Improve Jobs of Smallholder Women Farmers: Impact Evaluation Design

Joao Montalvao
Francisco Campos

World Bank

2018

The publication of this study has been made possible through a grant from the Jobs Umbrella Trust Fund, which is supported by the Department for International Development/UK AID, and the Governments of Norway, Germany, Austria, the Austrian Development Agency, and the Swedish International Development Cooperation Agency.

Jobs MDTF
Project Name: Improve Jobs of Smallholder Women Farmers
Country: Mozambique
TTL: Joao Montalvao, Francisco Campos
Output: Impact Evaluation Design
Date: April 2018

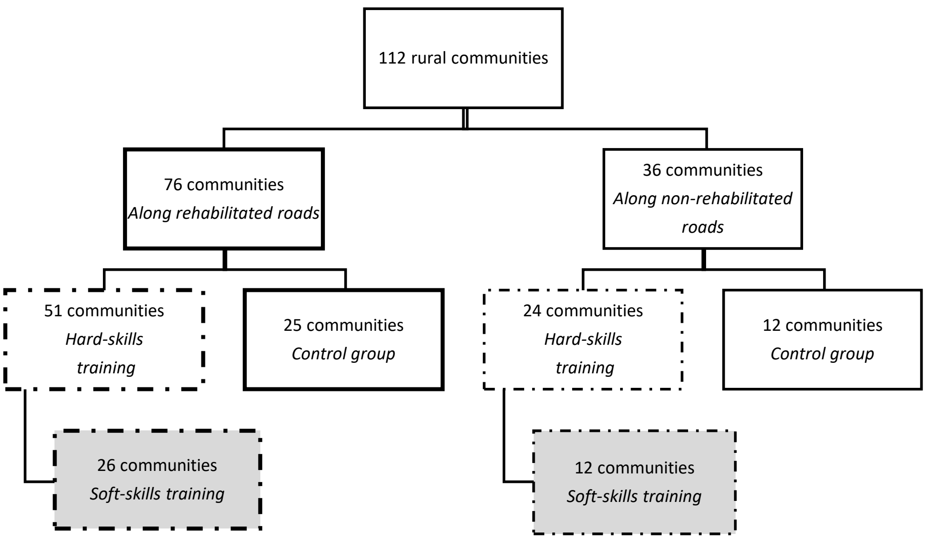
Part I. Agricultural extension program

1. Sample Selection

A rigorous randomized controlled trial (RCT) approach will be used to measure the impacts of the agricultural extension interventions in Mozambique. The sample is composed of 2,240 women spread across 112 rural communities (20 women in each community) in the Tete province of Mozambique. The communities are equally distributed along three feeder roads (R603, R604, R605) in the district of Angonia, Macanga, and Tsangano, with half of them located 0-2 km away from the roads and the other half located 2-10 km away from the roads[1]. Two of these roads (R604 and R605) will be rehabilitated under the Mozambique IGPP between February 2017 and October 2019. The other road R603 will not be rehabilitated.

The 112 communities were randomly assigned to one of three treatment groups: 37 communities were offered the hard-skills training only (Treatment 1), and 38 communities were offered the hard-skills training *and* the soft-skills training (Treatment 2). The third group of 37 communities (Control group) will not receive any training and will thus form a comparison group. The random allocation of communities across the three experimental arms will be stratified by both road and distance to road.

Figure 1. RCT Design



2. Econometric Specification

[1] Road R604 is situated between N304 close to Mphulu all the way through Tsangano and up to Ulongwe (100 km). Road R605 is situated between Ulongwe, through Domue and up to Furancungo (99.6 km). Road R603 starts from Furancungo and is 74 km long toward the west. Road N302 is 160 km long, but all the communities that are part of the RCT study are spread out over a distance of 50 km along the Zambian borders.

