FINAL REPORT (REVISED)

Impact Evaluation of the Farmer Training and Development Activity in Honduras

Millennium Challenge Corporation Contract MCC-10-0133-CON-20 TO01

SEPTEMBER 25, 2012

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, 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 Program 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 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

and insecticides, to the negotiation with buyers and the marketing of their high-value horticultural crops. In addition to TA, 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 ▪ Input expenditures on crops 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 a Quasi-Experimental Design

The impact evaluation design for this activity changed radically over the course of the evaluation due to implementation problems. In its original conception, NORC and MCA planned to use a rigorous experimental design for the impact evaluation. Following a series of implementation problems, the final approach used was a quasi-experimental design that relied on a model-based approach to impact evaluation.

The Original Intent: An Experimental Design

The planned randomized control trial consisted of an analytic survey design in which members of 200 matched pairs of aldeas (villages in Honduras) 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, thereby 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.

Implementation Problems

NORC encountered significant problems in implementing the experimental design described above. The fundamental issue we confronted was that the implementing agency's eligibility criteria for participant selection had a subjective component to it and, furthermore, kept changing as the program evolved. As a result, it was not possible for NORC and MCA to replicate the Program Farmer selection process on the basis of the objective and fixed criteria provided by the implementer. Of the almost 1,000 farmers screened using this objective set of criteria, the implementer chose to provide technical assistance to only 28 farmers. With this high rejection rate of potential program farmers, even after two attempts to replicate the screening and two rounds of large-scale data collection, our experimental treatment group was inadequate for a RCT evaluation.

These implementation problems posed serious threats to the experimental design. First, due to the very small treatment sample, there was a loss of precision and power. Second, selection bias was a problem because the implementer had rejected large numbers of screened farmers for reasons that could not be quantified. And finally, it also left us with a control group that had been picked solely according to objective criteria, which was not comparable to the treatment farmers. The absence of an adequate treatment group and the inability to construct a valid and reliable counterfactual effectively eliminated the possibility of conducting an experimental design impact evaluation.

A Quasi-Experimental Design: A Model-Based Approach

Our resolution to the problems described above was to use a model-based 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 the implementing agency from the sample of experimental treatment aldeas and a sample of the implementer's clients, with no relation to the evaluation design, who were randomly selected from their program lists. Baseline data was collected from these additional samples, which served to increase the sample size to achieve a satisfactory level of precision and power.

A key features of the quasi-experimental design that was ultimately used for the evaluation is that impact estimates are based on a statistical model and, as such, the impact estimator is a regression coefficient, and not a simple double difference. Additionally, covariate adjustments in the model are used to handle selection bias that occurs because in a non-randomized, non-matched design, treatment and control groups are likely to have different distributions for non-treatment variables. These covariate adjustments were implemented in two ways: first, as a "two-step" model, in which the first step estimated the selection probability, and the second step estimated the outcome conditional on the selection probability; and second, as a one-step regression model that included explanatory variables that affected selection or outcome.

We estimated five models, using the full set of available data. The impact estimators from these models fell into two groups: estimators based on selection models or propensity score-based estimators, and estimators based on outcome alone. Based on the outputs of these models and a close examination of the distribution of the estimated probability of participation in the program (estimated propensity scores) for treated and untreated households, which were widely different, we decided that the **modified regression-adjusted propensity-score-based estimator** was the best option for the FTDA evaluation.

The Results

The table below presents impact measures for the **modified regression-adjusted propensity-score-based estimator**[1].

Impact Estimates Using Modified Regression-Adjusted Propensity-Score-Based Estimator

Outcome Measure	Impact Estimate (standard error in parentheses)
Basic Grains (BG)	
- Income (IncBG)	-120 (837)
- Expenditures (ExpBG)	837 (393)*
- Net income (NetBG)	-957 (750)
- Labor expenditure (LabExpBG)	351 (264)
Other (Horticultural) Crops (OC)	
- Income (IncOC)	16,773 (4,298)*
- Expenditures (ExpOC)	5,413 (1,078)*
- Net income (NetOC)	11,360 (4,175)*
- Labor expenditures (LabExpOC)	1,911 (742)*
- % farmers growing horticultural crops (Horticulture)	-.0397 (.0194)
Aggregate Household Income and Expenditure	
- Employment income (IncEmp)	149 (733)
- Total household expenditures (TotHHExp)	204 (496)
- Net household expenditures (NetHHInc)	18,926 (13,306)

*Note: * Indicates significance; Income and Expense Measured in Honduran Lempiras (USD1 = L18.9).*

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. As well, 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. The results show 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, we do not see a corresponding increase in net household income nor household expenditures/consumption, as we might have expected given the increase in income from horticultural crops.

[1] Note that these impact measures do not represent an absolute increase or decrease in an indicator. Rather, they are the difference in change of a measure between the pretest and posttest times, between the treatment and control groups. For example, if 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.

Also, surprisingly, 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 implies 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, a broader positive impact on household income and expenditures was not detected.

In our adaptation of the original evaluation design after it was compromised and the subsequent analysis, we took numerous and significant measures to salvage the evaluation, and implement it as rigorously as possible, given the conditions and problems encountered. These approaches are described in detail in Annex 2. This analysis yielded results that show positive impacts that are largely consistent with the FTDA program logic. Our level of confidence in the results, particularly from the perspective of attributing causality to the FTDA activity, is less than it would have been had we been able to implement the planned experimental evaluation design from start to finish. However, given the circumstances, we believe that this analysis provides the best estimate possible of the impact of the FTDA activity.

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 pave approximately 65 km of secondary roads[2].

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[3]. 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

[2] 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.

[3] 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.

and its implementation. This section discusses the original experimental design, as it was developed in 2007, as well adaptations to that design necessitated by problems that were encountered during the implementation of the evaluation. These implementation problems had a major effect on the analysis of the impact evaluation and are described in detail 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. A much lengthier and technically detailed discussion of the impact analysis and results can be found in Annex 2. 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 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[4], 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

[4] 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.

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 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 and depth of land, 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 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 FTDA EVALUATION DESIGN

The evaluation design for the FTDA underwent a significant change since the inception of the project in 2007. Due to significant deviations from plan that occurred during the implementation of the design, we were compelled to abandon the originally developed experimental design (a design-based approach), and adopt an alternative, model-based, approach[5]. The impact evaluation results presented in Section E of this report pertain to this model-based approach. This section of the report describes the original experimental design and the alternative approach used for this evaluation as well as the implementation problems that led to the change.

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.

[5] 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.)

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 (villages) – into two groups: those that receive technical assistance now (the treatment communities) and those that receive it approximately 18 months later (control communities). (Comparison groups may be referred to as such, or as "control" groups. The term "control" used more often in experimental designs, and "comparison" in observational studies.) 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[6].

Box 6: The Double-Difference Estimator

The standard approach to calculating double differences with respect to projects is based on the two situations faced by households or communities: those that have an intervention – technical assistance in this case – and those that do not. The first difference is the comparison of average values for the outcome variables in the communities without the treatment (the control group) and the same variables in the treatment communities. The second difference is between the pre-treatment and post-treatment situations. 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 estimate of the impact of the program.

Under this randomized "pipeline" approach, the measure of impact is the interaction effect of treatment and time, or the double-difference estimate (see Box 6):

$$\textit{Estimate of impact} = (Y_{T,t2} – Y_{T,t1}) - (Y_{C,t2} – Y_{C,t1}),$$

where,

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

[6] 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 UC 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.

Since the sample design would be 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 evaluation design called for program treatment to be varied *among sample communities* rather than *among farmers within sample communities*[7]. This type of design is sometimes referred to as a "cluster-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 allowed us to make causal inferences about the effect of the program intervention.

The implementation of this rigorous randomized control design had operational implications for the implementation of the FTDA by Fintrac. However, 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 (municipalities) that Fintrac would expand into during 2008 and 2009[8]; NORC reached agreement with Fintrac on the *aldeas*, as a manageable community that would be used as the primary sampling unit (PSU) of a two-stage sample design.
- 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 200 pairs of matched aldeas, using an appropriate survey design[9].

[7] 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.

[8] The target population for this evaluation was the Fintrac expansion area for the three years following the start of this evaluation project. This was not the entire country and, as such, 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 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.

[9] 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.

- Random assignment of one aldea from each matched pair to the treatment group and the other to the control group.
- 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 required that a similar farmer selection process be used in both treatment and control communities, NORC and MCA Honduras decided to rely on 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. The screening forms were reviewed and approved by Fintrac. All measures were taken to ensure that identification of farmers for the sample aldeas mirrored Fintrac's selection process[10].
- 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 potentials 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 do not receive Fintrac assistance until approximately 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 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 actual program or intervention, allowing the calculation of unbiased estimation of the

[10] Note that the randomized design did not require that potential lead farmers be selected in the same way in the treatment and control aldeas. In other words, 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 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 sampling, it would have been possible to obtain an unbiased estimate of program impact, no matter how the farmers were selected. However, we strongly 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.

counterfactual effect. 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 EVALUATION DESIGN

Despite multiple attempts to do so, NORC and MCA Honduras were unable to replicate Fintrac's selection procedures and criteria for identifying eligible 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 explain the challenges we faced, as well as steps taken to address the problems.

C.2.1 Implementation Challenges

The validity of the original experimental design rested on maintaining the treatment/control status of the sample aldeas (i.e., that a treatment aldea would remain a treatment aldea and a control aldea would remain a control aldea), and that Fintrac follow the same procedures for selecting treatment farmers in the evaluation as it normally did. From the point of view of precision, it was desirable also that the same selection criteria be applied to farmers within treatment and control aldeas. Toward this end, NORC's preference from the outset of the evaluation was to engage Fintrac in the selection of eligible farmers within the treatment and control aldeas, thereby ensuring the replication of the Fintrac selection process in the evaluation sample. However, Fintrac expressed a strong reluctance to directly identify program farmers in control communities, since they were concerned about raising expectations in communities which they would not enter until 18 months later. Given this situation, in order to ensure an identical selection process in both treatment and control aldeas, NORC and MCA Honduras decided to rely on independent "screeners" to identify potential program farmers in all sample aldeas. Our intent was to replicate as closely as possible the procedures and criteria used by Fintrac in both treatment and control communities. Note that this approach was intended to increase *precision*, by assuring that the stratification of farmers within treatment and control aldeas would be the same. It had nothing to do with maintaining the *validity* (relating to bias) of the experimental design, which rested on maintaining the treatment/control status of the aldeas and on Fintrac's employing the same (aldea and farmer) selection procedures for the evaluation sample as for its normal program operations.

The initial screening, took place in February 2008, by a group of independent consultants (agronomists) hired by MCA Honduras, and trained by Fintrac representatives. The following steps were taken to ensure that the screening of farmers by independent (non-Fintrac) consultants would follow the Fintrac screening process to the letter:

- NORC developed the screening form in close cooperation and with input from senior Fintrac officials who reviewed several drafts of the form and provided in-person and written feedback.
- Fintrac's representative in Honduras and a senior Fintrac Field Technician led the training of supervisors (agronomists), who were charged with training the screeners. Screeners then accompanied the Field Technicians to observe the screening process. NORC, MCA Honduras, and MCC staff participated in the training of screeners.

