

NBER WORKING PAPER SERIES

IMPROVING LAST-MILE SERVICE DELIVERY USING PHONE-BASED MONITORING

Karthik Muralidharan
Paul Niehaus
Sandip Sukhtankar
Jeffrey Weaver

Working Paper 25298
<http://www.nber.org/papers/w25298>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2018

We are grateful to officials in the Government of Telangana, especially Mr. K Ramakrishna Rao and Mr. C Parthasarathi. This paper would not have been possible without the efforts and inputs of the J-PAL South Asia/UCSD project team in the Payments and Governance Research Program, including Kartik Srivastava, Avantika Prabhakar, Frances Lu, Vishnu Padmanabhan, Surya Banda, Mayank Sharma, and Burak Eskici. Finally, we thank the Strategic Impact Evaluation Fund (SIEF) at the World Bank (especially Alaka Holla) and the Bill and Melinda Gates Foundation (especially Dan Radcliffe) for the financial support that made this study possible. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2018 by Karthik Muralidharan, Paul Niehaus, Sandip Sukhtankar, and Jeffrey Weaver. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Improving Last-Mile Service Delivery using Phone-Based Monitoring
Karthik Muralidharan, Paul Niehaus, Sandip Sukhtankar, and Jeffrey Weaver
NBER Working Paper No. 25298
November 2018
JEL No. C93,D73,H53,O33

ABSTRACT

Improving "last mile" public-service delivery is a recurring challenge in developing countries. Could the rapid adoption of mobile phones provide a simple, cost-effective means to do so? We evaluate the impact of a phone-based monitoring system on improving the delivery of a program that transferred nearly a billion dollars to farmers in the Indian state of Telangana, using an at-scale experiment randomized across 5.7 million farmers. A randomly selected sample of officials were told that a representative sample of beneficiaries in their jurisdiction would be called to measure the quality of program implementation. This simple announcement led to a 1.5% increase in the number of farmers receiving their benefits, with a 3.3% increase among farmers in the bottom quartile of landholdings. The program was highly cost-effective, with a cost of 3.6 cents for each additional dollar delivered.

Karthik Muralidharan
Department of Economics, 0508
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0508
and NBER
kamurali@ucsd.edu

Paul Niehaus
Department of Economics
University of California, San Diego
9500 Gilman Drive #0508
La Jolla, CA 92093
and NBER
pniehaus@ucsd.edu

Sandip Sukhtankar
Department of Economics
University of Virginia
Charlottesville, VA 22904
sandip.sukhtankar@virginia.edu

Jeffrey Weaver
University of Southern California,
Department of Economics
3620 South Vermont Ave.
Kaprielian (KAP) Hall, 300
Los Angeles, CA 90089
jeffrey.b.weaver@gmail.com

A randomized controlled trials registry entry is available at
<https://www.socialscienceregistry.org/trials/2942>

1 Introduction

In developing countries, governments often find good policies easier to design than to implement (Pritchett, 2009). Managing the “last mile” of public service delivery has thus been a recurrent theme in recent research, from ensuring that employees show up to work (Muralidharan et al., 2017; Dufflo et al., 2012) to simply ensuring that food or money reach the people entitled to them (Muralidharan et al., 2016), among other examples.

In part the challenge is that, like any organization, a government can only manage what it measures. As Bloom and Van Reenen (2007) emphasize, collecting and analyzing measures of performance is one of the hallmarks of “good management.” Yet measuring service delivery is difficult. Front-line work typically takes place across thousands of communities, many of them remote. Internal reporting passes through layers of bureaucracy with incentives to apply spin – exaggerating its own performance, or overstating problems with initiatives it wants to undercut.¹ Those independent, representative surveys that are conducted (such as the Living Standards Measurement Surveys or India’s National Sample Survey) are typically too small, infrequent or both to be of much use for management.

In this paper we test a simple approach to improving last-mile service delivery by measuring whether people get what they are due: calling and asking. This approach leverages the rapid increase in mobile phone penetration in low-income countries, from 1 mobile subscription per 100 people in 2002 to 62 in 2017.² In many countries, mobile phone diffusion provides governments their first realistic opportunity to obtain quick, cheap, independent information about last-mile service delivery. If monitoring by phone works, this approach has the potential to be scaled across an unusually wide range of places, programs and outcomes.

We examine the potential for phone-based measurement to improve service delivery in the context of a high-stakes initiative of the state government of Telangana in India (GoTS). Between May and July of 2018, GoTS distributed \$0.9B, or around 7% of the state’s annual budget, as lump-sum payments to farmers with the aim of financing investments in seed and fertilizer. Responsibility for implementing the scheme rested primarily with Mandal (sub-district) Agricultural Officers (MAOs), who managed the process for physically delivering checks to farmers. Payments were issued in the form of special-purpose physical checks, which enables us to use bank administrative records of check encashment as a manipulation-proof measure of the program outcome.

¹In Andhra Pradesh, for example, the state government nearly shut down a highly-effective reform that significantly reduced corruption because officials (whose rents were threatened) reported a stream of cherry-picked negative anecdotes (Muralidharan et al., 2016).

²https://data.worldbank.org/indicator/IT.CEL.SETS.P2?end=2017&locations=XM&name_desc=true&start=1998, accessed 26 September 2018.

Working with the government, we implemented an experimental, at-scale test of phone-based performance measurement. We randomly assigned each MAO to either a treatment condition, in which they were told that we would call at least 100 of the recipients for whom they were responsible and produce reports visible to them and their supervisors, or to a control condition. This communication was conducted via a video-conference with senior officials and reinforced with a formal letter to treated MAOs. This design allows us to “experiment at scale” in the sense that we randomize treatment across (and observe outcomes for) all 5.7 million land-owning farm households in a state of 35 million people, the intervention was implemented by government at that scale, and the unit of randomization (one or more mandals) was large enough for treatment effects to be inclusive of local spillovers (Muralidharan and Niehaus, 2017).

The call center made 22,565 outbound calls in two weeks (during the peak four weeks of program implementation), and data from these calls were used to create reports of MAO-level performance (both absolute and relative). In practice, MAOs saw these reports only after most of their work on the program was completed, mainly because the program was implemented over a very short time window. Thus, the intervention only affected their awareness that their performance was being measured, but did not provide them with additional information about where best to focus their efforts. In this sense we view the results as a lower bound on the potential effects of the approach; in other contexts, the information collected by phone reports could be used by frontline workers to improve effort allocation, and could also be used for more explicit rewards or sanctions (which were not done in this case).

We find that phone-based monitoring significantly increased the rate of successful transfers by 1.3 percentage points relative to a base rate of 83% in control areas, or 1.5%. The difference in encashment rates between treatment and control groups peaks in the weeks immediately after check distribution began and declines somewhat thereafter, but appears to stabilize by September 2018 when we obtained the most updated set of administrative records. This corresponds to an increase in the money delivered to farmers of approximately one million dollars, and, if extended to the entire state, would result in \$8.6 million more being delivered annually.

