to the dependent variables of interest, and those that are simply candidate explanatory (independent) variables. Examples of the former are various measures of travel time and changes in travel time anticipated to be caused by the program intervention, from *caseríos* to various points of interest. Examples of the latter are demographic and administrative variables, Census variables, and physiographic variables produced by the GIS. The purpose of examining the Cramer coefficients is obtain information to guide combining explanatory variables that are highly correlated with each other (or delete some of them), and (to a lesser extent) to eliminate variables that show little relationship to the variables that are considered to be closely related to the dependent variable.

After conducting the analysis of the Cramer coefficients, it was decided that the large number of variables from the Census could be replaced by the Basic Necessities Index (NBI). Of the other variables, it was decided to retain all of the remaining demographic / administrative variables (agricultural region, urban/rural status, and population size), and most of the GIS variables. The sample selection was to be done in such a way as to ensure substantial variation on these variables, and low correlation among them. The procedures described in the "Sample Survey Design Details" memorandum were employed to select the sample.

The following is the final list of variables used in the survey design for the sample of *caseríos*. These variable names will appear in a number of tables to be presented to describe the sample design and sample selection process. The list presented below contains the definition of the variable as extracted from the GIS, and the definition of the recoding used in the sample-selection process. In the following, all travel times are in minutes, and all distances are in meters.

POP: population (individuals, not viviendas (households)
AUTO: [GIS definition?]; recoded in quintiles.
NRDIST2: distance (in meters) to nearest road (of any type); recoded in quintiles.
NRDSTT: travel time (in minutes) to nearest road (of any type); recoded in quintiles.
TTTEGUS1: travel time (in minutes) to Tegucigalpa before MCA improvements; recoded in quintiles.
TTEGUS1: same, after MCA improvements; recoded in quintiles.
TTEGUSD: reduction in travel time to Tegucigalpa after MCA improvements; recoded in quintiles.

TT133PP1: travel time to the nearest of the 133 largest Honduran cities and towns, before MCA improvements; recoded in quintiles.
TT133PP2: same, after MCA improvements; recoded in quintiles.
TT133PPD: reduction in travel time (to nearest 133 largest Honduran towns), due to MCA improvements; recoded in quintiles.
TT10CITY1: travel time to nearest of 10 largest Honduran cities, before MCA improvements; recoded in quintiles.
TT10CITY2: same, after MCA improvements; recoded in quintiles.
TT10CITYD: reduction in travel time due to MCA improvements; recoded in quintiles.
TTPC1: travel time to Puerto Cortes before MCA improvements; recoded in quintiles.
TTPC2: same, after MCA improvements; recoded in quintiles.
TTPCD: reduction in travel time due to MCA improvements; recoded in quintiles.
Region: agricultural region code (0-9); not recoded.
Urban: urban classification (1 = urban, 0 = rural); was recoded from 1 = urban and 2 = rural.
NBI; index of basic necessities; recoded in quintiles.
PRDIST: distance to nearest primary road; recoded in quintiles.
SRDIST: distance to nearest secondary road; recoded in quintiles.
CA5DIST: distance to nearest point on CA-5 highway; recoded in quintiles.
CITY5DIST: distance to the nearest of the five largest Honduran cities; recoded in quintiles.
MCAPSDIST: distance to the nearest point on MCA primary or secondary road improvement location; recoded in quintiles.
ELEVATION: elevation of the *caserío*, in meters; recoded in quintiles.
CLIMZONE: climate zone code; 0-9 unrecoded, values above 9 recoded as 9.
SOILCAP: soil capacity for the soil of the *caserío*; recoded in quintiles.
RAINREG: major rainfall rain code; evidently ordinal, so recoded in quintiles.
PREC_MM: median annual rainfall precipitation in millimeters; recoded in quintiles.
TEMP: median annual temperature in degrees Celsius; recoded in quintiles.
VEGCOVER: code for different types of major vegetation cover across Honduras; not recoded.
PROTAREAS: 1 if the *caserío* is in a nationally protected area or national park, 0 if not; not recoded.
HYDRODIS: distance in meters to the nearest major river (for all *caseríos*); recoded in quintiles.
TTMCAP1: travel time to nearest point on MCA-improvement primary road, before MCA improvements; recoded in quintiles.
TTMCAP2: same, after MCA improvements; recoded in quintiles.
TTMCAPD: reduction in travel time due to MCA improvements; recoded in quintiles.
TTMCAS1: travel time to nearest point on MCA-improvement secondary road, before MCA improvements; recoded in quantiles.
TTMCAS2: same, after MCA improvements; recoded in quantiles.
TTMCASD: reduction in travel time due to MCA improvements; recoded in quintiles.
TTC1000_1: travel time to nearest *caserío* with a population greater than 1,00 people (there are about 500 of these), before MCA improvements; recoded in quintiles.
TTC1000_2: same, after MCA improvements, recoded in quintiles.

TTC1000_D: reduction in travel time due to MCA improvements; recoded in quintiles.
TTMCAR1: travel time to nearest point on MCA tertiary improvement segment, before all road improvements; recoded in quintiles.
TTMCAR2: same, after MCA improvements; recoded in quintiles.
TTMCARD: reduction in travel time due to MCA improvements; recoded in quintiles.

All of the variables listed above were used as stratification variables in the design. The stratum categories are defined by the coding specified in the above list (e.g., the quintiles (values 0-4) or the agricultural region code (values 0-9)). In most cases, the stratum boundaries were determined by quantiles, so that the population frequencies are equal or similar for the various categories (they would normally be the same, but some variables contain large numbers of identical values, causing the numbers in the various quantile categories to differ from uniform). The use of quantiles to define the stratum boundaries is justified in cases where the nature of the relationship of impact to the variable is not well known. Since this is a "groundbreaking" study, the nature of the relationships is not known. The use of quantiles to define stratum boundaries is oriented toward "nonparametric" representations of relationships between impact and explanatory variables (since too little is known about the relationships to assume particular parametric forms (e.g., a linear relationship) at the present time). Another advantage of using quantile-defined stratum boundaries is that the stratum allocation results are unaffected by the particular measurement scales used for the variables. It is therefore particularly convenient when considering nonparametric relationships (it is also appropriate for parametric analyses, such as maximum likelihood estimation, which are invariant with respect to reparameterization (e.g., a scale transformation)). (Note that the methodology allows for completely arbitrary definitions of stratum boundaries. The choice of quantiles (e.g., quintiles) was guided strictly by conceptual desirability, and had nothing to do with any technical or programmatic constraints. The boundaries could have been specified using any points on any measurement scale. The use of quantiles is indicated because the nature of the relationship of impact to the design variables is not known. In a later study, the stratum boundaries may would likely be set using natural-scale values.)

Most of the stratum cells were defined by quintile quantiles, and the codes for the five quintile categories (0-20%, 20-40%, 40-60%, 60-80% and 80-100%) are 0-4 (shown in the column headings of the tables to be presented later). For the stratum cells that are not defined by quintiles (or other quantiles), the following criteria defined the stratum boundaries. Each of the columns of the tables corresponds to a different stratum category, or "cell," and the code value of the cell is the table heading.

Prior to beginning the analysis, all *caseríos* from the Islas de la Bahia and Gracias a Dios departments were deleted, since they are of little relevance to this program. All *caseríos* in a protected status were also deleted.

Agricultural Region ("Region")

0. Islas de la Bahia (all units deleted)
1. Sur
2. Centro Occidental
3. Norte
4. Litoral Atlantico

5. Norte Oriental
6. Centro Oriental
7. Occidental

Urban / Rural Status ("Urban")
Rural
Urban

Climatic Zone ("Climzone")
Values 1-8 as stored in the GIS; all values greater than 8 coded as 9.
Vegetation Cover ("Vegcover")
Code values 1-9 taken directly from GIS
Protected Areas ("Protareas")
Unprotected
Protected (all deleted from the sample frame)

Additional description of sample design details is presented in the document, *Sample Survey Design Details*, dated 27 December 2007 (revised 30 January 2008). Discussion of modifying the original experimental design to a revised design is presented at a conceptual level in the document *FTDARedesign20090828.ppt* and in greater detail in the document, *Alternative Design for FTDA/EDA Program Evaluation*, dated 12 May 2010.

