External Validity and Implementation at Scale: Evidence from a Migration Loan Program in Bangladesh

Harrison Mitchell[1], A. Mushfiq Mobarak[2], Karim Naguib , Maira Emy Reimão[3], and Ashish Shenoy[4*]

[1]UC San Diego
[2]Yale University
[3]Villanova University
[4]University of California, Davis
[*]Corresponding author. Email: shenoy@ucdavis.edu.

This version: October 2022
PRELIMINARY AND INCOMPLETE

Abstract

This paper presents results from an experimental evaluation of a large-scale migration loan program that offers a short-term low-interest migration loan to hundreds of thousands of landless rural workers during the agricultural lean season in northern Bangladesh. Pilot evaluations found the loan offer to increase the rate of temporary migration by 25–40 percentage points. At scale this effect falls to only 6 percentage points, with a 12 percentage point increase in migration in the regions where the pilot took place and no effect in newly treated regions despite participants being observationally similar across space. To account for treatment effect attenuation, we introduce a theory of delegation risk that leads implementing agents to systematically mistarget intended program beneficiaries. Mistargeting occurs because program benefits are concentrated among those induced to migrate by the loan offer, but capacity constraints at scale lead effort to be directed toward those already planning to migrate without a loan. We present evidence consistent with this theory that the characteristics that predict pre-loan migration are strongly correlated with the likelihood of remembering the loan offer, and show that delegation risk can quantitatively account for the diminished treatment effect. We rule out two alternative explanations: First, the geographically clustered randomization design reveals that there is, if anything, general equilibrium crowd-in rather than crowd-out of migrants. Second, we estimate conditional treatment effects and find that changes in population characteristics over time have little impact. The mistargeting identified in this study has the potential to undermine the effectiveness of a number of common development policies.

JEL Codes: O18, C93, J43, R23

1 Introduction

Program evaluation has become a prominent feature of policy design. Careful measurement of program effects can assist with cost–benefit analysis and ensure resources are directed to their most effective ends. In practice, impacts are frequently first established in small-scale pilot experiments before widespread adoption. Experimental evaluation is seen as the gold standard for generating unbiased, internally valid estimates of program effects. Both implementers and funders increasingly demand that evaluation strategies be included in the design of new policies.

Pilot programs that show success typically grow in scale to reach more beneficiaries. However, rigorous evaluation of policies scaled up from pilot has repeatedly found diminished program effectiveness. Failure to scale has been documented in education (Kraft et al., 2018; Bold et al., 2018; Andersen and Hvidman, 2020; Evans and Yuan, 2020; Ganimian, 2020; Kerwin and Thornton, 2021; Bellés-Obrero and Lombardi, 2022), public health (Fund, 2018; Cameron et al., 2019), early childhood intervention (Araujo et al., 2021; Bloem and Wydick, n.d.), microfinance (Giné et al., 2021), and behavioral nudges (DellaVigna and Linos, 2022; Rabb et al., 2022). Across a broad range of development policies, (Vivalt, 2020) reports a systematic negative relationship between a program's scale and the size of its impacts.

In this paper, we report experimental results from a large-scale implementation of a migration loan program in Bangladesh. The large positive effects demonstrated in pilot evaluations of this program fail to replicate at scale even though the policy itself remained unaltered. Implementation at scale was intentionally designed to measure changes caused by general equilibrium spillovers and by expansion to new geographic regions, but these two factors can explain at most a third of the decline in effectiveness. We introduce a new theory of delegation risk that causes attenuation of program impacts at scale due to management strategies commonly employed by development organizations that lead to systematic mistargeting of the intended beneficiaries.

The migration loan program is motivated by the prevalence of seasonal poverty in northern Bangladesh. Both food supply and labor demand in rural parts of the region are strongly tied to the seasonal agricultural cycle. This cycle generates an annual "hungry season" in September–October shortly before the main crop harvest when food prices are high and rural wages are low. Poor households that are unable to save in anticipation of this shock regularly experience sharp declines in consumption and food security during this period (Khandker, 2012).

Temporary migration is a common response to seasonal poverty. Many households in the region of study send a member to work elsewhere in the country during the hungry period. The No Lean Season (NLS) loan program aims to enable this option for households currently unable to finance it. The program offers low-interest, short-term loans to landless rural households ahead of the hungry season. The loan, sufficient to cover the cost of transportation and a few nights' lodging for one household member, is issued with the soft conditionality that recipients use it for labor migration.

Two rounds of pilot evaluation show NLS loans to have large, positive returns. In pilot experi-

ments conducted in 2008 and 2014 with 1,292 and 5,764 households, respectively, loan availability raised the fraction of households who sent a migrant during the lean season by 25–40 percentage points, from a baseline rate of around 35 percent. Local average treatment effect estimates of the return to migration indicate that households enabled by the loan earn nearly 50% more over the following months, and this income translates to consuming roughly 500 more calories per person per day during the lean season months (Bryan et al., 2014; Akram et al., 2017).

The first contribution of this paper is to report that these large program impacts unfortunately fail to replicate at scale. The NLS program expanded to encompass over 150,000 households per year in 2017 and 2018, encompassing 5% of the population in the implementation region. This expansion was rigorously evaluated in a randomized controlled trial, with village-level randomization into migration loan eligibility. The NLS program in the expansion rounds induced only a 6 percentage point increase in migration among those offered the loan, far below what was observed in piloting.

Decomposing the treatment effect by location reveals differences between the original two districts where the pilot took place and six new districts included at scale. In the most recent evaluation, NLS loan eligibility increases the propensity to migrate by 12 percentage points in villages from the original pilot districts, but has no effect on migration in new districts. While site-specific heterogeneity is commonly observed in development policy (e.g. Pritchett and Sandefur, 2015; Meager, 2019), we find this difference surprising in our setting due to the fact that expansion districts are geographically adjacent and the treated population is observably similar in baseline characteristics. There was no indication from (lack of) differences in income, education, household size, or migration history that treatment effects may be diminished in new districts. Site-based attenuation is consistent with other literature suggesting initial policy implementation frequently targets populations with the greatest propensity for success (Allcott, 2015; DellaVigna and Linos, 2022). Nevertheless, geographic heterogeneity only accounts for up to a third of the attenuation in treatment effect at scale.

The main contribution of this study is to quantify the reasons behind the decline in effectiveness at scale of the same migration policy in the same location as the pilot. Our experimental design enables us to explicitly measure general equilibrium spillovers, and we further explore the impact of changes in population characteristics over time. We find little evidence that these two factors play a role in program effectiveness. Instead, survey evidence from followup surveys on implementation quality reveals how program administration at scale may inadvertently lead implementers to systematically exclude those who would benefit most from a migration loan. Enough loan-eligible households are excluded at scale for mistargeting to account for all of the attenuation in treatment effect.

We present three pieces of evidence from debriefing surveys and administrative records consistent with systematic mistargeting of intended program beneficiaries. First, in a debrief of implementing agents after the 2017 round of implementation, we find strong evidence that loan officers targeted a specific number of loans to disburse. 85% of officers report being given a specific target by the

implementing organization, and the majority disburse close to their target number of loans. This behavior leads loan disbursement to nearly cease in the aggregate once the aggregate target is reached. In 2018, we gave explicit instructions to avoid numerical targets and see a greater rate of loan disbursement.

Second, we show evidence that capacity constraints limited program uptake at scale, even without explicit implementation targets. The biggest difference in loan-induced migration between pilot and the two years of at-scale implementation is in the fraction of the loan-eligible population that accept a loan. Rates of loan disbursement after acceptance and of migration conditional on disbursement are nearly identical across evaluation rounds, indicating that attenuation of program effects was primarily due to the loan offer/acceptance phase.

Third, among loan-eligible households, we find that the program was selectively advertised more intensely to those most likely to migrate. In a debrief survey of households assigned to treatment, only 40–60 percent can recall activity by the lending organization related to migration loans. This propensity to recall program implementation is strongly predicted by baseline characteristics such as migration history and number of adult males in the household, which are also strong predictors of the likelihood of migrating without a loan. To the extent that program recall is a proxy for outreach by implementing agents, this finding suggests that implementation at scale was selectively directed toward those that were most likely to be regular migrants.

These three facts point to a new theory of delegation risk that can lead to the systematic mistargeting of intended beneficiaries. This concern arises because program benefits are concentrated among the subset of the population who would not normally migrate but are enabled to do so with a loan. This subset, referred to as "induced migrants" or, in the language of program evaluation, "program compliers", are the intended beneficiaries for whom the program unlocks the high returns to migration during the lean season. In contrast, regular migrants (i.e. always-takers) who would have migrated anyway and never-migrants (i.e. never-takers) who would not migrate under any circumstances both derive little benefit from a small, short-term loan that does not alter their migration status.

This classification highlights the possibility for mistargeting because program compliance cannot directly be observed. In program evaluation, we can only statistically estimate the size of the complier population by measuring the difference in migration rates between treatment and control. For any given household that migrates with a loan, it is impossible to distinguish whether they are a program complier or always-taker. Likewise, it is impossible to distinguish compliers from never-takers among the set of untreated households that do not migrate. As a corollary, when implementation decisions are delegated to implementing agents, incentives cannot be explicitly tied to reaching program compliers. Instead, development organizations typically monitor and evaluate implementing agents based on implementation targets; in our case, loan officers were evaluated and retained based on the number of migration loans they awarded.

Mistargeting arises when implementation-based incentives for agents interact with resource

constraints during implementation. This was not a concern during piloting as the small scale allowed for advertising the migration loan to all eligible participants with sufficient intensity. In scaling up, the geographic spread, number of implementing agents, and caseload of eligible households per implementing agent all increased by 1–2 orders of magnitude. With the same short window for loan disbursement ahead of the lean season, loan officers could not advertise to all eligible households with equal intensity and instead had to be selective in their efforts.

In settings with high fractions of always-takers and never-takers, delegation of authority to resource-constrained agents motivated by implementation-based incentives will lead systematic undertreatment of program compliers in favor of always-takers. This is because a loan offer made to a household already planning to migrate (i.e. an always-taker) guarantees success for the implementing agent, while a loan offer made to a household not yet planning to migrate only succeeds if the household is an induced migrant (i.e. complier) but runs the risk of failure if the household turns out to be a never-migrant (i.e. never-taker). As a result, if implementing agents can observe any information about pre-loan migration plans, then they maximize their loan success rate by selectively directing effort toward always-takers.

This type of selection may also be induced by participant demand. Since regular migrants already have plans in place, they will likely be the first in line to take advantage of a low-interest loan. In contrast, induced migrants need time to scout opportunities and make plans before securing a loan. There is a risk that regular migrants exhaust program capacity before induced migrants have a chance to signal their demand, leaving some program compliers unable to access program benefits. In either case, systematic mistargeting will cause treatment effects to attenuate with program size as long as there exists some sort of resource constraint that gets tighter as a program grows, leading to more selectivity in treatment. We quantify the extent of such mistargeting necessary to generate the attenuation observed in our setting and show that it falls well within the range of plausibility.

The necessary conditions for delegation risk are quite general and common to many categories of development policy. Systematic mistargeting threatens implementation at scale when benefits are concentrated among program compliers, there are high fractions of always- and never-takers within a population, program compliance cannot be readily observed, and capacity constraints lead to selective variation in treatment or outreach intensity. In addition to directed lending such as NLS, these factors are present in microfinance more generally, conditional cash transfers, occupational training, agricultural extension, and a number of other common policies designed to enable or encourage beneficial behaviors. Importantly, systematic mistargeting can threaten any policy scale-up under these conditions independent of policy design or implementer quality.

We rule out two alternative explanations for why NLS treatment effects may be diminished at scale. First, there is no evidence of negative general equilibrium spillovers. In principle, large-scale policies may shift market prices or other macroeconomic conditions in ways that alter the policy impact (e.g. Cunha et al., 2018; Sraer and Thesmar, 2018; Egger et al., 2022; Khanna, 2022). In our study, this concern manifests as the possibility that mass migration lowers wages at the destination,

limits the number of available migration opportunities, or otherwise crowds would-be migrants out of the labor market. We explicitly test for this possibility with a novel clustered randomization design that leaves individual untreated villages amid intensely treated areas. We find that adjacency to treated villages has a small, positive effect on migration that is statistically indistinguishable from zero. This result indicates that local migration opportunities are not exhausted, destination labor demand remains sufficiently elastic, and migrants are if anything crowded in, not crowded out, by general equilibrium spillovers at the scale of this study.

Second, we find no evidence that the effect of the NLS program diminished due to changes in population characteristics over time. Rosenzweig and Udry (2019) demonstrate how the effects of economic policy can be sensitive to changes in the macroeconomic environment. In our setting, we observe two main sources of change over time. First, the two years of at-scale implementation saw substantially more rainfall ahead of the agricultural harvest. Second, the population of program participants grew wealthier. We evaluate the potential effect of these and other changes by using a machine learning algorithm to estimate treatment effect heterogeneity by baseline characteristics in the pilot experiment. We then simulate the treatment effect we would have observed at scale holding conditional treatment effects constant but allowing the distribution of population characteristics to evolve. This exercise generates counterfactual results very similar to the actual effect measured in pilot rounds, indicating that changes in observable population characteristics over time do not account for the decline in program effectiveness.

This paper contributes to the small but growing number of randomized evaluations of large-scale development programs. Other recent evaluations include conditional cash transfers (Schultz, 2004; Rivera et al., 2004), public assistance (Muralidharan et al., 2016, 2017; Cunha et al., 2018; Egger et al., 2022), education policy (de Ree et al., 2017; Bold et al., 2018; Khanna, 2022), police reform (Banerjee et al., 2021), and celebrity messaging (Alatas et al., 2019). We demonstrate the value of such large-scale evaluation by quantifying how treatment effects differ across evaluations of the same program implemented at different scales. Linking these various evaluations allows us to quantify the importance of sources of change from pilot to scale. Economic explanations for the relationship between effectiveness and scale typically center around changes in program implementation, changes in target population, or general equilibrium spillovers (see Banerjee et al., 2017; Al-Ubaydli et al., 2017, 2019). All three factors can threaten the external validity of pilot evaluations, limiting their informativeness for at-scale implementation.