The screening form contained the four eligibility criteria that Fintrac used to identify potential Program Farmers:

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

If the farmer did not have access to water, had land less that 50 hectares, had no interest in the program, or was already receiving technical assistance, he/she was ineligible to participate in the FTDA project. Only respondents/farmers who were eligible according to the four criteria would be listed as potential FTDA Program Farmers. The screening process yielded a list of 936 potential program farmers in 203 aldeas – 597 of these farmers were from treatment communities (Cohort 1). NORC provided the list of eligible farmers to Instituto Nacional de Estadisticas (INE) for baseline data collection and to Fintrac for one final test – "the three-visit test" - of eligibility among farmers treatment aldeas. This final test consisted of Fintrac Field Technicians returning to the potential program farmers and asking them to perform a series of three simple tasks to determine whether they would be suitable to receive technical assistance. The tasks were designed to determine the farmer's motivation and innovation. According to Fintrac, in the past, the rejection rate of farmers, who had been deemed eligible according to the four eligibility criteria, following the three-visit test was 8-10 percent.

Despite all measures taken to ensure that Fintrac's selection process was replicated in the study aldeas, upon visiting the screened farmers, in September 2008, we learned that Fintrac Field Technicians had visited 404 of the 597 farmers in treatment aldeas and rejected 341. Only 63 farmers were deemed eligible to participate in the "three-visit test[11]." In other words, according to Fintrac's assessment, 85 percent of these farmers should not have been screened as eligible. It was evident that there was a major problem with the screening undertaken by the independent consultants.

However, upon closer examination of the reasons listed for categorizing the 341 farmers as ineligible (provided to us in a Microsoft Excel spreadsheet by Fintrac), we found that many criteria used by Fintrac to exclude them had not been included on the original screening form or mentioned in the training. The new eligibility criteria included access to capital/economic resources, access to roads, flood zone, and ability to cultivate certain crop types. While some of the farmers had been incorrectly included on the list by screeners, most of them had been rejected by Fintrac staff for reasons that were never discussed as key eligibility criteria when

[11] Over the course of the next 12 months, the accepted sample of 59 farmers further diminished to 2, as Fintrac rejected more farmers because they failed to plant according to Fintrac technical assistance.

screening criteria were developed and shared with consultant agronomists in February 2008. Furthermore, Fintrac had rejected whole aldeas, the existence of an aldea-level screening in the farmer selection process had not been discussed previously.

This (rejection of entire aldeas) was the key event that effectively destroyed the original randomized experimental design. The ("cluster randomized") design was based on a randomized allocation of Fintrac-designated eligible aldeas to treatment and control. Program farmers could be selected by any means Fintrac desired within the sample aldeas. Unbiased estimation of impact was enabled in this design because *all* farmers in sample aldeas were subject to sampling (and selection for treatment) – a "certainty" stratum of program farmers and a random sample of 20 others (from a "noncertainty" stratum). (That is, every sample aldea was stratified into two strata – a certainty stratum of farmers selected for the program and a non-certainty stratum of all others.) When Fintrac chose to exclude aldeas that they had previously deemed eligible, the original randomized experimental design was no longer intact.

NORC and MCA Honduras responded immediately to the problem at hand by meeting again with Fintrac, discussing the new selection criteria, and selecting a second group of treatment and control aldeas to which the expanded list of eligibility criteria would be applied. The 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 survey firm, ESA Consultores, for this task. Once again, Fintrac led the training of screeners, and screeners accompanied Field Technicians into the field for on-site training. In this round, screening took place both at the aldea and farmer level. Separate forms were developed, with input and review from Fintrac, to include the vastly expanded criteria for the two levels of screening.

For the Cohort 2 screening, which took place February-March 2009, ESA screeners visited 251 sample aldeas of which 169 were deemed eligible according to the new aldea eligibility criteria. Seventy-two aldeas were rejected for the following reasons: less than 10 farmers; over two hours travel time to a paved road; lack of water for irrigation; not accessible during the entire year; farmers were dedicated to other activities – fisheries and livestock – and not interested in the FTDA; land with slope of over 47 degrees; and situated within protected forest lands.

The screeners visited and gathered information on 910 farmers in 85 treatment and 84 control aldeas. Of these, 658 farmers met all of the eligibility criteria; 343 of these farmers were located in the 85 treatment aldeas, while the remaining 315 came from control aldeas.

The expanded list of eligibility criteria used for screening Cohort 2 included the following:

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 the recommendations of the FTDA technician
4. Either possesses or can secure 70,000 lempiras per hectare to invest in crops
5. Not currently receiving technical agricultural assistance from another institution.

Following the screening of Cohort 2 aldeas and farmers, NORC provided the list of potential program farmers from the treatment aldeas (343 in all) to Fintrac for inclusion in the FTDA. INE proceeded with baseline data collection in the 169 sample aldeas.

In this second round (Cohort 2), in which the screening was very rigorous and based on an expanded and well-defined list of criteria provided by Fintrac, Fintrac initially rejected 256 farmers and accepted 87. This 75 percent rejection rate increased dramatically over subsequent months due to additional attrition, as farmers failed to follow Fintrac guidance in planting, or failed to show sufficient motivation.

Over two rounds of sample selection, NORC used selection criteria and processes employed by Fintrac to select over 900 potential program farmers in treatment and control aldeas. However, Fintrac rejected as "ineligible over 95 percent of these farmers screened and selected for the evaluation." After two separate screening efforts, and two rounds of full-blown data collection, which yielded less than 28 treatment farmers, NORC concluded that Fintrac's selection process could not be replicated, for two reasons: (1) it contains elements/criteria that cannot be quantified and depend on a 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 kept changing and evolving over time, based on lessons learned during implementation. As a result, despite our best efforts, we were unable identify aldeas and farmers using objective criteria that were acceptable to Fintrac. This problem led to very high rejection rates of aldeas and farmers by Fintrac technicians, and resulted in a very small sample of treatment aldeas and farmers for the evaluation. The arbitrary elimination of part of the aldea sample introduced selection bias into the impact estimates. The substantially reduced treatment sample size (aldeas and farmers) reduced the power of the sample to detect program effects. For both reasons (substantial selection of the sample and reduction of the sample size), the sample remaining from the original experimental design was not suitable for impact evaluation[12].

In an attempt to increase the sample size to a level that would support statistical analysis (even if not for a randomized experimental design), MCA Honduras and MCC prevailed upon Fintrac to return to the Cohort 2 treatment aldeas and recruit additional program farmers. Fintrac returned to the Cohort 2 aldeas in early 2010. The new recruitment yielded more farmers in the Cohort 2 aldeas. Following the supplemental recruitment in Cohort 2 aldeas, Fintrac submitted lists of new recruits to MCA and NORC. Interestingly, some of these farmers had been recruited into the

[12] The fact that the potential-program-farmer sample in the control group, which had been picked solely according to objective criteria, was clearly not comparable to the treatment sample, affected the precision of the design, but did not introduce bias or affect the validity of the design.

FTDA as early as June 2009. Others were recruited in January-March 2010. Baseline data collection for these new recruits occurred in April 2010. All respondents were asked to provide information on income and crops for the 12-month period before they entered the FTDA program. Therefore, the recall period for those who entered the program between June and December 2009 was far greater than for those who had entered in 2010, introducing concerns of recall error.

C.2.2 Implications of Implementation Challenges and Midstream Course Corrections for the Experimental Design

At the close of the MCA Honduras compact and the end of the FTDA intervention, Fintrac provided NORC with a final list of program farmers in Cohort 2 treatment aldeas. The final treatment sample consisted of 210 program farmers in 49 Cohort 2 aldeas for whom baseline data were available. INE collected follow-on data for these farmers, as well as for other rejected potential program farmers (from Cohort 2 list) in the 49 aldeas. Data were also collected for a probability sample of non-program farmer households in each of these aldeas.

This final sample of 210 program farmers in 49 aldeas was inadequate to form the basis for a sound program evaluation, for two reasons: the first concerned with precision and power, and the second concerned bias. With respect to precision and power, the reduced sample size was considered too small to have a high probability (power) of detecting a program effect of a size reasonably expected of the program, even if no bias were present. Power calculations showed that this sample size may be sufficient to detect a doubling of income among program farmers. With respect to bias, the concern was that subjective rejection of a large portion of the original Cohort 2 sample by Fintrac may introduce substantial bias into estimates of program impact, so that the validity of tests of hypothesis would be suspect. Small sample sizes from an ill-defined population (i.e., the "left-overs" resulting from massive nonresponse) are not a sound basis for a rigorous impact evaluation.

Below we discuss key concerns related to power, precision, and bias, as they pertained to the end-of-compact state of the experimental evaluation design.

Precision and Power. In order to apply statistical theory properly, it is necessary to have adequate sample sizes. Adequate sample sizes assure a high level of precision for sample estimates, and, in the absence of bias, a high level of power for tests of hypotheses[13]. The original sample sizes for the randomized control trial approach were determined in 2008 by means of a "statistical power analysis," which determined the size of samples required to enable the detection of program impact of a certain size with high probability (power).

The original sample sizes were estimated to be capable of detecting an increase in income (measured by a double-difference estimator) of 25 percent, with a probability (power) of 90 percent. The calculations were made assuming that there would be about three lead farmers and six beneficiary farmers per treatment aldea, and a comparable number of potential program

[13] Power is generally of greater interest than precision in evaluation studies, since the primary objective of an evaluation study is to detect a program impact of a certain size.

farmers per control aldea[14]. In addition to the program farmers, NORC recommended samples of 20 other households per aldea (so that all farmers within eligible aldeas would be subject to sampling, in order to enable unbiased estimation of impact, or average treatment effect). These original power calculations yielded the following sample sizes in 113 treatment and 90 control aldeas:

113 x 9 = 1,017 program farmers in treatment aldeas
113 x 20 = 2,260 other farmers or villages residents in treatment aldeas (probability sample)
90 x 9 = 810 potential program farmers in control aldeas
90 x 20 = 1,800 other farmers or village residents in control aldeas (probability sample)

Because of the high rejection rate of potential program farmers by Fintrac, the resultant sample sizes were much smaller than those recommended on the basis of the statistical power analysis. With these reduced sample sizes, the power of the sample to detect program impacts of the size originally specified is substantially reduced. Alternatively, for an effect size to be detected with 90 percent power (probability), the sample sizes must be much larger. NORC performed additional statistical power calculations and estimated that a sample of 30 treatment and 30 control aldeas would be the minimum number required to detect a *doubling* of income among program farmers, assuming no bias from nonresponse.

On the surface, it would appear that an available sample of 49 aldeas and 210 program farmers might suffice to detect a sizable program-caused change of income, such as an increase of 100 percent. Because of the nature of this sample, however, this is not the case. Sample size alone is not the sole consideration, in determining the quality of estimates and tests of hypothesis. Just as important as sample size is *how* the sample is selected.

Bias. In addition to consideration of *precision*, the second reason why the final sample is too low to form the basis for a sound program evaluation is *bias*. If a substantial portion of the sample is rejected, it is possible that a bias may be introduced into the sample estimates, and the integrity of the tests of hypothesis compromised. This is referred to as a non-response bias or a selection bias. Bias would not result from non-response if aldeas were rejected at random, or if they were determined to be out of scope; however, in the present application, it was not possible to quantify (using objective and measurable quantitative selection criteria) why Fintrac rejected aldeas. It is important to note here that the NORC survey experts met with individual Fintrac Field Technicians to conduct a case-by-case review of all potential program farmers that were rejected from the original Cohort 2 screening lists in an attempt to identify clearly defined reasons that might classify rejected farmers as out-of-scope. However, reasons provided by Field Technicians for rejecting potential program farmers varied widely across technicians, were subjective, and highlighted the fact that the criteria changed as the program evolved and workloads increased. Annex 1 presents a breakdown of reasons for high rejection rates gathered during these discussions.

