laterite TRANSFORMING RESEARCH IN AFRICA
Rwanda | Malawi | Burundi

TechnoServe Coffee Initiative: Socio-Economic Program Evaluation Tanzania

Endline Report | March 2014

TechnoServe International

Table of Contents

Introduction

. This report details the technical and implementation-related shortcomings of TechnoServe's socio-economic impact study for the coffee agronomy and wetmill program in Tanzania. These prevent us from providing any conclusive evidence on the effect of the program on beneficiary households. The findings presented here should be read in conjunction with a report written on the case of TechnoServe's agronomy and wetmill programs in Rwanda. The experimental design of these two evaluations was the same; the same consortium was responsible for design, data collection and analysis at the baseline; and these two program evaluations share many of the same design issues.

The purpose of this evaluation was to measure the effect of TechnoServe Tanzania's wetmill and agronomy programs on the socio-economic status of households in treatment areas. These programs were designed to boost and upgrade coffee production in the Meru and Hai districts in Northern Tanzania and in the Mbeya and Mbinga districts in the South. The evaluation focused in particular on the Northern districts and the district of Mbinga in South.

The wetmill and agronomy programs were complementary. Farmers were given training to produce more and better quality coffee through the agronomy training 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 better 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 responsabilities;

(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 equitable gender

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 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 question we are answering here is whether these two programs 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 in 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. The experimental design and the baseline data collection were handled by a consortium led by Tegemeo and Research International in 2009; endline data collection was sub-contracted by TechnoServe in 2013 to Daniel Mwero; finally, Laterite Ltd, based in Rwanda and that led the analysis for the Rwanda evaluation, was selected to complete the analysis of the evaluation.

We first briefly present the experimental design, before explaining some insurmountable technical issues in the design and implementation of this evaluation that undermine the validity of this experiment.

Experimental design

What follows is a brief description of the experimental design of this program evaluation adapted from Laterite's Rwanda report. For more details please refer to the Rwanda report by Laterite Ltd or the baseline report prepared by the consortium led by Tegemeo.

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

(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 early 2009. Randomization was stratified at the cooperative level. Households in the treatment area had the option to self-selecting into these two 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, although this was not specified at the baseline.

(iii) ***Two time periods.*** We have two data points in time for this evaluation. Baseline data was collected in 2009, before the start of project activities. An endline was conducted in 2013, after the completion of the programs in all 13 cooperatives. It is important to note that data was collected after four 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 as in 2013.

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 early 2013; 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 are 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 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;

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 farmers receiving or taking-up the treatment; and,

4) There should have been no other interventions – related or unrelated to coffee – that affected either the treatment or control group disproportionately.

In the next chapter 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.

Design and implementation shortcomings

In this chapter we explain the design and implementation failings of this program evaluation that undermine the validity of this evaluation and make it impossible to infer whether TechnoServe's agronomy and wetmill programs had an impact on the socio-economic status of beneficiary households or not. The main issues are: non-random attrition (50% in total), incomplete compliance in the treatment and control groups, a high risk of contamination of the control group from alternative TechnoServe supported cooperatives, selection bias at the baseline, data quality issues in particular for location data and the age variable, and finally negative impact estimates.

1. High and non-random attrition

Attrition is a major issue affecting this evaluation. Attrition was non-random and led to a reduction in the overall sample size of this evaluation by about 50%. Only 666 households were interviewed in the endline, compared to an initial sample size of 1,312 households in the treatment and control groups. There are two aspects to this attrition: (i) attrition by design, with the dropping of all treatment and control farmers from Northern Tanzania where the wetmill and agronomy programs were not fully implemented; and (ii) attrition amongst farmers in Southern Tanzania, which was non-random and disproportionately affected the control group (see table 1).

