laterite TRANSFORMING RESEARCH IN AFRICA
Rwanda | Malawi | Burundi

TechnoServe Coffee Initiative: Socio-Economic Study - Rwanda

Endline Report | January 2014

TechnoServe International

About the Authors

Laterite Ltd. (www.laterite-africa.com) is a research-consulting firm based in Rwanda and Burundi. We offer research services clustered around economic and social research, policy development and market research. Laterite's research team included Sachin Gathani, Maria Paula Gomez, Ricardo Sabates, and Dimitri Stoelinga.

Table of Contents

1. Introduction

The purpose of this evaluation is to measure the effect of TechnoServe Rwanda's Wetmill and Agronomy programs on the socio-economic status of households in treatment areas. Combined, the Wetmill and Agronomy programs were designed to boost and upgrade coffee production in multiple cooperatives throughout Rwanda. Farmers were given training to produce more and better quality coffee through the agronomy program; this coffee was then processed and fully-washed through newly established wetmills supported by TechnoServe's Wetmill program. Fully washed coffee, or what is often referred to as "specialty coffee", fetches significantly higher prices on global markets. The theory is that through increased production, higher selling prices, and support on improving management practices (be it at the farm, cooperative or wetmill levels), farmers that benefitted from the treatment would see their incomes rise, leading to improved socio-economic outcomes for them and their families.

The Wetmill program supported the creation of "greenfield" wetmill businesses in selected coffee growing areas. This involved:

(i) ***Capacity building***: helping farmers organize to set-up coffee cooperatives, providing training on book-keeping and wetmill operations to cooperative staff and managers, training cooperative members on their roles and responsibilities;

(ii) ***Brokering financing relationships:*** helping cooperatives access and manage loans for the construction of the wetmills (CAPEX) and working capital;

(iii) ***Washing Station construction:*** supporting cooperatives to construct low-cost, eco-friendly coffee washing stations;

(iv) ***Market access:*** developing and facilitating buyer relationships and coaching cooperatives to better understand the needs of buyers;

(v) ***Sustainability and gender equity:*** actively measuring the sustainability of wetmill businesses and supporting greater gender equity.

These activities impacted farmers primarily by: (i) offering them the opportunity to sell their coffee at a higher price by shifting from low-quality semi-washed coffee to high-quality fully washed coffee (an estimated 36% higher farm-gate price); (ii) monetizing the local economy by attracting financing for the construction and running of the wetmills; and (iii) providing capacity building for cooperative and wetmill members, staff and management. While this intervention primarily affected cooperative members, it is likely to have also had spill-over effects on the local economy and on farmers who were not members but nevertheless managed to sell their coffee to the wetmills.

The Coffee Agronomy program, otherwise referred to as the Coffee Farm College, provided coffee farmers with a two-year training program on best agronomic practices for coffee growing. This included training on everything from record keeping and knowledge on pest management, to mulching, weeding, composting, shade management, etc. The program was open to both cooperative members and non-members in the treatment areas and led to an increase in coffee yields and best agronomic practice adoption (see Laterite's Independent Assessment of TechnoServe's Rwanda Coffee Agronomy Program).

The question we are answering here is whether these two programs that were largely successful in achieving their coffee related objectives – improving farming methods, increasing yields and the production and processing of specialty coffee - also led to significant improvements in the socio-economic situation of beneficiary households. To measure the effect of the two programs TechnoServe invested in an independent evaluation that was designed and rolled out in 2009. This

socio-economic impact assessment is based on a quasi-experimental set-up which consists of comparing a randomly selected group of households from treatment areas to a randomly selected group of households from adjacent areas with similar agro-climatic conditions. We will explain and analyze the experimental design in depth in the following sections. Three separate firms/consortiums were recruited to collect and analyze the data. The experimental design and the baseline data was collected by Research International East Africa Limited (RIEAL) and Tegemeo Institute (Egerton University) in 2009; the midline by Incisive Africa in 2011: while Laterite Ltd was selected to collect the endline data and complete the analysis of the evaluation in 2013. What follows is an in-depth analysis of the results we obtain and how we got there.

We first deal with some significant technical issues in the design and implementation of this evaluation, many of which undermine the internal validity of the experiment and reduce its statistical power. We then describe in detail the proposed methodology to circumvent and attenuate some of these distortions, before presenting results. Results relate to six sets of indicators in particular: (i) education; (ii) nutrition and health; (iii) household decision-making; (iv) expenditures, savings and credit; (v) household assets; and finally (vi) access to utilities and housing quality.

2. Technical challenges and design flaws

This evaluation was designed as a quasi-experiment. Its basic building blocks are:

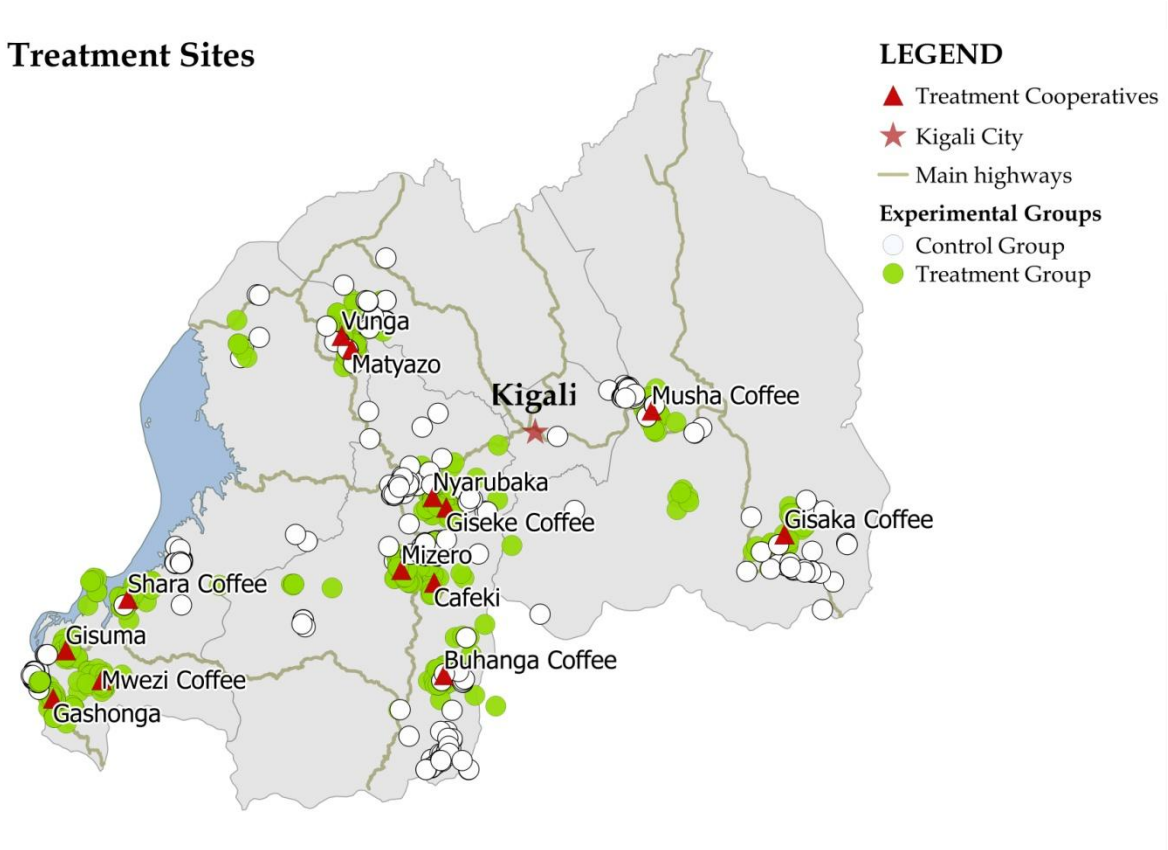
(i) ***A treatment group*** of randomly selected households in the vicinity of 13 different treatment sites scattered around the country. Households were randomly selected from the list of all prospective cooperative members in the 13 newly created cooperatives in late 2008, just as TechnoServe was about to launch its wetmill and coffee agronomy programs. Randomization was stratified at the cooperative level. Households in the treatment area were on this list because they had expressed interest in joining both programs (what we call the "treatment"), but did not necessarily take-up either treatment thereafter. Therefore, they had the option to self-select into the programs.

(ii) ***A control group*** of households that were selected using a "random walk" method in nearby coffee growing areas with similar agro-climatic conditions to the treatment sites, but beyond a 5km radius from the latter. For each treatment site, a list of areas with similar agro-climatic conditions was provided to the research consortium by TechnoServe. Households in that area were selected by "random walk". Stratification was again done at the cooperative level; i.e. for each treatment cooperative there was supposed to be a corresponding control group.

(iii) ***Three time periods.*** We have three data points in time for this evaluation. Baseline data was collected in 2009, before the start of project activities. Midline data was collected in 2011, two years in to the Wetmill and Agronomy programs, which were nearing completion. And an endline in 2013, after the completion of the programs in all 13 cooperatives. It is important to note that data was collected every two years, which means that results will not have been affected by coffee cycles. Coffee trees have a good year, followed by a bad year, and so forth. Farmers in 2009 were therefore in the same coffee cycle in 2011 and 2013.

Figure 1 depicts the distribution of treatment and control groups farmers throughout the country. The 13 cooperatives are indicated by name and highlighted by a small red triangle. Treatment farmers are highlighted by green points and control group farmers by white points.

Figure 1: Distribution of treatment and control groups farmers in Rwanda



We refer to this as a quasi-experiment because it differs from a typical randomized control trial (RCT) in that the treatment and control households were not randomly selected from the same sampling frame. Here we have two distinct sampling frames: (i) the sampling frame for treatment households, which consists of the list of all prospective cooperative members in 13 cooperatives in late 2008; and (ii) a separate sampling frame for control households, which consists of all households living in areas with similar agro-climatic conditions just beyond a 5km radius from the treatment site. This decision was probably made to ensure that this evaluation did not interfere with roll-out of TechnoServe's programs in the treatment areas, which were open to all farmers. Moreover, because of the sheer scale of the programs – which had hundreds of participants in each treatment site - there was a risk of spill-over effects. Selecting control group households from areas beyond the 5km mark reduced the risk that spill-over effects would affect control households as well.

The original intention of the researchers was to identify the treatment effect by studying the change in average outcomes over time in the treatment and control groups, or what is called a difference-in-difference approach. For the estimator to be a valid reflection of the treatment effect however a number of assumptions need to hold, namely:

1) The populations of the two sampling frames need to be very similar on average – i.e. average households from nearby coffee growing areas with similar agro-climatic conditions (altitude, soil quality, etc) need to have similar baseline socio-economic indicators to the treatment households;

2) Within each sampling frame, selection into the respective experimental groups must have been fully random or "as good as random";

3) The Wetmill and Coffee Agronomy programs should not have had any significant effect beyond a 5km radius from the treatment site and living beyond the 5km radium should have significantly reduced the likelihood of control farmers receiving or taking-up the treatment; and,

4) There should have been no other interventions – related or unrelated to coffee - in control areas.

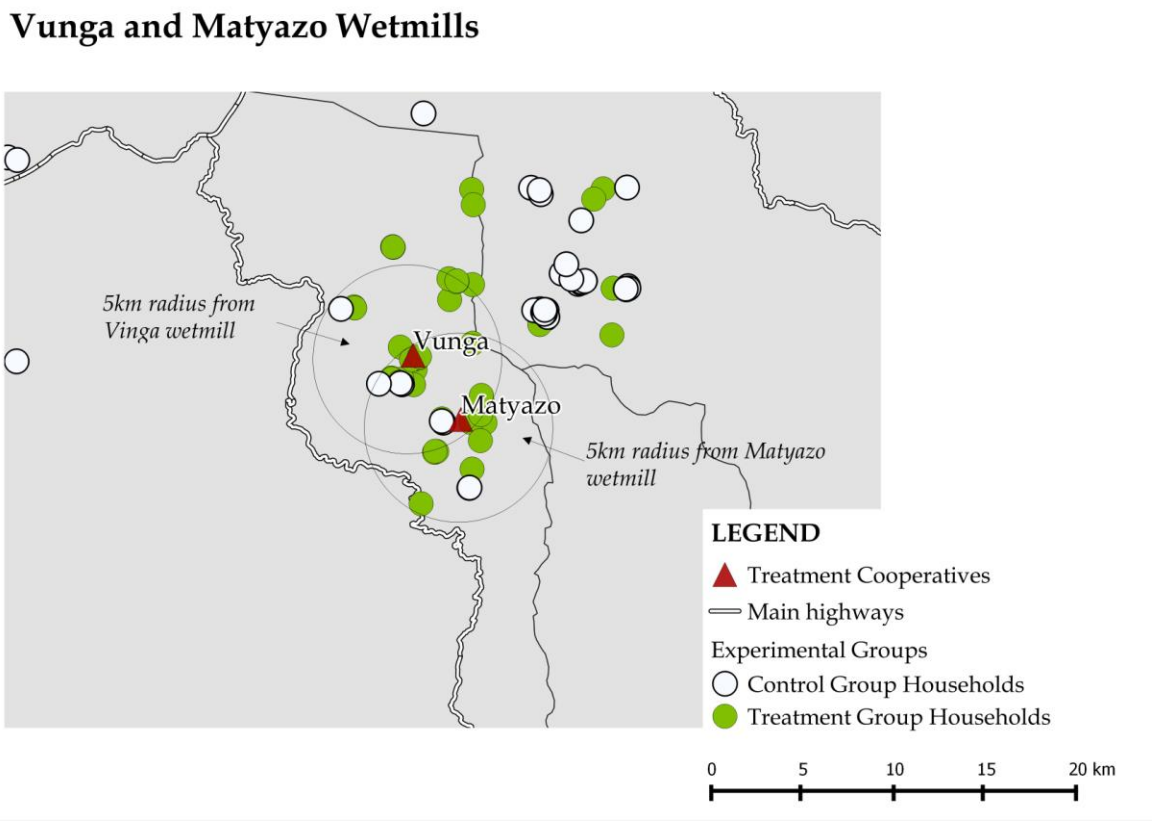
In the sub-sections that follow we test these assumptions and some other aspects of the internal validity of this experiment. As we will see there are many technical and implementation-related issues that have undermined the internal validity of this evaluation and significantly reduced its statistical power. In this section we will study these issues one-by-one, before proposing fixes and alternative estimation techniques in the methodology section.

Issue 1: Creation of Control Groups did not always follow experimental protocol

The first issue with the implementation and design of this evaluation is that the experimental protocols were not strictly adhered to at the baseline in terms of the creation of controls groups. The idea at the design phase was to have one specific control group for each treatment site. The geographic proximity of some of the treatment sites however made this very difficult to implement in practice (also see Annex 1 which lists distances between cooperatives calculated using GPS coordinates). Some treatment cooperatives were just a few kilometers away from each other, often with overlapping 5km radiuses (i.e. the area beyond which control farmers were selected). For these cooperatives the control groups were combined at the baseline. The reason for this is that the small distance between cooperatives conflicted with the research design, which stipulated that households in the control group needed to be selected beyond a 5km radius from the actual treatment site. This was the case for 4 groups of cooperatives: Cafeki and Mizero (about 9 kms apart); Gashonga, Mwezi and Gisuma (10-15 kms apart); Giseke and Nyarubaka (5kms apart); and finally Vunga and Matyazo (4kms apart).