We also find that the incidence of the intervention was mildly progressive, although the cash transfer program itself was regressive (since check sizes were proportional to landholdings). For farmers in the bottom quartile of landholdings, the increase in check encashment rate was 2.2 percentage points (3.3%), around twice the overall effect. A further noteworthy result is that although MAO performance could only be measured by calling beneficiaries with cell phones (around 60% of the population), we cannot reject that the measured improvements in performance in treated areas are similar across beneficiaries with and without

phones.

Finally, we estimate that phone-based measurement was highly cost-effective. Costing the intervention at the contracted cost GoTS paid to the vendor who ran the call center, we estimate that the cost per dollar of benefits delivered to beneficiaries was 3.6 cents, which is lower than the administrative cost of almost any anti-poverty program for which such data is available. To calculate the economic return to the program, we define benefits as the difference between the estimated return on capital held by farmers as opposed to the government. Even under relatively conservative assumptions we estimate a benefit of 4 times the cost incurred. This highlights the potential gains from using simple, generic technology already widely in use to monitor large-scale programs, where even small rate improvements can create substantial value.

Our investigation of performance measurement using a simple, general-purpose technology (phone calls) complements recent work testing more specialized approaches. There are obvious tradeoffs: monitoring technology developed for specific applications can presumably do more for those applications. For example, the time clocks examined by Banerjee et al. (2008) or the custom smart-phone app developed and tested by Callen et al. (2018) likely yield better information about whether front-line health workers are doing their jobs than would phone calls to random community members. Relative to these specialized approaches, measurement by phone has the advantages of (i) low fixed and variable costs and time to deploy, as call center services are typically available quickly and priced as an inexpensive service,³ and (ii) the flexibility to scale across an unusually wide number range of places, programs and outcomes.

Empirical evidence on service delivery in developing countries suggests that top-down administrative monitoring has typically been more effective than bottom-up community monitoring - partly for reasons of free-riding and partly due to asymmetry in power between citizens and officials (Olken, 2007; Muralidharan et al., 2017; Banerjee et al., 2010). However, a practical barrier to scaling up top-down monitoring has been the high cost of obtaining credible high-frequency data on last-mile service delivery at enough of a spatially disaggregated level to enable holding appropriate staff accountable. Our intervention and results suggest that using outbound call-centers to call representative samples of beneficiaries (who increasingly have access to a phone) provides a simple and scalable solution to this barrier.⁴

³The point is best illustrated by the example of this project: We were able to move from an in-principle agreement with the GoTS to implementing the full experiment within one month.

⁴The cost of measuring performance is also an important parameter in the literature on optimal level of decentralization (Mookherjee, 2015). Thus, changes in these costs may also have implications on this dimension.

2 Setting and intervention

Telangana is India’s newest state, created in 2014 from erstwhile Andhra Pradesh. It has a population of 35 million, with around 60% being rural. It is relatively well-off, with per capita income estimated to be 53% higher than the all-India average as of 2016-2017 (Government of Telangana, 2016). It is also thought to be relatively well-administered; Andhra Pradesh ranked 3rd out of 19 major states in the most recent Government Performance Index (Mundle et al., 2016).

2.1 The Rythu Bandhu scheme

The Government of Telangana (GoTS) introduced its flagship Rythu Bandhu (or “Friend of the Farmer”) Scheme (RBS) in May of 2018. The program aimed to provide capital to farmers for purchasing agricultural inputs such as seeds and fertilizer prior to the main agricultural season. The RBS was a response to widespread farm distress, and was hailed by economists as a more efficient way of helping farmers relative to other frequently deployed policies such as raising procurement prices, or loan waivers (Subramanian, 2018). It authorized payment of Rs. 4,000 (approximately \$55) per acre to every farmer registered as a landholder in the government’s land registry, which the government had updated and digitized in 2017.⁵

GoTS issued transfers in the form of “order checks” which could be exchanged for cash or deposits at any branch of the bank listed on the check, whether or not the beneficiary held an account with that bank. The Department of Agriculture managed the distribution of checks, with mandal agricultural officers (MAOs) responsible for their respective mandals. MAOs supervised teams of agricultural extension workers who conducted individual village meetings, originally scheduled for 8 to 12 May. To encash checks, banks required that holders produce an official form of identification matching the name printed on the check. The government allocated the 548 mandals in the state among 8 banks, assigning all farmers in a given mandal to the same bank.

Implementing RBS well was a priority for the government given the sum disbursed (approximately \$0.9B per cropping season or \$1.8B annually), the number of recipients (5.7 million), the high media profile of the scheme, the fact that the government had never before done anything comparable, and upcoming elections. Anticipated risks included (i) non-issue of checks, (ii) non-delivery of checks, (iii) late delivery of checks, which would force farmers to reduce investment or borrow at high rates to finance time-sensitive planting season

⁵Farmers with less than 1 gunta (0.025 acres) of land were issued checks of Rs. 100, and farmers with holdings of 100 or more acres were encouraged not to pick up their checks despite being entitled to them.

agricultural inputs, and (iv) corruption during the distribution process (e.g. bribe demands).

2.2 Phone-based monitoring intervention

The state government had previously collected phone numbers for farmers as part of land record digitization. Overall, 3.4 million of the 5.7 million entries (60%) in the registry list a contact number.

GoTS contracted a call center to collect data from beneficiaries between 29 May and 15 June.⁶ Using GoTS administrative records, we sampled approximately 150 farmers per treatment mandal and 50 farmers per control mandal from among those with a listed phone number, for a total of 46,007 targeted farmers.⁷ The vendor successfully completed audio calls with 22,565 of these farmers, a response rate of 49%.⁸ Calls collected information on whether and when the farmer received their check, whether and when they encashed it, any problems receiving or encashing the check (including time costs and bribes), and overall satisfaction with the program.

Prior to the distribution of cheques and calls to beneficiaries, the Telangana Department of Agriculture informed treatment MAOs that their mandals had been randomly selected to take part in a pilot of the phone-based monitoring system. During a special video conference, the state Commissioner of Agriculture explained the initiative and the data that would be collected. He informed MAOs that reports from the phone call data would be provided to them and their supervisors, including an implementation performance rating for their mandal. The MAOs were told which outcomes the report would cover, but not the specific formula for calculating ratings. On 10 May the Department of Agriculture sent treatment MAOs a follow-up letter containing the same information. To reduce the risk of spillovers, treatment MAOs were explicitly told the identity of other treatment MAOs in their district and that no other MAOs in their districts were part of the pilot.

Reports based on the phone data were issued to treatment MAOs and to their supervisors (district agricultural officers, DAOs) between 9 and 13 July. The reports listed five metrics: the proportion of farmers who reported receiving their checks, receiving them before 20 May, encashing their checks, being asked for a bribe, and being satisfied with the program overall. They showed performance on these metrics for the block in question, the district in which the block was located, and the state overall. They also showed a simple, color-coded categorical rating (“Poor,” “Fair,” “Good,” or “Excellent”) based on absolute performance, motivated

⁶The intervention was designed by the research team, but implemented by GoTS.

