Public Service Delivery and Free Riding: Experimental Evidence from India*

Alex Armand Britta Augsburg Antonella Bancalari
Maitreesh Ghatak

Abstract

This paper provides novel evidence on the mechanisms driving the combination of poor-quality public services and high prevalence of non-payment (free riding) in low- and middle-income countries. We implement a field experiment in the slums of two major Indian cities and in the context of a fee-funded public service provided by community toilets. Collecting original surveys, behavioral and objective measurements, we show that an exogenous boost in the maintenance quality of the service improves delivery and reduces free riding in a static and dynamic way, but excludes a share of residents from using the service. Providers react strategically to external rewards by shifting their efforts towards monitoring activities. Excluded users are forced to dispose human waste in common-property, generating large health externalities. Residents demand more public intervention in the service provision. (*JEL* C93, H40, I15, Q53)

Keywords: public service, infrastructure, maintenance, free riding, payment, water and sanitation, health.

---

*__Armand__: Nova School of Business and Economics – Universidade Nova de Lisboa, CEPR, NOVAFRICA, and Institute for Fiscal Studies (e-mail: alex.armand@novasbe.pt); **Augsburg**: Institute for Fiscal Studies (e-mail: britta_a@ifs.org.uk); **Bancalari**: School of Economics and Finance – University of St. Andrews, and Institute for Fiscal Studies (e-mail: antonella.bancalari@st-andrews.ac.uk); **Ghatak**: London School of Economics (e-mail: m.ghatak@lse.ac.uk). We are grateful to Nava Ashraf, Orazio Attanasio, Oriana Bandiera, Rachel Cassidy, Joan Costa-Font, Joshua Deutschmann, Fernanda Estevan, Greg Fisher, Sebastian Galiani, Anders Jensen, Michael Kremer, David Lagakos, Eliana La Ferrara, Meera Mahadevan, Berkay Ozcan, Rohini Pande, Imran Rasul, Hélène Rey, Sandra Sequeira, Sheetal Sekhri, Manisha Shah, Marcos Vera-Hernandez, and seminar/conference participants at the 2022 NBER Summer Institute, DSE Winter School, LSE, IFS, UCL, CSAE Oxford, University of St. Andrews, RES, 3ie Evidence Week, ADB Institute / Bill and Melinda Gates Foundation workshop, 6th Zurich Conference on Public Finance in Developing Countries, CEPR WE_ARE, and LISER for helpful comments. We would like to thank Tom Glinnan, Kerry Neitzel and Garima Rastogi for outstanding research assistance. This project would not have been possible without the insightful support of Abhijit Banerji, Sneha Santwana, Bhartendu Trivedi, and the teams at FINISH Society and Morsel Research and Development. We are also indebted to study participants for their willingness to contribute to the study. We gratefully acknowledge financial support from 3ie – International Initiative for Impact Evaluation (grant no. DPW1-1105), and from the ESRC's Centre for the Microeconomic Analysis of Public Policy at the IFS (ES/T014334/1). Armand acknowledges funding from Fundação para a Ciência e a Tecnologia (UIDB/00124/2020, UIDP/00124/2020 and Social Sciences DataLab - PINFRA/22209/2016), POR Lisboa and POR Norte (Social Sciences DataLab, PINFRA/22209/2016). Ethics approval was secured from University College London (application n. 2168/012). The pre-analysis plan was registered on the AEA RCT registry (Armand et al., 2018).

One of the most pressing issues of our time is how structural transformation is pushing the demand for basic services beyond capacity (Bryan et al., 2020; United Nations, 2022). Challenges to public service delivery, such as non-compliance with user fees or tax collection, are exacerbated in overcrowded cities and where institutional quality is poor (Burgess et al., 2020; Besley and Persson, 2013; Weigel, 2020; Jensen, 2022). When a large share of users do not pay for public services or 'free ride' (directly or indirectly through redistribution) effective prices of provision can fall below its marginal cost, enduring under-provision and inadequate service delivery. At the same time, user exclusion entails negative economic implications in our quest to provide basic services for all. Despite this dilemma, we know little about the mechanisms driving the combination of poor-quality public services and high prevalence of free riding in low- and middle-income countries (L&MICs).

We provide experimental evidence on how an exogenous boost in the maintenance quality of basic services impacts the behavior of both users and providers in one of the most challenging settings for public service delivery, informal settlements or *slums* (Marx et al., 2013). We show that improving service quality and payment comes at the expense of user exclusion, which generates high social costs if the outside option entails negative externalities.

The experiment is centered around community toilets (CTs). Present in many L&MICs, this public service provides access to hygiene and sanitation through shared complexes, serving mainly areas where access to safely-managed private toilets is constrained by limited housing.[1] In 2020, an estimated 3.6 billion people worldwide lacked access to safely-managed sanitation services, roughly half of it living in urban areas (WHO, 2021). In these settings, CTs often remain the only alternative to unimproved facilities or open defecation (OD), which generate large negative externalities.[2]

Our focus is on India, where only 37% of the 0.48 billion people living in urban areas has access to safely-managed sanitation services (WHO, 2021). In India, CTs are present across the country and operate with user fees, a model that characterizes the

[1] A *community* differs from a *public* toilet because it serves primarily a defined group of residents, rather than the general public.

[2] Unimproved sanitation and OD could be considered an alternative form of free riding. They damage environmental quality and impose health externalities (Coffey et al., 2018), increase mortality (Geruso and Spears, 2018) and jeopardize human capital accumulation (Miguel and Kremer, 2004; Bleakley, 2007; Adukia, 2017; Augsburg and Rodríguez-Lesmes, 2018; Orgill-Meyer and Pattanayak, 2020; Spears, 2020).

provision of most essential services in L&MICs and that often captures the largest part of the overall tax burden (see, e.g., Paler et al., 2017). Throughout India, the widespread non-payment of fees among users is combined with the extreme conditions of CTs: facilities are poorly maintained, dirty and with a widespread presence of bacteria harmful to health (National Geographic, 2017). The poor quality of the service is reflected in the very low willingness to pay (WTP) for the service among residents.

In partnership with governmental and non-governmental organizations (NGOs), we implemented a cluster randomized controlled trial (cRCT) in the slums of the two main cities of Uttar Pradesh, the largest state of India (Government of India, 2011). The study area exemplifies the urbanization process and the constraints to basic services that are commonly experienced by expanding cities in the poorest regions of the world (Section 1 provides a detailed discussion about external validity).

Following extensive efforts to map the universe of slums and CTs in the two cities, we randomly allocated 70 of the 110 CTs to a *maintenance* treatment group, with the remaining 40 serving as the control group. The intervention targeted the person in charge of fee collection and service delivery in the facility (the *caretaker*), and was structured in two subsequent components. In the first two months, labeled the *grant* period, a one-off grant was offered to primarily improve the structural quality of the facility. In the following 10 months, labeled the *incentive* period, caretakers were offered a large bimonthly financial reward (roughly 40 percent of their monthly salary), conditional on keeping the facility clean, to improve the quality of the routine side of service delivery.

To test whether citizens' mobilization can further improve public service delivery and reduce non-payment, in the catchment areas of a randomly-selected half of the *maintenance* treatment group, we implemented a sensitization campaign among potential users. The campaign was designed to raise awareness about the importance of payment for a well-maintained facility.

The study incorporates a unique set of measurements, combining observations, survey responses and incentivized behavioral measurements. From April 2018, over a period of 18 months, we gathered objective measures of the service's quality, the number of users, and the prevalence of non-payment among them. To map intervention effects into behavioral responses of both local service providers and potential users in slums, we supplement objective measures with panel survey data and behavioral measurements for both. While caretakers were identified in the process of listing CTs, a sampling frame for slum residents is generally rare, and was indeed missing in our setting. To overcome this limitation, one year before the begining of the intervention, we obtained a standard sampling frame by implementing a census of slum residents in both cities, collecting information on more than 30,000 households. Using this source, we interviewed users and potential users of CTs, i.e., households living in proximity to a CT, and without access to private toilets or reporting to be normally using the service. For both caretakers and the target population of slum residents, baseline data was supplemented with five rounds of panel data among caretakers, and three rounds among study participants.

We show that the maintenance treatment generates sustained improvements in the observed quality of the service. Quality improvements are achieved mostly by incentivizing caretakers, who improve routine maintenance and become more knowledgeable about best practices for maintaining the facility.

At the same time it reduces non-payment among users. The maintenance treatment reduces the share of users that do not pay the user fee by 17.8 percent as compared to the control group, accompanied by a small decrease in the total number of users. A dynamic analysis akin to Balboni et al. (2022) reveals that the maintenance treatment decreases the steady state at which CTs converge over time in terms of non-payment.

This pattern is not driven by an increased willingness of users to pay the service fee, but by the exclusion of users from using the service. In the maintenance treatment group, caretakers allocate a larger share of their time to monitoring (i.e., the collection of fees and the supervision of cleaning), which is sufficient to reduce non-payment among users. This behavioral change reduces the share of time allocated to operating activities, while leaving total labor supply unchanged.

The exclusion of residents is confirmed by potential users, who resort to the outside option. At endline, the average share of respondents who practice OD is 39.1 percent in the maintenance treatment group, as compared to 20.6 percent for the control group, an increase of 89.8 percent. The negative consequences for public health are confirmed by an increase in reported health issues in treated areas. Citizen mobilization increases as they demand further public intervention in the maintenance of the basic service. Using a purposefully designed structured community activity (SCA) to measure such demand in the study area, we show that the maintenance intervention increases the likelihood of residents asking local politicians to intervene in the operation and maintenance of

CTs (52.0 percent over the control mean), while it decreases the likelihood of asking them to address OD in their community (17.3 percent over the control mean). This result reflects OD becoming an acceptable practice in light of user exclusion, when at the same time awareness of the health externalities from OD also increases.

Our results provide novel insights to the literature on public service delivery in L&MICs. We shed light not only on the mechanisms driving their quality, but also on the obstacles to public infrastructure maintenance, an issue largely ignored in the literature (Duflo et al., 2012).[3] We show that improvements along these dimensions are driven by top-down incentives, complementing the theoretical evidence on the management of public goods when coordination among citizens fail (Banerjee et al., 2008). Within this line of incentives, we document the importance of decentralized mechanisms based on local providers, highlighting the role of performance-based incentives and of multi-tasking for these pro-socially motivated jobs (see Besley and Ghatak, 2018, for a review on incentives and pro-social motivation). These findings complement centralized mechanisms of public service delivery centered around state capacity and governance, which are more prevalent in the literature (Best et al., 2017; Burgess et al., 2017; Rasul and Rogger, 2018; Bandiera et al., 2021; Fenizia, 2022).

1 Background

Our experiment is implemented in the slums of Lucknow and Kanpur, the capital and the second largest cities in the Indian state of Uttar Pradesh. These cities provide a unique setting to study the constraints to public service delivery in L&MICs as they exemplify the urbanization processes experienced by growing cities in the poorest regions of the world. In 2015, Lucknow was the 129th largest city worldwide with 3.2 million inhabitants, and Kanpur was the 141st with 3.0 million inhabitants. Their populations are expected to grow by 2035 by 59 and 37 percent, respectively (United Nations, 2018). These prospects are similar to those of cities such as Accra (Ghana), Amman (Jordan), or Hyderabad (Pakistan), and of metropolises such as Karachi (Pakistan), Cairo (Egypt)

[3]Available evidence focuses on newly constructed infrastructure. For L&MICs, it covers the education (Duflo, 2001), electricity (Dinkelman, 2011; Rud, 2012; Lipscomb et al., 2013), housing (Galiani et al., 2017; McIntosh et al., 2018), water and sanitation (Devoto et al., 2012; Meeks, 2017; Alsan and Goldin, 2019; Bancalari, 2020), and transportation infrastructures (Gonzalez-Navarro and Quintana-Domeque, 2016; Donaldson, 2018; Asher and Novosad, 2020).

or Manila (the Philippines). The result of such a rapid growth is the proliferation of slums, home to more than 1 billion people worldwide and mostly located in L&MICs (United Nations, 2020). In Lucknow and Kanpur, slum residents are 13 and 15 percent of the population, respectively, comparable to the 15 percent of India's capital, Delhi.

Slums represent an extreme case of both poor access to public services and high prevalence of open defecation. In particular, the lack of access to water, sanitation and hygiene (WASH) is a major issue. While shared solutions are common in the slums of L&MICs (UNICEF, 2019), a large share of slum populations have access only to unimproved facilities or revert to open defecation (OD). Of the estimated half a billion people practicing OD worldwide, about 10 percent live in urban areas, with India being the most affected (WHO, 2021).

These conditions are common in our study area, where more than 40 percent of slum residents lack access to private toilets (Government of India, 2011). For these households, CTs provide the only access to WASH. Arranged in gender-specific areas, they offer sanitation, hand-washing and bathing facilities. The buildings are constructed by municipal corporations (or *Nagar Nigam*), i.e., the institution responsible for community services in cities with more than 1 million inhabitants. Facilities are connected either to sewerage systems or to septic tanks, providing an upgrade in the sanitation ladder compared with rudimentary private facilities or OD. Services are generally rendered on a long-term public–private partnership funded by user fees, with a single access costing a standard fee of 5 Indian rupees (INR, corresponding to US$ 0.07).[4]

Service delivery is performed by caretakers. They are in charge of the daily operation and management (O&M) of the CT, which includes collecting user fees, implementing routine maintenance by cleaning the facility or supervising cleaners, maintaining the stock of cleaning agents, and demanding and/or implementing repairs and maintenance of sanitation systems. Caretakers are hired centrally by the organization managing the CT and are supervised by zone managers who are charged with multiple facilities. They receive a fixed salary, equal on average to INR 5,000 (US$71) per month. Caretakers can be fired or moved to another facility in the case of poor performance. This option is not common as, in our sample, caretakers have on average 10 years of experience in the same job, and 4 years working in the same facility. Qualitative interviews with

[4]There is almost no variation of the fee across the two cities. Nominal INR are converted to nominal US$ using the 2019 average exchange rate of 70.42 US$/INR (IMF, 2020).

caretakers and with higher-level managers highlight that salaries do not include any performance-based incentive (Armand et al., 2020a).

The quality of the service rendered is poor and correlates with caretaker's characteristics and behavior.[5] On average, CTs serving slums are poorly maintained, reflected by low quality of construction, lack of functioning hand-washing facilities, and dirtiness. Less than 40 of facilities have finished walls for the compound, and hand-washing is available in only half of the facilities. Female areas are generally in a worse state. The low quality of service delivery is perceived by study participants with less than half reporting liking the services offered.

Low quality is accompanied by a high degree of non-payment among users (panel A in Figure 1). 34 percent of users do not pay the fee, with a higher percentage among female users (50 percent versus 24 percent among male users). Payments are only partly enforced by caretakers, with just 8 percent of study participants reporting having been prevented from using the facility for not been willing to pay the fee. At low levels of OD in the slum, non-payment and OD are negatively related, indicating that stricter payment monitoring may lead residents to practice OD. At high levels of OD in the slum, however, the relationship turns positive, which characterizes areas with a poor quality of the service and rampant non-payment.

Beyond the high free riding in service payment, there is also very low WTP for the service among potential users. An average household of four members could spend 8 percent of their average household income to use the service, a share that is smaller than the one spent on intoxicants. However, an incentivized measure of WTP for using the CT among potential users (detailed in Section 4.2) shows that, at baseline, WTP is particularly low (panel B in Figure 1). On average, WTP is INR 1.40 (28 percent of the official fee), higher for male respondents (INR 1.46 versus 1.36 for female respondents) and for households that use the service frequently (INR 1.53). At current conditions, WTP is unrelated to the quality of the facility, but study participants are, on average, willing to pay above the market price of INR 5 to access a hypothetical higher-quality CT.

[5] Recent improvements are more prevalent if caretakers spend more time in operations. Non-payment is lower if the caretaker is male, is pro-socially motivated, spends more time monitoring the payment, and implements better maintenance (Supplementary Material S.1).

2 The interventions

In partnership with city governments and NGOs,[6] we implemented two interventions with the objectives of improving the quality of the public service delivered by CTs and of raising awareness among potential users about the importance of eradicating OD in the community and of receiving a basic service of adequate quality. Supplementary Material S.2 provides the timeline and operational details.

Maintenance intervention. The first intervention was designed to boost the quality of public service delivery. Following Holmstrom (2017), we stimulated multiple components of service delivery by splitting the intervention in two subsequent periods. During the first two months of the intervention, labeled the *grant period*, we aimed at rehabilitating and improving the structural quality of the facility by providing a one-off grant. This component was offered to the caretaker who, according to the facility's needs, could choose to spend the grant in one of three equally-valued packages: repairs and/or refurbishments (chosen by 41 percent of caretakers), deep cleaning of the facility and the sanitation system (chosen by 41 percent), or the provision of tools and agents coupled with a training session on maintenance best practices (chosen by 18 percent). The average value of each package was INR 25,000 (US$ 355).

From the second to the twelfth months of the intervention, labeled the *incentive period*, we aimed at improving the routine side of service delivery. We provided a bimonthly financial reward to caretakers conditional on complying with various indicators of a clean and healthy facility. These indicators were selected based on the main determinants of inadequate service delivery at baseline. First, visible cleanliness of latrines rewarded caretakers with INR 500 (US$ 7.10). Second, the availability of soap in the hand-washing facilities rewarded caretakers with an additional INR 500 (US$ 7.10). Finally, bacteria counts kept to a minimum standard rewarded caretakers with a further INR 1,000 (US$ 14.20).[7]

We opted for an output-based absolute payment scheme with discrete incentive on the base of previous empirical evidence. First, we targeted own performance, rather than

[6]The partnership includes Lucknow and Kanpur Municipal Corporations, Sulabh International, and the zone and city managers of the CTs. The interventions were implemented by FINISH Society, a Lucknow-based NGO working in the WASH and waste management sectors.

[7]We selected a larger amount for the latter component because it reduces the exposure to pathogens (Flores et al., 2011), which is not guaranteed by visible cleanliness alone. The minimum standard is the baseline median value of the E. coli bacteria count in the cubicles of the study CTs.

relative performance or rankings, because social comparisons have been found to reduce performance in the health sector (Ashraf et al., 2014b). Second, in line with Bénabou and Tirole (2003) and Bandiera et al. (2015), we linked payments to contemporary performance to minimize gaming across periods, and we provided caretakers with information about their past performance during each round to help them estimating the effort required for the conditions. Third, to facilitate communication and implementation, the announcement of the scheme was made by a member of the implementing team, who provided a face-to-face explanation of the reward and shared a summary page with the main conditions, including a contact number to request further information. Finally, because in the context of pro-social tasks rewards have been found to be effective when their value relative to the baseline pay is high (Ashraf et al., 2014a), we set the magnitude of the bonus using baseline and qualitative data among caretakers and other service providers. In each round, the potential incentive is INR 2,000 (US$ 28.40), equivalent to 40 percent of their average monthly salary. In all rounds combined, this amount sums up to INR 8,000 (US$ 113.60), or 13 percent of the annual salary. The incentive is large as compared to other interventions that showed effects on exerted effort.[8]

Every two months and for a total of four times during the study, the conditions for the reward were verified by enumerators during random visits, and payments were delivered accordingly. Caretakers received on average INR 779 (US$ 11.06) in the first round of incentives, INR 1,036 (US$ 14.71) in the second round, INR 1,058 (US$ 15.02) in the third round, and INR 972 (US$ 13.80) in the last round. Paid amounts correspond to 39, 52, 53 and 49 percent of the potential reward, respectively. At the end of the intervention, caretakers were given a certificate signed by all implementing partners with the results achieved.

Sensitization campaign. The second intervention complemented the maintenance intervention with a sensitization campaign among potential users. The campaign aimed to increase awareness of negative externalities resulting from OD and contamination, the importance of paying the fee to fund the delivery of the public service offered by CTs, and the role of users in holding caretakers accountable for the quality of the service. The campaign was provided through four different means. First, door-to-door

[8]For India and in the context of education, Duflo et al. (2012) and Muralidharan and Sundararaman (2011) offer a reward equivalent to 1 percent and 3 percent of a typical teacher's annual salary, respectively. Larger magnitudes are more aligned to the incentives schemes provided to managers in the private sector (see, e.g., Bandiera et al., 2005).

visits were implemented three times in April–June 2018, July–September 2018 and January–March 2019. This component targeted all members of study households, and was implemented using a flip chart with pictures to allow participants with low literacy to process key messages. Second, at the end of the visit, a leaflet summarizing the main messages was left with the households. Third, posters highlighting messages provided during the door-to-door campaign were placed in the CTs. Fourth, voice message reminders were sent to study households' mobile phones with a monthly frequency.

3 Research design

The research design is a two-stage cRCT. In the first stage, the cluster is the CT serving slum residents and its catchment area. In the second stage, our unit of analysis is represented by potential users, i.e., slum residents that either use the service or dispose human waste in common-property areas. Supplementary Material S.4 provides further details about the experimental design and about sampling.

The lack of information for a sampling frame at both stages and the focus on a volatile population presented two main challenges for this research design. The first challenge was to locate CTs. For this purpose, we conducted a census of all CTs in the study area in 2017, followed by a geographical mapping of the slums surrounding each facility. For each facility, we gathered data on the location, the physical characteristics, the management of the service, maintenance practices, and the users. These data formed the basis for selecting as part of the study the active CTs offering the most common model of basic service delivery: a fee funded service used mostly by slum residents. To prevent potential risks of treatment contamination, we imposed distance requirements between CTs by removing areas served by multiple facilities. In addition, while caretakers work only in one facility, we limited their rotation to different facilities in agreement with service managers at the city level.[9] A total of 110 facilities were identified, 52 in Lucknow and 58 in Kanpur.

The second challenge was to obtain a sampling frame for potential users, which requires first identifying slum residents without access to adequate private sanitation infrastructure, and, among those, the ones that can potentially use the service rendered by a CT.

[9]During the study period, we do not observe rotation. In a share of facilities (ranging from 2 percent in follow-up 2 to 22 percent in follow-up 4), caretakers were replaced and the implementing team paid regular visits to inform the new caretakers about the intervention.

In the second half of 2017, we then conducted a census of all slum residents in the surrounding of the selected facilities. In total, we collected information on more than 30,000 households in both cities, covering demographics, dwelling characteristics (including geolocation) and access to basic services.

Using this information, we first define households without access to adequate private sanitation infrastructure as households where at least one member reported *not* to use a private toilet during the resident census, and with no intention to migrate in the 18 months following the census. To select potential users, we restrict this population to the households residing in the catchment area of a CT, defined as the space inside the slum borders and within a radius of not more than 250 meters from the facility. We fix this parameter after studying how service use among households without access to adequate private sanitation infrastructure is affected by the distance (computed using geolocation) from a resident's dwelling to the closest facility. We find that proximity is crucial: beyond 250 meters service use declines rapidly and OD becomes the main option (Supplementary Material S.1). Using these criteria, we identified 5,553 households with users and/or potential users of CTs from which we sampled 1,575 households for the study. The average characteristics of the sample are highly comparable to the population of slum residents in India (Supplementary Material S.4).

To create exogenous variation in the provision of the interventions discussed in Section 2, each CT was randomly allocated to one of two groups: the *maintenance* treatment group receives the maintenance intervention, and the control group does not receive any intervention. Within the *maintenance*, we cross-randomized the allocation of a *maintenance plus sensitization* treatment group, which receives the additional sensitization campaign in the catchment area. For randomization, we first stratified CTs according to the main organization managing the facility and the city. Using the rich census information, we then built blocks using Mahalanobis-distance relative proximity, and we randomly allocated each CT within a block to a treatment group using a lottery with equal probability of assignment. To further minimize the risk of treatment contamination, CTs that are within 400 meters from each other are allocated to the same treatment arm. As a result, 40 catchment areas were allocated to the control group, 70 catchment areas were allocated to the maintenance treatment and within these, 35 catchment areas were allocated to the maintenance plus sensitization treatment.

4 Data

To obtain information on both the service provision and potential users, we gathered a substantial amount of original data. Appendix A provides a definition of the variables used in the study. Supplementary Material S.4 provided detailed descriptions and scripts of each measurement.

4.1 CT survey and objective measurements

For each facility selected for the study, we administered a panel survey with the caretakers. The baseline survey was administered in April–June 2018, followed by five waves of follow-up data collection to document bimonthly the behavior of caretakers, starting one month after the baseline: in July–September 2018 (follow-up 1), October–November 2018 (follow-up 2), January–March 2019 (follow-up 3), April–May 2019 (follow-up 4) and July–September 2019 (follow-up 5). The questionnaire covered management and maintenance practices in the provision of the basic service. Appendix Table B1 presents descriptive statistics of facilities and their caretakers at baseline. In 80 percent of facilities, the CT is operated by a single caretaker; caretakers are generally male (82 percent), have roughly 10 years of experience in their job, and 44 percent live in the local community. Caretakers allocate 68 percent of their time to monitoring (collecting fees and supervising cleaners), while the remaining is allocated to conducting repairs, cleaning the facility, or spending time with the zone manager.

Survey data were supplemented with objective measurements of service delivery. First, in each wave of survey data, independent observers collected information about the quality, cleanliness and maintenance of facilities. Second, observers documented payment behavior at the facility by recording the number of users and the share of them paying the fee. This activity was performed for the duration of one hour between 4 and 6 am, when most residents of the community use the facility. Third, observers collected samples from randomly selected spots on the floor of facilities. These were then analyzed in a laboratory to identify the presence and counts of bacteria. On average, more than three types of hazardous bacteria were found in each facility in each round.

We further implemented an adapted dictator game played with caretakers to measure pro-social motivation for the cause. In each survey round, caretakers were provided with an endowment of INR 100 (US$ 1.42) with the option to donate all or part of it to a project improving access to WASH in disadvantaged areas of India, implemented by a partner NGO. Similar versions of this game have proven effective at identifying socially-motivated workers (see, e.g, Ashraf et al., 2014a).

Attrition was kept to a minimum between baseline and follow-up surveys. The average number of observations per facility equaled 4.9 out of 5 follow-up measurements, with no differential attrition across treatment groups (Appendix Table B4).

4.2 Household measurements

In conjunction with the baseline CT survey, a baseline survey was administered among study households in each catchment area. The survey covered a sample of 1,575 households. The main respondent was the household's main decision-maker, being in most cases the household head and always falling in the age range 18–64 years. The questionnaire covered the household's socio-demographic characteristics, the health status of family members, hygiene- and sanitation-related behavior and attitudes. Appendix Table B2 presents descriptive statistics for households at baseline. On average, household heads are 45 years old, mostly men, with primary education or less.

After the baseline, households were revisited three times in correspondence to the follow-ups 1, 3 and 5 of the CT survey. All follow-up surveys covered hygiene- and sanitation-related behavior and attitudes. In addition, follow-ups 3 and 5 collected data related to the health of family members. To measure OD among study households, in follow-up 5, the survey was supplemented with a list randomization question about sanitation behavior. This technique addresses potential stigma in questions related to sensitive behavior (see, e.g., Karlan and Zinman, 2012).[10]

Each baseline household was interviewed on average in 2.6 out of 3 follow-up measurements, implying attrition rates for individual follow-up surveys ranging from 9 to 19 percent. Only 2.7 percent of baseline households was never re-interviewed. To minimize sample loss, we interviewed additional households during follow-up surveys, which were randomly chosen from the baseline sampling frame. We observe no differential attrition across treatment groups (Appendix Table B5).

Survey data were supplemented with incentive-compatible behavioral measurements.

[10]Respondents were randomly allocated to report the number of true statements from either a list of statements on general behavior (short list) or the same list with an additional statement concerning the sensitive behavior (long list). The difference in the average number of true statements in the short and the long lists estimates the proportion engaging in the sensitive behavior.

First, in each wave of the household survey, we elicited individual-level WTP for service use for the most senior male and the most senior female decision-makers in the household, who are commonly spouses. Following extensive piloting and the low level of literacy in the sample, we opted for the incentivized version of the multiple price list (or take-it-or-leave-it) methodology, which performs well in settings where market prices are well-known (Andersen et al., 2006; Berry et al., 2020). The methodology prompts the participant to choose between different amounts of cash or a bundle of 10 tickets to use the CT in their catchment area. We do not focus on ability to pay because a single use is relatively cheap and highly recurrent.[11] One of the options is then randomly drawn and the decisions are realized. The game was played with each member alone, without other senior members present. While the market value of 10 tickets is INR 50 (US$ 0.71), we offered different amounts of cash, ranging from INR 0 to 60 (US$ 0.85, above the current market price to deal with truncation) with increases of INR 5 (US$ 0.07). We define the WTP for a single use as the point at which the participant switches from preferring the bundle of tickets to preferring the cash, divided by 10.

Second, similar to the game implemented with caretakers (Section 4.1), in each survey round we played with all participants an adapted dictator game to measure their willingness to contribute for the quality of service delivery. Each player was provided with an endowment of INR 50 (US$ 0.71) with the option to donate all or part of it to the purchase of cleaning products for the local facility. Within each area, the total amount donated was used to purchase cleaning products, then delivered to the caretaker.

Finally, we measured incentivized demand for public intervention using a novel SCA, labeled the *voice-to-the-people* initiative. In this activity, study participants were provided with an anonymous card containing a list of predefined issues in the community, and were given the opportunity to report the most pressing issue. Unique numerical identifiers allowed us to identify cards returned by study households. Each distributed card informed participants that the responses would be aggregated and provided to the city's municipal corporation. For the purpose of the study, we introduced in the list of issues the quality of the local CT and the contamination of the slum due to OD. We introduced these two issues in a random order among other pressing issues, which in-

[11] In the presence of new products, when ability to pay is more binding, price subsidies and microcredit have proven to be effective for adoption (Kremer and Miguel, 2007; Ashraf et al., 2010; Dupas, 2014; Ben Yishay et al., 2017).

cluded the poor health of children, the limited water availability, the poor quality of roads, the lack of trash collection, the absence of jobs, the limited access to healthcare and the poor lighting at night. We conducted this activity after follow-up 3, and reported the summary of issues to municipal corporations at the end of the study. A similar activity was implemented in different settings by Batista and Vicente (2011) and Armand et al. (2020b).