Figure 2 below maps out one such case for the Vunga and Matyazo cooperatives. As can be seen in the map the 5km radiuses of these two cooperatives overlap, making it difficult to create an independent control group for each. The bulk of the combined control group farmers are based north west of the two treatment sites, beyond their 5km radiuses.

Figure 2: Distribution of treatment and control groups farmers in Vunga and Matyazo wetmills



From an evaluation design perspective this creates a number of issues. First of all it means that the intensity of the treatment was not uniformly distributed across treatment sites and their immediate vicinity. Some control groups were in the influence area of two or three treatment sites, while others were only in the influence area of one. It also means that control groups were not selected in the same way – the sampling frame was slightly different for control groups that had multiple treatment cooperatives in their vicinity. This is likely to have been a source of bias although we have no way of testing how this specific issue might have affected outcomes.

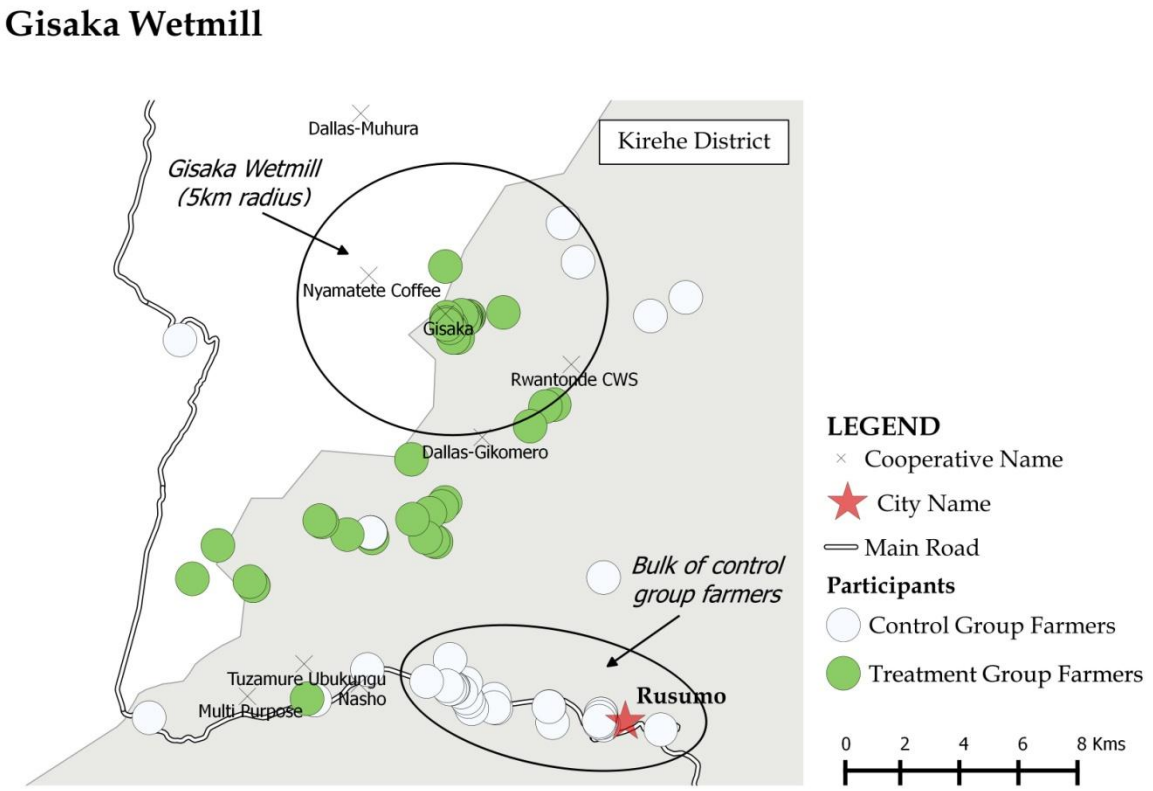
Issue 2: Selection Bias

The second and more worrying problem is significant selection bias in the composition of the control groups. When we take a close look at the composition of the control groups using GPS coordinates, we see clear patterns of selection bias emerging. Control group farmers were much more likely to be near main roads and cities, which we show, is synonymous to having better socio-economic outcomes. We illustrate this here using two examples.

The first is the example of the control group in the Gisaka cooperative. Figure 3 depicts the geographic distribution of treatment and control farmers around the Gisaka Wetmill in the Kirehe district (Eastern Province). As can clearly be seen from the map, while treatment group farmers were on average 5-10kms from the main road (the B3 highway to Tanzania) and at least 10kms from the district capital – Rusumo - the bulk of control group farmers were clustered on the main road within

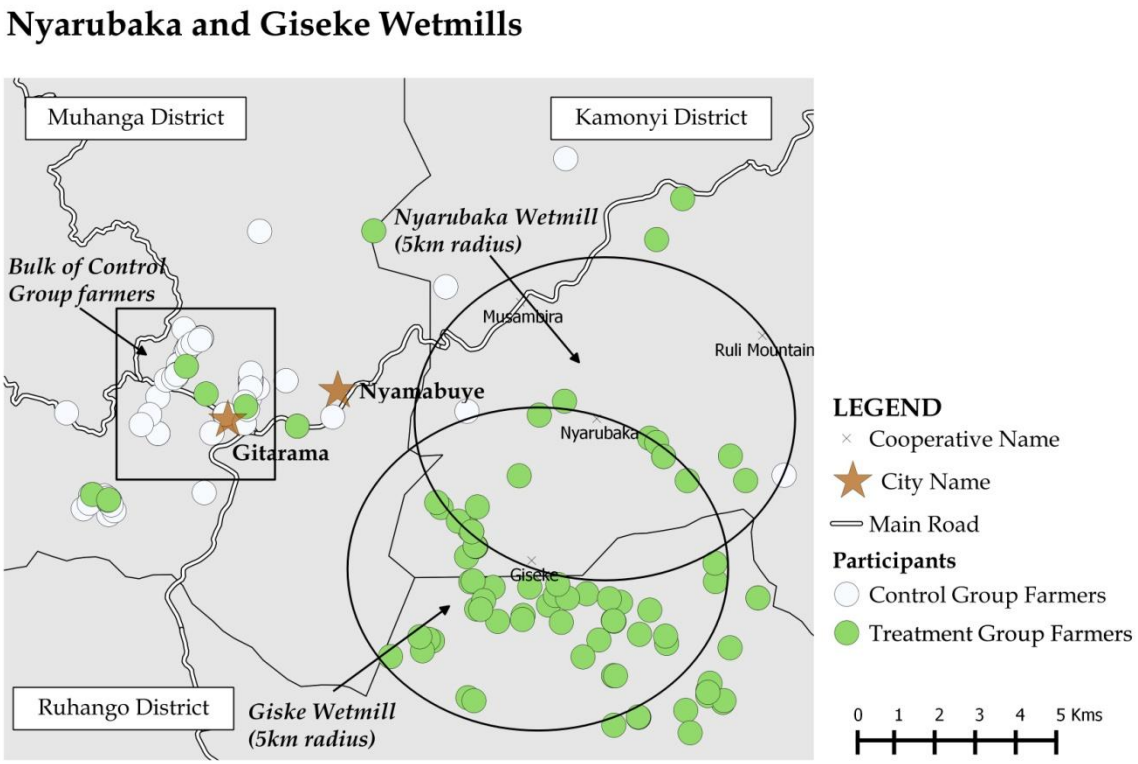
5kms from Rusumo. Even though this is a coffee producing area with other cooperatives such as Tuzamure, Ubukungu and Nasho a few kilometers further down the road, the location of control group farmers suggests that they are significantly more connected than their counterparts in the treatment group.

Figure 3: Distribution of treatment and control groups farmers in Gisaka wetmill



We find a similar pattern in the case of the Nyarubaka and Giseke cooperatives. Figure 4 depicts the geographic distribution of a subset of treatment and control farmers for the Nyarubaka and Giseke cooperatives (some farmers outside the map area have been excluded to enable us to zoom in). Again, the vast majority of treatment farmers are clustered around the two wetmills and are comparatively far from the highway and the two principle cities in the area, Gitarama and Nyamabuye. The bulk of control group farmers, however, are in the immediate vicinity of Gitarama, one of Rwanda's largest cities, and a mere 1-2km from the highway.

Figure 4: Distribution of treatment and control groups farmers in Nyarubaka and Giseke wetmills



These stark patterns signal a very high risk of selection bias; i.e. that receiving the treatment was not the only difference between treatment and control group households. To test whether survey data collected at the baseline confirms this bias, we compare the treatment and control groups on various baseline indicators. It is important to note that the baseline report did not report the possibility of any such bias. .

We start by comparing the treatment and control groups on declared distances to certain facilities and public infrastructure, such as schools, medical facilities, banks, roads, etc. We look at median distances (means can be misleading in this case because of outliers) and test the null hypothesis that the treatment and control groups were equidistant from the selected facilities/services. The results are unequivocal and confirm that control group households were significantly closer to cities and main roads than treatment group farmers. For example, the median control group household lived 3kms from a bank and 6kms from a tarmac road. The median treatment group household on the contrary lived 5kms from a bank and 10kms from a tarmac road. Median households in the control group lived relatively close to urban settlements and main roads, while households in the treatment groups were significantly more rural (see Table 1 below). This is a serious selection bias and omission from the baseline study that undermines the internal validity of this impact evaluation.

Table 1: Distance of treatment and control groups to facilities/public infrastructure

Distance to …	Control	Treatment	T-Statistic (equality test)
Primary School	1km	2 km	11.48***
Secondary School	3 km	4 km	3.47***

Hospital	7 km	8 km	1.33
Dispensary	3 km	4 km	1.73*
Clinic	4 km	5 km	1.38
Bank	3 km	5 km	4.19***
Agro-input supplier	2.5 km	4 km	3.00***
Electricity grid connection	6 km	6 km	0.00
Tarmac	6 km	10 km	5.43***
Motorable Road	1 km	2 km	5.56***
Market center	1.5 km	2 km	5.12***
Public Transport	3 km	5 km	2.80***

**p<0.1; **p<0.05; ***p<0.01*

Not surprisingly therefore, fewer control group households were dependent on coffee. At the baseline, they were less likely to state coffee as their main source of income, more likely not to list coffee as one of their top three income sources, they owned fewer coffee trees, and were investing less in the expansion of their coffee production (see Table 2). Median control group farmers for example owned close to 40% fewer coffee trees than the median treatment farmer.

Table 2: Coffee data for control and treatment groups

Indicator	Control	Treatment	T-Statistic (equality test)
Coffee is main source of income (mean)	55.6%	61.2%	2.2**
Coffee not in top 3 sources of income (mean)	17.0%	10.8%	-3.6***
Coffee acreage has increased over past 3y (mean)	11.9%	17.4%	2.7***
Number of coffee trees (median)	125	200	8.8***

**p<0.1; **p<0.05; ***p<0.01*

Their greater proximity to cities and main roads also implied that control group households were slightly better off at the baseline. Below are some examples of indicators where the difference between treatment and control group households were significant at the baseline. These included food intake, access to utilities and the quality of housing (see Table 3).

Table 3: Significant baseline differences in indicator estimates

Indicator	Control	Treatment	T-Statistic (equality test)
Meals per day (mean)	2.05	2.02	-1.8*
Household owns at least one mattress (mean)	56.4%	49.0%	-2.9***
Source of light: electricity bulb	2.8%	1.3%	-2.1**
Modern walls (not earth or bamboo)	22.3%	18.3%	-1.9*
Modern floors (concrete floors)	20.9%	16.8%	-2.0**
Modern toilet (with flush or with cement floor)	9.0%	3.6%	-4.4***

**p<0.1; **p<0.05; ***p<0.01*

These findings contradict the claim in the baseline report that the treatment and control groups were relatively similar on average. The main difference between the two groups was not that the one received the treatment and the other did not, but rather that control group households were

significantly less rural. This matters because it means that the control group and treatment groups are unlikely to have evolved in parallel over time, which is the basic identifying assumption of Difference-in-Difference estimators. Indeed, we would expect control group households to improve their living standards and socio-economic indicators faster than treatment group households. Given their proximity to tarmac roads, towns, and other services and facilities, they are more likely to have benefitted more from urban growth and investments than treatment households.

Issue 3: Multiple treatments

TechnoServe had two projects ongoing in the selected coffee cooperatives: the Wetmill Program and the Agronomy Program. Farmers in the treatment area could self-select into both programs, one of the two or neither. The treatment group is therefore divided into four separate types of households:

1. Households that participated in both the Wetmill and Agronomy programs:
2. Households that were Wetmill members but did not participate in the agronomy program;
3. Households that were not Wetmill members but did participate in the agronomy program;
4. Households that were neither Wetmill members nor participated in the Agronomy program.

The distribution of treatment group farmers across these four groups is summarized in table 4.

Table 4: Distribution of treatment group households

Treatment received	Number of households	Share
Both Wetmill and Agronomy programs	186	41.4%
Wetmill only	64	14.3%
Agronomy only	61	13.6%
No direct treatment received	138	30.7%
Totals	**449**	**100%**

On the positive side, this variation within the treatment group can give us some valuable insights on the type of farmers that self-selected into one treatment or the other. On the negative side, the small sample sizes within each group and the fact that treatment farmers self-selected into these programs and were not randomly assigned to them, implies that it will be virtually impossible to isolate the individual impact of either program on the socio-economic status of beneficiaries. This is further complicated by the fact that the interventions started at different points in time. While all the wetmills in the respective cooperatives were created in 2009, the Agronomy program started in 2010 for some cooperatives and in 2011 for others. The three cooperatives where the Agronomy program started in 2011 were Matyazo, Shara and Vunga. This adds a time-dimension to the already divided treatment group.

The difference-in-difference estimate of the treatment effect will therefore be a mix of these different levels of intervention. It will be an estimate of the **Intention-to-Treat**, regardless of whether the households were treated and which type of treatment they opted for.