It should be noted that the analytical survey design was constructed to assure adequate variability (balance, spread, orthogonality) in explanatory variables related to outcomes of interest. This approach is useful (but not necessary) for an experimental design, since it enables the estimation of the relationship of impact to explanatory variables. It is much more important for the revised design, which involves the use of covariate-adjusted regression models to estimate impact. It is very important to have adequate variation in explanatory variables in order to develop good regression models relating selection and outcome to explanatory variables. In retrospect, while the analytical survey design would have proved useful had the originally planned experimental design been implemented, it was the best possible choice for use with the revised design.

ANNEX 3. STATISTICAL POWER ANALYSIS

I. Introduction

The sample sizes (numbers of sample *aldeas* and number of sample households per sample *aldea*) were determined in consideration of the statistical power desired to detect impacts of specified magnitude.

The sample sizes that were decided on were 113 treatment *aldeas* and 90 control *aldeas*, with an expected sample size of 9 program farmers and 20 other farmers in treatment *aldeas*, and 9 potential treatment farmers and 20 other farmers in control *aldeas*, for a total sample size of 203 *aldeas* and (expected) 203 x 29 = 5887 households in each survey round.

Over the course of the design phase of the study, many power calculations were done, under various assumptions about population and design characteristics. As discussed in the main text of this report, the original evaluation design for estimation of impact was an experimental design, for which the estimate of impact would be the double difference estimator (the difference, between the treatment and control samples, of the difference in means before and after the program intervention. The sample sizes at the *aldea* and household levels were determined by means of statistical power analysis, in which sample sizes were determined to provide a specified level of power for detecting an impact (double difference measure) of specified size.

In the initial reviews of this *Final Report*, a concern was voiced that, because the evaluation was not finding impacts of the expected magnitude, the evaluation design may have been underpowered. As discussed in Annex 1 in detail (in the *ex post* power analysis), the study was not at all underpowered. In the design stage, much consideration was given to consideration of power in the determination of sample size. This Annex presents some of that analysis. It is presented in somewhat greater detail than would be customary, because of the concern raised over the power of the impact estimates. There was much consideration of and detailed discussion of power in determining the sample sizes (*aldeas* and households) for the study.

To conduct a statistical power analysis, information is required on a number of items: the impact estimator being used; the test parameters (power level, significance level); the minimum detectable effect; characteristics of the sampled (target) population (means, standard deviations, intra-unit correlation coefficients (if multistage sampling is used); and the sample design to be used for the sample survey. In the design phase, not a lot was known about the statistical properties of many of the outcome variables of interest in this study. For income, analysis of available data showed that the coefficient of variation (ratio of standard deviation to mean) of rural incomes in Honduras was about two. For estimation of proportions in the vicinity of .5, the coefficient of variation is also about one. This value (coefficient of variation equal to one) was assumed for the power calculations presented below.

The FTDA progam was implemented through *aldeas* (villages), and the *aldea* was selected as the first-stage sampling unit (primary sample unit, PSU). The Kish design effect (deff) used in the following formulas includes loss of precision associated with multistage sampling. The formula for deff is deff = 1 + (m-1)icc, where m denotes the number of sample households selected per sample PSU and icc denotes the intra-unit correlation coefficient. The value of icc depends on the outcome variable. For example, for m = 20 and icc=.2, the value of deff is 4.8. In the following, the value of deff is modified by a factor to reflect design effects additional to multistage sampling, such as the effects of stratification and matching of treatment and comparison roads.

In the course of the project planning, a number of different cases were examined, corresponding to a range of values of the parameters that affect power and sample size. The *Millennium Challenge Account-Honduras Monitoring and Evaluation Plan* (undated) estimated (page 5) that the economic rate of return (ERR) for the FTDA project would be 36% . The table shown below represents part of the "sensitivity analysis" that was conducted to determine the *aldea* and household sample sizes.

The material shown below is extracted from several documents and memoranda that were prepared in the design phase of the study. These include *Design Report* dated October 30, 2007; *Estimation of Power for Varying Sample Size* dated 31 October 2007; and *Revised Sample-Size Estimates*, FTDA Survey, dated 30 November 2007. In the course of the power analysis used to estimate sample size, a number of different cases were examined. The table presented below examines one selection of parameters.

The next section deals with determination of the household sample size within *aldeas*, and the section after that deals with determination of the *aldea* sample size.

II. Determination of Household Sample Size within Caseríos

The standard approach to determining sample size (for 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 (census segments) 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 not only for census segments, but also for two other area sampling units (*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. 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 subsample of clusters of similar size.)

It is expected that an interviewer could conduct two to four household interviews (lasting about an hour) per day, once present in the Census segment. 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 = $c_1n + c_2nm$
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 "power" 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 little 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.

It is noted that the formula for determining sample size using statistical power analysis can be expressed in terms of the coefficient of variation (CV) of the outcome variable. The coefficient of variation for a proportion equal to .5 is 1.0. The coefficient of variation of income in rural areas of developing countries is in the range .5-2. For Honduras, an analysis of Census data showed that the CV for income in Honduras was about 2. The CV varies for each outcome variable. The table presented below applies to any variable for which the coefficient of variation is 1.0.

Note that the total household sample size depends on the intra-unit correlation coefficient (icc) for the first-stage sample units (PSUs). The icc affects the within-PSU household sample size. Once the within-PSU sample size, m, has been specified, the PSU sample size depends only on the icc and m through the design effect (deff). The sample size formula to be presented is in terms of the deff, not icc and m.

[The material from the *Design Report* dealing with sample size estimation based on specification of precision is omitted here, since it was the power-based estimates that were used.]

III. Estimation of Sample Size Based on Specification of Statistical Power

The formula for the power of a test of hypothesis about a mean double difference is as follows:

$$\Pr\left(\frac{\hat{\mu}_1-\hat{\mu}_2-\hat{\mu}_3+\hat{\mu}_4}{(deff\ \mathrm{var}(\hat{\mu}_1-\hat{\mu}_2-\hat{\mu}_3+\hat{\mu}_4))^{1/2}}\right)>z_{1-\alpha}\left|\mu_1-\mu_2-\mu_3+\mu_4=D\right)=1-\beta$$

where

$$\mathrm{var}(\hat{\mu}_1-\hat{\mu}_2-\hat{\mu}_3+\hat{\mu}_4)=\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}+\frac{\sigma_3{}^2}{n_3}+\frac{\sigma_4{}^2}{n_4}-\frac{2\rho_{12}\sigma_1\sigma_2}{\sqrt{n_1n_2}}-\frac{2\rho_{13}\sigma_1\sigma_3}{\sqrt{n_1n_3}}+\frac{2\rho_{14}\sigma_1\sigma_4}{\sqrt{n_1n_4}}$$

$$+\frac{2\rho_{23}\sigma_1\sigma_3}{\sqrt{n_1n_3}}-\frac{2\rho_{24}\sigma_2\sigma_4}{\sqrt{n_2n_4}}-\frac{2\rho_{34}\sigma_3\sigma_4}{\sqrt{n_3n_4}}$$

where

μ_1 = mean for group 1 (treatment, time 1)
μ_2 = mean for group 2 (treatment, time 2)
μ_3 = mean for group 3 (comparison, time 1)
μ_4 = mean for group 4 (comparison, time 2)
n_1 = sample size for group 1
n_2 = sample size for group 2
n_3 = sample size for group 3
n_4 = sample size for group 4
σ_1 = standard deviation for group 1
σ_2 = standard deviation for group 2
σ_3 = standard deviation for group 3
σ_4 = standard deviation for group 4
ρ_{12} = correlation between items of groups 1 and 2
ρ_{13} = correlation between items of groups 1 and 3
ρ_{14} = correlation between items of groups 1 and 4
ρ_{23} = correlation between items of groups 2 and 3
ρ_{24} = correlation between items of groups 2 and 4
ρ_{34} = correlation between items of groups 3 and 4

(The correlation matrix should be positive definite.)
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 (The design effect is the ratio of the variance of an estimate for a specified survey design, compared to the variance using simple random sampling.)
D = (true) size of the mean double difference
and a caret (ˆ) over a symbol denotes a sample estimate.