Delegation risk can be seen as a counterpart to endogenous participant response. Chassang et al. (2012); Bulte et al. (2014) establish a relationship between program effectiveness and (often unobserved) effort taken by program participants. Such effort may be motivated by knowledge of the program and beliefs about its effectiveness. In this study we highlight the role of implementers in disseminating such knowledge, and present evidence of one channel through which their efforts may miss the intended beneficiaries.

Mistargeting in our setting arises from the inability to tie the incentives for implementing agents

directly to treatment of intended program beneficiaries. This concern arises in a number of other settings where implementer incentives have been shown to affect policy outcomes. Empirical studies show the importance of incentive design in banking (Hertzberg et al., 2010), health worker attendance (Dhaliwal and Hanna, 2017), environmental audits (Duflo et al., 2013), and tax collection (Khan et al., 2015; Balán et al., 2022). In each of these cases, the implementing agent mediates an adversarial relationship between the policymaker and the program participant. For instance, a tax collector balances the government's interest in raising revenue against the taxpayer's interest in minimizing payment. Our study extends this investigation to a setting in which policymakers' and participants' interests are perfectly aligned: both want to enable migration among induced migrants. We show that even with this alignment of interests, misaligned incentives for the implementing agent can undermine policy goals.

Finally, this paper sheds light on the household response to seasonal liquidity constraints. In rural areas around the world, seasonality in agriculture leads to regular periods of economic distress that lead households to take costly actions to satisfy immediate consumption needs. Seasonal constraints have been shown to limit agricultural investment (e.g. Duflo et al., 2011), food storage and sales (Stephens and Barrett, 2011; Basu and Wong, 2015; Burke et al., 2018), and labor market activity (Bryan et al., 2014; Akram et al., 2017; Fink et al., 2020). We show that the mere availability of credit is not sufficient to alleviate these constraints as lack of knowledge or outreach can substantially limit uptake.

2 Background and Data

2.1 Setting

This study takes place in rural parts of the Rangpur Division in northern Bangladesh. This is a poor and largely agrarian part of the country, with an urbanization rate under 15%. Among the rural population of Rangpur, 48% of households were classified as moderately or extremely poor in 2016, compared to only 24% for the country overall. A map of the region of study is provided in Figure 1. Pilot evaluation rounds included villages from the Kurigram and Lalmonirhat districts, and the at-scale implementation expanded to include these as well as the six other districts in the division.

Rural economies in this area are characterized by strong seasonality tied to the agricultural cycle. The primary crop season lasts from planting in June through harvest in November and December. Labor demand, and correspondingly the agricultural wage, peaks at these endpoints but remains low between the start and end of the season. Low wages are accompanied by high food prices as stocks dwindle ahead of the harvest. (Khandker, 2012)

Low wages and high prices combine to create an annual period of heightened food insecurity during the agricultural lean season. September through early November in this part of the country is locally referred to as "monga", which translates to "hunger season". As shown in Figure 2, more

than half of landless rural households report reducing meals or portion sizes during this period, and nearly a fifth do so for more than fifteen days per month. The fact that this level of deprivation occurs with annual regularity indicates that many vulnerable households have little capacity to use savings or other methods to smooth consumption over the year.

Many households turn to short-term, intra-national migration as a response to agricultural seasonality. A third to half of landless households in our region of study use migration to supplement earnings during the lean season months. The typical migration episode involves 1–2 household members, almost always male, traveling for work to destinations within Bangladesh. The typical migration episode lasts for 2–3 months, with migrants bringing their earnings home in cash. Comparably high rates of lean season migration can be observed in many rural parts of the world.

Agriculture is the most common destination sector of work for seasonal migrants from the Rangpur region. Roughly half of seasonal migrants find agricultural work in other parts of the country where the planting and harvest seasons are offset due to climate. Among the remaining half that travel to urban destinations, a third work in the transportation sector (i.e. pulling cycle rickshaws) and another third find employment in low-skill construction, both sectors that are far less sensitive to agricultural seasonality. The capital city of Dhaka accounts for nearly 30% of migration to urban areas, with the rest spread throughout other cities around the country.

2.2 No Lean Season Program

While migration during the agricultural lean season is fairly common in the Rangpur region, many households are unable to access this option due to liquidity constraints. The issue arises because migration requires up-front financing—for transportation and initial lodging—to realize a subsequent stream of labor market returns. This financing requirement comes at exactly the time of year when households that rely on agricultural labor and their local social network have the least available cash on hand. The challenge of migration is compounded by the fact that those who stand to benefit the most from migration fall closest to the threshold for subsistence. Therefore, even though they have the highest marginal return to increasing consumption, they also face the greatest risk from financing a migration attempt that turns out to be unsuccessful.

The No Lean Season (NLS) program aims to enable seasonal migration for a larger fraction of the rural landless population through credit access. NLS offers a short-term, zero-interest loan of BDT 1,000 (around $12 USD) for migration during the agricultural lean season. This value is sufficient to cover bus fare and a few nights' lodging for one household member to get established in a destination labor market. Loans are offered in early September at the start of the agricultural lean season, with a duration of 3–6 months for repayment.

Eligibility for NLS loans is based on exposure to the agricultural lean season. Households are eligible if they own under 0.5 acres of land, meaning their primary earnings must come from the labor market, and they self-report having reduced or missed meals in prior lean seasons. Roughly sixty percent of rural households satisfy these eligibility criteria in our region of study.

NLS loans are issued with soft conditionality. The loans are marketed as intended for the purpose of migration. Upon disbursement, loan officers follow up with recipients to inquire about migration status and encourage travel for those that have not yet departed. However, there is no penalty for failure to migrate nor any other formal enforcement of the migration requirement.

2.3 Program Implementation and Evaluation

The NLS program was implemented over three rounds of randomized evaluation that expand progressively in scale. The initial pilot to establish viability of migration loans took place in 2008. In this round, 1,292 households in 68 villages were offered migration loans[1], out of a total study population of 1,900. In villages where loan offers were made, offers comprised on average 14% of the eligible population. Eligibility surveys and outreach to participants were conducted by the same evaluation team that collected survey data on outcomes, as is typical in pilot studies. Full details and results are reported by Bryan et al. (2014).

The second evaluation round in 2014 maintained close to the same number of villages, but increased treatment intensity within village. In this round, loans were offered to 5,764 households in 95 villages with a comparably sized control group for comparison. Within villages where offers were made, treatment intensity as either 14% and 70% of eligible households. This variation was introduced to measure within-village spillovers as treatment intensity increased. Eligibility was determined by the evaluation team, and loan offers were made by a local microfinance organization with close coordination and oversight from the evaluation team. Full details and results are reported by Akram et al. (2017).

In this paper we present results from a third round of evaluation at scale that took place in 2017 and 2018. This round differs from prior pilots in two key ways. First, the scope of the program was greatly expanded, with 158,014 loan-eligible households in 734 villages in 2017 and 143,721 loan-eligible households in 2018, comprising just over 5% of rural households in the region each year. Within each village where loan offers were made, all eligible households were offered a migration loan. Second, implementation by the local microfinance organization was decoupled from evaluation. Implementers coordinated with evaluators only in identifying villages where loan offers would be made and setting thresholds for household eligibility. Once these design decisions were in place, the implementing organization conducted its own survey to determine household eligibility and make loan offers, and the evaluation team independently identified and surveyed a sample of households for evaluation.

Implementation at scale was conducted by a local microfinance organization with 110 branch offices spread throughout the Rangpur division. Each branch's catchment area was defined to be the set of villages within a one-hour bike ride from the branch office. Prior to implementation at scale, each branch conducted a census of villages in its catchment area, and this set of villages

[1] 598 households were offered loans and 703 offered conditional grants, but the treatment effect was early identical between these two groups.

makes up the study population. We randomly select a subset of these villages in which to make loan offers, and all eligible households in a selected village are offered a migration loan.

For evaluation purposes, we designate villages in which to offer loans using a two-level randomization design. At the first level, we randomly divide microfinance branch offices into a treatment group that makes loan offers and a control group that does not. 40 branch offices were assigned to treatment in 2017 and 50 in 2018, with treatment status reassigned between years. Villages in the catchment area for control branches are designated as "pure control" with no loan offers made in or nearby. Randomization at this level was stratified by district to enable comparison of treatment effects between pilot and expansion districts.

At the second level of randomization, we select a subset of villages within each treated branch's catchment area to make loan offers based on the branch's loan capacity. In this level of randomization we partition the catchment area into a treated and an untreated region, and isolate one untreated village in the middle of the treated region to test for spillover effects.

We introduce a novel design to preserve random assignment with high-intensity treated regions. Randomization proceeds by projecting villages in a microfinance branch's catchment area onto a circle with the branch office at the center. We first randomly select one village from this projection to be designated as the "spillover" village in which no offers will be made. Next, we assign an equal number of villages in either direction along the circle projection to the "treatment" group where loan offers are made. Between a third and a half of the villages in each branch's catchment area are assigned to the treatment, according to branch capacity. Finally, the remaining villages within the branch's catchment area are designated as "branch control", and no loan offers are made.

An example of the resulting village assignment is provided in Figure 3. This strategy effectively creates a pie-slice-shaped treated region originating from the branch office in the center of the catchment area. A single village close to the middle of the treated slice remains untreated to test for spillovers. Because all assignment is randomized according to projected circle order, the probability of treatment remains uncorrelated with density of villages, proximity to the branch, or other geographic characteristics.

After randomization, each treatment branch office hired two new employees as migration loan officers to conduct outreach to eligible households and handle the loan portfolio for the branch. These officers first administered a census of households in each village assigned to treatment ahead of the lean season. The census included questions about land ownership and history of food security to determine loan eligibility. Loan officers then contacted households deemed eligible to advertise and promote migration loans. Actual loan disbursement took place at the branch office, and loans were not disbursed to members of households not deemed eligible from the census to ensure compliance with treatment assignment.

2.4 Study Sample and Data

Village-level randomization generates four different categories: a treatment group where loan offers are made, a spillover group where no offers are made in the midst of a treated region, a branch control group where no offers are made in the catchment area of a branch that makes offers, and a pure control group where no offers are made in the catchment area of a branch that does not make loan offers. For evaluation purposes, we randomly select one treated and one branch-control village per treatment branch to conduct household surveys. We also conduct surveys in every spillover village as well as one randomly selected pure-control village from each untreated branch. Within each survey village, we randomly sample twenty households out of the eligible population for surveying. Figure 4 characterizes the full randomization and survey strategy for 2017; the only adjustment in 2018 was to increase to 50 treatment branches.

To identify households that would be loan-eligible in untreated villages for surveying, the survey team used a random walk sampling strategy. Surveyors followed a skip pattern to select households, asked about the eligibility criteria in each selected household, and stopped once they had identified twenty would-be-eligible households. In the 2017 evaluation, survey households in treated villages were drawn randomly from the census conducted by the microfinance organization. This asymmetry raised a concern that differences between the treated and untreated survey samples may have been induced by differential selection. As a result, in the 2018 evaluation, sample households in treated villages were selected by surveyors following the same random walk strategy as in untreated villages.

We conduct three surveys with each sample household in each evaluation year. First, we administer a short survey on loan eligibility, migration history, and household composition in August–September prior to the lean season and any potential migration. Second we conduct a longer survey in the following January with questions about migration, earnings, and food consumption during the lean season. Finally, we conduct a third survey in April–May 2017 and June–August 2018 focused on subsequent migration, subsequent earnings, and overall financial status to evaluate whether lean season earnings may have persistent effects.

Tables 1 and 2 report balance on baseline characteristics across treatment status in 2017 and 2018, respectively. These tables reveal evidence of imbalance across treatment arms, even in 2018 when sample selection was consistent across treatment assignment. In all that follows, we verify that results are robust to controlling for these baseline covariates.

We supplement the primary analysis with three additional sources of data. First, we conduct a debriefing survey with migration loan officers following the 2017 round of evaluation. Second, we include recall questions on loan offers and microfinance institution activities following the 2018 lean season. Third, we use administrative data on implementation from the microfinance institution itself.

It is worth noting there was strict separation between implementation and evaluation, and survey teams did not provide feedback or monitoring to implementers as is common in pilot evaluations. In this study, the research designed aimed to evaluate the policy as it would normally

operate without parallel research efforts. As an unfortunate outcome of this separation, we cannot match household survey data to administrative data at the household level. This prevents us from using administrative data to validate recall about loan offers, or from using survey data to validate administrative loan and migration records.

This study was preregistered in the AEA RCT Registry under ID No. AEARCTR-0002685.

3 Conceptual Framework

The main contributions of this paper are to evaluate the impact of the NLS program at scale and to quantify sources of difference in treatment effects between pilot and scale. One common concern when scaling up a policy is that implementers introduce design differences that alter the policy effect. Our evaluation avoids this concern because the migration loan on offer remains consistent across all rounds of evaluation. Instead, we focus on economic reasons why treatment effects may differ with the scale of implementation *for the same program.* We first describe commonly discussed external validity concerns that focus on population-level changes, and then introduce a new threat to external validity based on implementation targeting within a given population.

3.1 Classification of Migration Compliance

Differences between pilot and scale derive from the fact that the benefits of NLS loans are concentrated within a subset of the loan-eligible population. NLS works by enabling migration for those households that have high returns to migration but cannot access them due to credit constraints. We refer to this population as the induced migrants—those that would not migrate without a loan but are induced to do so by the loan offer. In the language of program evaluation, this group represents program compliers.

Nearly all NLS program benefits accrue to induced migrants. By contrast, never-migrants—those who do not migrate even when offered a loan, also known as never-takers—derive little benefit from a loan offer they refuse. Similarly, regular migrants—those who would migrate even without a loan, also known as always-takers—will enjoy the returns to migration whether they receive a loan or not. They may derive some additional direct benefit from a small, short-term, low-interest loan, but this direct effect is negligible compared to the return to migration. Therefore, economic explanations for differences in treatment effect between pilot and scale focus on changes in the frequency of and returns to program compliers.