Issue 4: Incomplete compliance in the treatment and control groups

Related to this is the problem of "incomplete compliance". We talk about incomplete compliance when individuals in the treatment group forego the treatment or on the contrary when individuals in the control group receive the treatment. This is a common issue with experiments, as the researcher often has no control over whether individuals in the treatment group decide to self-select into a program or alternatively to block individuals in the control group from receiving the treatment. The problem with incomplete compliance is that it significantly reduces the statistical power of the experiment. In fact, the sample size required to compensate for the lost statistical power is inversely proportional to the square of the difference in the take-up of the treatment between the treatment group and the control group[1]. For example if 75% of the treatment group had received the treatment and 25% of the control group as well, then the difference in take-up would be 50% (75%-25%=50%). In this case the required sample size to maintain the same level of significance as if there were full compliance would be the inverse of the square of 50%, which is 4 times the original sample size (i.e. $1/0.5^2=4$)

This example is quite close to the reality of this socio-economic impact survey, where about 76% of the treatment group received the treatment, as did 18% of the control group at the end-line stage of the study (see table 5):

Table 5: Take-up of Treatment and Control Groups at Endline phase

Intervention	Control Group	Treatment Group
No Treatment	81.9%	24.3%
Treatment	18.1%	75.7%
Total	**100.0%**	**100.0%**

The difference in take-up between the treatment group and the control group is therefore 57.4% (75.5%-18.1%). ***A sample three times the size of the current sample would have been needed to maintain the same statistical power as if there had been full compliance (1/0.574²=3.0).***

A different way to look at this question would be to ask what is the loss of statistical power resulting from this level of compliance? We illustrate this using the example of average meals per day. At the baseline average meals per day were about 2 per person per day, with a standard deviation of about 0.73. With full compliance and a baseline sample size of 750 individuals in the control group and 617 in the treatment group, the statistical power associated with detecting a 10% increase in meals per day was 99.9% (this is using a significance level of test, or alpha, of 5%). With incomplete compliance the statistical power of this experiment reduces to 82.4%. What this means is that with full compliance we would only have a one-in-a-thousand chance of rejecting the hypothesis that the treatment increased average meals per day by 10% if this were true; while with incomplete compliance, we would almost have a one-in-five chance of making that same error.

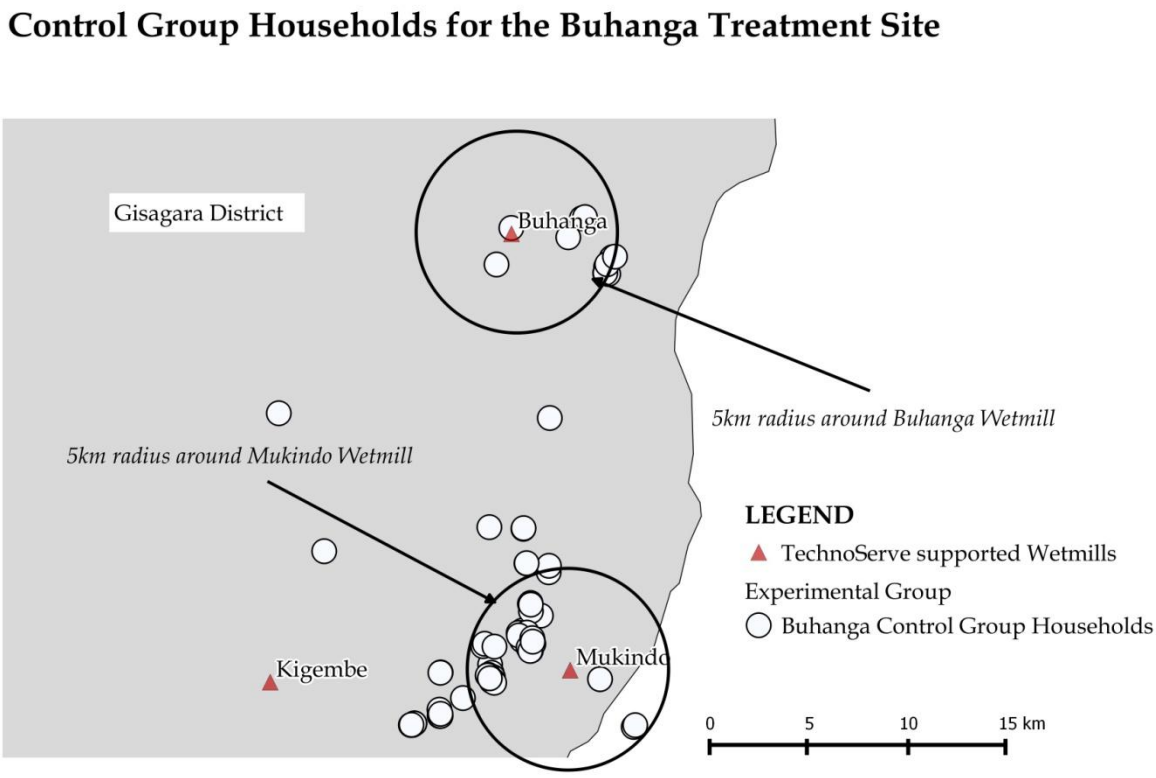
[1] Esther Duflo, Rachel Glennerster and Michael Kremer, 2007. "Using randomization in development economics research: a toolkit", Center for Economic Policy Research, Discussion Paper Series No6059, January 2007

Issue 5: Treatment and Control Group households were affected by other TechnoServe supported wetmills/Cooperatives

Adding to the issue of incomplete compliance is the fact that some control group households benefited from other TechnoServe supported wetmills/cooperatives in the vicinity of the treatment sites. Between 2009 and 2013, TechnoServe supported a total of 35 wetmills throughout the country, some very close to the treatment and control sites for this experiment. The Agronomy program was also rolled-out to a total of 32 cooperatives between 2009 and 2012 and benefited an estimated 30,000 farmers. With such an outreach at the national level, it is not surprising that some control group members would also benefit from other TechnoServe supported treatment sites.

We illustrate this "contamination" issue using two examples. The first relates to control group members for the Buhanga cooperative in Gisagara District. Figure 5 reveals that control group members for the Buhanga cooperative were selected from coffee growing areas that were in the direct vicinity of another TechnoServe supported wetmill called Mukindo. We can see that a significant share of control group households lived within a 5km radius from Mukindo wetmill. The likelihood that these households would have benefited from TechnoServe activities in and around the Mukindo wetmil - that had been ongoing for at least one and a half year at the time of the survey - is very high.

Figure 5: Control Group Households for Buhanga (Gisagara District) Treatment Site

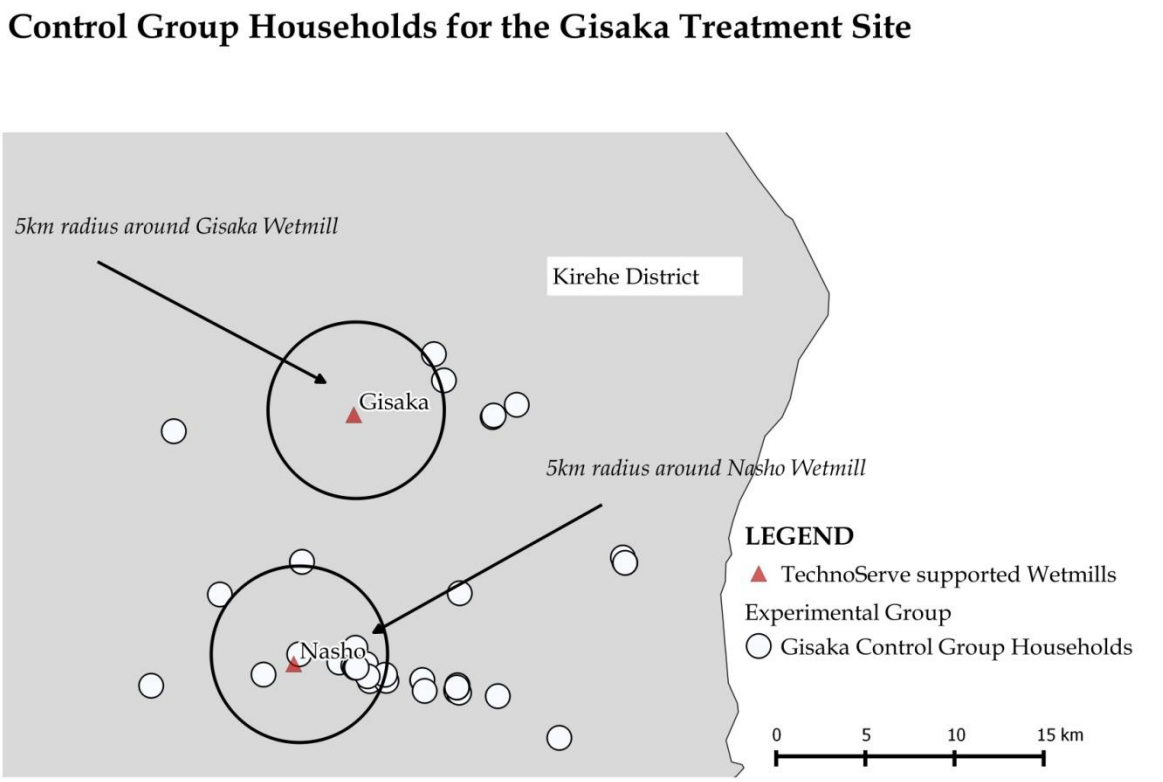


Another case where the risk of contamination was high is in the Kirehe district. Control group members associated with the Gisaka cooperative in Kirehe were also in the direct vicinity of another TechnoServe supported wetmill called Nasho. Figure 6 reveals that a significant number of the Gisaka control group households lived within the 5km radius from the Nasho wetmill. With more than 1,200 coffee farmers registered in TechnoServe's agronomy program in Nasho alone, the chance that these

control group members would be affected one way or another – through direct treatment or spill-over effects - is high.

This kind of "contamination" further reduces the statistical power of this experiment. Furthermore, in the survey, all farmers were asked whether they belonged to a TechnoServe supported wetmill or not, but some were not aware of the TechnoServe brand so answered no even if they were linked to a TechnoServe wetmill. This is an issue that was raised by the enumerator teams in the daily debriefs. The prevailing concern in this case is that we do not think we have fully captured this effect in the endline data-set.

Figure 6: Control Group Households for Gisaka (Kirehe District) Treatment Site



A related issue is that treatment farmers also benefited from other TechnoServe supported wetmills/cooperatives. These include Abangakurushwa, Coocafe, Gishyita, Koakagi, Kobakanya, Nyagihanga, Rusenyi, and Tuzamburubukungu, which were all not part of the experiment. In the case of the treatment group, this is less of an issue for the simple reason that these households were supposed to benefit from TechnoServe's Wetmill and Agronomy programs anyway. Nevertheless, the fact that this treatment might have come from a different treatment site blurs the lines of what effect we are actually measuring. What we call the treatment effect here is a medley of different levels of intervention, in different places and at different points in time.

Issue 6: High attrition from the survey

This issue of reduced statistical power is further exacerbated by the high attrition rate. Here we refer to attrition as targets in the treatment or control group that could not be reached in the endline (and/or midline) or that refused to co-operate and respond to the questionnaire. The attrition rate in the endline is 26% of households compared to the baseline in 2009 and about 13% compared to the midline in 2011. Attrition is a tricky challenge to handle technically as it can introduce bias into the composition of the treatment and control groups when it is not random (e.g. when attrition disproportionately affects certain types of people/groups).

In this case we find sufficient evidence to reject the hypothesis that attrition was random. Nevertheless, the imbalances introduced by attrition seem unlikely to have biased results significantly.

Attrition did not affect the treatment or control groups disproportionately and was relatively similar across cooperatives, except the Musha, Giseke and Nyarubaka areas, which were outliers (we could not find a valid explanation as to why Musha, Giseke and Nyarubaka were outliers). Attrition in the treatment group was 27.2% compared to 25.1% in the control, a difference that is not statistically significant. Table 6 summarizes attrition levels by cooperative.

Table 6: Attrition Levels by Cooperative

Cooperative	Attrition
Buhanga Coffee	24.8%
Cafeki and Mizero	25.1%
Gashonga, Mwezi and Gisuma	22.3%
Gisaka Coffee	23.7%
Giseke and Nyarubaka	32.1%
Musha Coffee	9.2%
Shara Coffee	22.5%
Vunga and Matyazo	24.4%

The two main characteristics for which we observe a statistically significant difference between households that dropped out of the survey and households that did not are age and number of household members. These are intricately inter-related – older people have larger households on average – and controlling for one cancels the effect of the other out. The heads-of households that dropped out were on average younger (47 years versus 51) and came from households with fewer members (5.4 versus 6.4 on average). Figure 7 plots the V-curved relationship between age and attrition; figure 8 plots a similar relationship between the number of household members at the baseline and attrition. In both cases attrition decreases with age or number of household members up to a certain cut-off point, and then slightly increases again. The slight increase after age 65 is most probably due to age-related issues - including death - that prevented older household-heads from responding to the questionnaire. Older heads-of-household also had larger households, hence the slight increase in the attrition rate for families with 10 or more household members at the baseline. The interesting point however is that attrition was significantly higher for households with younger heads-of-household and fewer household members: 41% for 20-29 year olds, compared to just 19% for 50-64 year-olds; and 54% for households with only 1 or 2 members, compared to 14% on average for households with 8 or 9 members. We interpret this negative relationship between attrition and age or attrition and number of household members as a direct result of mobility. Younger household-heads with smaller households seem to be more mobile, which would explain why it was more difficult to locate them during the survey period. Although we do not have data to show this, we

assume that this is a reasonable assumption to make - they are more likely to travel more, work in multiple jobs – potentially also off-farm, relocate to cities, etc.

Figure 7: Relationship between age and attrition

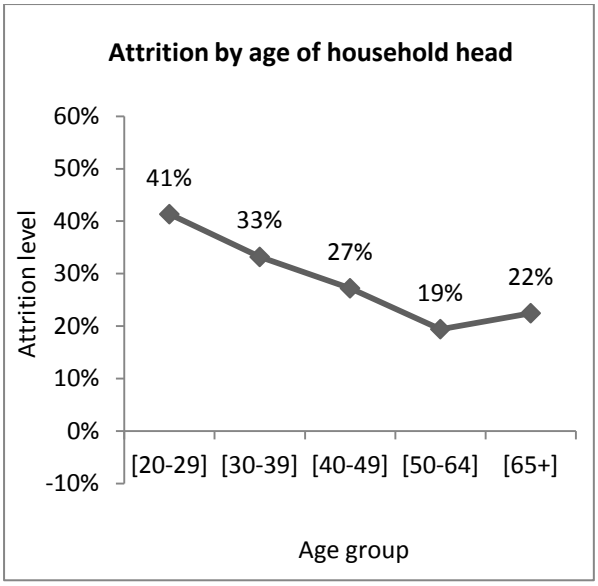
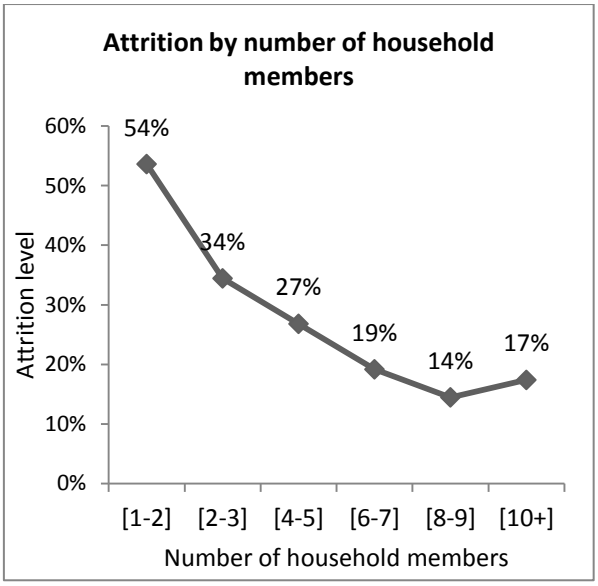


Figure 8: Relationship between attrition and number of household members



Aside from these imbalances in the age of the household-head and the number of household members, we do not find very large baseline differences between households that did not drop out of the survey and households that did. They are equi-distant from the treatment site and from other facilities (e.g. tarmac road, bank, public transport, etc.) and there are no clear patterns of socio-economic differences between them (see Annex 2).