As mentioned, this sample-size analysis will focus on determining the sample size for *aldeas* rather than households (farmers). It is estimated that there will be three lead farmers in each *aldea*, and two beneficiaries for each lead farmer, for a total of nine program farmers per *aldea*. It is known that the number of program farmers per *aldea* can vary substantially over the range 3-20. The standard deviation of means of samples of nine is one-third that of the individual elements. The original analysis assumed that the "panel" correlation (i.e., the correlation between a household at the beginning of the study and the same household at the end of the study) was .5. (The correlation of (matched) sample means is the same as the correlation of the elements comprising the mean.) This assumption is somewhat stringent / conservative; a less conservative value of .7 will be assumed for the present analysis.

In the previous analysis, the cluster subsample size (number of households randomly sampled per *aldea*) was set at the "optimal" value of 20. That number is still considered to be reasonable. The number 20 was determined by taking into account the relative costs of sampling *aldeas* versus households, and the intracluster correlation coefficient. It is not a "cast-in-stone" number – decreasing (or increasing) it somewhat from the value of 20 would not have a substantial effect on the sample-size estimates.

The principal impact measure of interest in this study is a double-difference estimator of program impact for the program participants, based on comparing treatment and control *aldeas*, before and after treatment (program intervention). We will also examine other estimates, such as the "spillover" effect – a double-difference estimator of program impact for the non-program-participants ("others").

The standard deviations referred to in the formula are the standard deviations of the mean income of farmers in an *aldea* – either program farmers or non-program farmers, depending on the analysis objectives. The number of program farmers may be about 9, and the number of non-treatment farmers will be fixed at 20. The sample size estimates will vary depending on which number is used

Sample Sizes and Power of Tests for Estimation of Program Impact for Program Participants

Initially, it will be assumed that the sample size for all four groups comprising the double difference estimate will be identical. Then, some sample-size estimates will be made assuming that the comparison group is smaller than the treatment group. (The before-and-after sample sizes are the same, since a panel survey is being done.)

In the previous analysis, the design effect was set equal to 3.85. This is the design effect for estimates based on a two-stage sample (*aldeas*/households, with an intracluster correlation coefficient of .15 and samples of 20 households per cluster), and (conservatively) assumes no precision improvement from other design features, such as stratification, matching or "blocking" of the *aldeas*. Recently (this past week), a series of regression analyses were conducted to determine the relationship of income to variables that may be used for stratification, matching or blocking, including urban/rural status, basic-necessities index, and agricultural region. (The regression analysis used data from the Encuesta Permanente de Hogares (EPHM).) The best regression had a coefficient of determination (R-squared, the proportion of the variance in the dependent variable (income) that is explained by the regression model) of .25 (i.e., 25 percent). For socioeconomic data, this is a high value. Because *aldeas* are generally larger than Census segments, it is expected that the value of R-squared would be somewhat lower for regressions of mean *aldea* income. These regression equations used the logarithm of the mean Census-segment income (samples of 10 households per Census segment) as the dependent variable, whereas in our data analysis we will use the household as the unit of analysis. The value of R-squared would be expected to be higher for the household-level analysis than for the Census-segment-level analysis. Based on this information, it is viewed that the effect of "blocking" the *aldeas* according to the listed variables will substantially reduce the design effect. For the present analysis, it will be assumed that the design effect (associated with sampling of clusters, and with stratification / matching / blocking, but apart from the effect of the correlation coefficients) is 1.0. The justification for this low value is the relatively strong relationship observed between (Census-segment) income and the variables to be used for stratification / matching / blocking, and the fact that the precision of the estimates in the data analysis (based on using the household as the unit of analysis) will be greater than that associated with the "raw" double-difference estimator used in the above formula.

As mentioned above, the value of the panel correlation coefficient will be assumed to be .7. Based on the regression analysis of income, it is expected that blocking of the *aldeas* (into stratum cells each of which contains one treatment *aldea* and one comparison *aldea*) will produce matched pairs for which the correlation is rather high. A value of .5 will hence be assumed for the correlation between *aldeas* of the treatment and comparison groups at time 1. Based on these two assumptions, the following values will be assumed for the various intergroup correlation coefficients: $\rho_{12} = .7$, $\rho_{13} = .5$, $\rho_{14} = .2$, $\rho_{23} = .2$, $\rho_{24}=.4$, $\rho_{34} = .7$.

In summary, the following values will be assumed:

μ_1 = 4,617 (current income, treatment group (i.e., time 1))
μ_2 = 9,234 (anticipated income after project intervention, treatment group (i.e., time 2))
μ_3 = 4,617 (current income, comparison group (i.e., time 1))
μ_4 = 4,617 (anticipated income after project intervention, comparison group (i.e., time 2))
σ_1 = 10,857 / sqrt(9) = 3,619 (current standard deviation of mean income (for samples of 9), treatment group (time 1))
σ_2 = 21,714 / sqrt(9) = 7,238 (anticipated standard deviation of mean income after project intervention (for samples of 9), treatment group (time 2)) [It is assumed that the standard deviation is double, since the income is assumed to double.]

σ_3 = 10,857 / sqrt(9) = 3,619 (current standard deviation of mean income (for samples of 9), comparison group (time 1))
σ_4 = 10,857 / sqrt(9) = 3,619 (anticipated standard deviation of mean income after project intervention (for samples of 9), comparison group (time 2))
$\rho_{12} = .7$
$\rho_{13} = .5$
$\rho_{14} = .4$
$\rho_{23} = .4$
$\rho_{24} = .4$
$\rho_{34} = .7$
$\alpha = .05$ (corresponding to $z_{1-\alpha} = 1.6449$)
$\beta = .1$
$1 - \beta = .9$ (the power of the test)
$z_{1-\alpha}$ = 1-α percentile point of normal distribution (e.g., 1.6449 for α=.05, or 1.2816 for α=.1)
deff = 1.0
D = 4,619 (i.e., we are determining the power of the sample for detecting a mean double difference in income equal to the mean income, i.e., the change in income for the treatment group is greater than the change for the comparison group by an amount equal to the current mean income).

Under the preceding assumption, the sample size is 14 *aldeas* in each group, for a total of 56 *aldeas* in both waves of the panel survey (28 at time 1 and 28 at time 2). As discussed above, the number of program farmers is approximately 9 per *aldea* (this number is what it is – it is not specified as part of the sample design), and the number of randomly selected other farmers is 20. The number of sample households interviewed is hence approximately 29 times the number of sample *aldeas*.

If the value of D is set equal to half the present income (i.e., we are determining the power of the sample for detecting a mean double difference equal to half the current mean income), D = 2,308.5, then the sample size is 56 *aldeas* per group.

In the preceding, the estimator under consideration is the overall mean double-difference estimate. If it is desired to compare estimates of the double-difference estimator for subpopulations, the sample sizes should be increased by a factor of 4 (this factor corresponds to random splits of the sample into two equal parts – the factor would be larger for smaller subsamples).

In the preceding, the sample sizes of the four groups are held constant. If it is desired to lower the sample size of the comparison group to .8 times the size of the treatment-group sample size, then the sample sizes will increase slightly for the treatment group, and for the overall sample (since this design is a little less efficient).