Table 1: Attrition by experimental group in the North and South of Tanzania

Experimental groups	Baseline	Endline	Attrition
Treatment group (North)	187	0	100%
Control group (North)	226	0	100%
Treatment group (South)	414	340	17.9%
Control group (South)	485	326	32.8%
Totals	1,312	666	50.8%

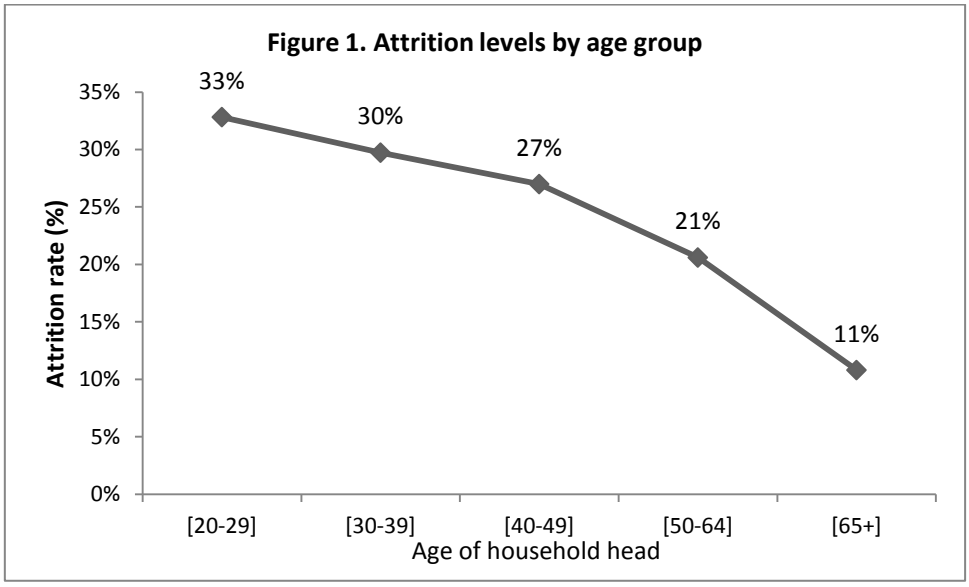
A total of 413 households from Northern Tanzania, or about 31.5% of the entire sample, were dropped from the evaluation because no single cooperative graduated from the wetmill and agronomy programs in the north of the country. This decision was justified for financial reasons and because the two geographic areas of focus were distinct – i.e. the effect of the agronomy and wetmill programs in Southern Tanzania was independent of their effect in Northern Tanzania. Dropping the entire Northern province from the evaluation significantly reduces the sampling power of the evaluation and makes it necessary to re-check whether treatment and control were similar at the baseline in the Southern Province (see point 4).

Attrition in Southern Tanzania affected 26% of the sample, which by program evaluation standards is considered very high. This is a very similar attrition rate to the case of Rwanda's socio-economic impact study, where attrition between baseline and endline was also about 26%. We define attrition here as targets in the treatment or control group that could not be reached in the endline or that refused

to co-operate and respond to the questionnaire. We find sufficient evidence to suggest that attrition in the Southern Province was non-random:

1) ***Attrition disproportionately affected households in the control group.*** Households in the control group were almost two times more likely to drop out from the survey, compared to households in the treatment group. Attrition was 32.8% in the control group and just 17.9% in the treatment group (see table 1). This is most probably due to the fact that treatment households were easier to locate given the support structures TechnoServe has on the ground. Control farmers were selected by random walk which could have led to less accurate location data.

2) ***As in the case of Rwanda, attrition affected younger household-heads more.*** The average age of household-heads that dropped out of the survey was 40 years-old, compared to 44 for households that stayed in the survey (this difference is highly statistically significant). As can be seen in figure 1, average attrition for households where the household head was younger than 30 years-old at the baseline was 33%, compared to just 11% for households where the household-head was past 65-years old. This age dynamic is also reflected in the average size of households: households that dropped out of the survey had on average 3.9 members at the baseline, compared to 5.7 for households that remained in the survey. We explain this dynamic by the fact that "younger" households and household-heads are more mobile and hence more difficult to locate.