5 Results

As described in Section 4, to study behavioral responses of both service providers and study households to interventions, we collected up to five follow-up measurements that can be used to estimate treatment effects (i.e., t ranges from 0 to 5, with 0 indicating the baseline measurement). When outcomes are measured with a varying degree of noise and present low serial correlation, like in our setting, the availability of multiple follow-up measurements allows averaging out noise in the outcome variables and increase power by pooling measurements (McKenzie, 2012). Exploiting the design of the experiment, we estimate treatment effects by restricting the sample to post-baseline observations. We begin by estimating the impact of the maintenance treatment on the outcome $Y_{ij,t}$ of individual/household/CT i in catchment area j at time t using the following specification:

$$Y_{ij,t} = \beta\, T_j + \alpha\, \mathbf{X}_{ij} + \delta_t + \epsilon_{ij,t} \tag{1}$$

where T_j is an indicator variable equal to 1 if the catchment area j received the *maintenance* intervention (with or without the additional sensitization campaign), and 0 otherwise. $\mathbf{X}_{ij}$ is a set of indicator variables capturing randomization strata, while δ_t are survey round indicator variables. The error term $\epsilon_{ij,t}$ is assumed to be clustered by catchment area. Results are robust to alternative assumptions about clustering, such as catchment area and data collection round when the outcome of analysis is at individual or household level, and by catchment area when the outcome of analysis is at catchment area level.

To estimate the effect of providing only the maintenance intervention or implementing the sensitization campaign in addition to the maintenance intervention, we estimate the following specification:

$$Y_{ijt} = \beta_1 T1_j + \beta_2 T2_j + \alpha \mathbf{X}_{ij} + \delta_t + \epsilon_{ij,t} \tag{2}$$

where $T1_j$ is an indicator variable for whether the catchment area j is in the *maintenance only* group, and 0 otherwise, and $T2_j$ is an indicator variable for whether the catchment area j is in the *maintenance plus sensitization* group.

We provide evidence that supports the interpretation of β, β_1 and β_2 as the causal effect of the interventions. First, randomization was successful in creating observationally-equivalent groups in the experiment for both household and catchment area characteristics (Appendix Tables B1–B2). Beyond the features we implemented in the design of the experiment to minimize the potential treatment contamination (Section 3), we show evidence against spillover effects across treatment arms (Appendix B.5).

Second, we support main estimates with estimates using alternative specifications, such as ANCOVA or correcting for attrition using inverse probability weights (Appendix C), including control variables selected with the post-double selection LASSO procedure (Appendix B.6), and using the Athey et al. (2019) causal forest procedure (Appendix B.7). Appendices B.7–B.9 present estimates of heterogeneous effects based on machine learning techniques, and by prespecified dimensions, by gender and religion.

Finally, we supplement standard inference with multiple hypothesis testing. In each table, we present both p-values for the significance of each individual coefficient and p-values adjusted for multiple hypotheses using the List et al. (2019) bootstrap-based procedure. The latter considers all hypotheses tested within a table.

Section 5.1 presents estimates of treatment effects using equation (1) focusing on the following groups of outcomes: the quality of service provision, the payment for service use, the extent of OD, and health, and the attitudes of potential users towards the service. Tables 1–5 present estimates of treatment effects by pooling all available follow-up measurements. The hypothesis of interest in these tables is therefore whether the *maintenance* treatment T has an average impact throughout the study.

Multiple follow-up measurements also allow to study the timing of treatment effects within the study period. To this purpose, we pool follow-up observations according to the two periods of the maintenance intervention (Section 2): the *grant period*, which includes only follow-up 1, and the *incentive period*, which includes follow-up surveys

2–5. Appendix B.11 shows estimates for individual follow-up measurements, including the evolution of mean values of y_{it} in the control group. Section 5.2 presents estimates of treatment effects using equation (2) and discusses whether the *maintenance only* and the *maintenance plus sensitization* treatments generate impacts, and whether these impacts differ.

5.1 The effect of the maintenance treatment

The successful implementation of the *maintenance* intervention is reflected in significant differences in exposure measures across experimental arms (Appendix Table B6). To measure exposure we use two indicators. First, we define *transfer to a CT* as the sum of the value of the initial grant, which only treated CTs received, and of the amounts transferred from the subsidized use of tickets from the WTP game and from the provision of products donated by study participants as part of the adapted dictator game, both of which were played in treated and control clusters (see Section 4). Second, we define a *transfer to a caretaker* as the sum of the financial incentive provided in treated CTs and the amounts kept by the caretaker in each round of the adapted dictator game. Over the study period, the average transfer to a treated CT amounted to INR 25,270 (US$ 358.84), 16 times larger than the average transfer to a control CT. Similarly, caretakers in the control group received on average INR 373 (US$ 5.30), while caretakers in the treatment groups received an additional INR 4,179 (US$ 59.34).

Quality of the service

Table 1 presents estimates of treatment effects related to the quality of the service. Column (1) report impacts on the likelihood of a facility to provide a service of higher quality. To build this indicator we aggregate into an index all objective measurements about service delivery, including the information about the rendered service as collected by observers, and the lack of harmful bacteria as collected with samples from the floors. We define a service of higher quality if the facility delivers a service that is in the top quartile of the sample distribution of the index across all follow-up measurements. Appendix B.4 details the construction of the index and discusses impacts on the index and its overall distribution. The remaining columns focus on inputs to the quality of service provision. Column (2) focuses on structural maintenance, measured with an indicator variable equal to 1 if the facility received repairs and/or deep cleaning in the month previous to the visit, and 0 otherwise. Column (3) focuses on routine maintenance, measured with the number of tools, equipment and cleaners used during the last routine cleaning of the facility. Column (4) refers to the number of hours worked daily by the caretaker, while column (5) focuses on the caretaker's knowledge about maintenance, measured by an indicator variable equal to 1 if the caretaker knows the recommended practices, and 0 otherwise.

The interventions consistently improved the quality of service delivery. On average, the maintenance treatment leads to an increase of 11.6 percentage points in the share of facilities providing a service of higher quality. The effect, which is robust to multiple hypothesis testing with a p-value of 0.04, corresponds to an increase by 63.4 percent over the control mean. The maintenance treatment pushes upwards the whole distribution of the quality of service delivery index, detectable already at low levels of quality (Appendix Figure B1). A Kolmogorov-Smirnov test of the equality of the distributions of the index in the control and in the maintenance treatment group is rejected at the 1 percent level of confidence (Appendix B.4). Underlying drivers of quality improvements are improvements in the structural quality of the facility and in visible cleanliness, while no effect is observed for the presence of harmful bacteria.

Improvements in quality are specific to the incentive period, while we do not observe any significant effect during the grant period (panel A of Figure 2). To further understand how quality of service delivery increases, in columns (2)–(5) we therefore look at inputs. We observe a positive but insignificant effect on structural maintenance when pooling all follow-up rounds, but a large and significant effect during the grant period, equal to an increase by 32.6 percentage points in the likelihood to have received this type of maintenance, as compared to the control group (panel B of Figure 2). This suggests that the rehabilitation of the infrastructure providing the basic services is strongly tied to external funding. Combining this result with the timing of the effect on the quality of service delivery, we show that structural maintenance does not translate into an overall improvement in quality.

Improvements in quality are instead accompanied by inputs to the routine side of the service delivery. In the overall period, routine maintenance increases by 3.4 percentage points, corresponding to a 5.6 percent increase over the control mean. In addition, the share of caretakers who are knowledgeable raises by 8.9 percentage points, as compared with an average share of 6.8 percent in the control group. This effect is mainly driven by caretakers with higher pro-social motivation at baseline (Appendix B.8). Correcting for multiple hypothesis testing, the effects on routine maintenance and knowledge have p-values of 0.13 and 0.06, respectively. The effect on routine maintenance is significant only in the incentive period, when caretakers are financially incentivized to keep the facilities clean (panel B of Table 1 and panel C of Figure 2). This highlights how the quality of the service is closely related to routine rather than structural maintenance. These results are not achieved by an increase in labor supply, as caretakers do not change their workload in terms of time input, continuing to work on average 12 hours a day in all treatment arms. Changes along this dimension might be limited by labor supply being closely aligned with the opening times of facilities.

The effect of the maintenance treatment on the quality of the service tend to be homogeneous across facilities (Appendices B.7–B.8). We highlight that the effect on routine maintenance in driven by catchment areas where the average WTP for the service is low, and positive effects on quality and routine maintenance are driven by areas where potential users are more prevalently of Hindu religion (Appendix B.9).

Payment for the service

Table 2 turns to outcomes related to the payment for the service. Column (1) documents the effect on the total number of users (reported in logarithms), while column (2) shows the effect on non-payment, defined as the share of users who use the service without paying the fee. Both outcomes rely on data collected by independent observers (Section 4.1). Columns (3)–(4) focus instead on the caretakers' efforts to obtain the payment. Column (3) refers to the enforcement of payment, computed as the share of study households that reported having been refused entry to the facility for not paying the fee. Column (4) reports the effect on monitoring, defined as the share of time the caretaker reports allocating to collecting fees and supervising cleaners, as opposed to conducting repairs, cleaning the facility, or interacting with the zone manager.

The maintenance treatment reduces the number of users by 5.9 percentage points, but the estimate is not statistically significant, although the point estimate is consistently below zero in the incentive period (panel B). On the contrary, the maintenance treatment leads to a significant reduction in non-payment by 7.5 percentage points on average (a

17.8 percent increase over the control mean). While the estimate for the overall period is not robust to multiple hypothesis testing (the corrected p-value is 0.13), it is robust if we restrict the sample to the incentive period, when the effect is present (in panel B, the estimated effect is -9.3 percentage points and the adjusted p-value is 0.06). In addition, these estimates are also robust to using machine learning techniques (Appendices B.6 and B.7).[12]

We next study dynamics in non-payment. To formalize this notion, define the transition equation as the function that relates public good $i's$ non-payment stock across two time periods:

$$k_{i,t+1} = \phi_i(k_{i,t}) + \epsilon_i \tag{3}$$

where $k_{i,t}$ denotes $i's$ non-payment at time t. A steady state is defined by a fixed point of $\phi_i(\cdot)$, that is, a level of non-payment, k_i^*, such that $k_i^* = \phi_i(k_i^*)$. We estimate the transition equation for non-payment between the grant period (t) and the incentive period ($t+1$). This dynamic analysis is critical to understanding if small differences in initial free riding can result in large, permanent differences that translate into higher revenues and better quality of public service. A causal interpretation of the transition equation would fail if, for example, $k_{i,t}$ was systematically correlated with unobservable characteristics that shape the response of CTs to the intervention. To alleviate this concern we control for operators' and caretakers' characteristics.

We use the maintenance exogenous shock to evaluate two views of the dynamics of non-payment. Figure 3 show the estimates of Equation 3 in the control and maintenance groups using a kernel-weighted local polynomial regression. We observe that all CTs initially below a level $\hat{k}$ end up with higher non-payment in $t+1$, while those initially above this level end up with lower non-payment. Over time, CTs reverse to a stable steady state of non-payment. The maintenance treatment generates a stochastic negative shift of the transition equation for non-payment, compared to that of control, and decreases the level of non-payment in equilibrium. We identify the steady state numerically by finding a point in the local polynomial plot just above and just below the intersection with the 45 degree line and averaging its coordinates. This yields a level

[12]We estimate that the reductions in users and non-payment translate into a small positive effect on revenues, but the estimate is not statistically significant (Appendix B.3). Costs and revenues at the facility level are not observable available. We estimate revenues using observers' data on users and payments, but restricted to the times during which these data are collected.

equivalent to 50 percent of users not paying in the control group and 42 percent of users in the maintenance.

Notice the non-monotonicity at the top of the curves. For control, the curve crosses again the 45 degree line at this point. This generates a threshold above which CTs can end up in a higher steady state. It is indeed the case that the estimated threshold falls exactly in the low-density range of the baseline distribution of non-payment (0.85). This observation should be interpreted with caution, however, as the potential unsteady and high steady states fall within the confidence intervals of the curve. It is encouraging that in the maintenance plot there is no potential for this high steady state.

The reduction in non-payment is driven mostly by monitoring of payments. Going back to the static effects, we observe that the maintenance treatment leads caretakers to spend a significantly larger share of their time on monitoring activities, which increase by 5.2 percentage points. This result suggests that the caretaker's presence at the facility's payment point is sufficient to disincentivize non-payment among users. At the same time, it does not induce caretakers to employ more stringent payment enforcement, which is rarely applied as only 8.0 percent of study households report being refused entry for not having paid the service fee. Increases in enforcement in response to the maintenance treatment are only observed in facilities characterized by a high degree of non-payment at baseline (panel D of Appendix B.8).[13]

In line with the effects on the quality of the service and on inputs to provision (Section 5.1), the effects on non-payment and on monitoring are specific to the incentive period, while no effect is recorded during the grant period (panel D and panel E in Figure 2). These results highlight how improving the quality of the service and reducing non-payment are closely related to the incentives of the providers delivering the service to the users.

OD and health outcomes

Because the maintenance treatment increases monitoring and reduces non-payment (Section 5.1), we study how it impacts the use of the outside option. Using the list randomization technique detailed in Section 4.2, Figure 4 shows the share of potential

[13] We also estimate effects on the caretaker's pro-social motivation for the cause, measured through the share of the endowment that the caretaker donates in the adapted dictator game. Caretakers donate on average 35 percent of the transfer each time the game is played during follow-up visits, with no statistically significant difference across treatment arms (Appendix Table B18).

users practicing OD at follow-up 5. The maintenance treatment increases the share of respondents who practice OD by 18.5 percentage points, compared to a share of 20.6 percent in control group.[14] These findings serve as evidence that incentivizing the quality of service delivery results in the exclusion of a share of potential users from the service. Excluded users are those not willing to pay the user fee, who then revert to practicing OD.

We provide further evidence on the exclusion of users by comparing the characteristics of study participants that reportedly stopped using the service between baseline and endline to those that did not change their sanitation behavior. Appendix Table B3 shows that characteristics that proxy poverty correlate significantly with a switch from service use to OD, such as female-headed households and those with fewer assets. Households that had access to a toilet at baseline but reported using the CT are also more likely to stop using the public facility.

To understand the impacts on health on study households, Table 3 focuses on morbidity and health expenditures. Column (1) refers to the (self-reported) morbidity, measured by an indicator variable equal to 1 if any household member had fever, diarrhea or cough during the two weeks prior to the interview, and 0 otherwise. Columns (2)–(5) focus on reported health expenditures, distinguishing between curative and preventive expenditures, and analyzing changes both at the extensive and intensive margins. Curative expenditures include costs associated with doctor visits during illnesses, medicines and diagnostics, and hospitalization. Preventive expenditures include all costs associated with CT fees, access to drinking water and hygiene, medical checks and preventive goods like vaccines, bed-nets, and anti-worm tablets.

Slums are a high-disease environment: the share of study households in the control group with positive curative expenditures in the two weeks previous to the interview is 64 percent. While, on average, we find no treatment effect on morbidity, the maintenance treatment increases the probability of spending a positive amount on curative healthcare by 4.6 percentage points (equivalent to an increase of 7 percent over the control mean and the adjusted p-value is 0.14). This effect is driven by a significant increase in the first follow-up of the incentive period. In line, we find that the mainte-

[14]Using the same technique, we estimate that, in the control group, the service is used by 58.0 percent of respondents and soap is used by 81.6 percent of respondents. We do not find any significant effect of the maintenance treatment for these variables (Supplementary Material S.4). We also do not observe any impact on self-reported behavior, which under-report sensitive behavior (Appendix B.10).

nance treatment increased self-reported morbidity by 7.6 percentage points in the same period (Appendix Figure B8). We find no effects on the intensive margin of curative expenditures, and no effects on preventive expenditures.

We highlight some important heterogeneity of the effect of the maintenance treatment on health (Appendix B.8). First, it leads to worse public health in slums with initially worse water quality, consistent with an increase in waterborne diseases from greater OD. Second, it increases morbidity in catchment areas with initially lower free riding in payment to use the service.

Attitudes towards the service

Table 4 focuses on the attitudes towards the service among potential users. Column (1) shows impacts on their incentivized WTP for a single service use. Column (2) refers to whether the respondent is aware of the health externalities linked to OD in the slum. Columns (3) to (5) analyze the effects on the demand for public intervention. For this we use the data collected through the voice-to-the-people initiative. Column (3) considers the likelihood of households demanding public intervention for improvements in the quality of basic services, in particular on the cleanliness of the CT, column (4) for a reduction in OD and the associated contamination, and column (5) for other issues in the slum, including the health of children, the availability of water, the quality of roads, the collection of trash, the lack of jobs, the limited access to healthcare, and the poor lighting at night.

We find no significant effect on the WTP for using the service. The average WTP for a single use is equal to INR 1.15 in both the control and the maintenance treatment groups, as compared with the market price of INR 5. Instead, the maintenance treatment increases the share of study households that are aware of the health externalities that OD generates to their family by 3.1 percentage points, an impact that loses significance when accounting for multiple hypothesis testing. Households are generally aware of the risks associated with contamination, with 66.0 percent of households in the control group aware of the externalities.

We next to the demand for public intervention. OD is a more pressing issue in the community as compared the quality of the service delivered by the CT. In the control group, the former issue is reported by 43.2 percent of households, as compared to 9.8

that report the latter. The maintenance treatment group leads to important changes in the demand for public intervention. First, it leads the likelihood of households reporting to local politicians that they need to intervene in the quality of CTs to increase by 5.1 percentage points, an increase of 52.0 percent over the control group. The effect is significant at the 5 percent level, and at the 12 percent level when correcting for multiple hypothesis testing. This effect is balanced by a reduction of 7.5 percentage points in the demand for addressing OD in the slum, while we do not observe any treatment effect on the demand for other issues in the slum.[15]

We highlight some important heterogeneity of the effect of the maintenance treatment (Appendix B.8). First, trust plays an important role in determining the demand for improving the quality of the service, as the increase is driven by catchment areas where potential users have lower trust in the community. Second, the payment for the service at baseline determines the shift in the demand to address OD. The maintenance treatment induces larger decreases in such demand where, at baseline, the average WTP for the service was higher and non-payment lower. These heterogenous results are consistent with OD becoming an acceptable practice in light of user exclusion in areas where free riding in payment for the service was initially low. Finally, political representation plays an important role in determining demand for public intervention, especially in light of the political salience of Hinduism in our study area and their discourse about sanition being a right. Decreases in the demand for addressing OD are driven by respondents of Hindu religion, and by areas where potential users are prevalently of Hindu religion (Appendix B.9).

5.2 The role of demand-side sensitization

The exposure to WASH campaigns is relatively high among study households. Nevertheless, the sensitization campaign was effective at reaching the targeted population (Appendix Table B.2). The maintenance plus sensitization treatment increases the share of study households reporting exposure to a WASH campaign previous to the interview

[15] We also estimate the effects on the study households' incentivized willingness to contribute for the quality of service delivery. We observe no effect of the maintenance treatment. In both treatment and in control catchment areas, the share donated by study households during the adapted dictator game is on average 21.2 percent of the transfer. In addition, we do not observe changes in attitudes toward cooperation among study households. During the public goods game, the contribution to the public pot is on average 17.4 percent of the endowment in both the maintenance treatment and the control groups (Appendix Table B18).

by 8.3 percentage points, as compared to an average share in the control group of 64.5 percent. The maintenance plus sensitization treatment group also raises the recall for all means used in the campaign, with the largest effect recorded for posters (15.8 percentage points compared to the control group), and the exposure to voice messages, which double as compared to the exposure in the control group (0.83 versus 0.19).

For all outcomes presented in Tables 1–3, Table 5 reports estimates of the effect of the maintenance only treatment and of the maintenance plus sensitization treatment, estimated with equation (2). Columns (1)–(2) and columns (4)–(5) report coefficients and standard errors, while p-values for the individual hypotheses are shown in columns (3) and (6). Column (7) tests the hypothesis that the impact of the two treatment arms do not differ.

For most outcome variables there is no differential effect of providing the sensitization campaign in addition to the maintenance intervention as compared to providing the maintenance intervention alone. This suggests that the quality of service delivery and the means to achieve it are mostly driven by top-down incentives. Nevertheless, we record two important effects of providing bottom-up incentives through the sensitization campaign in addition to top-down incentives.

First, demand-side sensitization has an impact in the number of service users. Providing only the maintenance treatment leads to a marginally significant decrease in the number of users as compared to providing the maintenance plus sensitization treatment. This pattern is observed in the first and last follow-up measurement of the incentive period (Appendix Figure B6). This effect is not observed in the maintenance plus sensitization treatment group, suggesting that sensitization compensates for the reduction in users caused by the maintenance only treatment. The larger reduction of users in the maintenance only treatment group is consistent with user exclusion during the incentive period.

Second, the increases in the awareness of the externalities of OD described in Section 5.1 are driven by the demand-side sensitization, in line with the aim of the campaign. Study households that were targeted by the sensitization campaign are 5 percentage points more likely to report being aware, as compared to a 1 percentage point, not statistically significant, in the maintenance only treatment arm. The effect remains significant when considering multiple hypothesis testing.

6 Funding basic services in L&MICs

The findings discussed in Section 5 highlight an important trade-off in a model in which basic service delivery is funded by user fees: it is possible to achieve better service quality and reduce non-payment, but at the expense of OD. In this section, we delve into this trade-off to further understand the mechanisms behind it and discuss implications of alternative instruments to reduce OD.

We consider a static framework where individuals can satisfy their basic needs by either accessing a service offered for fee (S) or rely on practices with negative externalities (F). Assume a slum is composed of a continuum of residents that are potential users, i.e., they have no alternative beyond the action set $a \in \{S, F\}$. We therefore define the share who chooses S as r, and the share who chooses F as $(1 - r)$. Residents are heterogeneous in the parameter θ, which captures tastes, norms or socio-economic characteristics that favor unsafe practices, OD in this context.[16] We assume that θ is distributed with a density function $f(\theta)$ uniform in the interval $[0, \bar{\theta}]$, where $\bar{\theta}$ indicates the resident's type with the largest value associated with OD.

Service funded by user fees. The resident i who chooses $a_i = F$ receives a utility θ_i at no direct cost. The resident i who chooses the service ($a_i = S$) obtains a fixed utility $u(f, q)$, where f is the structural quality of the infrastructure/facility, which is more time invariant, and q is the routine side of the service, which is closely related to service providers. To access the service the resident pays a service fee c (generally set at INR 5). To introduce non-payment, in line with the results discussed Section 5, we assume the fee is paid only if the provider monitors the payment, which happens with probability $\pi \in [0, 1]$. We define $\tilde{c} \equiv c \cdot \pi$ as the *effective fee* of the service.

Because OD contaminates the slum, we introduce an externality that is increasing in the share $(1 - r)$. While the externality negatively impacts all residents, we assume that its severity is larger for the choice F because it exposes to contamination more directly. This is standard in settings in which contamination leads to negative externalities, and proximity worsen their effects. For each action a, the externality is thus equal to $\gamma_a(1 - r)$, where $\gamma_a > 0$ captures its severity and $\gamma_F > \gamma_S$. While in other settings, residents might be uncertain about the presence and severity of externalities, we assume

[16] We do not model budget constraints explicitly because in our context the service is relatively cheap and we are interested in aggregating choices at the slum level.

no uncertainty with respect to both dimensions because slum residents are highly aware of their presence in our setting (Section 5.1).

The utility derived by resident i from action F is therefore equal to $U_i^F = \theta_i - \gamma_F(1-r)$, while the expected utility from action S is equal to $E[U_i^S] = u(f,q) - \tilde{c} - \gamma_S(1-r)$. For a given share r, resident i chooses the service if $E[U_i^S] \geq U^F$. We can rewrite the condition as:

$$\theta_i \leq u(f,q) - \tilde{c} + \Delta_\gamma(1-r) \equiv W \tag{4}$$

where $\Delta_\gamma = \gamma_F - \gamma_S$, and W indicates the resident's type that is indifferent between the two actions. To avoid the straightforward case in which the parameters set r equal to either 0 or 1, we assume that $0 < W < \bar{\theta}$. The share of residents who opt for the service is thus defined by $R(r) = \int_0^W f(\theta)d\theta = {}^W\!/_{\bar{\theta}}$. The (rational expectations) equilibrium is such that $r^* = R(r^*)$, which results in the following expression:

$$r^* = \underbrace{\frac{u(f,q)}{\bar{\theta}+\Delta_\gamma} + \frac{\Delta_\gamma}{\bar{\theta}+\Delta_\gamma}}_{\text{benefit from service}} - \underbrace{\frac{\tilde{c}}{\bar{\theta}+\Delta_\gamma}}_{\text{cost}} \tag{5}$$

Starting in a situation in which r^* is smaller than one, how can we aim at r^* approaching 1 by eradicating contamination? One option is to fund increases in quality by raising revenues, which demands increasing the effective fee. Conducting a simple accounting exercise based on the current cost of service delivery of a CT suggests that raising revenues by increasing users and reducing non-payment is sustainable. The current O&M costs for a CT with median characteristics is estimated at INR 10,200 (US$ 144.85) per month, while the cost of providing an improved service (conditional on the existing infrastructure) is estimated at INR 13,544–28,800 (US$ 192.33–408.97), or 1.3–2.8 times the current cost (refer to Supplementary Material S.3 for details).[17] Setting $r^* = 1$ and removing non-payment ($\pi = 1$) cover up fully the lower bound in the improved scenario and 71 percent of the upper bound of the improved scenario, which suggests that targeting non-payment can indeed sustain an improvement in the service. However, equation (5) highlights that raising both quality and revenues does not eradicate OD:

[17] An alternative model involves monthly passes, which provide unlimited access to the service for all family members at a fixed price. Only 8 percent of CTs in our sample provide this option at the (median) price of INR 80 (US$ 1.14). The Indian Government is considering the introduction of a monthly pass at the price of INR 200 (US$ 2.84). At follow-up 5, we measure the incentive-compatible WTP for monthly passes, which is found to be very low at INR 25 (US$ 0.35).

while increases in $u(f, q)$ attracts more users by raising the benefits of the service, the need of revenues to support operations demands higher effective fees, which in turn pushes users towards OD.

Our findings show that even raising q_I and q_P without changing the effective fee translates into the exclusion of a share of users, possibly due coordination failure between slum residents and service providers. To further understand users-providers coordination in our setting, we study how demand- and supply-side factors further affect non-payment by exploiting our experimental setting and decomposing the effect of the maintenance treatment T on non-payment Y using a mediation analysis. We follow the approach of Gelbach (2016) for the sample in the incentive period, when non-payment is mostly affected by interventions.[18]

Figure D9 in the Appendix plots the decomposition of the effect in these two components by grouping mediators according to Tables 1–3 and considering different sets of mediators. The negative net effect of the maintenance treatment on nonpayment is driven by supply-side factors, which tend to decrease non-payment. Increases in the quality of service delivery and in caretaker's effort explain the largest part of the decrease in non-payment (15.6 percent and 11.3 percent of the main effect, respectively). The negative effect would however be larger, were it not for demand-side factors that push non-payment in the opposite direction. Increases in the demand for public intervention is one of the key drivers of non-payment with a magnitude of 29.1 percent of the main effect. A share of the effect remains unexplained likely due to the role of unobserved mediators.

Free access to basic services. These hurdles towards users-providers coordination highlights an alternative solution to eradicate OD: free access to basic services.[19] To test whether this approach eradicates OD in our setting, we use the sample of catchment areas and exploit variation in time and across areas in r, q, and $\tilde{c}$ to provide structural

[18]Using a set of mediators $(m_1, .., m_k)$, i.e., outcome variables that can be affected by T and can indirectly induce a change in Y, the treatment effect $\frac{dY}{dT}$ is decomposed as the next equation shows, where the first part is explained by mediators, and R is the residual (unexplained) part:

$$\frac{dY}{dT} = \underbrace{\sum_{j=1}^{k} \frac{\partial Y}{\partial m_j} \frac{\partial m_j}{\partial T}}_{\text{explained by mediators}} + R \tag{6}$$

[19]Evidence on free basic services has shown so far limited effects on user behavior (Szabo, 2015). In addition, subsidized services can deter investment to modernize the infrastructure (Mcrae, 2015).

estimates of the parameters in equation (5). We define $r_{j,t}$ as the share of households in the slum j that reported not to have practiced OD at time t (see Section 4.2). To approximate $u(f, q)$ in equation (5), we first assume that f is equal to the sum between a basic benefit for using the service in general, α_0, and a benefit for using the service at a specific facility, a_j. In addition, we assume q is equal to a constant α_1 multiplied by a dummy variable, $higher_{j,t}$, for whether the service is observed to be of higher quality at time t (measured with the variable *higher quality* in Table 1). [20] For catchment area j at time t, we define $u(f_j, q_{j,t})$ with the following expression:

$$u(f_j, q_{j,t}) = (\alpha_0 + a_j) + \alpha_1 \cdot higher_{j,t} + \Omega_t + \epsilon_{j,t} \tag{7}$$

where time fixed effects Ω_t capture the fact that using the service might provide lower or higher benefits at different time of the year, and $\epsilon_{j,t}$ captures idiosyncratic differences in the benefit obtained from the service j at time t.

Variation in $\tilde{c}_{ij,t}$ is achieved at the CT level through π_j, determined by the share of time the caretaker spend in monitoring activities (variable *monitoring* in Table 2), and at the household level through c_i, determined by the distribution of tickets for free use of the CT as part of the incentivized WTP measurement (see Section 4.2). If households received the tickets as part of the WTP game in period $t-1$, then the fee c is reduced. Because tickets account for multiple uses and because we do not know ex-ante the times a household uses the facility, we need to consider alternative scenarios to compute the reduction in c. In *Scenario A*, receiving the tickets corresponds to setting c to zero, while in *Scenario B*, we assume that the household uses the facility once per day, and we calculate the exact discount introduced by the tickets in the time frame between two consecutive rounds of the WTP game. If households did not receive the tickets, then c is set at INR 5.

