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, Revised January 2019

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. We also thank Michael Callen, Markus Goldstein, and several seminar participants for helpful comments. 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.

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, Revised January 2019
JEL No. C93,D73,H53,O33

ABSTRACT

Improving “last-mile” public service delivery is a recurring challenge in developing countries. Could the widespread adoption of mobile phones provide a simple, cost-effective means for improvement?We use an at-scale experiment to evaluate the impact of a phone-based monitoring system on a program that transferred nearly a billion dollars to 5.7 million Indian farmers. In selected jurisdictions, officials were informed that program implementation would be measured via calls with beneficiaries. This led to a 3.9% increase in farmers receiving transfers on time, and a 1.5% increase overall. The program was highly cost-effective, costing 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
srs8yk@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). Improving 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; Duflo et al., 2012) to ensuring that beneficiaries receive food or money they are entitled to (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 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. Existing mechanisms for filing complaints about public service delivery are little-used or generate a non-representative picture of results for the most and least satisfied beneficiaries. 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 or infrequent to be of 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 (World Bank, 2018). 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 locations, programs and outcomes.

We examine whether phone-based measurement can improve service delivery in the context of a high-stakes government initiative in India. Between May and July of 2018, the government of the state of Telangana (GoTS) distributed \$0.9B, or around 3.5% of the state's annual budget, as lump-sum payments to farmers. Responsibility for implementing the scheme rested primarily with Mandal (sub-district) Agricultural Officers (MAOs), who managed the distribution of 5.7 million physical checks to farmers. This money was meant to finance investments in seed and fertilizer, and reduce farmer debt. Thus GoTS placed a high priority on distributing checks prior to planting in June.

[1] In India, for example, a state government nearly shut down a highly-effective reform because officials (whose rents were threatened) reported cherry-picked negative anecdotes (Muralidharan et al., 2016).

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 check 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 between treated MAOs and 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 spillovers (Muralidharan and Niehaus, 2017).

The call center completed 22,565 outbound calls with farmers in two weeks during the peak of program implementation, and data from these calls were used to create reports on MAO's absolute and relative performance. Because the program was implemented quickly, MAOs saw these performance reports only after most of their work on it was completed. The intervention thus affected their awareness that their performance was being measured, but did not provide them with usable information about where to focus their efforts. In this sense we view the results as a lower bound on the potential effects of the approach. In future iterations, information collected by phone could be used by frontline workers to improve effort allocation, and as the basis for explicit rewards or sanctions (which were not implemented here).

We use farmer-level bank administrative records of whether and when these checks were encashed as a reliable measure of MAO performance. Phone-based monitoring significantly improved the likelihood of farmers ever receiving their transfer, as well as receiving it on time. On-time delivery of transfers was 3.9% higher in treatment areas (2.3 percentage points higher on a base of 69% in control areas), and the likelihood of checks ever being delivered was 1.5% higher (1.3 percentage points higher on a base of 83% in control areas). These effects correspond to a $3.9 million increase in transfers that were delivered on-time and a $1 million increase in amount ever delivered.[2] If phone-based monitoring were extended to the entire state, it would yield $33.1 million more delivered on time and $8.6 million more ever delivered annually. The treatment-control gap does not change after performance reports were issued, consistent with the idea that the results reflect incentive rather than information effects.

[2]Among larger landowners, the treatment effect is larger for "on-time" delivery than "ever delivered". Since they received larger transfers, the impact on on-time delivery relative to ever delivered is bigger for *value* of funds than for *fraction* of checks.

We also find that the incidence of the intervention was mildly progressive, although the 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 price paid to the vendor who ran the call center, we estimate that the incremental cost per additional dollar of benefits delivered to beneficiaries was 3.6 cents, which is lower than the cost of almost any anti-poverty program for which such data is available. Turning to "on-time" delivery, the cost per dollar of benefits delivered on time was less than one cent. To calculate economic returns to the program, we define benefits as the difference between the estimated return on capital held by farmers rather than the government. Even under relatively conservative assumptions, we estimate a benefit of four times the cost.

Our paper complements recent work testing more specialized approaches such as monitoring worker attendance with time clocks Banerjee et al. (2010a) or with custom smart-phone applications Callen et al. (2018). 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, and (ii) the flexibility to scale across an unusually wide number range of places, programs and outcomes, and (iii) scope to adapt quickly as challenges and circumstances on the ground change. We also complement recent work by Aker and Ksoll (2018), who test a phone-based monitoring pilot implemented by an NGO in an adult education program in 134 villages in Niger and find significant learning gains.

Our results show that phone-based monitoring can be implemented by governments at scale, and deliver significant improvements in service delivery across millions of beneficiaries quickly and cost effectively. They highlight the potential of using a simple, widely-used, generic technology to monitor large-scale programs, where even relatively modest improvements can create substantial value. The speed and scalability of this approach are well-illustrated by this project, which moved from an in-principle agreement with GoTS to implementation at scale within one month, and impact at scale within two months.

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; Banerjee et al., 2010b). 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 a sufficiently spatially disaggregated level to hold appropriate officials 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.[3]

2 Setting and intervention

Telangana is India's newest state, created in 2014 from Andhra Pradesh. It has a population of 35 million, with around 60% living in rural areas. It is relatively well-off, with per capita income 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 ("Friend of the Farmer") Scheme (RBS) in May of 2018 to provide capital for the purchase of agricultural inputs such as seeds and fertilizer prior to the main agricultural season. The RBS was hailed by economists as a more efficient response to widespread farmer economic distress than common alternatives such as raising procurement prices or waiving loans (Subramanian, 2018). It authorized payment of Rs. 4,000 ($55) per acre to every farmer registered as a landholder in the government's land registry, which had been updated and digitized in 2017.

GoTS made the transfer through "order checks", which could be exchanged for cash at any branch of the bank listed on the check, whether or not the beneficiary held an account there (conditional on providing official ID matching the name 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. The Department of Agriculture managed the distribution of checks, with MAOs responsible for their respective mandals. MAOs supervised teams of agricultural extension workers, who held meetings in each village of the mandal to deliver checks to the farmers living there.