3.2 Population Differences and General Equilibrium Spillovers

Standard concerns regarding external validity in pilot evaluations center around differences in the fraction of program compliers—i.e. induced migrants—in the treated population at scale. Most directly, as the size of a program grows, it reaches new locations and populations that may differ from where the pilot took place. Program effects will vary with the fraction of compliers in newly

treated populations, and this variation can systematically attenuate treatment effects at scale if pilots strategically target populations with the greatest need or probability of success. We explicitly test for geographic population differences through stratification in the randomization design.

The fraction of program compliers may also change within the same population over time. To address the role of time-series changes in population characteristics, we first estimate conditional treatment effects from the 2008 and 2014 pilot evaluation rounds conditional on observable baseline characteristics. We then construct a counterfactual at-scale treatment effect that is the weighted average of conditional treatment effects weighted by the distribution of baseline characteristics in 2017 and 2018. By comparing this counterfactual to the actual estimated treatment effect at scale, we can quantify the importance of changes in observable population characteristics over time.

A second class of explanations for why there may be fewer compliers in the treated population at scale follow from general equilibrium spillovers. As the size of a program grows, market-level changes in prices or other sources of crowd-out can lower the value of a program to its beneficiaries. With respect to NLS, negative spillovers would exist if wages decline at migration destinations in response to a large-scale labor supply shock, or if there are a limited number of local migration opportunities that become saturated with broad loan availability. These factors would cause the individual return to migration to decrease with the overall size of the NLS program, shrinking the number of treatment compliers within the population.

We explicitly test for regional spillovers through the randomization design. Randomization generates a set of untreated high-spillover villages in the midst of an intensely treated area, as well as a set of untreated branch-control villages adjacent to microfinance branch offices offering loans. By comparing migration outcomes in these groups to the pure control, where no loans are offered nearby, we evaluate whether migrants induced to travel by the NLS program crowd out other migrants. This cross-village analysis complements the finding from the 2014 evaluation that migration compliance increases with village-level treatment intensity, indicating there is crowd-in rather than crowd-out of migrants within a village (Akram et al., 2017).

3.3 Delegation Risk and Targeting

The above explanations correspond to changes in the fraction of program compliers within the treated population. We introduce and quantify a new threat to scaling, which we refer to as delegation risk, based on program outreach to compliers within the same population. Even if the distribution of compliance status remains consistent as a policy grows, common management practices can lead compliers to be systematically undertreated at scale.

This risk stems from the fact that policy designers and managers delegate implementation duties to hired staff. The level of staff discretion in implementation grows with program scale for two reasons. First, if the number of implementing agents increases, then managers can perform less direct oversight per agent, leaving more scope for discretionary decisions. Second, if the number of assigned beneficiaries per agent grows, then agents must be more selective in how they allocate

effort between assigned beneficiaries. Agent effort cannot directly scale with the number of assigned beneficiaries because of constraints on the implementers' time, and selectivity may be exacerbated by other resource or program capacity constraints that further raise the shadow cost of effort per beneficiary.

In the scale-up of NLS, both administrative complexity grew as two new loan officers were hired per implementing branch, and loan-eligible households per branch increased by over an order of magnitude. Loan officers exercised discretion in their marketing efforts and enforcement of loan conditionality. At the start of each season, officers held a meeting announcing the NLS program for all eligible households in each treatment village, and then went door-to-door to inform anyone not in attendance. After this initial kickoff, officers conducted followup visits in their assigned regions to continue marketing and to encourage migration after loan disbursement. They exercised discretion in how to allocate their time between these activities and in which villages and households to follow up most intensely. By contrast, the pilot rounds were small enough that all study households were reached with high intensity.

Ideally, implementing agents exercising discretion would share policy designers' goal of maximizing effort toward program compliers. However, compensation schemes cannot directly incentivize this outcome because compliance status is unobservable for any given beneficiary. Even aggregate population frequencies can only be indirectly inferred from migration rates when an experimental control exists.

As an alternative, it is common to structure employee incentives around implementation. This metric can generate perverse incentives to focus effort on always-takers to the exclusion of program compliers. From the agent's perspective, a participant's pre-treatment behavior (i.e. if they are a regular migrant) is likely more observable than their hypothetical behavior after treatment (i.e. whether they would be induced to migrate with a loan). That is to say, it is easier to identify always-takers than to distinguish compliers from never-takers. Therefore, constrained agents seeking to maximize implementation quantity should direct efforts toward always-takers, who are guaranteed to take up the program, rather than reach out to prospective compliers and risk wasting time on never-takers.

This type of mistargeting may also be induced by participant demand. Since always-takers already have plans in place, they can more confidently signal their demand, even if the direct benefit of the program is relatively small. In contrast, compliers need to figure out their behavior after treatment. In our study, induced migrants need time to scout opportunities and make plans before requesting a loan. Therefore, implementers responding to demand signals may exhaust their capacity on always-takers to the exclusion of program compliers. In either case, the cost of agent effort is lower for always-takers even though program benefits primarily derive from compliers.

While we cannot directly measure loan officer effort, we provide suggestive evidence using survey data on households' recollection of loan offers. Using offer recall as a proxy for the level of effort directed toward a household, we show evidence that loan officers selectively recruited regular

migrants rather than induced migrants.

4 Evaluation at Scale

In this section we first present the estimated treatment effects of the NLS loan program at scale pooled cross all evaluation districts, and then break down the difference in estimated treatment effect between the original districts included in the pilot and new districts included at scale.

4.1 Pooled Evaluation

Our primary metric for evaluation is the fraction of the eligible population induced to migrate after receiving a loan offer. Estimated treatment effects across rounds of evaluation are presented in Table 3. The first two columns report results from the 2008 and 2014 pilot rounds, respectively. In isolation, being offered an NLS loan raises a household's propensity to send a migrant by 22–25 percentage points, from a base migration rate of 35%. This treatment effect size corresponds to the fraction of induced migrants in pilot rounds. As the saturation of loan offers within a village increases from 14% to 70% of households, the fraction of induced nearly doubles from 25% to 40%. This substantial increase indicates there is strong crowd-in of induced migrants as the within-village loan offer and migration rates increase.

These large migration impacts fail to replicate at scale. The final two columns of Table 3 report estimated treatment effects from the 2017 and 2018 evaluation rounds comparing treatment to pure control. In both years, the base rate of migration in control villages—i.e. the prevalence of regular migrants—remains close to its pilot level. From this base, eligibility for an NLS loan only increases a household's likelihood of sending a migrant by 6 percentage points despite complete within-village saturation of loan eligibility, and this change is not statistically distinguishable from zero[2]. We can both reject that the treatment effect at scale matches any pilot round and reject that the total migration rate in treated villages at scale (including regular migrants) matches the total migration rate in treated pilot villages at the 1% level.

Even without inducing new migration, the NLS program may benefit existing migrants by offering a low-cost alternative to finance migration. We test for direct program benefits by estimating the reduced-form effect of treatment on household earnings, consumption, and financial status, with results presented in Appendix A. Across a range of outcomes, reduced form effects are quantitatively small and statistically indistinguishable from zero. These results indicate that the direct welfare effects of a small, short-term loan are minimal compared to the return to migration for induced migrants in the pilot rounds of study.

[2]In Appendix A we verify robustness to the way migration is measured.

4.2 Evaluation by District

The NLS program was originally piloted in two districts within the Rangpur division of Bangladesh. In scaling up, the program expanded to incorporate the remaining six districts in the division. Tables 4 and 5 describe differences in baseline characteristics between the eligible population in old and new districts in 2017 and 2018, respectively. Populations are similar across geography on most characteristics, with the exception of land ownership. Notably, migration history among the eligible population is comparable between districts. This similarity on observables along with the geographic proximity of new districts sets the expectation that the distribution of program compliers, and as a corollary program treatment effects, will not differ substantially.

Despite the observable similarity, the eligible population in new districts may differ on unobservables related to program compliance. To quantify the importance of geographic expansion on NLS program effectiveness, we separately estimate the treatment effect at scale in old and new districts. Results from this exercise are presented in Table 6.

The regression table reveals a substantial difference between treatment effects in old and new districts. In the 2018 evaluation round, loan offers in the original pilot districts raised household migration by a statistically significant 12 percentage points[3]. By contrast, the effect in old districts was small, if anything negative, and statistically insignificant. Disappointingly, these geographic differences in treatment effect would not have been easy to predict given the observable similarity of new and old districts at baseline.

While this analysis indicates that geographic differences play some role in treatment effect attenuation from pilot to scale, even the twelve percentage point treatment effect measured in pilot districts in 2018 falls well short of the 25–40 percentage point increase in migration achieved during piloting. Half to two thirds of the difference in program effectiveness between pilot and scale remains unexplained even in the original pilot areas of study.

5 Delegation Risk and Implementation Targeting

We have shown that even though there may be fewer migration compliers in new program districts that were not a part of the pilot, even the at-scale treatment effect in the pilot districts falls well short of what was achieved in the pilot. In this section we present evidence that delegation risk led outreach efforts at scale to systematically undertarget program compliers, and can quantitatively explain the remaining treatment effect attenuation.

5.1 Leakage between Loan Eligibility and Migration

To motivate this analysis, we compare sources of leakage in program implementation between pilot and scale. Table 7 reports the fraction of households lost at each stage of the NLS process from loan

[3]The treatment effect in pilot districts is smaller and statistically indistinguishable from new districts in 2017. We discuss implementation differences between 2017 and 2018 in the next section.

eligibility to migration in administrative data from the implementing organization. The top row represents the fraction of the population in study villages that satisfied the eligibility criteria, and the second row represents the faction of qualified households assigned to receive an offer, which is around one tenth in 2008 because of experimental randomization. The remaining four rows describe implementation conditional on loan eligibility.

There is an increase in the qualified population in 2017 because the eligibility criterion of having missed meals in past lean seasons was dropped. The third column of Table 7 displays statistics for the subset of loan-eligible households that satisfy this additional criterion, though the fraction is low because this variable was poorly recorded. Nevertheless, administrative data within this subset is nearly identical to the full loan-eligible population, indicating that changes in eligibility criteria did not affect program administration. In 2018, past missed meals was reinstated as a condition of loan eligibility.

Table 7 reveals three notable facts about sources of leakage in program effectiveness. First, loan conditionality was, if anything, more stringently enforced at scale than in the pilot. Just under two thirds of loan recipients qubsequently migrated in 2008. This fraction was 70% in 2018, and actually reached over 90% in 2017. Second, in 2017, loan acceptance rates were substantially lower than in 2008, and fewer than two thirds of households that accepted a loan offer actually received a loan. Third, official program implementation metrics in 2018 look nearly identical to 2008. Notably, around 41% of loan-eligible households migrated with a loan in both years, which may have led implementers to mistakenly conclude the program was equally effective at scale were it not for the randomized evaluation.

5.2 Capacity Constraints and Implementation Targets

The low rate of loan acceptance and disbursement in 2017 was driven at least in part by binding capacity constraints. At the officer level, a debrief survey administered after the migration season reveals that 85% of officers were hired with an explicit target for the number of loans to disburse. Figure 5 plots the distribution of disbursements relative to the reported target for each officer. The figure shows that the majority of officers hit their target nearly exactly. The net result of these targets was that total loan disbursements across all branches in 2017 plateaued at the planned quantity of 40,000[4], explaining why so many households that accepted a migration loan never actually received one.

Quantity targeting was abandoned in the 2018 implementation year, and loan officer instructions shifted to use the language of program compliers and induced migrants. Correspondingly, the number of loan disbursements rose to nearly 90,000. However, while explicit loan requirements

[4]The planned number of loan disbursements fell far short of the 64% acceptance rate realized 2008 because the program made more loan offers than anticipated. At the time of implementation, microfinance branches realized they had the capacity to conduct eligibility surveys in more villages than they predicted, and expanded the scope of operations accordingly at the last minute. The anticipated number of loans had already been registered with the Bangladeshi government and could not be increased to match the new program size.

were dropped, incentives for implementation quantity implicitly remained in place as officers who did not disburse a sufficient number of loans were unlikely to be retained.

Household survey data following the 2018 evaluation corroborates the concern that migration officers directed different levels of effort to different households. Figure 6 reports the fraction of households that remember being offered a migration loan during the prior lean season. While this fraction is substantially higher in treated villages than untreated, it is not close to 1. Only 40–60% of eligible households recall being offered a migration loan. Recall is not a perfect measure of program outreach, but should be seen as a proxy for intensity of loan officer engagement.

Figure 7 verifies that this low fraction cannot entirely be attributed to faulty memory. The figure plots the fraction of households who remember being offered a loan by village on the y-axis against the fraction that remember loans being offered in their village on the x-axis. Nearly all the data lie below the 45-degree line, revealing a large number of eligible households who can recall NLS-related activity in their village but feel they were not invited to participate. The figure also shows there is no systematic difference in this metric between pilot and expansion districts.

5.3 Selectivity and Delegation Risk

We use the survey data on household recall of loan offers to test for selective effort by loan officers. To do this, we first calculate the frequency of baseline characteristics among migrants in control villages and among non-migrants in treated villages. The difference between these two populations reflects the difference between always-takers—regular migrants who travel even without a loan—and never-takers—never-migrants who do not travel even when a loan is available. For each characteristic, we compute the t-statistic for the difference as a measure of how informative the characteristic is about a household's propensity to accept a loan. If loan officers aim to maximize the number of loans disbursed without concern for reaching induced migrants, then they should adjust their targeting in accordance with these t-statistics.

The first two columns of Table 8 report differences and t-statistics for the difference between always- and never-takers in each characteristic available at baseline. Migration history is the strongest predictor of current migration plans by far, with households that migrated recently more likely to be regular migrants and those that did not more likely to be never-migrants. The presence of adult males in the household also increases the likelihood that a household will send a migrant, as does borrowing at baseline. Land ownership and education of the household head are both negatively associated with regular migration.