To address the potential endogeneity of $\tilde{c}_{ij,t}$, deriving from household's WTP and from caretaker's behavior, we instrument it with the random number extracted as part of the WTP game. This number determines whether the household receives the tickets or the cash. We perform a first stage regression and predict the household-level effective fee assuming exogeneity of the instrument and compute the slum level effective fee, $\tilde{c}_{j,t}$, by

[20] Estimates are robust to approximating $u(f, q)$ with a non-linear function in a continuous measure of observed quality to proxy for q.

averaging for each period the effective fee predicted from the first stage. Substituting equation (7) in equation (5), we obtain the following equation:

$$r_{j,t} = \lambda_0 + \lambda_1 q_{j,t} + \lambda_2 \tilde{c}_{j,t} + \tilde{\Omega}_t + \tilde{a}_j + \tilde{\epsilon}_{j,t} \quad (8)$$

where $\lambda_0 = \frac{\alpha_0+\Delta_\gamma}{\bar{\theta}+\Delta_\gamma}$, $\lambda_1 = \frac{\alpha_1}{\bar{\theta}+\Delta_\gamma}$, and $\lambda_2 = -\frac{1}{\bar{\theta}+\Delta_\gamma}$, while $\tilde{\Omega}_t$, $\tilde{a}_j$, and $\tilde{\epsilon}_{j,t}$ are Ω_t, a_j and $\epsilon_{j,t}$ rescaled by $\bar{\theta} + \Delta_\gamma$. Exploiting the panel dimension of our dataset, columns (1)–(4) in Table (6) provide estimates of equation (8) using a fixed effect model.[21] Columns (1)–(2) consider *Scenario A* to compute the effective fee, while columns (3)–(4) consider *Scenario B*. Columns (2) and (4) estimate equation (8) imposing $\lambda_1 = 0$ (i.e., higher quality does not contribute to r beyond the fixed benefit). For comparison, in columns (5)–(6), we estimate a reduced form regression at the household level of whether the respondent reported not to practice OD at time t on $\tilde{c}_{ij,t}$, instrumenting this variable with the random number extracted as part of the WTP game.

Table (6) provides a set of important results. First, OD is determined by the basic benefit derived from the service being smaller than the benefit obtained from OD for positive share of slum residents. The parameter λ_0 is estimated at 0.78–0.83, which represents the share of the (externality-adjusted) basic benefit α_0 as compared to the (externality-adjusted) benefit of OD for the resident that derives the larger benefit from OD, i.e., $\bar{\theta}$. This indicates that OD characterizes around 20 percent of households in our setting, in line with findings in Section 5.1.

Second, lowering the effective fee reduces OD. The effect of the effective fee on r, λ_2, is estimated to be negative, with an increase of INR 1 leading to an increase in OD equal to 3.1–3.8 percentage points. Exogeneity of our measure of effective fee is reinforced by estimates being comparable when removing the constraint on λ_1. However, in our sample characterized by highly-marginalized individuals, offering a higher quality service provides a reduction in the benefit from choosing S, as λ_1 is estimated to be negative. Such penalty captures the mechanisms associated with the exclusion of users from the service at higher level of quality.

Third, from the parameters of equation (8), we recover $\tilde{c}^* = \alpha_0 - \bar{\theta}$, the effective fee that would set OD to zero (assuming either $\lambda_1 = 0$ or a lower quality of service delivery).

[21] This approach assumes strict exogeneity, i.e., $E(\epsilon_{j,t}|q_{j,m}, \tilde{c}_{j,m}, a_j) = 0$ where $m = 0, .., T$. The constant term λ_0 is estimated by imposing the restriction $E[\tilde{a}_j] = E[a_j] = 0$.

This parameter is estimated to be negative at INR 4.8–6.3, indicating that eradicating OD requires subsidizing the use of the service beyond providing it for free.

7 Conclusion

Understanding the roots of inadequate public service delivery and of free riding in payment is fundamental to unleash the economic development of L&MICs. We provide novel insights by studying both supply- and demand-side incentives in the quality of public services. An exogenous boost in the quality of service delivery allows achieving sustained improvements in the observed quality of services and significant reductions in non-payment among users. However, these come at the cost of the exclusion of a share of users from the service, with subsequent increases in OD and worse public health. Users mobilize to demand more public intervention to solve this coordination failure.

Further analysis highlights an important puzzle in the delivery of public services in L&MICs: while externally funding improvements of services results in the exclusion of users, in line with the effects of policies enforcing financial sustainability through threats to disconnection (see, e.g., Coville et al., 2020), offering poor-quality services for free does not eliminate resorting to outside options with negative externalities.

These findings open new avenues for future research. First, if citizens treat access to basic services as a right, as highlighted by our results on the demand for public intervention, then fee-funded models of public service delivery fail and health externalities persists. As the international community aims to achieve universal access to safe and affordable basic services as part of the United Nations' Sustainable Development Goals, we need to enhance knowledge on how to provide sustainable basic services in the poorest settings. This requires further research on how to stimulate not only an improved quality of service delivery, but also the take-up of public services among citizens and the eradication of OD. For instance, conditional transfers have been used to reduce externalities in other settings such as deforestation (Jayachandran et al., 2017). Offering transfers to potential users conditional on using the service has been used in pilot projects in India.[22] However, offering such transfers demands monitoring and could result in overcrowding, leading to lower quality services at constant expenditure lev-

[22]For example, in the city of Ahmedabad in Gujarat, and in the Barmer district in Rajasthan (BBC News, 2015; The Times of India, 2017).

els. In line, free-to-use CTs in our study area are found in highly degraded status, a condition that is also observed in the rest of India (Armand et al., 2020a). Keeping quality of services constant while increasing users might demand larger investments in the infrastructure.

In addition, we also need to enhance knowledge on the design of effective mechanisms for tax collection and redistribution, which remain challenging in the poorest settings (see, e.g., Pomeranz and Vila-Belda, 2019 for a review of this research agenda). The optimal taxation policy given the outside option of OD could be adding a 'waste fee' to the property tax, but in informal settlements tax collection is extremely challenging. The alternative of introducing fines for OD is expected to be ineffective in slums because it requires further monitoring and could also lead to extortion (Ashraf et al., 2016). Imposing non-monetary fines for OD, such as working for community centers, is a more feasible solution in this setting.

Finally, while our results highlight the centrality of top-down incentives, improvements in public service delivery demand further evidence on the design of effective mechanisms that stimulate bottom-up incentives. In particular, understanding the limits to collective action in areas where coordination failures are more prevalent and social norms are more difficult to change is a crucial objective. For instance, further research is needed to understand the effectiveness of monitoring technologies in these environments by creating and reinforcing a new local norm of respect for the public good. In addition, the introduction of large upfront payments, such as monthly passes to use the sanitation facilities, can help achieving OD free slums.

References

Adukia, A. (2017). Sanitation and education. *American Economic Journal: Applied Economics*, 9(2):23–59.

Alsan, M. and Goldin, C. (2019). Watersheds in child mortality: The role of effective water and sewerage infrastructure, 1880–1920. *Journal of Political Economy*, 127(2):586.

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, 9(4):383–405.

Armand, A., Augsburg, B., and Bancalari, A. (2018). Community Toilet Use in Slums:

Willingness to Pay and the Role of Informational and Supply Side Constraints. AEA RCT Registry. June 20. https://doi.org/10.1257/rct.3087-1.0.

Armand, A., Augsburg, B., Bancalari, A., and Trivedi, B. (2020a). Community toilet use in slums: Willingness-to-pay and the role of informational and supply side constraints. *3ie Impact Evaluation Report 113*.

Armand, A., Coutts, A., Vicente, P. C., and Vilela, I. (2020b). Does information break the political resource curse? Experimental evidence from Mozambique. *American Economic Review*, 110(11):3431–3453.

Asher, S. and Novosad, P. (2020). Rural roads and local economic development. *American Economic Review*, 110(3):797–823.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014a). No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17.

Ashraf, N., Bandiera, O., and Lee, S. S. (2014b). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior and Organization*, 100:44–63.

Ashraf, N., Berry, J., and Shapiro, J. M. (2010). Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia. *American Economic Review*, 100(5):2383–2413.

Ashraf, N., Glaeser, E. L., and Ponzetto, G. A. M. (2016). Infrastructure, Incentives, and Institutions. *American Economic Review*, 106(5):77–82.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1179–1203.

Augsburg, B. and Rodríguez-Lesmes, P. A. (2018). Sanitation and child health in India. *World Development*, 107:22–39.

Balboni, C., Bandiera, O., Burgess, R., Ghatak, M., and Heil, A. (2022). Why Do People Stay Poor?

Bancalari, A. (2020). Can white elephants kill? Unintended consequences of infrastructure development in Peru. IFS working paper W20/32, Institute for Fiscal Studies, London, United Kingdom.

Bandiera, O., Barankay, I., and Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics*, 120(3):917–962.

Bandiera, O., Best, M. C., Khan, A. Q., and Prat, A. (2021). The allocation of authority in organizations: A field experiment with bureaucrats. *The Quarterly Journal of*

Economics, 136(4):2195–2242.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? A natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13–25.

Banerjee, A., Iyer, L., and Somanathan, R. (2008). Public action for public goods. *Handbook of Development Economics*, 4:3117–3154.

Batista, C. and Vicente, P. C. (2011). Do migrants improve governance at home? Evidence from a voting experiment. *World Bank Economic Review*, 25(1):77–104.

BBC News (2015). Paid to poo: Combating open defecation in India. https://www.bbc.com/news/health-33980904.

Ben Yishay, A., Fraker, A., Guiteras, R., Palloni, G., Shah, N. B., Shirrell, S., and Wang, P. (2017). Microcredit and willingness to pay for environmental quality: Evidence from a randomized-controlled trial of finance for sanitation in rural Cambodia. *Journal of Environmental Economics and Management*, 86:121–140.

Bénabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies*, 70:489–520.

Berry, J., Fischer, G., and Guiteras, R. (2020). Eliciting and utilizing willingness to pay: Evidence from field trials in northern ghana. *Journal of Political Economy*, 128(4):1436–1473.

Besley, T. and Ghatak, M. (2018). Prosocial motivation and incentives. *Annual Review of Economics*, 10:411–438.

Besley, T. and Persson, T. (2013). Taxation and development. *Handbook of Public Economics*, 5:51–110.

Best, M. C., Hjort, J., and Szakonyi, D. (2017). Individuals and organizations as sources of state effectiveness. NBER working paper 23350, National Bureau of Economic Research.

Bleakley, H. (2007). Disease and development: evidence from hookworm eradication in the american south. *The Quarterly Journal of Economics*, 122(1):73–117.

Bryan, G., Glaeser, E., and Tsivanidis, N. (2020). Cities in the developing world. *Annual Review of Economics*, 12:273–297.

Burgess, R., Greenstone, M., Ryan, N., and Sudarshan, A. (2020). The consequences of treating electricity as a right. *Journal of Economic Perspectives*, 34(1):145–169.

Burgess, S., Propper, C., Ratto, M., and Tominey, E. (2017). Incentives in the public sector: Evidence from a government agency.

Coffey, D., Geruso, M., and Spears, D. (2018). Sanitation, disease externalities and anaemia: evidence from nepal. *The Economic Journal*, 128(611):1395–1432.

Coville, A., Galiani, S., Gertler, P. J., and Yoshida, S. (2020). Pipe Dreams: Enforcing Payment for Water and Sanitation Services in Nairobi's Slums. NBER working paper 27569, National Bureau of Economic Research.

Devoto, F., Duflo, E., Dupas, P., Parienté, W., and Pons, V. (2012). Happiness on tap: Piped water adoption in urban Morocco. *American Economic Journal: Economic Policy*, 4(4):68–99.

Dinkelman, T. (2011). The effects of rural electrification on employment: New evidence from South Africa. *American Economic Review*, 101(7):3078–3108.

Donaldson, D. (2018). Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4–5):899–934.

Duflo, E. (2001). Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American economic review*, 91(4):795–813.

Duflo, E., Galiani, S., and Mobarak, A. M. (2012). Improving access to urban services for the poor: open issues and a framework for a future research agenda. *J-PAL Urban Services Review Paper. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab*, 102(4):1241–1278.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1):197–228.

Fenizia, A. (2022). Managers and productivity in the public sector. *Econometrica*, forthcoming.

Flores, G. E., Bates, S. T., Knights, D., Lauber, C. L., Stombaugh, J., Knight, R., and Fierer, N. (2011). Microbial biogeography of public restroom surfaces. *PLoS ONE*, 6(11):1–7.

Galiani, S., Gertler, P. J., Undurraga, R., Cooper, R., Martínez, S., and Ross, A. (2017). Shelter from the storm: Upgrading housing infrastructure in Latin American slums. *Journal of Urban Economics*, 98:187–213.

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, 34(2):509–543.

Geruso, M. and Spears, D. (2018). Neighborhood sanitation and infant mortality. *American Economic Journal: Applied Economics*, 10(2):125–162.

Gonzalez-Navarro, M. and Quintana-Domeque, C. (2016). Paving streets for the poor: Experimental analysis of infrastructure effects. *Review of Economics and Statistics*, 98(2):254–267.

Government of India (2011). Census of India 2011. Office of the Registrar General and Census Commissioner, Ministry of Home Affairs, Government of India, New Delhi, India.

Holmstrom, B. (2017). Pay for performance and beyond. *American Economic Review*, 107(7):1753–1777.

IMF (2020). Official Exchange Rate (LCU per US$, period average). Washington, D.C.: International Monetary Fund, International Financial Statistics.

Jayachandran, S., De Laat, J., Lambin, E. F., Stanton, C. Y., Audy, R., and Thomas, N. E. (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science*, 357(6348):267–273.

Jensen, A. (2022). Employment structure and the rise of the modern tax system. *American Economic Review*, 112(1):213–34.

Karlan, D. S. and Zinman, J. (2012). List randomization for sensitive behavior: An application for measuring use of loan proceeds. *Journal of Development Economics*, 98(1):71–75.

Kremer, M. and Miguel, E. (2007). The illusion of sustainability. *Quarterly Journal of Economics*, 122(3):1007–1065.

Lipscomb, M., Mobarak, A. M., and Barham, T. (2013). Development effects of electrification: Evidence from the topographic placement of hydropower plants in Brazil. *American Economic Journal: Applied Economics*, 5(2):200–231.

List, J. A., Shaikh, A. M., and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4):773–793.

Marx, B., Stoker, T. M., and Suri, T. (2013). The Economics of Slums in the Developing World. *Journal of Economic Perspectives*, 27(4):187–210.

McIntosh, C., Alegría, T., Ordóñez, G., and Zenteno, R. (2018). The neighborhood impacts of local infrastructure investment: Evidence from urban mexico. *American Economic Journal: Applied Economics*, 10(3):263–86.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*, 99(2):210–221.

Mcrae, S. (2015). Infrastructure quality and the subsidy trap. *American Economic Review*, 105(1):35–66.

Meeks, R. C. (2017). Water works: The economic impact of water infrastructure. *Journal of Human Resources*, 52(4):1119–1153.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1):159–217.

Muralidharan, K. and Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, 119(1):39–77.

National Geographic (2017). Nearly a billion people still defecate outdoors. Here's why. National Geographic Magazine. London, UK: August 2017.

Orgill-Meyer, J. and Pattanayak, S. K. (2020). Improved sanitation increases long-term cognitive test scores. *World Development*, 132:104975.

Paler, L., Prichard, W., Sanchez de la Sierra, R., and Samii, C. (2017). Survey on total tax burden in the drc. Final Report, Department for International Development.

Pomeranz, D. and Vila-Belda, J. (2019). Taking state-capacity research to the field: Insights from collaborations with tax authorities. *Annual Review of Economics*, 11:755–781.

Rasul, I. and Rogger, D. (2018). Management of bureaucrats and public service delivery: Evidence from the nigerian civil service. *The Economic Journal*, 128(608):413–446.

Rud, J. P. (2012). Electricity provision and industrial development: Evidence from India. *Journal of Development Economics*, 97(2):352–367.

Spears, D. (2020). Exposure to open defecation can account for the indian enigma of child height. *Journal of Development Economics*, 146:102277.

Szabo, A. (2015). The Value of Free Water: Analyzing South Africa's Free Basic Water Policy. *Econometrica*, 83(5):1913–1961.

The Times of India (2017). Use toilet and earn Rs 2,500 per month. https://timesofindia.indiatimes.com/city/jaipur/use-loo-and-earn-rs-2500-per-month/articleshow/56880690.cms?from=mdr.

UNICEF (2019). Global Framework for Urban Water, Sanitation and Hygiene. United nations children's fund, New York: US.

United Nations (2018). World Urbanization Prospects: The 2018 Revision. New York, US: United Nations Department of Economic and Social Affairs, Population Division.

United Nations (2020). Progress towards the Sustainable Development Goals. United Nations Economic and Social Council, United Nations Economic and Social Council, New York.

United Nations (2022). Progress in the implementation of the New Urban Agenda. General Assembly Economic and Social Council, United Nations Economic and Social

Council, New York.

Weigel, J. L. (2020). The participation dividend of taxation: how citizens in congo engage more with the state when it tries to tax them. *Quarterly Journal of Economics*, 135(4):1849–1903.

WHO (2021). Progress on household drinking water, sanitation and hygiene 2000-2020: five years into the SDGs. Joint monitoring programme for water supply, sanitation, and hygiene, World Health Organization and UNICEF.

Table 1: Quality of the service

Dep. variable:	Service delivery			Caretaker's input	
	Higher quality	Maintenance		Hours worked	Caretaker's knowledge
		Structural	Routine		
	(1)	(2)	(3)	(4)	(5)
Panel A: All rounds					
Maintenance (T)	0.116	0.045	0.034	0.175	0.089
	(0.043)	(0.049)	(0.017)	(0.356)	(0.035)
	[0.01 ; 0.04]	[0.36 ; 0.60]	[0.05 ; 0.13]	[0.62 ; 0.62]	[0.01 ; 0.06]
Mean (control group)	0.183	0.623	0.608	11.754	0.068
Observations	542	542	542	542	542
Catchment areas	110	110	110	110	110
Observation rounds	5	5	5	5	5
Panel B: Incentive period					
Maintenance (T)	0.146	-0.027	0.049	0.077	0.115
	(0.044)	(0.053)	(0.018)	(0.391)	(0.038)
	[0.00 ; 0.01]	[0.62 ; 0.60]	[0.01 ; 0.03]	[0.84 ; 0.85]	[0.00 ; 0.01]
Mean (control group)	0.164	0.625	0.578	11.388	0.059
Observations	434	434	434	434	434
Catchment areas	110	110	110	110	110
Observation rounds	4	4	4	4	4

Note. Estimates based on CT-level OLS regressions using equation (1). Standard errors clustered by catchment area are reported in parentheses. P-values are presented in brackets, the first from individual testing, the second adjusting for testing that each treatment is jointly different from zero for all outcomes presented in the table (see Section 5 for details). Dependent variables by column: (1) *Higher quality*: indicator variable equal to 1 if the quality index is above the 75th percentile, and 0 otherwise; (2) *Structural maintenance*: indicator variable equal to 1 if the CT received repairs and/or deep cleaning of the infrastructure in the month previous to the visit, and 0 otherwise; (3) *Routine maintenance*: number of tools, equipment and cleaners used during the last routine maintenance of the facility, normalized to be between 0 and 1; (4) *Hours worked*: number of hours worked by the caretaker; (5) *Awareness*: indicator variable equal to 1 if the caretaker knows the recommended practices of the cleaning routine and the need for deep cleaning, and 0 otherwise. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A.

Table 2: Payment for the service

Dep. variable:	Traffic		Caretaker's effort	
	Users	Non-payment	Enforcement	Monitoring
	(1)	(2)	(3)	(4)
Panel A: All rounds				
Maintenance (T)	-0.059	-0.075	0.002	0.052
	(0.050)	(0.042)	(0.024)	(0.028)
	[0.24 ; 0.25]	[0.07 ; 0.13]	[0.92 ; 0.92]	[0.07 ; 0.15]
Mean (control group)	3.477	0.422	0.080	0.686
Observations	542	542	110	542
Catchment areas	110	110	110	110
Observation rounds	5	5	1	5
Panel B: Incentive period				
Maintenance (T)	-0.057	-0.093	0.002	0.060
	(0.050)	(0.042)	(0.024)	(0.032)
	[0.25 ; 0.26]	[0.03 ; 0.06]	[0.92 ; 0.92]	[0.07 ; 0.14]
Mean (control group)	3.443	0.444	0.080	0.707
Observations	434	434	110	434
Catchment areas	110	110	110	110
Observation rounds	4	4	1	4

Note. Estimates based on CT-level OLS regressions using equation (1). Standard errors clustered by catchment area are reported in parentheses. P-values are presented in brackets, the first from individual testing, the second adjusting for testing that each treatment is jointly different from zero for all outcomes presented in the table (see Section 5 for details). Dependent variables by column: (1) *Users*: total number of users observed (reported in logarithms); (2) *Non-payment*: observed share of users who do not pay the entry fee; (3) *Enforcement*: share of surveyed households in the catchment area that reported having been refused entry to the CT by the caretaker for not paying the fee; (4) *Monitoring*: share of worked hours allocated by the caretaker to collecting fees and supervising cleaners, rather than conducting repairs, cleaning the facility, or spending time with the manager. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A.

Table 3: Health outcomes

Dep. variable:	Morbidity	Curative expenditure		Preventive expenditure	
		Extensive	Intensive	Extensive	Intensive
	(1)	(2)	(3)	(4)	(5)
Maintenance (T)	0.006	0.046	-35.152	-0.002	5.571
	(0.022)	(0.025)	(194.683)	(0.003)	(56.657)
	[0.79 ; 0.79]	[0.07 ; 0.14]	[0.86 ; 0.86]	[0.51 ; 0.77]	[0.92 ; 0.92]
Mean (control group)	0.401	0.637	1701.871	0.992	741.117
Observations	4793	3276	3276	3301	3300
Catchment areas	110	109	109	109	109
Observation rounds	3	2	2	2	2
Level of analysis	Household	Household	Household	Household	Household

Note. Estimates based on household-level OLS regressions using equation (1). Standard errors clustered by catchment area are reported in parentheses. P-values are presented in brackets, the first from individual testing, the second adjusting for testing that each treatment is jointly different from zero for all outcomes presented in the table (see Section 5 for details). Dependent variables by column: (1) *Morbidity*: indicator variable equal to 1 if any household member had fever, diarrhea or cough during the two weeks previous to the interview, and 0 otherwise; (2) *Curative expenditure - extensive*: indicator variable equal to 1 if the respondent had positive curative expenditures, and 0 otherwise; (3) *Curative expenditure - intensive*: level of curative healthcare expenditures (in Rupees); (4) *Preventive expenditure - extensive*: indicator variable equal to 1 if the respondent had positive preventive expenditures, and 0 otherwise; (3) *Preventive expenditure - intensive*: level of preventive healthcare expenditures (in Rupees). Columns (2)–(5) include only 109 catchment areas in the sample because the dependent variables were measured only in rounds 3 and 5, after a catchment area was displaced. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A.

Table 4: Attitudes of potential users

	WTP	Awareness	Demand for public intervention		
Dep. variable:	for service use	of externalities	Quality of basic service	OD free	Other issues
	(1)	(2)	(3)	(4)	(5)
Maintenance (T)	-0.042	0.031	0.051	-0.076	-0.021
	(0.074)	(0.018)	(0.026)	(0.037)	(0.024)
	[0.57 ; 0.57]	[0.10 ; 0.20]	[0.05 ; 0.12]	[0.04 ; 0.12]	[0.38 ; 0.40]
Mean (control group)	1.149	0.660	0.098	0.432	0.844
Observations	8635	4793	1551	1551	1551
Catchment areas	110	110	109	109	109
Observation rounds	3	3	1	1	1
Level of analysis	Respondent	Household	Household	Household	Household
Incentivized measurement	Yes	No	Yes	Yes	Yes

Note. Estimates based on respondent- and household-level OLS regressions using equation (1). Standard errors clustered by catchment area are reported in parentheses. P-values are presented in brackets, the first from individual testing, the second adjusting for testing that each treatment is jointly different from zero for all outcomes presented in the table (see Section 5 for details). Dependent variables by column: (1) *WTP for service use*: incentivized willingness to pay for a single CT use (in Rupees), elicited for a bundle of ten tickets and divided by 10 to get at single use WTP; (2) *Awareness of externalities*: indicator variable equal to 1 if the respondent reports that OD generates a health externality for their family, and 0 otherwise; (3) *Quality of the basic service*: indicator variable equal to 1 if the household asks for public intervention in the CT's O&M as incentivized through the voice-to-the-people initiative (Appendix S.4.8), and 0 otherwise; (4) *OD*: indicator variable equal to 1 if the household asks for public intervention in keeping the community OD free as incentivized through the voice-to-the-people initiative (Appendix S.4.8), and 0 otherwise; (5) *Other issues*: indicator variable equal to 1 if the household asks for public intervention in any other issues in the community captured on the 'voice-to-the-people' cards as incentivized through the voice-to-the-people initiative (Appendix S.4.8), and 0 otherwise. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Specifications where the level of analysis is the respondent also include gender. Additional details about the variables are presented in Appendix A.

Table 5: The effect of sensitization

	Maintenance only (T1)			Maintenance + sensitization (T2)			T1=T2
	β	se	p-value	β	se	p-value	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quality of the service							
Higher quality	0.14	0.06	0.01 ; 0.04	0.09	0.05	0.06 ; 0.20	0.44
Physical maintenance	0.02	0.05	0.68 ; 0.69	0.07	0.06	0.22 ; 0.23	0.32
Routine maintenance	0.04	0.02	0.05 ; 0.11	0.03	0.02	0.13 ; 0.24	0.80
Hours worked	-0.03	0.42	0.95 ; 0.95	0.38	0.38	0.32 ; 0.32	0.28
Caretaker's knowledge	0.08	0.04	0.08 ; 0.19	0.10	0.05	0.04 ; 0.11	0.66
Payment for the service							
Users	-0.10	0.06	0.10 ; 0.21	-0.02	0.06	0.76 ; 0.77	0.23
Non-payment	-0.06	0.05	0.20 ; 0.20	-0.09	0.05	0.06 ; 0.12	0.17
Enforcement	-0.02	0.02	0.53 ; 0.53	0.02	0.03	0.50 ; 0.52	0.50
Monitoring	0.04	0.03	0.20 ; 0.38	0.06	0.03	0.04 ; 0.09	0.12
Health outcomes							
Morbidity	0.00	0.02	0.83 ; 0.83	0.01	0.02	0.66 ; 0.66	0.90
Curative health expenditure:							
Extensive	0.03	0.02	0.23 ; 0.40	0.04	0.02	0.05 ; 0.09	0.15
Intensive	53.16	195.21	0.79 ; 0.90	-170.01	163.43	0.30 ; 0.68	0.42
Preventive health expenditure:							
Extensive	-0.00	0.00	0.25 ; 0.80	-0.00	0.00	0.59 ; 0.85	0.52
Intensive	65.95	66.81	0.33 ; 0.75	38.01	53.18	0.48 ; 0.87	0.58
Attitude of potential users							
WTP for service use	0.03	0.08	0.72 ; 0.87	-0.11	0.09	0.25 ; 0.25	0.37
Awareness of externalities	0.01	0.02	0.53 ; 0.71	0.05	0.02	0.01 ; 0.02	0.01
Demand for public intervention:							
Quality of the basic service	0.05	0.03	0.11 ; 0.23	0.05	0.03	0.12 ; 0.24	0.15
Free riding	-0.08	0.04	0.06 ; 0.15	-0.08	0.04	0.08 ; 0.22	0.12
Other issues	-0.03	0.03	0.26 ; 0.27	-0.01	0.03	0.64 ; 0.67	0.53

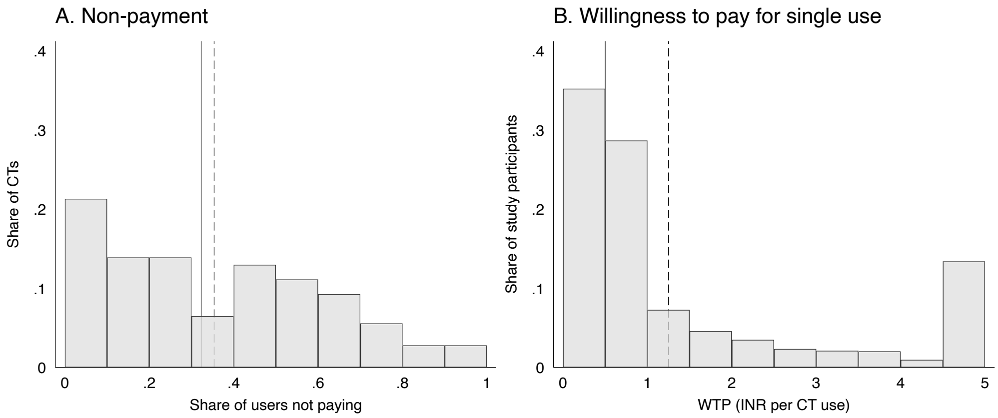
Note. In columns (1)–(6), estimates are based on CT-, respondent- or household-level OLS regressions using equation (2) for each outcome. $p-$values are presented in columns (3) and (6), the first from individual testing, the second adjusting for jointly testing that each treatment is different from zero for all outcomes presented in the table. Column (7) presents a test based on equality of coefficients of the effects of T1 and T2. Standard errors are clustered by catchment area for CT-level outcomes and by catchment-area–round for respondent- and household-level outcomes. The dependent variables are indicated in the rows and are defined in Appendix A. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT.