3) ***Finally, households that dropped out of the survey also differed from households that remained on some other key characteristics.*** Attrition was higher for households that had invested less / divested more from coffee production in the three years preceding the baseline. 14.4% of remaining households in the survey had increased their coffee acreage in the three years preceding the survey, compared to just 6.4% for households that dropped out of the survey (this difference is statistically significant at the 1% level controlling for individual and households characteristics). Likewise, 9% of households that dropped out of the survey decreased their coffee acreage in the three years before the baseline, compared to just 6.9% for households that remained in the survey. Even though this does not hold for all indicators of interest, we find that households that dropped out of the survey were on average slightly

worse off on the socio-economic indicators of interest. This was particularly true for indicators relating to access to finance. Attrition was significantly higher for households that had not taken any credit in the past twelve months. Only 1.3% of households that dropped out of the survey had taken a credit in the 12 months prior to the survey, compared to 5.1% of households (a difference that is statistically significant at the 1% level after controlling for individual and household characteristics). In the same way only 5.2% of households that dropped out of the survey had a savings account, compared to 8.9% of households that remained in the survey. By affecting households that were slightly worse-off on average and by disproportionately impacting the control group, attrition introduced bias into the experiment.

Attrition therefore impacted this evaluation in two ways: (i) it significantly reduced the statistical power of this experiment; and (ii) it introduced bias into the results, potentially undermining some of the main identifying assumptions that would make this program evaluation valid.

2. Incomplete compliance in treatment and control

Incomplete compliance in both the treatment and control groups adds to the problem of attrition. We talk about incomplete compliance in the treatment group when its members end-up not receiving the treatment or decide not to take it; incomplete compliance in the control group happens when some control group members receive the treatment. Incomplete compliance is a very common issue in program evaluation because the researcher typically does not have full control of whether someone receives the treatment or not. While dealing with incomplete compliance is relatively straightforward form an analytic perspective, it does have a very negative effect on statistical power. Here we consider that a household has received the treatment if at least one household member participated in either program or if the household sold coffee to a TechnoServe supported wetmill.

Using this definition of treatment we find that about 80.4% of remaining treatment farmers (after attrition) received the treatment, as did 12.5% of control group farmers (see table 2). How does this affect the statistical power of this experiment? 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]. Here the difference in take-up between treatment and control is 67.9% (80.4%-12.5%=67.9%). The required sample size to maintain the same level of statistical power as if there were full compliance is therefore the inverse of the square of 69.7%, which is 2.2 times the original sample size (i.e. $1/0.679^2=2.2$). In other words, with this level of incomplete compliance a sample size of 1,465 households would have been required to maintain the same level of statistical power as with a sample of 666 households in which there is full compliance.

[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

Table 2: Share of treatment and control group households that reported to have received the treatment at the endline (2013)

Intervention	Control Group	Treatment Group
No Treatment	87.5%	19.6%
Treatment	12.5%	80.4%
Total	100.0%	100.0%

Let us illustrate the combined loss in statistical power due to both attrition and incomplete compliance using an example. Assume we are interested in estimating the statistical power of this experiment associated with identifying a 5 percentage point increase in the number of households with a savings account. We know that at the baseline 8.3% of households had a savings account, with a standard deviation of about 28%. With full compliance and based on the original sample size of 601 households in the treatment group and 711 in the control group, the statistical power associated with identifying a treatment effect of 5 percentage points is about 90.5% (assuming an alpha of 5%). This means that we would only have a 9.5% chance of rejecting the hypothesis that the treatment effect on owning a savings account was at least 5 percentage points if this hypothesis were true. Attrition reduces the statistical power of this test to 64.4%. After attrition and with a remaining sample size of 666 individuals, we would have a much larger chance of rejecting the hypothesis even if it were true. Finally, after taking incomplete compliance into account, the statistical power of this experiment would be reduced to 36.9%. In this scenario we would fail to identify the treatment effect more than 3 times out of 5.