[14] Note that during the course of the evaluation, the distinction between lead and beneficiary farmers became “blurred,” with both lead farmers and beneficiary farmers being viewed as “program farmers.

The problem of selection bias cannot be resolved simply by increasing the sample size – it is the absence of a portion of the sample that causes it. The problem cannot be solved simply by adding replacements to the sample (to bring the sample size back up to the intended size). The issue is that the rejected sample units (aldeas) may differ in some respects from the accepted ones, with respect to variables that may affect outcomes of interest. Therefore, simply replacing the rejected units with acceptable ones, without understanding and quantifying the reasons for their rejection, is not a satisfactory solution.

Because of the possibility of selection bias, design-based impact estimates derived from the sample data, even if they were found to be of adequate precision (despite the small responding sample size), would be of limited interest – it would be possible for the selection bias to be high, and it would not be evident what target population we were making inferences about. The presence of bias corrupts the tests of hypothesis, so that erroneous conclusions may result. (Note that these remarks refer to *design-based* estimates, not to *model-based* estimates.)

In conclusion, a high level of power (high precision and low bias) in an evaluation design requires that the sample size be adequate and that the rejection (non-response) rate be very low. By the end of the MCA Honduras compact, neither of these conditions held true for the experimental design originally envisioned for the evaluation of the FTDA. In light of these issues and events, at MCC’s request, NORC proposed a modified design (moving from a “design-based” approach to a “model-based” approach) that would allow us to complete a satisfactory evaluation despite the challenges described above.

C.3 AN ALTERNATIVE DESIGN: MODEL-BASED EVALUATION 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 model-based approach that would make use of data that had been collected for the original design, and supplement it with additional sample data.

The additional sample data were collected from a random sample of 545 Fintrac clients who had entered the program around May/June 2009, which coincides with data collection for Cohort 2 farmers. These additional sample units were intended to increase the sample size such that it was adequate to achieve a satisfactory level of precision and power. Since this sample was not determined by randomized allocation to treatment, design-based estimates based on it may have selection bias. The magnitude of the selection bias can 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. This approach maintains a considerable amount of the structure of the original experimental design (since all of the data collected under the original design were retained).

This alternative approach is a “model-based” approach, as contrasted with the original “design-based” approach. With the design-based approach, unbiased estimates of program impact are determined by using the probabilities of selection of the sample units (e.g., in a Horvitz-Thompson estimate). With that (design-based) approach, the estimation formulas depend on the structure of the sample design, not on the relationship of outcome to explanatory variables. With the model-based approach, the estimate of program impact is based on a statistical model that describes the relationship of treatment outcome to explanatory variables[15]. Under these two approaches, the form of the impact estimate and the procedures for constructing it are quite different. For the design-based approach using a highly-structured experimental design, the impact estimate can be represented as a simple double difference in sample means of the four design samples (treatment before, treatment after, control before, and control after). For the model-based approach, the estimate is more complicated, and is usually constructed using multiple regression analysis[16].

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)):

Outcome measure = f(treatment indicator variable, other explanatory variables)

Under this approach, 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 influence of all of these variables has not been removed by randomization, it is necessary to adjust for them in the analytical model. The outcome variables of interest are referred to as response (dependent or explained) variables. Variables that have an effect on the response variables are referred to as explanatory (or independent) variables. The explanatory variables include design variables, treatment variables and non-treatment variables, called covariates. The average measure of program impact is obtained by determining a regression-equation (or other) model showing the relationship of program outcome to explanatory variables, and 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.

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 have collected for this study, it is possible to develop models of the relationship of program impact to explanatory variables and obtain 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. 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

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

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

of the model, and that the correlation among them is low. This is exactly what was done in the original sample design.

The double-difference impact-estimation formula presented in Section C1.2 is appropriate for the experimental design, but not for the modified design. In the experimental-design case, no covariate adjustment is necessary since, because of randomization the distributions of the treatment and control samples are the same for all variables other than treatment. For the model-based evaluation design, it is necessary to modify the formula to reflect the covariates. In this case, 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[17].

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

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
δ = impact (interaction effect of treatment and time)
u_t = model error term.

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 logistic-regression formulation (which represents the probability of participation as a logistic function of a linear form such as shown above).

In summary, the proposed alternative design was considered a valid approach that was feasible to implement for estimating the average treatment effect of the program. It made full use of the data already collected from the experimental design and the supplementary sample of Fintrac clients. It makes full use of the data on the large number of variables contained in the survey questionnaire.

The impact evaluation findings presented in Section E of this report are based on the model-based approach described above.

[17] 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).

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) was never used in this analysis, since Fintrac rejected all but two potential program farmers from this cohort. For the baseline survey, the local data collector, INE, under NORC guidance was obliged to recruit and train interviewers, field data collection, and conduct data processing for multiple separate data collection efforts. Hence, while the original study design required only pre- and post-intervention data collections, the problems described in Section C.2 above meant that NORC and INE had to return to the field to collect data at five different times (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 1. 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-	Qualitative questions on impact of program activities,

Table 1. Household Survey Elements	
Key Elements	**Item Description**
Program[18] elements	negative consequences, etc.

In the first phase of questionnaire development, NORC conducted a systematic review of existing questionnaires that gathered 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. Through an iterative process, the team carefully reviewed all items and instructions in the questionnaire, and tested and revised question wording, instructions, skip patterns, and response categories.

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.

As the evaluation design changed over the course of its implementation, numerous new rounds of baseline data collection (in addition to the first round collected among Cohort 1 aldeas and farmers in July 2008) were conducted among screened farmers in Cohort 2 aldeas, new Fintrac recruits in Cohort 2 aldeas, the supplemental sample of 545 program farmers from Fintrac's own lists, and a probability sample of other households in Cohort 2 aldeas. Each of these rounds was conducted in an effort to increase the size of a baseline sample that was constantly diminishing due to the rejection of treatment aldeas and farmers.

[18] 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.

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.

Finally, for the follow on (endline) survey instrument, we included a series of qualitative questions regarding technical assistance, whether from Fintrac or other sources. This series of items were asked of only program farmers.

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. As such, 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," 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)

During the week of May 25, 2008, INE trained 10 field staff to conduct a pilot test of the survey instrument. The pilot test occurred on May 28 and 29, following which INE, together with NORC, conducted a debriefing with the field staff to discuss the training and data collection experience. Based on feedback from the debriefing, a series of minor modifications were made to the survey instrument. Most of these modifications were changes or additions to interviewer instructions to make them clearer. Changes to the survey instrument were finalized on June 10, 2008.

It is important to note that the description above refers to the Cohort 1 data collection, which took place in July 2008. Each subsequent data collection consisted of interviewer training; however, given that the instrument largely remained the same, additional pilot-testing was not conducted in every baseline round.

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 were 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.[19]

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 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.

[19] 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.

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

Three 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 only inducted a handful of the potential program farmers identified into the FTDA. To try and shore up the treatment sample and salvage the 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 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 once again, Fintrac rejected most of the potential farmers. Eventually, Fintrac agreed to return 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 subset of 545 program farmers from Fintrac's own lists was conducted in three 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 during four week data collection periods. Based on the length of the survey instrument, we estimated that each interviewer should complete an average of 2.5 surveys per field day. INE organized the field teams and recruited requisite numbers of field staff for each round so as to complete the survey within a maximum 30 day field period. Field staff was organized into teams of 5 key staff: one supervisor, one editor and three field interviewers, together with one driver.

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 that the NORC field team witnessed 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, village, caserio 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 was 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.

E. PROGRAM IMPACT: SUMMARY OF RESULTS

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 the implementation phase (described above in Section C.2) meant that our approach, and analysis, switched from a straightforward pre-test post-test 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 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. Information about all of the estimators is included in Annex 2, and results are summarized for one of them in this section. (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.)

E.1 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 Fintrac's stated 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 (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 – these 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
Endline

Based on these various combinations of cohort, aldea type and farmer type, 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 producer 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 2. 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.

E.2 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.

Table 2 presents means and standard deviations for a selected set of these outcome variables for basic grains, other crops, and total household for baseline and follow-on rounds of the survey. Income and expenditure is presented in Honduran lempira. The current exchange rate is 18.9 lempiras to the dollar. More detailed tables with basic characteristics of distribution (means, standard deviations, and ranges) for all variables listed above are presented in Annex 2. Note that income and expense amounts for crops are annual, household incomes and expenses (IncEmp, TotHHExp) are monthly, and NetHHInc (which is based on both annual and monthly figures) is annualized.

Table 2. Means and Standard of Outcome Variables (Honduran Lempiras)				
	Baseline (N=4526)		Endline (N=2736)	
Indicator	Mean	Std. Dev	Mean	Std. Dev
Income, basic grains (IncBG)	8976.86	39483.47	9703.24	33967.62
Total expenses, basic grains (ExpBG)	4362.98	17053.47	5058.422	16016.66
Net income, basic grains (NetBG)	4613.87	29894.96	4644.819	26480.76
Income, other crops (IncOC)	24245.63	152281.1	34221.16	191685.90
Total expenses, other crops (ExpOC)	9111.02	59633.56	20544.11	239307.2
Net income, other crops (NetOC)	15134.61	135858.1	13677.05	282061.6
Labor market income (IncEmp)	6939.36	15994.66	9587.845	25095.99
Total hhold expenditures (TotHHExp)	5375.21	4921.943	7760.885	11043.65
Net household income ((NetHHInc)	113914	263066	143183	465696

E.3 IMPACT ESTIMATORS AND SUMMARY RESULTS

The standard measure of impact for socio-economic programs is the average treatment effect (ATE), which is defined as the expected difference in outcome caused by the program for a randomly selected household. Since randomization was applied at the aldea level, a “randomly selected household” means a program household in a randomly selected program-eligible aldea. This simple definition is formalized in Annex 2, based on a “counterfactuals” (or “potential outcomes”) model for estimating impact.

As Section C of this report describes in great detail, the FTDA evaluation design started out as a pretest-posttest-control-group design in which treatment was randomly assigned to aldeas. After difficulties were encountered in implementing this design, data from the original design were supplemented by additional sampling from Fintrac’s client pool and a model-based approach was utilized.

For a pretest-posttest-control-group design, the average treatment effect (ATE) is the difference, between the treatment and control populations, of the difference in means between the posttest and the pretest. This quantity is called the double-difference measure of impact. For a highly structured randomized experimental design, the *sample* double-difference estimator is a consistent estimate of the population double-difference *measure* (and hence of the ATE). For more complicated designs, such as the present one, the intention remains to estimate the *double-difference measure* (or ATE), but the *sample double difference estimate* (observed treatment effect, or OTE) is no longer a consistent estimate of the *population value*.

For the current impact evaluation, we estimated program impact using five impact estimators, listed below, to assess the impact of the FTDA on impact indicators of interest. Under certain assumptions, these five estimators provide consistent estimates of the double-difference measure (or ATE) for the project at hand. They reduce the bias caused by a lack of randomization in different ways, but all of them are based on regression models that take into the account of covariates on selection or on outcome.