As training participation is voluntary, not all eligible women will take-up the offer of receiving the training, and we therefore focus on intent-to-treat (ITT) impacts. Given the random assignment of communities to treatment arms, it is straightforward to estimate the ITT impacts of the training interventions using the following OLS specification,

$$y_{\{i\}hv,t} = \gamma_0 + \gamma_1 T_v^1 + \gamma_2 T_v^2 + \gamma_3 y_{\{i\}hv,0} + \gamma_4 x_v + \varepsilon_{\{i\}hv,t} \quad (1)$$

where $y_{\{i\}hv,t}$ is an outcome of interest for household h in village v measure at midline or endline ($t = 1$ or 2). Index i denotes the individual (man or woman) for individual level outcomes. T_v^1 is a dummy variable equal to 1 for villages assigned to Treatment 1 group (hard skills training), and 0 otherwise. T_v^2 is a dummy variable equal to 1 for villages assigned to Treatment 2 group (hard skills training and soft skills training), and 0 otherwise. To improve the power of our dataset to detect meaningful treatment effects we control for the baseline level of each outcome $y_{\{i\}hv,0}$ [McKenzie 2012], as well as for the community-level randomization strata (road and distance to road) x_v [Duflo *et al.* 2007]. We allow the error term $\varepsilon_{\{i\}hv,t}$ to be clustered by community v. The parameters of interest are γ_1 and γ_2, which identify the standalone impact of the hard skills training and the combined impact of the hard skills training and the soft skills training. The difference, $\gamma_2 - \gamma_1$, identifies the marginal impact of incorporating soft skills training into an agricultural extension program.

3. Data

The main source of data for the impact evaluation will be annual surveys to collect community, household, and parcel level data. The community questionnaire focus on village-level crop prices and access to markets. Surveys are administered to the household head and his or her spouse.

Households and communities were first surveyed in April-June 2016 (baseline). The same households and communities will be interviewed again in June-July 2018 (midline), and in June-July 2019 (endline). Funds permitting, another round of data collection will take place post-2019 in order to detect longer-term effects.

The main outcomes of interest collected through these surveys are:

(i) cognitive and non-cognitive skills;
(ii) women empowerment and intra-household bargaining;
(iii) employment including off-farm;
(iv) feeder road usage and access to markets;
(v) agricultural production and sales, crop choices, input usage, and farming practices;
(vi) household and farm assets;
(vii) consumption.

The module on non-cognitive skills will be designed and validated by a dedicated team of psychologists. It will comprise multiple items tapping the construct of personal initiative specific to poor women in rural settings.

Part 2. Road rehabilitation works

The research project is built on an ongoing impact evaluation of a feeder roads rehabilitation program under an active World Bank project (IGPP - P127303). The roads rehabilitation evaluation combines aspects of the (non-random) placement of the feeder roads rehabilitation program with baseline and follow-up survey data – in a difference-in-differences framework – to measure the effectiveness of better road infrastructure aimed at improving the ability of farmers (both male and female) to sell their produce in the markets. We explain in this section the design of the rural roads rehabilitation impact evaluation. The project will use the following four alternative quasi-experimental empirical frameworks, which differ in terms of their identification assumptions and data requirements. Together they will provide a rigorous, yet approximate, set of estimates of the causal impact of rehabilitating roads R604 and R605.

1. Double-Difference using Time and Comparison Roads

First, a popular strategy is to collect baseline and follow-up data and use a difference-in-difference (DD) methodology [e.g. Mu and Van de Walle 2007, Khandker *et al.* 2009, Ali 2011, and Aggarwal 2015]. This method compares the change over time in outcomes between communities in close proximity to rehabilitated roads and communities in close proximity to non-rehabilitated roads. In discussions with the project team and counterparts at the ANE (Associacao Nacional de Estradas) and FE (Fundo de Estradas), roads R302 (between Farracungo and Mualadze (≈100 km)) and N603 (between Farracungo and N9 (≈100 km)) were selected for the comparison group. These roads are in the vicinity of roads R604 and R605, share the same agro-ecological characteristics, and are in an equally bad state.

This method identifies the causal impact of road rehabilitation under the assumption that unobservable factors driving a wedge in underlying time-trends between communities along roads R604/R605 and communities along roads N302/R603 are constant across time. In practice this method is estimated with the following specification using pre- and post-intervention data on households in communities along roads R604/R605 and roads N302/R603:

$$\Delta y_{hcr} = \alpha_0 + \alpha_1 R_{cr} + \alpha_2 X_{ct=0} + \Delta\varepsilon_{hcr} \qquad \text{(B.1)}$$

where $\Delta y_{hcr} = (y_{hcrt=1} - y_{hcrt=0})$ is the change in the outcome of interest between baseline and follow-up for household h, in community c close to road r. R_{cr} is a rehabilitation dummy that equals 1 if road r is either R604 or R605, or 0 if either N302 or R603. $X_{ct=0}$ is vector of community level controls measured at baseline. The parameter of interest is α_3 which measures the differential change in mean outcomes between communities along rehabilitated roads and communities along non-rehabilitated roads. The inclusion of vector $X_{ct=0}$ controls for observable factors driving a wedge between these two types of communities in their underlying time trends. Throughout all specifications standard errors will be clustered at the community level.