⁷Details of the sampling algorithm are in the online appendix.

⁸The vendor also piloted automated calls (IVR), but these performed poorly with a high error rate in capturing and encoding response, and so were discontinued.

in part by the finding of Callen et al. (2018) that simple “flagging” of high or low performers can make performance data more readily actionable. A redacted example report is in the online appendix.

In principle, the treatment included both a monitoring and an information component. In practice, the information provided by the reports came too late to meaningfully affect performance. This was mainly because of the very compressed time-frame of the entire program that aimed to distribute all checks in 4-6 weeks starting in early May (whereas the reports above were distributed 8 weeks later in early July). Thus, our estimates should be interpreted as reflecting the impact of MAO’s knowing they were being monitored, but without having the information from the calls that they could have used to potentially do their job better.

The Department of Agriculture did not explicitly inform control MAOs about the existence of the pilot. If they asked, they were told that the initiative might be extended to their areas in the future, but not during the current agricultural season. While the call center collected phone data from control mandals, it did not generate or issue reports using these data or inform control MAOs of their existence. Of course, the interpretation of reduced form intent-to-treat effects ultimately depends on treatment and control MAOs’ beliefs about what was happening, which we discuss below.

MAOs and their staff could potentially react to monitoring in several ways. They could improve scrutiny of processes to ensure that checks were issued to all eligible beneficiaries and that names on checks were not misspelt. They could work harder to find recipients – both before the village meetings by publicizing more thoroughly, or after them by following up on those who did not attend. They could also handle marginal cases in which an applicant’s documentation was incomplete more leniently.

That said, one might reasonably expect phone-based monitoring to have limited impact in this setting. Unlike many of its other programs, the government expected to receive timely and detailed administrative data on RBS implementation: MAOs recorded check distribution, and banks reported check encashment in order to claim reimbursement. Our intervention added an incremental, independent source of information, and synthesized reporting. Officials may also have interpreted the intervention as a signal that the government was taking RBS implementation seriously and was evaluating the performance of individual officers, rather than simply monitoring aggregate statistics. All told, however, we would generally expect to find larger effects in settings where phone monitoring was the only source of performance data, and where the program itself was not so high-profile and being actively monitored by top government officials.

3 Research methods

Our design and methods follow a registered pre-analysis plan.⁹

3.1 Experimental design

The study population consists of essentially all households eligible to receive RBS, i.e. all landholding households in Telangana. We excluded one largely urban district (Hyderabad) from the intervention as it did not contain a significant number of program beneficiaries, leaving 30 remaining districts.

Within these 30 districts, we randomly assigned treatment at the level of the MAO (a single MAO occasionally oversees multiple mandals). We selected approximately 25% of MAOs for treatment, yielding a total of 122 treatment MAOs and 376 control MAOs.¹⁰ This corresponded to 132 treatment mandals and 416 control mandals. We stratified randomization within each district on an indicator for whether an MAO oversees multiple mandals, the only MAO-level covariate available to us at the time of randomization. Figure 1 shows the geographical distribution of treatment and control mandals across the entire state.

Table 1 describes balance between treatment and control farmers on pre-specified landholder-level characteristics from the landholder registry as well mandal-level characteristics from the 2011 census. Columns 1 and 2 report means in the control and treatment groups, respectively, while column 3 reports the coefficient from a regression of the covariate on a treatment indicator and strata fixed effects. Of 11 tests, one (for share of the population that is Scheduled Tribe) is significant at the 10% level, consistent with what we would expect to see by chance. Further, since we randomized across nearly the universe of mandals in the state (outside Hyderabad), the study sample was representative of the rural population of the state.

3.2 Data

We primarily use administrative data provided by GoTS. These include (i) the complete register of agricultural landholders in the state, including names, village names, landholdings in acres, and in 61% of entries a contact phone number; (ii) an individual-level record of check distribution maintained by the MAOs; and (iii) individual-level records of check encashment uploaded by participating banks. Our analysis focuses on encashment, which is the ultimate outcome of interest to policy-makers. Bank reports of encashment were recorded in real-time

⁹See <https://www.socialscisceregistry.org/trials/2942>.

¹⁰Details of the randomization algorithm are in the online appendix.

and were the basis for reimbursement from the government, meaning that misrepresentation of encashment status would constitute serious fraud and could jeopardize a bank’s license to operate. We work with copies of these registries current as of 26 September, 2018. This is after the last date (15 August) on which the initial batches of RBS checks were valid for encashment and thus should well approximate the final result of the distribution of checks for this season.¹¹

We also use data from phone calls conducted by the call center itself as a secondary source. The call center data were collected over the phone from program beneficiaries as described above. The vendor attempted to reach 46,007 farmers, completed surveys with 49%, began but did not complete surveys with another 24%, had 10% decline to participate, and could not reach the remaining 17% for various reasons.

Finally, we use data from a short phone survey of MAOs to assess their beliefs about the intervention. We surveyed 88 of 122 treatment MAOs and a small sample of 54 control MAOs.¹² Surveys covered their awareness of the pilot and beliefs about their own treatment status. To minimize any potential Hawthorne effects, we conducted these surveys on 9 June after 84% of checks had already been encashed, and within only a small sample of control MAOs.

3.3 Estimation

We report intent-to-treat estimates, which compare average outcomes in treatment and control areas. The interpretation of these depends on MAO beliefs, which we discuss below. We thus estimate

$$y_{ivmsd} = \alpha + \beta T_{msd} + \delta_{sd} + \gamma X_{ivmsd} + \epsilon_{ivmsd} \quad (1)$$

where y is an outcome, T an indicator for assignment to treatment, and X a vector of pre-determined covariates (in practice, landholdings). Indices denote individual i in village v in mandal m in stratum s in district d .¹³ Note that treatment is strictly exogenous

¹¹Checks were initially printed in four tranches, on 19 April, 1 May, 10 May and 15 May, and were valid for three months from the date of printing. The government subsequently began proactively re-printing checks as they expired on a case-by-case basis. By 26 September encashment activity had largely ceased, with the checks encashed during the trailing month just 0.5% of all checks ever encashed.

¹²We attempted surveys with all of the treatment MAOs (72% response rate), but only attempted surveys with a random sample of 2 control MAOs per mandal (60% response rate).

¹³Technically, we estimate models using the administrative data at the individual x check level, as a handful (0.8%) of individuals in the database were issued multiple checks. According to the Revenue Department, amounts above Rs. 50,000 (corresponding to 12.5 acres of land) had to be split into multiple checks. Anecdotally, we also heard that multiple checks were given to farmers if they held land in different areas. In the data, 89% of farmers with multiple checks have more than 12.5 acres of land. Outcomes are essentially perfectly correlated within individual: farmers either picked up and encashed all or none of their checks.

conditional on the randomization stratum fixed effects δ_{sd} . We cluster standard errors at the level of treatment assignment (i.e. the MAO) and also conduct randomization inference as a robustness check. When using call center data, we reweight estimation by the inverse probability of being sampled using the algorithm above.