Impact Estimators:

1. Basic propensity-score-based estimator of average treatment effect (ATE)
2. Regression-adjusted propensity-score-based estimator of ATE
3. Modified regression-adjusted propensity-score-based estimator for ATE
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.

More confidence was placed in the first three models, both because the models on which these estimators are based have greater face validity than the last two models (i.e., the structural representation of the selection process is a closer representation to a casual model) and because the standard errors of the impact estimates are substantially smaller. Results are presented here for the third estimator, the "modified regression-adjusted propensity-score-based estimator," which includes more explanatory variables than the first two. This estimator showed that the program had a positive effect on outcomes of interest. The last two estimators showed weak effects or no effects, or effects of unexpected sign.

Table 3 presents impact estimates for the **modified regression-adjusted propensity-score-based estimator**. Each estimate is followed by its estimated standard error in parentheses. Roughly speaking, an estimate is considered statistically significantly different from zero if it exceeds its standard error in magnitude by a factor of two. More precisely, an effect is considered statistically significantly different from zero if it differs from zero by more than 1.95 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). Annex 2 presents results for all five impact estimators.

The impact measures presented in Table 3 should not be interpreted as a level or change in level. They do not represent increases or decreases in an indicator. Rather, they are the difference in change of a measure between the pretest and posttest times, between the treatment and control groups. For example, if 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 table presented below.

Table 3. Impact Estimates Using Modified Regression-Adjusted Propensity-Score-Based Estimator

Outcome Measure	Impact Estimate (standard error in parentheses)
Basic Grains (BG)	
IncBG	-120 (837)
ExpBG	837 (393)*
NetBG	-957 (750)
LabExpBG	351 (264)

Other (Horticultural) Crops (OC)	
IncOC	16,773 (4,298)*
ExpOC	5,413 (1,078)*
NetOC	11,360 (4,175)*
LabExpOC	1,911 (742)*
Horticulture	-.0397 (.0194)
Aggregate Household Income and Expenditure	
IncEmp	149 (733)
TotHHExp	204 (496)
NetHHInc	18,926 (13,306)
Note: * Indicates significance; Income and Expense Measured in Honduran Lempiras (USD1 = L18.9).	

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.

The program does not appear to have had a significant impact on household expenditures or net household income. The estimated effects on these two indicators are positive, but not statistically significant.

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 implies 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.

Overall, the propensity score-based estimator provides 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).

In our adaptation of the original evaluation design after it was compromised and the subsequent analysis if the data, we took numerous and significant measures to salvage the evaluation, and implement it as rigorously as possible, given the conditions and problems encountered. These approaches are described in detail in Annex 2. This analysis yielded results that show positive impacts that are largely consistent with the FTDA program logic. Our level of confidence in the results, particularly from the perspective of attributing causality to the FTDA activity, is less than it would have been had we been able to implement the planned experimental evaluation design from start to finish. However, given the circumstances, we believe that this analysis provides the best estimate possible of the impact of the FTDA activity.

E.4. KEY OBSERVATIONS ON THE CHOICE OF IMPACT ESTIMATORS

The summary results presented in the preceding section are based on a complex estimator called the modified regression-adjusted propensity-score-based estimator. As mentioned in Section E.3, we started the analysis by examining a total of five estimators: the basic propensity-score-based estimator of average treatment effect (ATE); the regression-adjusted propensity-score-based estimator of ATE; the modified regression-adjusted propensity-score-based estimator for ATE; the regression estimator for ATE, not based on the estimated propensity score; and the instrumental-variable (IV) regression estimator for ATE, based on the estimated propensity score.

The results of these five estimators, which are presented in Annex 2, differ from one another. Furthermore, the impact estimates differ somewhat from the observed treatment effect (OTE, the “raw” double difference estimator, equal to the difference, between the treatment and control samples, of the difference in means before and after treatment). This section presents some descriptive analysis of the sample data, to better understand why the various estimators did not present a single, uniform set of results.

The main reason why the results varied by estimator is that there were substantial differences between the treated and untreated households, with respect to variables that had a signficant effect on outcome. These differences were accounted for (represented) better in the logistic-regression “selection” model that was central to the first three estimators than in the ordinary linear-regression “outcome” model that was central to the last two estimators. They were not accounted for at all in the “raw” double-difference estimator of impact, or OTE.

Data on a large number of variables were collected in the questionnaire, but only those with a strong relationship either to outcome or to selection were included in the models as explanatory variables. Because correlation exists among the explanatory variables, the representation of an outcome variable to explanatory variables is not unique. Alternative model versions may be obtained simply by replacing explanatory variables in a model by other variables with which they are correlated. For this reason, little substantive significance should be attributed to the statistical significance of any particular variable in a regression model. In statistical terminology, the variables are not orthogonal, and the data were not obtained by making forced changes in the explanatory variables. Another reason for not attributing significance to estimated coefficients in a model is Simpson’s Paradox: any statistical relationship between two variables may be reversed by including additional variables in the model.

To understand the reasons for differences in the various estimators (the five estimators listed above, plus the OTE), we present below a table that compares for the treatment and control populations selected variables that may affect selection or outcome. Table 4 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 variables be the same for the treatment and control populations (not just a single distribution attribute such as the mean). However, it is cumbersome to compare distributions, and therefore, what is often done is simply to compare the “supports” (ranges of observed values) of the distributions of the variables. If the supports are similar, then regression methods may be used to take into account differences in the shape of the distributions in obtaining valid (unbiased, consistent) estimates of impact. The supports are of interest in the analysis, but they will not be described here.

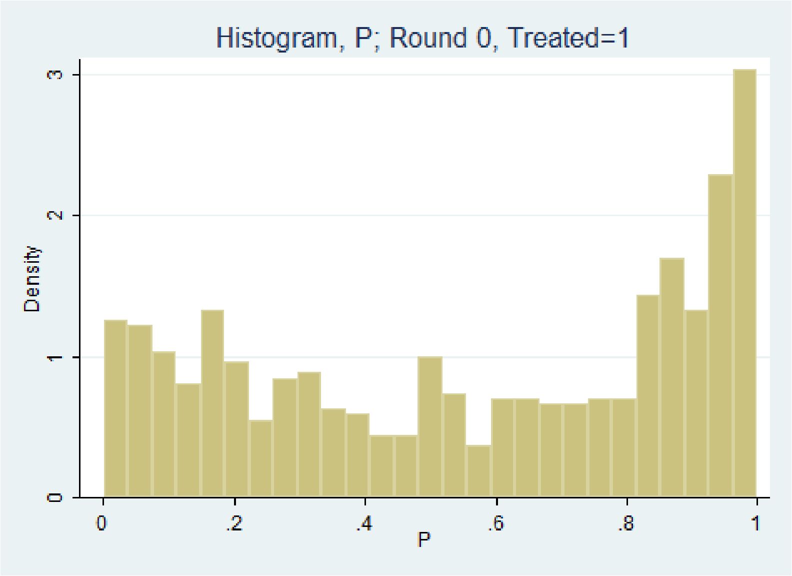
Table 4. Basic Characteristics of the Distribution of Key Variables for Treated and Untreated Populations

Indicator		Obs	Mean	Std. Dev	Min	Max
UNTREATED						
Household size		5881	4.94	2.32	1	17
Agricultural employees		5881	0.55	0.59	0	5
Total hectares of farm		5883	2.91	11.98	0	312.4
Mean education (years)		5881	3.69	2.29	0	20
Equipment value (lempiras, L)		3878	18994.56	83738.37	0	3007000
Rental value of installation (L/month)		3876	138.44	638.41	0	19052
Travel time to school (minutes)		5874	10.50	13.26	0	300
Travel time to hospital (minutes)		5811	117.52	70.08	0	1440
Total household expenditure (L/month)		5883	5590.89	5244.96	0	100000
Income - basic grains (L/month)		5883	6744.73	12902.03	0	96680
Labor expenditure-basic grains (L/month)		5883	1082.27	3112.90	0	31500
Income – other crops (L/month)		5883	12043.64	48312.26	0	498825
Labor expense - other crops (L/month)		5883	2085.27	8882.50	0	112500
TREATED						
Household size		1371	5.09	2.387	1	19
Agricultural employees		1371	0.98	0.81	0	7
Total hectares of farm		1376	4.35	33.24	0	1065
Mean education (years)		1371	5.07	2.89	0	20
Equipment value (lempiras, L)		1337	76917.11	408188.5	0	8505000
Rental value of installations (L/month)		1336	474.67	3796.33	0	120000
Travel time to school (minutes)		1368	12.73	13.35	1	180
Travel time to hospital (minutes)		1366	85.66	63.39	0	480
Total household expenditure (L/month)		1376	8862.33	8285.38	0	43446.31
Income - basic grains (L/month)		1376	12340.21	19612.45	0	96680
Labor expenditure-basic grains (L/month)		1376	1118.74	3007.22	0	288
Income – other crops (L/month)		1376	61785.98	118542	0	498825
Labor expense - other crops (L/month)		1376	3662.08	10923.31	0	101250

The preceding table shows the substantial differences that exist between the treatment and control samples, with respect to variables that were determined to be important in models of selection and outcome. (To obtain an approximate estimate of the standard error of a mean, divide the standard deviation by the square root of the sample size. If two means (i.e., the Treated=0 mean and the Treated=1 mean) differ by more than the sum of their standard errors, the difference may be regarded as "statistically significant.")

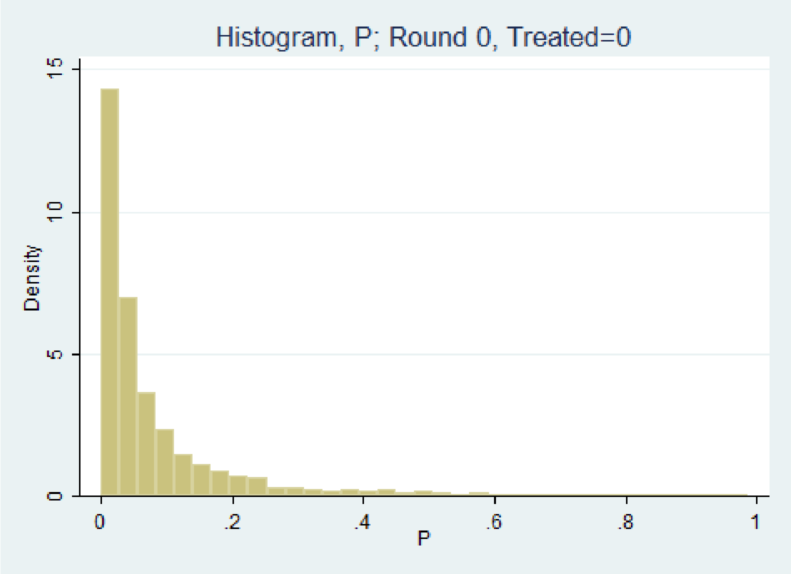
Another, simpler, way to compare the treatment and control samples is to compare the distributions of their estimated propensity scores (i.e., of the estimated probability of participation in the program). This comparison is much simpler than comparing the samples with respect to a large number of variables. 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 the same. The following two graphs show the distribution of estimated propensity scores for the treated and non-treated household samples[20].

Figure 1. Distribution of Estimated Propensity Score (P) for Households Treated by Fintrac, Full Data Set (Original Experimental Design and Additional Sample of "600" Fintrac Clients)



[20] 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 2. 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 distribution of estimated propensity score is quite different for the treatment and control samples.