Implementing RBS well was a priority for the government given the sum disbursed (approximately $0.9B per cropping season or $1.8B annually – accounting for 7% of the annual

[3]The cost of measuring performance is also an important determinant of the optimal level of decentralization (Mookherjee, 2015), so lowering this cost may have further organizational implications.

state budget), 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 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, 3.5 million (61%) of the 5.7 million entries in the registry listed a contact number. GoTS contracted a call center to collect data from beneficiaries between 29 May and 15 June.[4] The call center attempted to reach a random sample of 46,007 farmers representative of those with listed phone numbers in the GoTS administrative records.[5] It successfully completed calls with 22,565 (49%) of these farmers.[6] 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 checks and calls to beneficiaries, the Telangana Department of Agriculture informed treatment MAOs that their mandals had been selected by lottery to take part in a pilot of the phone-based monitoring system. During a special video conference with the treatment MAOs, the state Commissioner of Agriculture explained the initiative and the data that would be collected. He informed them 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) between 9 and 13 July. The reports listed five metrics: the proportion of farmers who reported receiving their check, receiving it before 20 May (to measure speed of delivery), successfully encashing it at the bank, being asked for a bribe,

[4] The intervention was designed by the research team, but implemented by GoTS.

[5] The sample included approximately 150 farmers per treatment and 50 per control mandal. See pre-analysis plan for details.

[6] The vendor also piloted automated calls (IVR), but these had a high error rate in capturing responses, so were discontinued.

and being satisfied with the program overall. They showed performance on these metrics for the mandal in question, relative to other mandals within the same district, and relative to the state overall. They also showed a simple, color-coded categorical rating ("Poor," "Fair," "Good," or "Excellent") based on absolute performance, motivated in part by the finding of Callen et al. (2018) that "flagging" of high or low performers can make performance data more accessible. 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 due to the program's compressed time-frame. The program aimed to distribute all checks between early May and mid-June, whereas reports were issued in early July. Our estimates thus reflect the impact of MAOs knowing they were being monitored, but not using information from the calls to do their jobs better.

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

MAOs and their staff could potentially react to monitoring in several ways. They could improve processes to ensure that checks were distributed to all eligible beneficiaries. They could work harder to find recipients – both before the village meetings by publicizing them more thoroughly, or after them, by following up with those who did not attend. They could also demand fewer bribes.[7]

That said, one might reasonably expect phone-based monitoring to have limited effects in this setting, as government scrutiny of RBS implementation was already high: MAOs recorded check distribution, banks reported check encashment in order to claim reimbursement, and MAOs were broadly aware that data of this sort were being recorded. The availability of high-quality administrative data on outcomes makes the RBS an unusually low-cost setting in which to measure effects of phone-based monitoring, but also means those effects could be lower than in other settings where phone data are the *only* performance information available. Our estimates should thus be interpreted as the effects of adding an incremental, independent source of monitoring, and making this salient.

[7]MAOs were only responsible for check distribution. A different government department (Revenue) printed checks *before* our intervention, after verifying farmer eligibility, and banks independently checked farmer identity before cashing the checks. Improvements in benefit receipt are thus unlikely to have come at the cost of lower scrutiny of eligibility requirements.

3 Research methods

Our design and methods follow a registered pre-analysis plan.[8]

3.1 Experimental design

The study population consists of nearly all households eligible to receive RBS, i.e. all landholding households in Telangana. We excluded one largely urban district (Hyderabad) as it had very few program beneficiaries, leaving 30 remaining districts.

Within these districts, we randomly assigned treatment at the level of the MAO (who occasionally oversees multiple mandals). We randomly selected approximately 25% of MAOs for treatment, yielding a total of 122 treatment MAOs and 376 control MAOs. This corresponded to 132 treatment 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 (further details of the randomization algorithm are in the pre-analysis plan). Figure 1 shows the geographical distribution of treatment and control mandals.

Table A.1 reports means and balance tests across treatment and control groups on landholder-level characteristics from the landholder registry as well mandal-level characteristics from the 2011 census. Of 11 tests, one (Scheduled Tribe population share) is significant at the 10% level, as we would expect to see by chance. 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, including (i) the register of all agricultural landholders in the state, including names, village, acres held, and a contact phone number; (ii) a farmer-level record of check distribution maintained by the MAOs; and (iii) farmer-level bank records of check encashment. Our analysis focuses on encashment, as getting the money is the ultimate outcome of interest to policy-makers. Bank reports of encashment were recorded in real-time and were the basis for reimbursement from the government; manipulating them would constitute serious fraud and could jeopardize a bank's operating license. We find that they closely match encashment as reported by the surveyed farmers.

We use encashment data at two dates. The first (8 June) captures on-time delivery. This was exactly a month after the start of distribution, and reflects the government's goal of

[8]See `https://www.socialscienceregistry.org/trials/2942`.

ensuring that farmers had funds in place at the start of planting to buy seeds and fertilizers.[9] This was a high priority for GoTS since a key goal of the program was to break the cycle of farmer debt, which was widely believed to be a driver of farmer suicides. The second (26 September) captures if the checks were *ever* encashed. This is after the last date (15 August) on which the checks were valid for encashment and thus should well approximate the final distribution of checks.[10]

We also use data from phone calls conducted by the call center as a secondary source. These 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 other reasons.

Finally, we use data from a short phone survey of MAOs. We surveyed 88 of 122 treatment MAOs and a sample of 54 control MAOs.[11] Surveys covered their awareness of the pilot and beliefs about their treatment status. To minimize potential Hawthorne effects, we conducted these surveys within a small sample of control MAOs after the distribution was mostly complete.[12]

3.3 Estimation

We report intent-to-treat estimates, comparing mean outcomes in treatment and control areas. We discuss MAO beliefs and their implications for interpretation in our cost-benefit analysis below. We thus estimate

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

where y is an outcome, T an indicator for assignment to treatment, and X a vector of pre-specified covariates (in practice, only one variable, the size of landholdings). Indices denote individual i in village v in mandal m in stratum s in district d. Treatment is strictly exogenous conditional on the randomization stratum fixed effects δ_{sd}. We cluster standard errors at the level of treatment assignment (the MAO) and conduct randomization inference as a robustness check. When using call center data, we reweight estimation by the inverse probability of being sampled.