4 Results

4.1 Effects on overall program performance

Overall, RBS implementation was imperfect but relatively successful compared to other similar interventions. After 5 months, checks had been distributed to 3.7M farmers, or 87% of the target. Banks had encashed checks for 83% of farmers. Among checks ever encashed, the average time from the intended scheme launch date (8 May) to distribution was 11 days, and the average time to encashment was 20 days.

Phone-based monitoring nevertheless significantly improved implementation. Figure 2 summarizes the main effects visually. The top panel plots the proportion of checks encashed by date in the treatment and control groups separately. The bottom panel reports the estimated differences between these rates (to be precise, coefficients from a regression of an indicator for having encashed on a treatment indicator). The treatment effect widens to a peak of roughly 2.8 percentage points on 25 May ($p = 0.008$) and then narrows somewhat after this, asymptoting to 1.3 percentage points by 26 September.¹⁴

Panel A of Table 2 reports the corresponding average effects on check encashment metrics. Treatment increased the overall probability a farmer encashed their check by 1.3 percentage points ($p = 0.06$). Conditional on ever encashing, treatment also lowered by three-fourths of a day the mean number of days that passed before recipients encashed their checks ($p = 0.05$).¹⁵

4.2 Distributional consequences

The baseline allocation of benefits under RBS was regressive, as check size was proportional to registered landholdings: a farmer with 10 acres of land received ten times as much money

¹⁴Note that because significant treatment effects begin to appear only after the date on which phone calls were made, these calls were of limited value in examining mechanisms. In the appendix we show that treatment had small and insignificant effects on the likelihood that phone call respondents received their check at a Gram Sabha (village meeting), were asked to pay a bribe, or were satisfied overall with the scheme (Table A.2).

¹⁵For completeness we also report effects on check distribution (Table A.1). We treat these data with caution as they were uploaded by MAOs with substantial lags, causing dates to be mismeasured, and were not subject to penalties for misreporting like those banks faced. However, we find the same broad pattern of results here as in the encashment data.

as one with one acre. This pattern was exacerbated by differences in distribution and encashment rates. As of 25 September, 89% of farmers in the top quartile of the landholding distribution (those holding more than 3.1 acres of land) had encashed their checks, declining monotonically to 68% of farmers in the bottom quartile (those holding fewer than 0.4 acres). These patterns could reflect differences in the effort made by government officials, or differences in farmers' motivation to collect and encash their checks. At the bottom of the distribution a farmer with a twentieth of an acre of land would have received a check worth just Rs. 200 (approximately \$3), possibly less than the time and money costs of encashing.

Measuring performance via a phone-based monitoring system could differentially benefit either larger or smaller landholders. If distributing checks to larger landholders was easier or more rewarding on average for MAOs, then it might be easier on the margin as well. On the other hand, the availability of a large pool of smaller landholders who had not encashed checks might mechanically imply greater scope for improvement among them.

In the data, treatment had positive effects across the landholding distribution, but with different effects for different groups. The middle panel of Table 2 suggests that the effect on overall encashment rates was strongest for the smallest landholders (many of whom never encashed their checks in the control group), while the effect on time to encash was strongest for the largest landholders (most of whom did eventually encash their checks in the control group). However, the differences in outcomes across quartiles are not statistically significant at conventional levels. One way to summarize and compare these disparate effects is to compare the estimated total benefits by group, using methods we define below (Section 4.3). When we do this we estimate that the intervention increased value received by 0.088% for landholders below the median and 0.078% for those above the median, suggesting the overall incidence of benefits was mildly progressive.

One specific concern about measurement by phone is that it could skew MAO effort towards those who own phones or have listed phone numbers. MAOs had access to the land registry and so could potentially have targeted their efforts towards farmers with listed numbers. We test this in bottom panel of Table 2. The estimated impact on encashment rates is slightly larger for those with phones (1.3% v.s. 0.7%), but the difference is not significant ($p = 0.68$); a 95% confidence interval for the difference is $[-1.8\%, 2.6\%]$. If anything, time to encashment seems to have improved more for the group without a phone ($p = .05$). Thus, an important finding is that despite MAO performance being measured based only on the population with mobile phones, the resulting increase in MAO effort appears to have improved program performance for all beneficiaries (regardless of phone access).

4.3 Tallying costs and benefits

We next examine whether the intervention was a cost-effective means of delivering money to farmers, and what the overall welfare consequences were.

We cost the intervention at the price the Government of Telangana contracted to pay its vendor to operate the call center, pro-rated for the fraction of calls made to treatment areas, which was Rs. 2.5 million (\$36,000). This is a conservative estimate, as the government paid a premium to run its tender process quickly rather than wait for additional bids. Based on our conversations with the vendor we estimate that the call center could be staffed and run for roughly half this cost. On the other hand, this figure does not include time spent by government employees or members of the research team designing the intervention (e.g. sampling protocols), as these costs were relatively small and are sunk costs from the perspective of a continuation decision.¹⁶

At this cost we estimate that phone-based monitoring was a highly cost-effective means of delivering cash transfers. The estimated impact on money ultimately delivered to farmers was Rs. 69 million, or roughly \$1 million, and would have led to \$8.6 million more being delivered to farmers if the monitoring were extended to the entire state. The cost per incremental rupee delivered was 3.6%, gross of any banking fees, a small number in India where major social protection schemes often spend an order of magnitude more (Niehaus and Sukhtankar, 2013).

To estimate an overall cost-benefit ratio we must price the value of putting capital in the hands of farmers during the planting season as opposed to leaving it on the government's books. We assume that farmers who do not receive the transfer finance the purchase of inputs at rate r_f , either by borrowing or by substituting capital from alternative uses. Capital held by the government earns a lower return r_g . Time runs from the start of the program ($t = 0$) to the last date T on which farmers' higher-yielding investments pay off. The total value of a unit of capital held by the government until time t and then by the farmer from time t until T is thus

$$v(t) = e^{r_g t} e^{r_f (T-t)} \quad (2)$$

Given a distribution F of check encashment dates, the total social value created is

$$W(F) = \int v(t) dF(t) \quad (3)$$

¹⁶We do not attempt to estimate the costs of incremental effort exerted by MAOs to increase encashment rates. In principle, measuring these would require structural estimation of the disutility of effort. In practice it seems reasonable to expect they are small. For example, if 122 MAOs earning \$1,000 a month worked an extra half day then this is equivalent to roughly \$2,000 worth of their time.

In this framework, higher rates of distribution and faster distribution create social benefit by increasing the amount of capital earning the higher rate r_f at any point in time. If F_c is the distribution of dollar encashment dates in the control group and F_m the distribution in the monitoring group then the total value of treatment is $W(F_m) - W(F_c)$. We calculate this quantity directly using administrative records and conduct hypothesis testing using randomization inference.