2. Double-Difference using Time and Distance to Roads

Second, an alternative DD approach exploits proximity to the rehabilitated roads. This method only uses data from households in communities along rehabilitated roads, and avoids the common time-trend assumption between rehabilitated and non-rehabilitated roads used in the DD approach described above. It compares changes in outcomes between communities that are in close proximity to the rehabilitated roads and communities that are further away [Ghani *et al*. 2015]. The identification assumption here is that had the rehabilitation intervention not occurred these two types of communities would have been on a similar pattern of evolution.

In practice this method is estimated with the following specification using pre- and post-intervention data on household in communities along roads R604/R605:

$$\Delta y_{hcr} = \beta_0 + \beta_1 D_{cr} + \beta_2 X_{ct=0} + \Delta\varepsilon_{hcr} \tag{B.2}$$

where D_{cr} is a distance dummy that equals 1 if community c is 0-2 km from the road, and 0 if 2-10 km away. The parameter interest is β_1 which measures the differential change in mean outcomes between communities in close proximity to roads R604/R605 and communities at some distance from roads R604/605. As before, the inclusion of vector $X_{ct=0}$ controls for observable factors driving a wedge between these two types of communities in their underlying time trends.

3. Triple-Difference using Time, Comparison Roads and Distance to Roads

Third, a more robust method is to use a triple-difference (DDD) approach that combines variation across time, across rehabilitated and non-rehabilitated roads, and across distance to roads. This method compares changes in outcomes between communities that are in close proximity to roads and communities that are further away, both along rehabilitated and non-rehabilitated roads. It relaxes the common time-trend identification assumptions of the two DD models described above. Under this approach the identification assumption is instead that in the absence of the project, the effect of proximity to roads on changes in outcomes over time is constant across rehabilitated and non-rehabilitated roads.

This method is estimated with the following specification using pre- and post-intervention data on household in communities that are close and not so close to both roads R604/R605 and roads N302/R603:

$$\Delta y_{hcr} = \gamma_0 + \gamma_1 R_{cr} + \gamma_2 D_{cr} + \gamma_3 [R_{cr} \times D_{cr}] + \gamma_4 X_{ct=0} + \gamma_5 [X_{ct=0} \times R_{cr}] + \gamma_6 [X_{ct=0} \times D_{cr}] + \Delta\varepsilon_{hcr}. \tag{B.3}$$

The coefficient of interest is γ_3 which can be shown to equal the change across time in mean outcomes for households in communities located 0-2 km from roads R604/R605, net of changes in mean outcomes for households in communities located 0-2 km from roads N302/R603 and of changes in mean outcomes for households in communities located 3-10 km from roads R604/R605.

4. Straight-Lines Instrumental Variables

Finally, as a robustness check we will follow an approach similar to Banerjee *et al.* [2012] and Ghani *et al*. [2015] and use an instrumental variable (IV) approach based on straight-line distances between termini communities. Specifically we will instrument for being 0-2 km from roads R604/R605 with being 0-2 km from a straight line between the termini communities of these roads. The identification assumption of this method is that proximity to the straight line only affects communities in the post-intervention period due to the likelihood of the community being in close proximity to the roads experiencing the rehabilitation upgrade.

In practice this method is estimated with the following system of equations:

$$D_{cr} = \pi_0 + \pi_1 L_{cr} + \pi_2 X_{ct=0} + \varepsilon_{cr} \tag{B.4}$$

$$y_{hcrt=1} = \delta_0 + \delta_1 D_{cr} + \delta_2 X_{ct=0} + \varepsilon_{hcrt=0} \tag{B.5}$$

where L_{cr} in equation (C.4) is a dummy that equals 1 if community c is 0-2 km from the straight line, and 0 if 2-10 km away. This method only uses data from communities along roads R604/R605. Equation (B.4) only uses community level pre-intervention data, and equation (B.5) only uses household and community level post-intervention data. We use equation (B.4) as a first-stage to predict how distance to the straight-line predicts distance to the roads. Our parameter of interest is δ_1 in equation (B.5) in which D_{cr} has been instrumented with L_{cr} using equation (B.4) as our first-stage.