[9] While the optimal planting date depends on monsoon arrival, planting typically takes place in June.

[10] Checks were 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. By 26 September encashment activity had largely ceased.

[11] We attempted surveys with all of the treatment MAOs (72% response rate) and a random sample of 2 control MAOs per district (60% response rate).

[12] On the survey date, 84% of checks that would ever be encashed had already been encashed.

4 Results

4.1 Effects on overall program performance

Overall, RBS implementation was imperfect but fairly successful compared to other similar interventions. Checks were successfully encashed by 4.03 million farmers (69% of target) within the government-targeted 1-month window from the start of the program (Table 1: Column 2). After 5 months, this figure rose to 4.8 million farmers or 83% (Column 4). Corruption was not a major issue, with only 2% of farmers reached by phone reporting that they had to pay a bribe to obtain their checks.[13]

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, while the bottom panel plots regression estimates of the treatment effect by date. The treatment effect peaks at 2.8 percentage points on 25 May ($p = 0.008$) and then narrows somewhat, asymptoting to 1.3 percentage points by 26 September. There is no evidence of a differential change in encashment rates following the distribution of the reports themselves (5 to 9 July). This is not surprising given most encashment had already taken place by that time, and suggests that the results are driven by the incentive effects of MAO's knowing that they were being monitored and anticipating these reports rather than by the information that the reports contained.[14]

Table 1 reports average treatment effects on check encashment. Treatment increased the probability of on-time check encashment by 2.3 percentage points ($p = .004$), and the probability the farmer ever encashed the check by 1.3 percentage points ($p = 0.054$). As seen in Table 2, conditional on ever encashing, treatment lowered the mean number of days that passed before recipients encashed their checks by three-fourths of a day ($p = 0.051$).[15] [16] [17]

[13] We also find slightly higher reported encashment rates in our phone call data than in the corresponding administrative records, suggesting that officials did not collude with banks to encash beneficiary's checks without their knowledge.

[14] In hazard models, an indicator for post 9 July is not a significant predictor of encashment. Results available on request.

[15] For completeness we also report effects on check distribution (Tables A.2 and A.3). We treat these data with caution as they were uploaded by MAOs with substantial lags, causing date of distribution to be mismeasured, and were not subject to penalties for misreporting like those banks faced. See the online appendix for further description of the issues with the MAO data.

[16] In the smaller sample of phone data, treatment had insignificant effects on the likelihood that phone call respondents were asked to pay a bribe or were satisfied, with a slight increase in likelihood of receiving their check at the village meeting when distribution was supposed to occur (Table A.5).

[17] As pre-specified, we test whether these results could be explained by supervisors of MAOs focusing more attention on treatment MAOs. Table A.7 finds no evidence of this.

4.2 Distributional consequences

The baseline allocation of benefits under RBS was regressive, as check size was proportional to registered landholdings. This pattern was exacerbated by differences in distribution and encashment rates. As of 26 September, 89% of farmers in the top quartile of the landholding distribution (holding more than 3.1 acres of land) had encashed their checks, declining monotonically to 68% of farmers in the bottom quartile (holding fewer than 0.4 acres). This 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 0.05 acres of land would receive a check worth just Rs. 200 ($3), possibly less than the time and money costs of encashment.

Turning to distributional effects, we find that the effect on on-time delivery was significant and nearly identical across farmers of different landholding sizes (middle panel of Table 1). By the end of September, the treatment effect continued to be statistically significant for farmers in the lower three quartiles of landholdings, but not for farmers in the top quartile. We reject equality of treatment effects between the top and bottom quartiles ($p = 0.051$), but do not reject a joint test of equality across all four quartiles ($p = 0.13$). We interpret this as poorer farmers having a harder time claiming their checks if they did not initially receive it, allowing the initial gap between treatment and control to persist. Wealthier farmers were eventually able to claim their checks, but the treatment sped up the process and lessened their cost of accessing the transfer (Table 2). We present the full pattern of treatment effects by quartile of land holding over time in Figure A.1.

One concern about measurement by phone is that it could skew MAO effort towards those who own phones or have phone numbers (especially since MAOs had access to the land registry and could see which farmers had numbers listed and thus could be called). However, we find significant positive impacts on on-time encashment for both those with and without phones, and cannot reject that these effects are the same (Table 1: Bottom panel). The difference in ever-encashed between those with and without phones is also not statistically significant ($p = 0.68$), but time to encashment seems to have improved more for the group without a phone ($p = .05$). Thus, despite MAO performance being measured only in the population with mobile phones, the resulting increase in MAO effort appears to have led to improvement in program performance for all beneficiaries.

4.3 Tallying costs and benefits

We next examine cost-effectiveness of the intervention at delivering money to farmers, and its overall welfare consequences.

We cost the intervention at Rs. 2.5 million ($36,000), the price GoTS paid the call-center vendor pro-rated for the proportion of calls made to treatment areas. This is conservative, as the government paid a premium to complete the procurement process quickly; conversations with the vendor indicate that the call center could be operated for roughly half this cost. On the other hand, this figure does not include the (relatively small) sunk costs of time spent by government employees or members of the research team designing the intervention (e.g. sampling protocols).[18]

The estimated impact on money ultimately delivered to farmers was roughly $1 million and on money delivered on time was $3.9 million).[19] The cost per incremental dollar delivered was 3.6 cents, a much smaller administrative cost of delivering benefits than most social protection schemes (Niehaus and Sukhtankar, 2013). Focusing on the government's objective of getting transfers to farmers on time, the cost per dollar delivered on time was less than one cent.