These results are summarized in the following tables.

We note that the sample size estimates are in some cases rather sensitive to the values of the various parameters. Since these values are uncertain, it is advisable to examine a variety of alternative parameter values, and select a sample size that is adequate for a wide range of likely parameter values.

Estimates of Aldea Sample Size, FTDA Evaluation

Scenario 1: Power = .9; Sample designed to estimate an *overall* mean double difference for program farmers; equal-sized treatment and comparison groups.

Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment *Aldeas,* Time 1	Number of Comparison *Aldeas,* Time 1	Number of Treatment *Aldeas,* Time 2	Number of Comparison *Aldeas,* Time 2	Total *Aldea* Sample Size (both waves of panel survey)
.5	56	56	56	56	**224**
1.0	14	14	14	14	**56**

Estimates of Aldea Sample Size, FTDA Evaluation

Scenario 2: Power = .9; Sample designed to enable comparison of mean double differences of program farmers *for population subgroups*; equal-sized treatment and comparison groups.

Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment *Aldeas,* Time 1	Number of Comparison *Aldeas,* Time 1	Number of Treatment *Aldeas,* Time 2	Number of Comparison *Aldeas,* Time 2	Total *Aldea* Sample Size (both waves of panel survey)
.5	224	224	224	224	**896**
1.0	56	56	56	56	**224**

Estimates of Aldea Sample Size, FTDA Evaluation

Scenario 3: Power = .9; Sample designed to estimate an *overall* mean double difference for program farmers; number of comparison-group *aldeas* = .8 times number of treatment-group *aldeas*.

Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment Aldeas, Time 1	Number of Comparison Aldeas, Time 1	Number of Treatment Aldeas, Time 2	Number of Comparison Aldeas, Time 2	Total Aldea Sample Size (both waves of panel survey)
.5	58	46	58	46	**208**
1.0	15	12	15	12	**54**

Estimates of Aldea Sample Size, FTDA Evaluation

Scenario 4: Power = .9; Sample designed to enable comparison of mean double differences of program farmers *for population subgroups*; number of comparison-group *aldeas* = .8 times number of treatment-group *aldeas*.

Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment Aldeas, Time 1	Number of Comparison Aldeas, Time 1	Number of Treatment Aldeas, Time 2	Number of Comparison Aldeas, Time 2	Total Aldea Sample Size (both waves of panel survey)
.5	232	184	232	184	**832**
1.0	60	48	60	48	**216**

The following table shows the statistical power associated with a sample of 100 treatment *aldeas* (at time 1) and varying sizes of comparison-group *aldeas* (at time 1).

Statistical Power Estimates (for program farmers), FTDA Evaluation			
Scenario 5: Sample of 100 treatment *aldeas* (at time 1) and varying numbers of comparison-sample *aldeas* (at time 1)			
Size of Double Difference to Detect, as a Fraction of the Mean	Number of sample *aldeas* in comparison group (time 1)	Power of a test of the overall mean double-difference estimate of program impact	Power for comparing the mean double-differences of subpopulations
.5	100	.99	.63
1.0	100	.999	.99
.5	90	.99	.62
1.0	90	.999	.99
.5	80	.99	.62
1.0	80	.999	.99
.5	70	.99	.60
1.0	70	.999	.99
.5	60	.98	.59
1.0	60	.999	.98
.5	56	.98	.57
1.0	56	.999	.98
.5	50	.975	.56
1.0	50	.999	.975

Sample Sizes and Power of Tests for Estimation of Program Impact for Non-Program Farmers (the "Spillover" Effect)

The preceding analysis was directed to determining the (*aldea*) sample size required to achieve a statistical power of .9 (90%), for tests involving the mean double difference of income for *program participants* (lead farmers and beneficiaries). We shall now conduct a similar analysis to determine the sample size required to achieve a power of .9 for tests involving the mean double difference of income for *non-program farmers*.

It will be assumed that the within-cluster sample size will be set at 20 households, and the problem is to determine the *aldea* sample size subject to this condition.

There are just a few parameter values that need to be changed for this analysis, from the values used above. First, we shall consider the case in which it is desired to detect a change of 5 percent and 10 percent of current income, rather than 50 percent or 100 percent as for program participants (i.e., D = .05 x 4,617 = 230.85 and .10 x 4,617 = 461.9, instead of .5 x 4,617 = 2,308.5 and 1.0) x 4,617 = 4,617. Second, the standard deviation of income is not expected to increase in the second wave of the survey (i.e., it remains at 10,857, and does not increase to 21,714). Third, the standard deviation of the cluster means is the element standard deviation divided by sqrt(20), rather than sqrt(9), so that the standard deviation of the cluster means is 10,857 / sqrt(20) = 10,857/4.472 = 2,428.

Estimates of Aldea Sample Size (*for non-program farmers*), FTDA Evaluation					
Scenario 6: Power = .9; Sample designed to estimate an overall mean double difference *for non-program farmers*; equal-sized treatment and comparison groups.					
Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment *Aldeas,* Time 1	Number of Comparison *Aldeas,* Time 1	Number of Treatment *Aldeas,* Time 2	Number of Comparison *Aldeas,* Time 2	Total *Aldea* Sample Size (both waves of panel survey)
.05	948	948	948	948	**3792**
.1	237	237	237	237	**948**

These sample sizes are "massive" – because of the substantial variation in income, very large samples are required to detect small changes in income (i.e., mean double differences on the order of 5-10 percent of the pre-treatment mean).

The following table shows the power associated with a test of the mean double difference for non-program farmers, using the sample size depicted in the first table presented above (for estimating the mean double difference of program farmers).

Statistical Power Estimates (*for non-program farmers*), FTDA Evaluation			
Scenario 7: Power of tests involving the mean overall double difference *for non-program farmers*, using the *aldea* sample sizes considered above			
Size of Double Difference to Detect, as a Fraction of the Mean	Number of sample *aldeas* in treatment group (time 1)	Power of a test of the overall mean double-difference estimate of program impact	Power for comparing the mean double-differences of subpopulations (not calculated – small)
.05	14	.10	
.1	14	.17	
.05	56	.17	
.1	56	.41	
.05	100	.24	
.1	100	.60	
.05	224	.41	
.1	224	.88	

The preceding tables show that it will be difficult to detect a "spillover" impact (change in mean double difference of income for non-program farmers), using the sample sizes determined to produce satisfactory power for impact estimates for program farmers. The situation is not "hopeless," since it is possible to determine substantially more precise estimates of impact using analytical techniques that are more sophisticated than the raw double-difference estimate assumed for the above calculations (i.e., statistical regression models involving various explanatory variables, instead of the mean double difference used above). (Also, the significance level of the test could be increased from α = .05 to α = .10. This would increase the power (1 – probability of making a Type II error) of the tests, at the expense of increasing the probability of making a Type I error.)

Additional material on sample size estimation is presented in the documents, *Additional Sample Size Estimates*, dated 21 December 2007 and *Estimation of Sample Size*, dated 18 July 2008. The former document contains the sample sizes actually used for the survey, and is reproduced here.

This memorandum presents estimates of sample sizes for the FTDA household survey corresponding to assumptions that were discussed in yesterday's telephone conference. The cases examined here are extension of the "Scenario 3" case presented in the memo of 30 November 2007 entitled, "Revised Sample Size Estimates, FTDA Survey." The main characteristics of that case are that the comparison group is .8 as large as the treatment group, and the principal analysis objective of the survey is to achieve high power for tests of hypotheses about an overall double-difference of program impact. Since the four cases presented here are extensions of the Scenario 3 case, they will be named Scenarios 3b, 3c, 3d and 33.

The following tables present estimates of the sample sizes, in terms of number of *aldeas*, required to achieve a certain level of statistical power for a test of hypothesis about the size of a double-difference estimate of program impact. The estimate requires specification of the values of a number of parameters in a sample-size formula. The formula is presented in the referenced memorandum. All of the parameter values specified on page 5 of that memo will be assumed here, except as noted below.