3. Potential for contamination

Another issue that undermines the internal validity of this experiment is the potential for contamination, spill-over effects and unidentified non-compliance in the control groups. This experiment was carried out in the vicinity of nine treatment sites in the district of Mbinga in Southern Tanzania. By design control group farmers were selected from coffee growing areas that were at least 5km away from a given treatment site. But these nine treatment sites were not the only wetmills/cooperatives that TechnoServe worked with in the area. During the 2008-2013 period TechnoServe worked with a total of 50 cooperatives in Mbinga district, some very close to the treatment sites that were included in the socio-economic impact study. We find that seven of the nine treatment sites (including Barazani, Kimate, Kipando, Kungiro, Mkuka, Nguvumali, Tupendane, Ukombozi, Unango) had between 15 to 23 other TechnoServe supported cooperatives operating within a 10km radius from them[2]. This suggests that control group farmers in the vicinity of these cooperatives were likely to be in the direct catchment area of multiple TechnoServe supported cooperatives/wetmills.

[2] This is based on distance calculations using GPS coordinates on the location of wetmills/cooperatives provided to Laterite by TechnoServe.

Figure 2: Proximity of control group households to TechnoServe supported Wetmills in Mbinga

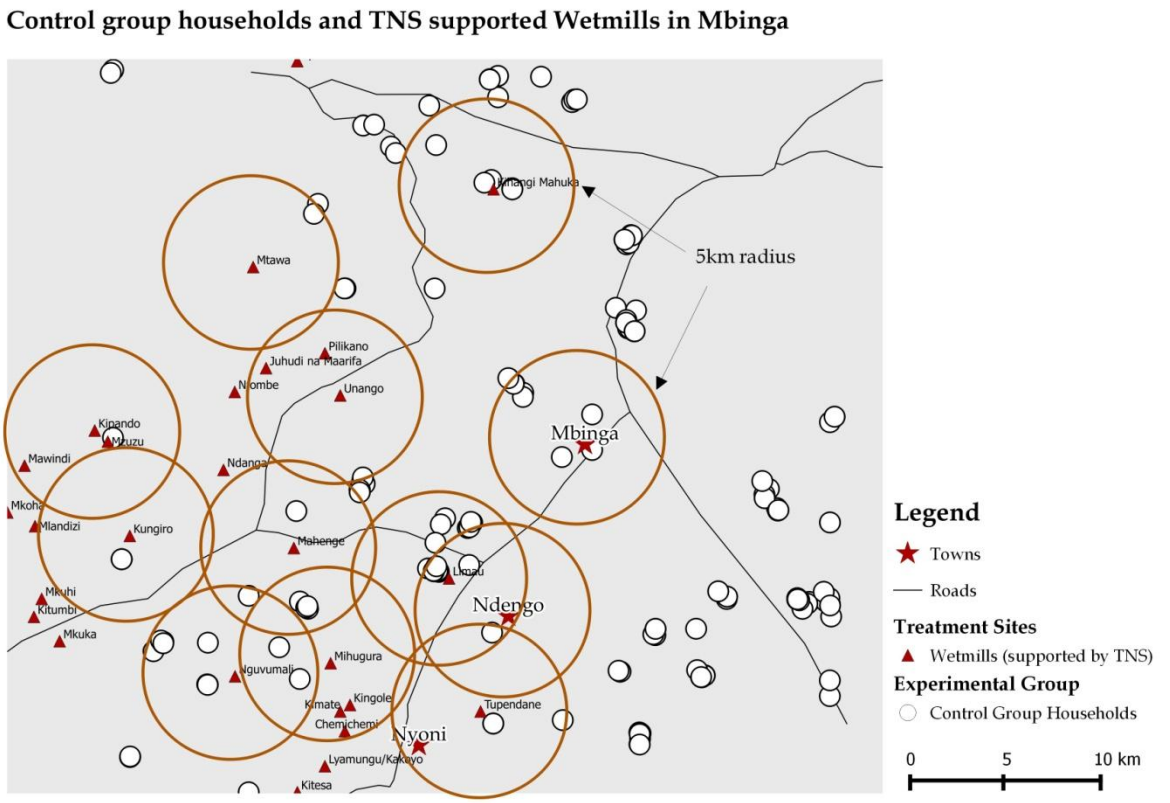


Figure 2 depicts the location of some control group households and TechnoServe supported wetmills in Mbinga district. As expected, the map reveals that many control group households either fall within the 5km radius of one or multiple TechoServe supported wetmills or are very close to towns such as Mbinga, Ndengo and Nyoni. The location of these control group households suggests that the risk of spill-over effects and unidentified non-compliance in the control group is high. This map is incomplete because GPS location data for many households was found to be inaccurate, but the location of many of the control households speaks for itself. We would expect households that are close to cooperatives to benefit either directly or indirectly from the agronomy and wetmill programs, and households that are close to cities to have greater access to urban facilities, services, and investments. Both proximity to cooperatives and urban areas would therefore make control group households better off..

4. Selection bias at the baseline