The economic returns to the program depend on the value of capital in the hands of farmers during the planting season as opposed to on the government's books. We assume that farmers who do not receive the transfer finance input purchases by borrowing at rate r_f. Capital held by the government earns a lower return r_g. Time runs from the start of the program ($t = 0$) to the date T on which farmers' investments pay off and debt is repaid. 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

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

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

$$W(F) = \int v(t) dF(t) \tag{3}$$

Faster and broader distribution shifts F (as seen in Figure 2), increasing the amount of capital earning the higher rate r_f. We calculate $W(F)$ for both treatment and control groups using administrative records and conduct hypothesis testing using randomization inference.

We value capital on the government's books at the rate it earns on deposits ($r_g = 5\%$ annu-

[18]We also do not cost incremental MAO effort, which is likely to be small or at least below the wage premium enjoyed by public employees (Finan et al., 2017).

[19]The treatment effects on amount ever delivered and amount delivered on time were Rs. 54 and Rs. 203 per farmer respectively.

ally),[20] and capital held by farmers at the going rate for short-term farm loans ($r_f = 25\%$).[21] We conservatively assume that investments are realized and debt is repaid immediately at harvest, so T equals 4 months.

Using these estimates, phone-based monitoring generated Rs. 10.6M ($140,000) in benefits, or roughly four times its cost. We reject the null of no benefit ($p = 0.04$) using randomization inference. This result is reasonably robust to variation in T and δ. At $\delta = 20\%$, benefits exceed costs for any T longer than 26 days, while at $T = 4$ months benefits exceed costs for any $\delta \in [5\%, 25\%]$ (Figure 3). Even under conservative parameter assumptions, the intervention was cost-effective.

These calculations may also be conservative in the sense that they reflect intent-to-treat estimates, while awareness in the treatment group was incomplete. Among treatment MAOs we surveyed, 90% had heard of the intervention, but only 28% were sure that the initiative had rolled out in their area; 28% were unsure and 35% thought it had not. This may partly reflect strategic misrepresentation, such as if MAOs believed they could excuse poor results by feigning ignorance. In the control group, 52% of MAOs had heard about the intervention, but only 4% believed themselves treated, with another 8% unsure. While the control group was relatively "uncontaminated" by misperceptions of being treated, treatment effects may have been even larger if awareness of phone-based monitoring were universal.

Overall, these benefit-cost estimates suggest that phone-based monitoring can cheaply be applied to large-ticket programs at scale. Consequently, even modest improvements in performance can create substantial economic value.

4.4 Comparing call center with administrative records

The fact that MAOs responded to phone-based monitoring implies that they believed it would at least partially reflect their true performance. We now examine the accuracy of phone data by comparing measured MAO performance in phone call data to the administrative data.

Phone call and administrative data agree on whether a given check was encashed in 88.6% of cases. At an aggregated level, we examine the reliability of phone-based measurement of MAO performance to see if the data can be reasonably used for personnel management. We calculate how often phone and administrative data rank the relative performance of a pair (m, m') of MAOs within a district the same way.[22] These rankings disagree in 31% of

[20] In principle, the government could use funds for other productive investments. In practice funds appropriated for the program would not be reallocated till the next fiscal year and would only earn interest.

[21] This is the rate charged by registered micro-finance organizations; informal moneylenders typically charge more.

[22] For example, suppose the call center rates MAO A as 3rd and MAO B as 4th best. If the administrative data rates them as 2nd and 3rd best respectively then the sources agree; if it rates them as 3rd and 2nd best

cases. However, 22% can be explained by sampling variation, with the underlying rate of disagreement between the two data sources being 9% (see notes to Table A.4 for details on calculation). Finally, for the 20% of MAOs who were ranked as the worst performers in the phone data, 47% are also among the worst 20% of MAOs in the administrative data, while 80% are in the bottom 50%.

Overall, these results suggest that managers could reasonably use phone data to help decide which officials to push for more effort or acknowledge for good performance. However, data reliability may not be high enough to justify using them to determine more serious administrative actions (e.g. suspensions) without data over multiple cycles and years.

5 Conclusion

We find evidence that a cheap, simple, and flexible approach to monitoring beneficiary experiences can be a cost-effective tool for improving last-mile service delivery. While the approach we studied here is itself adaptable to other settings and programs, this does not mean that its effects will be the same. It would therefore be useful to test phone-based monitoring in other settings. For instance, it may perform better for outcomes that beneficiaries experience more directly (e.g. check distribution) than indirectly (e.g. public good maintenance). It would also be useful to test this approach in a setting where the scope for improvement is greater than in the RBS, which was relatively well-implemented.

Similarly, it would be valuable to examine how effects evolve over time in settings where bureaucrats perform similar functions repeatedly. As with all monitoring technologies, the officials being monitored would learn about the consequences of performing at different levels and might develop new strategies – both productive and counterproductive – to influence their ratings. But over time, phone-based monitoring could also inform officials in real-time on what locations are most in need of their targeted intervention, as well as motivating them to increase effort. It could provide inputs for improving personnel management, which has been identified as the most important component of organizational management quality, and is systematically worse for public organizations (Bloom and Reenen, 2010). It could be tuned in many ways to improve performance, evolving statistical protocols for different types of follow-up action reflecting the cost of different kinds of Type I and Type II errors. Optimal monitoring protocols would take into account the need for whistleblower protection in small samples (Chassang and i Miquel, 2018) and the motivations of the respondents answering the phone (Fiorin, 2018). One could even consider making the results publicly available, trading off the costs and benefits of transparency.

respectively then they disagree.

Historically, better measurement has been a foundation for improved productivity in several settings by enabling better coordination, management, and contracting (Landes, 1983; Baker and Hubbard, 2004). High-frequency and low-cost measurement of last-mile service delivery using phone-based monitoring could similarly enable productivity improvements in the delivery of public services.

References

Aker, Jenny C. and Christopher Ksoll, "Call Me Educated: Evidence from a Mobile Monitoring Experiment in Niger," *Center for Global Development Working Paper 406*, 2018.