It is necessary to take this fact into account in estimation of impact and, therefore, these impact estimates turn out to be quite different from the raw double-difference estimates (OTE). As it turned out, the two-step logistic-regression model / linear regression model (used for the first three estimators listed on page 26) was able to represent the selection process better than the last two estimators. The preceding graphs show that the selection probabilities are very different for the treatment and control samples, indicating that the selection process must be taken into account in impact estimation. The large difference in the distributions suggests that the two estimation approaches might lead to rather different estimates of impact, as was the case, but it does not provide detailed information to explain exactly why the estimates are as different as they turned out to be.

To understand why the last two estimators failed to show any positive program effects, it is helpful to examine the raw double-difference estimator, or Observed Treatment Effect (OTE). The following list presents the OTE for all of the outcome measures, taking into account the design feature that the same households are interviewed in both rounds, but no other explanatory variables:

For basic grains (BG):
IncBG: -2300, se = 672
ExpBG: 1201, se = 257
NetBG: -3502, se = 585
LabExpBG: 1310, se = 171

For other crops (OC):
IncOC: 4301, se = 3205

ExpOC: 6512 se = 738
NetOC: -2211, se = 2893
LabExpOC: 5873, se = 490

For labor-market employment and household income and expenditures:
IncEmp: 336, se = 517
TotHHExp: -560, se = 319
NetHHInc: 237, se = 7792

Production of horticultural crops:
Horticulture: -.0427, se = .0183

The salient feature of the preceding table is that, at first look, it appears that the program has no positive effects. Even worse, although the effects 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 (i.e., the first three estimators) are based on a strong logistic-regression model of selection, and they estimate positive results for the program. The other two estimators do not represent the selection process well, and they fail to show positive results. The instrumental-variable regression estimator (the last one of the five) fails to show statistically significant results for either NetOC or NetHHInc. Because these models are weak, they reflect the OTE.

For these reasons, we opted to focus on the three propensity-score-based models, and in particular on the "modified regression-adjusted propensity-score-based estimator," which includes more explanatory variables than other two propensity score-based models.

F. SUGGESTIONS FOR ADDITIONAL ANALYSIS

In the analysis presented in this report, we assessed the overall impact of the FTDA program as reflected in a variety of indicators (outcome variables) of interest, at the household level. The household questionnaire from which the household-level data were obtained contains considerable detail on within-household characteristics, such as data on individual family members, components of income and expense, and crop types. These detailed data are contained in the data files that record detailed question responses for each of the questionnaire modules, prior to aggregation for the analysis presented in this report. The detailed data from the questionnaires may be analyzed further to investigate issues related to program impact at a more detailed level than intended for this evaluation. Below, we list some examples of issues that may be further analyzed using the detailed (disaggregated) questionnaire data, as well as additional issues that may be addressed, beyond the basic project goal of assessing overall program impact.

Analysis of Changes in Crop Prices

The impact analysis presented in this report did not take into account the effects of changes in crop prices. This approach is reasonable when a double-difference measure of impact is used, under the assumption that the price changes are similar for treated and untreated sample units. This assumption holds under the assumption of conditional independence (i.e., that the response pair (y_0, y_1) is independent of treatment (w) given the covariates (**x**)), which is the general assumption on which the impact analysis was based. In the event that this assumption does not hold, it may be of interest to estimate impact taking into account changes in crop prices. As a first step in this analysis, price indices may be calculated for the total sample and for subsamples of interest (e.g., treatment and controls). Two popular price indices are the Laspeyres and Paasche indices. The Laspeyres index measures the effect of price changes holding quantities fixed at the baseline levels, and the Paasche index measures the effect of price changes holding quantities fixed at endline levels.

Amortization of Investment Costs

The FTDA program requires a considerable investment by the program participant on agricultural inputs such as fertilizer, herbicides, pesticides, hybrid varieties, and irrigation equipment. In the analysis, all of these costs – even for durable equipment such as pumps, poles, pipes and irrigation tapes – were recorded in the form of current expenses. This procedure may cause the program impact to be substantially reduced in the year in which these expenditures are made. To obtain a better estimate of the economic rate of return of the program, one could amortize the cost of equipment having a useful life greater than a year. The questionnaire was not designed so that this may be easily done, since it did not distinguish between program-related and non-program-related expenditures. Nevertheless, it may be possible to identify some equipment items (e.g., pumps) that are likely to be associated with the program, and to amortize the cost of these items. Impact would then be re-estimated, using the amortized expense in place of cost in the calculation of net income.

Cost-Benefit Analysis

The analysis presented in this report assesses the magnitude of the economic impact of the program in terms of income and expense, but it does not investigate the issue of whether the positive effects of the program justify the program cost. In order to determine whether to replicate the FTDA program, it is necessary to conduct a detailed cost-benefit analysis of the program. This analysis would assemble complete data on the cost of the FTDA program and use the results of this evaluation to estimate the net present value of the program, the economic (internal) rate of return, and the cost-benefit ratio.

ANNEX 1: REASONS FOR HIGH FINTRAC REJECTION RATES – A SUMMARY OF DISCUSSIONS WITH FINTRAC FIELD TECHNICIANS

In February 2010, following a second round of rejections by Fintrac of potential program farmers in Cohort 2 aldeas, NORC staff conducted interviews with Fintrac Field Technicians to understand why they were rejecting farmers that were eligible according to objective selection criteria. Below, we summarize observations from these interviews.

- Most of the technicians reported that the manner in which new farmers are identified has changed dramatically over the course of the project. In most cases (80 of 100 new farmers according to one technician), new farmers were not really selected, but “self-identified.” These self-selected farmers then called the technician about participating in the Fintrac program. Technicians report that these new recruits are both very motivated and more aware of the level of effort involved in implementing the program.Their awareness has led to fewer dropouts because they know the requirements upfront and have either visited demonstration plots or have witnessed their neighbor’s efforts. One technician reported that the quality of the program farmers has actually risen with the advent of self-selection and the number of those dropping out along the way has fallen.
- In some instances, technicians report that they are now also recruiting farmers who have already planted crops. If during their visits to aldeas, they notice a plot where a farmer has done a reasonable job planting, the technician will try and recruit him into the program and help him “clean it up” and introduce key components of the Fintrac program that are missing from his practice.
- Technicians reported that several of the aldeas identified by NORC would surely contain farmers suitable for the program and that they would have worked to try and recruit them into the program a year or two ago. But now with their large caseloads and tight schedules, they do not have the time to make visits and try to sell the program to prospective farmers in distant locations over a period of months.   According to several technicians, the trend towards self-identification has increased the quality of the program farmers and reduced the once heavy recruitment burden on technicians, enabling them to focus more on technical assistance.
- Since TAs work in clusters of aldeas and have a very full work schedule, it is difficult to add many of the NORC sampled aldeas because they are too far off their now established routes (1.5 to 2 hours) and they would require a significant number of new program farmers (20 or more) to make it worth their while to travel the long distance.
- The technicians reported that they have between 180 and 280 program farmers and that they are able to visit most of these groups of farmers (groups range from 8 to 20 farmers) only once every 15 days (instead of weekly as the program requires). Their high caseload has led technicians to create very organized group visit schedules that take them to as many as 5 groups in one day, leaving before 6 a.m. and not arriving back home until after 7 p.m. These

training groups are usually located across clusters of 2 and 3 villages that lie no more than 20 to 30 minutes distance from one another. One technician reported that the ideal work load would be 35 to 50 farmers; at the time of the interview Field Technicians already had about 4 times that number.

- When a program farmer "officially" enters the program continues to evolve, as we have seen during the course of our field visits. It appears that those technicians who have been working longest in the field have made the greatest adjustments to these criteria to mirror the reality that many farmers do not follow through with the many components of the program and end up dropping out. TAs with more experience now tend to view a producer as a “program farmer” only after he has either constructed the raised beds or even completed planting his first crop. Not surprisingly, the newer technicians tend to adhere more to the initial 3-visit test we heard about when the program began, but even they report making these "steps" more elaborate and rigorous as they struggle to identify the best point at which to convert a potential farmer into a program farmer and add them to their official client list. More and more, one seems to become a Program Farmer only when s/he has learned and adopted many elements of the program up until the planting stage.
- The introduction of the Para Tecnicos seems to have helped alleviate some of the burden. Several technicians have delegated as many as 100 farmers to the Para Tecnicos. Technicians report that since these individuals are recruited locally, they are the best opportunity to provide some sustainability to the program once the Technicians leave the area.
- Technicians reported in several areas that they have recruited quite a few individuals who do not farm and that this experience has been successful because these individuals do not have to “unlearn” bad practices. These individuals have other employment, but have approached the technicians with a strong interest and they have been engaged by them. Originally, having other full-time employment was a criterion for elimination.

ANNEX 2: ESTIMATION OF IMPACT: A DETAILED DESCRIPTION OF THE ANALYSIS AND RESULTS

I. Conceptual Framework for Estimation of Impact (Potential Outcomes; Counterfactuals)

I.A Measures of Impact for Pretest-Posttest Evaluation Designs

The conceptual framework adopted for the impact evaluation was an evaluation design that would support estimation of standard measures of impact. For the FTDA Project, our original intent was to use a double-difference measure of program impact, and the evaluation design corresponded to this plan (e.g., a binary treatment variable, and a design with before-and-after treatment and control groups).

A double-difference estimate is the difference, between the treatment sample and the control sample, of the difference in means of an outcome measure between the baseline and endline surveys. The standard evaluation design used to obtain data for constructing double-difference estimates of program impact is the pretest-posttest-randomized-control-group design, and this was the design proposed for the FTDA evaluation. Ordinarily, the design involves data collected at two points in time – the baseline time (“time 0”, “before”, “pretest”) and the endline (“time 1”, “after”, “posttest”, “follow-up”).

It is important to distinguish between the double-difference *measure* of impact and the double-difference *estimator*. The double-difference *measure* 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 sample means. That is, the double-difference measure is a population characteristic (parameter), and the double-difference estimator is a statistic based on a sample. We further describe these concepts below.

For a pretest-posttest-comparison-group design, the (unadjusted, “raw”) double-difference estimate is given by the following formula:

$$\mathrm{DD}_{\mathrm{raw}} = (\bar{y}_{t1} - \bar{y}_{t0}) - (\bar{y}_{c1} - \bar{y}_{c0})$$

where

$\mathrm{DD}_{\mathrm{raw}}$ = 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.

The preceding statistic is also called the observed treatment effect (OTE). In the preceding, the variable $\bar{y}_{ij}$ refers to any outcome variable of interest, such as income. The means (averages) referred to are "design-based" ("weighted") sample estimates that take into account the nature of the probability sampling used in the sample survey used to collect the data (e.g., stratification, multi-stage sampling, and selection with varying probabilities). For example, the means and their estimated variances may be estimated using Horvitz-Thompson estimation procedures.

For a pretest-posttest-randomized-control-group design, the double-difference estimator is an unbiased 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 estimators, such as regression estimators, must be used to obtain an unbiased estimate of the double difference measure.

For design-based estimates, the mathematical (statistical) model used as a basis for constructing estimates of interest and conducting tests of hypotheses of interest describes the sample design and sample selection procedures used to collect data for the evaluation design. Note that no causal model (economic model) is specified in the design-based approach. If a randomized experimental design is used (and if it is assumed that no variables that affect outcome change over time differently for the treatment and control samples), there is no need to specify a causal model to obtain causal estimates – causal estimates of impact may be estimated directly from the sample, using design-based estimates.

