Appendix D: Pre-Analysis Plan, Study Design, and Methods

In what follows, we report the Pre-Analysis plan for the study.

Introduction

Previous studies of microfinance, which focused on microcredit alone, have not found large effects on poverty alleviation (see for example, Banerjee et al.,2014) or households' ability to cope with shocks. In contrast, non-experimental evaluations of rural bank branches in India report significant poverty reduction (for example, Burgess and Pande, 2003). The intervention we implement provides a unique opportunity to undertake rigorous experimental evaluation of the impacts of bank branch expansion in rural areas at both the household and village level. Importantly, we will evaluate a financial service delivery model that uses bank branches in villages to provide a full range of credit, saving, and insurance services to entire communities. Our evaluation is also unique in the breadth of its outcome measurement and covers asset ownership, business investment, farming, migration, trust relationships, and health outcomes. An in-depth understanding of the impact of using rural bank branches to provide comprehensive financial services will inform policy on financial inclusion both in India and globally.

Our implementing partner is a large financial institution in rural South India (referred to as LFI below). It comprises of a group of Indian non-banking financial companies with the mission to “maximize the financial well-being of every individual and every enterprise in remote rural India by providing complete financial services”. LFI offers financial products spanning loans, savings, and insurance in addition to tailored financial advice through local village branches, in order to effectively reach individuals in financially marginalized rural communities. The LFI model is an alternative to the standard microfinance movement in India, which has focused primarily on microcredit.

The intervention to be evaluated is being implemented by a company founded by the LFI which operates, amongst other areas, in our study districts Thanjavur, Ariyalur and Puddukottai in Tamil Nadu. The base of the intervention is the expansion of bank infrastructure across villages. The construction and ongoing operation of the 50 bank branch locations that will constitute the intervention’s treatment group is funded by the LFI.

In this pre-analysis plan, we will focus on outcomes in six areas: Financial Access, Income and Wealth, Consumption Smoothing, Human Capital Investment, Labor Market Outcomes and Female Empowerment. Additional survey components connected to this set of surveys look at Social Network, Health and Farming Technology Outcomes.

The structure of this pre-analysis plan is as follows: section 2 gives an overview of the experimental design, sampling and the key data sources used. Section 3 specifies the main regression model, how standard errors will be computed and which basic and extended controls will be used. Section 4 will lay out the various hypotheses we aim to test while section 5 deals with the analysis of treatment heterogeneity. We do not exclude the possibility of running further analyses for the final paper, but will make clear which estimations were specified in the plan and which were not (cf. Casey, Glennerster and Miguel, 2012).

Study design

Sampling

We will use data from 50 service area pairs[1] across three districts. The average service area spans several villages in a radius of 3-5 km from the branch and covers a population of an estimated 10,000 people. Below we first describe randomization and then surveying.

The selection of potential branch sites and randomization across them proceeded as follows: In conjunction with the bank, potential location sites were identified using a global position system (GPS)-based population survey which determined relevant political, administrative and social boundaries. Once all feasible branch locations in the district had been designated, we used Edmond's algorithm for minimum distance matching to construct pairs of service areas. This matching for treatment and control allows the study to overcome issues in seasonality and geographic correlation in outcomes by minimizing differences between paired branches. It also improves balance across treatment and control villages on observed and unobserved factors, and provides a strong service-area-level control variable. For several 2001 census village outcomes (including caste composition, number of primary schools, water facilities and proportion of irrigated land), we find that controlling for pair fixed effects explains roughly 70% of the variance. One service area in each pair was then randomly selected to receive a bank branch first (treatment area). Expansion in the other area (control) will be delayed for 36 months. Bank employees are not informed about the study or whether their branch is a study branch or not. We will assess the impact of increased formal financial access by comparing treatment areas with control areas two years after branch opening in treatment areas. Treatment and control areas of the same pair will be surveyed simultaneously. Surveyors do not know the treatment status of villages and are rotated across treatment and control.

The opening of bank branches happened in three rounds due to operational constraints following the Indian microfinance crisis in late 2010. We will account for this fact in the empirical analysis and may, at times, restrict the analysis to certain rounds only, for example when looking at the influence of the Pradhan Mantri Jan Dhan Yojana (PMJDY) scheme which started in 2014.

Selection of households within service areas

In each service area, a total of 46 households were selected for inclusion in the household survey. The selection of households generally followed a two-stage design to account for clustering of households in villages, while ensuring that the sample is representative of the chosen service areas.