Next, we regress household recall of receiving a loan offer on these baseline characteristics for households in treated villages. To the extent that this recall measure proxies for the effort that a migration officers spent in advertising migration loans to a household, this regression identifies the characteristics that predict effort level. Regression results are reported in the final column of Table 8.

We plot regression coefficients against t-statistics in Figure 8, and the results are striking. There

is a nearly monotonic relationship between how informative a characteristic is about always-takers versus never-takers and how strongly that characteristics predicts whether a household remembers being offered a migration loan. This relationship is consistent with the idea that in the presence of capacity constraints, migration officers actively seek out always-takers and avoid never-takers to maximize their rate of loan disbursement.

This exercise shows selectivity by propensity to migrate, but leaves open the question of whether effort is directed toward or away from induced migrants. Because induced migrants eventually accept a loan, effort toward this group would in principle achieve implementation-based goals. Delegation risk poses a threat to implementation specifically when implementing officers prioritize other lower-cost alternatives to maximize loan disbursement.

To measure selectivity, we first need to extend the classification of treatment compliance to separately account for loan acceptance and migration. Beyond the standard categories of Always Takers (AT) who take the loan when offered and migrate with or without a loan, Compliers (C) who take the loan when offered and migrate if and only if they receive a loan, and Never Takers (NT) who decline the loan when offered and never migrate, we describe two additional behaviors types. First, there are Self Sufficient (SS) households that decline the loan offer but still migrate on their own, and fall into the broader category of "regular migrants". Second, we observe Time Wasters (TW) who accept a loan when offered but never migrate and remain "never migrants". Table 9 presents a full description of types.

The fifth column of Table 9 reports the implied type distribution in the population according to the 2008 implementation data under the assumption that treatment was administered with uniform effort intensity. For instance, 36% of households in the control group migrated, corresponding to the population fraction of regular migrants. In the treatment group, 16% declined a loan and still migrated, corresponding to the portion of Self-Sufficient, meaning the other 20% must be standard Always-Takers. Other population frequencies are calculated similarly.

Normalizing the effort intensity in 2008 to 1, we can compute the implied effort intensity for each compliance type in 2018 treating loan offer recall as a proxy for implementer effort. In the data, we observe loan offer recall, loan acceptance conditional on recall, and migration with or without a loan for every sample household in treatment villages. The fraction of the population in each data cell can be written as a function of the type distribution and the effort directed to each type according to a system of equations.

$$\begin{aligned}
Recall,\, Accept,\, Migrate &= P_{AT}AT + P_C C \\
Recall,\, Accept,\, Remain &= P_{TW}TW \\
Recall,\, Decline,\, Migrate &= P_{SS}SS \\
Recall,\, Decline,\, Remain &= P_{NT}NT \\
No\, Recall,\, Migrate &= (1-P_{AT})AT + (1-P_{SS})SS \\
No\, Recall,\, Remain &= (1-P_C)C + (1-P_{TW})TW + (1-P_{NT})NT
\end{aligned} \tag{1}$$

Where capital letters X represent population frequencies by type and P_X represents the average treatment intensity of that type.

Given population frequencies by type, the system (1) represents five equations[5] with five unknowns corresponding to treatment intensities. We solve this system for treatment intensity in 2018 under the assumption that the type distribution remained constant in 2008.

The implied effort intensity by type is reported in the final column of Table 9. The table reaffirms strong selectivity by migration status: effort intensity is over three times greater for always-takers than never-takers. However, this fact seems to be linked to loan acceptance rather than migration. Effort toward self-sufficient households, who migrate without a loan, closely resembles that of never takers. Inversely, time-wasters who accept a loan without migrating experience similar levels of implementer effort as always-takers. In all, the data indicate that loan officers exercising discretion selectively target groups that will accept a loan offer.

The deleterious effects of delegation risk are apparent in the estimated effort intensity for compliers. The implied value of 0.49 is substantially lower than for always-takers and time-wasters, and is slightly lower than the unconditional offer recall rate of 51% among all households assigned to treatment. That is to say, induced migrants, who benefit the most from the program and are responsible for the promising pilot results, receive the least implementer focus among those who would accept a loan. In fact, induced migrants receive slightly less implementer focus than they would were effort assigned across groups at random.

6 Other Population-Based Explanations

We have provided suggestive evidence that delegation risk can account for the majority of decline in program effectiveness between pilot and at-scale implementations of NLS. In this section we rule out two other possible reasons why the fraction of program compliers in the population may have diminished from pilot to scale.

6.1 General Equilibrium Spillovers

We test for cross-village crowd-out of migration using the spillover and branch-control groups from the randomization design. Spillover villages are those located in the middle of a group of treated villages where loans were offered. Branch control villages are those in the same catchment area of a branch that made loan offers. Both groups are in close geographic proximity to treated villages and are more likely to send migrants to similar destinations. If NLS implementation at scale crowds out migration, we would expect to see depressed rates of migration in these villages relative to pure control.

The second and third rows of Table 6 reveal that general equilibrium spillovers have little effect on migration rates. Differences in migration between spillover/branch-control and pure control

[5]There are six equations, but one is redundant because outcome frequencies must sum to one.

are quantitatively small in magnitude and generally statistically indistinguishable from zero. The positive sign of the point estimates indicates that, if anything, the NLS program induces slight crowd-in of migrants across village. In fact, crowd-in across villages is strongest exactly in the original pilot districts where the program induces the greatest migration. This finding suggests that, even at the current scale, migration opportunities have not been exhausted and labor demand at migrant destinations is sufficiently elastic to absorb a further supply shock.

6.2 Time Series Differences

The population under study in the at-scale evaluation does not just differ from the pilot population in geography, it also differs in time. In particular, we observe two key factors that differentiate the at-scale evaluation years from pilot that may have altered the prevalence of program compliers. First, there has been a secular trend of economic growth, leading the later evaluation rounds to take place among a generally wealthier population. Second, both 2017 and 2018 saw substantially greater rainfall than 2008 and 2014. These and other changes may affect the population responsiveness to NLS loan offers.

6.2.1 Household Characteristics

We first explore the importance of wealth and other population characteristics by estimating conditional treatment effects by baseline characteristics. As a demonstrative exercise, consider household calorie consumption. Figure 9 plots the distribution of per capita calorie consumption across control households in 2008 and 2018. The figure reveals a nearly 25% increase in average calorie consumption over the decade.

To evaluate the importance of calorie consumption in responsiveness to treatment, we proceed in three steps. First, we divide households from the 2008 evaluation into k bins using data on calorie consumption collected before migration loan offers. Second, we estimate a bin-specific treatment effect in the 2008 data. Third, we construct a counterfactual 2018 treatment effect by taking a weighted average of bin-specific treatment effects weighted by the share of the population in each bin in the 2018 control group. This counterfactual describes the average migration response we would have observed if conditional treatment effects remained constant over time and the only change was the distribution of calorie consumption in the study population. We compare this counterfactual to the true treatment effect from 2018 to quantify how much of the gap can be explained by changes in calorie consumption as a proxy for compliance status.

We present results from this exercise in Figure 10. The x-axis plots the number of bins used to estimate conditional treatment effects. The y-axis plots the fraction of the difference between 2008 and 2018 that can be attributed to changes in the calorie distribution over time. The figure shows that with a small number bins, this exercise can only explain on average only 10–15 percent of the attenuation in treatment effect, and the explanatory power shrinks as the number of calorie bins grows.

We generalize this exercise using a machine learning algorithm on all baseline characteristics available in 2008. We implement a random forest algorithm following Wager and Athey (2018) to select variables with the greatest explanatory power for conditional treatment effects in the 2008 evaluation. The algorithm selects from covariates among baseline calories, education of household head, cultivable land owned, household size, number of adult males, and number of adult females. We then interact conditional treatment effects with the distribution of the selected variables in the 2018 control group to construct a counterfactual 2018 treatment effect.

The machine learning algorithm generates a counterfactual 2018 treatment effect of 21.9%. This small attenuation accounts for only 4% of the measured difference in treatment effect between 2008 and 2018. As a validity check, we verify that weighting the 2008 marginal treatment effects by the 2008 distribution of population characteristics matches the 2008 treatment effect nearly exactly at 22.6%.

6.2.2 Rainfall

We are unfortunately unable to include rainfall in this conditional treatment effect exercise because there is little overlapping support between rainfall in 2008 and in 2017/2018. Both 2017 and 2018 saw above-average levels of rainfall before and during the agricultural lean season, with historically high levels of flooding experienced in 2017. In contrast, 2008 was an abnormally dry year with most study village experiencing below-average rainfall.

We provide two pieces of evidence that suggest changes in rainfall had minimal impact on the change in estimated treatment effect over time. First, we interact treatment status with satellite rainfall data from the National Oceanic and Atmospheric Administration. For each village, we create a dummy for whether the village received above- or below-average rainfall during the lean season relative to the period from 2001 to 2019.[6] We then interact this dummy variable with treatment assignment. Results, presented in Table 10, reveal that above-average rainfall depresses responsiveness to treatment by only 3 percentage points, and is not statistically distinguishable from zero.

Second, we use administrative data from the implementing microfinance organization to test for a cross-sectional relationship between rainfall relative to average and the fraction of households that accept a migration loan. This (lack of) relationship, plotted in Figure 11, confirms there is little correlation between rainfall realization and program effectiveness, indicating that variation in rainfall is unlikely to affect the distribution of program compliers in the population..

Together, these facts suggest that observable changes in population characteristics over time add little explanatory power in accounting for the attenuation in treatment effect between evaluation rounds. Of course, it would be impossible to fully rule out the possibility that our results driven entirely by time series changes in population unobservables that lower program compliance.

[6] Recall survey data shows a strong correlation between this measure and self-reported flooding over 2014–2018.

7 Conclusion

In this paper, we report results from an evaluation of a large-scale migration loan program that fails to replicate the success achieved in pilot. We show that this failure to replicate cannot be attributed to crowd-out in general equilibrium, and that differences in the treated population at scale can explain at most a third of the attenuation in treatment effect. We introduce and provide evidence supporting a new theory of delegation risk caused by capacity-constrained implementers seeking to maximize measurable implementation outcomes.

The conditions that lead to delegation risk are fairly general. Any program with an increasing marginal cost or shadow cost of effort[7] will lead implementers to be more selective as scale grows. There is room for this selectivity to be mistargeted when benefits are concentrated among a subset of the population that are compliers, but the population also consists of always-takers and never-takers with no verifiable or contractible way to distinguish between types.

In addition to directed lending programs such as NLS, many types of policy satisfy these conditions. For example, other forms of microfinance, conditional cash transfers, occupational or technological training, and agricultural extension all seek to enable or promote behaviors that some subset of the population would already engage in. In such cases, contracts for implementing agents cannot directly reward effort focused on compliers, and alternative performance metrics can drive a wedge between agent incentives and program intent.

A common management practice in many development organizations is to focus on implementation quantity. This metric is both used to evaluate employee performance as well as reported to donors and other benefactors as a measure of impact. Unfortunately, contracts and career incentives built around implementation quantity can induce selection in exactly the wrong direction if it is easier to identify always-takers than program compliers.

It remains an open question how best to design agent incentives. The ideal contract would reward program impact. In the case of NLS, this would mean paying and retaining loan officers based on the induced migration in their catchment area. Unfortunately, this is only possible with a credible counterfactual, such as from independently collected data on an experimental control group. More feasible alternatives may use performance bonuses and competition across agents, but such schemes risk triggering fairness concerns as outcomes are influenced by unobserved cross-sectional heterogeneity. Our experience also indicates that replacing implementation targets with intrinsic motivation can recover some, but not all, of a program's intended effect.

The presence of delegation risk adds complications to cost–benefit analysis. Pilot studies often draw a distinction between the cost of evaluation and implementation, and report only the latter for policy analysis. Our work suggests that evaluation plays an important oversight role by providing detailed feedback on implementation quality. This monitoring cannot be divorced from the implementation itself and should be factored into program costs.

[7] Delegation risk would be minimal, for example, in a text messaging campaign where the marginal cost of messaging remains constant with scale.

This lesson also applies when evaluating aid effectiveness. Rating agencies for charitable giving commonly focus on the fraction of an organization's budget devoted to beneficiaries.[8] We find that resources for effective targeting can be equally important as poorly targeted programs can fail to deliver promised impacts.

More broadly, the analysis in this paper highlights the complementary roles played by pilot experimentation and evaluation at scale. A pilot can be considered a proof-of-concept evaluating whether a market failure exists and if remedying it generates returns to beneficiaries. These are necessary but not sufficient conditions for policy success. Evaluation at scale reveals whether the remedy can be sustained as a general policy. Ideally, pilot experimentation would provide insight into the potential for scaleup. However, we would not want to sacrifice evidence about market failures by designing pilots that are too frequently derailed by implementation challenges.

[8]For instance, Charity Navigator's methodology page explicitly states that charities " fulfill the expectations of givers when they allocate most of their budgets towards their charitable missions."

References

Akram, Agha Ali, Shyama Chowdhury, and Ahmed Mushfiq Mobarak, "Effects of Emigration on Rural Labor Markets," NBER Working Paper 23929 2017.

Al-Ubaydli, Omar, John A. List, and Dana L. Suskind, "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results," *American Economic Review*, May 2017, *107* (5), 282–86.

_ , John A List, and Dana Suskind, "The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments," Working Paper 25848, National Bureau of Economic Research May 2019.

Alatas, Vivi, Arun G Chandrasekhar, Markus Mobius, Benjamin A Olken, and Cindy Paladines, "When Celebrities Speak: A Nationwide Twitter Experiment Promoting Vaccination In Indonesia," Working Paper 25589, National Bureau of Economic Research 2019.

Allcott, Hunt, "Site Selection Bias in Program Evaluation," *The Quarterly Journal of Economics*, 2015, *130* (3), 1117–1165.