We calibrate parameters as follows. We value capital on the government’s books at the rate of interest it earns on deposits, or $r_g = 5\%$ annually, and value capital held by farmers at $r_f = 25\%$, which is the going annual rate for short-term farm loans,¹⁷ for a differential return of $\delta = 20\%$. We also examine sensitivity to $\delta \in [5\%, 25\%]$. We interpret the end date T conservatively as the end of the kharif planting and fertilizing season, setting it to 4 months (when our encashment data end) and examine sensitivity to end dates between 0 and 5 months. If farmers paid the higher interest rate for a longer period (such as till the end of harvesting and sale of crops) this would increase our estimate of the benefits.

Using our preferred parameter estimates, phone-based monitoring generated Rs. 10.6M (\$140,000) in benefits, or roughly four times the cost. We reject the null of no benefit ($p = 0.04$). The result is also reasonably robust to plausible variation in T and δ . Estimated benefits are 0 by definition at $T = 0$ and then increase steadily as we increase T , as implied by equation 2, and at $\delta = 20\%$ exceed the costs of the intervention for any T after 26 days (Figure 3). At $T = 4$ months, benefits exceed costs for any $\delta \in [5\%, 25\%]$. Overall, even under conservative parameter assumptions the intervention was cost-effective.

These calculations are further conservative in the sense that they are based on intent-to-treat estimates. Awareness of the intervention in the treatment group was incomplete. Only 90% of treatment MAOs attended the video conference that explained the intervention; consistent with this, only 90% of treatment MAOs in our MAO survey reported that they had heard of the intervention. Moreover, only 28% were sure that the initiative had rolled out in their area, while 28% were unsure and 35% thought it had not. Some of this may reflect strategic misrepresentation, for example if MAOs believed they could excuse poor results by feigning ignorance. But at a minimum, it implies that we should interpret the ITT results as lower bounds on the effect of treating all MAOs. In the control group, on the other hand, while 52% of MAOs had heard about the intervention (likely through peers in treated areas or media mentions of the pilot of call-center based measurement), only 4% believed themselves treated, with another 8% unsure. This suggests that the control group was relatively “uncontaminated” by misperceptions of being treated.

¹⁷This is conservative and uses interest rates charged by registered micro-finance organizations. Informal loans from local moneylenders would carry a much higher interest rate.

Overall, the benefit-cost performance we estimate here reflects the fact that phone-based monitoring can cheaply be applied to very large-ticket programs (in this case, 7% of the state budget). Consequently, even modest improvements in performance rates can create substantial economic value.

4.4 Comparing call center with administrative records

The fact that phone-based monitoring affected their behavior suggests that MAOs expected it to at least partially reflect their true performance. This incentive could become stronger or weaker in the longer run as they learn more about its accuracy. For example, if they learn it is more accurate than they expected, the incentive effect will strengthen. Anticipating this, this section examines how accurate phone-based performance ratings were by comparing them to the administrative records.

At the farmer level, phone call and administrative data on encashment agree in 88.6% of cases, suggesting that phone data is relatively accurate (Columns 4 and 5 of Table A.3). However, the key for performance management is whether the two data sources evaluate MAO level service delivery similarly. To quantify this, we examine the relative performance of each pair (m, m') of MAOs within a district, calculating whether the relative ranking of check encashment levels in the call center data agrees with the ranking in the administrative data.¹⁸ Some disagreement in MAO rankings will result mechanically from sampling variability: one MAO may perform better than another in the population, but worse in a selected sub-sample. To benchmark disagreement due to sampling variation, we first calculate the typical agreement rate for two representative samples of the administrative data that are the same size as our phone samples. The mean agreement rate in 1,000 Monte Carlo simulation is 78%, meaning that we would expect to see disagreement in 22% of cases due purely to sampling variation.¹⁹

Comparing the rankings from the full administrative data to our call center data, the agreement rate is 69%. On net, call center data thus yield agreement rates that are 9% lower than a representative sample of the administrative records themselves (row 1 of 3). This suggests that they can be a useful guide for managers deciding whom to push for more effort or reward for good performance. On the other hand, measures based on the call center may not be precise enough to form the basis for serious administrative actions such as suspension. For these actions, the relevant parameter is not overall agreement, but

¹⁸For example, if the call center data rates MAO A as 3rd and MAO B as 4th best, while the administrative data rates them as 2nd and 3rd best respectively, they agree. If the administrative data instead ranked them as 3rd and 2nd best respectively, they disagree.

¹⁹The mean agreement rate using only the sample of the administrative data that we reached by phone is virtually the same.

whether the data sources are agreement on which MAOs are the worst performers. Among the MAOs ranked among the bottom 20% in the phone data, 80% (47%) are among the bottom 50% (20%) of MAOs in the administrative data; this is informative and likely to be useful for identifying MAOs for checking in on, but arguably not sufficient grounds on its own for administrative penalties.²⁰

5 Conclusion

Overall, we find encouraging evidence that a cheap, simple, flexible approach to monitoring beneficiary experiences of government benefits/services using telephone calls can be a cost-effective tool for improving last-mile service delivery. Our results complement recent work on specialized approaches to monitoring in developing countries that measure a unitary outcome such as showing up to work. An advantage of phone based monitoring is its applicability in a broader range of settings. It can measure any outcome that a person can describe, including some that are currently difficult or impossible to measure automatically such as demands for bribes, police interactions with citizenry, or delivery of publicly provided goods or services. It does not require separate infrastructure to measure separate outcomes, lowering costs. Finally, it can capture relatively open-ended feedback useful for identifying problems that the system designer did not anticipate, and then adapt to capture new types of information.

While the approach we studied here is itself easily adaptable to other settings and programs, this does not mean that its effects will be the same. It would be useful to test phone-based monitoring in a broader array of settings, particularly those where bureaucrats take repeated, rather than one time actions, so that the system can be incorporated to provide feedback of a more continuous nature. The current project only explored one dimension of high frequency monitoring, the monitoring effect. Providing a flow of real-time information to supervisors, it has the potential to improve performance by changing how supervisors manage. Over time it could also provide inputs for improving personnel management, which has been identified as the most important component of organizational management quality, and also shown to be systematically worse for public organizations (Bloom and Van Reenen, 2010). It would also be useful to test this approach in the context of a program with more serious baseline implementation problems than the RBS we study here, as the scope for improvement may be even greater.

²⁰Over time, a scaled up high-frequency monitoring system could evolve statistical protocols for different types of follow up action after including parameters for the cost of different kinds of Type I and Type II errors.