The first stage employed a probability proportional to size (PPS) sampling of villages within service areas. That is, villages were drawn to be included in the sample according to their relative population size. Additionally, the center village with the intended branch location was included. Each service area was allocated 46 baselines which were divided evenly into portions, and villages were drawn to be included in the sample according to their relative population size. Additionally, the center village with the intended branch location was always included in the baseline selection.

In stage two, listing was conducted with a 5-household skip in all villages sampled during stage one, collecting residential addresses and information for identification purposes, such

[1] 101 service areas are covered, due to one triplet with two control areas.

as names and occupations of household members. We dropped all households that did not include a woman between the ages of 18 and 55. We then randomly selected the number of households in each village that had been determined in stage one.

Key data sources

Baseline surveys (prior to the intervention, starting in September 2010):

The baseline household survey occurred in each pair prior to the opening of the treatment branch. Attrition has been below 5% and not differential across treatment and control group. After an initial wave of 8 branch openings in 2010, branch openings only continued at the start of 2013 and lasted until September 2015 due to operational constraints and delay in permission to open following the microfinance crisis in late 2010. Delay on the side of branch openings resulted in a further, shorter break in branch openings. Overall, 8 service areas were opened during round one, 34 during round two and 8 during round three. In total, 4,391 households were approached over the three rounds of baseline surveys.

Besides the baseline household survey, we conducted a short survey to the Panchayat head about basic village characteristics.

The Household survey itself consists of two parts with several sections each. For the first part, the head of the household or their spouse is interviewed, while for the second section part, we interview the female spouse. The second part is administered to women as it contains questions on female empowerment and child health amongst others.

Endline surveys (18-24 months after branch opening in treatment areas, starting in September 2013. Scheduled to run until mid-2017):

In the endline household survey, the same data is collected from the respondents interviewed at baseline, with adjustments for relevant external changes (e.g. the introduction of Aadhaar identification cards or voter ID cards that led to additional questions) or to clarify meaning.

LFI Customer Management System data

We will also use administrative LFI client data to augment the quality of self-reported survey measures. Monthly data on enrollment (registering in a branch), product take-up, frequency of loan renewal and other outside loans taken can be used to give a more detailed view of development of treatment over time.

Methods

Estimation of treatment effects

Our main specification will model the effect of the randomized treatment, which is increased access to formal finance through the opening of a LFI bank branch in a service area. Given the use of pair-wise matching described earlier in assigning randomization, we will control for pair-wise fixed effects. Drawing on the endline household survey, we will thus estimate the following model:

$$Y_{ik} = \propto_0 + \propto_1 T_k + \propto_2 S_k + \delta_{pk} + \varepsilon_{ik}$$

In the regression above, *i* indexes the individual or household and *k* indexes her service area.

Y_{ik} is a given outcome (e.g. amount of formal savings for instance) for individual or

household *i* in service area *k*. T_k is the service area treatment dummy, such that $\propto_1$ gives the intent-to-treat effect. S_k are survey round dummies, δ_{pk} are pair fixed effects and ε_{ik} is the idiosyncratic error term.

Standard errors will be clustered at the level of randomization, i.e. at the service area level. In case we can draw on a subset of pairs/clusters only, we will compute wild bootstrapped standard errors clustered at the service area level following Cameron et al. (2009).

Additionally, we will draw on both baseline and endline household survey data and estimate intent-to-treat (ITT) effects in a difference-in-differences (DID) framework, using the following specification that considers changes over time in our panel:

$$Y_{ik\theta} = \propto_{0\theta} + \propto_1 T_k + \propto_2 S_k + \propto_3 \theta_k + \beta\theta_k * T_k + \delta_{pk} + \varepsilon_{ik}$$

where θ is the endline/baseline indicator and β the DID coefficient. Otherwise notation follows the basic cross-section regression above. We will additionally consider a specification using a vector of controls and pair fixed effects:

$$Y_{ik\theta} = \propto_{0\theta} + \propto_1 T_k + \propto_2 S_k + \propto_3 \theta_k + \beta\theta * T_k + \delta_{pk} + X_{ik\theta} + \varepsilon_{ik}$$

Individual, household and service area level controls

Besides estimating the model as described above, we will estimate it controlling, additionally, for several baseline characteristics in order to improve the precision of our estimates. In order to ensure that we do not obtain spurious results based on the inclusion or exclusion of controls, we will conduct a robustness analysis.