The sample size in terms of households (farmers) is equal to the *aldea* sample size times the sum of the number of program farmers (leads plus beneficiaries) plus nonprogram farmers sampled. Based on consideration of the relative costs of sampling *aldeas* versus households, and on the likely value of the intra-*aldea* correlation coefficient, the optimal (efficient) nonprogram farmer sample size is about 20. This value is somewhat larger than the expected number of program farmers (estimated to be about 9 – an average of three lead farmers and six beneficiary farmers (two per lead)). It could be reduced slightly with little loss in efficiency (or power or precision).

Scenario 3b: This case will estimate the sample size required to achieve a power of .9 for the test of the hypothesis that the mean double difference of income *of program lead farmers* increases by an amount equal to 0.5 times the base-year income. It will be assumed that the average number of lead farmers per *aldea* is three.

For this case, the following values will be assumed:

μ_1 = 4,617 (current income, treatment group (i.e., time 1))
μ_2 = 6,926 (anticipated income after project intervention, treatment group (i.e., time 2))
μ_3 = 4,617 (current income, comparison group (i.e., time 1))
μ_4 = 4,617 (anticipated income after project intervention, comparison group (i.e., time 2))
σ_1 = 10,857 / sqrt(3) = 6,268 (current standard deviation of mean income (for samples of 3), treatment group (time 1))
σ_2 = 1.5 x 10,857 / sqrt (3) = 16,286 / sqrt(3) = 9,403 (anticipated standard deviation of mean income after project intervention (for samples of 3), treatment group (time 2)) [It is assumed that the standard deviation of income is proportional to the mean.]
σ_3 = 10,857 / sqrt(3) = 6,268 (current standard deviation of mean income (for samples of 3), comparison group (time 1))
σ_4 = 10,857 / sqrt(3) = 6,268 (anticipated standard deviation of mean income after project intervention (for samples of 3), comparison group (time 2))
ρ_{12} = .7
ρ_{13} = .5
ρ_{14} = .4
ρ_{23} = .4
ρ_{24} = .4

$\rho_{34} = .7$
$\alpha = .05$ (corresponding to $z_{1-\alpha} = 1.6449$)
$\beta = .1$
$1 - \beta = .9$ (the power of the test)
$z_{1-\alpha}$ = 1-α percentile point of normal distribution (e.g., 1.6449 for α=.05, or 1.2816 for α=.1)
deff = 1.0
D = 2,308 (i.e., we are determining the power of the sample for detecting a mean double difference in income equal to one-half the mean income, i.e., the change in income for the treatment group is greater than the change for the comparison group by an amount equal to one-half the current mean income).
n2 = 1.0 n1
n3 = .8 n1
n4 = .8 n1

These results are summarized in the following table.

Estimate of Aldea Sample Size, FTDA Evaluation

Scenario 3b: Power = .9; Sample designed to estimate an overall mean double difference for *lead program farmers*; number of comparison-group *aldeas* = .8 times number of treatment-group *aldeas*.

Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment *Aldeas*, Time 1	Number of Comparison *Aldeas*, Time 1	Number of Treatment *Aldeas*, Time 2	Number of Comparison *Aldeas*, Time 2	Total *Aldea* Sample Size (both waves of panel survey)
0.5	106	85	106	85	**382**

Scenario 3c: This case will estimate the sample size required to achieve a power of .9 for the test of the hypothesis that the mean double difference of income *of program beneficiary farmers* increases by an amount equal to 0.25 times the base-year income. It will be assumed that the average number of beneficiary farmers per *aldea* is six. This case uses all of the same parameter values specified for Scenario 3b, except that the sqrt(3) divisor in the standard deviations is replaced by sqrt(6), and the time-2 standard deviation is 1.25 x 10,857 = 13,571 (yielding standard errors of 4,432, 5,540, 4,432 and 4, 432), and the income at time 2 for beneficiary farmers is 1.25 x 4,617 = 5,771 (so that D = 1,154).

Estimate of Aldea Sample Size, FTDA Evaluation

Scenario 3c: Power = .9; Sample designed to estimate an overall mean double difference for *program beneficiary farmers*; number of comparison-group *aldeas* = .8 times number of treatment-group *aldeas*.

Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment *Aldeas*, Time 1	Number of Comparison *Aldeas*, Time 1	Number of Treatment *Aldeas*, Time 2	Number of Comparison *Aldeas*, Time 2	Total *Aldea* Sample Size (both waves of panel survey)
0.25	169	135	169	135	**608**

The required sample size for beneficiaries is substantially larger than the required sample size for lead farmers since we are seeking to detect a much smaller difference (i.e., .25 times current income vs. .5 times current income).

Note that Scenarios 3b and 3d refer to making different estimates from the same sample *aldeas*. Hence the sample size that would satisfy both requirements (for the power of tests involving lead farmers and the power of tests involving beneficiary farmers) is the maximum of the two sample sizes (i.e., the sample sizes for Scenario 3c).

Scenario 3d: This case will estimate the sample size required to achieve a power of .9 for the test of the hypothesis that the mean double difference of income *of program farmers (leads plus beneficiaries)* increases by an amount equal to 0.33 times the base-year income (.33 is the weighted average, 1/3 x .5 + 2/3 x .25). It will be assumed that the average number of program farmers per *aldea* is nine (i.e., three lead farmers plus six beneficiary farmers). This case uses all of the same parameter values specified for Scenario 3b, except that the sqrt(3) divisor in the standard deviations is replaced by sqrt(9), and the time-2 standard deviation is 1.33 x 10,857 = 14,476 (yielding standard errors of 3,619, 4,825, 3,619 and 3,619), and the income at time 2 for program farmers is 1.33 x 4,617 = 6,156 (so that D = 1,539).

Estimate of Aldea Sample Size, FTDA Evaluation

Scenario 3d: Power = .9; Sample designed to estimate an overall mean double difference for *program farmers (leads & beneficiaries)*; number of comparison-group *aldeas* = .8 times number of treatment-group *aldeas*.

Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment *Aldeas*, Time 1	Number of Comparison *Aldeas*, Time 1	Number of Treatment *Aldeas*, Time 2	Number of Comparison *Aldeas*, Time 2	Total *Aldea* Sample Size (both waves of panel survey)
0.33	68	54	68	54	**244**

Comparing the third table to the first two, we see that the sample size required to achieve the specified level of power is substantially less for tests about *all program farmers*, than for lead program farmers and beneficiary program farmers separately. This is because the within-*aldea* sample sizes for leads and beneficiaries (three and six, respectively) are substantially less than the sample size for all program farmers combined (leads plus beneficiaries, or nine).

Scenario 3e: This case will estimate the sample size required to achieve a power of .9 for the test of the hypothesis that the mean double difference of income *of program farmers (leads plus beneficiaries)* increases by an amount equal to 0.25 times the base-year income. This scenario is identical to Scenario 3d, except for changing the difference to be detected from .33 to .25. This case uses all of the same parameter values specified for Scenario 3b, except that the sqrt(3) divisor in the standard deviations is replaced by sqrt(9), and the time-2 standard deviation is 1.25 x 10,857 = 13,571 (yielding standard errors of 3,619, 4,524, 3,619 and 3,619), and the income at time 2 for program farmers is 1.25 x 4,617 = 5,771 (so that D = 1,154).

Estimate of Aldea Sample Size, FTDA Evaluation

Scenario 3e: Power = .9; Sample designed to estimate an overall mean double difference for *program farmers (leads & beneficiaries)*; number of comparison-group *aldeas* = .8 times number of treatment-group *aldeas*.