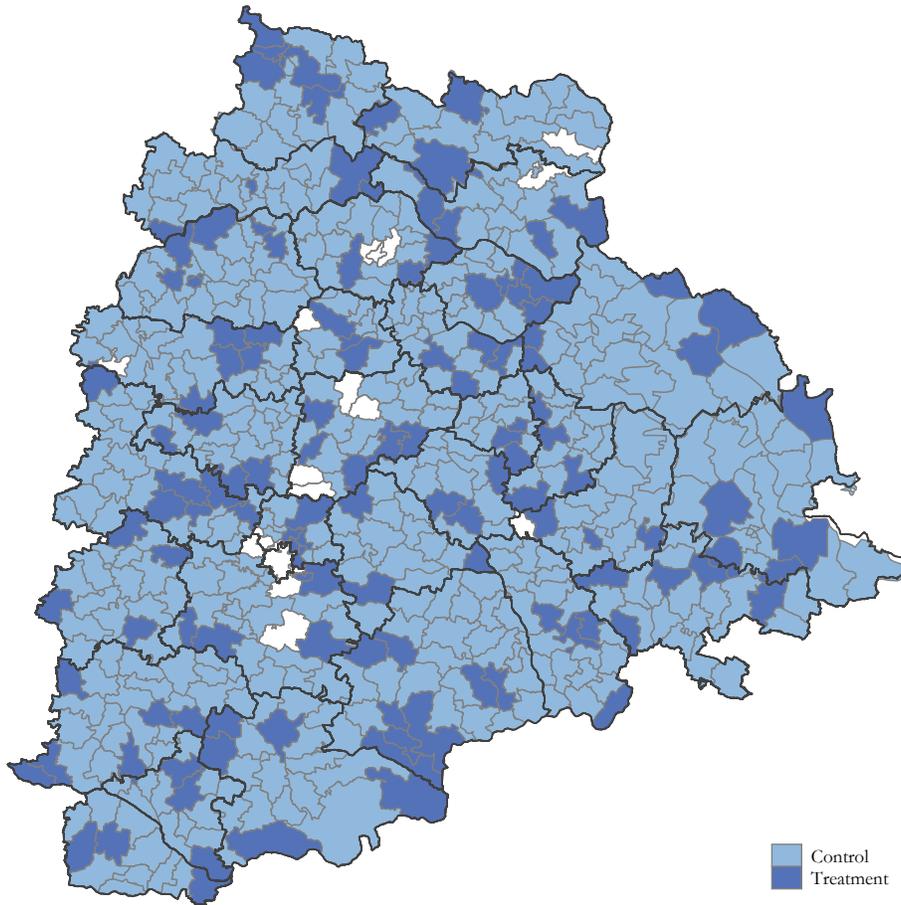
References

- Banerjee, Abhijit V., Rachel Glennerster, and Esther Duflo**, “Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System,” *Journal of the European Economic Association*, 2008, 6 (2-3), 487–500.
- , **Rukmini Banerjee, Esther Duflo, Rachel Glennerster, and Stuti Khemani**, “Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India,” *American Economic Journal: Economic Policy*, 2010, 2 (1), 1–30.
- Bloom, Nicholas and John Van Reenen**, “Measuring and Explaining Management Practices Across Firms and Countries,” *Quarterly Journal of Economics*, 2007, 122 (4), 1351–1408.
- and – , “Why Do Management Practices Differ across Firms and Countries?,” *Journal of Economic Perspectives*, 2010, 24 (1), 203–224.
- Callen, Michael, Saad Gulzar, Ali Hasanain, Muhammad Yasir Khan, and Arman Rezaee**, “Data and Policy Decisions: Experimental Evidence from Pakistan,” Technical Report, Stanford Institute of Economic Policy Research (SIEPR) 2018.
- Duflo, Esther, Rema Hanna, and Stephen P. Ryan**, “Incentives Work: Getting Teachers to Come to School,” *American Economic Review*, 2012, 102 (4), 1241–1278.
- Government of Telangana**, “Telangana Socio Economic Outlook 2017,” Technical Report, Planning Department, Government of Telangana 2016.
- Mookherjee, Dilip**, “Political Decentralization,” *Annual Review of Economics*, 2015, 7, 231–249.
- Mundle, Sudipto, Samik Chowdhury, and Satadru Sikdar**, “Governance Performance of Indian States 2001-02 and 2011-12,” Technical Report, National Institute of Public Finance and Policy 2016.
- Muralidharan, Karthik and Paul Niehaus**, “Experimentation at Scale,” *Journal of Economic Perspectives*, 2017, 31 (4), 103–124.
- , **Jishnu Das, Alaka Holla, and Aakash Mohpal**, “The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India,” *Journal of Public Economics*, 2017, 145, 116–135.
- , **Paul Niehaus, and Sandip Sukhtankar**, “Building State Capacity: Evidence from Biometric Smartcards in India,” *American Economic Review*, 2016, 106 (10), 2895–2929.
- Niehaus, Paul and Sandip Sukhtankar**, “Corruption dynamics: The golden goose effect,” *American Economic Journal: Economic Policy*, 2013, 5 (4), 230–69.
- Olken, Benjamin A.**, “Monitoring Corruption: Evidence from a Field Experiment in Indonesia,” *Journal of Political Economy*, 2007, 115 (2), 200–249.

Pritchett, Lant, “Is India a Flailing State?: Detours on the Four Lane Highway to Modernization,” Technical Report, HKS Faculty Research Working Paper Series RWP09-013, John F. Kennedy School of Government, Harvard University 2009.