Baker, George P. and Thomas N. Hubbard, "Contractibility and Asset Ownership: On-Board Computers and Governance in U.S. Trucking," *The Quarterly Journal of Economics*, 2004, *119* (4), 1443–1479.

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*, 2010, *6* (2-3), 487–500.

_ , Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani, "Pitfalls of Participatory Programs: Evidence from a Randomizaed 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," *The Quarterly Journal of Economics*, November 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," *Stanford Institute of Economic Policy Research (SIEPR) Working Paper No. 1022*, 2018.

Chassang, Sylvain and Gerard Padró i Miquel, "Crime, Intimidation, and Whistleblowing: A Theory of Inference from Unverifiable Reports," *Review of Economic Studies*, 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.

Finan, Frederico, Benjamin A Olken, and Rohini Pande, "The Personnel Economics of the Developing State," *Handbook of Economic Field Experiments*, 2017, *2*, 467–514.

Fiorin, Stefano, "Reporting Peers' Misbehavior: Experimental Evidence from Afghanistan," *UCSD Working Paper*, 2018.

Government of Telangana, "Telangana Socio Economic Outlook 2017," *Planning Department, Government of Telangana*, 2016.

Landes, David S., *Revolution in Time: Clocks and the Making of the Modern World*, Harvard University Press, Cambridge, Mass., 1983.

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," *National Institute of Public Finance and Policy Working Paper 16/164*, 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*, January 2017, *145*, 116–135.

_ , Paul Niehaus, and Sandip Sukhtankar, "Building State Capacity: Evidence from Biometric Smartcards in India," *American Economic Review*, October 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.

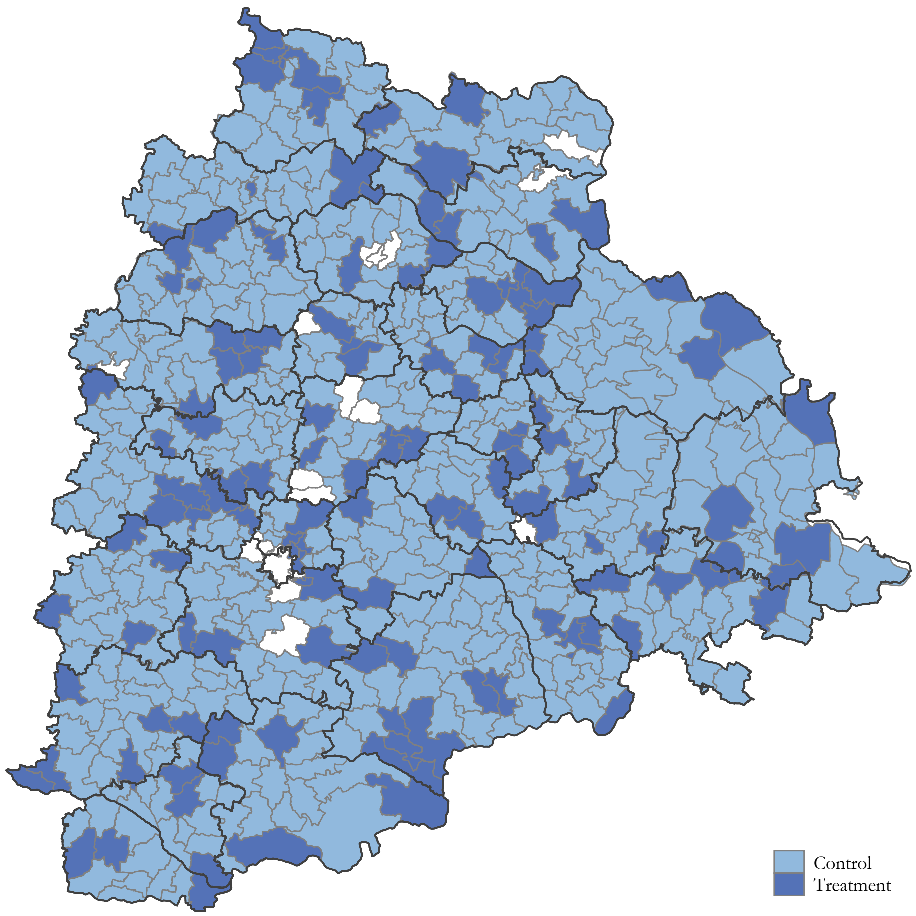
Olken, Benjamin A., "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, April 2007, *115* (2), 200–249.

Pritchett, Lant, "Is India a Flailing State?: Detours on the Four Lane Highway to Modernization," *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.