In addition to the issues of attrition, incomplete compliance, and the potential for contamination, we find evidence of selection bias at the baseline in the sample limited to the Mbinga district. Selection bias here is driven by unobservables; i.e. we do not have an explanation as to the source of this bias. In particular we find that the control group is significantly better off than the treatment group on a number of key variables of interest, including enrolment rates and meals per day for children. These differences persist when controlling for differences in basic characteristics such as gender, age, number of households members, etc. The difference for average enrolment rates for children aged 6 to 20 years old is very large and inexplicable. We also find that treatment group members are more coffee intensive – in particular they were much more likely to have increased their coffee acreage in

the years leading to evaluation – and less likely to work off-farm. Table 3 compares average baseline values for selected variables in the treatment and control groups.

Table 3: Comparison of average baseline values in treatment and control for selected indicators (2009)

Indicator	Treatment	Control	Difference
Coffee main source of income (% households)	90.6%	88.8%	+1.8p
Increased coffee acreage in 3 years before survey (% households)	16.1%	11.7%	+4.4p***
Enrolment (% children aged 6-20)	51.7%	66.7%	-15.0p***
Meals (# meals per day)	2.61	2.66	-0.05**
Occupation (age 18+, % households)			
On farm	86.0%	83.7%	Chi-squared test: p<0.000
Off-farm	1.9%	4.9%	
Unpaid work	4.4%	2.3%	
Student	7.8%	9.0%	

significant at 10% level; ** significant at 5% level; * significant at 1% level*

Even though treatment and control appear to be balanced on other variables of interest, these differences – in particular differences in school enrolment rates – suggest that there was selection bias at the baseline. These biases further undermine the internal validity of this experiment.

5. Quality of data

Data quality is an additional impediment to this evaluation. There are two main issues: the quality of location data and time-inconsistencies in the age variable, which is an important determinant for the majority of the socio-economic variables we are studying, in particular education, health, assets, etc. These data quality issues were made worse by the fact that two separate agencies were involved in the data collection, while a third, Laterite, was responsible for the final data analysis. This has resulted 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.