Table 6: Effective fee

Dependent variable:	**Share choosing S (r)**				**Household chose S**	
Scenario for fee c:	*A*		*B*		*A*	*B*
	(1)	(2)	(3)	(4)	(5)	(6)
Effective fee (λ_2)	-0.031	-0.034	-0.034	-0.038	-0.031	-0.049
	(0.014)	(0.014)	(0.018)	(0.020)	(0.017)	(0.024)
	[0.03]	[0.02]	[0.06]	[0.06]	[0.06]	[0.04]
Higher quality (λ_1)		-0.040		-0.040		
		(0.020)		(0.019)		
		[0.04]		[0.04]		
λ_0	0.812	0.833	0.786	0.803		
	(0.057)	(0.058)	(0.050)	(0.056)		
	[0.00]	[0.00]	[0.00]	[0.00]		
$\tilde{c}^*$	-6.033	-4.833	-6.263	-5.209	.	.
$\alpha_0 + \Delta_\gamma$	26.109	24.159	22.947	21.269	.	.
α_1	.	-1.145	.	-1.049	.	.
$\bar{\theta} + \Delta_\gamma$	32.143	28.992	29.211	26.478	.	.
Observations	428	428	428	428	3723	3716
Measurements	0, 1, 3, 5	0, 1, 3, 5	0, 1, 3, 5	0, 1, 3, 5	0, 1, 3, 5	0, 1, 3, 5
Level of analysis	Slum	Slum	Slum	Slum	Household	Household

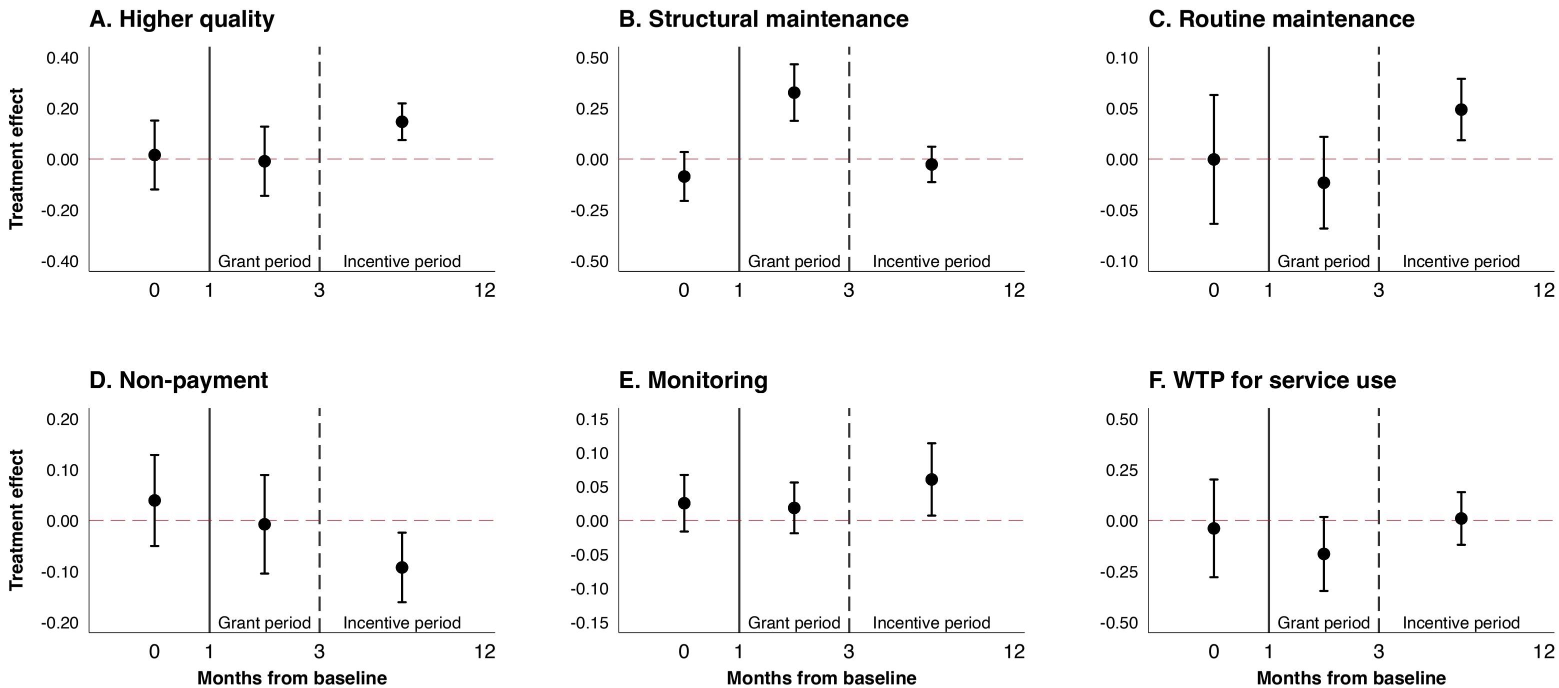
Note. In columns (1)–(4), estimates are based on fixed effects regressions using equation (8) and standard errors (reported in parentheses) are computed using stratified bootstrap with 1000 repetitions. The dependent variable is the share of households in the slum that reported not to have practiced OD at the time t of the interview. Columns (1)–(2) and (5) use an effective fee computed assuming that receiving the tickets in the WTP games corresponds to setting c to zero. Columns (3)–(4) and (6) use an effective fee computed assuming that receiving the tickets in the WTP games corresponds to a discount in c that is calculated assuming a daily use of the facility. In columns (5)–(6), estimates are based on a 2SLS regression and standard errors are clustered by catchment area–round. P-values are presented in brackets. The dependent variable is an indicator variable equal to 1 if the household reports not to have practiced OD at the time t of the interview, and 0 otherwise. All specifications include indicator variables for data collection rounds. In column (4), controls also include strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A. The effective fee is instrumented with the random number extracted as part of the WTP game. $\tilde{c}^*$ is the effective fee that set free riding to zero (assuming either $\lambda_1 = 0$ or a lower quality of service delivery).

Figure 1: Non-payment and willingness to pay for CT use, at baseline



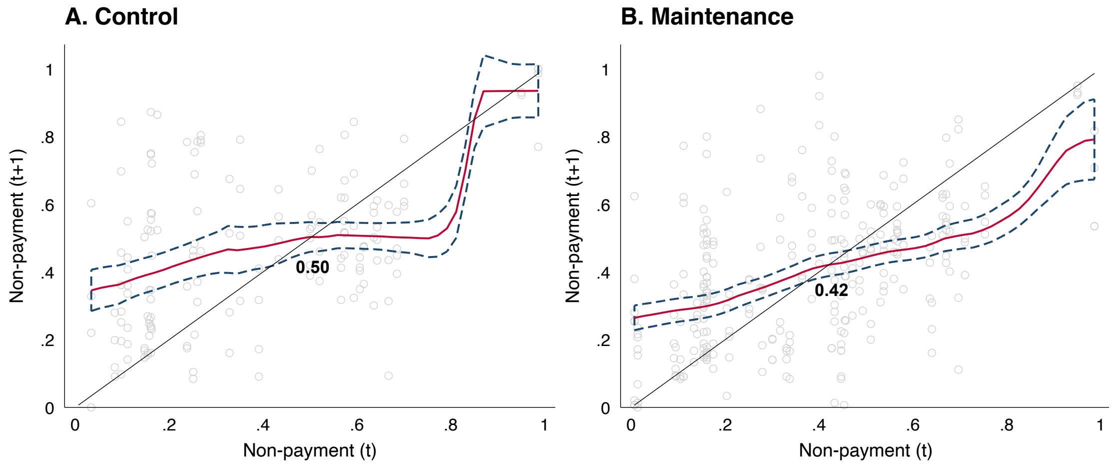
Note. Data collected at baseline. Panel A reports the share of users who do not pay the fee for the use of the CT during 1 hour at dawn, measured by observers. Panel B shows the distribution of the WTP for a single use of the service among study participants, measured using the incentivized elicitation of WTP. The distribution is censored at INR 5, the most common market price for a single CT use. The solid vertical lines represent the sample median, and the dashed vertical lines represent the sample mean. Additional details about the variables are presented in Appendix A. Supplementary Material S.4 provides details about measurement.

Figure 2: The timing of the maintenance treatment effects: grant versus incentive periods



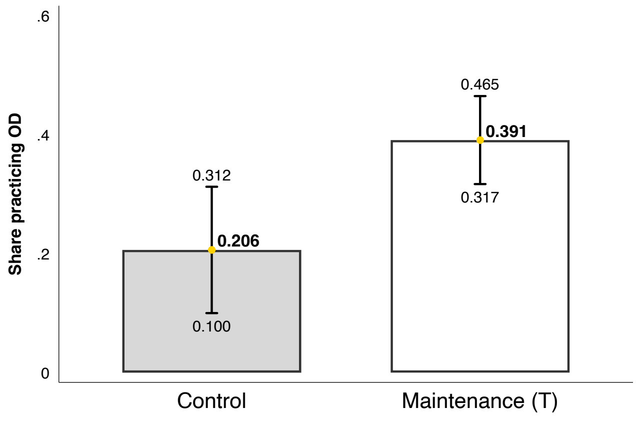
Note. Each panel presents estimates of treatment effects based on OLS regressions using equation (1) at the CT level or at the respondent level. Confidence intervals are built using statistical significance at the 10 percent level. *Baseline* includes the measurement at baseline, *Grant period* includes the measurement from follow-up 1, and *Incentive period* pools all subsequent follow-up measurements. See Section 1 for details about each intervention. When the regression is based on a single measurement period, robust standard errors are used for CT-level outcomes, and standard errors clustered at the catchment area are used for respondent level outcomes. When multiple measurement periods are pooled, standard errors are clustered at the catchment area for CT-level outcomes, and at the catchment area by collection round for respondent-level outcomes. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. For outcomes at the respondent-level, gender is also included as a control variable. Additional details about the variables are presented in Appendix A. Estimates pooling all post-baseline measurements are presented in Tables 1–3.

Figure 3: Local polynomial estimate of the transition equation for non-payment



Note. Non-payment is defined as the observed share of users who do not pay the entry fee to use the public service. The solid red line plots the smoothed values of a local polynomial regression, after controlling for strata indicators for the city and the provider of the CT, and caretakers'pro-social motivation for the cause. They area between the dashed blue lines depicts the 95 percent confidence intervals. The solid black line represents the 45 degree line at which non-payment in t equal non-payment in $t+1$. t is the grant scheme period and $t+1$ is the incentive period. The dataset is built stacking follow-up rounds 2 to 5. Standard errors are clustered by catchment area. The stable steady state is 0.42 for the maintenance treatment and 0.50 for the control.

Figure 4: Environmental contamination driven by OD



Note. The figure shows the share of study participants practicing OD in the day previous to the interview. We estimate shares using a list randomization technique., in which shares are the difference in the number of items reported by respondents who faced a long list of itimes (which includes OD), and the respondents who faced the short list (which excludes OD). We compute this average separately in the control group and in the maintenance treatment group. Confidence intervals are built using statistical significance at the 10 percent level and assuming errors are clustered at the level of the catchment area. Randomization of lists was performed at individual level, and data were collected during follow-up 5. Supplementary Material S.4 provides details about measurement.

ONLINE APPENDIX

Public Service Delivery and Free Riding: Experimental Evidence from India

Alex Armand, Britta Augsburg, Antonella Bancalari

A Definition of variables

Variable	Description
Higher quality	Indicator variable equal to 1 if the quality index is above the 75th percentile, and 0 otherwise. Quality index computed aggregating indicator variables about the structural quality of the facility, its cleanliness and the lack of bacteria. The index is re-scaled to be between zero and one. The variable aggregate survey responses from the CT survey, data from observers, and data from laboratory tests.
Structural maintenance	Indicator variable equal to 1 if facility received structural maintenance (repairs and/or deep cleaning intervention) in the month previous to the visit, and 0 otherwise. The variable aggregates responses from the CT survey, and administrative data from the implementing team.
Routine maintenance	Number of tools, equipment and cleaners used during the last routine maintenance for the CT. The variable aggregate survey responses from the CT survey. The number is normalized to be between 0, indicating that no tools reported in the questionnaire were used, and 1, indicating that all tools reported in the questionnaire were used. Tools include broom, mop, and safety equipment. Liquid tools include water, pressurized water and disinfectants. The baseline survey asks for information only on use of the broom, and disinfectants, while the full list is available for the following rounds of follow-ups.
Hours worked	Number of hours worked by the caretaker, self-reported during each CT survey.
Caretaker's knowledge	Indicator variable equal to 1 if the caretaker knows the recommended practices for cleaning routine and the need for deep cleaning, and 0 otherwise. The variable evaluates the correctness of questions about routine maintenance. These questions are asked during each CT survey.
Users	Total number of users entering the CT (reported in logarithms). The variable uses data from observers.
Non-payment	Share of users who do not pay the entry fee. The variable uses data from observers.
Enforcement	Share of surveyed households in the catchment area that reported having been refused entry to the CT by the caretaker for not paying the fee.
Monitoring	Share of worked hours allocated by the caretaker to collecting fees and supervising cleaners. Alternative activities include conducting repairs, cleaning the facility, and meeting the manager. The variable is self-reported by the caretaker during each CT survey.
WTP for service use	Willingness to pay for a single CT use (in INR). The variable is incentivized and elicited for a bundle of ten tickets, and is collected for both the household head and any partner separately during the household survey. We divide the WTP for the bundle by 10 to get at measure of single use WTP.
Awareness of externality	Indicator variable equal to 1 if the respondent reports that OD generates a health externality for their family, and 0 otherwise. The variable is self-reported by the household head during the household survey.
Demand for public intervention	*Service* is an indicator variable equal to 1 if the household asks for public intervention in the CT's O&M, and 0 otherwise. *Free riding* is an indicator variable equal to 1 if the household asks for public intervention to keep the community OD free, and 0 otherwise. *Other issues* is an indicator variable equal to 1 if the household asks for public intervention in categories unrelated with sanitation (i.e. children are frequently ill, water availability is limited, the quality of roads is poor, there is no waste collection, jobs are missing, access to healthcare is limited, and lighting at night is poor), and 0 otherwise. Since up to two participants per household could participate, the indicator variable is equal to 1 if any household member asks for public intervention in the CT's O&M. The information is incentivized and collected during the SCA voice-to-the-people initiative at the time of follow-up 3.
Pro-social motivation	Share of the endowment donated by the caretaker in the adapted dictator game. The variable is incentivized and is measured for each caretaker.
Preference for improvement	Share of the endowment that is donated by the slum resident in the adapted dictator game. The variable is incentivized and is measured for both the household head and the partner separately during the household survey.

(continued on next page)

Variable	Description
PGG contribution	Share contributed by the participant in the public good game. The variable is incentivized and is measured at the end of all the activities of the experiment.
Morbidity	Indicator variable equal to 1 if any household member had fever, diarrhea or cough during the two weeks previous to the interview, and 0 otherwise. The variable is self-reported by the household head during the household survey.
Health expenditures	We consider *curative expenditures*, which include costs associated with doctor visits when the person is ill, with the purchase of medicine, with hospitalization, and with x-rays, and include travel costs associated with these expenses, and *preventive expenditures*, which include all expenses associated with regular doctor checks, vaccines, anti-worm tablets, bed-nets, and prenatal tests, and travel costs associated with these expenses. *Extensive* (margin) is an indicator variable equal to 1 if the respondent had positive expenditures, and 0 otherwise. *Intensive* (margin) is the level of expenditures (in rupees). The variable is self-reported by the household head during the household survey, but is not collected during follow-up 1.
Open defecation	Aggregate share of study participants who practiced open defecation the day before the interview. Data are obtained using the list randomization technique. Information is obtained from both the household head and any partner in conjunction with follow-up 5.
CT use	Aggregate share of study participants who used the CT the day before the interview. Data are obtained using the list randomization technique. Information is obtained from both the household head and any partner in conjunction with follow-up 5.
Hand-washing with soap	Aggregate share of study participants who washed their hands with soap after defecating on the day before the interview. Data are obtained from using the list randomization technique. Information is obtained from both the household head and any partner in conjunction with follow-up 5.
Implementation	
Transfer to the CT	Transfer provided to the CT in the corresponding period as part of the intervention (in thousands of rupees). This includes the value of the initial grant to treated CTs, the subsidized use of tickets from the WTP game to both treated and control CTs, and the value of products bought with the transfer from study participants as part of the adapted dictator game to both treated and control CTs. Information is based on administrative data from the implementing team.
Transfer to the caretaker	Transfer provided to the caretaker in the corresponding period as part of the intervention (in thousands of rupees). This includes the financial incentive provided in treated CTs and the amounts kept from each round of the adapted dictator game. Information is based on administrative data from the implementing team.
Interactive activities	Indicator variable equal to 1 if the respondent is aware of any personal or community activities about WASH, and 0 otherwise. The variable is self-reported by the household head during the household survey.
Posters at CT	Indicator variable equal to 1 if the respondent is aware of any messages about WASH posted in the CT, and 0 otherwise. The variable is self-reported by the household head during the household survey.
Voice messages	Proportion of voice messages about WASH and CTs listened to by the participant. The variable is built from administrative data derived from the implementation of voice messages as part of the sensitization campaign.

Note. Supplementary Material S.4 provides further details about the measurement of these variables.

B Additional analysis

B.1 Balance in observable characteristics and attrition

Tables B1 and B2 present the balance test for characteristics at baseline.

Table B1: CT characteristics at baseline, by treatment group

	Descriptive statistics		**Differences from control group, by treatment group**			
	All	Control	Any treatment	Maintenance	Maintenance + sensitization	p-value joint test (4)-(5)
	(1)	(2)	(3)	(4)	(5)	(6)
Year of construction	1996.98	1995.26	2.78	2.34	3.23	0.32
	[8.85]	[9.29]	(1.88)	(2.11)	(2.19)	
Distance to closest CT	0.54	0.58	-0.06	-0.04	-0.07	0.76
	[0.44]	[0.66]	(0.11)	(0.11)	(0.11)	
Surrounding market	0.33	0.35	-0.04	-0.01	-0.06	0.82
	[0.47]	[0.48]	(0.10)	(0.11)	(0.11)	
Surrounding road	0.84	0.88	-0.06	-0.05	-0.08	0.67
	[0.37]	[0.33]	(0.07)	(0.09)	(0.09)	
Surrounding government office	0.25	0.20	0.07	0.08	0.06	0.69
	[0.43]	[0.41]	(0.08)	(0.10)	(0.10)	
Only residents use CT	0.12	0.07	0.07	0.07	0.07	0.53
	[0.32]	[0.27]	(0.06)	(0.07)	(0.07)	
Single caretaker	0.80	0.82	-0.04	0.03	-0.11	0.28
	[0.40]	[0.39]	(0.07)	(0.08)	(0.09)	
Share of female caretakers	0.18	0.22	-0.06	-0.02	-0.10	0.42
	[0.37]	[0.39]	(0.07)	(0.08)	(0.08)	
Caretaker is also cleaner	0.27	0.28	-0.02	-0.02	-0.03	0.96
	[0.45]	[0.46]	(0.09)	(0.10)	(0.10)	
Caretaker is from local community	0.44	0.49	-0.07	-0.11	-0.02	0.60
	[0.50]	[0.51]	(0.10)	(0.12)	(0.12)	
Caretaker's experience (months)	125.28	129.91	-5.43	1.37	-11.53	0.86
	[103.45]	[109.34]	(22.81)	(26.60)	(25.96)	
CT is cleaned frequently	0.86	0.87	-0.02	-0.02	-0.02	0.97
	[0.35]	[0.34]	(0.07)	(0.08)	(0.08)	
Time allocated to managing	0.68	0.66	0.03	0.03	0.02	0.58
	[0.14]	[0.11]	(0.03)	(0.03)	(0.03)	
Capacity	13.00	13.21	-0.32	-0.46	-0.17	0.94
	[5.57]	[5.52]	(1.11)	(1.27)	(1.34)	
Daily opening hours	17.76	17.88	-0.19	-0.35	-0.02	0.53
	[1.49]	[1.59]	(0.28)	(0.36)	(0.27)	
Share of functioning toilets	0.90	0.88	0.03	0.05	0.01	0.47
	[0.22]	[0.23]	(0.04)	(0.04)	(0.05)	
WTP (avg. catchment area)	1.41	1.44	-0.05	-0.03	-0.06	0.95
	[0.83]	[0.65]	(0.15)	(0.17)	(0.20)	
Distance from CT (avg. catchment area)	128.71	128.77	-0.01	-2.22	2.21	0.94
	[49.56]	[43.87]	(9.26)	(10.21)	(12.25)	

Note. Columns (1) and (2) report sample mean with standard deviation in brackets for the whole sample and for the control group, respectively. Column (3) reports the difference from the control group with any treatment group. Columns (4) and (5) report the difference from the control group for each treatment group. Differences in columns (3)–(5) are estimated using OLS and controlling for strata indicators for the city and the provider of the CT. Robust standard errors are reported in parentheses. Column (6) presents a joint test of significance of the coefficients for each treatment dummy. Statistical significance denoted by *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Concerning attrition, we limit our analysis at verifying whether attrition created differences across treatment arms, while details about the collection of the CT survey and of the slum resident surveys are provided in Supplementary Material S.4.2 and S.4.4. For the CT survey, Table B4 shows estimates of treatment effects on the number of CT observations and caretaker surveys and the opening of new CTs in study catchment

Table B2: Household characteristics at baseline, by treatment group

	Descriptive statistics		**Differences from control group, by treatment group**			
	All	Control	Maintenance	Maintenance only	Maintenance + sensitiza-tion	p-value joint test (4)-(5)
	(1)	(2)	(3)	(4)	(5)	(6)
Household head is male	0.75	0.73	0.02	0.04	0.01	0.30
	[0.43]	[0.44]	(0.02)	(0.03)	(0.03)	
Household head is married	0.77	0.76	0.01	0.01	0.01	0.93
	[0.42]	[0.43]	(0.03)	(0.03)	(0.03)	
Age of household head	45.44	46.04	-0.87	-0.89	-0.84	0.55
	[12.82]	[13.42]	(0.80)	(0.97)	(0.86)	
Age of spouse	39.14	39.39	-0.33	-0.75	0.07	0.61
	[11.39]	[12.00]	(0.76)	(0.94)	(0.79)	
Household head has no education	0.54	0.55	-0.02	-0.07	0.03	0.05
	[0.50]	[0.50]	(0.04)	(0.05)	(0.04)	
Spouse has no education	0.45	0.45	-0.00	0.01	-0.01	0.91
	[0.50]	[0.50]	(0.03)	(0.04)	(0.04)	
Household members	4.94	4.94	0.00	0.01	-0.00	1.00
	[1.99]	[2.08]	(0.13)	(0.15)	(0.14)	
Household members (0-5 y.o.)	0.47	0.50	-0.06	-0.05	-0.07	0.59
	[0.77]	[0.82]	(0.06)	(0.06)	(0.07)	
Household members (older than 5 y.o.)	4.47	4.44	0.06	0.05	0.07	0.85
	[1.83]	[1.92]	(0.11)	(0.13)	(0.12)	
Muslim	0.17	0.12	0.09**	0.11*	0.06	0.12
	[0.37]	[0.32]	(0.04)	(0.06)	(0.04)	
Spent on religious items	0.94	0.94	-0.01	-0.01	-0.00	0.84
	[0.25]	[0.24]	(0.01)	(0.02)	(0.02)	
General caste	0.07	0.05	0.03	0.03	0.02	0.25
	[0.26]	[0.23]	(0.02)	(0.02)	(0.02)	
Asset index	0.53	0.53	0.00	0.01	-0.00	0.77
	[0.15]	[0.16]	(0.02)	(0.02)	(0.02)	
Household members per room	3.99	3.90	0.14	0.05	0.21	0.31
	[1.86]	[1.94]	(0.14)	(0.16)	(0.15)	
Access to piped water	0.71	0.70	0.01	-0.01	0.04	0.67
	[0.45]	[0.46]	(0.05)	(0.06)	(0.06)	
Access to private toilet	0.08	0.07	0.01	0.01	0.02	0.67
	[0.27]	[0.26]	(0.02)	(0.02)	(0.02)	
Expenditure on CT use (INR)	180.75	173.72	11.09	-2.37	23.85	0.65
	[244.60]	[221.49]	(23.04)	(23.00)	(30.62)	
Prevalence of diarrhea (last 15 days)	0.08	0.07	0.02	0.01	0.03	0.25
	[0.28]	[0.26]	(0.02)	(0.02)	(0.02)	
Prevalence of fever (last 15 days)	0.18	0.18	-0.01	-0.01	-0.01	0.89
	[0.38]	[0.39]	(0.02)	(0.03)	(0.03)	
Distance to CT (meters)	126.22	126.42	-1.08	-2.09	-0.12	0.97
	[79.90]	[80.42]	(8.74)	(9.63)	(11.55)	

Note. Columns (1) and (2) report sample mean with standard deviation in brackets for the whole sample and for the control group, respectively. Column (3) reports the difference from the control group with any treatment group. Columns (4) and (5) report the difference from the control group for each treatment group. Differences in columns (3)–(5) are estimated using OLS and controlling for strata indicators for the city and the provider of the CT. Standard errors clustered at slum level are reported in parentheses. Column (6) presents a joint test of significance of the coefficients for each treatment dummy. Statistical significance denoted by *** p<0.01, ** p<0.05, * p<0.1.

areas. For all outcomes, we do not observe any significant impact of the interventions. Concerning the slum resident survey, columns (1)–(5) in Table B5 estimates the probability of attrition as a function of the treatment status. Attrition does not differ between treatment and control groups for any of the attrition indicators. In order to maintain a comparable sample size in all follow-up surveys, we handled attrition with replacements at random using the sampling frame used for the baseline sampling. Column (6) tests whether the replacement was introduced differently across treatment arms, showing no statistical difference across treatment arms.

Table B3: Selection in sanitation behavior, by baseline characteristics

Dep. variable:	**Stopped using the service**		**Continued to OD**	
Sub-sample:	Maintenance treatment	Control	Maintenance treatment	Control
	(1)	(2)	(3)	(4)
Household head is male	-0.154***	-0.227***	0.159	0.121
	(0.056)	(0.056)	(0.127)	(0.159)
Household head is married	0.093	0.127	-0.225	-0.401**
	(0.074)	(0.078)	(0.153)	(0.145)
Age of household head	-0.000	0.002	-0.004	-0.012**
	(0.002)	(0.002)	(0.003)	(0.005)
Household head has no education	0.037	-0.049	-0.011	0.110
	(0.043)	(0.049)	(0.094)	(0.141)
Spouse has no education	-0.000	-0.044	0.013	0.328***
	(0.041)	(0.053)	(0.091)	(0.107)
Household members	0.016	-0.008	0.046	0.063
	(0.014)	(0.015)	(0.028)	(0.038)
Muslim	0.011	-0.006	0.062	-0.118
	(0.061)	(0.069)	(0.102)	(0.214)
Spent on religious items	0.020	-0.028	0.123	0.078
	(0.090)	(0.089)	(0.164)	(0.189)
General caste	0.014	0.081	-0.160	0.102
	(0.082)	(0.118)	(0.238)	(0.425)
Asset index	-0.529***	-0.599***	-1.138***	-0.586*
	(0.168)	(0.177)	(0.246)	(0.332)
Household members per room	-0.038**	-0.012	-0.042	-0.067*
	(0.015)	(0.019)	(0.029)	(0.037)
Access to piped water	-0.018	-0.005	0.027	-0.091
	(0.058)	(0.072)	(0.092)	(0.126)
Access to private toilet	0.168*	0.248*	0.042	-0.119
	(0.084)	(0.132)	(0.164)	(0.205)
Expenditure on CT use (INR)	-0.000***	-0.000***	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Distance to CT (meters)	0.001***	0.001**	0.000	-0.000
	(0.000)	(0.000)	(0.001)	(0.001)
WTP for CT use	-0.016	0.002	-0.008	0.004
	(0.010)	(0.010)	(0.017)	(0.028)
Preference for CT clean	0.000	0.002	-0.005	0.002
	(0.002)	(0.002)	(0.004)	(0.007)
Awareness of externalities	0.025	-0.080	-0.000	0.124
	(0.041)	(0.057)	(0.082)	(0.109)
Morbidity	-0.003	-0.084	0.107	0.066
	(0.040)	(0.060)	(0.104)	(0.108)
Health expenditure (extensive)	-0.065	-0.039	-0.117	0.006
	(0.045)	(0.053)	(0.078)	(0.190)
Observations	658	420	175	103

Note. Dependent variables by column: (1) *Stopped using the service*: indicator variable equal to 1 if used the CT at baseline and stopped using it at endline, and equal to 0 if continued using CT at endline; (2) *Continued to OD*: indicator variable equal to 1 if at least one member of the household practices OD at baseline and endline, and equal 0 if practiced OD only at baseline. Sample restricted to households interviewed at baseline. All specifications include strata indicators for the city and the provider of the CT. Standard errors clustered at slum level are reported in parentheses. Statistical significance denoted by *** p<0.01, ** p<0.05, * p<0.1.

Table B4: Attrition in CT observations and caretaker surveys

	Observations collected FU rounds (1)	**Caretaker surveyed** FU rounds (2)	**New CT in catchment area** (3)
Panel A			
Maintenance (T)	0.100	0.125	0.014
	(0.100)	(0.135)	(0.010)
	[0.32]	[0.36]	[0.16]
Panel B			
Maintenance only (T1)	0.100	0.052	0.028
	(0.100)	(0.170)	(0.020)
	[0.32]	[0.76]	[0.15]
Maintenance + sensitization (T2)	0.100	0.198	0.001
	(0.100)	(0.124)	(0.002)
	[0.32]	[0.11]	[0.82]
T1 = T2 (p-value)	0.942	0.238	0.153
Mean (control group)	4.900	4.775	0.000
Catchment areas	110	110	110

Note. Estimates based on CT-level OLS regressions using equation (1) in panel A, and equation (2) in panel B. Robust standard errors in parentheses. p-values are presented in brackets. Dependent variables by column: (1) *Observations collected*: number of follow-up surveys where CT observation were collected; (2) *Caretaker surveyed*: number of follow-up surveys where the CT caretaker was surveyed; (3) *New CT in catchment area*: indicator variable equal to 1 if an additional CT opened in the same catchment area and hence was included in the study later on, and 0 otherwise. All specifications include strata indicators for the city and the provider of the CT.

Table B5: Attrition in the household measurements

	Follow-up interviews per baseline household (1)	**Interviewed at baseline and not re-interviewed in ...** Any follow-up (2)	Follow-up 1 (3)	Follow-up 3 (4)	Follow-up 5 (5)	**Household is a re-placement** (6)
Maintenance only (T1)	0.025	0.006	0.013	-0.024	-0.013	0.001
	(0.074)	(0.011)	(0.025)	(0.037)	(0.035)	(0.015)
	[0.74]	[0.61]	[0.62]	[0.52]	[0.70]	[0.97]
Maintenance + sensitization (T2)	0.009	0.012	0.012	-0.013	-0.008	-0.007
	(0.079)	(0.014)	(0.023)	(0.042)	(0.034)	(0.014)
	[0.91]	[0.38]	[0.59]	[0.75]	[0.81]	[0.62]
T1 = T2 (p-value)	0.824	0.661	0.985	0.738	0.867	0.616
Attrition rate	2.525	0.027	0.110	0.209	0.156	0.171
Observations	1555	1555	1555	1555	1555	6614

Note. Figure S.2.1 provides the timing of each follow-up survey. Dependent variables by column: (1) indicator variable equal to 1 if the household was interviewed at baseline and was not re-interviewed in any of the follow-ups, and zero otherwise; (2) indicator variable equal to 1 if the household was interviewed at baseline and was not re-interviewed in two out of three follow-ups, and 0 otherwise; (3)–(5) indicator variable equal to 1 if the household was interviewed at baseline and was not re-interviewed at follow-up 1 or follow-up 2 or follow-up 3, and 0 otherwise; (6) indicator variable equal to 1 if the household is part of the replacement sample (it was interviewed in any of the follow-ups, but it was not interviewed at baseline), and 0 otherwise. In columns (1)–(5), the sample is restricted to baseline observations, while in column (6) the sample is restricted to follow-up observations. All specifications include strata indicators for the city and the provider of the CT. Standard errors clustered by catchment area are presented in parenthesis in columns (1)–(5). Standard errors clustered by catchment area and follow-up round are presented in parenthesis in column (6).