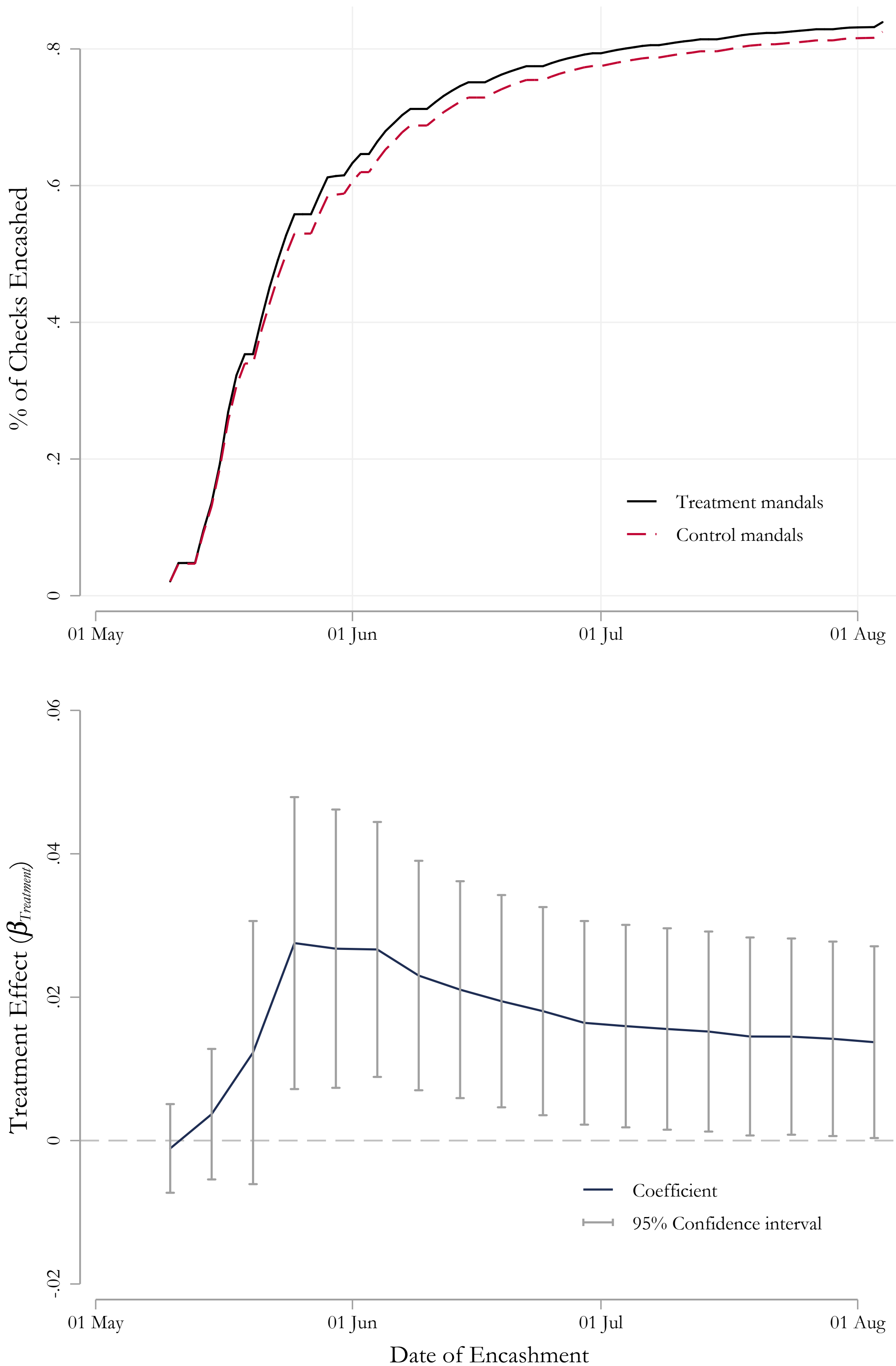
World Bank, "Mobile cellular subscriptions (per 100 people).," `https://data.worldbank.org/indicator/IT.CEL.SETS.P2?end=2017&locations=XM&name_desc=true&start=1998` 2018. Accessed: 2018-09-26.

Figure 1: Study areas with treatment and control mandals



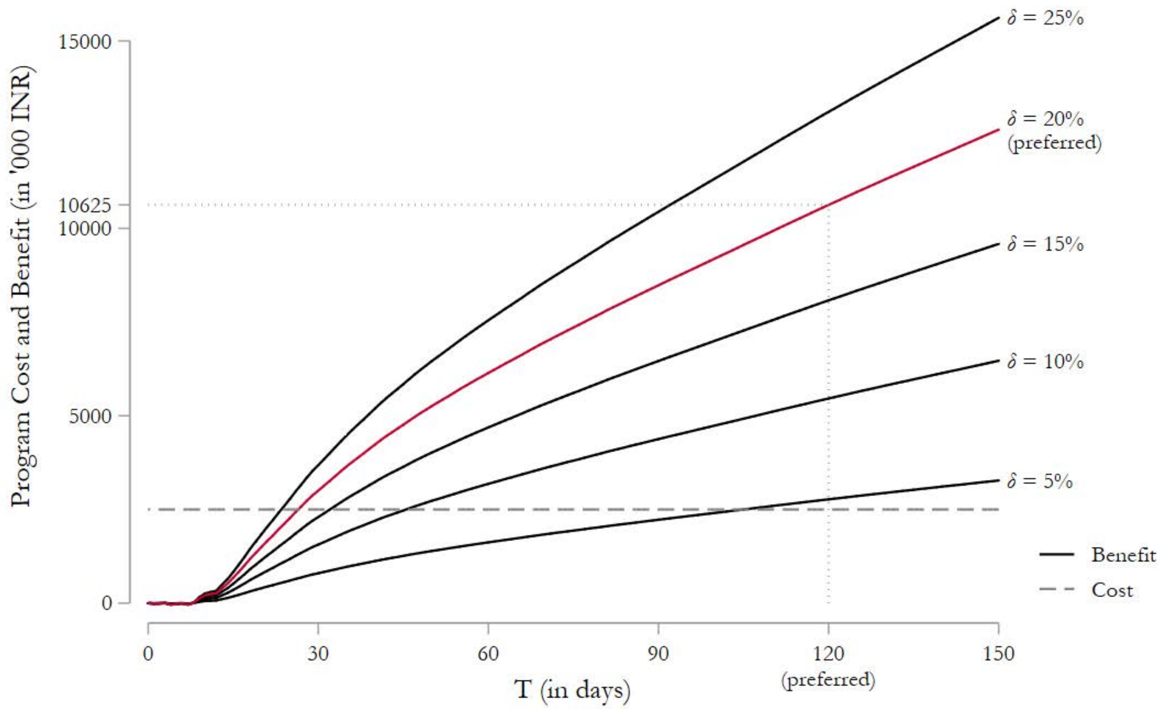
This map shows the geographical distribution of treatment and control mandals (sub-districts) across the entire state. Dark black lines indicate district boundaries, whereas gray lines are mandal boundaries. Randomization was stratified by district, and occurred at the mandal agricultural officer level. 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. Less than 1% of checks were encashed after August 4 or before 10 May, so the axis is restricted to those time periods.