Andersen, Simon Calmar and Ulrik Hvidman, "Implementing Educational Interventions at Scale," Working Paper 2020-039, Human Capital and Economic Opportunity Working Group May 2020.

Araujo, M. Caridad, Marta Rubio-Codina, and Norbert Schady, "70 to 700 to 70,000: Lessons from the Jamaica Experiment," Working Paper IDB-WP-1230, Inter-American Development Bank April 2021.

Balán, Pablo, Augustin Bergeron, Gabriel Tourek, and Jonathan L. Weigel, "Local Elites as State Capacity: How City Chiefs Use Local Information to Increase Tax Compliance in the Democratic Republic of the Congo," *American Economic Review*, March 2022, *112* (3), 762–97.

Banerjee, Abhijit, Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh, "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy, and Training," *American Economic Journal: Economic Policy*, 2021, *13* (1), 36–66.

Banerjee, A.V., S. Chassang, and E. Snowberg, "Decision Theoretic Approaches to Experiment Design and External Validity," in Abhijit Vinayak Banerjee and Esther Duflo, eds., *Handbook of Field Experiments*, Vol. 1 of *Handbook of Economic Field Experiments*, North-Holland, 2017, pp. 141–174.

Basu, Karna and Maisy Wong, "Evaluating seasonal food storage and credit programs in east Indonesia," *Journal of Development Economics*, 2015, *115* (C), 200–216.

Bellés-Obrero, Cristina and María Lombardi, "Teacher Performance Pay and Student Learning: Evidence from a Nationwide Program in Peru," *Economic Development and Cultural Change*, 2022, *70* (4), 1631–1669.

Bloem, Jeffrey and Bruce Wydick, "All I Really Need to Know I Learned In Kindergarten? Evidence from the Philippines," *Economic Development and Cultural Change*, *forthcoming.*

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, 2018, *168*, 1–20.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak, "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh," *Econometrica*, 2014, *82* (5), 1–43.

Bulte, Erwin, Gonne Beekman, Salvatore Di Falco, Joseph Hella, and Pan Lei, "Behavioral Responses and the Impact of New Agricultural Technologies: Evidence from a Double-blind Field Experiment in Tanzania," *American Journal of Agricultural Economics*, 2014, *96* (3), 813–830.

Burke, Marshall, Lauren Falcao Bergquist, and Edward Miguel, "Sell Low and Buy High: Arbitrage and Local Price Effects in Kenyan Markets*," *The Quarterly Journal of Economics*, 2018, *134* (2), 785–842.

Cameron, Lisa, Susan Olivia, and Manisha Shah, "Scaling up sanitation: Evidence from an RCT in Indonesia," *Journal of Development Economics*, 2019, *138*, 1–16.

Chassang, Sylvain, Gerard Padró I Miquel, and Erik Snowberg, "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 2012, *102* (4), 1279–1309.

Cunha, Jesse M, Giacomo De Giorgi, and Seema Jayachandran, "The Price Effects of Cash Versus In-Kind Transfers," *The Review of Economic Studies*, 2018, *86* (1), 240–281.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia*," *The Quarterly Journal of Economics*, 2017, *133* (2), 993–1039.

DellaVigna, Stefano and Elizabeth Linos, "RCTs to Scale: Comprehensive Evidence From Two Nudge Units," *Econometrica*, 2022, *90* (1), 81–116.

Dhaliwal, Iqbal and Rema Hanna, "The devil is in the details: The successes and limitations of bureaucratic reform in India," *Journal of Development Economics*, 2017, *124*, 1–21.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, "Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India*," *The Quarterly Journal of Economics*, 2013, *128* (4), 1499–1545.

_ , Michael Kremer, and Jonathan Robinson, "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya," *American Economic Review*, October 2011, *101* (6), 2350–90.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael W Walker, "General Equilibrium Effects of Cash Transfers: Experimental Evidence from Kenya," *Econometrica*, 2022, *forthcoming.*

Evans, David K. and Fei Yuan, "How Big Are Effect Sizes in International Education Studies?," Working Paper 545, Center for Global Development August 2020.

Fink, Guünther, B. Kelsey Jack, and Felix Masiye, "Seasonal Liquidity, Rural Labor Markets, and Agricultural Production," *American Economic Review*, 2020, *110* (11), 3351–92.

Fund, Global Innovations, "Does "Sugar Daddies" replicate? The preliminary results are in for Botswana," Technical Report 2018.

Ganimian, Alejandro J., "Growth-Mindset Interventions at Scale: Experimental Evidence From Argentina," *Educational Evaluation and Policy Analysis*, 2020, *42* (3), 417–438.

Giné, Xavier, Jessica Goldberg, and Dean Yang, "Fingerprinting in Malawi: The Challenges of Scaling Up a Complicated Technological Solution in a Resource-Constrained Setting," Technical Report, Innovations for Poverty Action 2021.

Hertzberg, Andrew, Jose Maria Liberti, and Daniel Paravisini, "Information and Incentives Inside the Firm: Evidence from Loan Officer Rotation," *The Journal of Finance*, 2010, *65* (3), 795–828.

Kerwin, Jason T. and Rebecca L. Thornton, "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures," *The Review of Economics and Statistics*, 2021, *103* (2), 251–264.

Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken, " Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors," *The Quarterly Journal of Economics*, 2015, *131* (1), 219–271.

Khandker, Shahidur R., "Seasonality of income and poverty in Bangladesh," *Journal of Development Economics*, 2012, *97* (2), 244–256.

Khanna, Gaurav, "Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India," *Journal of Political Economy*, 2022, *forthcoming.*

Kraft, Matthew A., David Blazar, and Dylan Hogan, "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence," *Review of Educational Research*, 2018, *88* (4), 547–588.

Meager, Rachael, "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," *American Economic Journal: Applied Economics*, 2019, *11* (1), 57–91.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar, "Building State Capacity: Evidence from Biometric Smartcards in India," *American Economic Review*, 2016, *106* (10), 2895–2929.

_ , _ , and _ , "General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence from India," Working Paper 23838, National Bureau of Economic Research 2017.

Pritchett, Lant and Justin Sandefur, "Learning from Experiments When Context Matters," *American Economic Review*, 2015, *105* (5), 471–75.

Rabb, Nathaniel, Megan Swindal, David Glick, Jake Bowers, Anna Tomasulo, Zayid Oyelami, Kevin H. Wilson, and David Yokum, "Evidence from a statewide vaccination RCT shows the limits of nudges," *Nature*, 2022, *604* (7904), E1–E7.

Rivera, Juan A., Daniela Sotres-Alvarez, Jean-Pierre Habicht, Teresa Shamah, and Salvador Villalpando, "Impact of the Mexican Program for Education, Health, and Nutrition (Progresa) on Rates of Growth and Anemia in Infants and Young ChildrenA Randomized Effectiveness Study," *JAMA*, 2004, *291* (21), 2563–2570.

Rosenzweig, Mark R and Christopher Udry, "External Validity in a Stochastic World: Evidence from Low-Income Countries," *The Review of Economic Studies*, 2019, *87* (1), 343–381.

Schultz, T. Paul, "School subsidies for the poor: evaluating the Mexican Progresa poverty program," *Journal of Development Economics*, 2004, *74* (1), 199–250. New Research on Education in Developing Economies.

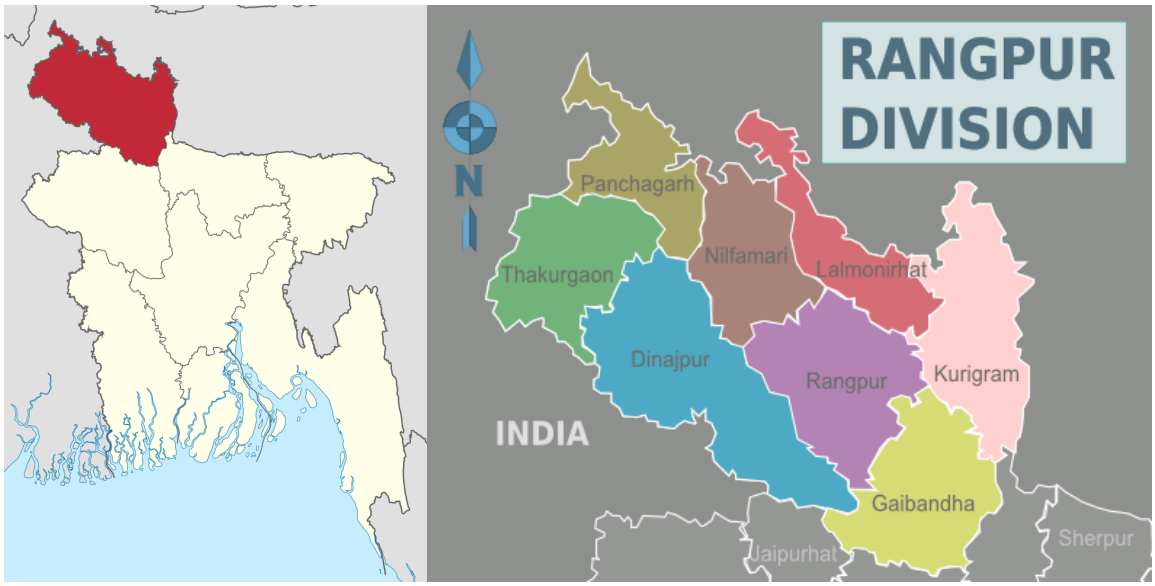
Sraer, David and David Thesmar, "A Sufficient Statistics Approach for Aggregating Firm-Level Experiments," Working Paper 24208, National Bureau of Economic Research January 2018.

Stephens, Emma C. and Christopher B. Barrett, "Incomplete Credit Markets and Commodity Marketing Behaviour," *Journal of Agricultural Economics*, 2011, *62* (1), 1–24.

Vivalt, Eva, "How Much Can We Generalize From Impact Evaluations?," *Journal of the European Economic Association*, 2020, *18* (6), 3045–3089.

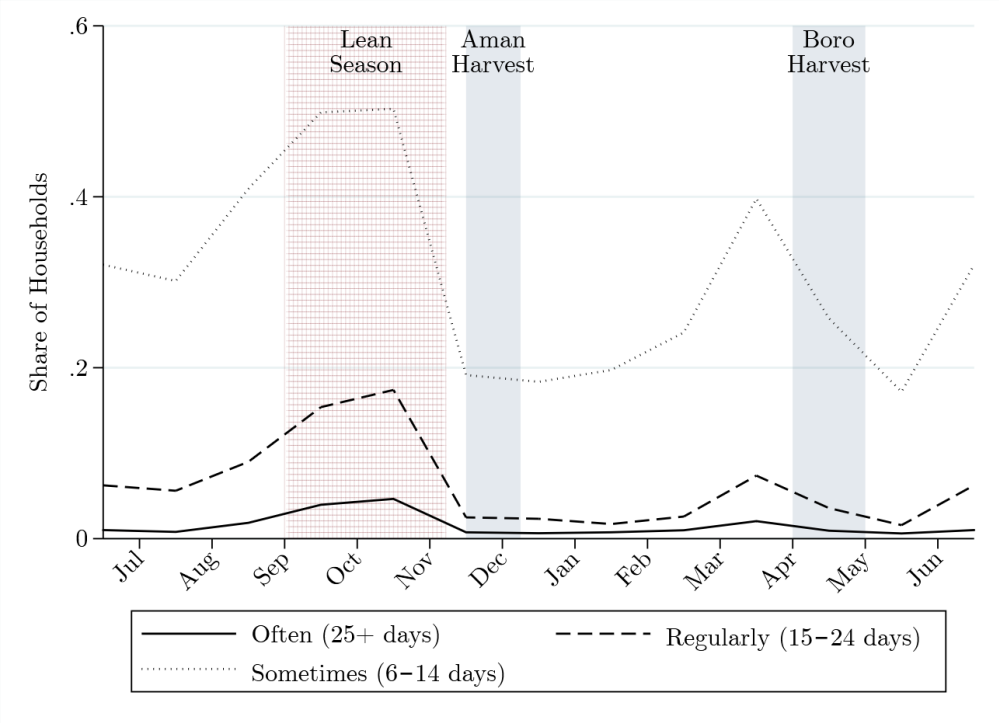
Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

Figure 1: Map of Rangpur Division, Bangladesh



Source: TUBS (left); James Adams (right) accessed from Wikipedia.

Figure 2: Food Insecurity among Landless Rural Households



Self-reported monthly frequency of reducing meals or portion sizes from a 12-month recall survey.

Figure 3: Within-Branch Treatment Assignment According to Circle Order

Example of village assignment in a treated branch according to circle order. Top left panel shows geographic distribution of villages in catchment area, with branch office represented by +. Top right panel shows circle projection of villages around branch office. Bottom panel shows resulting treatment assignment. Square represents randomly selected "spillover" village. Triangles represent "treated" villages according to circle order around spillover. Circles represent untreated villages designated as "branch control". Randomization generates treated and untreated regions, with one untreated village in midst of treated region. Shaded shapes represent villages selected for evaluation surveys.