B.2 Implementation of interventions across treatment groups

Table B6 shows the effect of treatments on indicators of exposure to the interventions. We focus on transfers as part of the maintenance intervention, and of indicators of the sensitization campaign.

Table B6: Exposure to the interventions, by component

	Maintenance		Sensitization campaign		
	Transfer to the ...		Recall of WASH campaign		Voice messages
	CT	Caretaker	Interactive activities	Posters at CT	Exposure
	(1)	(2)	(3)	(4)	(5)
Panel A					
Maintenance (T)	4.739	0.761	0.054	0.089	0.403
	(0.060)	(0.034)	(0.020)	(0.026)	(0.059)
	[0.00]	[0.00]	[0.01]	[0.00]	[0.00]
Panel B					
Maintenance only (T1)	4.645	0.746	0.023	0.017	-0.038
	(0.081)	(0.045)	(0.024)	(0.030)	(0.047)
	[0.00]	[0.00]	[0.33]	[0.58]	[0.42]
Maintenance + sensitization (T2)	4.839	0.776	0.083	0.158	0.827
	(0.074)	(0.047)	(0.023)	(0.029)	(0.086)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]
T1 = T2 (p-value)	0.063	0.636	0.014	0.000	0.000
Mean (control group)	0.315	0.063	0.645	0.327	0.188
Std. dev. (control group)	0.358	0.025	0.479	0.469	0.347
Observations	560	560	4793	3301	4793
Catchment areas	110	110	328	218	328
Observation rounds	5	5	3	2	3

Note. In columns (1) and (2), estimates are based on CT-level OLS regressions using equation (1) in panel A, and equation (2) in panel B. Standard errors clustered by catchment area are reported in parentheses. Transfers are reported in thousands of INR. In columns (3)–(8), estimates are based on household-level OLS regressions using equation (1) in panel A, and equation (2) in panel B. Standard errors clustered by catchment area–round are reported in parentheses. p-values are presented in brackets, the first from individual testing, the second adjusting for jointly testing that each treatment is different from zero for all outcomes presented in the table. See Section 5 for details. Dependent variables are reported in the column header and are defined in Appendix A. Dependent variables by column: (1) *Transfer to the CT*: total transfers provided to the CT as part of the intervention (in thousands of Rupees); (2) *Transfer to the caretaker*: total transfers provided to the caretaker as part of the intervention (in thousands of Rupees); (3) *Interactive activities*: is an indicator variable equal to one if the respondent participated or is aware of a water, sanitation and hygiene interactive activities, and zero otherwise; (4) *Posters at CT*: is an indicator variable equal to one if the respondent has seen or is aware of posters promoting safe sanitation behaviour placed in the CT, and zero otherwise; (5) *Exposure to voice messages*: Share of voice message listened multiplied by the number of words in the message related to water, sanitation and hygiene. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT.

B.3 Treatment effects on revenues

Table B7 provides estimates of treatment effects on monthly revenues estimated using observation during the rush hour. Revenues are imputed using information from observers about the number of people using the CT and the share of them who are is paying the fee (assuming a standard fee of INR 5).

Table B7: Service revenues

	Monthly revenues			
	All periods		Incentive period	
	Extensive	Intensive	Extensive	Intensive
	(1)	(2)	(3)	(4)
Panel A				
Maintenance (T)	0.033	267.073	0.033	267.073
	(0.033)	(254.664)	(0.033)	(254.664)
	[0.32]	[0.29]	[0.32]	[0.29]
Panel B				
Maintenance only (T1)	0.020	127.399	0.020	127.399
	(0.041)	(293.467)	(0.041)	(293.467)
	[0.62]	[0.66]	[0.62]	[0.66]
Maintenance + sensitization (T2)	0.045	407.935	0.045	407.935
	(0.029)	(290.198)	(0.029)	(290.198)
	[0.13]	[0.16]	[0.13]	[0.16]
T1 = T2 (p-value)	0.379	0.330	0.379	0.330
Mean (control group)	0.948	3027.202	0.948	3027.202
Std. dev. (control group)	0.222	1870.703	0.222	1870.703
Observations	542	542	542	542
Catchment areas	110	110	110	110
Observation rounds	5	5	5	5
Follow-ups	1–5	1–5	2–5	2–5

Note. In columns (1), (3), (5) and (7), estimates are based on CT-level OLS regressions using equation (1) in panel A, and equation (2) in panel B. In columns (2), (4), (6) and (8), estimates are based on CT-level tobit regressions using equation (1) in panel A, and equation (2) in panel B, and imposing censoring at zero. Standard errors clustered by catchment area are reported in parentheses. p-values are presented in brackets. See Section 5 for details. Dependent variables are reported in columns. *Extensive* is an indicator variable equal to 1 if the revenues are larger than zero, and 0 otherwise. *Intensive* is the revenues reported in levels. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A.

B.4 Effect on the quality of the service

To construct a measure capturing the overall **quality of service delivery**, we use all the observed indicators related to the facility's structural quality and cleanliness, and to the lack of harmful bacteria. Since quality is unobserved and multidimensional and varies over time, we build the index using item response theory (IRT), a technique used to describe the relationship between individual responses to questionnaire items and an unobserved latent trait (Gordon et al., 2012; Kline, 2014). We build the index using a two parameter IRT model with the two parameters being an ability score, which could be used as a weight in constructing the index, and a discrimination score, which measures how well the indicator differentiates between low- and high-quality. The index is re-scaled to be between 0 (lowest quality) and 1 (highest).[2] Table B8 provides the list of all indicators included.

Figure B1 shows the distribution of the quality of service delivery index by treatment

[2] We compute the index separately for baseline and for all follow-up measurements due to the fact that the baseline survey includes a lower number of indicators. At baseline, due to convergence, we adopt a one parameter IRT model.

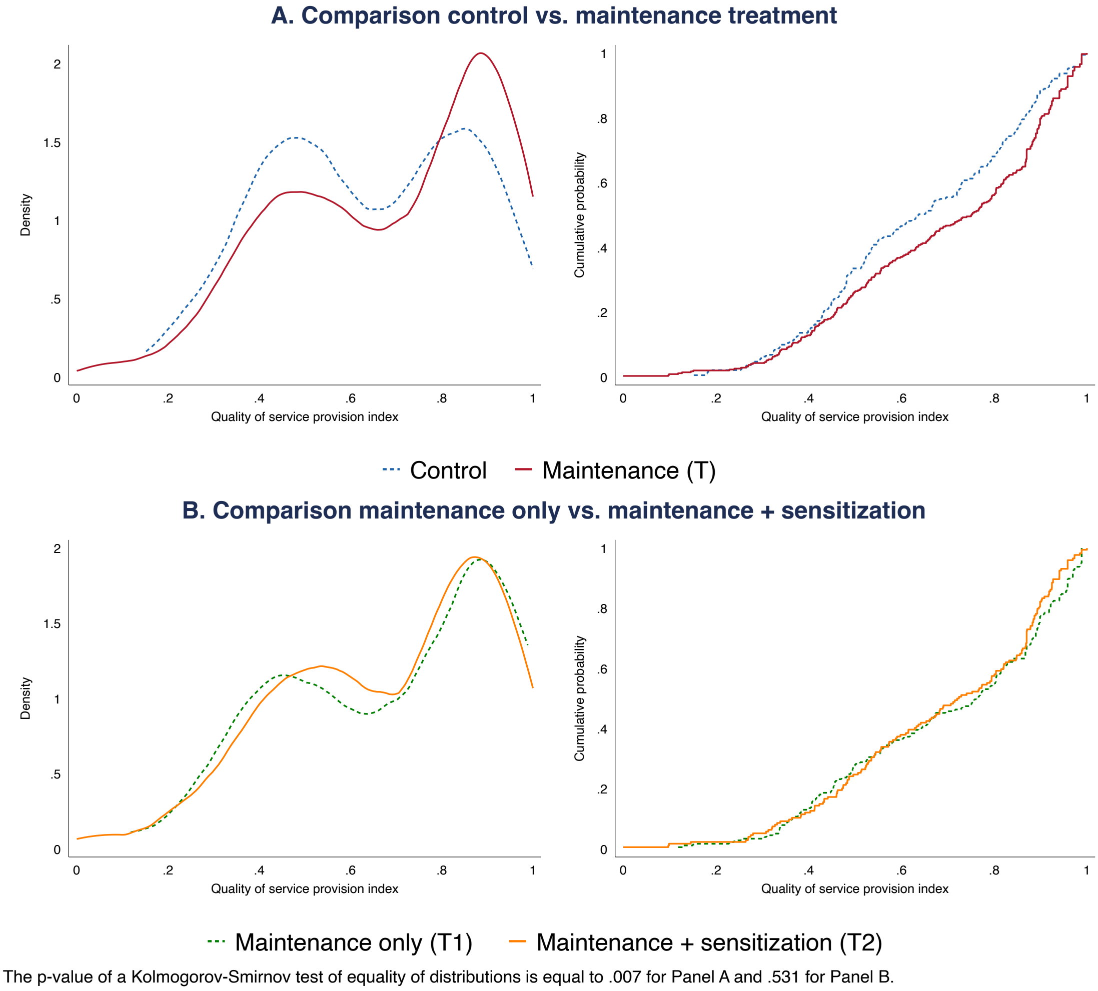
Table B8: Indicators used for the construction of the quality index

Indicator variables	Ability score (1)	Discrimination (2)
Structural quality		
All cubicle doors are functioning	1.971	0.247
All locks are functioning	-0.603	0.435
Compound has finished walls	2.259	0.412
Internal walls are in good condition	3.156	0.294
Soap is available and visible for both genders	1.731	0.572
Hand-washing facility available for both genders	1.667	0.811
Female area has lighting	1.842	1.002
Male area has lighting	1.751	1.059
Common area has lighting	2.960	0.762
Cleanliness		
Toilets in female area are not dirty	0.699	3.705
Toilets in female area do not stink	0.640	4.121
Flies not present in the female area	0.837	3.904
Toilets in male area are not dirty	0.570	4.843
Toilets in male area do not stink	0.771	3.431
Flies not present in the male area	0.525	5.990
Feces not visible inside the latrine in the female area	1.009	5.186
Feces not visible outside the latrine in the female area	1.200	4.523
Feces not visible inside the latrine in the male area	0.987	3.699
Feces not visible outside the latrine in the male area	1.192	3.134
Common area is not dirty	1.276	2.924
Common area does not stink	1.254	3.254
Flies not present in the common area	1.272	2.764
No visible sewage leaks inside the compound	2.449	2.235
Lack of bacteria		
Bacteria count of E. coli is low	-0.379	-0.196
Bacteria of bacillus are not detected	2.148	-3.145
Bacteria of staphylococcus are not detected	-25.405	-0.097
Bacteria of salmonella are not detected	38.091	0.025
Bacteria of klebsiella are not detected	10.820	-0.123
Mold is not detected	3.537	-0.455

Note. All indicator variables are equal to 1 if the condition is true, and 0 otherwise. The table reports the main parameters in the index build using IRT, with the ability score reported in column (1) and the discrimination reported in column (2). Observations are restricted to follow-ups 1–5 for computing the index. The manual for observers defines the rules for the visual evaluation of CTs (Supplementary Material S.4.2). *Finished walls* are defined as built in cement, and bricks, with no cracks or crumbles on the paintwork or tiles. *Dirt* is reported as the presence of mud, mold, red spitting, urine or feces on floors or walls. *Stink* is reported as the presence of an unpleasant smell from urine or feces. *Sewage leaks* are identified by fecally contaminated black waters leaking from a septic tank, pit/cesspool or pipes.

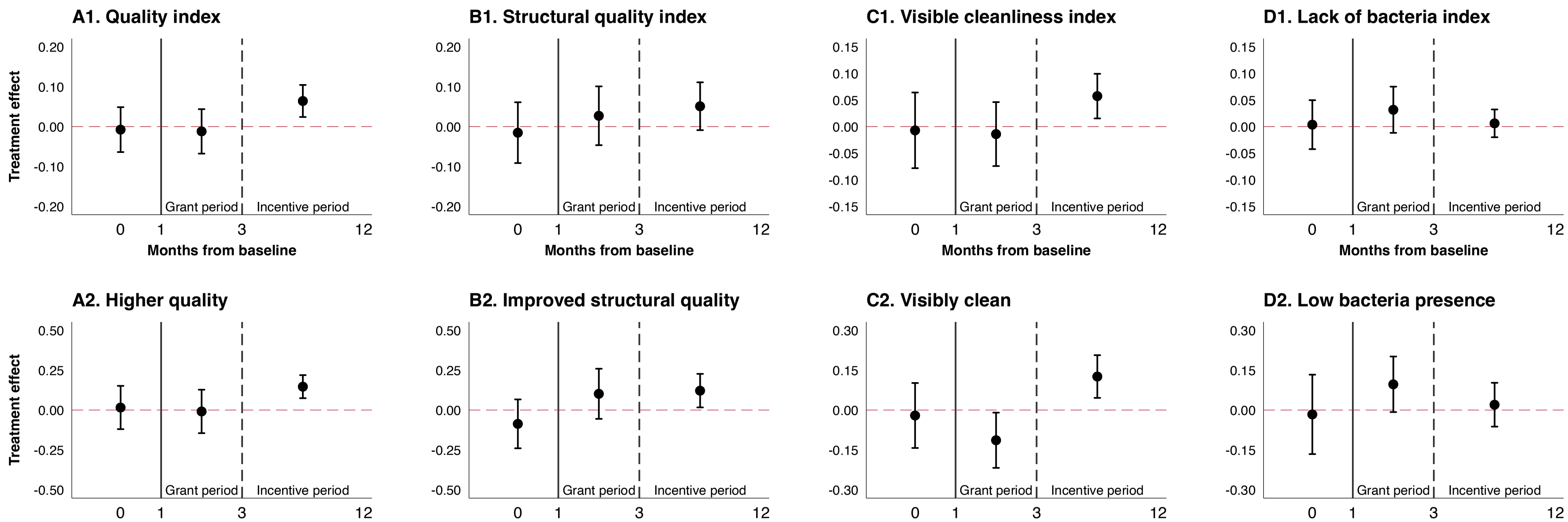
arm. Panel A compares the control group with the maintenance treatment group, while panel B restricts the sample to the maintenance treatment group and compares the maintenance only treatment group and the maintenance plus sensitization treatment group. The left figures present distribution fits, while the right figures show the empirical cumulative distribution functions. In addition to the quality of service delivery index, we build three separate indices using IRT to measure the structural quality of the facility, the cleanliness of the CT, and the lack of bacteria. Figure B2 shows the effect of the maintenance treatment on the oveall quality index and by component.

Figure B1: Distribution of the quality of service delivery index at follow-up



Note. The distributions include all follow-up measurements. The quality of service delivery index is built using a two parameter IRT model. Panel A shows the comparison between the control group and any treatment group. Panel B shows the comparison between the two treatment groups individually (maintenance and maintenance plus sensitization). The left figures show the distribution fits estimated non-parametrically using kernel density estimation assuming an Epanechnikov kernel function. Bandwidths are estimated by Silverman's rule of thumb (Silverman, 1986). The right figures show the empirical cumulative distribution functions.

Figure B2: Effect on CT quality by component: grant versus incentive period



Note. Each panel presents estimates of treatment effects based on OLS regressions using equation (1) at the CT level. Confidence intervals are built using statistical significance at the 10 percent level. *Baseline* includes the measurement at baseline, *Grant period* includes the measurement from follow-up 1, and *Incentive period* pools all subsequent follow-up measurements. See Section 1 for details about each intervention. When the regression is based on a single measurement period, robust standard errors are used. When multiple measurement periods are pooled, standard errors are clustered at the catchment area. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Supplementary Material S.4 provides details about measurement.

B.5 Spillover analysis

Table B9 shows a test for contagion or spillover effects by estimating heterogeneous treatment effects according to the average distance of a CT or catchment area to another treated CT or catchment area. We define a catchment area to be close to (far from) another treated catchment area if the distance is below or equal to (above) the sample median. Among all outcome variables, we do not observe any heterogeneous effect, suggesting the absence of spillover effects.

Table B9: Contagion and spillover effects

	Effect of any treatment, by average distance to another treatment						**Het. test**
	β	se	N	β	se	N	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Close to T			**Far from T**			
Higher quality	0.12*	0.06	266	0.12*	0.06	276	0.97
Structural maintenance	0.02	0.05	266	0.07	0.09	276	0.62
Routine maintenance	0.03	0.02	266	0.04	0.03	276	0.86
Hours worked	-0.25	0.40	266	0.66	0.59	276	0.20
Caretaker's knowledge	0.08***	0.03	266	0.09	0.06	276	0.95
Users	-0.04	0.07	265	-0.06	0.07	277	0.85
Non-payment	-0.08*	0.04	265	-0.07	0.07	277	0.89
Enforcement	0.03	0.02	54	-0.01	0.04	57	0.36
Monitoring	0.01	0.02	266	0.10*	0.05	276	0.13
WTP for service use	-0.07	0.12	4309	-0.01	0.08	4326	0.72
Awareness of externalities	0.06**	0.03	2425	0.01	0.02	2368	0.13
Demand for pub. int. (service)	0.05	0.04	797	0.05	0.03	754	0.92
Demand for pub. int. (free riding)	-0.10*	0.05	797	-0.04	0.05	754	0.42
Demand for pub. int. (other)	-0.01	0.04	797	-0.03	0.03	754	0.81
Morbidity	0.02	0.03	2425	-0.01	0.03	2368	0.42
Curative exp. (extensive)	0.07**	0.03	1677	0.02	0.04	1599	0.41
Curative exp. (intensive)	-254.09	264.40	1677	181.19	279.43	1599	0.27
Preventive exp. (extensive)	-0.00	0.00	1692	-0.00	0.00	1609	0.48
Preventive exp. (intensive)	11.24	78.12	1691	-1.04	82.00	1609	0.90

Note. Close to (*far from*) indicates whether the average distance is below or equal to (above) the sample median. Variables referring to catchment areas are averages of the corresponding variable within the catchment area. In columns (1)–(6), estimates are based on CT-, respondent- or household-level OLS regressions using equation (1) separately for each category. Column (7) presents a heterogeneity test based on CT-level OLS regressions using equation (1) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. Standard errors are clustered by catchment area for CT-level outcomes and by catchment-area–round for respondent- and household-level outcomes. The dependent variables are indicated in the rows and are defined in Appendix A. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1.

B.6 Robustness to the inclusion of control variables

Table B10 presents estimates of the effect of any treatment (T) using equation (1) in columns (1)–(3), and the post-double selection LASSO (PDSL) procedure (Tibshirani, 1996; Belloni et al., 2013) in columns (4)–(6). The PDSL procedure provides a method for model selection in the presence of a large number of control variables.

Table B10: Effect of any treatment: comparison between main estimates and PDSL

	No control variables			**Post-double selection LASSO**			
	β	se	p-value	β	se	p-value	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CT-/caretaker-level outcomes							
Higher quality	0.12	0.04	0.01	0.09	0.04	0.02	542
Structural maintenance	0.04	0.05	0.36	0.04	0.05	0.45	542
Routine maintenance	0.03	0.02	0.05	0.03	0.01	0.03	542
Hours worked	0.17	0.36	0.62	0.14	0.36	0.69	542
Caretaker's knowledge	0.09	0.04	0.01	0.09	0.03	0.01	542
Users	-0.06	0.05	0.24	-0.06	0.05	0.21	542
Non-payment	-0.08	0.04	0.07	-0.07	0.04	0.10	542
Enforcement	0.00	0.02	0.83	0.01	0.02	0.79	111
Monitoring	0.05	0.03	0.07	0.05	0.02	0.05	542
Respondent/household-level outcomes							
WTP for service use	-0.04	0.07	0.57	-0.03	0.07	0.65	8635
Awareness of externalities	0.03	0.02	0.10	0.03	0.02	0.09	4793
Demand for pub. int. (service)	0.05	0.03	0.05	0.05	0.03	0.04	1551
Demand for pub. int. (free riding)	-0.08	0.04	0.04	-0.08	0.04	0.03	1551
Demand for pub. int. (other)	-0.02	0.02	0.38	-0.02	0.02	0.49	1551
Household-level outcomes							
Morbidity	0.01	0.02	0.79	0.01	0.02	0.64	4793
Curative exp. (extensive)	0.05	0.02	0.07	0.05	0.02	0.06	3276
Curative exp. (intensive)	-35.15	194.68	0.86	-27.80	193.80	0.89	3276
Preventive exp. (extensive)	-0.00	0.00	0.51	-0.00	0.00	0.64	3301
Preventive exp. (intensive)	5.57	56.66	0.92	-7.57	55.60	0.89	3300

Note. Columns (1)–(3) show estimates using equation (1), while columns (4)–(6) show estimates using the PDSL procedure (Tibshirani, 1996; Belloni et al., 2013), with selection over a large number of baseline-level control variables. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. N indicates the sample size. In order to have the same sample size of estimates as in the main tables, missing values are replaced by the value 0 and an indicator variable equal to 1 if the observation had a missing value is introduced for all variables. To build the set of potential control variables, we include the following observable characteristics in the procedure (all continuous variables are also included in their squared term and are standardized): *CT characteristics* (variables describing the facility at baseline included in Table B1); *caretaker characteristics* (variables related to caretakers at baseline included in Table B1); *catchment area characteristics* (for CT- and caretaker-level outcomes, we include the catchment-area average at baseline for the household head's gender, education, marital status, religion and caste, WTP for service use, trust of the community, bacteria contamination in water sources, share practicing OD, and distance from the CT); *individual characteristics* (for household- and respondent-level outcomes, we include the baseline characteristics of the household and of the respondent included in Table B2); *outcome variables* (for CT- and caretaker-level estimates, we include the baseline value of outcomes presented in Tables 1–2, while for household- and respondent-level outcomes, the baseline values of outcomes are presented in Tables 4–3). Additional information about outcome variables is provided in Appendix A.

B.7 Robustness to estimation of treatment effects via causal forest

Table B11 presents estimates of ATE of any treatment on all outcome variables using the causal forest procedure of Athey et al. (2019) and following the cluster-robust procedure of Basu et al. (2018) and Athey and Wager (2019). In the procedure, we use the set of variables from Appendix B.6. Columns (1)–(3) present estimates of the ATE and the p-value of a two-sided test for the ATE being different from 0. To verify the overall presence of heterogeneity in the impacts, Columns (4)–(5) implement a calibration test based on the best linear predictor method of Chernozhukov et al. (2017). Column (4) presents the p-value for the equality to 1 of the coefficient on the mean forest prediction, with 1 indicating that the mean forest prediction is correct. Column (5) presents the p-value for the equality to 1 (no heterogeneity) of the coefficient on the quality of the estimates of treatment heterogeneity.

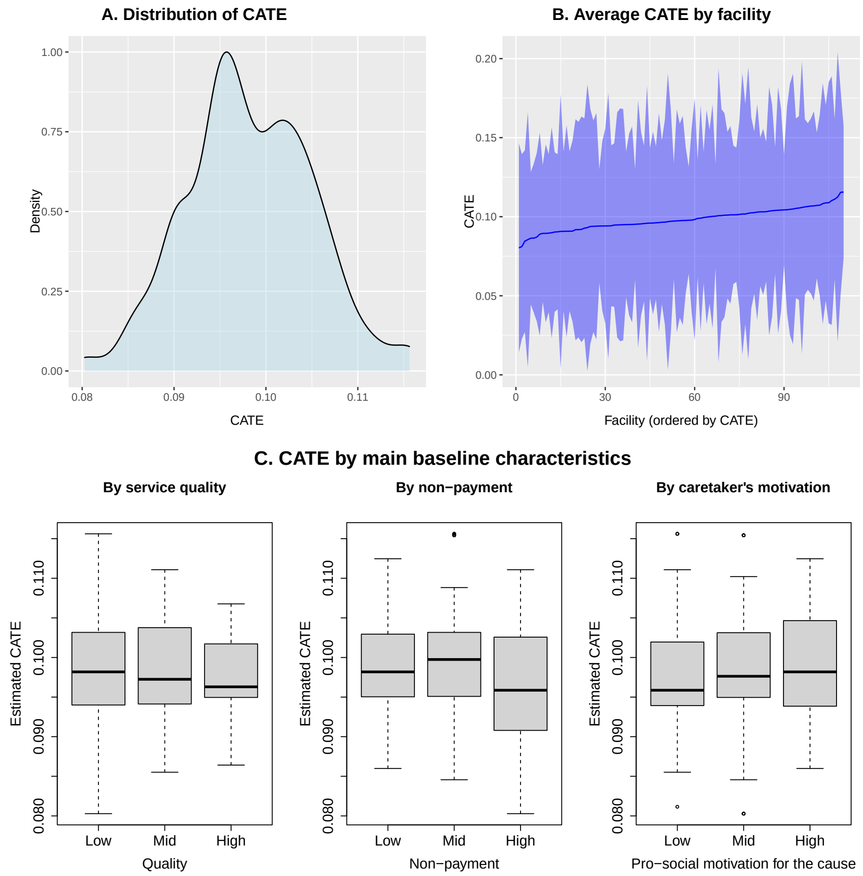
Table B11: Effects of maintenance treatment: causal forest procedure

	ATE via causal forest procedure			Calibration test	
	β	se	p-value	Mean prediction (p-value)	Heterogeneity (p-value)
	(1)	(2)	(3)	(4)	(5)
CT-/caretaker-level outcomes					
Higher quality	0.109	0.044	0.014	0.002	1.000
Structural maintenance	0.037	0.050	0.461	0.202	1.000
Routine maintenance	0.036	0.016	0.027	0.001	1.000
Hours worked	0.403	0.461	0.381	0.179	0.921
Knowledge of maintenance	0.094	0.037	0.012	0.003	1.000
Non-payment	-0.092	0.043	0.034	0.000	1.000
Users	-0.051	0.051	0.318	0.172	0.501
Enforcement	0.007	0.023	0.761	0.314	1.000
Monitoring	0.042	0.029	0.151	0.003	1.000
Household-/respondent-level outcomes					
WTP for service use	-0.027	0.059	0.642	0.373	0.968
Awareness of externality risks	0.025	0.017	0.14	0.116	1.000
Demand for pub. intervention (CT)	0.05	0.029	0.088	0.038	0.057
Demand for pub. intervention (OD)	-0.077	0.041	0.061	0.010	1.000
Demand for pub. intervention (other)	-0.019	0.026	0.471	0.185	1.000
Morbidity	0.007	0.017	0.669	0.389	0.949
Curative expenditure (extensive)	0.042	0.023	0.069	0.052	0.925
Curative expenditure (intensive)	-16.481	176.526	0.926	0.399	1.000
Preventive expenditure (extensive)	-0.002	0.003	0.484	0.136	1.000
Preventive expenditure (intensive)	-1.722	48.353	0.972	0.482	0.893

Note. Estimates presented in the first column are based on the cluster-robust causal forest procedure of Athey et al. (2019). We use the set of variables used in Appendix B.6, and we maintain the same assumptions about clustering implemented in Tables 1–3. Columns (1)–(3) present estimates of the ATE and the p-value of a two-sided test for the ATE being different from zero. Columns (4)–(5) implement a calibration test based on the best linear predictor method of Chernozhukov et al. (2017). Column (4) presents the p-value for the equality to 1 of the coefficient on the mean forest prediction, with 1 indicating that the mean forest prediction is correct. Column (5) presents the p-value for the equality to 1 of the coefficient on the quality of the estimates of treatment heterogeneity, with 1 indicating that the forest has captured heterogeneity in the underlying signal. Additional information about outcome variables is provided in Appendix A.

Figure B3 summarizes the causal forest results on heterogeneity of the effect on non-payment. Panel A shows the distribution of the Conditional ATE (CATE), while panel B averages the CATE at CT level and includes the 90 percent confidence interval. Panel C shows instead how CATE estimates vary according to three baseline characteristics of the CT: quality, non-payment, and caretaker's pro-social motivation for the cause. Results show the relatively homogeneous impact of the interventions on non-payment.

Figure B3: Conditional ATE of any treatment on non-payment



Note. Panel A shows the distribution of the Conditional ATE (CATE) of any treatment on non-payment computed using the cluster-robust causal forest procedure of Basu et al. (2018) and Athey and Wager (2019). Panel B shows the average CATE at CT level with the 90 percent confidence interval. Panel C shows variation of the CATE by baseline characteristics of the facility using a box plot. Low, mid and high indicates the first, second and third terciles in the distribution of the characteristic. In the box plot, each rectangle represents the inter-quartile range, with the top indicating the upper quartile, the bottom the lower quartile, and the middle line the median. The vertical line indicates the whiskers, i.e., the smallest value greater than the lower quartile minus 1.5 times the inter-quartile range, and the largest value less than the upper quartile plus 1.5 times the inter-quartile range. Additional information about the variables is provided in Appendix A.

B.8 Treatment heterogeneity by pre-specified dimensions

This section presents estimates of heterogeneous effects by a series of pre-registered variables. Table B12 presents an analysis of heterogeneity for CT- and caretaker-level outcomes. Tables B13 and B14 refer instead to respondent- and household-level outcomes.