These comparisons suggest that while attrition was not random it is unlikely to have biased the results of this evaluation significantly one way or the other. We nevertheless propose to use weights to correct for the age-related imbalances created by attrition. What we cannot change unfortunately is the fact that attrition significantly reduces the statistical power of this experiment, which means that we are more likely to reject a given hypothesis even if it is true. The smaller the treatment effect, the less likely we are to detect it.

Issue 7: Data quality and comparability over time

One additional impediment to the successful roll-out of this evaluation is data inconsistency over time. Most of these inconsistencies are due to the fact that three different agencies/consortiums were involved in data collection - one at the baseline, another at the midline and yet another at the endline. The initial research team that was overseeing this project from TechnoServe's side has also moved on, resulting in little institutional memory about the details of how the evaluation was conducted, the reliability of some of the indicators, how the data was entered, cleaned, etc. There are three main issues that affect the quality of data over time: (i) different approaches to data entry; (ii) a different research instrument; and (iii) respondent-driven error leading to time-inconsistencies, in particular for the age variable. These issues undermine the accuracy of comparisons over time, including difference-in-difference estimators calculated using 2 or 3 time periods.

Different approaches to data entry. We do not have a clear understanding of how data was cleaned in the baseline and midline. For some indicators at the baseline, for example, we are unable to

distinguish between missing variables and zeroes. These include indicators such as sickness (whether a member of the households has been sick in the past year), up-take of credit or a savings account, ownership of assets such as livestock, household expenditures, etc. Although we have re-cleaned baseline and midline data to adjust for these differences across years, we cannot exclude the possibility that there might be data-entry induced errors in the estimates. Moreover, for questions where we find that the data is not comparable over time due to differences in data-entry methods, we make use of one-period estimation techniques rather than difference-in-difference methods.

Different research instrument. In conjunction with TechnoServe, a decision was made to conduct the endline using a much shorter and refined questionnaire than the original questionnaire used at the baseline and midline. The time required to deliver the questionnaire was cut in half, from about 3h to 1.5h. The initial questionnaire focused not only on socio-economic indicators of interest, but also on covariates relating to agriculture, coffee production and household income. While most of the socio-economic questions were kept unchanged, some that were deemed confusing or misleading were re-adjusted. This means that on some indicators – e.g. decision making in the household – it is not possible make any comparison over time. For these indicators we can only rely on alternative estimation methods (e.g. matching, regression discontinuity, instrumental variables) to measure the treatment effect.

Respondent-driven errors. The questionnaires that were used at the midline and endline were pre-populated. There was one unique questionnaire sheet per household, which contained the household ID, the name of each household member, their corresponding identification number, and some additional details such as their gender, relationship to the household head, etc. One detail that was not contained in this pre-populated sheet, but that was asked each time, was the age or year of birth of each individual household member. This makes it possible for us to test the consistency of responses over time on age, which is a key covariate of interest given that so many factors (e.g. education, nutrition, etc) are very sensitive to age. While on average the results appear to be perfectly consistent, a careful look at the breakdown of the age variable is not very encouraging. If we say a given person's age is correctly specified if it is within two years of what we would expect it to be given that person's age at the baseline (which is a relatively large margin) we find that 56% of people's ages were correctly specified at the midline and just 38% at the endline. The standard deviation of ages around the expected value is in both cases 6 to 7 years. Part of this can be explained by the fact that people round up age to the nearest 5-year mark (e.g. 5, 10, 15, etc.) and some of these differences could simply be due to data entry mistakes. A more likely explanation is that respondents felt uncomfortable sharing precise personal details about other household members, leading to inaccuracies. This kind of caution with personal details is something we have encountered in other surveys in Rwanda as well.

We do not know which of the three estimates of age – at the baseline, midline or endline - is the most accurate. We attempted to adjust the age variable by anchoring age either at the baseline or the endline and then calculate forwards or backwards, depending on whether age at the midline was more similar to the baseline or the endline. However, this introduced additional biases especially when we focus the analysis on a specific age sub-group (e.g. for education or nutrition). For this reason, we have not adjusted the age estimates.

.

3. Methodology

For all the reasons listed above, we have to be extremely careful with how we go about estimating the treatment effect, if it is even possible to do so. The theoretical base underpinning this entire evaluation is very fragile: we have selection bias, interference from alternative TechnoServe programs, incomplete compliance within the program, attrition, multiple treatments at different points in time and in different places, time-inconsistencies in the data, and as a results of all of this significantly lower statistical power than initially thought. No estimation method can overcome this entire set of constraints at once. Given that we cannot rely on the estimates of one-single-method, as there will always be an explanation as to why a particular identification assumption is likely not to hold, we decide to explore several different alternatives to calculating the effect of the program. If all the estimators pull in one direction it can give us a signal, albeit not statistically significant or conclusive, about what the effect of the program was or might have been.

The alternative approaches we propose are well established methods that are meant to attenuate the effect of some of the design issues discussed above, in particular selection bias, incomplete compliance and time-inconsistencies in the data. These are second-best solutions, less efficient than a perfect experiment, but that enable exact inference **if the underlying identifying assumptions are met**.

3.1 What are we measuring?

The objective of this impact evaluation is to measure to what extent TechnoServe's interventions in treatment areas have impacted socio-economic indicators, ranging from education, health, nutrition, to access to finance, housing, assets ownership and gender balance. We are measuring the effect of two sets of complimentary programs: (i) the wet-mill program, which aims to organize farmers into cooperatives and support them in setting-up and running a wet-mill to upgrade from semi-washed to fully washed coffee; and (ii) the agronomy program, which trains farmers on certain agronomical best practices in coffee growing. The treatment varies by household: some households in the treatment areas did not participate in any program, some participated in only one of the two programs, while others participated in both. Moreover, as explained earlier, the treatment was not synchronized across cooperatives. This was the case of the agronomy program, which was delayed in three of the treatment cooperatives. Despite these differences between interventions and the timing of thereof, we measure impact using a binary variable, *Treatment*, that takes the value 1 when a household belongs to the treatment group and 0 when the household belongs to the control group. What we are measuring is therefore not a uniform treatment, but rather a combination of different levels of treatment delivered at different points in time.

In the context of imperfect compliance, we conduct the analysis at two levels. For all variables of interest we provide:

(i) Estimates of the *Intention-to-Treat* (ITT) effect, which is the average effect of the treatment on all households in the treatment group, regardless if they were treated or not (the intention was to treat them) or if control group households were treated or not (the intention was not to treat them). The ITT measures not the average effect of the treatment, but rather the effect of having a higher probability of receiving the treatment (i.e. assignment to the treatment group). Assignment to the treatment group was not always successful, as some treatment group members did not take the treatment, whereas some

control members did. As such it is easy to see why this will be an under-estimation of the actual treatment effect.

(ii) An estimate of the effect of the intervention on households that were actually treated, which for simplicity we will call the *Average Treatment Effect on Compliers* (ATEC).

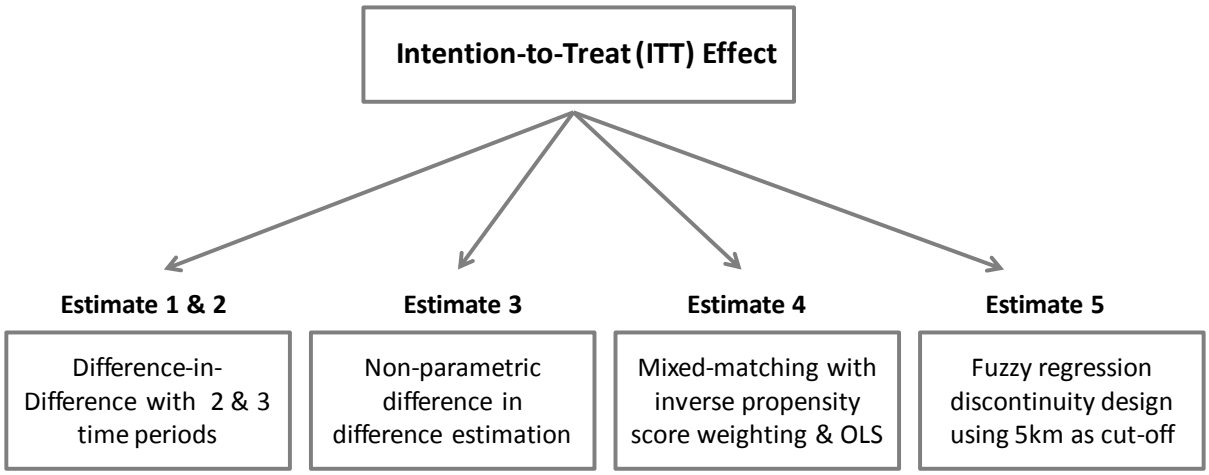
3.2 How do we measure these effects?

Estimating the Intention-to-Treat Effect

Starting with the Intention-to-Treat Effect, we propose to use multiple program evaluation techniques to estimate the effect of the programs on treatment households. No single technique – including options such as matching techniques, regression discontinuity design, non-parametric regressions, etc – stands out as an ideal alternative for the purposes of this evaluation. All options in this case come with certain advantages and disadvantages. The objective in using different approaches is to compare the results we obtain for consistency and thereby increase the robustness of the conclusions.

We propose five different estimates of the Intention-to-Treat Effect (see figure 8). These include two difference-in-difference estimators - one with two time periods and the other with three (i.e. panel data) - and three alternatives that apply to only one time period.

Figure 8: Estimates of the ITT Effect



The five techniques we propose to focus on include:

(i) ***Difference-in-difference estimation with two time periods,*** controlling for baseline differences, weighting to adjust for attrition, and limiting the sample to an area of "common support";

(ii) ***Difference-in-difference estimation with panel data*** (baseline, midline and endline) over an area of "common support", controlling for baseline differences and weighting to adjust for attrition ;

(iii) ***A non-parametric machine-learning estimation technique,*** called Kernel Regularized Least Sqaures, over an area of "common support";

(iv) ***A mixed matching-methods approach*** combining propensity score weighting with OLS regression over a specific area of "common support"; and

(v) ***A fuzzy regression discontinuity design*** approach where we exploit the fact that control group members were selected from coffee growing areas at least 5km away from the treatment site. Before giving a brief overview of the methodology for each of these

techniques, we first define the propensity score and the region of common support and explain how we weight households to correct for imbalances created by attrition.

Two criteria that we will be using moving forward and that we define here are (i) comparison over the area of common support and (ii) the conditional independence assumption. In order to identify the area of common support, we first need to determine the propensity score (i.e. the probability that a household belongs to the treatment group), which enables the researcher to test whether the treatment and control groups are comparable across a number of characteristics. Measuring the propensity score enables the researcher to reduce the sample to include only households in both groups that are comparable. This is what we refer to as the "area of common support". We provide details in Section 3.3 and 3.4 below.

Secondly, all of these estimation methods - except the fuzzy regression discontinuity design - require the ***conditional independence assumption*** to be met, which states that controlling for certain covariates, the assignment of households to the treatment and control groups is "as good as random". Given good baseline location data and a variety of possible controls, this should in theory be a reasonable assumption to make. The main source of selection bias at the baseline was that enumerator teams ended up recruiting more control households in the vicinity of main roads and towns rather than in less connected locations. This could have been a deliberate strategy, we cannot know for sure, but could also simply be a reflection of the fact that given the time allocated enumerator teams were more successful in recruiting control households in areas close to roads and towns. These are typically more dense than in rural areas where the hilly terrain and the sparse housing makes it difficult and time-consuming for enumerators to move between households, recruit and effectively implement a "random walk" approach. "Random walk" techniques are now widely accepted as suboptimal assignment techniques that are rarely fully random [Duflo et al, 2008]. Here we assume that controlling for distance from a tarmac road, distance from a town (which we proxy for using distances from various facilities, e.g. a bank, public transport, etc.) and differences between cooperative groups, the assignment of households to the treatment or control groups was ***"as good as random"***. If the experimental protocols were deliberately not observed at the baseline – which is difficult to ascertain even though GPS data suggests that in some areas this might have been the case - then this assumption does not hold and would also mean that it is impossible to draw any meaningful inference from this program evaluation. One way to test whether this assumption holds is to look at the consistency of the estimators and whether they make sense.

3.2.1 Estimating the Propensity Score

In the context of program evaluation, what researchers refer to as ***the propensity score is the probability that an individual - or in this case a household - belongs to the treatment group***. Sometimes, this probability is known in advance. For example if by design TechnoServe had randomly selected 50% of treatment households from a pool of households headed by women and the remaining 50% from a pool of male-headed households, then we would know that the propensity score for both female- and male-headed households was 50%. In most cases however the propensity score is not know in advance and needs to be estimated from the data. There are a number of ways to calculate the propensity score, but the most commonly used technique is to predict the probability of belonging to the treatment using a simple logistics regression where the dependent variable is the *Treatment* dummy.

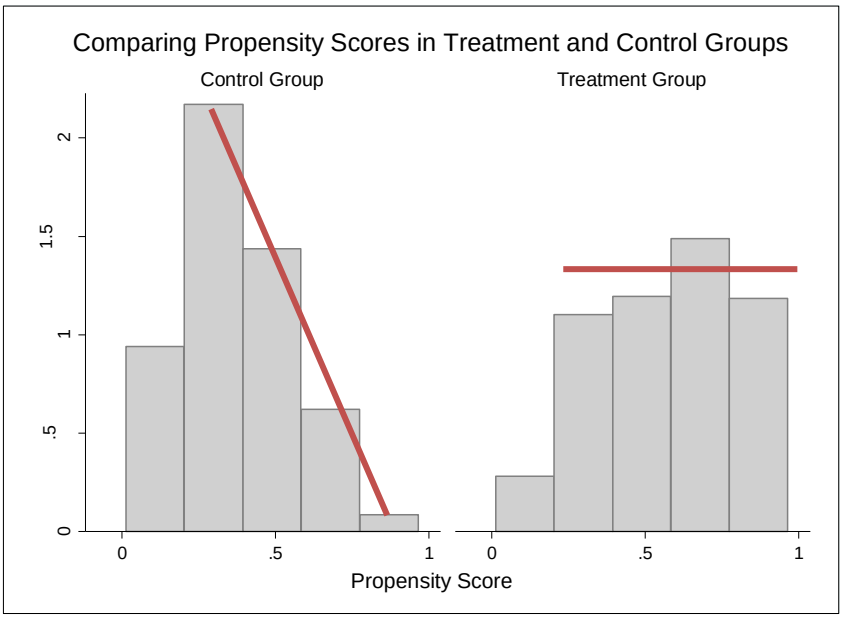
Here we estimate the propensity score at the baseline using the following equation which we estimate using a logistics regression:

$$Treatment_t = B_{0,t} + B_{1,t}Individual\_Cov + B_{2,t}Household\_Cov + B_{3,t}Location\_Cov + \varepsilon$$

In this equation *t* represents the baseline year (in this case 2009); $Treatment$ the treatment dummy which takes the value 1 if the household belongs to the treatment group and 0 otherwise; $Individual\_Cov$ represents covariates relating to the household head (including age, baseline education level, marital status, gender of household head); $Household\_Cov$ covariates relating to the household (including number of household members, whether coffee is the main source of income or not, whether the household has recently increased coffee acreage or not); and finally $Location\_Cov$ includes variables that capture the location of the households (including distance from the nearest tarmac road, a bank, public transport, and hospital as well as the cooperative areas the household belongs to). Together these baseline variables should yield a relatively good estimate of the probability that a certain household belongs to the treatment group or not. Details of the regression can be found in Annex 3.