Figure 4: Village Randomization and Survey Assignment for 2017

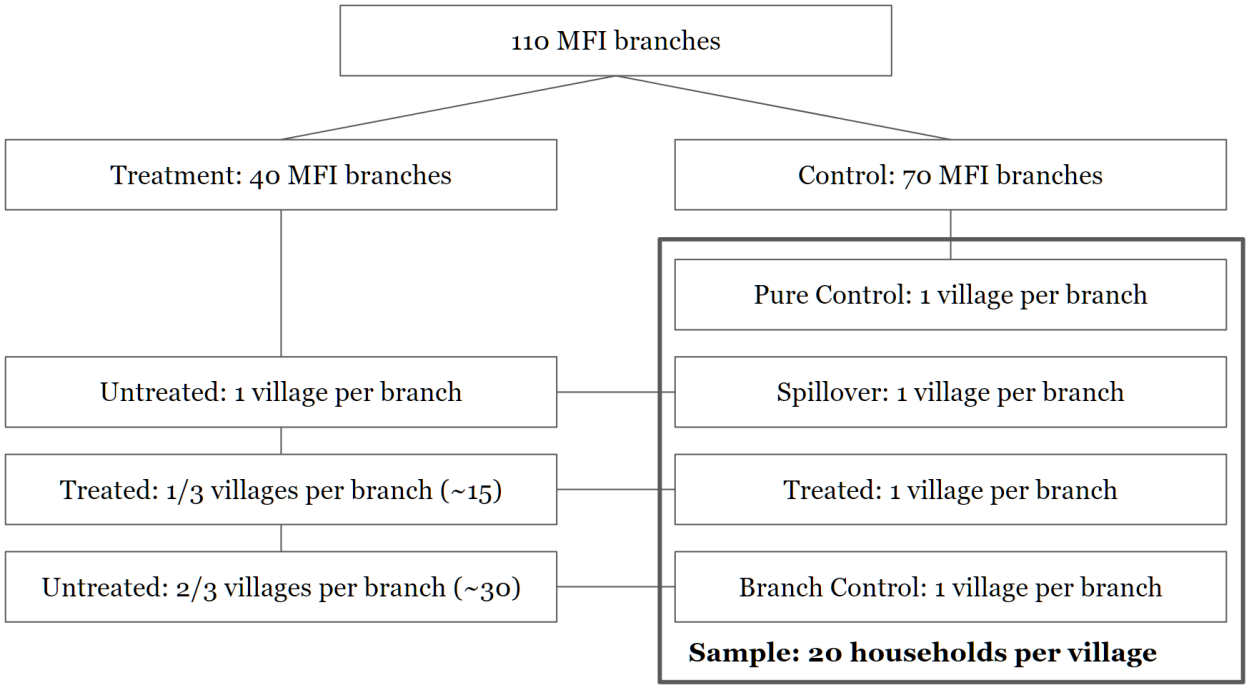
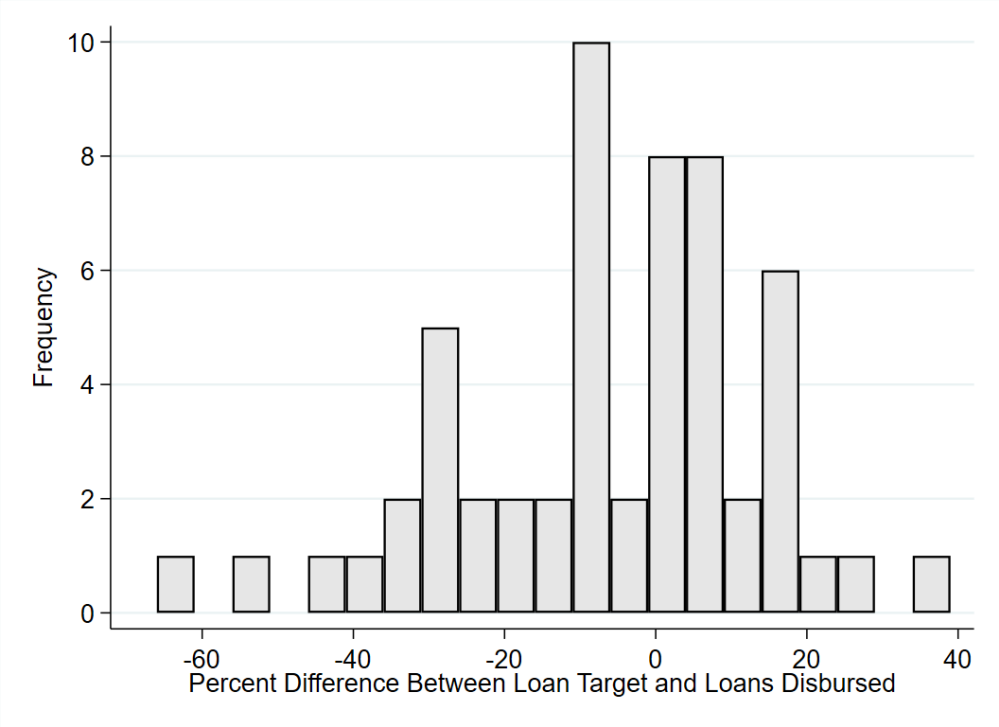
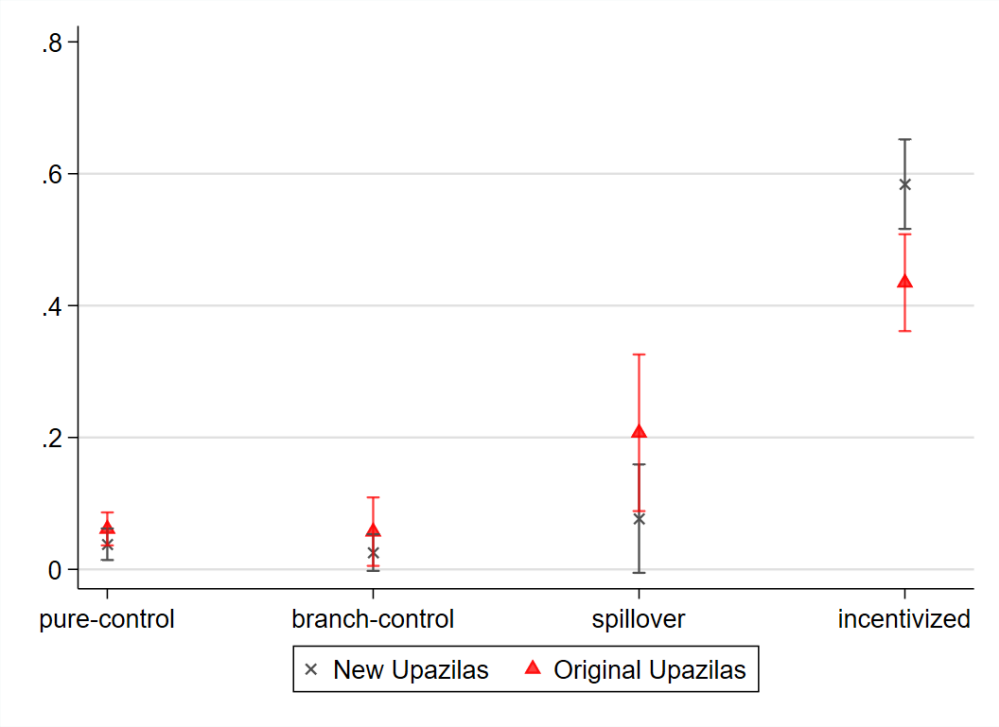


Figure 5: 2017 Loan Disbursements Relative to Target



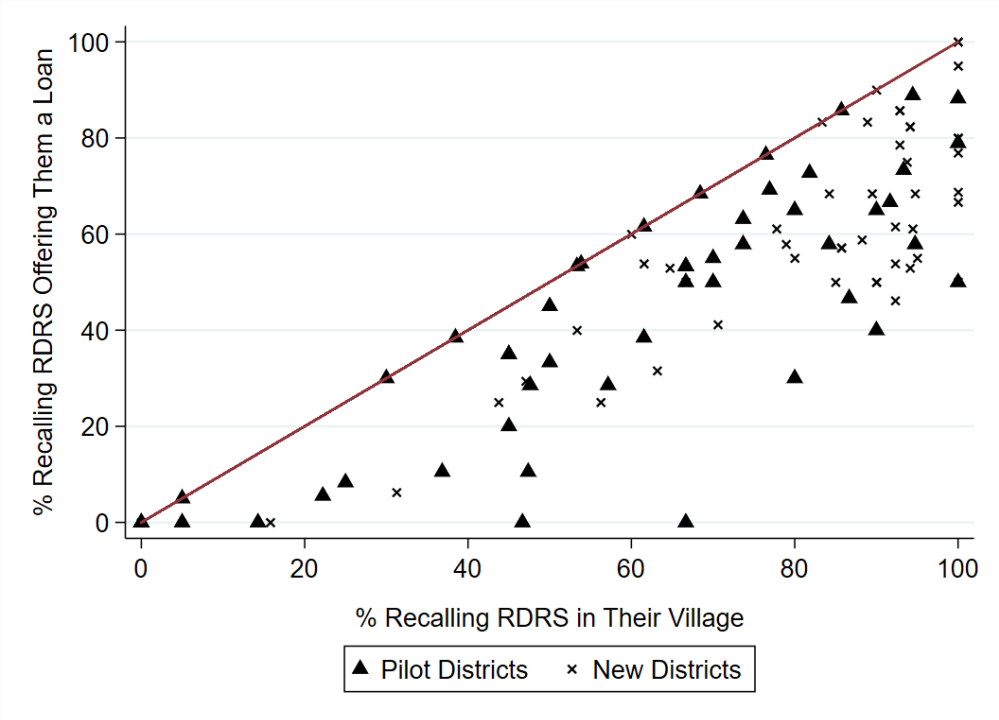
Loans disbursed by migration officer relative to target. Data from debrief survey with loan officers following 2017 migration season.

Figure 6: Fraction of Households that Remember Loan Offer by Village



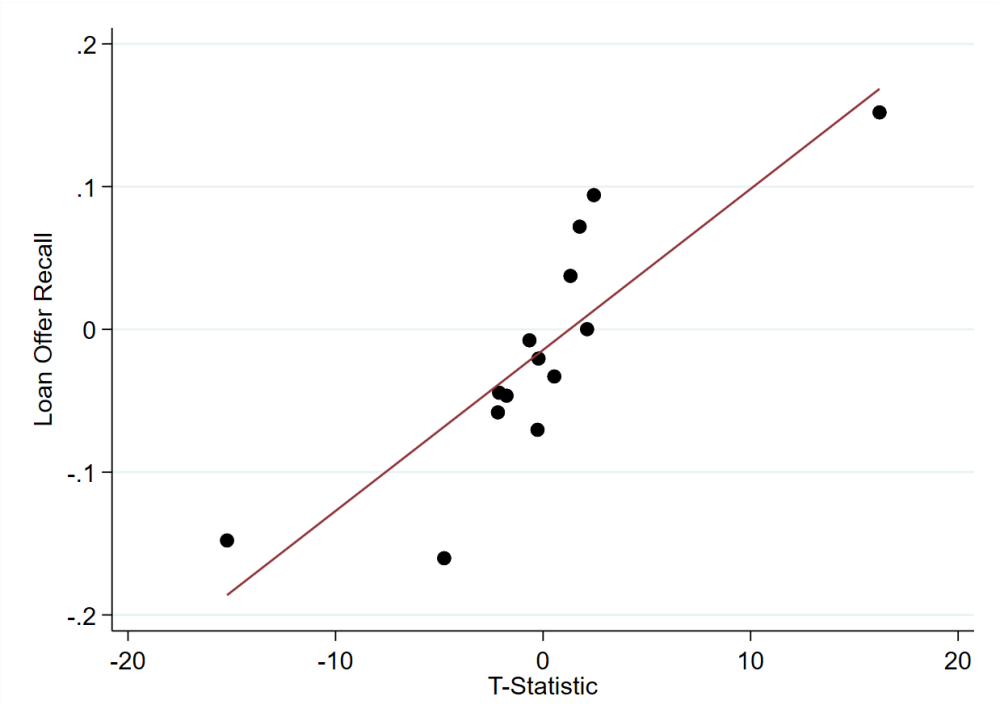
Fraction of households per village that remember receiving loan offer.

Figure 7: Households that Remember Own Offer and Offers in Village



Data from household recall survey following 2018 implementation. Each datapoint represents a village. x-axis: fraction of eligible households that remember loan offers made in their village. y-axis: fraction of eligible households that recall receiving a loan offer. 45-degree line in red.

Figure 8: Loan Offer Targeting to Always-Takers



Each datapoint represents a household characteristic. x-axis: t-statistic for difference in that characteristic between always-takers and never-takers. y-axis: impact of that characteristic on likelihood of remembering a loan offer.