Although randomization was involved in the assignment of treatment for the FTDA Project, it is possible that the distribution of explanatory variables is different for the different groups (e.g., because of events that occurred between the baseline and endline surveys), in which case it is necessary to adjust the impact estimator to account for these differences. These estimates are referred to as "model-based" estimates. Because regression analysis is usually used to make the adjustment, the procedure is often called "regression adjustment." Often, the estimate of impact (i.e., of the double-difference measure) is the coefficient of a variable in a regression model. The estimate may be referred to as a "regression-adjusted" estimate or a "covariate-adjusted" estimate.

I.B The Counterfactuals Model

The standard conceptual framework for conducting a rigorous impact evaluation is called the "Neyman-Rubin causal model," or the "potential outcomes model," or the "counterfactuals model." This framework assumes that each experimental unit has two potential outcomes, depending on whether it is treated or untreated. Each of these two outcomes is called a counterfactual for the other. For each experimental unit, only one of these two outcomes may be observed, depending on whether the unit is treated[21]. For a randomized experimental design,

[21] The counterfactual approach to impact estimation is described in *Causality: Models, Reasoning, and Inference* by Judea Pearl (Cambridge University Press, 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)).

there is no need to consider the concept of counterfactuals to make causal inferences. The need for consideration of counterfactuals arises in analysis of observational data, where allocation to treatment is not determined by randomization[22].

Morgan and Winship op. cit. present four basic approaches to causal analysis: associational analysis; conditional associational analysis; mechanism-based analysis; and all-cause structural analysis. The approach of structural-equation modeling that prevails in economics exemplifies the last approach. Judea Pearl op. cit. promotes the view that we should strive to get as close as possible to the goal of all-cause structural models. A recognized weakness of structural analysis based on the use of covariates and instrumental variables is that the assumptions involved (e.g., conditional independence) are difficult to justify and impossible to prove, and that the approach produces estimates of marginal causal effects that cannot unequivocally be extrapolated to the general population of interest, beyond the particular sample under consideration.

For the FTDA evaluation, we adopt the potential outcomes framework and strive to achieve all-cause structural models, while making use of unexplained associational relationships if a satisfactory all-causal model cannot be determined[23].

It may appear that inordinate attention is being focused on the conceptual framework and methodological approach to be used in estimating impact. A principal reason for this emphasis is that there are alternative methodologies available for analyzing the survey data, i.e., a model-based approach and a design-based approach. (The model-based approach may be further split into a "model-dependent" approach, which is based almost solely on a causal model and not on the survey design model, and a "model-assisted" approach, which involves consideration of both the causal model and the sample design model.) The design-based approach is used for descriptive sample surveys for which the goal is to produce unbiased estimates of characteristics of the finite, extant population being surveyed (or subpopulations), such as means, proportions or totals. The model-based approach is used for analytical surveys, in which the objective is to develop a mathematical model of a process, such as to assess the impact of a development

[22] Some researchers strongly object to the counterfactuals approach. Objections to consideration of counterfactuals is discussed by A. P. Dawid in the article "Causal Inference without Counterfactuals," *Journal of the American Statistical Association*, Vol. 95, pp. 407-24. The problem with counterfactuals is that they are "metaphysical" (hypothetical) in nature, and that this approach departs from the traditional positivist-empiricist approach of science, viz., reliance on observable quantities. Quoting from Morgan and Winship op. cit., "The reliance on what-if potential outcomes and the consideration of unobservable characteristics of the treatment assignment / selection process consigns the counterfactual model to the post-positivist model of science generally labeled realism." The strongest argument in favor of the potential-outcomes approach is that it is useful, and a more useful alternative has not been proposed for analysis of observational data.

[23] The models used for this analysis are described in the book, *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, by Jeffrey M. Wooldridge (Massachusetts Institute of Technology Press, 2010, 2002). For additional information, see *Mostly Harmless Econometrics* by Joshua D. Angrist and Jörn-Steffen Pischke (Princeton University Press, 2009); *Counterfactuals and Causal Inference: Methods and Principles for Social Research* by Stephen L. Morgan and Christopher Winship (Cambridge University Press, 2007); *Micro-Econometrics for Policy, Program, and Treatment Effects* by Myoung-Jae Lee (Oxford University Press, 2005; *Analysis of Panel Data* 2nd edition by Cheng Hsaio (Cambridge University Press, 1986, 2003); *Econometric Analysis of Panel Data* by Baki Baltagi, 4th ed., (Wiley, 2008); and *Econometric Analysis* 7th ed. by William H. Greene (Prentice Hall, 2011).

program or policy, or to describe the relationship of impact to explanatory variables. Under this approach, the population is conceptually infinite, and the model of interest is a causal model that describes how the extant finite population is affected by it (or produced from it).

In this impact evaluation, we originally planned to use a design-based approach, but for the reasons explained, ended up using a model-based approach. The theory of design-based estimation is well established and presented in standard texts on sample survey design and analysis. The theory of model-based estimation involving potential outcomes is less well known (although it, too, is well established and presented in standard texts), and it is considered desirable to summarize that approach. In the discussion that follows, it is assumed that samples are selected from a conceptually infinite population, and that, in general, explanatory variables are viewed as random variables (rather than fixed, nonstochastic variables)[24].

A fundamental aspect of the present application is that the evaluation design involves a panel sample for which, in most instances, the *same* households are interviewed at the baseline and endline times. This type of design purposely introduces a correlation into the "before" and "after" observations, which substantially increases the precision of difference estimates (either first-difference or double-difference estimates or regression-based estimates, depending on the particular design and model used).

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)[25]. (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 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

[24] See "Appendix A: Population and Data Generation Models" on pp. 51-53 of Morgan and Winship op. cit. for additional discussion of this point. See also Wooldridge op. cit. (2002 or 2010) pp. 4-7.

[25] 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.

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 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 334 for Round 2. 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 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 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 estimate of program impact.

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

AldeaTrt = 1 if PC = 1, 2, 3, 4, 8, 10 or 11; 0 otherwise
AldeaCtrl = 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
AldeaTrt	0	2310	2076	1765	1597
	1	2216	660	970	660
AldeaCtrl	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. 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.

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

Note: All units of measure for the 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

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 most instances, missing values in regression models were imputed by substitution of the mean of the non-missing values.

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[26]. 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.)

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

[26] 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
FactorOC = .094 IncOC
TranspOC = .0065 IncOC
OthCostOC = .017 IncOC
LabExpOC = .067 IncOC.

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

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

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. 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 1.95 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).)

III.A Impact Estimators of Interest[27]

We use the following impact estimators for this evaluation analysis:

1. Basic propensity-score-based estimator of average treatment effect (ATE)
2. Regression-adjusted propensity-score-based estimator of ATE
3. Modified regression-adjusted propensity-score-based estimator for ATE
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.

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 Design, Treated, or DesSelForTrt. For the more complex estimates, the regression models include both design parameters and other explanatory variable (covariates such as family size, education of head or household, or 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 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 here is restricted to the analysis.)

It may be asked why this analysis involves several estimators of impact. The general approach to modeling used in this analysis is “structural equation modeling” based on causal modeling and counterfactuals. (For a description of this approach, see *Causality: Models, Reasoning, and*

[27] 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.

Inference by Judea Pearl (Cambridge University Press, 2000).) An essential aspect of structural equation modeling is assessment of the goodness of fit of a model, leading to selection of a preferred one. A number of approaches and tools are used for assessing goodness of fit, including the principle of conditional error, in which the variance of the model residuals is compared for different model formulations. A variety of measures of goodness of fit are available, including the coefficient of determination ("R^2"), the Akaike Information Criterion (AIC), the Schwarz Bayesian Information Criterion (BIC), the root mean square error of approximation (RMSEA), and a variety of statistical tests can be conducted to assess goodness of fit (chi-squared, F and t). These tools involve analysis of model residuals (differences between observed values and values predicted by the model), and are produced by Stata's "postestimation" procedures. Some of them place a premium on model parsimony of parameterization, i.e., on representations that use relatively few parameters. In general, considerable weight is given to the "face validity" of a statistical model with respect to the underlying causal model. Based on subjective judgment of the reasonableness of the causal model and examination of the various goodness-of-fit measures, tests and graphs, most confidence was placed in the modified regression-adjusted propensity-score-based estimator of ATE (i.e., the third estimator listed above). The analysis that follows presents results for all of the five estimators listed above.

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).

In the main text, we present summary results for just one of the preceding estimators of ATE, viz., the modified regression estimator of ATE based on the estimated propensity score. We note that in the case of a descriptive survey (intended to provide estimates of population means and totals), the formulas (or numerical algorithms) for the estimates are fixed (determined) by the survey design. In an analytical survey (as in the present application), however, the objective is to estimate a causal model, and there is no formula for accomplishing this. A number of estimators are available, and their performance depends on the application. While it may be confusing to examine a number of different estimators, doing so is considered desirable (since some work better than others in different situations, and which ones work better in the present application is not generally known in advance of conducting the analysis). It is because no fixed formula (or numerical algorithm) is available for constructing an analytical model that detailed description is presented in this report about the model development and model-based estimation process.

Regardless of the estimator used, it is necessary that it take into account the design features. 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, 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 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. (It is noted that there are other tests for specification, such as the "principal of conditional error" or the so-called "Hausman" test, which compares model parameters for alternative specifications.) 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. Unfortunately, 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 (identified).

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 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.

Because of the similarity of the impact estimators to double-difference estimates, 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.

Some regression models (e.g., logistic regression models) were developed with and without sample “weights” (reciprocals of the probabilities of selection). No significant differences were observed between the weighted and unweighted estimates. This is evidence that the models are well specified (identified). The goal in this project is to construct models that are sufficiently well specified that the use of weights does not make much difference. As mentioned, Stata panel-data-analysis procedure, *xtreg*, does not allow the use of weights.

III.B Assumptions about the Stochastic Nature of Explanatory Variables

The following sections present estimates of impact for the impact measures and estimators listed above. These impact estimates and their estimated standard errors were determined using the Stata regression progam *xtreg* (along with bootstrapping). This program may be used in two modes, depending on the stochastic nature of the explanatory variables, i.e., whether it is assumed that the explanatory variables are random variables or non-random specified numbers. The first mode is called the “random effects” assumption and the second mode is called the “fixed effects” assumption. In evaluation research, it is usually assumed that explanatory variables (other than design variables) are random, since that assumption corresponds to the (analytical-survey) conceptual framework of making inferences about the effect of the program on the population being surveyed, rather than the (descriptive-survey) conceptual framework of describing the surveyed population. In addition, it is standard practice in econometric analysis to assume that most model explanatory variables are random variables, since that formulation allows for correlations among the variables (which is often the case in economic applications), whereas the fixed-effects assumption does not.