Location data

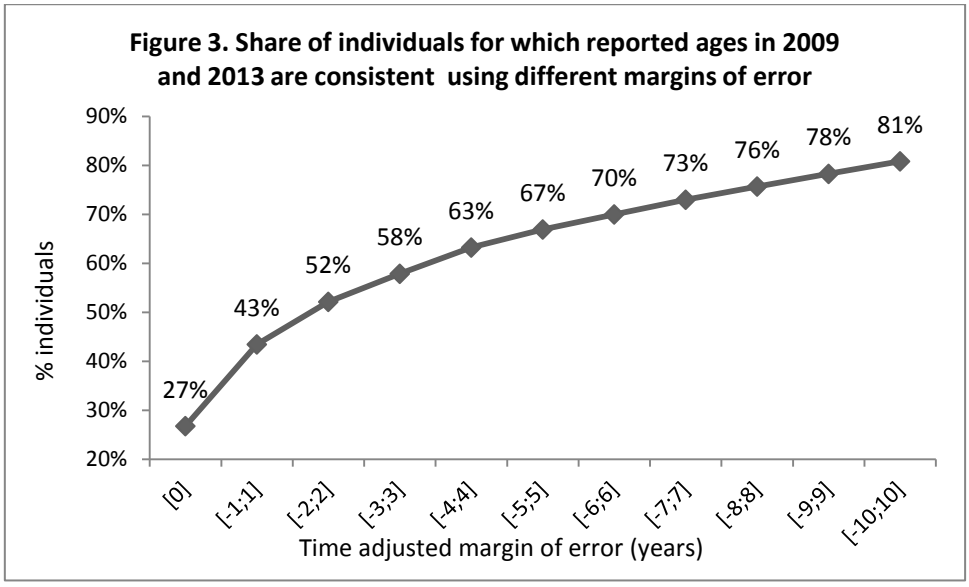
Inconsistent location data has made it impossible for the research team to control for location variables and to capture the sampling design in the analysis. There are two ways to take location into account in this evaluation: (i) either by focusing on the location of the target wetmills, in which case we need to know which treatment and control group households are associated to which wetmill; or (ii) to focus on the administrative location of households. GPS data is inaccurate for half of the sample, placing households in Malawi, rather than the Mbinga district. This means that we cannot use GPS data to calculate distances from the treatment site or to associate a given control household to the catchment area of a wetmill (in the baseline dataset control farmers are not assigned to any treatment site / wetmills in particular). While administrative data on the location of households is excellent, there are too many subdivision within the district of Mbinga to make geographic controls at the "Division" level meaningful[3]. Including "Division" dummies creates multicollinearity in the

[3] Districts in Tanzania are divided into "Divisions". The district of Mbinga has almost 50 "Divisions" or wards.

regressions and perfectly identifies multiple households. Moreover, we find that using administrative data (for which we do not have accurate GPS coordinates) to associate control group households to treatment wetmills leads to inaccuracies. This is because some wetmills are in the same "Division" or in neighboring "Divisions", while some control group households are too far away from any particular treatment site to link them to a particular wetmill or cooperative. This leaves us with no way to control for the location of households or to reflect the sampling strategy in the analysis. Many of the observed biases could have been explained by the geographic distribution of treatment and control group households, but we have no way of analytically testing this.