Figure 9: Distribution of Per Capita Calorie Consumption by Year

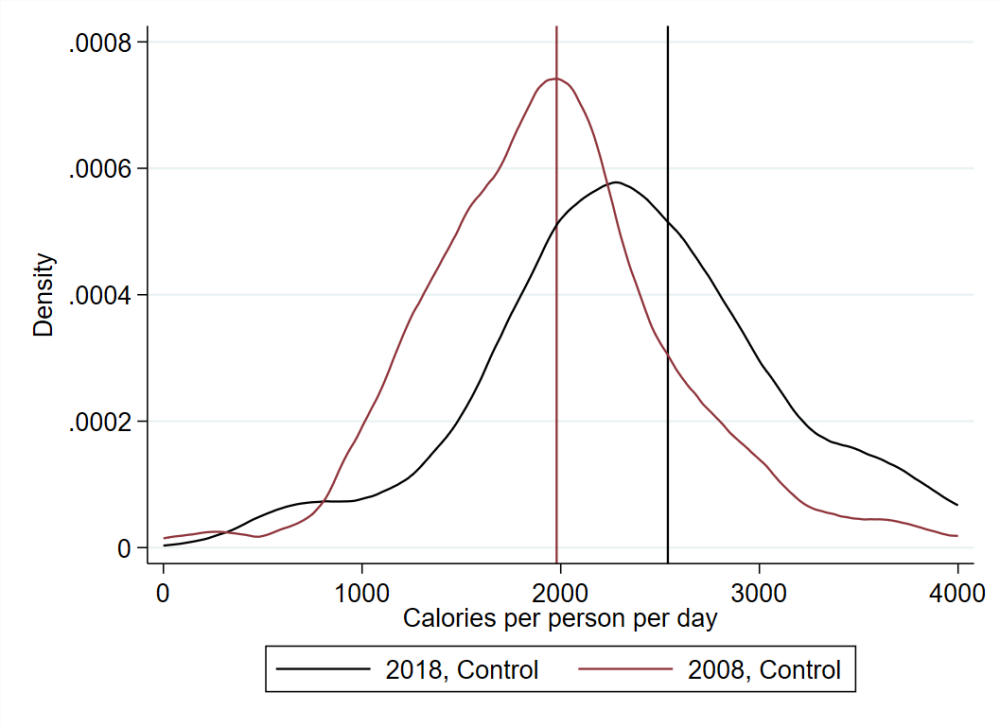
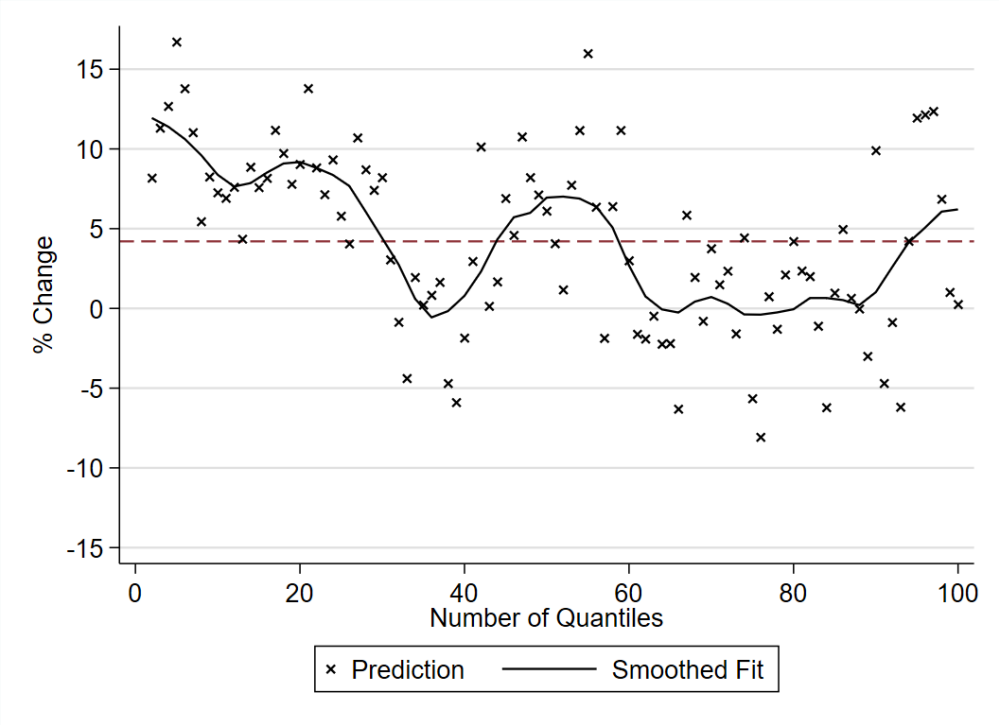
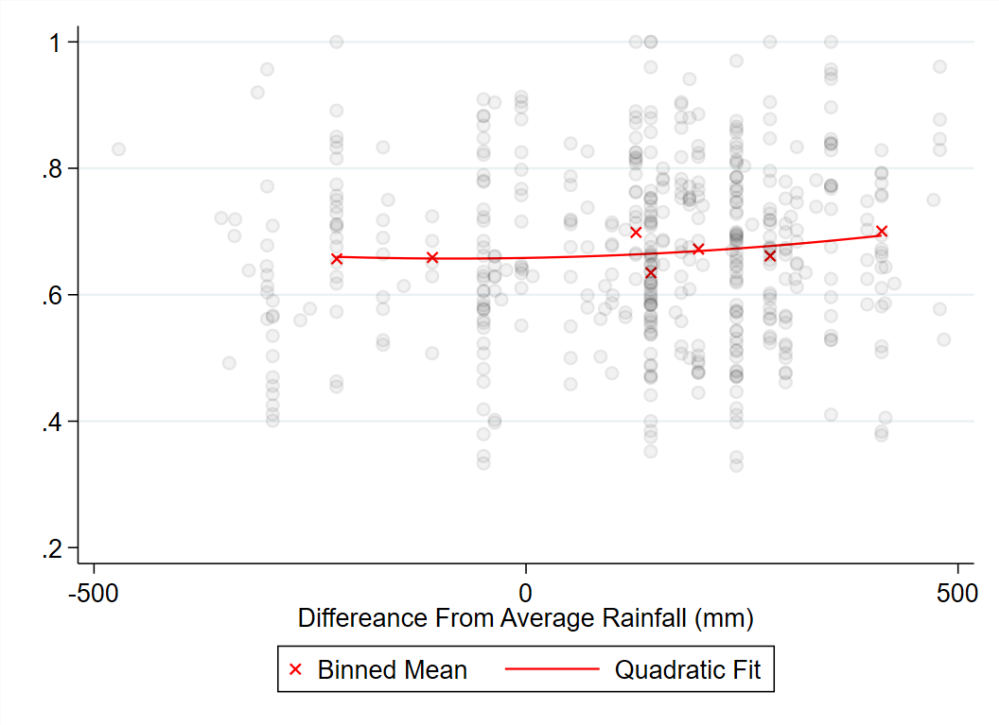


Figure 10: Treatment Effect Attenuation Predicted by Change in Calorie Consumption over Time



x-axis plots number of quantile bins used to compute conditional treatment effects. y-axis plots fraction of difference between pilot and at-scale treatment effect explained by reweighting 2008 conditional treatment effects using 2018 population shares.

Figure 11: 2018 Loan Acceptance Rate by Rainfall Relative to Average



Loan acceptance among eligible households based on administrative data from microfinance institution. x-axis: rainfall during 2018 lean season relative to 2001–2019 average. y-axis: fraction of eligible households in village that accept a migration loan.

Table 1: Balance across Treatment Arms, 2017

Variable	(1) pure-control Mean/SE	(2) branch-control Mean/SE	(3) spillover Mean/SE	(4) incentivized Mean/SE	T-test Difference (1)-(2)	(1)-(3)	(1)-(4)	(2)-(3)	(2)-(4)	(3)-(4)
Never migrated or migrated over 3 years ago	0.548	0.602	0.617	0.626	-0.053	-0.068*	-0.078**	-0.015	-0.024	-0.009
	(0.025)	(0.029)	(0.032)	(0.027)						
Migrated 2-3 years ago	0.087	0.073	0.075	0.069	0.014	0.012	0.018	-0.002	0.003	0.005
	(0.008)	(0.011)	(0.011)	(0.011)						
Migrated a year ago	0.365	0.326	0.309	0.305	0.039	0.056	0.060*	0.017	0.021	0.004
	(0.023)	(0.028)	(0.029)	(0.022)						
No land	0.442	0.423	0.502	0.513	0.019	-0.060	-0.071	-0.079	-0.090	-0.011
	(0.031)	(0.040)	(0.035)	(0.040)						
Below Med | on $0 <$ land ≤ 50	0.201	0.162	0.150	0.158	0.040	0.051**	0.043*	0.011	0.004	-0.007
	(0.016)	(0.019)	(0.017)	(0.017)						
Above Med | on $0 <$ land ≤ 50	0.165	0.168	0.136	0.154	-0.003	0.030	0.012	0.032	0.014	-0.018
	(0.017)	(0.021)	(0.017)	(0.019)						
Below Med | on land > 50	0.105	0.116	0.097	0.087	-0.011	0.008	0.018	0.019	0.029	0.010
	(0.012)	(0.013)	(0.013)	(0.012)						
Above Med | on land > 50	0.087	0.132	0.115	0.088	-0.045*	-0.028	-0.002	0.017	0.043	0.026
	(0.013)	(0.022)	(0.019)	(0.019)						
Completed Some Education	0.449	0.405	0.446	0.479	0.044	0.003	-0.031	-0.041	-0.074**	-0.033
	(0.017)	(0.027)	(0.030)	(0.024)						
Num of Adult Males in the Household	1.647	1.604	1.590	1.508	0.042	0.057	0.139***	0.014	0.097*	0.082
	(0.027)	(0.041)	(0.036)	(0.041)						
Num of Children in the Household	1.536	1.458	1.513	1.445	0.079	0.024	0.092	-0.055	0.013	0.068
	(0.040)	(0.052)	(0.060)	(0.043)						
Baseline Food Insecurity	0.566	0.545	0.549	0.731	0.021	0.017	-0.165***	-0.004	-0.186***	-0.182***
	(0.041)	(0.045)	(0.042)	(0.048)						
N	1288	743	751	780						
Clusters	70	40	40	40						
F-test of joint significance (p-value)					0.107	0.239	0.000***	0.636	0.002***	0.100
F-test, number of observations					2031	2039	2068	1494	1523	1531

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are clustered at variable village. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 2: Balance across Treatment Arms, 2018

Variable	(1) pure-control Mean/SE	(2) branch-control Mean/SE	(3) spillover Mean/SE	(4) incentivized Mean/SE	T-test Difference (1)-(2)	(1)-(3)	(1)-(4)	(2)-(3)	(2)-(4)	(3)-(4)
Never migrated or migrated over 3 years ago	0.554	0.483	0.465	0.452	0.070*	0.089**	0.102***	0.018	0.031	0.013
	(0.020)	(0.031)	(0.036)	(0.020)						
Migrated 2-3 years ago	0.091	0.102	0.119	0.103	-0.011	-0.027*	-0.011	-0.016	-0.000	0.016
	(0.007)	(0.015)	(0.013)	(0.008)						
Migrated a year ago	0.355	0.414	0.417	0.446	-0.059*	-0.061	-0.091***	-0.002	-0.031	-0.029
	(0.020)	(0.029)	(0.035)	(0.020)						
No land	0.825	0.809	0.796	0.818	0.016	0.029	0.007	0.013	-0.009	-0.023
	(0.011)	(0.024)	(0.019)	(0.012)						
Below Med | on $0 <$ land ≤ 50	0.095	0.119	0.106	0.096	-0.024	-0.011	-0.001	0.013	0.023	0.010
	(0.008)	(0.016)	(0.013)	(0.009)						
Above Med | on $0 <$ land ≤ 50	0.080	0.072	0.098	0.086	0.008	-0.018	-0.006	-0.026	-0.014	0.012
	(0.008)	(0.012)	(0.015)	(0.009)						
Completed Some Education	0.451	0.369	0.373	0.433	0.082***	0.078**	0.018	-0.004	-0.064**	-0.060*
	(0.017)	(0.023)	(0.025)	(0.017)						
Num of Adult Males in the Household	1.525	1.518	1.480	1.496	0.007	0.045	0.029	0.038	0.023	-0.015
	(0.025)	(0.033)	(0.037)	(0.024)						
Num of Children in the Household	1.554	1.509	1.513	1.546	0.044	0.040	0.007	-0.004	-0.037	-0.033
	(0.036)	(0.050)	(0.055)	(0.040)						
Borrowed Money at Baseline	0.806	0.849	0.785	0.809	-0.043*	0.021	-0.003	0.064*	0.040	-0.024
	(0.014)	(0.020)	(0.028)	(0.016)						
Baseline Food Insecurity	0.731	0.648	0.738	0.760	0.083*	-0.007	-0.028	-0.090*	-0.111**	-0.022
	(0.020)	(0.041)	(0.034)	(0.019)						
Village flooding during the last year	0.322	0.462	0.521	0.434	-0.139	-0.199**	-0.112	-0.059	0.027	0.087
	(0.050)	(0.082)	(0.084)	(0.051)						
N	1411	654	641	1618						
Clusters	100	40	39	99						
F-test of joint significance (p-value)					0.001***	0.024**	0.086*	0.028**	0.013**	0.439
F-test, number of observations					2065	2052	3029	1295	2272	2259

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are clustered at variable village. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 3: Estimated Treatment Effect on Household Migration

	Evaluation Round			
	2008	2014	2017	2018
Low Offers	0.18	0.25		
	(0.03)	(0.04)		
High Offers		0.40	0.06	0.06
		(0.03)	(0.03)	(0.03)
T for $H_0 : \beta = \beta_{08,Low}$			2.86	3.13
T for $H_0 : \beta = \beta_{14,High}$			7.21	7.82
Control Mean	0.36	0.34	0.36	0.38
HH Controls	Yes	No	Yes	Yes
Upazila FEs	Yes	Yes	Yes	Yes
N	1826	3600	3678	4324

Outcome is a dummy for whether any member of the household migrated during the lean season. Standard errors clustered at the village level.