In general, the choice between the random-effects and fixed-effects assumption in econometric analysis may not be clear-cut. The random-effects assumption is more natural, since, as mentioned, in economic theory most variables are viewed as random variables, and in evaluation research the populations being surveyed are viewed as particular samples from a conceptually infinite population that may be affected by the program intervention. The random-effects model handles unobserved variables simply by including them in the model error term. A difficulty with the random-effects model is that assumptions must be made about the correlation between unobserved variables (even if they are constant in time) and the observed explanatory variables (i.e., $E(c|\mathbf{x}) = E(c) = 0$, where c denotes an unobserved variable). This difficulty disappears under the fixed-effects assumption. The unobserved variables are simply numbers, and may be arbitrarily related to **x**. A difficulty that arises under the “fixed-effects” assumption, however, is that it is not possible to distinguish the effects of time-invariant observed variables from time-invariant unobserved variables (since the unobserved variables may be arbitrarily correlated with the observed variables). Variables that have no over-time variation must be dropped from the fixed-effects model. There must be some over-time variation in variables in order to estimate their effects. Since many variables within a household remain the same (e.g., the gender and level of education of the head of household, and the home type of construction and location), estimation of the temporal effect of those variables on outcome is not possible from within-household data (i.e., they are confounded with unobserved time-invariant variables).

In summary, the fixed-effects assumption allows us to ignore the effects of time-invariant unobservables, but there must be over-time variation (not just cross-sectional variation) in the observables to estimate their effects over time. If it is desired to assess the effects of variables that are time-invariant in a fixed-effects model, then they must be represented in the model as interaction effects with time (i.e., their effect is estimated from cross-sectional (between-unit) variation).

The Stata *xtreg* procedure requires that all explanatory variables be random, or all be fixed, but it does not handle the "mixed-effects" model in which some variables are random and some are fixed. In the present application, some of the design parameters, such as survey round, are fixed effects; some, such as aldea and household, are random; and some, such as treatment variables, may be viewed either way. Covariates such as household size and farmer level of education are all random. Apart from Round and Treated, the most important design feature of this application is the fact that households are (by and large) matched between survey rounds. A Stata procedure, *xtmixed*, is available that handles mixed effects, but its running time is very slow compared to *xtreg* – far too slow for extensive use in the present application. It was applied in some examples, and the results were similar to those for *xtreg* under either the all-fixed-effects or all-random-effects assumptions.

In the present application, similar results are obtained under any of the three assumptions about randomness (all fixed effects, or all random effects, or mixed effects). Because the fixed-effect models in this application are less sensitive to model specification errors than random-effects models, most models were run assuming all fixed effects. (Actually, although it is the estimates based on the all-fixed-effects models that are presented here, virtually all models were run under both assumptions (all fixed or all random effects).)

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 to obtain detailed information for the other outcome measures.

III.C. Calculation of Estimates

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

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 were used, 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) 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).

The following model was developed to estimate the probability of selection of a household for treatment (provision of 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. 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.

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 $\mathbf{x}'\boldsymbol{\beta}$ 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.

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

$$\mathbf{x}'\boldsymbol{\beta} = -10.87894 - .1250273*\text{HouseholdSize} + .1594562*\text{FormalEducHead} + .868967*\text{AgEmployees} - .0870384*\text{TotHaOwnFarm} + .0211418*\text{TimeToSchool} - .0103442*\text{TimeToHosp} + .9303271*\text{LogTotHHExp} + .2160815*\text{LogIncBG} - .2096164*\text{LogLabExpBG} + .2420389*\text{LogIncOC} - .2358687*\text{LogLabExpOC}$$

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. 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.

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 coefficients became 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.)

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.

The output of the Stata procedure for determing 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.

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**). Note that this assumption "takes care of" the problem of p values of 0 and 1. Development of a strong binary response model, based on a wide range of observables, provides high assurance that condition (1) holds. Since the questionnaire contains so many variables, it is considered unlikely that there are important "hidden" farmer-level variables that affect participation. 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 questionnaire in its rejection of aldeas. On the other hand, 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 (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 $0 < P(w=1|\mathbf{x}) < 1$, 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.”)

Implementation of the procedure for estimating the correct standard error is implemented in Stata by defining an “ado” file that calculates the estimate, and applying the *bootstrap:* procedure to this ado file. (It is noted here that it is somewhat presumptuous to use the term “correct” in referring to the recommended procedure for estimating the standard error, although this (or similar adjectives, such as “proper” or “valid”) 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.”)

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 (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.

For basic grains (BG):

IncBG:-758, se = 606
ExpBG: 635, se = 278
NetBG: -1394, se = 530
LabExpBG: 522, se = 180

For other crops (OC):

IncOC: 7745, se = 3373
ExpOC: 4401, se = 911
NetOC: 3344, se = 3018
LabExpOC: 2791, se = 601

For labor-market employment and household income and expenditures:

IncEmp: -157, se = 682
TotHHExp: -193, se = 363
NetHHInc: 3378, se = 8472

Production of horticultural crops:

Horticulture: -.040, se = .022

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.[28] The estimated effect on NetHHInc is positive, but not statistically significant.

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 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

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. The results obtained from this estimator are similar to those for the basic propensity-score estimator.

[28] 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.)

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.2. Figure A.2 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 (.0598) 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.2. Regression-Adjusted Propensity-Score-Based Estimate of ATE, for ExpOC

```
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
```

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.

For basic grains (BG):

IncBG: -147	naïve se = 939	bootstrap se = 834
ExpBG: 923	naïve se = 359	bootstrap se = 391
NetBG: -1070	naïve se = 815	bootstrap se = 735
LabExpBG: 435	naïve se = 238	bootstrap se = 260

For other crops (OC):

IncOC: 13205	naïve se = 4478	bootstrap se = 4096
ExpOC: 4972	naïve se = 1031	bootstrap se = 1097
NetOC: 8233	naïve se = 4037	bootstrap se = 3950
LabExpOC: 2098	naïve se = 678	bootstrap se = 745

For labor-market employment and household income and expenditures:

IncEmp: -44	naïve se = 724	bootstrap se = 750
TotHHExp: 344	naïve se = 446	bootstrap se = 464
NetHHInc: 11796	naïve se = 10934	bootstrap se = 13254

Production of horticultural crops:

Horticulture: -.0348	naïve se = .0245	bootstrap se = .0193

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.

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.)

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, $\hat{p}(\boldsymbol{x})$, Treated($\hat{p}(\boldsymbol{x})$ - $\hat{\mu}_p$) and RoundTreated($\hat{p}(\boldsymbol{x})$ - $\hat{\mu}_p$), where $\hat{\mu}_p$ denotes the mean of the estimated propensity scores. The additional assumption required for use of this estimator is that E(y_0|p(**x**)) and E(y_1|p(**x**)) are linear in p(**x**).

The modified regression analysis of ExpOC is shown in Figure A.3.

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

```
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
 TreatedPstd |  (dropped)
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) =     1.94          Prob > F = 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).

Income and expense measured in Honduran lempiras.

For basic grains (BG):

IncBG: -120	naïve se = 979	bootstrap se = 837
ExpBG: 837	naïve se = 375	bootstrap se = 393
NetBG: -957	naïve se = 851	bootstrap se = 750
LabExpBG: 351	naïve se = 248	bootstrap se = 264

For other crops (OC):

IncOC: 16773	naïve se = 4665	bootstrap se = 4298
ExpOC: 5413	naïve se = 1075	bootstrap se = 1078
NetOC: 11360	naïve se = 4206	bootstrap se = 4175
LabExpOC: 1911	naïve se = 707	bootstrap se = 742

For labor-market employment and household income and expenditures:

IncEmp: 149	naïve se = 755	bootstrap se = 733
TotHHExp: 204	naïve se = 465	bootstrap se = 496
NetHHInc: 18926	naïve se = 11411	bootstrap se = 13306

Production of horticultural crops:

Horticulture: -.0397	naïve se = .0258	bootstrap se = .0194

The results are similar to the earlier estimates, and the conclusions drawn are the same.

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.

4. 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. This approach was examined, but not found to be not as useful (precise) as the propensity-score approach. Nevertheless, the results are provided in this subsection and the next, for information. 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 logistic 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 specifications of the causal model.

This subsection presents a regression model in which the treatment variable, Treated, is assumed to be an exogenous variable, not correlated with the model error term. This assumption is tenuous, and the following subsection drops this assumption. The results of this subsection are presented just for information, as background for the results of the next section.

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

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

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

δ = impact (interaction effect of treatment and time)

u_t = model error term.

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 treatment effect is the coefficient 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 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 or educational level of the head of household, but also design parameters, such as 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).

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.

Figure A.4 shows the regression analysis of ExpOC. It is not apparent from the variable names regression program output that the covariates of the model (the RT* terms) were demeaned, but they were.

Figure A.4. Regression Estimate of ATE, Not Based on the Estimated Propensity Score, for ExpOC

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

R-sq:  within  = 0.0794                         Obs per group: min =         1
       between = 0.0753                                        avg =       1.6
       overall = 0.0787                                        max =         2

                                                F(12,2721)         =     19.56
corr(u_i, Xb)  = 0.0098                         Prob > F           =    0.0000

------------------------------------------------------------------------------
       ExpOC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Round |   377.8979   356.6774     1.06   0.289     -321.488    1077.284
     Treated |  (dropped)
RoundTreated |   4592.003   829.6086     5.54   0.000     2965.276    6218.729
    MeanEduc |   288.4949   201.8661     1.43   0.153    -107.3315    684.3213
 AgEmployees |   1258.665   463.1651     2.72   0.007      350.474    2166.856
 TotHaOwnFarm |  8.531448   24.25812     0.35   0.725    -39.03475    56.09765
   EqptValue |   .0157555   .0027145     5.80   0.000     .0104328    .0210782
InstRental~e |  -.2558013   .1305267    -1.96   0.050    -.5117428    .0001402
  RTMeanEduc |   1093.457   234.6361     4.66   0.000     633.3743     1553.54
RTAgEmploy~s |   51.67062    883.066     0.06   0.953    -1679.877    1783.218
RTTotHaOwn~m |   23.64938   28.58886     0.83   0.408    -32.40868    79.70745
 RTEqptValue |  -.0022424   .0017931    -1.25   0.211    -.0057584    .0012737
RTInstRent~e |   .7992302   .2857984     2.80   0.005     .2388264    1.359634
       _cons |   2127.976   861.0994     2.47   0.014     439.5008    3816.451
-------------+----------------------------------------------------------------
     sigma_u |   13246.94
     sigma_e |  11357.732
         rho |  .57633244   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4525, 2721) =     1.93          Prob > F = 0.0000
```

Listed below are regression estimates of the ATE, not based on the estimated propensity score, along with standard errors calculated from the regression model, for all of the outcome indicators of interest. (The full regression output is shown above just for ExpOC.) Note that this regression is estimated by the method of ordinary least squares and does not involve a propensity score. For this reason, the standard errors may be taken directly from the regression output, without the need for bootstrapping.

For basic grains (BG):

IncBG: -2974, se = 747
ExpBG: 1121, se = 287
NetBG: -4095, se = 659
LabExpBG: 1237, se = 194

For other crops (OC):

IncOC: -5315, se = 3565
ExpOC: 4592, se = 830
NetOC: -9907, se = 3235
LabExpOC: 4841, se = 558

For labor-market employment and household income and expenditures:

IncEmp: 146, se = 585
TotHHExp: -942, se = 354
NetHHInc: -11196, se = 8798

Production of horticultural crops:

Horticulture: -.0490, se = .0210

In contrast to the case for the propensity-score-based estimators, the preceding results do not show that the program had any positive effect, either for basic grains or for other crops or for net household income. In fact, NetOC is statistically significantly negative.