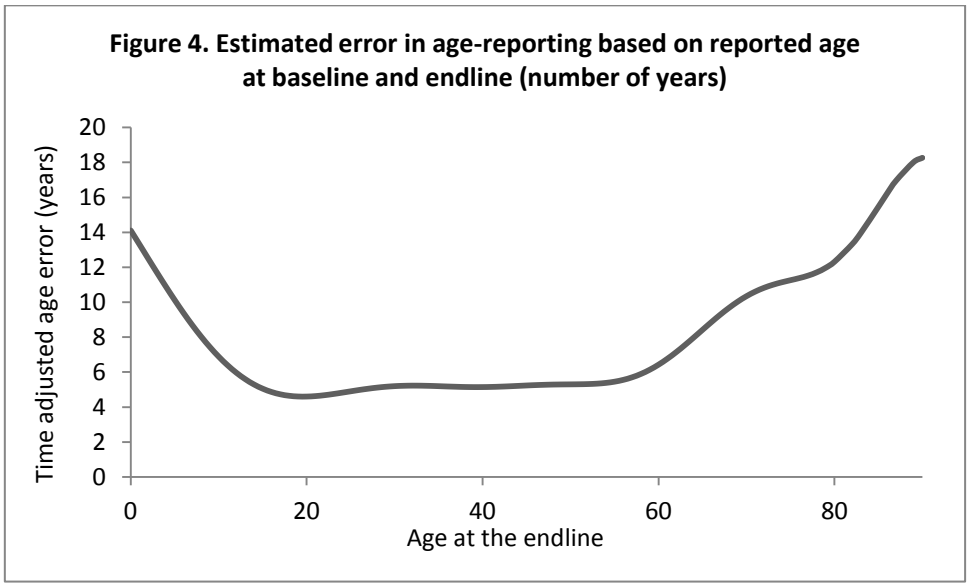
Quality of data - age

One way to test the quality of the data we are working with is to study the age variable, which also happens to be a very important explanatory variable for many of the socio-economic variables of interest at both individual and household levels. Age is a perfect variable to test, because given someone's age at the baseline we know what that person's age should have been at the endline. Age at the baseline was captured using the "year-of-birth" variable. At the endline, questionnaires 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 households head, etc. One detail that was not contained in this pre-populated sheet, but that was asked each time, was the age of each individual household member. This makes it possible for us to test the consistency of responses over time on age.



As in the case of the socio-economic impact study for Rwanda, we find that only about 30-40% of ages were correctly specified (see figure 3). We say that the age variable is correctly specified when the age reported at the endline is about 4 years older than the age at the baseline (which we calculate based on the reported year of birth). With a margin of error of +/-1 years, we find that 43% of ages were correctly specified. This increases to about 52% with a margin of error of +/-2 years. This means that for about half of our sample the error on the age variable was greater than 2 years. For 1/3rd of the sample the error in the age variable is greater than +/-5 years, which is an interval of 10 years. Errors were significantly larger on average for children and older people above the age of 60 (see figure 4, where we estimate the error in age reporting using a Kernel-based Regularized Least Squares). Specified ages for children below the age of 20 were off by 5 to 14 years on average, which is very

large, especially if we are interested in the effects of the agronomy and wetmill programs on education or the health of children.



The sheer scale of the age discrepancies suggests that the coefficients we obtain when controlling for age or when limiting the sample to a certain age-group will be inaccurate. This could be an issue limited to the age variable, but this is not something we can check.

6. Negative impact estimates confirm biases

Selection bias at the baseline and attrition-induced bias have resulted in a control group consisting of households that were better off at the baseline than treatment households. This mismatch suggests that had there not been any intervention, treatment and control group households would most likely not have evolved in parallel. This is one of the most important identifying assumptions of any difference-in-difference strategy. If treatment and control would have evolved in the same way had there been no intervention, then it would have been possible to interpret any post-treatment difference as a measure of the impact of the intervention. In this case, where control farmers are better off than treatment farmers to start off with, we would expect control group households to perform better over time than treatment group households. This is because they are starting from a better base.

Difference-in-difference estimates for a number of target variables seem to confirm this hypothesis, even if it is not the case for all indicators. For variables of interest relating to food intake, access to savings, the ownership of certain assets, we find that belonging to the treatment group is associated with a negative and statistically significant effect – i.e. control group households performed better on average between 2009 and 2013 than treatment households. Table 4 presents difference-in-difference estimates for these variables of interest after controlling for individual characteristics of the household head (age, age squared, gender), household characteristics (number of household members, whether the main source of income is coffee or not, education of household head, occupation of household

head, marital status of household head), and location indicators (distance from a bank, a hospital, and a public transport point[4]).