Table 4: Baseline Characteristics of Eligible Population by District, 2017

Variable	(1) New Districts Mean/SE	(2) Original Districts Mean/SE	T-test Difference (1)-(2)
Net Income	39924.426	39292.854	631.572
	(1336.293)	(1444.903)	
Never migrated or migrated over 3 years ago	0.565	0.527	0.039
	(0.034)	(0.039)	
Migrated 2-3 years ago	0.096	0.076	0.019
	(0.012)	(0.011)	
Migrated a year ago	0.339	0.397	-0.058
	(0.029)	(0.036)	
No land	0.361	0.541	-0.179***
	(0.040)	(0.043)	
Below Med | on $0 <$ land ≤ 50	0.222	0.175	0.047
	(0.024)	(0.020)	
Above Med | on $0 <$ land ≤ 50	0.181	0.146	0.036
	(0.023)	(0.024)	
Below Med | on land > 50	0.128	0.076	0.052**
	(0.017)	(0.014)	
Above Med | on land > 50	0.107	0.062	0.045*
	(0.020)	(0.014)	
Completed Some Education	0.457	0.438	0.019
	(0.024)	(0.026)	
Num of Adult Males in the Household	1.678	1.608	0.070
	(0.036)	(0.041)	
Num of Children in the Household	1.533	1.541	-0.008
	(0.058)	(0.054)	
Baseline Food Insecurity	0.505	0.641	-0.136*
	(0.054)	(0.061)	
N	711	577	
Clusters	38	32	
F-test of joint significance (p-value)			0.005***
F-test, number of observations			1288

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are clustered at variable village. All missing values in balance variables are treated as zero.***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 5: Baseline Characteristics of Eligible Population by District, 2018

Variable	(1) New Districts Mean/SE	(2) Original Districts Mean/SE	T-test Difference (1)-(2)
Net Income	41586.615	37833.030	3753.585
	(1831.982)	(1644.920)	
Never migrated or migrated over 3 years ago	0.556	0.549	0.007
	(0.027)	(0.031)	
Migrated 2-3 years ago	0.082	0.105	-0.023
	(0.009)	(0.013)	
Migrated a year ago	0.361	0.346	0.016
	(0.026)	(0.031)	
No land	0.812	0.844	-0.032
	(0.015)	(0.015)	
Below Med | on $0 <$ land ≤ 50	0.107	0.077	0.030*
	(0.011)	(0.010)	
Above Med | on $0 <$ land ≤ 50	0.081	0.079	0.002
	(0.010)	(0.012)	
Completed Some Education	0.452	0.449	0.003
	(0.019)	(0.031)	
Num of Adult Males in the Household	1.501	1.561	-0.061
	(0.031)	(0.042)	
Num of Children in the Household	1.560	1.544	0.016
	(0.048)	(0.056)	
Borrowed Money at Baseline	0.805	0.807	-0.002
	(0.019)	(0.023)	
Baseline Food Insecurity	0.731	0.732	-0.000
	(0.027)	(0.029)	
Village flooding during the last year	0.203	0.498	-0.295***
	(0.056)	(0.083)	
N	841	570	
Clusters	60	40	
F-test of joint significance (p-value)			0.010**
F-test, number of observations			1411

Notes: The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are p-values. Standard errors are clustered at variable village. All missing values in balance variables are treated as zero.***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 6: Geographic Differences in Treatment Effects

	2017 Districts			2018 Districts		
	All	Pilot	New	All	Pilot	New
Treated	0.06	0.04	0.06	0.06	0.12	-0.03
	(0.03)	(0.05)	(0.04)	(0.03)	(0.04)	(0.03)
Spillover	-0.03	-0.01	-0.05	0.02	0.13	-0.09
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.04)
Branch Control	-0.02	-0.02	-0.02	0.02	0.09	-0.07
	(0.04)	(0.06)	(0.04)	(0.03)	(0.05)	(0.04)
Control Mean	0.36	0.40	0.33	0.38	0.39	0.38
HH Controls	Yes	Yes	Yes	Yes	Yes	Yes
Upazila FEs	Yes	Yes	Yes	Yes	Yes	Yes
N	3,678	1,537	2,141	4,324	1,901	2,423

Outcome is a dummy for whether any member of the household migrated during the lean season. Standard errors clustered at the village level.

Table 7: Sources of Administrative Leakage from Loan Eligibility to Migration

	2008	2017	2017 alt	2018
% Qualified from Population	56.4	76.5	21.6	61.6
% Eligible from Qualified	9.8	100.0	100.0	100.0
% Accepted Offer from Eligible	64.2	42.8	48.4	62.6
% Received Loan from Accepted	.	62.4	61.0	99.1
% Migrated from Disbursed	65.9	91.3	91.5	70.0
% Migrated with Loan from Eligible	42.3	23.3	25.7	41.0
Population	21,902	206,655	206,655	232,916

Notes: Loan offers and migration from initial census population according to administrative implementation data. "2017 alt" refers to subset of 2017 qualified population that satisfied the 2008/2018 eligibility criteria according to the (poorly recorded) history of missing meals. In 2017 and 2018, 5% of loan recipients were not surveyed for migration status, so %Migrated from Disbursed may vary by up to 5%.

Table 8: Baseline Differences between Always- and Never-Takers

Household Characteristics	Always Taker vs Never Taker: Difference	Always Taker vs Never Taker: T-statistic	Remember Offer from RDRS
Never migrated or migrated over 3 years ago	-0.3723	-15.228	-0.148
			(0.024)
Migrated 2-3 years ago	-0.0385	-2.118	-0.044
			(0.042)
Migrated a year ago	0.4108	16.214	0.152
			(0.024)
No land	0.0381	1.762	0.072
			(0.028)
Below med | on $0 <$ land ≤ 50	-0.0045	-0.267	-0.070
			(0.038)
Above med | on $0 <$ land ≤ 50	-0.0336	-2.177	-0.058
			(0.040)
Completed some education	-0.0485	-1.760	-0.046
			(0.023)
Zero adult males	-0.0533	-4.764	-0.160
			(0.068)
One adult male	-0.0181	-0.657	-0.008
			(0.026)
Two adult males	0.0326	1.320	0.037
			(0.028)
Three or more adult males	0.0388	2.122	0.000
			(0.034)
Borrowed money at baseline	0.0539	2.449	0.094
			(0.035)
Baseline food insecurity	-0.0054	-0.221	-0.020
			(0.029)
Village flooding during the last year	0.0145	0.541	-0.033
			(0.052)
Obs	1363	1363	1618
Sample	Always Takers and Never Takers	Always Takers and Never Takers	Incentivized Villages

Table 9: Distribution and Effort Intensity by Compliance Status

Type	Accept Offer	Migrate w/ Loan	Migrate w/o Loan	2008 Freq.	2018 Effort Intensity
Always Taker	Yes	Yes	Yes	0.20	0.76
Self Sufficient	No	–	Yes	0.16	0.17
Complier	Yes	Yes	No	0.22	0.49
Time Waster	Yes	No	No	0.21	0.84
Never Taker	No	–	No	0.20	0.22

Notes: Population frequencies in Column 5 assume uniform treatment intensity in 2008 normalized to 1. Effort intensity in Column 6 assume type distribution remains constant from 2008 to 2018.

Table 10: 2018 NLS Treatment Effects by Rainfall

	(1) Sent Migrant	(2) Sent Migrant	(3) Sent Migrant
Treated	0.07**	0.06	0.02
	(0.03)	(0.12)	(0.09)
Treated × Above Avg Rain=1			0.05
			(0.09)
Spillover	0.05	-0.01	-0.05
	(0.04)	(0.12)	(0.09)
Spillover × Above Avg Rain=1			0.08
			(0.10)
Branch Control	0.05	0.00	-0.07
	(0.04)	(0.12)	(0.09)
Branch Control × Above Avg Rain=1			0.11
			(0.10)
Above Avg Rain=1			-0.04
			(0.08)
Mean Rainfall	2222.021	1758.971	2122.001
R-Squared	0.140	0.144	0.132
Obs	3390	934	4324
Controls	Yes	Yes	Yes
Upazila FE	Yes	Yes	Yes
Sample	Above Average Rainfall	Below Average Rainfall	Full Sample

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Supplementary Appendix for "External Validity and Implementation at Scale"

A Supplemental Results

A.1 Alternate Migration Measures

The measure of migration reported in the paper uses the self-reported departure date for each migration episode to construct a dummy variable that covers the same months as 2008 and 2014. Tables S1 and S2 report results using all migration reported at endline and midline, respectively. Although the control mean changes because the endline survey covers a larger migration window and the midline survey covers a smaller one, the estimated treatment effects remain consistent across measures.

A.2 Reduced Form Outcomes

In Tables S3–S11 we present reduced-form estimates of the effect of treatment assignment on various measures of earnings and consumption, including all measures prespecified in the prea-analysis plan. The estimates are uniformly small in magnitude and statistically indistinguishable from zero. We reject the possibility that the large program benefits observed in pilot resulted from replacing other forms of credit for regular migrants rather than enabling induced migrants to access the returns to migration.

Table S1: Estimated Treatment Effect on Household Migration

	2017 Districts			2018 Districts		
	All	Pilot	New	All	Pilot	New
Treated	0.04	0.06	0.01	0.05	0.10	-0.02
	(0.04)	(0.05)	(0.04)	(0.03)	(0.04)	(0.03)
Spillover	0.00	0.07	-0.06	0.04	0.14	-0.06
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.05)
Branch Control	0.02	0.03	0.00	0.04	0.13	-0.06
	(0.04)	(0.06)	(0.05)	(0.04)	(0.06)	(0.05)
Control Mean	0.47	0.50	0.45	0.47	0.47	0.47
HH Controls	Yes	Yes	Yes	Yes	Yes	Yes
Upazila FEs	Yes	Yes	Yes	Yes	Yes	Yes
N	3,678	1,537	2,141	4,324	1,901	2,423

Table S2: Estimated Treatment Effect on Household Migration

	2017 Districts			2018 Districts		
	All	Pilot	New	All	Pilot	New
Treated	0.04	0.04	0.03	0.08	0.13	0.01
	(0.03)	(0.05)	(0.04)	(0.03)	(0.04)	(0.03)
Spillover	-0.05	-0.03	-0.08	0.05	0.14	-0.03
	(0.03)	(0.05)	(0.04)	(0.04)	(0.06)	(0.04)
Branch Control	-0.03	-0.01	-0.05	0.02	0.07	-0.04
	(0.03)	(0.06)	(0.04)	(0.04)	(0.05)	(0.05)
Control Mean	0.31	0.35	0.28	0.34	0.36	0.33
HH Controls	Yes	Yes	Yes	Yes	Yes	Yes
Upazila FEs	Yes	Yes	Yes	Yes	Yes	Yes
N	4,428	1,839	2,589	4,324	1,901	2,423

Table S3: Effect on Net Income at Endline

	Evaluation Round			
	2017		2018	
Treated	1052.98	-243.30	398.22	880.09
	(1742.00)	(1368.08)	(1605.94)	(1713.60)
Spillover	424.29	-1417.60	896.56	-403.95
	(1799.97)	(1423.02)	(2181.67)	(2137.09)
Branch Control	1553.92	23.43	1077.10	1603.45
	(1490.58)	(1450.29)	(2085.95)	(2027.17)
Control Mean	39712.55	39712.55	40030.49	40030.49
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,477	3,477	4,324	4,324

Table S4: Effect on Food Expenditure at Endline

	Evaluation Round			
	2017		2018	
Treated	-88.08	-2.35	39.31	103.76
	(142.36)	(134.44)	(187.68)	(184.80)
Spillover	129.91	222.47	-251.29	125.92
	(141.23)	(132.03)	(264.96)	(220.53)
Branch Control	-62.00	15.46	15.12	339.58
	(153.06)	(148.57)	(236.37)	(187.74)
Control Mean	5147.39	5147.39	2582.21	2582.21
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,536	3,536	3,516	3,516

Table S5: Effect on Food Insecurity at Midline

	Evaluation Round			
	2017		2018	
Treated	0.09	0.23	0.40	0.03
	(0.29)	(0.24)	(0.21)	(0.18)
Spillover	0.02	0.07	0.78	0.42
	(0.30)	(0.24)	(0.32)	(0.22)
Branch Control	0.20	0.25	1.16	0.72
	(0.28)	(0.23)	(0.30)	(0.24)
Control Mean	3.51	3.51	9.87	9.87
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,678	3,678	4,064	4,064

Table S6: Effect on Food Insecurity at Endline

	Evaluation Round			
	2017		2018	
Treated	0.11	0.10	0.13	0.02
	(0.21)	(0.16)	(0.22)	(0.21)
Spillover	-0.04	0.01	-0.02	0.06
	(0.22)	(0.17)	(0.27)	(0.25)
Branch Control	0.11	0.14	0.40	0.20
	(0.20)	(0.16)	(0.28)	(0.23)
Control Mean	3.71	3.71	12.01	12.01
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,678	3,678	4,155	4,155

Table S7: Effect on Working at Midline

	Evaluation Round			
	2017		2018	
Treated	-0.01	-0.01	-0.03	-0.02
	(0.01)	(0.01)	(0.01)	(0.01)
Spillover	0.00	0.00	-0.02	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Branch Control	0.00	0.00	-0.02	-0.01
	(0.01)	(0.01)	(0.02)	(0.02)
Control Mean	0.94	0.94	0.95	0.95
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,678	3,678	4,324	4,324

Table S8: Effect on Working at Endline

	Evaluation Round			
	2017		2018	
Treated	0.00	0.00	0.02	0.03
	(.)	(.)	(0.03)	(0.02)
Spillover	0.00	0.00	0.02	0.01
	(.)	(.)	(0.03)	(0.03)
Branch Control	0.00	0.00	-0.02	-0.01
	(.)	(.)	(0.04)	(0.03)
Control Mean	0.00	0.00	0.74	0.74
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,678	3,678	4,324	4,324

Table S9: Effect on Earning at Midline

	Evaluation Round			
	2017		2018	
Treated	-206.93	-49.89	-126.98	-120.16
	(138.45)	(105.01)	(127.77)	(140.82)
Spillover	72.26	202.64	-55.97	-123.31
	(178.88)	(166.24)	(147.71)	(156.94)
Branch Control	-250.55	-112.32	-193.89	-128.24
	(134.76)	(109.53)	(142.07)	(148.10)
Control Mean	2425.00	2425.00	1933.28	1933.28
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,678	3,678	4,324	4,324

Table S10: Effect on Earning at Endline

	Evaluation Round			
	2017		2018	
Treated	0.00	0.00	13.92	-9.76
	(.)	(.)	(111.88)	(133.81)
Spillover	0.00	0.00	-14.19	-58.59
	(.)	(.)	(110.95)	(131.91)
Branch Control	0.00	0.00	-122.88	-157.07
	(.)	(.)	(110.80)	(130.96)
Control Mean	0.00	0.00	1359.57	1359.57
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,678	3,678	4,324	4,324

Table S11: Effect on Per Capital Daily Calories at Endline

	Evaluation Round			
	2017		2018	
Treated	0.00	0.00	1.43	25.47
	(.)	(.)	(59.34)	(59.70)
Spillover	0.00	0.00	-166.67	-157.44
	(.)	(.)	(60.06)	(60.66)
Branch Control	0.00	0.00	-41.52	-19.41
	(.)	(.)	(59.41)	(64.03)
Control Mean	0.00	0.00	2502.72	2502.72
HH Controls	Yes	Yes	Yes	Yes
Upazila FEs	No	Yes	No	Yes
N	3,678	3,678	4,324	4,324