At first glance, it might be thought that this outcome model should produce results similar to the selection model, because it is based on similar variables as the selection model and because those variables may be represented in the model in any way, not just through a propensity score. (The variables are not exactly the same, since the selection model is based on Round 0 observables, and the outcome model is based on Round 1 observables that are not considered endogenous.) The fact is that the logistic-regression-based selection model fits the Round 0 data very well (the structural representation of the selection process is represented better), and standard linear “kitchen-sink” regressions (relating outcomes of interest to exogenous observables) do not. (It should be kept in mind that the outcome regression model has a relatively simple structure (linear in the parameters), whereas the selection model (the logistic binary-response model) is, by comparison, quite complex. The logistic binary-response model represented the selection (participation) process well, and led to good propensity-score-based estimators of impact, but the simple linear-regression model did not represent the program process well.)

A second question that might be raised is why the outcome model actually indicates *negative* impact. A partial explanation for this is that the Observed Treatment Effect (OTE – the treatment effect estimated from a model that includes only the Treated variable and the design parameters, but no explanatory variables) is negative for NetOC. The OTEs are as follows (these estimates are the "raw" double-difference estimates, taking into account the design feature that the same households were interviewed in both survey rounds, but no other explanatory variables):

For basic grains (BG):

IncBG: -2300, se = 672
ExpBG: 1201, se = 257
NetBG: -3502, se = 585
LabExpBG: 1310, se = 171

For other crops (OC):

IncOC: 4301, se = 3205
ExpOC: 6512 se = 738
NetOC: -2211, se = 2893
LabExpOC: 5873, se = 490

For labor-market employment and household income and expenditures:

IncEmp: 336, se = 517
TotHHExp: -560, se = 319
NetHHInc: 237, se = 7792

Production of horticultural crops:

Horticulture: -.0427, se = .0183

While the NetOC is not statistically significantly negative for the OTE, it is in fact negative. This is an unusual and unexpected situation for a program such as FTDA, which appears to have very positive results – one would expect strong positive program to be reflected in the observed treatment effect. It is a partial explanation for why the regression models not based on the propensity score produce the negative results that they do – the models are "weak," and hence they simply reflect the "raw" observed treatment effects. There may be economic explanations for this situation, such as the initial capital investment required by the program, but those explanations are not readily apparent from the available data.

The selection model is relatively strong, but it is not reflected in outcome regression models that do not include it. Absent the estimated propensity score, the outcome models are of relatively little use in this application. They may be useful for additional econometric analyses, but they do not reflect the impact of this program.

As mentioned in the beginning of this subsection, the models of this subsection are based on the assumption that Treated is an exogenous variable, and this assumption may not be justified. The following subsection drops this assumption. Since the results for the OLS regression are weak, and it is expected that the instrumental-variable regression model would be similar.

5. 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. 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 estimated propensity score is considered to be uncorrelated with the model error term since it is based solely on Round 0 data ("selection on observables"), and the effects of interest occur after Round 0.

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 regression analysis of ExpOC is shown in Figure A.5. (Note that this regression model is based on a "fixed-effects" assumption, and that any household variables that are constant between survey rounds are dropped from the model. For this reason, the variables P and Treated are dropped from the model. The printout shows both the "first-stage" regression of the instrumented variables Treated and RoundTreated on their instruments P and RoundP, as well as the complete two-stage regression. P is dropped from the first-stage regression and Treated is dropped from the two-stage regression.) The R^2 for the first-stage regression of RoundTreated on RoundP is .5684, which is relatively high. This value is in fact higher than the .44 R^2 value for the logistic-regression propensity score model. It is higher because this model is in terms of the original variables (e.g., income, expense), whereas the logistic-regression model was in terms of logarithmic transformations of income and expense. The value of R^2 for the second-stage regression is low (.0879), but this is of little interest. What is of greater interest is the standard error of the coefficient of RoundTreated, which is the estimate of impact.

Figure A.5. Estimate of ATE of ExpOC Based on Instrumental-Variable Regression

```
First-stage within regression

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

R-sq:  within  = 0.0000                         Obs per group: min =         1
       between = 0.0000                                        avg =       1.6
       overall = 0.0000                                        max =         2

                                                F(6,2727)          =      0.00
corr(u_i, Xb)  = -0.0000                        Prob > F           =    1.0000

------------------------------------------------------------------------------
     Treated |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    MeanEduc |   5.27e-27   3.62e-16     0.00   1.000    -7.10e-16    7.10e-16
 AgEmployees |   5.19e-27   8.23e-16     0.00   1.000    -1.61e-15    1.61e-15
TotHaOwnFarm |   1.24e-29   2.13e-17     0.00   1.000    -4.18e-17    4.18e-17
   EqptValue |  -3.18e-33   3.86e-21    -0.00   1.000    -7.56e-21    7.56e-21
InstRental~e |  -2.11e-31   2.21e-19    -0.00   1.000    -4.33e-19    4.33e-19
           P |  (dropped)
      RoundP |   1.77e-27   1.68e-15     0.00   1.000    -3.30e-15    3.30e-15
       _cons |   .1895578   1.55e-15  1.2e+14   0.000     .1895578    .1895578
-------------+----------------------------------------------------------------
     sigma_u |  .36824566
     sigma_e |  2.079e-14
         rho |          1   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4525, 2727) =        .          Prob > F =      .

First-stage within regression

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

R-sq:  within  = 0.6137                         Obs per group: min =         1
       between = 0.5338                                        avg =       1.6
       overall = 0.5684                                        max =         2

                                                F(6,2727)          =    722.02
corr(u_i, Xb)  = -0.0912                        Prob > F           =    0.0000

------------------------------------------------------------------------------
RoundTreated |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    MeanEduc |   .0041443   .0037187     1.11   0.265    -.0031476    .0114361
 AgEmployees |   .0275022   .0084595     3.25   0.001     .0109145    .0440898
TotHaOwnFarm |   .0009777   .0002192     4.46   0.000     .0005478    .0014075
```

```
   EqptValue |   8.95e-08   3.96e-08     2.26   0.024     1.18e-08    1.67e-07
InstRental~e |   4.67e-06   2.27e-06     2.05   0.040     2.14e-07    9.12e-06
           P |  (dropped)
      RoundP |    1.07813   .0172676    62.44   0.000     1.044271    1.111989
       _cons |  -.0352659   .0159298    -2.21   0.027    -.0665015   -.0040302
-------------+----------------------------------------------------------------
     sigma_u |  .12036611
     sigma_e |  .21357608
         rho |  .24105389   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4525, 2727) =     0.61          Prob > F = 1.0000

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

R-sq:  within  = 0.0649                      Obs per group: min =            1
       between = 0.0921                                     avg =          1.6
       overall = 0.0879                                     max =            2

                                                F(4532,2727)    =       30.40
corr(u_i, Xb)  = 0.0268                         Prob > F        =       0.0000

------------------------------------------------------------------------------
       ExpOC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Treated |  (dropped)
RoundTreated |   7620.935   857.4776     8.89   0.000     5939.563    9302.306
   MeanEduc |    445.854   199.2412     2.24   0.025     55.17491     836.533
 AgEmployees |   1448.512    448.435     3.23   0.001      569.205    2327.818
TotHaOwnFarm |   30.09843   11.86848     2.54   0.011      6.82631    53.37055
   EqptValue |   .0145942   .0021353     6.83   0.000     .0104072    .0187812
InstRental~e |  -.1636868   .1218605    -1.34   0.179     -.402635    .0752615
       _cons |   1340.535   850.2306     1.58   0.115    -326.6266    3007.696
-------------+----------------------------------------------------------------
     sigma_u |  13136.716
     sigma_e |  11434.429
         rho |  .56894896   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(4525,2727) =     1.92       Prob > F    = 0.0000
------------------------------------------------------------------------------
Instrumented:   Treated RoundTreated
Instruments:    MeanEduc AgEmployees TotHaOwnFarm EqptValue InstRentalValue P RoundP
------------------------------------------------------------------------------
```

Below, we present IV-regression estimates of the ATE and its standard error from the regression output and from bootstrapping, for all of the impact measures of interest. (The regression output is shown, above, only for ExpOC.) Estimates of the standard error are presented for both the regression-model output (which does not take into account the fact that the propensity score is an estimate), and from bootstrapping (which does account for the fact that the propensity score is an estimate). Note that this model is not a two-step M-estimator under ignorability of treatment, and so the remarks made earlier about the naïve estimator of the standard error being conservative do not apply. In this case, the two-stage IV regression ignores the fact that the propensity score is an estimator, and the naïve estimate of the standard error from the regression is not conservative. In this case, the “proper” standard errors obtained by bootstrapping are likely to be larger than the naïve standard errors from the regression output, as indeed they are seen to be. (In this application, the basic conclusions do not differ when moving from the naïve estimator of the standard error to the proper estimate – statistically significant positive results were not obtained using the naïve estimator, and the results are even less signficant with the larger standard errors.)

In the following list, "regression-model se" denotes the standard error as calculated by the regression program without taking into account that the propensity score used in the model is an estimate, and "bootstrap se" denotes the "proper" estimate taking into account the fact that the propensity score is an estimate.

For basic grains (BG):

IncBG: -4094, regression-model se = 773; bootstrap se = 1220
ExpBG: 1515, regression-model se = 297; bootstrap se = 334
NetBG: -5609, regression-model se = 679; bootstrap se = 1022
LabExpBG: 2216, regression-model se = 201; bootstrap se = 205

For other crops (OC):

IncOC: -1878, regression-model se = 3676; bootstrap se = 7881
ExpOC: 7621, regression-model se = 857; bootstrap se = 1488
NetOC: -9499, regression-model se = 3332; bootstrap se = 7112
LabExpOC: 8971, regression-model se = 579; bootstrap se = 817

For labor-market employment and household income and expenditures:

IncEmp: 3381, regression-model se = 607; bootstrap se = 987
TotHHExp: 1668, regression-model se = 376; bootstrap se = 417
NetHHInc: 21209, regression-model se = 9066; bootstrap se = 11862

Production of horticultural crops:

Horticulture: -.0329, regression-model se = .0198; bootstrap se = .0171

These results are similar to the regression estimates not based on the propensity score (discussed in the preceding subsection), but generally "weaker" (larger standard errors). The preceding results do not show positive results for the program[29]. An impact of primary interest, NetOC, is in fact of negative sign, but this result is not statistically significant. (In assessing statistical significance, compare the estimate to the "bootstrap se".) The estimate of the effect NetHHInc is positive, and is statistically significant (test statistic = 21209/11862 = 1.79, which exceeds 1.645, the one-sided critical value). The standard errors for the instrumental-variable regression estimator are generally larger than for the propensity-score-based estimates.

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

[29] All of the regressions examined are included in the Do14FTDAEstimation.do Stata command file.

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.D Summary of Program Impact

The preceding sections have presented estimates of program impact for a variety of outcome measures, using several different impact estimation techniques. We believe that greater confidence should be placed in the propensity-score-based estimates (i.e., the first three estimators of the estimator list), both because those models have greater face validity than the other two models (i.e., the structural representation of the selection process is a closer representation to the casual model, or "better specified") and because the standard errors of the estimates are substantially smaller. In the main text, we present results for the modified regression-adjusted propensity-score-based estimate (i.e., the third propensity-score-based model discussed). Based on that estimator, this analysis provides evidence that the FTDA program has a positive effect on income, expense, net income and expenditures for labor for other crops (the category that includes those addressed by the FTDA program). For one estimator (the IV regression estimator), the effect of NetHHInc was statistically significantly positive.