Table B12: Heterogeneity by catchment area or CT characteristics

	Effect of maintenance treatment, by category						Het. test
	β	se	N	β	se	N	p-value
Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. WTP in catchment area		**Lower**			**Higher**		
Higher quality	0.15**	0.06	272	0.09	0.06	270	0.48
Structural maintenance	0.08	0.06	272	0.01	0.07	270	0.39
Routine maintenance	0.08***	0.03	272	0.01	0.02	270	0.05
Hours worked	0.28	0.54	272	0.08	0.45	270	0.77
Caretaker's knowledge	0.08	0.06	272	0.03	0.03	270	0.39
Users	0.05	0.07	273	-0.13*	0.07	269	0.07
Non-payment	-0.08	0.06	273	-0.07	0.05	269	0.90
Enforcement	-0.05	0.04	57	0.07*	0.04	54	0.02
Monitoring	0.07**	0.03	272	0.03	0.04	270	0.38
B. Quality of the service		**Lower**			**Higher**		
Higher quality	0.11*	0.06	294	0.11*	0.06	248	0.98
Structural maintenance	0.13**	0.06	294	-0.06	0.07	248	0.05
Routine maintenance	0.02	0.02	294	0.04*	0.02	248	0.48
Hours worked	-0.16	0.56	294	0.46	0.40	248	0.37
Caretaker's knowledge	0.13**	0.05	294	0.04	0.04	248	0.16
Users	-0.05	0.06	294	-0.07	0.08	248	0.85
Non-payment	-0.07	0.06	294	-0.08	0.06	248	0.91
Enforcement	0.01	0.03	60	-0.00	0.03	51	0.85
Monitoring	0.05	0.05	294	0.06**	0.03	248	0.87
C. Users		**Lower traffic**			**Higher traffic**		
Higher quality	0.08	0.07	204	0.13**	0.05	338	0.56
Structural maintenance	0.06	0.08	204	0.05	0.06	338	0.98
Routine maintenance	0.01	0.03	204	0.05**	0.02	338	0.29
Hours worked	-0.01	0.54	204	0.33	0.45	338	0.59
Caretaker's knowledge	0.16**	0.06	204	0.05	0.04	338	0.12
Users	-0.21**	0.10	203	0.03	0.05	339	0.05
Non-payment	-0.08	0.06	203	-0.07	0.05	339	0.92
Enforcement	0.01	0.03	41	0.01	0.03	70	0.94
Monitoring	0.03	0.03	204	0.06	0.04	338	0.71
D. Non-payment		**Lower**			**Higher**		
Higher quality	0.11	0.07	270	0.14***	0.04	272	0.74
Structural maintenance	0.04	0.07	270	0.06	0.08	272	0.88
Routine maintenance	0.02	0.02	270	0.05*	0.03	272	0.56
Hours worked	-0.33	0.36	270	0.75	0.59	272	0.12
Caretaker's knowledge	0.15***	0.05	270	0.01	0.05	272	0.04
Users	-0.08	0.07	270	-0.04	0.07	272	0.62
Non-payment	-0.07	0.05	270	-0.10*	0.06	272	0.65
Enforcement	-0.04	0.03	55	0.05*	0.03	56	0.04
Monitoring	0.03	0.02	270	0.07	0.05	272	0.50
E. Caretaker's motivation		**Lower**			**Higher**		
Higher quality	0.11**	0.05	264	0.09	0.07	278	0.72
Structural maintenance	0.16**	0.08	264	-0.08	0.06	278	0.01
Routine maintenance	0.02	0.03	264	0.04	0.02	278	0.71
Hours worked	0.60	0.60	264	-0.25	0.40	278	0.23
Caretaker's knowledge	0.03	0.05	264	0.15***	0.05	278	0.08
Users	0.04	0.08	265	-0.14**	0.06	277	0.07
Non-payment	-0.08	0.07	265	-0.06	0.05	277	0.77
Enforcement	0.02	0.03	54	-0.01	0.03	57	0.42
Monitoring	0.04	0.05	264	0.05*	0.03	278	0.95

Note. Categories for heterogeneity analysis are defined at baseline, with *lower* (*higher*) indicating whether the variable is smaller than or equal to (larger than) the sample median. Variables referring to catchment areas are averages of the corresponding variable within the catchment area. In columns (1)–(6), estimates are based on CT- or caretaker-level OLS regressions using equation (1) separately for each category. Column (7) presents a heterogeneity test based on CT- or caretaker-level OLS regressions using equation (1) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. The p-value is relative to the significance of the coefficient on the interaction term. Standard errors clustered by catchment area. The dependent variables are indicated in the rows and are defined in Appendix A. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Statistical significance is denoted by *** $p<0.01$, ** $p<0.05$, * $p<0.1$. *Caretaker's motivation* refers to the caretaker's pro-social motivation for the cause.

Table B13: Heterogeneity by individual characteristics: household-level outcomes

	Effect of maintenance treatment, by category						Het. test
	β	se	N	β	se	N	p-value
Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. WTP for service use		**Lower**			**Higher**		
WTP for service use	-0.07	0.08	4045	-0.02	0.09	4590	0.61
Awareness of externalities	0.01	0.02	2307	0.05**	0.02	2486	0.18
Demand for pub. int. (service)	0.07**	0.03	717	0.04	0.03	834	0.50
Demand for pub. int. (free riding)	-0.04	0.04	717	-0.10**	0.05	834	0.26
Demand for pub. int. (other)	-0.01	0.03	717	-0.04	0.03	834	0.44
Morbidity	-0.00	0.03	2307	0.02	0.03	2486	0.61
Curative exp. (extensive)	0.04	0.04	1487	0.03	0.04	1789	0.76
Curative exp. (intensive)	321.12	337.09	1487	-230.92	288.95	1789	0.20
Preventive exp. (extensive)	-0.01*	0.01	1499	0.00	0.01	1802	0.07
Preventive exp. (intensive)	-24.58	80.10	1499	58.67	77.86	1801	0.35
B. Awareness of externality		**Lower**			**Higher**		
WTP for service use	0.01	0.13	2474	-0.07	0.08	6161	0.57
Awareness of externalities	0.01	0.03	1359	0.04**	0.02	3434	0.21
Demand for pub. int. (service)	0.04	0.04	435	0.05*	0.03	1116	0.75
Demand for pub. int. (free riding)	-0.03	0.05	435	-0.09**	0.04	1116	0.28
Demand for pub. int. (other)	-0.03	0.04	435	-0.02	0.03	1116	0.76
Morbidity	0.03	0.03	1359	-0.00	0.03	3434	0.58
Curative exp. (extensive)	0.05	0.04	880	0.03	0.03	2396	0.62
Curative exp. (intensive)	253.62	439.98	880	-71.43	231.34	2396	0.48
Preventive exp. (extensive)	-0.01	0.01	887	-0.00	0.00	2414	0.64
Preventive exp. (intensive)	23.50	82.30	886	16.20	71.58	2414	0.94
C. Trust in the community		**Lower**			**Higher**		
WTP for service use	-0.06	0.08	7041	-0.00	0.13	1594	0.59
Awareness of externalities	0.02	0.02	3897	0.08**	0.04	896	0.07
Demand for pub. int. (service)	0.08***	0.03	1218	-0.03	0.04	333	0.01
Demand for pub. int. (free riding)	-0.07	0.04	1218	-0.09	0.07	333	0.66
Demand for pub. int. (other)	-0.04*	0.02	1218	0.03	0.05	333	0.12
Morbidity	0.00	0.03	3897	0.05	0.04	896	0.40
Curative exp. (extensive)	0.03	0.03	2510	0.05	0.05	766	0.83
Curative exp. (intensive)	-24.18	256.10	2510	229.48	444.50	766	0.65
Preventive exp. (extensive)	-0.01	0.00	2526	0.01	0.01	775	0.10
Preventive exp. (intensive)	2.25	67.32	2525	83.10	123.00	775	0.48
D. Distance to CT		**Shorter**			**Longer**		
WTP for service use	-0.16	0.10	4256	0.07	0.09	4379	0.06
Awareness of externalities	0.03	0.02	2370	0.03	0.03	2423	0.87
Demand for pub. int. (service)	0.07*	0.04	774	0.03	0.03	777	0.40
Demand for pub. int. (free riding)	-0.06	0.05	774	-0.09**	0.04	777	0.58
Demand for pub. int. (other)	-0.01	0.03	774	-0.03	0.03	777	0.64
Morbidity	0.02	0.03	2370	-0.01	0.03	2423	0.51
Curative exp. (extensive)	0.03	0.04	1615	0.04	0.03	1661	0.96
Curative exp. (intensive)	-406.57	349.23	1615	435.42	268.26	1661	0.05
Preventive exp. (extensive)	-0.00	0.01	1626	-0.00	0.00	1675	0.85
Preventive exp. (intensive)	-3.32	83.44	1625	41.79	80.59	1675	0.68

Note. Categories for heterogeneity analysis are defined at baseline, with *lower* (*higher*) indicating whether the variable is smaller than or equal to (larger than) the sample median. In columns (1)–(6), estimates are based on respondent- and household-level OLS regressions using equation (1) separately for each category. Column (7) presents a heterogeneity test based on CT-level OLS regressions using equation (1) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. The p-value is relative to the significance of the coefficient on the interaction term. Standard errors are clustered by catchment-area–round of observation. The dependent variables are indicated in the rows and are defined in Appendix A. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Specifications where the level of analysis is the respondent also include gender. Statistical significance denoted by *** p<0.01, ** p<0.05, * p<0.1. *Trust in the community* refers to the trust in the community to keep the CT clean.

Table B14: Heterogeneity by catchment area characteristics: household-level outcomes

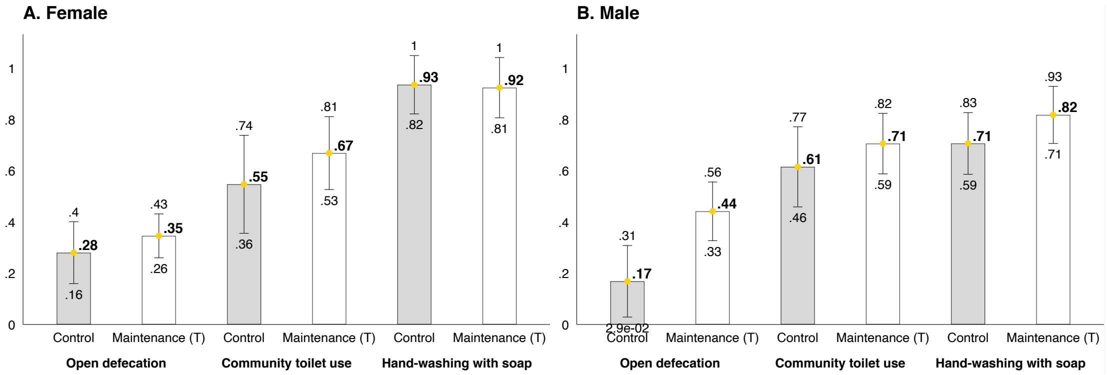
	Effect of maintenance treatment, by category						Het. test
	β	se	N	β	se	N	p-value
Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Water quality		**Lower**			**Higher**		
WTP for service use	-0.03	0.10	4257	-0.05	0.11	4378	0.90
Awareness of externalities	-0.00	0.02	2357	0.06**	0.02	2436	0.10
Demand for pub. int. (service)	0.03	0.04	756	0.07**	0.03	795	0.38
Demand for pub. int. (free riding)	-0.02	0.05	756	-0.12**	0.05	795	0.19
Demand for pub. int. (other)	-0.04	0.03	756	0.00	0.03	795	0.32
Morbidity	0.05	0.04	2357	-0.03	0.03	2436	0.09
Curative exp. (extensive)	0.06	0.04	1598	0.01	0.04	1678	0.37
Curative exp. (intensive)	597.79**	296.02	1598	-464.56	296.42	1678	0.02
Preventive exp. (extensive)	-0.00	0.01	1610	-0.00	0.01	1691	0.79
Preventive exp. (intensive)	147.57*	81.00	1610	-97.19	93.24	1690	0.05
B. Quality of the service		**Lower**			**Higher**		
WTP for service use	-0.19**	0.09	4571	0.08	0.12	4064	0.08
Awareness of externalities	0.05*	0.03	2517	0.02	0.02	2276	0.46
Demand for pub. int. (service)	0.08*	0.04	818	0.04	0.03	733	0.45
Demand for pub. int. (free riding)	-0.06	0.05	818	-0.09*	0.06	733	0.68
Demand for pub. int. (other)	-0.05*	0.03	818	0.01	0.04	733	0.27
Morbidity	0.03	0.03	2517	-0.02	0.04	2276	0.30
Curative exp. (extensive)	0.01	0.03	1709	0.07	0.05	1567	0.26
Curative exp. (intensive)	-426.43	294.68	1709	471.68	331.17	1567	0.04
Preventive exp. (extensive)	-0.00	0.01	1721	-0.00	0.01	1580	0.80
Preventive exp. (intensive)	-32.22	88.99	1720	73.73	96.15	1580	0.44
C. Non-payment		**Lower**			**Higher**		
WTP for service use	-0.03	0.10	4184	-0.06	0.11	4451	0.85
Awareness of externalities	0.06***	0.02	2330	0.00	0.03	2463	0.10
Demand for pub. int. (service)	0.05	0.04	754	0.05	0.04	797	0.86
Demand for pub. int. (free riding)	-0.13**	0.05	754	-0.02	0.05	797	0.11
Demand for pub. int. (other)	-0.03	0.04	754	-0.01	0.03	797	0.69
Morbidity	0.06*	0.03	2330	-0.04	0.04	2463	0.03
Curative exp. (extensive)	0.03	0.05	1608	0.04	0.04	1668	0.86
Curative exp. (intensive)	-158.36	330.20	1608	239.20	314.62	1668	0.38
Preventive exp. (extensive)	-0.00	0.01	1617	-0.00	0.01	1684	0.91
Preventive exp. (intensive)	55.10	99.48	1617	-5.91	90.48	1683	0.69
D. Caretaker's motivation		**Lower**			**Higher**		
WTP for service use	0.12	0.09	4124	-0.20*	0.11	4511	0.03
Awareness of externalities	0.03	0.03	2292	0.04*	0.02	2501	0.73
Demand for pub. int. (service)	0.03	0.04	731	0.08**	0.04	820	0.31
Demand for pub. int. (free riding)	-0.04	0.06	731	-0.10**	0.04	820	0.47
Demand for pub. int. (other)	-0.02	0.04	731	-0.03	0.03	820	0.82
Morbidity	0.01	0.04	2292	0.01	0.03	2501	0.89
Curative exp. (extensive)	0.03	0.05	1561	0.07*	0.04	1715	0.54
Curative exp. (intensive)	145.83	349.65	1561	-126.53	310.43	1715	0.55
Preventive exp. (extensive)	0.00	0.01	1573	-0.01	0.00	1728	0.20
Preventive exp. (intensive)	37.24	94.75	1573	21.52	89.16	1727	0.90

Note. Categories for heterogeneity analysis are defined at baseline, with *lower* (*higher*) indicating whether the variable is smaller than or equal to (larger than) the sample median. In columns (1)–(6), estimates are based on respondent- and household-level OLS regressions using equation (1) separately for each category. Column (7) presents a heterogeneity test based on CT-level OLS regressions using equation (1) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. The p-value is relative to the significance of the coefficient on the interaction term. Standard errors are clustered by catchment area–round of observation. The dependent variables are indicated in the rows and are defined in Appendix A. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Specifications where the level of analysis is the respondent also include gender. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1. *Caretaker's motivation* refers to the caretaker's pro-social motivation for the cause.

B.9 Treatment heterogeneity by gender and religion

Supplementary Material S.1 discusses the differences by gender for non-payment. Figure B4 shows estimates of the effect of interventions on hygiene- and sanitation-related behavior using the list randomization technique distinguishing by the gender of the respondent. Tables B15–B16 presents an analysis of heterogeneity by religion for CT- and caretaker-level outcomes and for respondent- and household-level outcomes.

Figure B4: Sanitation- and hygiene-related behavior, by gender



Note. The figure shows the share of study participants practicing each behavior in the day previous to the interview, estimating using a list randomization technique. Panel A restricts the sample to female respondents, while panel B restricts the sample to male respondents. Randomization of lists was performed at individual level, and data were collected during follow-up 5 only. Supplementary Material S.4 provides details about measurement.

Table B15: Heterogeneity by of religion of potential users: CT-level outcomes

	Effect of maintenance treatment, by category						Het. test
	β	se	N	β	se	N	p-value
Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Share of Hindu in catchment area		**Lower**			**Higher**		
Higher quality	0.04	0.06	282	0.19***	0.06	260	0.10
Structural maintenance	0.05	0.08	282	0.05	0.06	260	0.99
Routine maintenance	-0.01	0.03	282	0.07***	0.02	260	0.02
Hours worked	0.14	0.54	282	0.21	0.49	260	0.92
Caretaker's knowledge	0.09**	0.04	282	0.12**	0.06	260	0.66
Users	-0.13*	0.07	281	0.01	0.07	261	0.17
Non-payment	-0.08	0.07	281	-0.07	0.06	261	0.90
Enforcement	-0.01	0.04	57	0.00	0.02	54	0.80
Monitoring	0.05	0.05	282	0.05	0.03	260	0.99

Note. Categories for heterogeneity analysis are defined at baseline, with *lower* (*higher*) indicating whether the share of potential users that are of Hindu religion is smaller than or equal to (larger than) the sample median. Variables referring to catchment areas are averages of the corresponding variable within the catchment area. In columns (1)–(6), estimates are based on CT- or caretaker-level OLS regressions using equation (1) separately for each category. Column (7) presents a heterogeneity test based on CT- or caretaker-level OLS regressions using equation (1) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. The p-value is relative to the significance of the coefficient on the interaction term. Standard errors clustered by catchment area. The dependent variables are indicated in the rows and are defined in Appendix A. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1. *Caretaker's motivation* refers to the caretaker's pro-social motivation for the cause.

Table B16: Heterogeneity by religion of potential users: household-level outcomes

	Effect of maintenance treatment, by category						Het. test
	β	se	N	β	se	N	p-value
Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Respondent's religion		**Other religion**			**Hindu**		
WTP for service use	-0.28	0.18	1139	-0.02	0.08	5623	0.22
Awareness of externalities	0.02	0.04	634	0.03	0.02	3112	0.76
Demand for pub. int. (service)	0.14**	0.06	196	0.07**	0.03	973	0.45
Demand for pub. int. (free riding)	0.07	0.08	196	-0.09*	0.05	973	0.10
Demand for pub. int. (other)	-0.04	0.07	196	-0.04	0.03	973	0.89
Morbidity	-0.03	0.05	634	0.02	0.03	3112	0.46
Curative exp. (extensive)	0.04	0.05	404	0.03	0.04	2011	0.87
Curative exp. (intensive)	170.27	740.93	404	-120.17	292.46	2011	0.72
Preventive exp. (extensive)	-0.02***	0.01	406	-0.00	0.01	2025	0.10
Preventive exp. (intensive)	11.67	159.09	405	-10.77	65.07	2025	0.88
B. Share of Hindu in catchment area		**Lower**			**Higher**		
WTP for service use	-0.16	0.10	4236	0.04	0.11	4399	0.25
Awareness of externalities	0.03	0.03	2345	0.04	0.03	2448	0.80
Demand for pub. int. (service)	0.09**	0.04	760	0.02	0.04	791	0.30
Demand for pub. int. (free riding)	-0.02	0.04	760	-0.12**	0.06	791	0.18
Demand for pub. int. (other)	-0.06*	0.03	760	0.01	0.04	791	0.18
Morbidity	0.03	0.03	2345	-0.00	0.04	2448	0.51
Curative exp. (extensive)	0.03	0.05	1600	0.03	0.03	1676	0.98
Curative exp. (intensive)	-24.68	323.98	1600	63.12	322.95	1676	0.83
Preventive exp. (extensive)	-0.01*	0.00	1613	0.00	0.01	1688	0.28
Preventive exp. (intensive)	-5.05	90.31	1612	30.71	97.16	1688	0.84

Note. In panel A, categories for heterogeneity analysis are defined at baseline, with *Hindu* indicating whether the respondent's religion is Hindu. In panel B, categories for heterogeneity analysis are defined at baseline, with *lower* (*higher*) indicating whether the share of potential users that are of Hindu religion is smaller than or equal to (larger than) the sample median. In columns (1)–(6), estimates are based on respondent- and household-level OLS regressions using equation (1) separately for each category. Column (7) presents a heterogeneity test based on CT-level OLS regressions using equation (1) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. The p-value is relative to the significance of the coefficient on the interaction term. Standard errors are clustered by catchment-area–round of observation. The dependent variables are indicated in the rows and are defined in Appendix A. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Specifications where the level of analysis is the respondent also include gender. Statistical significance denoted by *** p<0.01, ** p<0.05, * p<0.1. *Trust in the community* refers to the trust in the community to keep the CT clean.

B.10 Sanitation and hygiene behavior and attitudes

Table B17 shows estimates of treatment effects on self-reported sanitation and hygiene behavior.

Table B18 shows estimates of treatment effects on incentivized behavior measured by the modified dictator game (see Section S.4.6) and the public goods game (see Section S.4.7).

Table B17: Self-reported sanitation and hygiene behaviour

	Open defecation						Soap use
	Respondent	Spouse	Male child		Female child		Respondent
			> 14yo	6–14yo	> 14yo	6–14yo	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A							
Maintenance (T)	-0.013	-0.011	-0.010	-0.004	0.006	-0.001	0.014
	(0.040)	(0.035)	(0.023)	(0.022)	(0.019)	(0.018)	(0.007)
	[0.75]	[0.75]	[0.68]	[0.86]	[0.78]	[0.96]	[0.06]
Panel B							
Maintenance only (T1)	0.019	0.022	0.007	0.006	0.024	-0.000	0.017
	(0.053)	(0.046)	(0.031)	(0.030)	(0.028)	(0.022)	(0.008)
	[0.73]	[0.63]	[0.81]	[0.85]	[0.39]	[0.99]	[0.03]
Maintenance + sensitization (T2)	-0.042	-0.042	-0.026	-0.013	-0.012	-0.001	0.011
	(0.044)	(0.036)	(0.026)	(0.025)	(0.021)	(0.022)	(0.008)
	[0.34]	[0.25]	[0.31]	[0.58]	[0.58]	[0.95]	[0.17]
T1 = T2 (p-value)	0.287	0.159	0.316	0.555	0.254	0.965	0.301
Mean (control group)	0.148	0.116	0.090	0.089	0.057	0.058	0.967
Std. Dev. (control group)	0.355	0.321	0.286	0.285	0.232	0.234	0.178
Observations	9588	7908	9588	9588	9588	9588	9588
Catchment areas	110	110	110	110	110	110	110

Note. Estimates based on household-level OLS regressions using equation (1) in panel A, and equation (2) in panel B. Standard errors clustered by catchment area are reported in parentheses and p-values in brackets. Dependent variables by column: (1)–(6) *Open defecation*: is an indicator variable equal to 1 if the household member (by demographic group) is reported to have practiced open defecation the last time they defecated, and 0 otherwise; (7) *Soap*: is an indicator variable equal to 1 if the respondent reports washing her/his hands with soap, and 0 otherwise. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A.

Table B18: Behavioural games

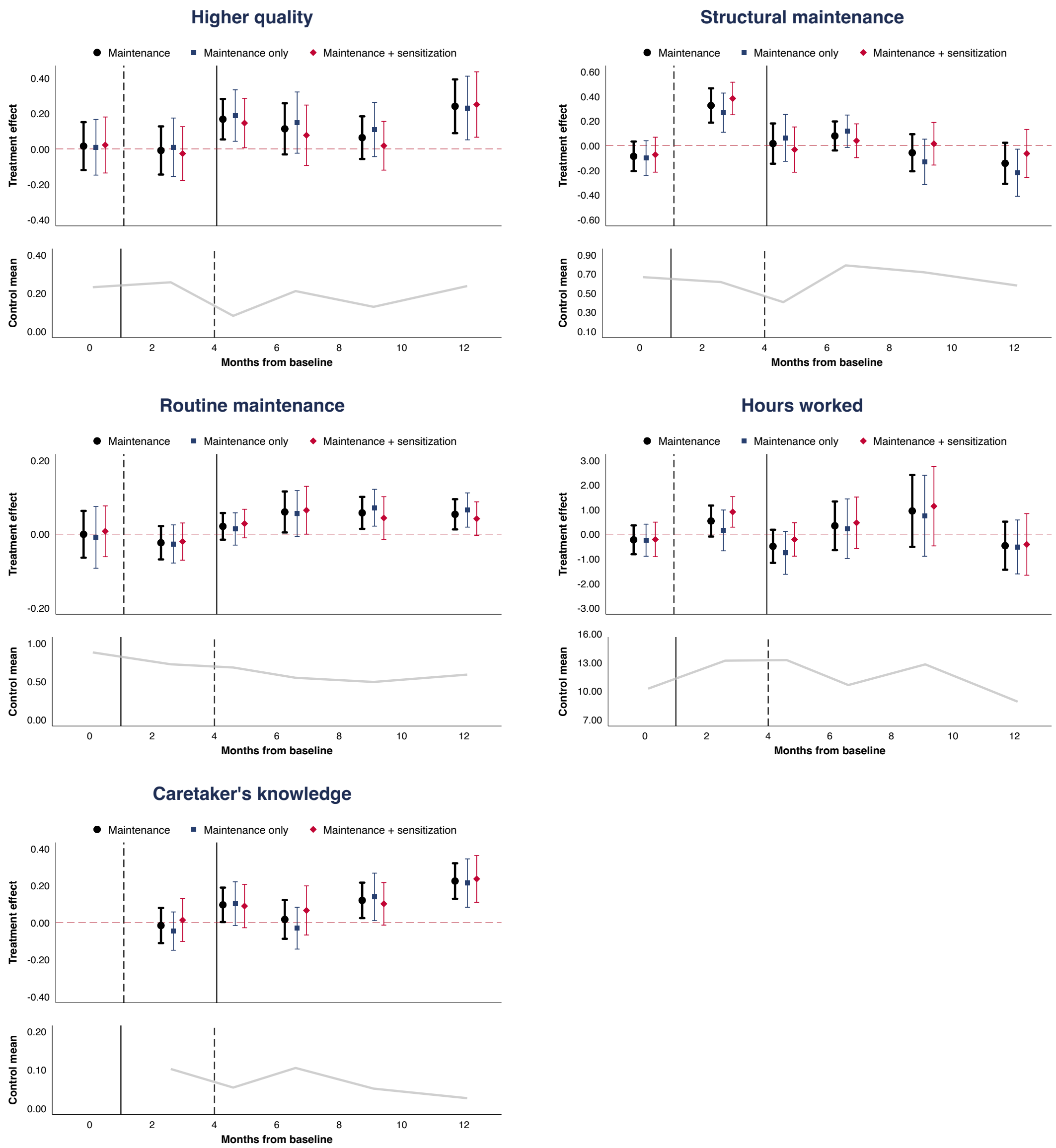
	Caretaker	Potential users	
	Pro-social motivation	Contribution to service quality	PGG contribution
	(1)	(2)	(3)
Maintenance (T)	-0.024	-0.005	0.001
	(0.025)	(0.007)	(0.013)
	[0.33]	[0.47]	[0.92]
Mean (control group)	0.345	0.212	0.174
Observations	542	8635	1228
Catchment areas	110	110	109
Observation rounds	5	3	1

Note. Estimates based on household-level OLS regressions using equation (1). Standard errors clustered by catchment area are reported in parentheses and p-values in brackets. Dependent variables by column: (1)–(3): (1) *Pro-social motivation (for the cause)*: share of the endowment that is donated by the caretaker in the adapted dictator game (Appendix S.4.6); (2) *Contribution to service quality*: share of the endowment that is donated by the respondent in the adapted dictator game (Appendix S.4.6): (3) *PGG contribution*: share contributed in the public good game (Appendix S.4.7). All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A.

B.11 Estimates of treatment effects by survey

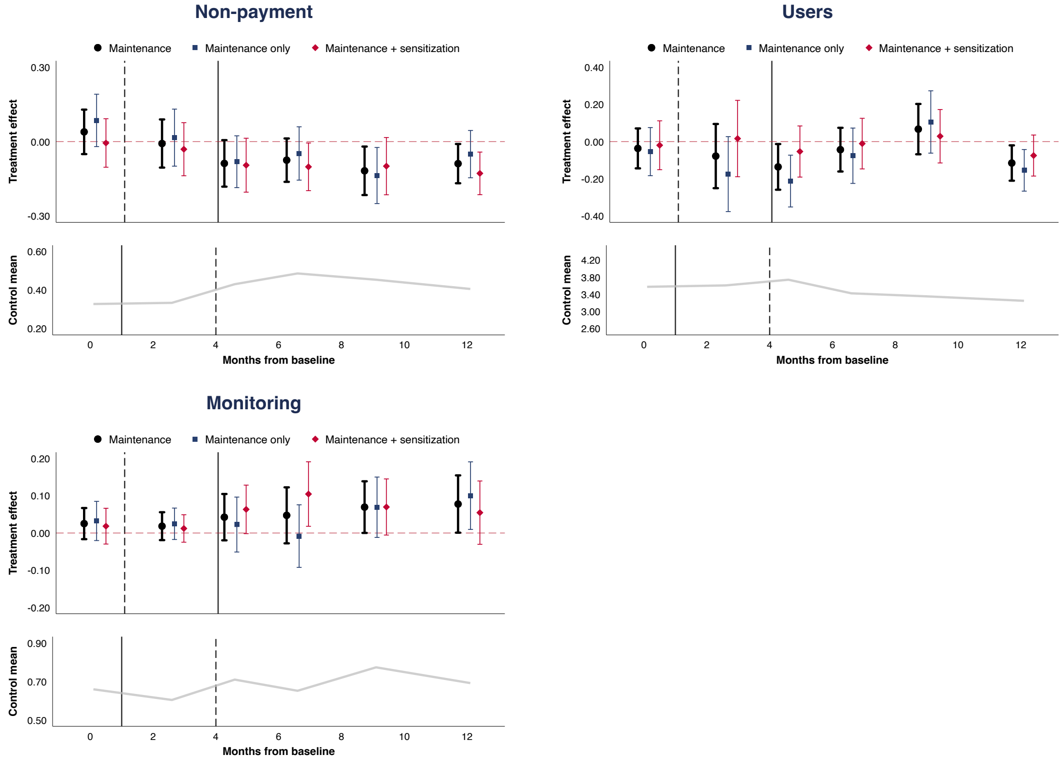
For the outcomes presented in Tables 1–3, this section presents estimates of equation (1) and equation (2) separately for each survey. Estimates are presented in Figures B7–B8. The upper part of each panel presents estimates of treatment effects on the corresponding variable, while the lower part reports the evolution over time of the average of the corresponding variable in the control group. Figures B7–B8 do not report variables that were measured only once.