But why measure the propensity score in the first place? The first answer is that the propensity score enables the researcher to test whether two groups – here the treatment and control - are comparable across a number of characteristics. If this program evaluation was perfectly random – i.e. there was no selection bias in the assignment of households to the treatment or control groups – we would expect the probability of a household ending up in the treatment or the control group to be the same on average. Figure 9 compares the distribution of the propensity scores in the treatment and control groups within the context of this program evaluation. The figure clearly shows that the distribution of the propensity score in both groups is different. Control group households have a much lower propensity score on average. This is simply another way of showing that the treatment and control groups had significantly different baseline characteristics and that there was selection bias in the assignment of households to the treatment and control groups.

Figure 9: Comparing Propensity Scores in Treatment and Control Groups



3.2.2 Defining the area of common support

The second reason why measuring the propensity score is useful is that it enables the researcher to reduce the sample to include only households in both groups that are comparable. This is what we call

the ***area of common support***. Households are comparable if they have a relatively similar propensity score – i.e. if the probability of them having been assigned to the treatment group was similar. If, for example, there were many treatment group households with a propensity score of 80-100%, but no corresponding control households, we would have to exclude all households with a propensity score between 80-100% from the area of common support. There would be no households in the control group to "support" a comparison with the selected treatment households. The intuition behind the area of common support is that it is one way of reducing selection bias. We exclude households for which there is no equivalent in the opposite group. By doing this we lose statistical power in terms of sample size, but reduce bias. This is the cost of having selection bias.

Figure 10: Area of Common Support & Propensity Scores

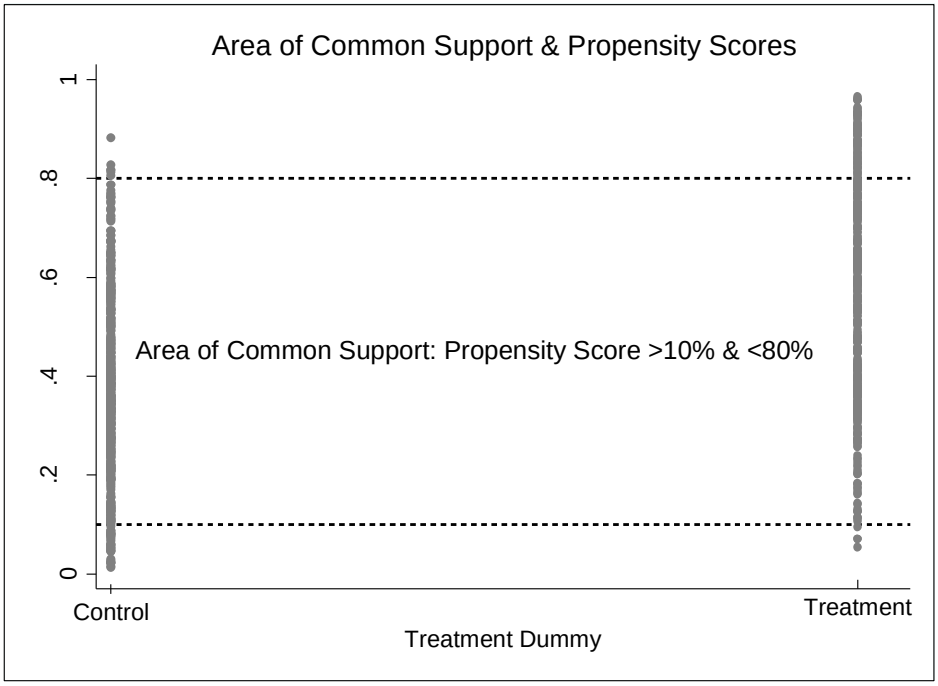


Figure 10 plots household propensity score in the treatment and control groups in a way that makes it intuitive to compare propensity scores across groups and identify the area of common support. As can be seen, in this case, there are just 3-4 treatment households with a propensity score of less than 10% and 3-4 control households with a propensity score of more than 80%. In both groups however, there are many households with a propensity score between 10% and 80%. In other words for any treatment household with a propensity score between 10% and 80%, we can find a corresponding control group household with a relatively similar propensity score. We therefore define the area of common support as all household with a propensity score between 10% and 80%. We will use this definition for the first four estimations of the Intention-to-Treat effect.

The third reason propensity scores are useful is that they enable a number of different matching estimation techniques designed to overcome selection bias at the baseline. Here we will be using the inverse propensity score matching combined with OLS regression as one of the possible alternatives to a simple difference-in-difference approach (see Section 3.2.5).

Dealing with attrition – weights to correct for age and family size imbalances

In the section on design-related issues we established that attrition from the sample was not random and affected household with younger household-heads and fewer members more. While this does not

seem to have affected the treatment or control groups disproportionately or, for that matter, created imbalances in the socio-economic outcome variables of interest, it is nevertheless something we need to adjust for. We do this by applying weights such that the share of households with X members in the midline and endline is the same as in the baseline. By adjusting for number of household members, we automatically also correct for age of the head-of-household. The table in Annex 3 summarizes the weights applied based on the size of the household. These weights only slightly adjust results and do not create any major distortions. We apply these weights for estimations 1, 2, 4 and 5.

3.2.3 Estimations 1&2: Difference-in-difference estimators with two and three time periods

Difference-in-difference (DID) estimators measure the differential change in outcomes over time in the treatment and control groups. If assignment to the treatment and control groups is perfectly random, then any incremental change (positive or negative) in the outcomes of the treatment group compared to the control group can be interpreted as the average treatment effect of the program. In other words, had there been no intervention, the treatment and control groups would have evolved in parallel. Here, given that we are working with selection bias, this does not work as differences between the treatment and control groups could be affected by other non-program related effects.

However, we assume that controlling for location indicators (such as distance from a tarmac road and distances from various services usually available only in towns or larger settlements – e.g. banks, hospitals, public transport services) and other individual and household covariates of interest, the assignment of households to the treatment and control groups is as good as random. Furthermore, we decide to focus this analysis on the sub-sample of households that belong to the area of common support in order to slightly increase the comparability of the treatment and control groups. Lastly, we cluster standard errors at the cooperative level as observations for households from the same cooperative are likely to be correlated in time.

The first difference-in-difference estimator is based on two time periods, comparing endline values in the treatment and control groups to the baseline. We measure it using a linear regression (or linear probability model with heteroskedasticity-robust standard errors in the case of binary variables) based on the following equation:

$$Y = \beta_{0,t} + \beta_{1,t} Period_t + \beta_{2,t} Treatment + \delta Period \times Treatment + \gamma_i X_i + \varepsilon$$

where Y is the dependent variable, $Period$ is a time dummy that takes the value 0 in 2009 (at the baseline) and 1 in 2013 (at the endline), $Treatment$ is the treatment dummy, δ the coefficient that multiplies the interaction term $Period \times Treatment$ and is the difference-in-difference estimator, and finally X_i a list of covariates to control for baselines differences between treatment and control. We focus throughout this exercise on three sets of controls:

- ***Individual controls*** (including age, gender, marital status, education and occupation);
- ***Household level controls*** (number of members, whether coffee main source of income and whether the household expanded coffee acreage in the three years prior to the intervention); and finally,
- ***Location controls*** (cooperative area, distance from tarmac road, form a bank, from a hospital and from public transport).

The second estimator is based on a similar difference-in-difference strategy, only this time using three time periods instead of two: the baseline, the midline and the endline. Panel data methods are usually preferred as they enable us to control for time-invariant effects (even when they are unobservable).

Again we weight to adjust for the effect of attrition, cluster standard errors at the cooperative level and focus the analysis on the area of common support. We estimate the difference-in-difference estimator using a pooled OLS regression with random effects. We test for the random effects hypothesis using a Hausmann test – if the test fails then the random effects estimator is deemed not consistent and we do not report it. The estimation is based on the following equation:

$$Y_t = \beta_{0,t} + \beta_{1,t} Period2_t + \beta_{2,t} Treatment + \delta_t During_t \times Treatment + \gamma_{i,t} X_i + u_t + \varepsilon_t$$

where $Period2$ is a time variable that takes the value 0 at the baseline, 1 at the midline and 2 at the endline; $During_t$ a dummy variable that takes the value 0 at the baseline, and 1 during and post intervention in the midline and endline; δ_t the coefficient that multiplies the interaction term $During_t \times Treatment$ and is the difference-in-difference estimator at time *t*.

Under the assumption that controlling for all the selected covariates (denoted X_i), the assignment of households to the treatment or control groups is random and **that in the absence of the intervention the treatment and control groups would have evolved in parallel**, these difference-in-difference methods provide an unbiased estimate of the Intend-to-Treat Effect. We test and discuss this hypothesis in the results section. The other issue with using two or three time periods is that data was not collected or entered consistently across years – see section 2 above for a discussion of time-inconsistency issues. This is part of the reason why using alternative estimators – based on one time period - to check for the consistency of the results we obtain is an attractive option in this case.

3.2.4 Estimation 3: Kernel Regularized Least Squares

The third estimate will be based on a non-parametric machine learning method called Kernel Regularized Least Squares (KRLS). It is a very new technique developed by Hainmueller and Hazlett in 2013 that enables us to tackle inference problems in a way that is both easy to interpret and not based on any assumptions of additivity or linearity[2]. This method, not to be confused with non-parametric "kernel regressions", is not based on strong parametric assumptions and yields estimators that are unbiased and consistent. Non-parametric methods are a good alternative to matching when there are baseline differences in the covariate distributions of the treatment and control groups as linear regressions, which rely on extrapolations, can be sensitive to these differences. Even though this method is relatively new and does not yet permit heteroskedasticity-, autocorrelation-, and cluster-robust estimators for standard errors, we propose to use it instead of established non-parametric estimation methods - such as kernel methods proposed by Heckman et al (1997, 1998) or other estimators proposed by Imbens et al (2005) and Chen et al (2005) – which tend to be more difficult to implement and interpret. We use the KRLS estimator here as an additional point of comparison to the difference-in-difference estimators, the matching estimators and the fuzzy regression discontinuity estimators.

The Intention-to-Treat estimator in this case is: $E[Y|T = 1, X_i]$- $E[Y|T = 0, X_i]$, which is the expected difference between the outcomes of the treatment and control groups, keeping all the X_i covariates constant. As in the case of the difference-in-difference estimators, we measure the KRLS estimator on the sub-sample of households that belong to the area of common support. Due to software limitations we cannot weight results to adjust for the effect of attrition.

[2] Jens Hainmueller and Chaz Hazlett (2013)

3.2.5 Estimation 4: Mixed matching methods with inverse propensity score weighting

Inverse Propensity Score Weighting (IPSW) is an established matching technique first introduced by Imbens et al (2000)[3]. As the name suggests the method consists in weighting the treatment and control groups by the inverse of the probability of being included in the treatment group. It is a slightly atypical matching technique in that it does not compare outcomes for an individual in the treatment group to one or multiple similar individuals in the control group; rather it balances out treatment and control on average, such that after applying the weights the treatment and control groups are similar on average. To achieve this balance, the weighting for the treatment and the control groups needs to be slightly different. Treatment group households are weighted by the inverse of the propensity score ($\frac{1}{e(Treatment)}$), while control group households are weighted by the inverse of one minus the propensity score (i.e. $\frac{1}{1-e(Treatment)}$). Otherwise the weights would just accentuate the differences between the two groups.

The advantage of this technique is that it can easily be combined with linear regression in what we call mixed-matching methods. In practice it translates into a simple linear regression with weights. The weights basically ensure that the covariates are uncorrelated with the treatment indicators, thereby making the weighted estimator consistent (see Imbens, 2000). Researchers call this a "doubly robust" technique as only one of the two estimation methods needs to be correctly specified for the method to yield consistent and unbiased estimators. We estimate the Intention-to-Treat effect using the following equation:

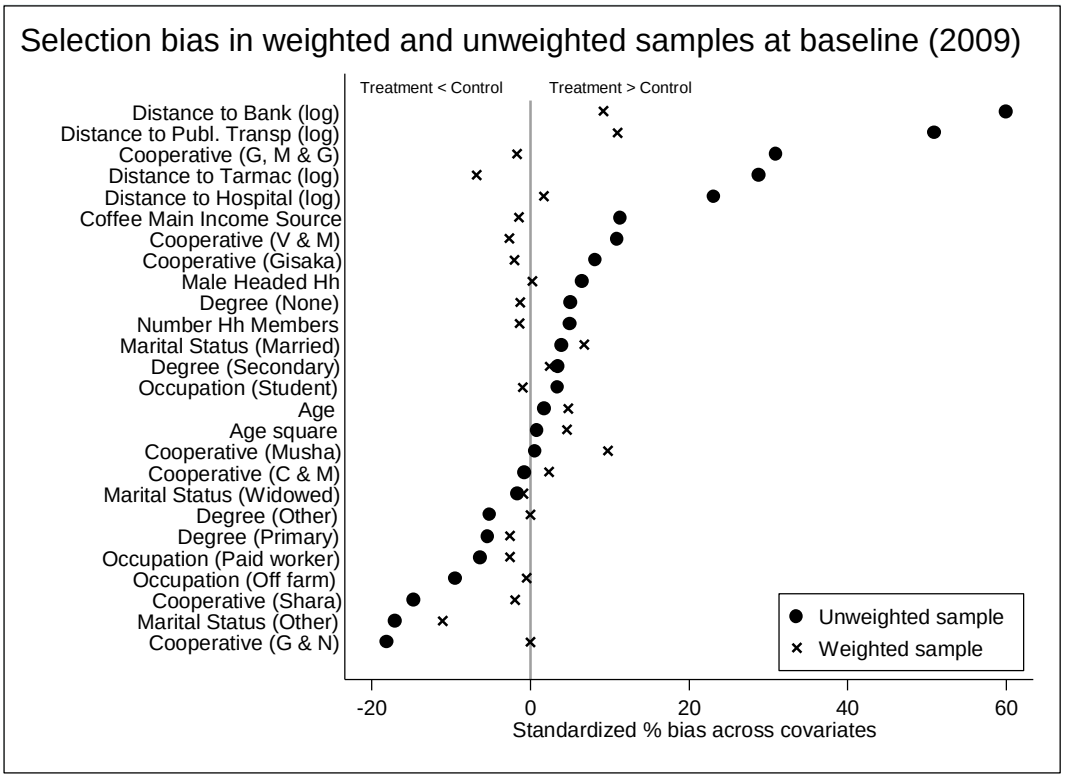
$$Y = \beta_0 + \delta Treatment + \beta_i X_i + \varepsilon$$

where δ is our Intention-to-Treat estimator and all the X_i are the individual, household and location controls. We limit this estimate to all households that belong to the area of common support, adjust the inverse propensity score weights so that they also account for attrition, and cluster standard errors at the cooperative level.