Size of Double Difference to Detect, as a Fraction of the Mean	Number of Treatment *Aldeas*, Time 1	Number of Comparison *Aldeas*, Time 1	Number of Treatment *Aldeas*, Time 2	Number of Comparison *Aldeas*, Time 2	Total *Aldea* Sample Size (both waves of panel survey)
0.25	113	90	113	90	**406**

When the decision was made to modify the original experimental design to a revised design by supplementing the responding sample with a sample of approximately 600 (actually 545) Fintrac-selected clients, additional consideration was given to the power of the impact estimates. These considerations are documented in the memorandum, *Alternative Design for FTDA/EDA Program Evaluation*, dated 12 May 2010. The number 600 was taken because it would (when combined with those from the experimental design) increase the number of program (treatment) farmers to somewhat more than the number proposed for the experimental design (a larger number was desired because the new design would require the use of covariate-adjusted regression models, and it was considered that the sample size should be somewhat larger for such models than for the original experimental design).

ANNEX 4: DOCUMENTATION OF COMMENTS FROM PARTICIPANTS AT THE PRESENTATION OF INDEPENDENT EVALUATION RESULTS OF THE FARMER TRAINING AND DEVELOPMENT ASSISTANCE ACTIVITY, MAY 2, 2013, AND NORC RESPONSES

TECHNICAL AND FACTUAL ACCURACY AND GAPS

1. MCC expressed a lack of confidence in the findings of the FTDA evaluation, claiming that the results are not robust. While MCC did not present an explanation of all the reasons underlying this claim, in the discussion following this comment, Jack Molyneaux agreed that this view was based on the fact that the five estimators used and presented in an annex to the report did not produce similar results. In order to be able to respond fully to MCC's assertion, NORC would like to receive any and all additional reasons for MCC's claim that the evaluation results are not robust.

   NORC Response:

   We did not receive further clarification on MCC's specific concerns about questionable validity of the evaluation findings. As such, we have taken the following measures to bolster the technical strength of the evaluation report:

   (a) Refocused the report to emphasize a single preferred estimate of impact, rather than presenting detailed results for five alternative estimators. We mention in the main text that several different estimators were examined in the course of the analysis, we only present results in the main text for one single estimator. Detailed results are included in a technical annex (Annex 1) for three closely related estimators, but not for the remaining two estimators that were presented in previous versions of this report, because they did not show statistically significant results. We discuss that while the various estimators differ in some respects, they are not contradictory or inconsistent. We note that the lack of statistically significant results (failure to reject the null hypothesis of zero impact) does not imply that the impact results produced by the model are inconsistent with the statistically significant results produced by the preferred model. In statistical theory, a lack of evidence to reject a null hypothesis is not equivalent to acceptance of an alternative hypothesis.

   (b) Addressed the issue of the results' robustness. We believe the results are "robust" in the sense that they are based on sound econometric methodology / causal modeling (based on the Neyman-Fisher-Cox-Rubin causal model (potential outcomes model, counterfactuals model)), statistical modeling and estimation relevant to this causal modeling approach, and a good sample design for constructing this type of (causal) model and (causal) estimates, and adequate statistical power. Much more is included in the report on the sample design and the statistical power analysis that was used to determine sample size. We also include a discussion on economic interpretation of estimated models and estimators of impact.

(c) *Included an analysis of the relationship of impact to explanatory variables. Sample sizes were not adequate to construct estimates by occupational category, but a detailed analysis is included on estimating the relationship of outcomes of interest to the estimated propensity score, which is highly correlated with explanatory variables related to selection for treatment. We do not present estimates of the relationship of impact to individual explanatory variables since experimental control was exercised on individual explanatory variables.*

(d) *Conducted and included ex post power analysis. The sample sizes for the original experimental design and the revised design were based on "ex ante" statistical power analysis. We discussed this at length in the Inception Report, but not in the prior version of the Final Report. We want to make clear that the sample sizes were determined by a detailed statistical power analysis, that the power of the estimators is quite sufficient to detect impacts of anticipated magnitudes, and that the lack of substantial results is attributable to other factors (such as weak impact or insufficient time to see longer-term effects), not to an underpowered design. Toward this end, we conducted an ex post (post hoc) statistical power analysis to estimate the power for detecting impacts of specified size, using the model developed from the analysis rather that the rough assumptions made prior to conducting the surveys. This analysis is included in Annex 1.*

(e) *Reinserted the tables that we originally included in the annex on the unadjusted ("raw") double difference and the Observed Treatment Effect (OTE).*

(f) *Included more discussion about the selection model to address the peer reviewers' concerns, including discussion of model specification, effect identification, and estimation. More discussion is included about the variables affecting selection for treatment, and the assumptions required for conditional independence (of treatment and counterfactual outcomes). We also included discussion of economic interpretation in the main text of the report (not just in Annex 1).*

(g) *Included a discussion about how to extend the impact estimate to the full target population. We clearly state that the population to which the impact estimate applies is the population of program-eligible farmers, and that the estimate of impact is the expected impact for a randomly selected program-eligible farmer.*

2. Fintrac pointed out the following inaccuracies in NORC's report and presentation:

- The farmer selection criteria were not, as described in NORC's report, Fintrac's criteria for selecting program farmers. These were criteria spelled out by MCC and MCA-H in Fintrac's contract; as such, Fintrac was contractually obliged to use them. Fintrac added some criteria along the way, but the majority were defined by MCA-H/MCC and included in Fintrac's contract.

  ***NORC Response:** We have specified in the report that these were criteria spelled out by MCC and MCA, which Fintrac was contractually obliged to adhere to.*

- The criteria did not change over time. Fintrac provided a streamlined set of criteria to NORC for the purpose of replicating Fintrac's selection process. Note that this was the extent of the discussion around criteria for selecting program participants.

NORC Response: *We have chosen not to address this issue, since the criteria provided to NORC by Fintrac did change and expand between Cohorts 1 and 2. The screening forms and PowerPoint presentations that were reviewed and approved by Fintrac's senior staff attest to this fact.*

3. Fintrac stated that NORC's report does not mention how the sample aldeas were selected

   NORC Response: *We have elaborated further on the sample frame construction and the selection of sample aldeas from that frame.*

4. A former MCA-H staff member stated that it would be useful to see the relationship of impact to explanatory variables, such as farm size.

   NORC Response: *We have included a detailed discussion on the relationship of impact to the estimated propensity score (estimated probability of participation in the program). Because experimental control was not exercised at the level of the individual explanatory variable, and because of a high degree of intercorrelation ("confounding") among explanatory variables, it is preferred not to highlight estimated relationships of impact to individual explanatory variables. The report contains a detailed description of the relationship of the estimated propensity score to a number of explanatory variables, including household size, education and farm size. For the propensity-score-based estimates, the models used for impact estimation contain single covariate, the estimated propensity score, which combines the influence of all other covariates (by assuring conditional independence of treatment and the counterfactual responses, given the propensity score).*

TONE OF THE REPORT

Fintrac objected to the tone of the report stating that the report lays all the blame for the problems with the evaluation at Fintrac's feet; MCC and NORC should take some of the blame.

NORC Response: *We have rewritten some of the text to be present a more objective picture.*

OTHER CONCERNS AND OBSERVATIONS

1. Former MCA-H staff brought up the fact that the report did not include discussion of "conditions at entry." They believe that those conditions influenced the results of the FTDA and, as such, need to be addressed in the report. NORC understands that "conditions at entry" is not synonymous with information about the structure and nature of the Fintrac program. We would appreciate it if MCA-H would please elaborate further on these entry conditions that affected the FTDA and, as such, should be addressed in the report.