Figure B5: Timing of effects for outcomes in Table 1



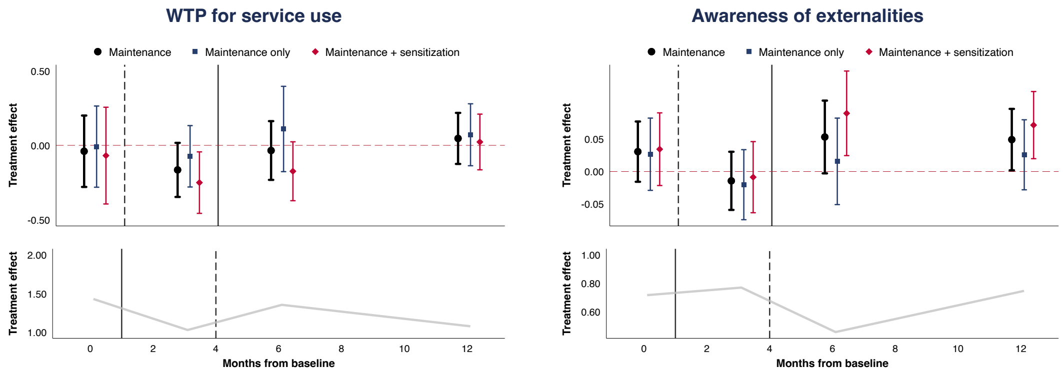
Notes. Estimates based on CT-level OLS regressions using equation (1) and equation (2) separately for each data collection period. Confidence intervals are computed at the 10 percent level of confidence using robust standard errors. Outcome variables are defined in Appendix A. All specifications include strata indicators for the city and the provider of the CT.

Figure B6: Timing of effects for outcomes in Table 2



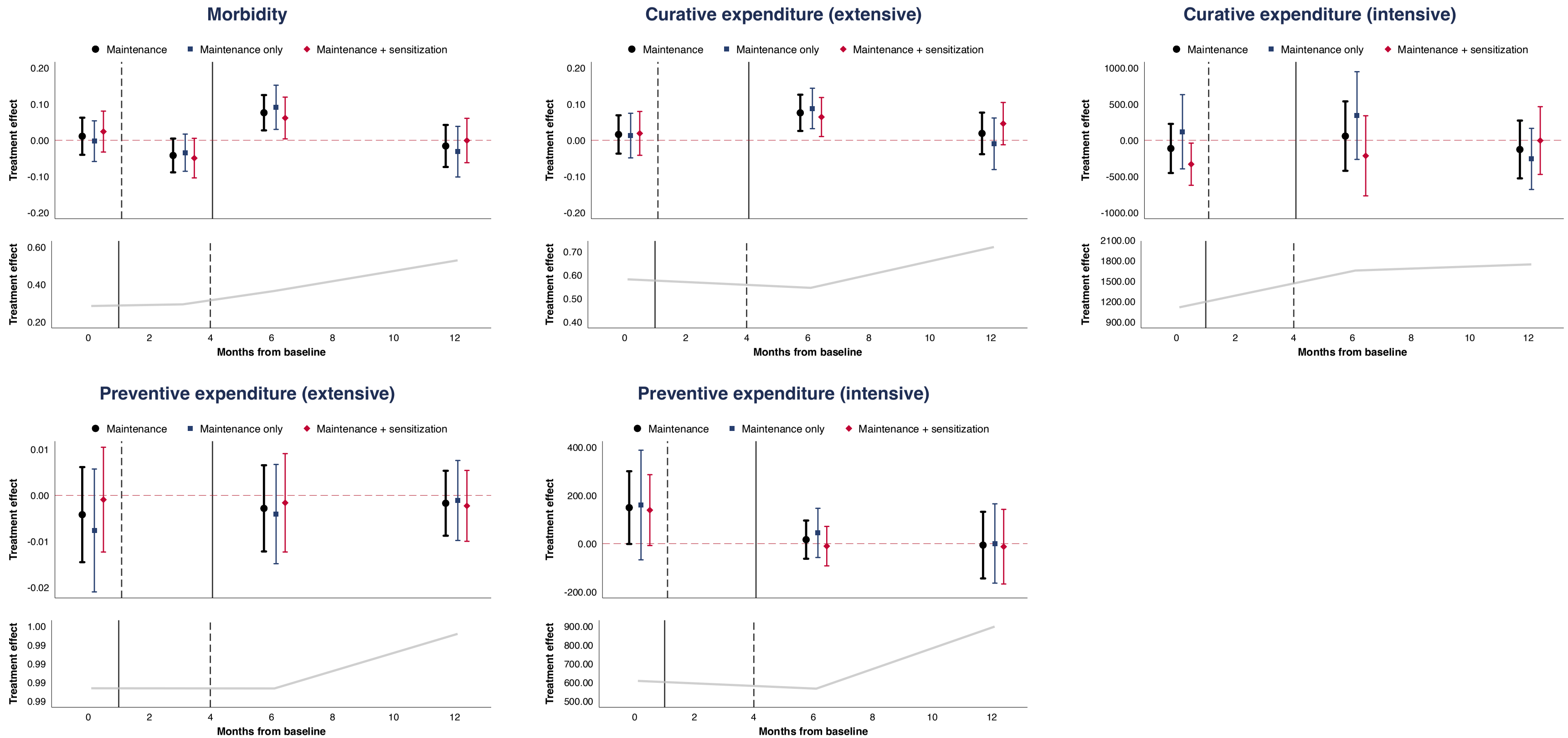
Notes. Estimates based on CT-level OLS regressions using equation (1) and equation (2) separately for each data collection period. Confidence intervals are computed at the 10 percent level of confidence using robust standard errors. Outcome variables are defined in Appendix A. All specifications include strata indicators for the city and the provider of the CT. *Enforcement* is excluded since the measurement is available for only one round.

Figure B7: Timing of effects for outcomes in Table 4



Notes. Estimates based on respondent- and household-level OLS regressions using equation (1) and equation (2) separately for each data collection period. Confidence intervals are computed at the 10 percent level of confidence using standard errors clustered at the catchment area. Outcome variables are defined in Appendix A. All specifications include strata indicators for the city and the provider of the CT. Respondent-level regressions include a control for the gender of the respondent.

Figure B8: Timing of effects for outcomes in Table 3



Notes. Estimates based on household-level OLS regressions using equation (1) and equation (2) separately for each data collection period. Confidence intervals are computed at the 10 percent level of confidence using standard errors clustered at the catchment area. Outcome variables are defined in Appendix A. All specifications include strata indicators for the city and the provider of the CT. Respondent-level regressions include a control for the gender of the respondent.

C Estimates using ANCOVA and IPW specifications

Tables C19–C20 present estimates of treatment effects using equations (1) and (2) adding the value at baseline of the dependent variable as a control variable (ANCOVA specification). The organization of the results and the order of the variables are the same as in Tables 1–3 in the main text. Table C20 also present estimates of treatment effects using equations (1) and (2) weighting observations by inverse probability weights (Wooldridge, 2002).

Table C19: Tables 1–2, estimates with ANCOVA specification

	β (1)	se (2)	p-value (3)	ANCOVA specif (4)
Higher quality	0.11	0.04	0.01	Yes
Structural maintenance	0.06	0.05	0.23	Yes
Inputs	0.03	0.01	0.02	Yes
Hours worked	0.19	0.32	0.55	Yes
Awareness	0.09	0.04	0.01	-
Non-payment	-0.08	0.04	0.04	Yes
Users	-0.05	0.05	0.30	Yes
Enforcement	0.00	0.02	0.92	-
Monitoring	0.05	0.03	0.09	Yes

Note. Estimates based on CT-level OLS regressions using equation (1), controlling for the baseline value of the dependent variable if available (see *ANCOVA specification* column). All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A.

Table C20: Tables 4–3, estimates with ANCOVA and IPW specifications

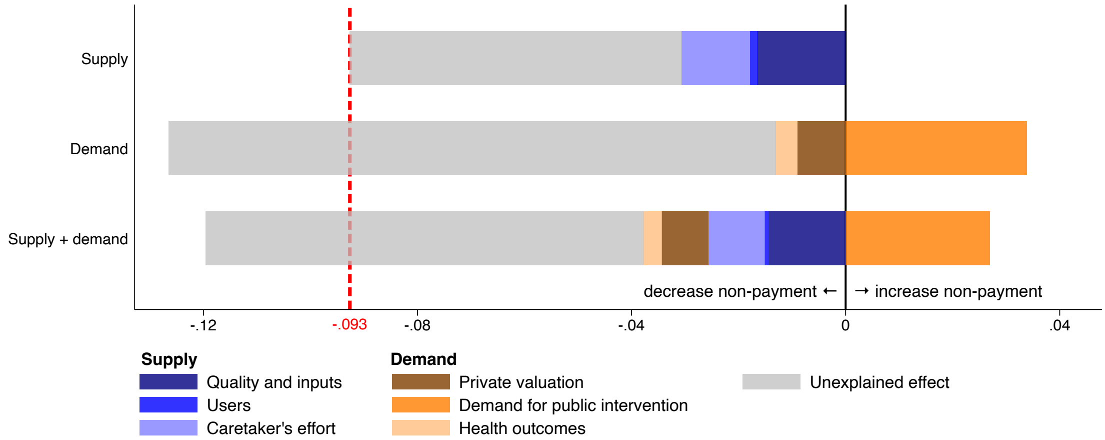
	ANCOVA				**IPW**		
	β	se	p-value	ANCOVA specif	β	se	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
WTP for CT use	-0.04	0.06	0.52	Yes	-0.05	0.07	0.43
Awareness of OD externality	0.03	0.02	0.09	Yes	0.03	0.02	0.08
Demand for pub int (CT)	0.05	0.03	0.05	-	0.05	0.03	0.05
Demand for pub int (OD)	-0.07	0.04	0.04	-	-0.07	0.04	0.06
Demand for pub int (Other)	-0.02	0.02	0.38	-	-0.02	0.02	0.39
Morbidity	0.01	0.02	0.78	Yes	0.01	0.02	0.78
Curative exp (extensive)	0.05	0.02	0.05	-	0.05	0.02	0.05
Curative exp (intensive)	-34.27	190.05	0.86	-	-54.28	197.03	0.78
Preventive exp (extensive)	-0.00	0.00	0.53	-	-0.00	0.00	0.48
Preventive exp (intensive)	0.62	49.47	0.99	-	2.59	47.01	0.96

Note. Estimates based on respondent- and household-level OLS regressions using equation (1), controlling for the baseline value of the dependent variable if available (see *ANCOVA specification* column) in columns (1) to (3), and weighting observations by inverse probability weights in columns (4) to (6). Weights are estimated at baseline using a probit regression on indicator variables for attrition at different follow-ups on observable characteristics of the household and of the catchment area where the household resides. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Specifications where the level of analysis is the respondent also include gender. Additional details about the variables are presented in Appendix A.

D Mediation analysis

Figure D9 shows the decomposition of the effect of the maintenance treatment on non-payment.

Figure D9: The effect on non-payment: a mediation analysis



Note. The figure shows the decomposition of the effect of the maintenance treatment on non-payment. We follow the procedure of Gelbach (2016) applied to non-payment in the incentive period (follow-ups 2–5). We include as mediators all outcome variables (with the exception of non-payment) included in Tables 1–3, grouped by table. We distinguish two groups of mediators: *supply* mediators, which include CT and caretaker outcomes, and *demand* mediators, which include the median value in the catchment area of slum resident outcomes. The decomposition is presented by including only *supply* mediators in the top bar, only *demand* mediators in the middle bar, and both groups in the bottom bar. The dashed vertical line indicates the estimate of the effect of the maintenance treatment on non-payment in the incentive period and estimated using equation (1). The shaded gray areas represent the part of the effect not explained by mediators. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix A.

SUPPLEMENTARY MATERIAL – NOT FOR ONLINE PUBLICATION

Public Service Delivery and Free Riding: Experimental Evidence from India

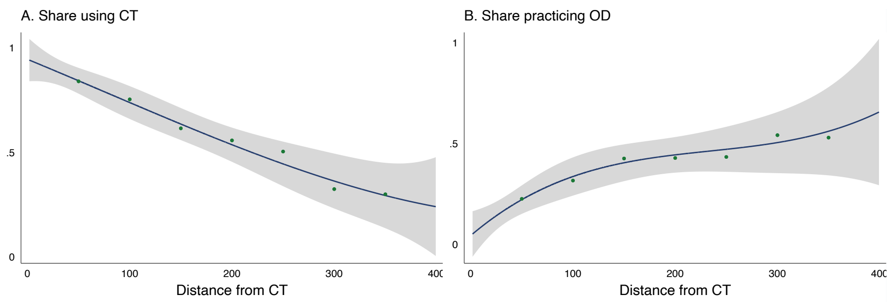
Alex Armand, Britta Augsburg, Antonella Bancalari

This supplementary material provides detailed information about the context in which the study is implemented (S.1), a detailed description of interventions (S.2), their cost (S.3), sampling, data collection and measurement (S.4).

S.1 An analysis of CTs in the study area

The location of a CT in the slum is a strong predictor of sanitation-related behavior among eligible households. Using self-reported data from the census of slum residents (see Supplementary Material S.4.1), we study how distance from a facility affects the use of the service versus free riding among study households. Figure S.1.1 presents cubic fits for the relationships between the distance from the facility and self-reported use of the service (panel A), or OD (panel B). The use of a CT reduces rapidly with distance from the facility: at 200 meters, only 50 percent of eligible households use the facility, and more than 40 percent practice OD. At 400 meters from a CT, very few households use the service, while over 50 percent of respondents practice OD.

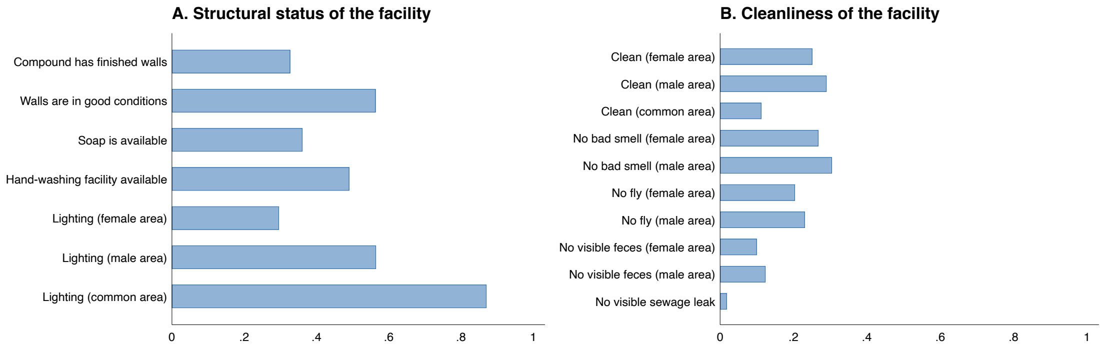
Figure S.1.1: Sanitation behavior in the slum, by distance from a facility



Note. Data source is the slum resident census (see Supplementary Material S.4). The figures present cubic fits of the share using CT (panel A) and of the share practicing OD (panel B) on distance from the closest CT. Dots show the average of the variable indicated on the vertical axis for equally-spaced intervals on the variable indicated on the horizontal axis. The shaded area presents the 90 percent confidence intervals, assuming standard errors are clustered at the slum level. The sample includes all households in the census that are considered eligible for the study (see Supplementary Material S.4.1 for details).

In the urban slums of Uttar Pradesh, CTs are often found in poor condition. Figure S.1.2 summarizes the average status of the facilities as collected by observers at baseline (Supplementary Material S.4.2). Panel A refers to the structural status of the facility, while panel B refers to the cleanliness of the facility. CTs are characterized by the poor quality of the construction, by the lack of hand-washing facilities, and by a general lack of cleanliness.

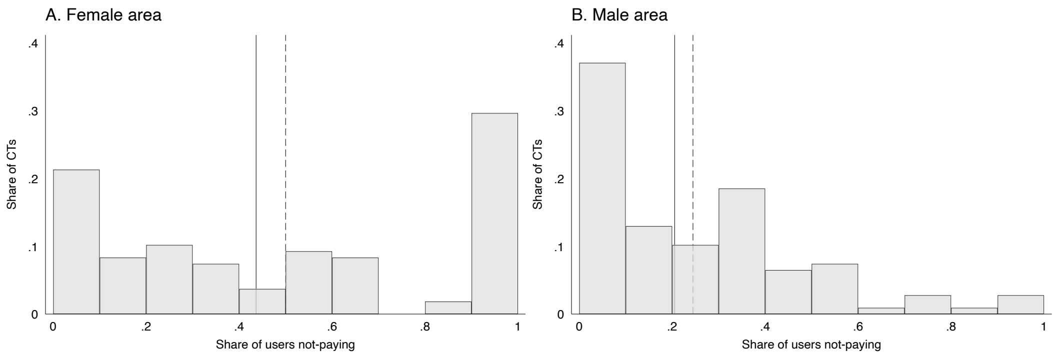
Figure S.1.2: Status of facilities at baseline



Note. Share of CTs that have or have access to the corresponding characteristic. Information is measured at baseline by observers. Supplementary Material S.4 provides details about measurement.

The role of potential users. Less than half of households reporting that they like the services offered in the local CT, 36 percent reporting the CT is clean, 15 percent reporting that they like the facility, and 28 percent reporting they considered it safe. Figure S.1.3 documents the distribution of the share of women and men who use the CT without paying during the rush hour. On average, only 66 percent of users pay the CT fee. This is mainly driven by women, among whom 50 percent do not pay the fee, as compared with 24 percent among men.

Figure S.1.3: Non-payment at baseline, by gender-specific area

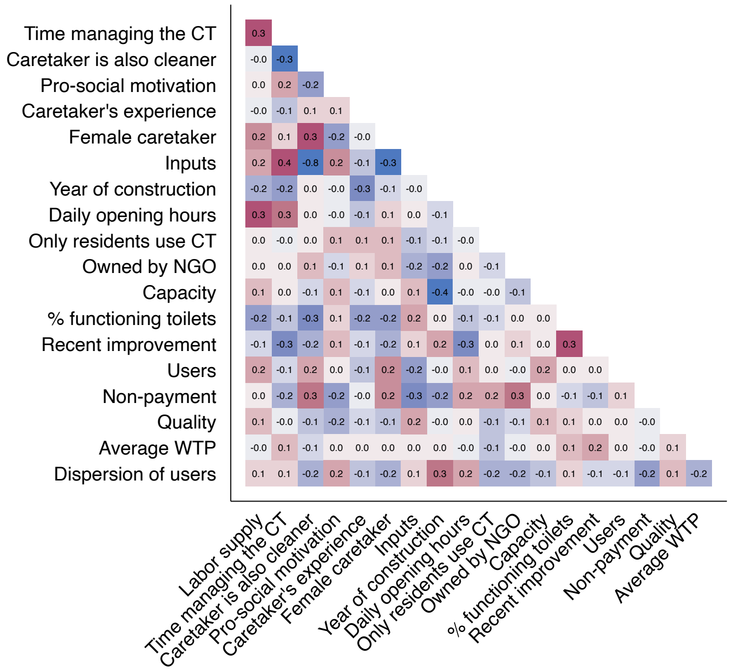


Note. Data collected at baseline. Panel A (panel B) reports the share of female (male) users who do not pay the fee for the use of the female (male) area of the CT during 1 hour at dawn (rush hour), measured by observers. The solid vertical lines represent the sample median, and the dashed vertical lines represent the sample mean. Additional details about the variable are presented in Appendix A. Supplementary Material S.4 provides details about measurement.

The role of caretakers. Figure S.1.4 presents an unconditional correlation matrix measured at baseline for a variety of indicators associated with the CT, including caretakers' characteristics and average characteristics of study participants in the catchment area.

Caretakers' labor supply is positively related to the opening hours of the CT and the share of time allocated to managing the CT. Caretakers who are also working as cleaners tend to be female, have lower pro-social motivation and work in worse CTs (as measured by the share of functioning toilets). Recent improvements positively correlate with the share of functioning toilets, which is higher where caretakers have higher experience. Non-payment is especially concentrated in CTs owned by an NGO, and in CTs used only by residents, but is lower where caretakers are male, they are more pro-socially motivated, they spend a higher share of time managing the CT, they do not work as cleaners, and they use better inputs. Overall, these statistics indicate the importance of the caretaker for the quality of service delivery.

Figure S.1.4: Correlation matrix for caretaker and CT characteristics



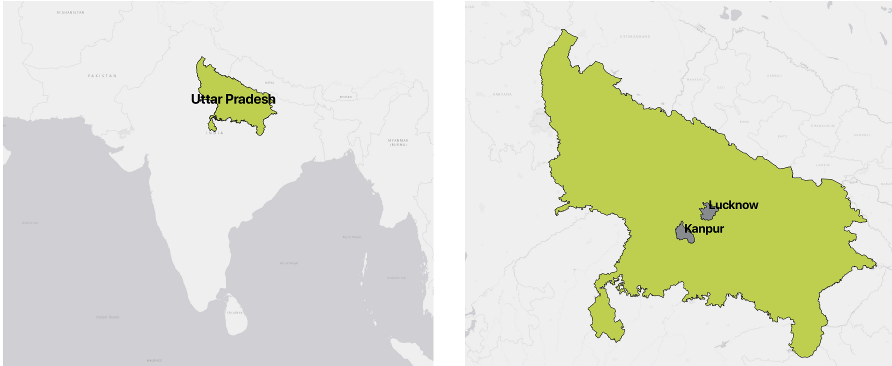
Note. The figure presents unconditional correlations between CT indicators. Variables are measured at baseline. Refer to Appendix A for variable definitions. *Caretaker's experience* is the number of months the caretaker has worked in the CT. *Dispersion of users* is the square of the average distance between a potential user and the CT in the catchment area.

S.2 Details about the interventions

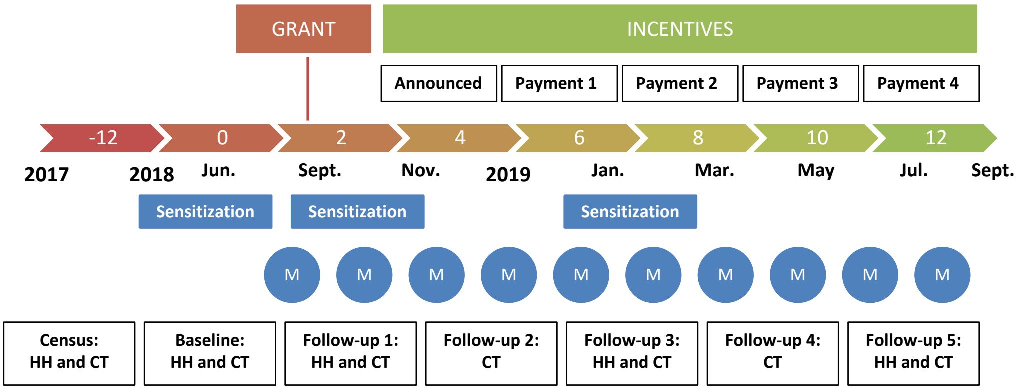
Figure S.2.1 shows the study location and the timeline of activities.

Figure S.2.1: Location and timeline of the study

A. Location



B. Timeline



Note. Panel A shows the location of the state of Uttar Pradesh and of Lucknow and Kanpur within the state. Panel B shows the timeline of activities. M indicates the delivery of voice messages. HH and CT indicates the collection of the household and CT surveys, respectively. Basemap source: Esri (see Supplementary Material S.4.9 for details and attributions). Details about data collection activities are reported in Supplementary Material S.4.

S.2.1 The maintenance intervention

The maintenance intervention consisted of two subsequent components: a grant and a financial reward. The **grant** offered three packages of similar monetary value from which the caretaker(s) could select one. *Deep cleaning* includes septic tank sewage removal, unclogging latrines and sewerage pipes, and cleaning walls, floors and inside toilets. *Repairs* includes sanitation/water connection repairs and/or infrastructure refurbishment. *Cleaning tools and agents* included four pairs of gloves, five floor cleaners, four toilet disinfectants, five liquid soaps, four toilet-cleaning brushes, two wipes, four nose masks, two brooms, two bucket and mop sets, three detergents, two hand-washing dispensers, two dustpans and two dustbins. The training was provided as a theoretical session followed by a practical session about cleaning practices. For CTs that selected repairs or deep cleaning, pictures of the CT area to be improved were taken before the work was done. Also, in this visit, a date was set for the works to be conducted.

Based on this information, our partner FINISH arranged and supervised the work with an external contractor, which was used in all CTs.

Figure S.2.2: Examples of grant use

A. Pre-grant B. Post-grant

Note. Example of deep cleaning of walls and repair of locks in a CT in Lucknow. Panel A shows the status before the intervention, while panel B shows the status after the deep cleaning. Source: Antonella Bancalari.

The **financial reward** was introduced in order to improve the quality of the service rendered by the caretaker. Two months after completion of the grant scheme works, we announced the financial reward scheme to caretakers in order to incentivize them to keep the CT clean. Caretakers could receive the following rewards: INR 500 conditional on soap availability in hand-washing facilities for both genders; INR 500 conditional on visible cleanliness of latrines, defined by whether cubicles were free from visible feces (both inside and outside the latrines); INR 1,000 conditional on bacteria counts being kept to a minimum standard (i.e., being below the median of the demeaned baseline distribution by city). Caretakers were informed that an external agent would return to measure each condition on a random day and time within the following two months, and that we would pay the financial reward depending on what the external agent measured. In CTs with more than one caretaker (20 percent of the sample), the financial reward was split among them. After two months and with a bi-monthly frequency, the conditions were verified by observers following a manual, and the incentives paid accordingly. In each round, we reminded the caretaker(s) of the conditions to be awarded the financial reward. In each payment round, we informed caretakers of their past cleanliness performance.

S.2.2 The sensitization campaign

The sensitization campaign targeted all members of study households, in particular the heads of participant households and their spouses. The campaign was designed in conjunction with our NGO partner FINISH to provide information that was accessible to participants with low literacy levels. We provided key messages regarding the risks of unsafe sanitation behavior and the importance of paying the fee to fund operation and maintenance of the CTs through four different means. First, **door-to-door visits** used a flip chart with cartoons and messages targeted at all household members, especially household heads and spouses. This session covered the following sections: how open defecation affects your community; how open defecation affects your family; benefits of using CTs; what you and your family can do to make the CT better; your rights when you pay the fee for using the CT. The cartoons were made by a local graphic designer considering the context of urban slums in India. Figure S.2.3 shows the flip chart cover used for the campaign, and an example of delivery.

Figure S.2.3: Door-to-door campaign

A. flip chart cover | B. Delivery of the campaign

Note. Panel A shows the cover of the flip chart used to communicate key messages to residents in slums. It translates from Hindi as 'Awareness campaign to encourage CT use and maintenance in India". Panel B shows a moment of the sensitization campaign, in which a household head and spouse pay attention to the flip chart during a household visit in Lucknow. Source: Morsel.

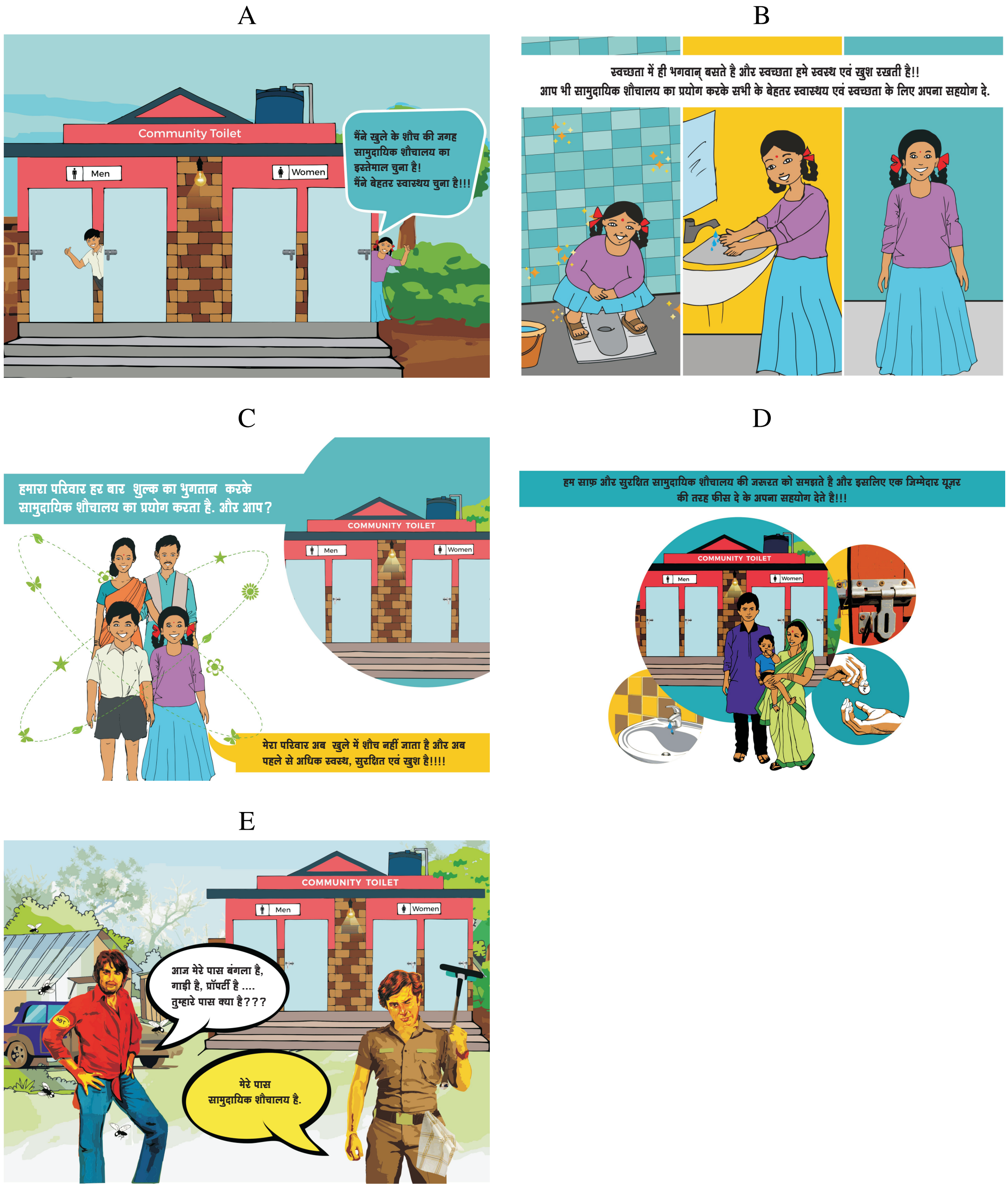
Second, the main messages of the door-to-door campaign were summarized into a four-page **leaflet** (Figure S.2.4) distributed among study households. The key messages provided during the door-to-door visits were also summarized in a series of **posters** (Figure S.2.5). We placed three medium-sized and two large posters in the entrance to CTs, in the area close to the hand-washing facilities and in each gender-specific area.