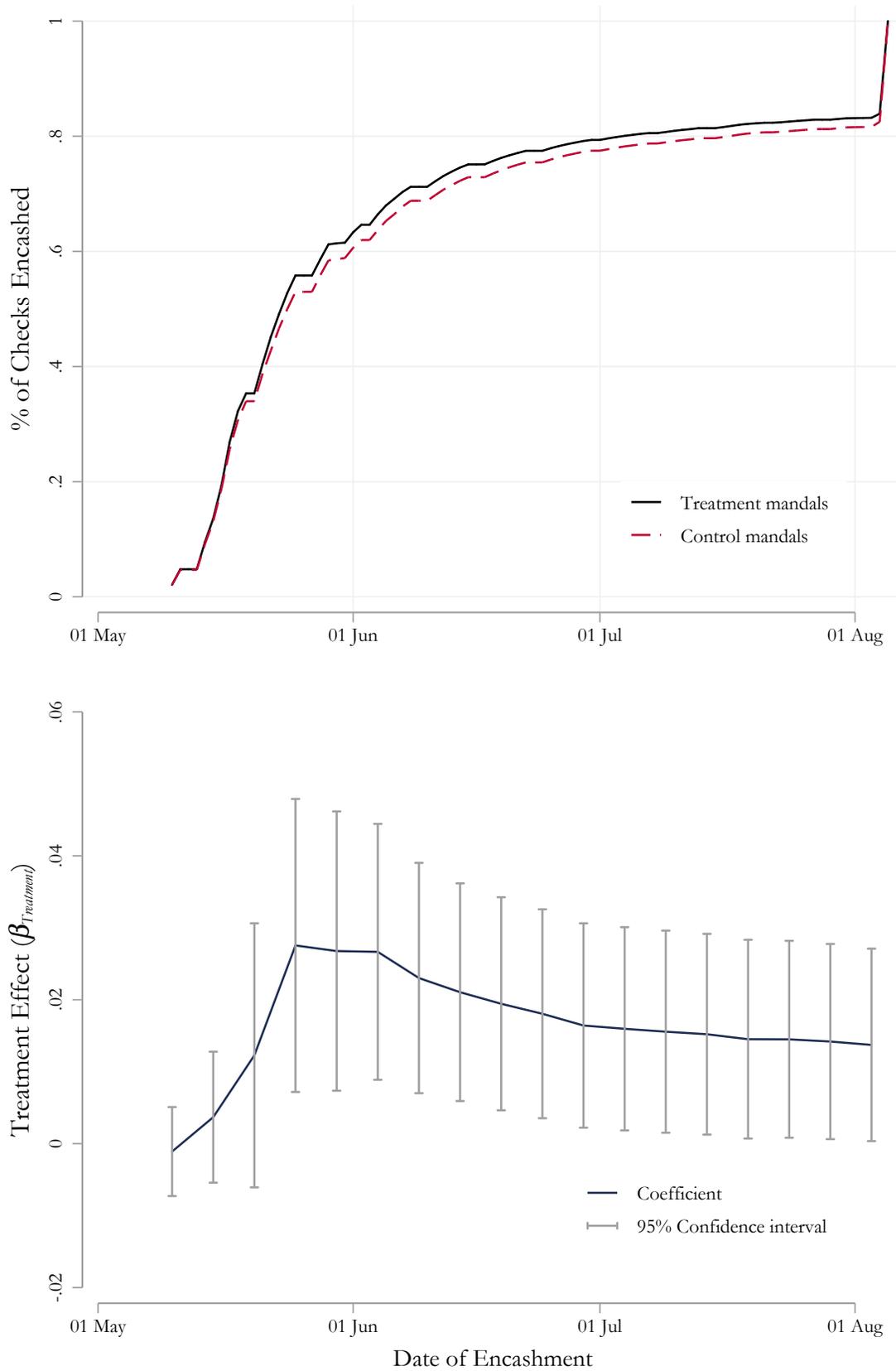
Subramanian, Arvind, “QUBI can wipe off farmers’ tears,” *The Hindu Business Line*, July 2018.

Figure 1: Study areas with treatment and control mandals



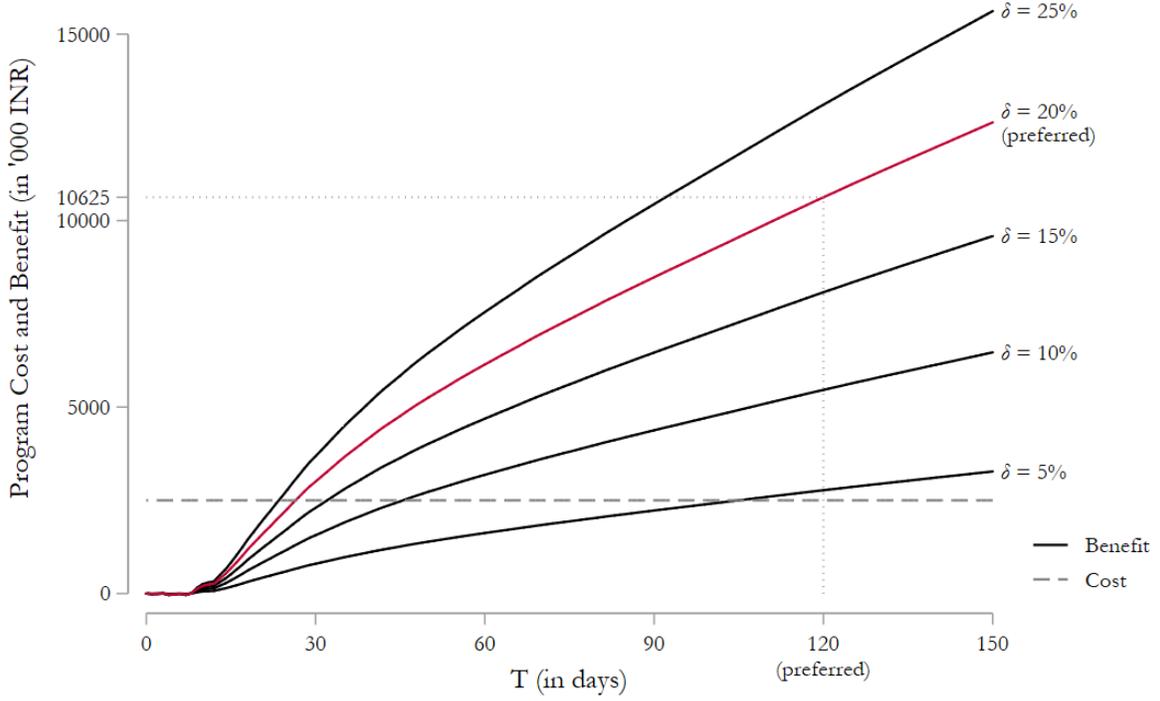
This map shows the geographical distribution of treatment and control mandals (sub-districts) across the entire state. Dark black lines indicate district boundaries, whereas gray lines are mandal boundaries. Randomization was stratified by district, and occurred at the mandal agricultural officer level. Mandals in white were not included in the randomization and study. This typically occurred because the mandal is urban, such as those around Hyderabad, or did not have an MAO assigned to it, so it was not possible to implement the treatment. Note that since there are 10 cases where a treatment MAO oversees multiple geographically contiguous mandals, there is slightly more geographical clustering of treatment mandals than would occur due to chance.

Figure 2: Treatment effect, by date



The two graphs in this figure report (a). the cumulative rate of encashment in treatment and control mandals by day, and (b). the coefficient of treatment effect on the cumulative rate of encashment over the period of check distribution in our data. The coefficient in the bottom graph are estimated through regressions with fixed effects at the randomization strata level and standard errors clustered at the MAO level.

Figure 3: Sensitivity of cost-effectiveness estimates



Sensitivity of cost-effectiveness estimates tested with respect to the total time period of consideration (T) and the differential rate of return (δ , i.e. $r_f - r_g$). The interest earned by the government (r_g) is 5% annually, and the short-term annual interest rate for farmers (r_f) varies from 10% to 30% annually. The preferred specification for these parameters is $T = 120$ days and $\delta = 20\%$.