To illustrate how matching using the inverse propensity score can balance the treatment and control groups we compare average values in the treatment and control groups on the selected individual, household and location controls at the baseline. As can clearly be seen in figure 11, the weighted sample is much more balanced than the original sample. In the original sample, the biggest baseline differences are in the distance of households to various public services or infrastructure (e.g. bank, public transport, hospital, tarmac road) where the standardized bias across covariates reaches more than 60%. In the weighted sample these standardized differences are significantly reduced. Based on a Chi-Squared test of equivalence on the selected covariates, we can reject the hypothesis that the treatment and control groups are the same in the original sample with a very high degree of certainty (p-value<0.001). We cannot however reject the hypothesis that the treatment and control groups are the same of average in the weighted scenario (p-value=0.98). Inverse propensity score weighting reduces that average standardized bias on the selected covariates from 12.8% on average, to just 3.5%.

[3] Keisuke Hirano, Guido W. Imbens, Geert Ridder. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score". NBER 2000.

Figure 11: Selection bias in weighted and unweighted samples at baseline (2009)



In mixed-matching methods we rely on a linear regression with covariates to control for the remaining imbalances between treatment and control. Here we control only for baseline differences in location covariates, which are the main drivers of bias.

3.2.6 Estimation 5: Fuzzy regression discontinuity design

The final proposed Intention-to-Treat estimator will be based on a method that is called fuzzy regression discontinuity design. Regression discontinuity design methods allow for causal inference when the assignment to the treatment is partially or completely determined by the value of a certain variable (in the statistics jargon it is called the "forcing" variable) – i.e. when the probability of assignment to the treatment group jumps discontinuously at a certain cut-off point. Households above the threshold are assigned to one group (either treatment or control), while households below the threshold are assigned to the other. The basic idea is that households or individuals very close to that threshold – just above or below it - are likely to be very similar on average. Under certain assumptions, any discontinuity in the conditional distribution of an outcome variable at the cut-off point, can be interpreted as the treatment effect. In the case where assignment to the treatment group is fully determined by the cut-off point we talk about ***sharp regression discontinuity design***; when assignment to the treatment is only partially determined by the cut-off value of the "forcing" variable – i.e. there are non-compliers, potentially on both sides of the cut-off point – then we talk about ***fuzzy regression discontinuity design***.

The experimental set-up of this evaluation can be construed as an atypical form of fuzzy regression discontinuity design in that assignment to the treatment group was not determined by any variable in particular, but assignment to the control group was. Control group members were selected from beyond a 5km distance from the treatment site. On this metric, experimental protocols were quite strictly adhered to at the baseline. Calculations using GIS data suggest that the vast majority of control farmers (96%) were selected from beyond a 5km radius from the treatment site (see Table 7). This is

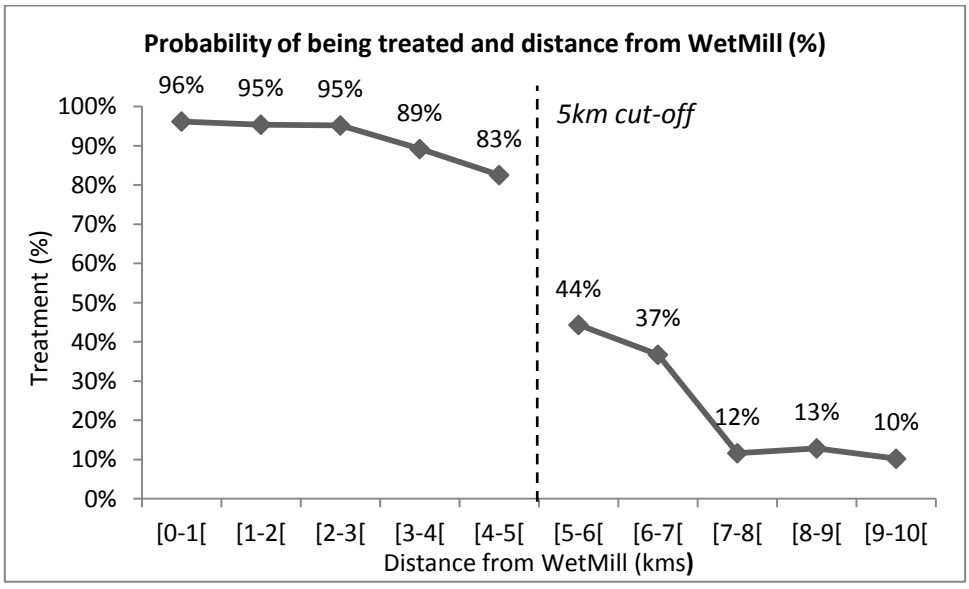
in contrast to more than 60% of treatment farmers who live within 5kms of the site. The median distance from the treatment site was 3kms for treatment farmers, compared to 9.6kms for control farmers. Treatment farmers were mostly clustered around the treatment site, while control group farmers were more spread out. One of the conditions required for a regression discontinuity design that the assignment variable (i.e. distance from treatment site) could not be manipulated, which is the case here.

Table 7: Distance from Wet-Mills

Distance from Wet mill	0-2 kms	2-5kms	5-10kms	>10kms
Treatment	31%	33%	16%	13%
Control	1%	3%	43%	43%

Even though there is non-compliance on both sides of the cut-off (about 40% of treatment farmers are beyond a 5km radius from the treatment site, and about 4% of control farmers live within 5km of the treatment site), there is a large jump in the probability of receiving treatment at the 5km mark (see figure 12). Households that live within 5km of the treatment site were on average 50% more likely to have been included in the treatment group, than households that lived beyond the 5km mark. We exploit this jump in probability at the cut-off point, to propose an estimate of the local treatment effect.

Figure 12: Probability of being treated and distance from the Wetmill (%)



The assumptions under which the fuzzy regression discontinuity design will yield a consistent local estimator of the Intention-to-Treat effect are called "monotonicity" and "excludability" (see Imbens and Lemieux, 2008). In the case of this set-up monotonicity would translate into the following assumption: crossing the 5km mark (i.e. moving closer to the treatment site) makes a household more likely to take the treatment. This seems to be a reasonable assumption to make: living 4.5km away from the treatment site rather 5.5km from the treatment is not going to make any less likely to take the treatment, on the contrary and the numbers show that. The "excludability" requirement would translate into the following assumption: being within a 5km range of the treatment site does not impact the outcome variable in any other way than through the receipt of the treatment. Here we would need to control for distance from various services (e.g. nearest bank, hospital and public transport point) and local infrastructure (e.g. tarmac roads) for that to be true. This is because we

know that not only are control households further away from the treatment site, but they are also closer to larger settlements and major roads.

We measure the fuzzy regression discontinuity design estimator using a two equation system as specified below:

$$Y = \beta_0 + \delta Treatment + \beta_1(Dist - 5) + \beta_i X_i + \varepsilon \qquad (1)$$

$$Treatment = \alpha_0 + \gamma Km5 + \alpha_1(Dist - 5) + \alpha_i X_i + \theta \qquad (2)$$

where $Dist$ is the distance from the Wetmill; 5km is the cut-off point; $Km5$ is a dummy which takes the value 1 when a household is within the 5km radius from the treatment site and a 0 otherwise ($Km5$ is what we call the "forcing" variable); and where $\delta \times \gamma$ is the Intention-to-Treat effect. This amounts to a two-stage least square estimation or what is otherwise referred to as an instrumental variable approach. Here we instrument $Treatment$ with the forcing variable, $Km5$, and weight to correct for the effect of attrition. Given that regression discontinuity design forces us to work with smaller sample sizes we control only for selected location covariates.

The last and possibly most tricky issue to deal with in regression discontinuity design is the selection of the appropriate bandwidth. The idea behind this whole method is to compare individuals or households that are just below or just above the cut-off, as the latter are likely to be very similar. The question is how far below and above the cut-off can we go? The answer is it that it is a question of balance. The smaller the interval to the left and right of the cut-off that we are working with, the smaller the bias but also the smaller the statistical power of the estimate. The larger the interval, the larger the bias, and the better the statistical power of the estimate. There is no optimal answer, but here for sample size considerations we decide to focus the analysis on an area of -4km/+4km from the cut-off. One important point to mention is that because we are focusing the analysis on a sub-sample of households within a certain distance from the 5km cut-off, this is a local treatment effect.

In theory, regression discontinuity design methods are thought to have very high internal validity. In practice we find that this estimator is quite jumpy compared to others and highly sensitive to bandwidth specifications. The reader should keep in mind that the fuzzy regression discontinuity design estimator reported here is a very local treatment effect - it is the effect of the treatment on households that are within a certain distance (or in statistical jargon "bandwidth") from the 5km mark. Given their proximity to each other and to the cut-off point these are expected to be quite similar on average.

3.3 Estimating the Average Treatment Effect on Compliers (ATEC)

The final piece of the puzzle is to provide an estimate of the average treatment effect of the program on the households that actually participated in the programs, received the treatment and were meant to receive the treatment (i.e. compliers). One option would have been to focus only on this estimate, because it relates directly to the effect of the program on people that received the treated and not on the less palpable concept of the Intention-to-Treat. However, given some of the design issues we have encountered and the potential for spill-over from households that received the treatment to households that did not, we are obliged to caveat the results. Moreover, focusing on the Intention-to-Treat estimates provides us with many more alternative estimation methods to test the robustness of the results obtained.

Moving from the Intention-to-Treat effect to the Average Treatment Effect on Compliers is not very difficult. The main challenge we need to overcome is the fact that there are bound to be baseline differences between compliers and non-compliers. We would expect for example that households in the treatment group that were given the opportunity to self-select into the TechnoServe program but decided not to had different characteristics to the potentially more motivated households that did participate in the program. Likewise, we would expect households that were in the control group but that nevertheless put in the effort to walk the extra mile and participate in the program, to be different on average than those that did not. Simply comparing the outcomes of households that received the treatment to households that did not would lead to bias, because these are likely to be very different households on average to start with.

The solution to this problem is an instrumental variables approach, where we instrument for receiving the treatment (*TNS*) with the treatment dummy (*Treatment*). The basic idea is to use the variation in the treatment dummy – which in a perfect experiment is fully random and therefore only affects the outcome through its impact on a certain household getting treated – to infer something about the causal effect of the actual treatment. This approach was first used proposed by Imbens and Angrist (1994) and can be summarized by the following WALD estimator:

$$ATEC = \frac{E(Y \mid Treatment = 1) - E(Y \mid Treatment = 0)}{E(TNS \mid Treatment = 1) - E(TNS \mid Treatment = 0)}$$

This estimator – which can be measured using an instrumental variables approach - consists of the Intention-to-Treat estimator, divided by the difference between the share of compliers in the treatment group (treatment group members that received the treatment) and non-compliers in the control group (control group members that took the treatment). For this estimator to be valid three assumptions are required:

- ***Assumption 1 – Independence of the instrument (1).*** The first condition under which an instrument is independent is when the "comparisons between outcomes for households exposed to different values of the instrument identify the causal impact of the instrument" (see Duflo et al, 2008). This condition is automatically met when the assignment to the treatment and control groups is fully random. Here we chose to weight by the inverse of the propensity score to ensure that this assumption is met. Under inverse propensity score weighting, we assume that the assignment to treatment and control groups is "as good as random".

- ***Assumption 2 – Independence of the instrument (2***): The second condition of independence, or what is often referred to as excludability, is that the instrument does not directly affect any of the outcome variables for non-compliers in the treatment group (i.e. people who did not self-select into the treatment). While we cannot test this assumption, we have reason to believe that belonging to the treatment group can also affect the outcomes of a household, even if the latter did not receive any treatment. We think this might be the case for two reasons: first because households in the treatment group are quite concentrated geographically around the treatment site; and second because of the nature of both treatments (the Wetmill program and the agronomy program). Spill-overs could have occurred through both programs: (i) from the Wetmill program because this intervention monetizes the local economy, connects it to sources of funding, and incentivizes farmers to enhance their coffee production; and (ii) through contagion/imitation of best coffee growing practices in the agronomy program. We have no way to check whether there are spill-over effects and how

large these effects are. For the purposes of this estimation we have to assume that these are minimal, but this is a potential issue that can lead to a biased Wald estimator.

- ***Assumption 3 – Monotonicity***. As in the case of regression discontinuity, the monotonicity assumption requires that the treatment dummy makes every household more likely to actually participate in the program. This is a reasonable assumption to make here and that is closely matched by the data.

Here we consider that receiving the treatment means any of the following:

- Being member of the cooperative that manages the Wetmill;
- Having received Wetmill training;
- Having received Agronomy training; or,
- Having sold coffee to a TechnoServe supported wetmill.

This is what we could call a "light version" of the treatment, because we cast the net wide and include households that might not have been very affected by the treatment. This would be the case for example of households that have only sold some coffee to a TechnoServe wetmill, even though they might not have been cooperative members and benefited in any other way from the program. This specification of treatment also includes households that benefited from one program but not the other. While a tighter definition of treatment might have been preferable (for example looking at just the households that benefited directly from both programs), the estimates obtained would not be accurate as they would further undermine the "excludability" requirement.

4. Results

With the methodological framework in place, we can now report results. We calculated all 5 Intervention-to-Treat (ITT) estimators and the Average Treatment Effect on Compliers (ATEC) for all variables of interest, grouped under the following headings: (i) education; (ii) nutrition and health; (iii) household decision-making; (iv) expenditures, savings and credit; (v) assets and infrastructure. This entailed running more than 250 different regressions. For all indicators, the idea was to compare the results obtained using each impact estimator and discuss implications.

The good news is that in the vast majority of cases the estimators yielded consistent results – i.e. the point-estimates had the same sign, with similar coefficients and levels of significance. It was not always possible to calculate all six estimators because of time inconsistencies, but for most variables this was not an issue. It is important to note also that given the breadth of the design issues and low statistical power, what we were looking for was not an exact estimate or level of statistical significance, but rather a broad signal as to the direction and scale of the program effect.