Table 4: Difference-in-difference estimates for selected variables of interest

Indicator	Difference in difference estimate	t-Statistic
Had serving of meat the day before (% sample)	-20.4p	-8.5***
Household has a savings account (% household)	-11.8p	-3.2***
Household has taken credit (% households)	-2.2p	-0.6
Household has television (% households)	-5.7p	-2.0**
Household has bicycle (% households)	-8.5p	-1.7*
Number of mobile phones (# per household)	-0.15	-1.7*
Number of mattresses (# per household)	-0.36	-2.3**

significant at 10% level; ** significant at 5% level; * significant at 1% level*

Conclusion

The negative impact estimates we obtain are not reflective of a negative program effect, but rather of the inherent flaws in the design and implementation of this evaluation. We have also tested alternative techniques, such as various matching estimators using propensity scores to explore whether we could overcome some of the issues created by selection bias and attrition, but they do not work, suggesting that biases are driven by unobservables. Even if we were able to overcome some of these biases we would most probably not have enough statistical power to identify the effects of both programs on socio-economic indicators of interest due to incomplete compliance. Lastly the estimates we obtain could be inaccurate because of geographical imbalances that we cannot control or weight for or because of unidentified contamination from alternative TechnoServe-supported wetmills in the vicinity of the selected treatment sites.

As in the case of Rwanda's socio-economic impact survey, we conclude that there are too many experimental design and implementation flaws to make any meaningful inference on the effect of TechnoServe's agronomy and wetmill programs on the socio-economic status of beneficiary households. In terms of future evaluation design, the lessons that TechnoServe can take away from this experience are the same as in the case of Rwanda. We repeat these and adapt them slightly below:

- ***Commit to an established experimental design.*** One crucial aspect of the Tanzania and Rwanda evaluations 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

[4] Location data were selected on the basis of the quality of the location data.

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. In this case there were some elements of regression discontinuity design which were not fully followed through (only control group members were assigned using the 5km cut-off point) and an attempt at creating a quasi-experiment with two almost identical groups selected from different sampling frames.

- ***Ensure that the treatment and control groups are selected in the same way, from the same sampling frame.*** In this experiment households in the treatment and control groups were selected randomly, but from a different sampling frame. Treatment households were randomly selected from a list of potential beneficiaries in treatment cooperatives, while control farmers were selected by random walk from areas with similar agro-climatic conditions. This design is based on some very strong assumptions: (i) that farmers that signed-up to treatment cooperatives, would be similar to farmers randomly selected in nearby coffee growing areas with similar agro-climatic conditions; and (ii) that the list of areas with similar agro-climatic conditions, selected by TechnoServe, would be similar on average to the selected treatment sites (which were selected not only based on their agro-climatic situation, but also cooperative management indicators). These assumptions are unlikely to hold, as one would expect prospective cooperative members to be different to the random coffee farmer selected from nearby coffee growing areas who might not be member of cooperative. Also, unless the treatment and cooperative sites were selected in the same way, using the same set of criteria, there are likely to be geographic differences between treatment and control sites as well. A better strategy would have consisted in selecting treatment and control sites in the same way, and then within these, selecting treatment and control farmers using an identical strategy.

- ***Take the risk of incomplete compliance and attrition into account when estimating required sample sizes***. Combined, incomplete compliance and attrition dissipated the statistical power of this experiment. As we have shown, statistical power reduces very fast as the rate of non-compliance and attrition increases. While it is impossible to know in advance what share of people will take-up a certain experiment and what share of people will drop out of an experiment, it is useful to build some broad assumptions into sample size calculations at the baseline.

- ***To avoid contamination, it is important have a good idea of future expansion plans before designing an experiment.*** The risk of contamination turned out to be a major issue in the roll-out of the Tanzania evaluation. TechnoServe and the research team failed to take the future expansion of TechnoServe support to other cooperatives into account in the experimental design phase. The result was that many control group households ended-up being in the immediate vicinity of new TechnoServe supported wetmills. This lead to greater incomplete compliance and increases the risk of contamination / spill-over effects.

- ***Beware of the random walk approach***. Random walks are never fully random. This is especially true in hilly coffee growing areas, where housing can be quite sparse and the walk from one house to the next very time-consuming. The incentives for enumerators not to respect the random walk approach in these settings are quite high. Why climb a hilly terrain in an uncertain direction, when you can just walk down a road a bit and knock on a random door?

- ***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.