   NORC Response: *Based on MCA-H's response below, we did not further address this point:*

   "As far as we understand the comment, we consider these "conditions at entry" are already addressed in the report, specifically in the baseline data analysis and in the comparison of treatment and control groups of farmers with respect to key variables that may affect selection or outcome." (E-mail communication from Marco Brogan on June 25, 2013)

2. Former MCA-H staff stated that there is a lot of evidence in the field showing that the impact of the FTDA project was very positive. For example, income gains were greater than expected; moreover, prior to the implementation of the FTDA, Honduras was importing most consumed products, but now it has started exporting some of the products it previously imported for consumption; and finally, the project helped build the basis to change farmers' attitudes towards agriculture. These impacts are not captured in the impact evaluation. [Note: Could MCA please elaborate on the statement made by Daisy Avila regarding imports and exports, and why they believe that such macro-level changes are due to the FTDA.]

   NORC Response: *Based on MCA-H's response below, we did not further address this point:*

   MCA-H recognizes there is a shared perception among government agencies and the public opinion in general that the FTDA had a great impact and changed farmer´s attitudes and behaviors towards agriculture. Also, MCA-H is aware there may be macro-level effects derived from the FTDA, but it was not within the scope of the impact evaluation design to measure those perceptions and effects. Nevertheless, MCA-H believes the estimated impact on annual crop incomes of participants is a remarkable hint of the accomplishments of FTDA that supports the favorable perception among stakeholders in Honduras. (E-mail communication from Marco Brogan on June 25, 2013)

3. Many participants stated that the evaluation should have consolidated quantitative data and qualitative approaches. Supplementing the impact evaluation with qualitative methods would have added value to the evaluation and should be considered in future evaluations.

   NORC Response: *We agree. This was not within the scope of work specified by MCA and MCC for the evaluation.*

4. Fintrac expressed an opinion that the impact evaluation team should have included an agricultural expert; NORC's team did not include this expertise.

   NORC Response: *An agricultural expert was not required under the RFP issued first by MCA-H and subsequently by MCC.*

5. Fintrac and a former MCA-H staff member concurred with NORC's observation that it is critical for impact evaluations to start before implementation begins. In the case of the FTDA evaluation, NORC's impact evaluation commenced 2 years after program implementation had begun.

   NORC Response: *We agree.*

6. Fintrac stated that NORC should have had someone from their team on the ground full-time, embedded at MCA-H.

   NORC Response: *This is not a typical requirement of MCC Impact Evaluation contracts. Given the peaks and lulls in workload for an impact evaluation, such an investment may not be efficient. However, we believe that this is a decision for MCC to make as it issues future impact evaluation contracts.*

7. A former MCA-H staff member stated that impacts are usually visible after longer periods of time than those allowed for in this evaluation. A Fintrac staff member elaborated on this point, stating that there are varying rates of absorption/uptake of recommended practices among Program Farmers. In the FTDA group there were high achievers, others that learned

with some help, and slow learners (those that needed a lot of extra support). The Fintrac staff member observed that to really work, this program should have spanned 6-8 years. Many farmers needed further experience and support to consolidate their learning. Even though farmer received 52 classes in a year, the “slow learners” need more support time to ensure that they continue (do not abandon) newly learned practices.

NORC Response: *We agree; however, this was not the nature of the program nor the evaluation.*

ANNEX 5: RESPONSE TO REVIEWER COMMENTS

The following tables summarize the reviewer's comments in the left-hand column and describe NORC's response in the right-hand column. The table does not include positive reviewer comments, which do not require a response.

Comments from Michael Carter	
Comment	**Response**
1. Factual accuracy	
Given the peculiar history of this evaluation, it is not surprising that the report is eager to convince its readers that results from a nonexperimental study can usefully inform about causal impacts. While I am fully convinced of this point (and indeed have published papers on the topic), I think the report would be better served if it acknowledged the maintained assumptions of the analysis that temper its claim to unbiased estimates, rather than simply insisting on the general validity of non-experimental work.	Additional description and discussion has been added on causal models, statistical model specification, parameter / effect identification, and estimation. Assumptions relating to estimation are now clearly identified. The soundness of the results are based on the accuracy of the assumptions.
2. Technical rigor	
My biggest concern about the paper is that while it seeks identification through a "model-based" approach, it really does not thoroughly address the 'model' that underlay assignment to treatment. While we are given appropriate detail on how FINTRAC implemented the program (e.g., we learn on page 16 that FINTRAC selected participants on "non-quantifiable" criteria, including "subjective assessment" of "farmers' motivation and ability to learn and grow"), we are presented with an ad hoc propensity score model in which program assignment is explained with the usual suspects: farmer location, assets, and wealth. If this had been the treatment process, the original randomized design presumably would not have collapsed! The report does itself a disservice by failing to seriously move from what it learned about assignment to treatment by FINTRAC and how it actually approaches the problem econometrically. This is not a trivial issue. On page 22 (first full paragraph), the report has a confusing if not completely misleading set of statements. Its propensity score modeling methods do rely on the notion that treatment assignment is based on observables, as opposed to assignment based systematically on unobservables (such as "farmer motivation and ability to grow and learn). The offending paragraph on page 22 says that "unobserved variables [are] not highly relevant	A detailed description of the causal model has been added, including detailed description of the selection process. The discussion of selection on unobservables has been clarified.

Comments from Michael Carter	
Comment	**Response**
since they are no doubt reflected in the explanatory variables in the regression model." However, the very fact that they are unobservable means that they cannot be included/reflected in the explanatory variables. Indeed, the only explanatory variable with which these unobserved variables are correlated is the treatment assignment dummy variable, which is precisely the identification problem we face when selection is based on unobservables!	
I am not making any new or novel observations in making this point. In comments on an earlier version, I suggested that James Heckman had suggested an approach to the problem of selection on unobservables many years ago. Besides winning a Nobel Prize for his efforts, he puts forward a method that merits use in the current report. I say not based on my predilections, but on what I learned from the report about the 'model' that generated treatment. There is more to learn from the data about this whole problem. On pages 68-69, we are appropriately given the predicted propensity scores for the treated and untreated populations. The paper claims that there is strong overlapping support between the two populations. I am used to seeing balance tests based on blocking the observations into quintiles of the estimated propensity scores. I do not see any of that kind of testing here. More to the point, the two distributions are quite odd and in fact seem consistent with the parts of the report that describe what FINTRAC actually did. It appears that almost no-one who had a high propensity score (based on the short list of observables) was denied the program). However, the long left tail of the PS distribution for the treated makes it appear that FINTRAC went fishing for farmers with low PS (poor, isolated) and pulled out the ones that they somehow thought were motivated and were ready to grow. I have never seen a set of PS distributions that look like the ones in this paper, and it really makes me want to think careful about the model that created treatment assignment. This is not a point that only randomization is valid, but rather that if we want to take a model-based approach, then we need to seriously think about what the treatment assignment model was, and then proceed with appropriate econometric methods.	The models and estimation methods presented in the report draw on the "statistical" models of Rosenbaum and Rubin and the "econometric" models of Heckman. Additional discussion of this has been added to the text. Tables of means of explanatory variables were constructed, by propensity-score quintile and treatment. The similarity of the means was substantially greater within quintile categories, but substantial differences remain.
I have been perhaps a bit harsh in this section, but I am trying to clearly communicate what I see to be the fundamental issue about what we infer from the report's results. Let's imagine that say 25% of the	With the expanded discussion of the selection model, and in particular on the issue of selection on unobservables, the need for consideration of Rosenbaum bounds is obviated.

Comments from Michael Carter	
Comment	**Response**
sample was selected based on unobservables. Can we put some bounds on the seriousness of the bias this could create. As I recall, Rubins (?) has done some recent work on putting biasedness bounds on estimates from non-experimental data based on some reasonable conjectures on key forces (e.g., how much more productive could a highly motivated farmer be than a 'standard' famer). I do not know if it is worth further effort at this stage, but I do think the statistical methods need to be better nuanced (and please fix page 22).	
Summary	
Prior to publication on the MCC web site, I think at least some of the issues raised in section 2 should be addressed. This need not be a lot more work, but at a minimum the paper cannot misrepresent the weight of its assumption that treatment assignment was based on observables. Later sections of the paper (in the appendix) appear more sanguine on this point. This needs tightening up in the sections of the paper that are likely to be most heavily read by specialists.	Additional description of model specification, parameter / effect identification and estimation has been added. Assumptions relating to identification and estimation are clearly spelled out.