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: Effect on encashment outcomes

	Encashed before June 8th		Ever encashed		
	(1) Treatment	(2) Control mean	(3) Treatment	(4) Control mean	(5) Obs.
Overall	0.0231	0.69	0.0126	0.83	5,645,937
	(0.00807)		(0.00655)		
Land quartiles					
Quartile 1	0.0278	0.52	0.0224	0.68	1,449,482
	(0.00960)		(0.00932)		
Quartile 2	0.0248	0.71	0.0145	0.85	1,460,294
	(0.00791)		(0.00631)		
Quartile 3	0.0241	0.76	0.0113	0.88	1,443,788
	(0.00755)		(0.00601)		
Quartile 4	0.0208	0.77	0.00699	0.89	1,443,836
	(0.00803)		(0.00621)		
Test of H_o : $\beta_{Q1} = \beta_{Q2} = \beta_{Q3} = \beta_{Q4}$	*0.64 (0.59)*		*1.72 (0.16)*		
Phone coverage					
No listed phone	0.0229	0.57	0.00691	0.72	2,254,142
	(0.0116)		(0.116)		
Listed phone	0.0202	0.76	0.0128	0.90	3,543,258
	(0.00821)		(0.00554)		
Test of H_o : $\beta_{No-Phone} = \beta_{Phone}$	*0.04 (0.84)*		*0.17 (0.68)*		

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. Models are estimated using administrative data at the individual 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 (12.5 acres of land) were split into multiple checks. Outcomes are essentially perfectly correlated within individual, as farmers either picked up and encashed all or none of their checks, which is accounted for by clustering at the mandal level. Farmers with less than 0.025 acres of land (less than 1% of the sample) were still issued checks, but in the amount of Rs. 100.

Table 2: Effect on time to encashment

	Days till encashed		
	(1) Treatment	(2) Control mean	(3) Observations
Overall	-0.759	20.16	4,663,678
	(0.388)		
Land quartiles			
Quartile 1	-0.655	23.99	984,251
	(0.511)		
Quartile 2	-0.676	20.08	1,239,604
	(0.383)		
Quartile 3	-0.842	18.71	1,278,096
	(0.359)		
Quartile 4	-0.982	18.79	1,284,734
	(0.367)		
Test of H_o : $\beta_{Q1} = \beta_{Q2} = \beta_{Q3} = \beta_{Q4}$		*0.80 (0.50)*	
Phone coverage			
No listed phone	-1.295	22.14	1,614,180
	(0.475)		
Listed phone	-0.475	19.13	3,172,505
	(0.396)		
Test of H_o : $\beta_{No-Phone} = \beta_{Phone}$		*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. Models are estimated using administrative data at the individual 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 (12.5 acres of land) were split into multiple checks. Outcomes are essentially perfectly correlated within individual, as farmers typically encashed all of their checks at the same time, which is accounted for by clustering at the mandal level.

Appendix A

Table A.1: Balance tests

Variable	(1) Control mean	(2) Treatment mean	(3) Difference (SE)
Land registry data			
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)
Census 2011 data			
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 (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.

Table A.2: Effect on check distribution outcomes (MAO reports)

	Distributed before June 8th		Ever distributed		
	(1) Treatment	(2) Control mean	(3) Treatment	(4) Control mean	(5) Obs.
Overall	0.00924 (0.00653)	0.81	0.00793 (0.00468)	0.87	5,645,937
Land quartiles					
Quartile 1	0.0177 (0.00984)	0.67	0.0165 (0.00878)	0.74	1,449,482
Quartile 2	0.00955 (0.00634)	0.83	0.00910 (0.00417)	0.89	1,460,294
Quartile 3	0.00742 (0.00568)	0.87	0.00654 (0.00319)	0.92	1,443,788
Quartile 4	0.00546 (0.00569)	0.87	0.00371 (0.00334)	0.93	1,443,836
Test of H_o : $\beta_{Q1} = \beta_{Q2} = \beta_{Q3} = \beta_{Q4}$	*0.43 (0.73)*		*1.18 (0.32)*		
Phone coverage					
No listed phone	0.00814 (0.0114)	0.69	0.00673 (0.0104)	0.76	2,254,142
Listed phone	0.00544 (0.00536)	0.89	0.00498 (0.00269)	0.94	3,543,258
Test of H_o : $\beta_{No-Phone} = \beta_{Phone}$	*0.10 (0.75)*		*0.08 (0.78)*		

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.

Table A.3: Effect on time to distribution (MAO reports)

	Days till distributed		
	(1) Treatment	(2) Control mean	(3) Observations
Overall	-0.125	11.70	4,930,113
	(0.310)		
Land quartiles			
Quartile 1	-0.220	13.55	1,082,824
	(0.386)		
Quartile 2	-0.0543	11.53	1,302,380
	(0.312)		
Quartile 3	-0.104	18.71	1,334,261
	(0.299)		
Quartile 4	-0.232	11.23	1,343,004
	(0.297)		
Test of H_o : $\beta_{Q1} = \beta_{Q2} = \beta_{Q3} = \beta_{Q4}$		*0.61 (0.61)*	
Phone coverage			
No listed phone	-0.128	13.85	1,729,723
	(0.403)		
Listed phone	-0.0826	10.57	3,332,746
	(0.286)		
Test of H_o : $\beta_{No-Phone} = \beta_{Phone}$		*0.05 (0.83)*	

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.

Table A.4: Agreement between phone and administrative data on MAO performance

	(1) Actual agreement rate	(2) Agreement rate from sampling variation	(3) Residual disagreement rate
Pair-wise order of rankings	68.6%	77.6%	9.0%
Bottom 20% in PD found in bottom 20% of AD	43.0%	61.7%	18.7%
Bottom 20% in PD found in bottom 50% of AD	83.0%	92.7%	9.7%

AD (Administrative Data). PD (Phone Data). The actual rate of agreement between phone and administrative data is reported in (1). Next, a comparison is made between the entire population of administrative data and 1,000 random draws of farmers sampled from the administrative data, where each draw is the size of the phone call sample. The mean of these 1,000 agreement rates is reported in (2), showing the amount of disagreement that we would expect due simply to sampling variation in which farmers were selected for the phone call sample. The residual disagreement rate after accounting for (2) is reported in (3).