Table 1: Balance tests

Variable	(1) Control mean	(2) Treatment mean	(3) Difference (SE)
<i>Land registry data</i>			
Land size (acres)	2.21	2.18	-0.01 (0.05)
Median land size	1.57	1.56	0.00 (0.05)
Land size - 25th percentile	0.65	0.66	0.02 (0.04)
Land size - 75th percentile	2.96	2.93	-0.03 (0.06)
Registered mobile numbers	0.61	0.61	0.01 (0.01)
Farmer population	11345	10935	-249 (389)
<i>Census 2011 data</i>			
Literacy rate	0.60	0.60	-0.00 (0.01)
Share of rural population	0.86	0.85	0.01 (0.02)
Share of working population	0.51	0.51	0.01 (0.00)
Share of SC population	0.18	0.18	-0.00 (0.00)
Share of ST population	0.13	0.14	0.02* (0.01)
Observations	4,299,904	1,312,199	5,612,104

Differences in column (3) are estimated through regressions on a treatment indicator, with fixed effects at the randomization strata level. Standard errors are clustered at the MAO level and reported in parentheses.
 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Effect on encashment outcomes

	Ever encashed			Days till encashed		
	(1) Treatment	(2) Control	(3) Obs.	(4) Treatment	(5) Control	(6) Obs.
	mean			mean		
Overall	0.0126* (0.00655)	0.83	5,645,965	-0.759* (0.388)	20.16	4,663,586
<i>Land quartiles</i>						
Quartile 1	0.0224** (0.00932)	0.68	1,449,482	-0.655 (0.511)	23.99	984,251
Quartile 2	0.0145*** (0.00631)	0.85	1,460,294	-0.676* (0.383)	20.08	1,239,604
Quartile 3	0.0113* (0.00601)	0.88	1,443,788	-0.842** (0.359)	18.71	1,278,096
Quartile 4	0.00699 (0.00621)	0.89	1,443,836	-0.982*** (0.367)	18.79	1,284,734
Test of $H_o :$ $\beta_{Q1} = \beta_{Q2} =$ $\beta_{Q3} = \beta_{Q4}$	1.72 (0.16)			0.80 (0.50)		
<i>Phone coverage</i>						
No listed phone	0.00691 (0.0116)	0.72	2,254,142	-1.295*** (0.475)	22.14	1,614,180
Listed phone	0.0128** (0.00554)	0.90	3,543,258	-0.475 (0.396)	19.13	3,172,505
Test of $H_o :$ $\beta_{No-Phone} =$ β_{Phone}	0.17 (0.68)			3.75* (0.05)		

All specifications include fixed effects at the randomization strata fixed level. Standard errors are clustered at the MAO level and reported in parentheses. The bottom row of each panel reports the F-statistic and p-value from a test of the null that coefficients are statistically similar across categories. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Agreement between phone and administrative data on MAO performance

	Actual agreement rate	Agreement rate			Residual disagree- ment rate
	(1)	(2)	(3)	(4)	(5)
		between full and finite sample of AD	between AD and PD among those called	from sampling variation	
Pair-wise order of rankings	68.6%	73.0%	77.7%	77.6%	9.0%
Bottom 20% in PD found in bottom 20% of AD	43.0%	54.0%	58.0%	61.7%	18.7%
Bottom 20% in PD found in bottom 50% of AD	83.0%	88.0%	92.0%	92.7%	9.7%

AD (Administrative Data). PD (Phone Data). The actual rate of agreement between phone and administrative data is reported in (1). In (2), the same comparisons are made between the entire population of administrative data and a smaller finite sample of administrative data, matching the size of the phone call sample. In (3), the same comparisons are made with rankings constructed using phone and administrative data, but restricting administrative data sample to those farmers who were called by phone. Lastly, the same comparison is made between the entire population of administrative data and 1,000 Monte Carlo simulations of farmers sampled from the population, matching the size of the phone call sample. The mean of 1,000 agreement rates is reported in (4). The residual disagreement rate after accounting for tests in (2)-(4) is reported in (5).

Appendix A

Table A.1: Effect on distribution outcomes

	Ever distributed			Days till distributed		
	(1) Treatment	(2) Control mean	(3) Obs.	(4) Treatment	(5) Control mean	(6) Obs.
Overall	0.00793* (0.00468)	0.87	5,645,965	-0.128 (0.309)	11.70	4,930,356
<i>Land quartiles</i>						
Quartile 1	0.0165* (0.00878)	0.74	1,449,482	-0.220 (0.386)	13.55	1,082,824
Quartile 2	0.00910** (0.00417)	0.89	1,460,294	-0.0543 (0.312)	11.53	1,302,380
Quartile 3	0.00654** (0.00319)	0.92	1,443,788	-0.104 (0.299)	18.71	1,334,261
Quartile 4	0.00371 (0.00334)	0.93	1,443,836	-0.232 (0.297)	11.23	1,343,004
Test of H_o :						
$\beta_{Q1} = \beta_{Q2} =$		$1.18 (0.32)$		$0.61 (0.61)$		
$\beta_{Q3} = \beta_{Q4}$						
<i>Phone coverage</i>						
No listed phone	0.00673 (0.0104)	0.76	2,254,142	-0.128 (0.403)	13.85	1,729,723
Listed phone	0.00498* (0.00269)	0.94	3,543,258	-0.0826 (0.286)	10.57	3,332,746
Test of H_o :						
$\beta_{No-Phone} =$		$0.08 (0.78)$		$0.05 (0.83)$		
β_{Phone}						

All specifications include fixed effects at the randomization strata fixed level. Standard errors are clustered at the MAO level and reported in parentheses. The bottom row of each panel reports the F-statistic and p-value from a test of the null that coefficients are statistically similar across categories. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: Impact on beneficiary experience

	(1) Correct amount on check	(2) Received at Gram Sabha	(3) Asked to pay bribe	(4) Satisfied with scheme
<i>Panel A: Original</i>				
Treatment	-0.000916 (0.00787)	0.00391 (0.00380)	-0.0000828 (0.00190)	0.00603 (0.00373)
Control Mean	0.85	0.94	0.02	0.93
<i>Panel B: Re-weighted</i>				
Treatment	-0.00907 (0.00854)	0.00759* (0.00457)	0.00108 (0.00230)	0.00232 (0.00359)
Control Mean	0.86	0.94	0.02	0.93
Observations	19,834	19,890	19,830	22,329

Outcomes in header. (Inversed) sampling probability used to construct weights in Panel B. All specifications include randomization strata fixed effects. Standard errors are clustered at the MAO level and in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: Impacts across data sources

	Administrative data				Phone data
	(1) All	(2) With phones	(3) With phones, sampled	(4) With phones, reached	(5) Reached
<i>Panel A: Distribution status</i>					
Treatment	0.801	0.874	0.873	0.878	0.885
Control	0.793	0.870	0.876	0.884	0.880
Difference	0.00901 (0.00744)	0.00614 (0.00609)	0.00384 (0.00669)	0.00326 (0.00719)	0.00389 (0.00644)
<i>Panel B: Encashment status</i>					
Treatment	0.657	0.727	0.731	0.743	0.757
Control	0.630	0.700	0.711	0.732	0.754
Difference	0.0254*** (0.00912)	0.0229** (0.0101)	0.0221* (0.0115)	0.0128 (0.0115)	0.00204 (0.0102)
Observations	5,536,538	3,356,249	44,690	21,835	21,835

Outcome variables reflect distribution and encashment status as of the date the call was made to the respondent, which is why the means differ and the effects are larger than in the main specifications. For respondents who were not called in the survey, the median date of calls made to their district is used as the cut-off date. All specifications include fixed effects at the randomization strata levels. Standard errors are clustered at MAO level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.