Comments from Marcus Goldstein	
They do not seem to understand the inherent problems in their method and hence make claims that are not backed up by either theory or the data.	Additional description of the causal model and selection process has been added, and of the statistical models that were used to estimate impact.
They seem to have done a significant amount of work to understand the selection process. They do not adequately deal with the factors they have identified.	A detailed discussion of the selection process has been included, including discussion of selection on unobservables.
As it is currently written, this paper should not have an impact on policymakers. Conclusions are drawn based on what, in my opinion, are currently insufficient grounds.	Additional description of the causal model and associated statistical model have been added

Comments from MCC reviewers (unidentified), as presented in the document "MCC Comments on NORC FTDA Evaluation Report_2013.8.20.docx	
1. **Two concerns with the evaluation method presented**. This evaluation asserts that it uses valid econometric causal modeling, but there are two closely related weaknesses to the evaluation method presented. The first is a failure to formally state the theoretical model	A detailed description of the causal model and the selection process has been added, including consideration of unobserved variables affecting selection. Additional discussion has been included about factors and variables that affect selection, including identification of unobserved factors, which

<table>
<tr><th colspan="2">Comments from MCC reviewers (unidentified), as presented in the document “MCC Comments on NORC FTDA Evaluation Report_2013.8.20.docx</th></tr>
<tr><td>used to justify the proposed statistical specification. This formal modeling is an essential part of any econometric model. It is commonly used to distinguish <i>exogenous</i> variables – those that are plausibly unaffected by the farmers or the implementers decisions related to the evaluated program – from <i>endogenous</i> variables – especially those that are subject to the choices and behaviors that the program is intended to influence. The formal modeling is also essential to understand, and carefully specify, what behaviors are being modeled.
For this program, there are at least two distinct sets of actors whose decisions are important to the outcomes of interest: the implementers, as they both decide which farmers they will eventually work with and they also determine what kind of training and technical assistance they will provide; and the farmers themselves, who decide whether they will participate in the program, how much effort they are willing to devote to participation and how they will respond to the training and assistance they receive. With two distinct sets of actors with their own sets of incentives, resources and constraints and with their own limited abilities to affect the necessary investment and farming decisions, it is important to understand what can and cannot influence each groups’ decisions. Yet this potentially complex model is glossed over with an assertion that this is a valid causal model.

The second, closely related weakness is the evaluator’s inability to understand and reproduce the selection process that lies at the heart of their causal modeling. Despite multiple rounds of data collection aimed at reproducing the selection, and detailed follow-up data collection attempting to understand the reasons for rejection of farmers, there is remarkably little effort devoted to understanding the selection, particularly the relative roles of the evaluators and the farmers in this selection process. As a result, the selection regression provides correlations, with no clear sense that the evaluator understands the selection process, and no clear understanding of why the selection specification chosen contains the variables it does. Given the multitude of alternative specifications that could have been chosen, why do we rely on this specification? And how</td><td>are generally assumed to be time-invariant. Endogeneity of explanatory variables in the selection model has been addressed.</td></tr>
</table>

Comments from MCC reviewers (unidentified), as presented in the document “MCC Comments on NORC FTDA Evaluation Report_2013.8.20.docx	
robust are these results to alternative specifications? These questions are largely unanswered. The statistical challenge that this evaluation needs to overcome is that both the outcomes of interest and the selection into the program are influenced by unobserved factors. Consequently the selection and the outcomes may be correlated for reasons unrelated to the assumed causal effect of the program on farming outcomes. Typical examples of this measurement problem cite farmer motivation, but in this setting it could also include any number of fixed or time-varying characteristics encompassing the farmers’ knowledge and experience with improved techniques, access to and experience with capital investments, or simply the ease of communication and collaboration between the farmer and the implementer. Rather than seek to purge the selection of the effects of these unobservable factors – as an instrumental variables approach would – the evaluator simply inserts a predicted probability of selection. This prediction equation includes plausibly exogenous variables, but it also includes clearly endogenous variables (inputs purchased, area farmed, etc.) As a result, we have a poorly understood selection probability estimate that needs to eliminate the selection variable of its correlation with unobserved factors that influence the outcomes. Together, these two weaknesses yield estimates of impacts that do not instill confidence in this reader. It is not at all clear that the effect of unobserved factors has been adequately addressed. But more worrying is the concern that we do not really know what impact is being modeled. Do the measured correlations reflect actual impacts of the program? Or are we measuring the implementer’s effectiveness at finding farmers who would have raised their farm incomes over the period of the program anyway? Either of these is a plausible causal relationship, but without understanding how the selection process works and modeling it appropriately, we have no way to distinguish these two alternative relationships. Just as with modeling supply and demand curves to properly predict whether price increases will increase or decrease market quantities, econometrics	

Comments from MCC reviewers (unidentified), as presented in the document "MCC Comments on NORC FTDA Evaluation Report_2013.8.20.docx	
requires a theoretical model to identify the relevant equations, and appropriate methods to distinguish the behavioral responses of different actors.	
2. **Summary/data on reasons for farmer rejection (pg 3, 12, 14, etc)** : Prior to the May 2 workshop, MCC has requested on multiple occasions that NORC provide a summary of the reasons for FINTRAC's rejection of farmers in the two rounds of farmer screening. Given the NORC report asserts that Fintrac's higher than normal rejection rate of farmers compromised the experimental design, could NORC please provide data on the reasons provided for determining 'out of scope' farmers, particularly reasons grouped by (i) possible screening error and (ii) changes in originally defined selection criteria, as well as any other group necessary.	
3. **Variation in exposure to treatment for additional sample of FINTRAC recruits**: It is MCC's understanding that the 545 farmers in the survey that came from the supplemental sample were exposed to treatment (receiving the technical assistance) between 18-24 months; while the remaining farmers in the survey- the 185 program farmers from Cohort 2- were exposed to treatment for only 1 crop season (received only about 4 months of technical assistance). Is this correct? If so, does the analysis account for this variation in exposure to treatment? We did not see a discussion in the analysis/results section around this, so please clarify. a. Could NORC please clarify the average exposure to treatment?	Regressions were run including an indicator variable that specified whether a household was included in the sample from the original experimental design (Design=1) or in the Fintrac sample (Design=0). The coefficient on this variable (for net household income) is statistically significant, with incomes larger for the Fintrac sample. These regressions are included in the .log file. The propensity score selection model includes this effect.
4. **Testing increased employment on farms**. The indicator that was to be used for the third main hypothesis of the evaluation (page 2) is Labor Expenses (LabExpOC, LabExpBG), but results are not presented on this indicator. Can you please clarify/explain?	Labor expenses (LabExpOC and LabExpBG) are included in the results tables.
5. **Discuss external validity**: How can these results be generalized given geographic and socioeconomic characteristics of treatment/comparison farmers?	A comment has been added that the evaluation addressed half of the area treated by Fintrac, and that it is considered that this area is agriculturally similar to the other half.
6. **MOU (pg 13)**: Can NORC confirm the MOU was executed between Fintrac, MCA, NORC?	We do not have a fully executed copy of the MOU. Please check with MCA-H.

Comments from MCC reviewers (unidentified), as presented in the document “MCC Comments on NORC FTDA Evaluation Report_2013.8.20.docx	
We did not find any evidence of this and it should be confirmed before it is stated in the report.	
7. **Typo**: There appears to be one typo in the Table on page 5 - Is Variable Net household expenditures (NetHHInc) supposed to be Net household Income? Expenditures is already in the table, the concept of net is odd with expenditures and the variable id ends with Inc.	Corrected.