Table A.5: Impact on beneficiary experience

	(1) Correct amount on check	(2) Received at Gram Sabha	(3) Asked to pay bribe	(4) Satisfied with scheme
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. Estimates are weighted using (inverse) sampling probability, as pre-specified, based on the probability that an individual was sampled for an attempted call. All specifications include randomization strata fixed effects. Standard errors are clustered at the MAO level and in parentheses. The number of observations varies due to lower rates of response on some questions, which were asked later in the phone survey.

Table A.6: Heterogeneity by land holdings

	(1) Ever distributed	(2) Days till distributed	(3) Ever encashed	(4) Days till encashed
Treatment	0.00917	-0.137	0.0141	-0.759
	(0.00459)	(0.314)	(0.00644)	(0.394)
Log land size	0.0625	-0.766	0.0721	-1.670
	(0.00157)	(0.0429)	(0.00153)	(0.0504)
Interaction	-0.00278	-0.0256	-0.00299	-0.112
	(0.00304)	(0.0868)	(0.00300)	(0.116)
Constant	0.862	11.92	0.815	20.66
	(0.00257)	(0.153)	(0.00351)	(0.200)
Observations	5,645,937	4,930,113	5,645,937	4,663,678

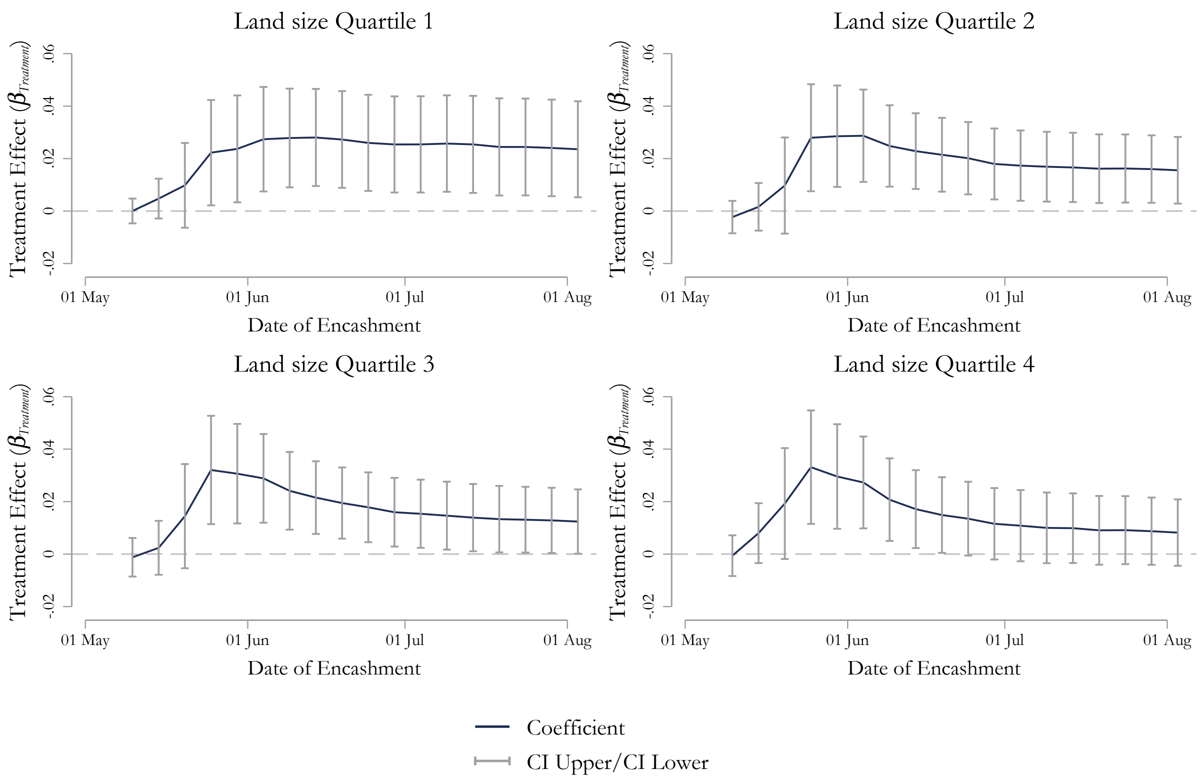
Outcome in header. Interaction: (Treatment)*(Log land size). Land size winsorized for the bottom 1% and top 1%, and logged. All specifications include randomization strata fixed effects. Standard errors in parentheses and clustered at the MAO level.

Table A.7: Testing for spillovers

	(1) Ever distributed	(2) Ever encashed
Number of treatment mandals in revenue division	0.000679	0.00847
	(0.00473)	(0.00551)
Constant	0.874	0.818
	(0.00696)	(0.00807)
Observations	399	399

As pre-specified, this table tests for the possibility that these results could be explained by supervisors of MAOs focusing more attention on treatment MAOs. Districts in Telangana are divided into "revenue divisions", which each contain several mandals. Although roughly the same fraction of mandals were treated in each district, we did not stratify the randomization at the revenue division level. As a result, there is random variation in the fraction of MAOs within each revenue division that are treated. If there were diversion of revenue division supervisor-level attention and attention matters for performance, we should expect worse performance among control MAOs with more treated MAOs in their revenue division, as these control MAOs would get less attention paid to them. This table does not find this to be the case. Outcome in header. All specifications include fixed effects for districts and number of mandals in the revenue division. Standard errors in parentheses and clustered at the revenue division level. 17 mandals could not be matched to revenue divisions, so were not included.

Figure A.1: Treatment Effect Over Time, by Landsize Quartile



The graphs in this figure report the coefficient of treatment effect on the cumulative rate of encashment over the period of check distribution, across the four landsize quartiles, as well as the 95% confidence interval. The first graph is among the quartile of farmers with the smallest farms, while the fourth is for the quartile of farmers with the largest farms. 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. Less than 1% of checks were encashed after August 4 or before 10 May, so the axis is restricted to those time periods.