The basic set of controls at the individual and village level will be:

Basic demographic controls: **{**age, education, land ownership, caste, religion**}**

Basic village level controls**: {**village size, distance to branch village (real – for treated village; or hypothetical – for control villages)**}**

To explain heterogeneity in treatment results, we will employ an extended set of controls, namely:

Extended demographic controls: **{**risk aversion, experience of income shocks or major health shocks, business ownership, number of children, gender of respondent**}**

Extended village level controls**: {**road connectivity, distance to next factory, cost of travel to next city, number of financiers, number of moneylenders, average (agricultural and non-agricultural) wage level, share of land irrigated in village**}**

If any of these control variables has an equal to or larger than 95 percent share of uniform answers at baseline, we will only include them if we have reason to believe that they have a strong connection to the specific outcome indicator and mark this exception in the analysis.

Procedure for accounting for attrition, non-response, questions with limited variation and extreme values

Survey attrition

If a sampled household is not found after at least three attempts, it is not replaced and counted towards general attrition. Rates are calculated as the share of all non-surveyed households (whether not found or non-consenting) of the initially sampled households.

We will then test if attrition is differential across treatment and control areas. If attrition is found to be related to treatment status at the 5 percent significance level, we will employ a bounding method to obtain ranges on our treatment estimates which are robust to this attrition.

Missing data from non-response to individual questions

No imputation for missing data from item non-response at follow-up will be performed. We will check whether item non-response is correlated with treatment status following the same procedures as for survey attrition, and if it is, we will construct bounds for our treatment estimates that are robust to this. In the regression analysis, we will replace missing observations by 0 and generate an associated dummy.

Questions with limited variation

Questions for which 95 percent of observations have the same value within the relevant sample will be omitted from the analysis and will not be included in any indicators or hypothesis tests, in order to limit noise in the analysis. Should this omission rule result in the exclusion of all relevant variables for an indicator, we will not calculate the indicator.

Extreme values or outliers

Extreme values or outliers are identified using the three-sigma rule (also called the 68-95-99.7 rule). We top code variables at 3 standard deviations, meaning that for values outside an interval defined as +/- 3 standard deviations of the mean, values are set at the upper or lower bound of the interval. When trimmed, values outside this interval are set to missing. Another alternative method is to top code variables at the 99^{th} percentile. We use both methods in order to ensure the robustness of our results. In the latter method, we top code the top 1% of the distribution, meaning that values at the top are set to the value of the 99^{th} percentile. When trimmed, values at the top of the distribution are set to missing. These usual statistical procedures are used in order to ensure that outliers do not drive the results.

Procedure for dealing with multiple outcomes

We will aim to account for the effects of multiple, correlated outcomes by grouping our outcome measures into domains, based on the idea that items within a domain are measuring an underlying common factor. Our six domains are detailed below. Then we will sign the outcomes within each domain, so that the hypothesized effects go in the same direction, and take a standardized treatment effect within that group (compare Kling et al., 2007, Finkelstein et al., 2010).

Hypotheses

We will group our hypotheses into six broader categories for which we will try to account for effects of multiple hypothesis testing (see section 3.3 above):

1. *Financial Access*
   1.1 Increase in formal financial activity
   1.2 Change in informal borrowing

2. *Income and Wealth*
   2.1 Increase in riskier but higher return activities and asset investments
   2.2 Increase in savings and wealth
   2.3 Change in diversification of financial activity
3. *Consumption smoothing*
   3.1 Increase in expenditure on durable goods
   3.2 Change in expenditure on non-durable goods
   3.3 Increase in formal borrowing or savings used in response to a shock
   3.4 Change in informal borrowing or savings used in response to a shock
   3.5 Smaller consumption decrease and volatility in response to a shock
4. *Educational Investment*
   4.1 Increase in expenditure on education
5. *Labor Market outcomes*
   5.1 Increase in entrepreneurship/ business ownership
   5.2 Reduction in unemployment
   5.3 Change in permanent migration
   5.4 Change in seasonal migration
6. *Female empowerment*
   6.1 Increase in women reporting having a source of income
   6.2 Increase in women reporting being able to make joint or individual decision in financial household issues
   6.3 Increase in women reporting being able to make joint or individual decision in non-financial household issues
   6.4 Increase in subjective well-being of female household members

Outcome response to treatment heterogeneity

A key interest of the study is to find out mechanisms of impact. We therefore aim to identify the channels through which household and village-level outcomes change for different groups, which will help improve the design of products and services for specific household types. Using the specification outlined above, we will examine treatment effects for the following subgroups:

- By household type: households with landholding, female-headed households, households with school-aged children, households with a literate household's head, income quintile analysis.
- By occupation: farmers, wage laborers, business owners.
- By treatment status (treatment-on-treated) or likelihood of take-up as predicted by baseline variables at endline.