The bad news is that these estimators confirm our fears that there is too much baseline selection bias and contamination, which makes it impossible to make any causal inference with relation to the intervention. Even after multiple corrections, adjustments and the use of alternative estimation methods, most of the impact estimates are negative suggesting that the underlying identifying assumptions do not hold – these include the conditional independence assumption and the assumption that after controlling for distance and other covariates the treatment and control groups would have evolved in parallel in the absence of any treatment. The negative impact estimates are not consistent with the nature of the two interventions which we know from alternative work have led to an increase in coffee production and better prices for farmers. Rather they reflect the fact that control group households lived closer to main roads and urban settlements and were also exposed to the benefits of TechnoServe's Wetmill and Agronomy programs through alternative treatment sites that were not part of this evaluation. We have shown this already with maps and we will illustrate it again with data below. In our opinion, we have exhausted all possible options to correct for these unexpected biases, which were, to the best of our knowledge, not mentioned or identified either in the baseline or the midline studies.

In this section we present the results obtained for each of socio-economic areas of focus and make the case as to why identifying assumption do not hold. Finally we discuss what the characteristics are of farmers that ended up taking the treatment versus farmers that did not, a useful take-away from this experiment. Households that did take-up treatment had very different baseline characteristics to households that did not. Having an understanding of what these characteristics are can help TechnoServe better target households that are less inclined to take-up the treatment.

4.1 Results by socio-economic indicator

One way to test whether the underlying identification assumptions of an estimator are valid or not is to look at the results we obtain using that estimator. Based on readings from the six different estimators, we consistently find that for many indicators of interest the estimated program effect is inexplicably negative. We argue that the only possible explanation is that the underlying identification assumptions do not hold and that the adjustments and alternative estimation techniques adopted were insufficient to overcome selection biases at the baseline and other design and implementation related issues.

The set of indicators that most convincingly imply that baseline differences between the treatment and control groups are driving the estimates - rather than the effect of the treatment itself - are indicators relating to "access to utilities and housing quality". We start by presenting results on this set of indicators, before moving on to other indicators, which also confirm the hypothesis that the identification assumptions do not hold.

4.1.1 Access to utilities and housing quality indicators

In the section on design flaws and implementation issues, we showed that at the baseline there were significant differences between the treatment and control groups in terms of household infrastructure and access to services such as electricity and piped water. The impact estimates we present suggest that these baseline differences worsen over the duration of the program (see table 8). The ITT and ATEC estimators would imply that TechnoServe's Wetmill and Agronomy programs made households holds significantly worse off than they would have been had there been no treatment in terms of their quality of housing and access to utilities.

Here we define having access to electricity as any form of electricity – i.e. connection to the grid, a generator, solar power, a battery, etc.;. "piped water" as access to piped water from a protected source (e.g. covered well); we refer to "modern walls" as walls that are not made of bamboo or earth; "modern floors" as floors that are made of concrete; and "modern toilets" as toilets with a flush or a concrete floor.

The results obtained reveal that the same baseline differences we were trying to eliminate and circumvent using alternative identification techniques, persist at the endline. It does not matter how we specify the propensity scores, which covariates we control for, or what bandwidth we select for the fuzzy regression discontinuity estimator, these differences persist. This means there are "unobservables" driving the results. Here we find that belonging to the treatment group was associated with households being 4 to 18 percentage points less likely to have access to electricity, 2.8 to 22.4 percentage points less likely to use an electricity bulb; and 4.6 to 12.7 percentage points less likely to have good walls. The only indicator where the estimators yield a positive signal is on sanitation (i.e. toilet facilities).

Table 8: Access to utilities and housing quality indicators

Indicators	Intention-to-Treat effect					ATEC
	DiD (2)	DiD (3)	NP	IPSW	FRDD	
Access to electricity source (% households)	-5.4p*	n/a	-4.0p	-7.7p***	-18.0p***	-14.0p***
Lighting source=electricity bulb (% households)	-5.2p**	n/a	-2.8p	-5.6p**	-22.4p***	-10.5p**
Access to piped water (% households)	n/a	n/a	-0.1p	-1.1p	-8.9p**	-2.9p
Modern walls (% households)	-2.8p	n/a	-4.6p	-7.4p**	-12.7p*	-14.5p**
Modern floors (% households)	-1.9p	n/a	-1.9p	-2.1p	-10.8p*	-4.2p
Modern toilet (% households)	4.6p*	n/a	0.8p	1.8p	-6.9p*	3.8p

**p<0.1; **p<0.05; ***p<0.01 ; "p" stands for percentage points*

Color Legend:

	Significant improvement		Non-significant improvement		Significantly worse		Non-significantly worse

Not only were there baseline differences to start with, but these differences become larger over time. We explain this by the fact that control group households were more connected and closer to urban settlements, making them likely to benefit from investments and improvements in standards of living in and around these areas. This implies that the assumption that after controlling for certain covariates, the treatment and control groups would have evolved in parallel in the absence of any treatment does not hold.

Table 9 illustrates this with the example of access to electricity. Households in both treatment and control areas experienced a very rapid increase in their access to electricity over the course of the evaluation. This increase was however much larger in the control group. Access to a source of electricity in the control group increased from a base of 3.2% in 2009 to 23.1% in 2013; the equivalent increase in the treatment group was from 1.3% in 2009 to just 11.0% in 2011.

Table 9: Access to electricity indicator changes (baseline & endline)

Group	Baseline	Endline	Difference
Treatment	1.3%	11.0%	**+9.7p**
Control	3.2%	23.1%	**+19.9p**

4.1.2 Credit and Savings indicators

At a first glance the ITT and ATEC estimators for access to finance indicators, including access to credit, savings and use of mobile money, are non-conclusive (see table 10). Even though most of the point-estimates obtained are negative - pointing towards a "negative" program effect - none are statistically significant. Indeed, they would suggest that the Wetmill and Agronomy programs did not have a detectable effect on access to finance levels for members of the treatment group or more specifically program beneficiaries. This is in line with average access levels observed in both groups at the endline: in 2013 access to a savings account was slightly higher in the control group, 65.4% vs. 62.2%; access to credit was higher in the treatment group, 34.8% vs. 33.7%; while the use of mobile money was 18.7% in the control group, compared to 17.3% in the treatment group. None of these differences are statistically significant, even after controlling for covariates.

Table 10: Credit & Savings indicators

Indicators	**Intention-to-Treat effect**					**ATEC**
	DiD (2)	DiD (3)	NP	IPSW	FRDD	
Access to credit (% households)	1.3p	n/a	1.6p	-0.6p	-9.0p	-2.3p
Access to savings (% households)	-0.9p	-1.2p	0.7p	-0.6p	1.0p	-1.9p
Use mobile money (% households)	n/a	n/a	-0.4p	0.7p	-6.9p	-0.5p

**p<0.1; **p<0.05; ***p<0.01 ; "p" stands for percentage points*

When taking a closer look at the breakdown of these variables however, we see a different picture emerging. While the treatment and control groups had similar levels of access to finance overall, this access was different in nature. For example, out of households that had access to a savings account in the treatment group, 25% had their account in a bank, compared to 37% in the control group (a difference that is highly statistically significant). Likewise, 17% of treatment households got their loan from a bank compared to 22% in the control group. 24% of treatment group households used their credit for agriculture related expenses/investments, compared to just 19% in the control group. The amount of credit that households in the treatment group received was smaller on average. The median loan in the control group was RWF100,000 compared to RWF 84,600 in the treatment group. All of these statistics are consistent with the conclusion that households in the control group were more connected and closer to urban settlements (and had better access to credit/savings) than households in the treatment group.

4.1.3 Assets indicators

Except for bicycles, we find that being in the treatment group or benefitting from the Wetmill and Agronomy programs was associated with a decrease in the likelihood of households owning items such as radios, televisions, mobile phones and mattresses (see table 11). Estimators for which we obtained statistically significant differences included: (i) the ownership of televisions – belonging to the treatment group was associated with a 0.8 to 10.2 percentage point drop in the likelihood of households owning a TV set; and (ii) the number of mattresses in a household, controlling for number of household members and all the other individual, household and locational covariates of interest. We would expect households that live closer to main roads and cities to own more consumer goods such TVs, mobile phones, etc. than households in more rural areas. These point-estimates are therefore consistent with the hypothesis that the identification assumptions do not hold. Impact estimates for livestock indicators are more balanced, which is also to be expected if the treatment group is more rural (see table 11).

Table 11: Asset Indicators

Indicators	**Intention-to-Treat effect**					**ATEC**
	DiD (2)	DiD (3)	NP	IPSW	FRDD	
Household items						
Owns radio (% households)	0.2p	0.3p	-0.3p	-3.3p	-0.9p	-6.8p
Owns TV (% households)	-2.6p	-1.6p	-0.8p	-2.3p*	-10.2p**	-4.9p
Owns motorbike (% households)	-1.6p	n/a	-0.0p	-1.7p	-0.2p	-3.8p
Owns bicycle (% households)	0.8p	n/a	8.9p***	9.1p***	-5.0p	16.9p***
Number of mobile phones (# per household)	-0.06	-0.05	-0.00	-0.02	-0.21	-0.07
Number of mattresses (# per household)	-0.11	-0.04	-0.09	-0.18**	-0.60***	-0.40**
Livestock						

Number of hand hoes (# per household)	0.13	-0.02	0.04	0.12	-0.06	0.20
Number of cattle (# per household)	-0.18	-0.03	0.05	0.02	-0.26*	-0.01
Number of goats (# per household)	0.16	-0.07	0.08	0.10	-0.15	0.26
Number of chicken (# per household)	0.14	0.23	0.07	0.19	-0.20	0.41
Number of pig (# per household)	0.08	0.07	0.00	-0.17	-0.15	-0.23

**p<0.1; **p<0.05; ***p<0.01 ; "p" stands for percentage points*

4.1.4 Health and nutrition indicators

Results for health and nutrition related indicators tell a similar story. Most ITT estimators and the ATEC estimator show that within the framework of this experiment, belonging to the treatment group or receiving the treatment was associated with a slight decrease in the health status of households (see table 12). Although they are mostly negative, these differences are very small and not statistically significant. Endline values for health and nutrition related indicators in the treatment and control groups are almost identical: on average 79.7% of treatment group members have access to health insurance, compared to 79.8% of control group households; 27.6% of individuals in the treatment group were sick at some point in the 12 months prior to the endline survey, compared to 27.4% in the control group; average meals per day for a family in the treatment group is 1.997, compared to 2.001 in the control group. We also do not find any significant differences in the food intake patterns of treatment and control group households (see table 13). Signals that link control group households to slightly more urban settings – despite these similarities - include the fact that they were more likely to visit a hospital when sick than their counterparts in the treatment group (17% vs. only 11% in the treatment group) and less likely to report getting malaria when falling sick (48.5% vs. 43%).

Table 12: Health Indicators

Indicators	**Intention-to-Treat effect**					**ATEC**
	DiD (2)	DiD (3)	NP	IPSW	FRDD	
Health insurance (% households)			0.3p	-3.3p	-1.7p	-6.3p
Sick in past 12 months, 0 to 18 year-olds (% households)			0.5p	1.0p	-1.9p	1.8p
Meals per day, average household (# per day)	0.01	0.03	0.00	-0.03	-0.02	-0.07

**p<0.1; **p<0.05; ***p<0.01 ; "p" stands for percentage points*

Table 13: Nutrition Indicators

Indicators (endline values)	**Treatment**	**Control**	**Difference**
Meat (% households)	12.5%	`13.1%	-0.6
Starch (% households)	97.5%	97.4%	+0.1%

Vegetables (% households)	79.6%	84.5%	-4.9%
Fruit (% households)	33.0%	36.6%	-3.6%
Beans (% households)	94.0%	95.1%	-1.1%
Eggs (% households)	5.3%	4.6%	+0.7%
Milk (% households)	28.2%	26.5%	+1.7%

**p<0.1; **p<0.05; ***p<0.01; "p" stands for percentage points*

4.1.5. Household decision-making

We obtain negative estimates for household decision-making indicators, suggesting that belonging to the treatment group was associated with a decrease in female decision-making (see table 14). We ask respondents who in the household makes key decisions on certain items, ranging from decisions on food expenditures, to decisions on consumer durables, education, health, investments, etc. Responses can include one or multiple individuals, therefore also allowing for "joint" decision-making. These estimates show that female household members in the control group were more likely to make key decisions alone (compared to female household members in the control group). Some of the estimates obtained are statistically significant.

Table 14: Household Decision-Making Indicators

Indicators	**Intention-to-Treat effect**					**ATEC**
	DiD (2)	DiD (3)	NP	IPSW	FRDD	
Enrolment by age-group						
Female household member makes decisions on food alone (% households)	n/a	n/a	-0.5p	-4.0p*	-8.1p**	-6.5p
Female household member makes decisions on education expenditures (p households)	n/a	n/a	0.5p	-2.1p	-2.1p	-4.1p
Female household member makes decisions on clothing (p households)	n/a	n/a	-1.6p	-3.7p	-3.4p	-7.7p*
Female household member makes decisions on consumer durable items (p households)	n/a	n/a	0.2p	-2.6p	-6.3p*	-7.1p*
Female household member makes decisions on home purchase, improvement or repair (p households)	n/a	n/a	-2.4p	-3.2p	5.6p	-4.8p
Female household member makes decisions on surplus investments (p households)	n/a	n/a	-0.7p	-3.4p*	-5.5p*	-6.6p*
Female household member makes decisions on health expenditures (p households)	n/a	n/a	-0.4p	-2.7p	-4.2p	-5.0p

Table 15: Female Decision-Making Indicators

Female member decides about …	Treatment	Control
Food expenditure	29.2%	33.9%
Education	28.4%	28.5%
Clothing	28.4%	33.5%
Consumer durables	25.8%	27.5%
Home purchase, improvements and repair	25.1%	28.7%
Surplus investments	25.2%	28.3%
Health expenditures	26.9%	29.7%

4.1.6 Education indicators (for children/students aged 6 to 23)

The only set of indicators for which the ITT and ATEC estimators yield mostly positive are education indicators. These results do not imply any causality for a number of reasons: (i) the estimates obtained are not statistically significant; (ii) given that the identification assumptions do not hold, even if these results were statistically significant we would not be able to reject the hypothesis that results are driven by "unobservables" (e.g. disproportionate government investments in education in more rural areas); and lastly (iii) the "theory-of-change" leading to such outcomes in terms of education, but not in terms of other socio-economic indicators, is difficult to justify.

The ITT and ATEC point-estimates we obtain for education indicators suggest that belonging to the treatment group was associated with a 0.7 to 4.3 percentage point increase in the probability of children/students aged 6 to 23 being enrolled in school (either primary, secondary, or tertiary/vocational). These effects seem to have been slightly more pronounced for girls, according to difference-in-difference estimators (see table 16). As can be seen in figure 12, these positive estimates are not uniform across age-groups. For younger children aged 6 to 11 we do not find any major differences between the treatment and control groups. There is a gap in enrolment rates between treatment and control between the ages of 12 to 18, but we do not find a valid explanation as to why this might be the case. It is important to note also that enrolment-related estimates are highly sensitive to the age variable, which as we have seen is not very reliable.

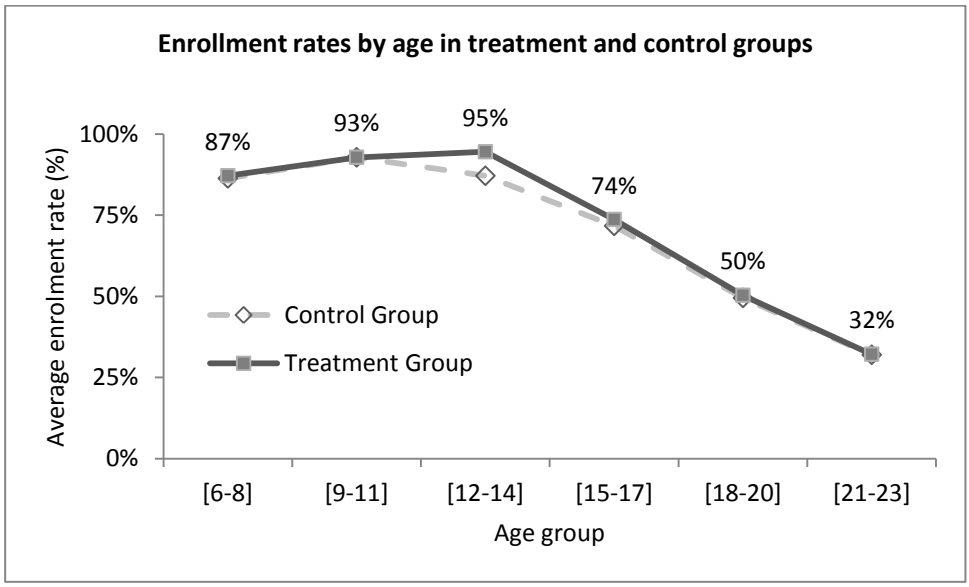
Table 16: Education Indicators

Indicators	Intention-to-Treat effect					ATEC
	DiD (2)	DiD (3)	NP	IPSW	FRDD	
Enrolment by age-group						
Enrolled, age 6-23 (p children/students)	4.3p	3.8p	0.7p	3.1p	2.4p	4.9p
Enrolled, girls age 6-23 (p children/students)	8.5p**	4.9p*	0.3p	2.0p	1.3p	3.7p
Enrolled, age 6-11 (p children)	1.6p	3.2p	-3.7p	-2.0p	-1.3p	-2.4p
Enrolled, age 12-17 (p teenagers)	7.9p*	3.6p	2.3p	4.3p	4.8p	8.3p
Enrolled, age 18-23 (p children/students)	10.4p*	3.6p	0.9p	2.9p	3.1p	5.6p

Net enrolment primary, age 6-12 (p children)	1.7p	1.0p	-2.6p	0.1p	0.7p	-2.8p
Net enrolment secondary, age 12-18 (p children)	1.8p	1.8p	-1.9p	1.5p	5.5p	2.1p
Days missed at school						
Child has missed days of school in last 3 months (p of enrollees)	n/a	n/a	0.2p	0.0p	-2.7p	-0.3p

**p<0.1; **p<0.05; ***p<0.01 ; "p" stands for percentage points*

Figure 13: Enrollment rates by age in treatment and control groups



4.2 Comparing compliers and non-compliers in the treatment group

An interesting question to ask at the endline - now that we know which treatment group members self-selected into the Wetmill and Agronomy program and which did not - is what the baseline characteristics of compliers and non-compliers were. We would expect compliers and non-compliers to have slightly different characteristics. Non-compliers might have not taken the treatment because coffee was not an important crop for them, because they were less motivated, or maybe because they lived further away from the treatment site, etc.

Baseline data shows that self-selection into the treatment group had very little to do with location data and much more to do with the status of coffee production and the socio-economic status of households (see table 17 below):

- In terms of coffee production, the strongest predictor of whether a household would end-up taking up the treatment or not was not whether coffee was their main source of income, but rather whether coffee acreage in the years leading to the program had increased or decreased. As would be expected, households that had invested in the expansion of their coffee acreage in the three years prior to the program were more likely to take-up the treatment than households that had

divested: 85% of households that were investing in expansion took-up the treatment, compared to 72% of households that were in the process of divesting.

- Socio-economic differences between compliers and non-compliers were also surprisingly large. Female headed households (80% of which were widowed) and older household-heads were much more likely to drop out of the program. Household that took-up the treatment were significantly better off: they were better educated, had much higher access to finance levels – both in terms of savings and credit, and also had slightly better quality housing. Their children were also more likely to be enrolled.

Due to multi-collinearity, it is difficult to identify which of these covariates is the strongest predictor of compliance or non-compliance. A Kernel Regularized Least Square regression suggests that four indicators stand-out, including: the age and education levels of the head-of-household, access to finance (in particular credit) and finally whether acreage dedicated to coffee production was trending upwards or downwards in the years leading to the program. These figures suggest that TechnoServe's Wetmill and Agronomy programs were successful in attracting households that were already relatively better-off.

Table 17: Baseline indicators for compliers and non-compliers in the treatment group (2009 data)

Indicator	Compliers	Non-compliers	Difference
Distance from treatment site (median, kms)	3.2kms	2.5kms	+0.7kms
Distance from bank (median, kms)	6kms	5kms	+1kms
Distance from tarmac road (median, kms)	10kms	10kms	0kms
Coffee main income source (% households)	61.3%	60.1%	+1.2%
Coffee acreage **increase in three years before program (% households)**	19.6%	10.5%	+8.2%**
Male-headed household (% households)	82.0%	72.0%	+10.0%**
Average age	50.2	55.5	-5.3***
Number of household members (#)	6.3	5.6	+0.7***
Household head has completed primary (% household heads)	36.1%	28.1%	+8.0
Children aged 6-23 enrolled in school (% households)	75.0%	67.5%	+7.5%**
Has received credit (% households)	38.9%	22.4%	+16.5%***
Has a savings account (% households)	28.3%	14.9%	+13.4%***
Modern floor (% households)	16.9%	14.1%	+2.8%
Modern walls (% households)	17.0%	16.2%	+0.8%

**p<0.1; **p<0.05; ***p<0.01 ; "p" stands for percentage points*

5. Conclusions

This program evaluation suffered from a number of design and implementation issues that have made it impossible to make any causal inference about the effect of TechnoServe's Wetmill and Coffee Agronomy programs on the socio-economic status of beneficiary households. These issues included: (i) the non-compliance to the experimental protocol in terms of assignment to the control group; (ii) selection bias in the assignment of households to the control group, creating a bias towards households that were closer to main roads and larger settlements; (iii) multiple levels of treatment that were delivered in different locations at different points in time; (iv) incomplete compliance in both the treatment and control groups; (v) contamination from other TechnoServe treatment sites, affecting both control and treatment group households; (vi) high non-random attrition; and finally (vii) low data quality and comparability over time, due to changes in the survey instruments, different service providers for the baseline, midline and endline surveys and finally respondent bias. Combined, these effects led to biased estimators, reduced statistical power and low comparability of results over time. Despite attempts to overcome these biases using difference-in-difference techniques, non parametric regressions, inverse propensity score matching with regression, fuzzy regression discontinuity design and instrumental variables, it was not possible to make any causal inference about the effect of the program on the socio-economic status of beneficiaries. In fact, the vast majority of the estimates point towards a negative treatment, which is consistent with the hypothesis that control group farmers lived significantly closer to roads and urban settlements than their more rural counterparts in the treatment group and inconsistent with the nature and operational success of the Wetmill and Agronomy programs.

In terms of experimental design, there some important lessons to take away from this experience. These include the following:

- ***Commit to an established experimental design.*** One crucial aspect of this evaluation that was not clear from any of the baseline documentation was what the exact experimental or quasi-experimental approach was meant to be. What we know for a fact is that a treatment and control group were created using two distinct methods with in-built elements of randomization. Having a randomly selected control group and a separate randomly selected treatment group however does not make a valid program evaluation. A clear identification strategy is also needed, be it difference-in-difference in a randomized-control-type setting, regression discontinuity design, instrumental variables, etc.

- ***Working off one single sampling frame is always better.*** In designing an experiment, it is important to use the same sampling frame for the selection or assignment of households to the treatment and control groups. Here treatment households were randomly selected from a list of prospective farmers, while control group households were selected from coffee-growing surrounding areas with similar agro-climatic conditions. The first mistake in the design of this experiment was assuming that these two sampling frames were similar enough on average. As the data clearly reveals, these two populations might have had similar agro-climatic conditions, but they did not have similar socio-economic levels. The scale of selection bias that we experienced in this evaluation would have been avoided if households had been selected/assigned to the treatment or control groups from one-single sampling frame.

- ***Beware of the random walk approach***. Random walks are never fully random. This is especially true in the rigid hilly terrain of Rwanda, where housing in rural areas can be quite sparse and the walk from one house to the next very time-consuming. The GPS maps we present show this very convincingly. While treatment farmers were scattered around rural areas, the random walk approach more often than ended-up in straight line along roads that were easy to navigate. The incentives for enumerators not to respect this approach are also very high. Why climb a hilly terrain in an uncertain direction, when you can just walk down the road and knock on a random door?

- ***Use the collected GPS data***. A missed opportunity that would have avoided the costs of doing a midline and an endline when in fact assignment to the treatment and control groups was not random or not done following a clear evaluation and identification strategy, was the use of GPS data to visualize the location of treatment and control households. The GPS data was immensely useful in helping us understand the structure of the data and the nature of the biases we could see in the data. It also helped us identify cases of potential contamination from other TechnoServe treatment sites that had not yet been identified at the time when this evaluation was designed.

- ***Take incomplete-compliance and attrition into account***. Incomplete compliance is a very normal issue in evaluations because researchers more often than not do not have full control of whether treatment and control group members will be compliant or not. If assignment to the treatment group is fully random, then circumventing the problem of non-compliance is not very complicated from an identification-strategy perspective. What is problematic though is the effect of non-compliance on the statistical power of an evaluation. This is also true for attrition. As we have shown, statistical power reduces very fast as the rate of non-compliance and attrition increases. It is therefore important to adjust for expected compliance and attrition rates in power calculations aimed at determining the desired sample size for a certain experiment.

- ***Use one provider or make sure there is sufficient institutional memory***. One of the main challenges we faced in this evaluation was trying to understand exactly what had been done at the baseline and the midline and trying to figure out how we could make the endline data more compatible and comparable to the latter. This was made more difficult by the fact that the entire TechnoServe team that had worked on this evaluation had moved on to other jobs/countries and that three different providers were used for the baseline, midline and endline studies respectively. The unnecessary loss of time was significant.

The last take-away from this exercise is that the TechnoServe program was more successful in attracting better-off and more educated households and coffee farmers in treatment areas, rather than poorer coffee growers that might potentially have benefitted even more from the program. We cannot speculate on the ideal target-group for these type of interventions, but this useful knowledge for TechnoServe to keep in mind for future interventions. Key predictors of whether a household would take-up the treatment or not when given the opportunity were the age and education of the household-head, their level of access to credit and savings, and finally whether they were expanding their coffee production or divesting away from it.

References

Esther Duflo, Rachel Glennerster and Michael Kremer, 2008. "Using randomization in development economics research: a toolkit", Center for Economic Policy Research, Discussion Paper Series No6059, January 2007

Hainmueller, Jens and Hazlett, Chad, Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach (Sept 2013). Political Analysis (2013 Forthcoming).

Imbens, Guido W. & Lemieux, Thomas, 2008. "Regression discontinuity designs: A guide to practice," Journal of Econometrics, Elsevier, vol. 142(2), pages 615-635, February.

Imbens, Guido W & Angrist, Joshua D, 1994. "Identification and Estimation of Local Average Treatment Effects," Econometrica, Econometric Society, vol. 62(2), pages 467-75, March.

Keisuke Hirano, Guido W. Imbens, Geert Ridder. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score". NBER 2000.

ANNEX 1: Distances between Cooperatives

Cooperative	BUH	CAF	MIZ	GSH	GIS	GKA	NYA	GKE	MUS	MWE	SHA	VUN
Buhanga (BUH)	-											
Cafeki (CAF)	24	-										
Mizero (MIZ)	30	9	-									
Gashonga (GSH)	103	104	97	-								
Gisuma (GIS)	99	98	90	13	-							
Gisaka (GKA)	97	93	101	197	191	-						
Nyarubuka (NYA)	47	22	21	113	104	93	-					
Giseke (GKE)	44	20	20	115	107	89	5	-				
Musha (MUS)	88	73	78	174	166	48	62	60	-			
Mwezi (MWE)	90	91	84	14	12	183	99	101	161	-		
Shara (SHA)	85	80	72	33	21	173	84	87	146	22	-	
Vunga (VUN)	93	69	63	122	110	127	49	53	84	110	89	-
Matyazo (MAT)	89	65	59	120	109	123	44	48	80	109	88	4

Annex 2: Indicator estimates between compliers vs. non-compliers

Indicators	Households that didn't drop out	Households that dropped out	T-statistic (equality test)
Location indicators			
Distance from treatment site (median)	7.36	7.04	-1.31
Distance from tarmac (median)	8	8	0
Distance from bank (median)	4	4	0
Distance from public transport (median)	4	4	0
Distance from hospital (median)	7	8	1.17
Coffee intensity indicators			
Coffee main sources of income (mean)	57.6p	61.4p	1.24
Coffee not in top 3 sources of income (mean)	14.6p	11.4p	-1.53
Coffee acreage has increased over past 3y (mean)	16.2p	11.6p	-2.16**
Number of coffee trees (median)	160	190	2.57**
Food intake indicators			
Meals per day (mean)	2.02	2.04	-1.30
Full serving of meat previous day (mean)	9.6p	11.3p	0.88
Full serving of starch previous day (mean)	93.6p	93.1p	-0.36
Full serving of eggs previous day (mean)	3.0p	3.1p	0.07
Full serving of fruit previous day (mean)	30.9p	30.3p	-0.24
Full serving of vegetables previous day (mean)	81.3p	86.6p	2.38**
Full serving of beans previous day (mean)	94.5p	94.1p	-0.29
Household assets indicators			
Household owns pig	17.7p	22.0p	1.55
Household owns chickens	21.9p	21.4p	-0.18
Household owns goats	60.5p	50.8p	-2.97***
Household owns cattle	61.7p	64.7p	0.93
Household owns mobile phone	14.1p	14.2p	0.06
Household owns mattress	53.0p	51.4p	-0.53
Household owns TV	2.2p	2.4p	0.23
Household owns radio	73.4p	76.7p	1.28
Housing facilities indicators			
Better walls (not earth or bamboo)	19.7p	22.1p	0.95
Better floors (concrete floors)	19.2p	19.3p	0.04
Better toilet (with flush or with cement floor)	6.0p	6.5p	0.34

**p<0.1; **p<0.05; ***p<0.01*

Annex 3: Endline Adjustment Factor (weights) for # of Household Members

Number of household members	Endline adjustment factor
1	2.958
2	1.342
3	1.177
4	1.101
5	1.065
6	0.964
7	0.882
8	0.959
9	0.858
10	0.880
11	0.896
12	0.924
13 or more	0.863