Figure S.2.4: Leaflet

Note. The figure presents the leaflet circulated during the sensitization campaign. The first page from the left presents the 'benefits of CTs' and includes: (1) improved sanitation facilities; (2) operation and maintenance of infrastructure; (3) safety with doors, locks and lights; (4) hand-washing facilities; and (5) gender-specific areas. The second page presents 'duties of users' and includes: (1) paying the fee to use the CT; (2) not throwing trash into the latrines; (3) flushing after using; (4) not spitting; (5) helping the elderly in the family; (6) accompanying females in the family during darkness; and (7) keeping the facility clean. The third page presents the 'rights of users' and includes: (1) caretakers not allowing free riders; (2) regular cleaning; (3) repairs; (4) respecting opening hours; (5) functional doors, locks and lights; (6) keeping men out of female areas; and (7) respecting and giving priority to females with children and the elderly. The final page, the cover, is the same as the one provided in the flip chart, shown in Figure S.2.3, and provides the title of the campaign.

Third, reminders in the form of **voice messages** were sent to participants' mobile phones. We sent a total of 10 rounds of voice message between month 1 and 11 of the study. The monthly frequency was chosen because it has been shown to be adequate to induce behavioral change (Cortes et al., 2021). This component was implemented using an purposely designed tracking app pre-populated with all mobile phone numbers. Households listened on average to 7 of the 10 monthly rounds of messages (panel A in Figure S.2.6), and listened to a good proportion of the message (panel B). More than 20 per-

Figure S.2.5: Posters placed on CT walls

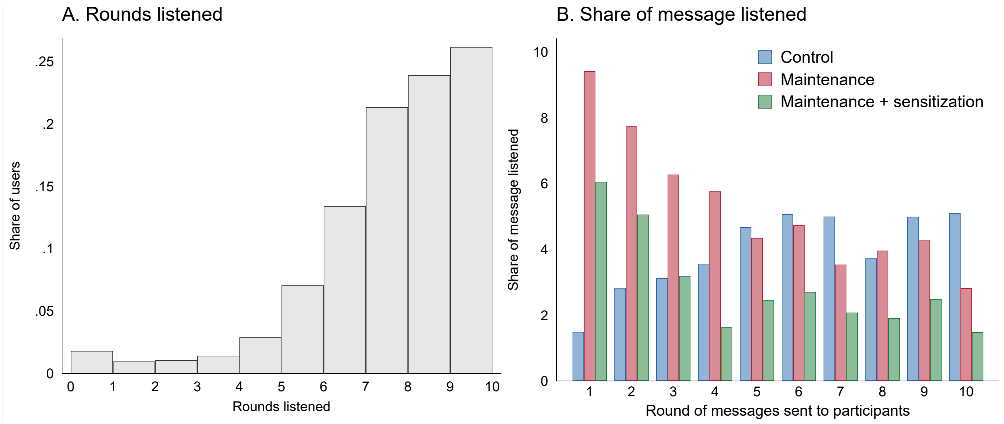


Note. The five posters placed on the walls of CTs read in Hindi: A, 'I choose to always defecate in CTs, I choose better health'; B, 'Health is happiness and cleanliness is godliness. Do your bit by using CTs'; C, 'We always pay and use CTs, do you? My family moved away from open defecation and now is healthier, safer and happier'; D, 'I value a clean and safe CT, that's why I pay the fee'; and E replicates a Bollywood scene but replacing the words to make it relevant to CTs. The villain, depicted as a dirty man says 'I have buildings, properties, vehicles, what do you have?' and the hero replies 'I have my CT'.

cent of the information messages highlighting public and private health risks of unsafe sanitation, as well as supply-side messages, were heard. Concerning voice message reminders, to disentangle the effects of receiving voice messages and the effect of receiving messages about the sensitization campaign, all study participants received the following voice message with no content related to the sensitization campaign: *the community toilet is open from early morning until late evening*. In addition, study participants in the maintenance treatment group received the message: *your community toilet has been granted aid to improve its quality. We hope you get to enjoy this better service.* Study participants in the maintenance plus sensitization group received the following messages:

> *Do you know open defecation is one of the biggest causes of diarrhea which can even kill your children? Adopting good sanitation behavior will ensure a healthier future for your family. / Open defecation is a big risk for your family's as well as your neighbors' health. Use community toilets to defecate instead of polluting and contaminating your community with open defecation. / Health is wealth! By not defecating in the open you are keeping your health safe and reducing expenses on medicines and treatment. / Cleanliness is godliness! By using community toilets, you are contributing towards the cleanliness and health of your community. / Do you know how unsafe it is for women and girls in your family to go for open defecation? Be the change and adopt the use of community toilets. / Using community toilets ensures dignity of women in your community. Women should not feel ashamed of going to community toilets… It is way better than open defecation. / Using community toilets improves the health of your children and keeps medicines and doctors away.*

Figure S.2.6: Voice messages



Note. Panel A indicates the number of rounds in which at least one household member answered the phone. Panel B shows the average share of the message that is listened to by participants, distinguishing across treatment arms and rounds. We correct for the differential duration of messages per treatment arm by multiplying the mean share of the message listened to by the maximum length per treatment arm.

S.3 Cost of intervention and quality scenarios

Table S.3.1 presents a summary of the cost associated with each activity falling under the maintenance (panel A) and sensitization interventions (panel B). Note that these are total costs throughout the project, while individual components have different timelines for implementation. Based on input from our implementing partner FINISH Society, as well as Lucknow Municipal Corporation, Table S.3.2 provides information on O&M costs for the median CT in our study sample, defined by its age (20 years), size (four female WCs, six male WCs and two urinals), and number of daily users (average of 150). Cost items include salaries for a caretaker and cleaner(s), cleaning supplies, as well as electricity and costs for minor repairs. Notice that an eligible household has the potential to provide monthly revenues of at least INR 600 if all members over 5 years of age use the CT once per day and pay the market fee of INR 5.

Table S.3.1: Cost of interventions

	Total expenditure		Cost per facility	
	INR	**US$**	**INR**	**US$**
A. Maintenance intervention				
Management	324,000	4,601	4,629	66
Implementation of grant scheme	1,688,500	23,678	24,121	343
Incentives for caretakers	267,000	3,792	3,814	54
Laboratory tests	210,000	2,982	3,000	42.60
Total	**2,489,500**	**35,352**	**35,564**	**505**
B. Sensitization intervention				
Management	81,000	1,150	2,314	32.86
Design and printing of material	50,000	710	1,429	20
Door-to-door campaign	440,770	6,259	12,593	179
Voice messages	21,662	308	619	8.79
Total	**593,432**	**8,427**	**16,955**	**241**

Note. For conversion of Indian rupees into US$, we assume an exchange rate of 70.42 INR/US$. The implementation of the grant component includes subcontracting, material for repairs, human resources, transportation and the overall management of the intervention. Door-to-door campaign includes transportation costs. Cost per facility is computed assuming 70 CTs in the maintenance intervention, and 35 in the sensitization intervention. Details about the interventions are provided in Supplementary Material S.2.1 and S.2.2 .

The monthly maintenance cost for the current scenario (which we term as 'status quo') is INR 10,200 (US$ 144.85). Under the current scenario, salaries represent 78 percent of the total budget, and cover the costs for a full-time caretaker and for one cleaner performing a daily routine clean. We consider one alternative cost scenario that was deemed to support an 'improved' maintenance level. Under this scenario, we assume that the number of users remains constant. The scenario introduces a higher salary for the caretaker (which allows hiring a more experienced caretaker), higher input costs, and a yearly investment into cleaning machinery, such as a pressurized water cleaner, which costs about INR 20,000 (US$ 284.01). This scenario leads to a total of INR 28,800 (US$ 408.97) per month, with salaries representing 63 percent of the total. It is important to note that we do not claim that this scenario is optimal, and it can be improved further. The table also shows cost per eligible household (see Supplementary Material S.4.1 for eligibility and proximity criteria), of which there are 34 in the median

Table S.3.2: Monthly O&M costs and grant and incentive costs per CT

	Maintenance level			
	Poor (status quo)		Improved	
	INR	US$	INR	US$
Panel A. O&M COSTS				
Salaries				
Caretaker (full-time)	5,000	71.00	12,000	170.41
Cleaner(s)	3,000	42.6	6,000	85.2
Supplies				
Cleaning agents	500	7.10	4,000	56.80
Cleaning equipment	200	2.84	2,200	31.24
Other				
Electricity	500	7.10	2,600	36.92
Minor repairs	1,000	14.20	2,000	28.40
Total	**10,200**	**144.85**	**28,800**	**408.97**
Total per eligible household	**300**	**4.26**	**847**	**12.03**
Panel B. INTERVENTION				
Maintenance grant				
Implementation	2,010	28.54		
Management	193	2.74		
Incentive scheme				
Amount paid to caretaker	477	6.77		
Management	289	4.11		
Laboratory tests	375	5.33		
Total	**3,344**	**47.49**		
Total per eligible household	**98**	**1.40**		
TOTAL (A + B)	**13,544**	**192.33**	**28,800**	**408.97**
TOTAL (A + B) per eligible household	**398**	**5.66**	**847**	**12.03**

Note. For conversion of INR into US$, we assume an exchange rate of 70.42 INR/US$. We assume that the grant is provided once a year and that incentives are provided on an ongoing basis every two months. We allocate 50 percent of total management cost to the maintenance grant implementation and 50 percent to the incentive scheme. To compute the total per eligible household, we consider the median number of households in the catchment area (34), and we assume no other household is using the CT.

CT. In panel B of Table S.3.2 we convert the total intervention expenditures of the maintenance intervention (Table S.3.1) into monthly expenditures. Adding these costs, the total monthly costs become INR 13,544 (US$ 192.33) per CT.

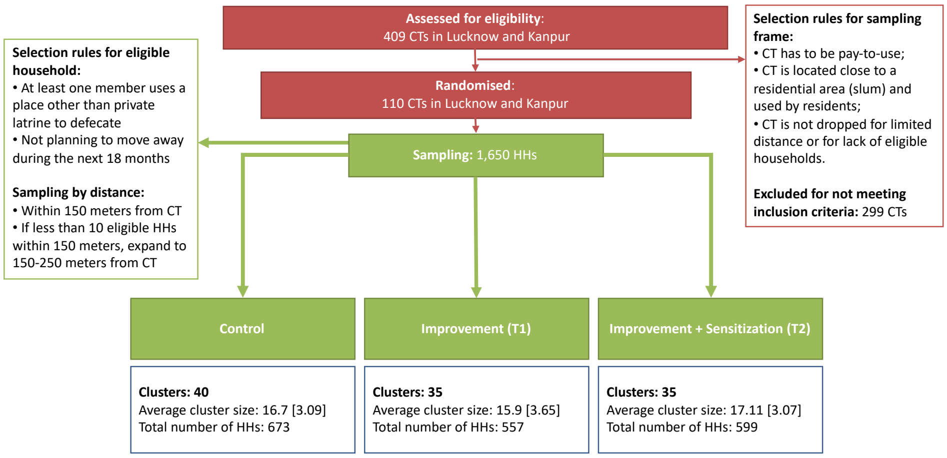
S.4 Sampling, data collection and measurement

S.4.1 CT census, slum resident census, and sampling strategy

Figure S.4.1 summarizes the sampling procedure for CTs and households. In order to obtain the sampling frame, during the first half of 2017, we performed mapping of slums and a census of all CTs in both study cities. The definition of *slums* follows official criteria used in the Indian Census. According to Government of India (2011), an *identified* slum is 'a compact area of at least 300 people or about 60-70 households of poorly built congested tenements, in unhygienic environment usually with inadequate infrastructure and lacking in proper sanitary and drinking water facilities.'

The census questionnaire was administered to caretakers and/or supervisors, and collected information on the geolocation and the main characteristics of CTs, such as main users, building characteristics, ownership, management structure, and payment system. We identified a total of 201 CTs in Lucknow and 208 CTs Kanpur. Out of these, we dropped CTs with the following characteristics: free to use, located outside slum areas, and used primarily by non-residents (generally located near market areas). To avoid cases in which residents can choose between different CTs, we drop CTs that are closer than 300 meters to each other, and CTs that have two other CTs closer than 350 meters. In addition, we also dropped very small complexes, i.e. CTs in whose catchment areas are living fewer than eight eligible households. This resulted in a total of 110 CTs, which were all selected for the study.

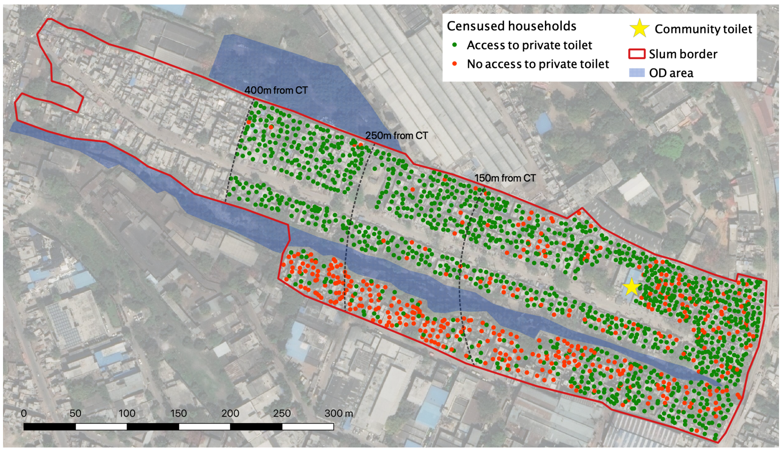
Figure S.4.1: Sampling frame definition and sampling procedure



Note. The flowchart summarizes the procedure followed for the selection of CTs and the sampling of households within their catchment areas. Details of the procedure are discussed in Appendix S.4.1.

The second level of the sampling frame is characterized by all households living in proximity to the CT without access to a private toilet. To identify these households, during the second half of 2017, we performed a census of all households living within slum borders and within 400 meters of the selected CTs. The distance bound was selected based on qualitative evidence about the maximum distance one person would walk to opt for CT use versus open defecation (Armand et al., 2020a). The questionnaire was administered to household heads and gathered information about demographic and dwelling characteristics (including geolocation), and sanitation-related behavior for more than 30,000 households. To identify potential users, we defined a household to be *eligible* for the study if all these conditions are met: the household lives in the catchment area of a selected CT, defined by the area within the slum and within 150 or 250 meters in straight distance from the CT building; at least one household member reports using a CT or a shared toilet, or practicing open defecation; the household does not intend to migrate during the 18 months following the census interview. Figure S.4.2 shows an example of this selection process. Figure S.4.3 provides the spatial distribution of CTs in the study.

Figure S.4.2: Definition of sampling frame: an example



Note. The figure shows an example of the selection process for constructing the sampling frame using a hypothetical slum. Each dot represents a censused household. The area within the slum border but more than 400 meters from the CT was not covered by the census. Distance bounds are computed as straight distance from the CT. Basemap source: Esri (see Appendix S.4 for details and attributions).

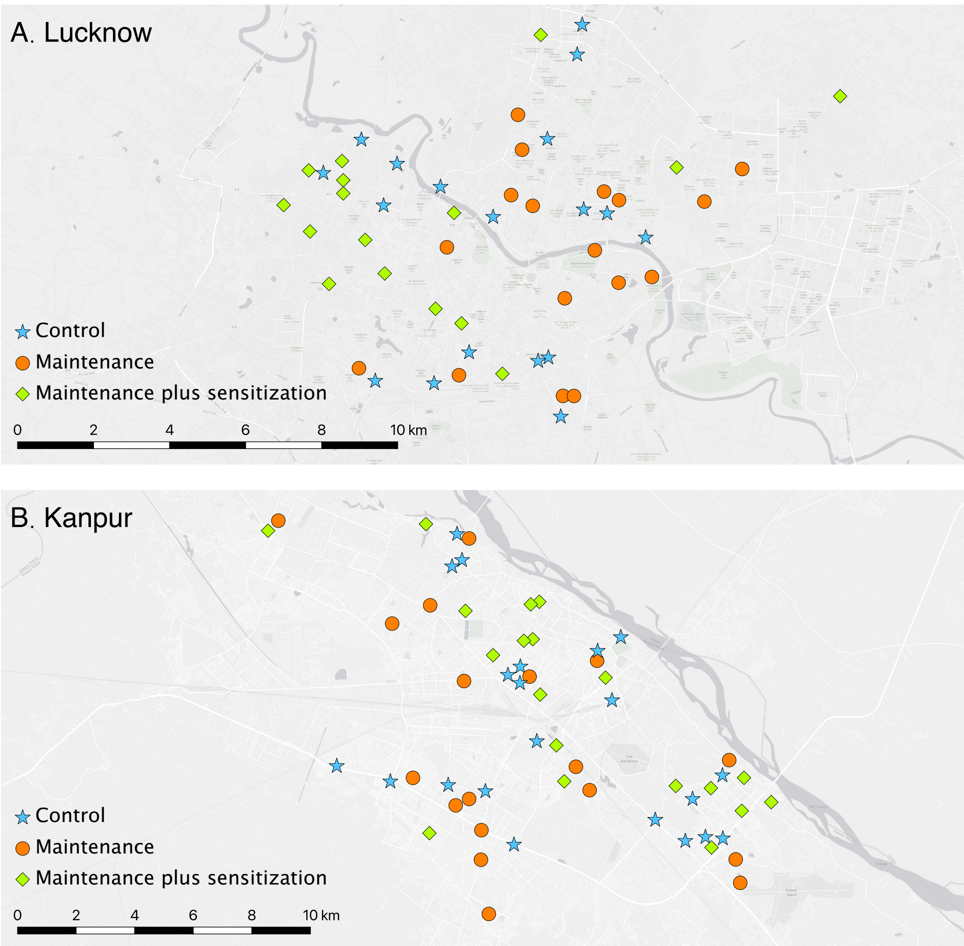
Within each of the 110 catchment areas, we sampled up to 17 eligible households. For catchment areas with fewer than 10 eligible households available within 150 meters, we selected all households within this bound, and randomly selected the remaining ones (up to 17 households) from the households living between 150 and 250 meters from the CT. In total, we obtained a sample of 1,650 households living in 110 catchment areas. Table S.4.3 provides a comparison between sampled households and the average characteristics of slum residents across all states of India and in Uttar Pradesh.

Table S.4.3: Descriptive statistics of slum populations

	2011 Census of India		**Study sample**
	India	Uttar Pradesh	Lucknow and Kanpur
	(1)	(2)	(3)
A. Share of the population			
Male	0.52	0.53	0.53
Female	0.48	0.47	0.47
Children (0-6 y.o.)	0.12	0.14	0.09
Scheduled caste	0.20	0.22	0.45
B. Other characteristics			
Sex ratio (female to male)	1.08	1.12	1.12
Literacy rate	0.78	0.69	0.46

Note. The table provides descriptive statistics for the slum population in India in Column (1), for the slum population in Uttar Pradesh in Column (2), and for the study sample in Column (3). The source for Columns (1) and (2) is the 2011 Indian Slum Population Census (Government of India, 2011).

Figure S.4.3: Geographical distribution of CTs, by city and treatment group



Note. Panel A shows the geographical distribution of CTs selected for the study in the city of Lucknow. Panel B shows the geographical distribution of CTs selected for the study in the city of Kanpur. Details about the procedure to select CTs is provided in Supplementary Material S.4.1. Basemap source: Esri (see Supplementary Material S.4.9 for details and attributions).

Table S.4.4: Selected CTs and households, by treatment arm and city

	Control		**T1**		**T2**		**Total**	
	N (1a)	% (1b)	N (2a)	% (2b)	N (3a)	% (3b)	N (4a)	% (4b)
A. CTs								
Lucknow	19	36.5	17	32.7	16	30.8	52	100
Kanpur	21	36.2	18	31.0	19	32.8	58	100
Total	40	36.4	35	31.8	35	31.8	110	100
B. Households								
Lucknow	255	35.5	225	31.3	239	33.2	719	100
Kanpur	321	37.5	262	30.6	273	31.9	856	100
Total	576	36.6	487	30.9	512	32.5	1,575	100

Note. The table presents the distribution of selected CTs (panel A) and households (panel B) by treatment arm and city. *T1* indicates the ‘maintenance’ treatment group, and *T2* indicates the‘maintenance + sensitization’ treatment group.

S.4.2 CT surveys and observation

During regular unannounced visits to the CTs, we administered a questionnaire to the caretaker to collect data on cleaning practices, CT management and time allocation to different tasks. In addition to self-reported information from the caretakers, we also gathered information about the condition and cleanliness of CTs using observers. To provide uniform reports from observers, the data collection manual defined conditions for the visual evaluation of the state of the CT. Observers also recorded the number of users and the share of users who pay the fee for the duration of 1 hour using manual counters. We collected CT-level data in a sequence of six waves.

We collected data for all of the 110 selected CTs at the baseline, but only for 108 in follow-up 1, 109 in follow-up 2, 107 in follow-up 3, 105 in follow-up 4 and 106 in follow-up 5, given that some CTs closed temporarily/permanently for refurbishment, and one slum was completely displaced after follow-up 2. In some cases, we were able to collect observations and bacteria swabs, while not being able to survey caretakers. In 92 percent of CTs, we surveyed caretakers in all five follow-up rounds. In addition, two new CTs opened very close to the study toilets, both allocated to the maintenance (T1) treatment arm. Because households in the catchment area also used these new CTs, we collected data from these toilets during follow-ups 2 to 5. These new CTs did not increase the number of clusters, since we consider them part of the same cluster as the old CTs given their close proximity. Table B4 shows that the number of CT observations and caretaker surveys and the addition of CTs in study catchment areas are orthogonal to treatment allocation.

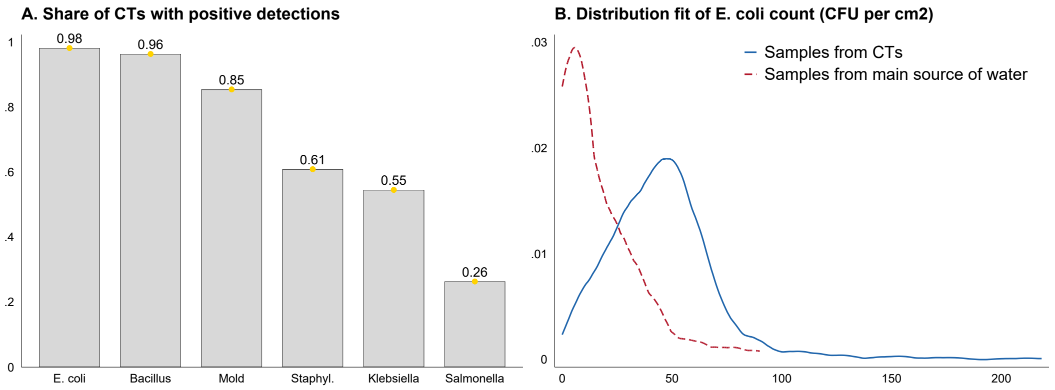
S.4.3 Laboratory tests

We collected data about bacteria and mold presence using samples analyzed in the laboratory. We first focus on the presence of the species *Escherichia coli* (*E. coli*) of genus *Escherichia*, an indicator of fecal contamination, measured as bacteria count (CFU per cm^2) using the arithmetic mean among all samples collected in a CT during a measurement round (see, e.g., WHO, 2017). We also compute the presence of potentially harmful bacteria of the genus *Bacillus*, genus *Staphylococcus*, genus *Klebsiella*, and genus *Salmonella*. For further information on the effect of bacteria on human health, see Jenkins and Maddocks (2019). In addition, we test for the presence of mold, which can cause allergic reactions and respiratory problems (Gent et al., 2002).

To implement these measurements, we prepared a protocol in conjunction with a laboratory based in Lucknow, which analyzed the samples. For each CT and during each survey round, three samples were collected using swabs in specific locations of the facility based on evidence about the microbial bio-geography in public toilets (Flores et al., 2011). CTs were first randomized into two groups: a *male* group, in which the swabs were collected in the male area of the CT throughout the study, and a *female* group, in which the same was performed in the female area of the CT. During each visit, the enumerator collected three samples. The first two samples were collected from the floor of the cubicles at the mid-point between the entrance wall and the latrine/water. Cubicles were randomly selected by the research team in each round to avoid the caretaker focusing on a specific point in the CT. A third sample, aimed at collecting information about the area where most people walk, was collected from the floor where one would take one's first step to enter the cubicle hallway.

At baseline, we also collected information about access to clean water. We collected and analyzed two samples of water for each catchment area. During the baseline survey, we asked households about their main source of water, and we then collected water samples from up to two randomly selected sources. Figure S.4.4 shows descriptive statistics at baseline for these measurements.

Figure S.4.4: Bacteria and mold detection at baseline



Note. Panel A presents the share of CTs where each bacteria type or mold was detected in at least one of the three samples. Panel B shows the distribution of the E. coli count from CT and water samples. The distribution fits are estimated non-parametrically using kernel density estimation assuming an Epanechnikov kernel function. Bandwidths are estimated by Silverman's rule of thumb (Silverman, 1986).

S.4.4 Slum resident survey

We collected household-level data in a sequence of four waves (refer to Figure S.2.1 for the timing and label of each wave). This was a standard household survey, collecting information on demographic and socio-economic characteristics, such as household composition, dwelling characteristics, assets, income and expenditure. This information was supplemented by a section on health and sanitation behavior, including attitudes and expectations associated with different sanitation practices. At baseline, we further collected information on child health for children under the age of 6, and on intra-household dynamics between spouses. The instrument, including all modules, has an average duration of one hour. All follow-up surveys aimed to recollect some of the same information collected at baseline. Some information collected at baseline was not collected during the follow-up surveys, such as detailed information on child health and childcare, and intra-household decision-making. Some new information was collected after the baseline survey, such as information related to exposure to the interventions.

The respondent is the main decision maker in the household. We select the respondent using the following rules: if the household head is present, then the respondent is the household head; if the household head is absent, then the respondent is the spouse of the household head; if the household head and spouse are both absent, the household is revisited; if the household head and spouse are both absent during the revisit, then the respondent is the most senior member (over 18 years old) who is actively participates in the household's decision-making. At baseline, we also interviewed the spouse of the household head to gather information about intra-household decision-making, and the primary caregiver to collect information on child health for children aged 5 years or younger.

In total, we interviewed 1,575 households at baseline (an average of 12 households per cluster), 1,532 households during follow-up 1, 1,578 households at follow-up 3, and 1,772 households in the follow-up 5. Each baseline household was interviewed in 2.6 out of 3 follow-up measurements, with only 2 percent of baseline households that was never re-interviewed at follow-up. In terms of attrition from baseline to a specific follow-up surveys, the rate is 9 percent for follow-up 1, 19 percent for follow-up 3, and 14 percent for follow-up 5.[2] Columns (1)–(5) in Table B5 estimates the probability of attrition for each of these indicators as a function of the treatment status. Attrition does not differ between treatment and control groups for any of the attrition indicators. In order to maintain a comparable sample size in all follow-up surveys, we handled attrition with replacements at random using the sampling frame used for the baseline sampling. Column (6) tests whether the replacement was introduced differently across treatment arms. In total, 16 percent of follow-up observations are replacements, with no statistical difference across treatment arms.

The questionnaire for follow-up 5 was supplemented with a list randomization tech-

[2]The attrition rate was the highest at follow-up 3 because the survey coincided with school vacations, a period when a share of study participants goes back to their native villages.

nique. Respondents were randomly allocated to one of four groups. Depending on the group, respondents received a different list of statements, and were asked to report how many of them were true. Table S.4.5 provides the list of statements. Group A received only a list of statements related to general behavior. Groups B–D received the same list and one extra statement capturing sensitive behavior.

Table S.4.5: Statements used for list randomization

Group A	Group B	Group C	Group D
- I cooked yesterday	- I cooked yesterday	- I cooked yesterday	- I cooked yesterday
- I bought milk yesterday	- I bought milk yesterday	- I bought milk yesterday	- I bought milk yesterday
- I watched TV yesterday	- I watched TV yesterday	- I watched TV yesterday	- I watched TV yesterday
	- I defecated in the open yesterday	- I used the CT to defecate yesterday	- I washed my hands with soap yesterday

Note. Group A reports a list of statements related to general behavior. Groups B–D provide the same list, but adding one extra statement capturing sensitive behavior (OD, use of CT, or hand-washing).

Using the information from the list randomization questions, Figure S.4.5 shows the share of study participants practicing OD, using the community toilet and hand-washing with soap, respectively, in the day previous to the interview. We observe a small insignificant increase in service use driven mostly by the sensitization campaign. Although the list randomization technique does not identify individual behavior, this result can be explained by a shift away from service use and into OD, balanced by a shift away from unimproved (private) sanitation and into CT use. In addition, users could have practiced both.

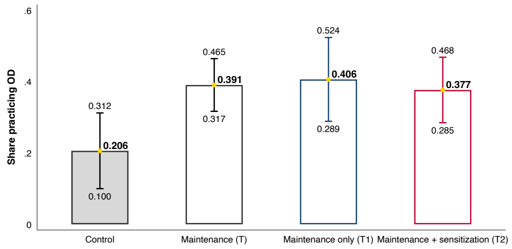
Self-reported sanitation behavior was measured by asking survey respondents where each demographic group defecated the last two times. To prevent under-reporting of open defecation due to social stigma, we included the following prelude: ‘I've been to many similar communities and I've seen that even people owning latrines and having nearby community toilets defecate in the open.’

S.4.5 Lab-in-the-field experiment: WTP for service use

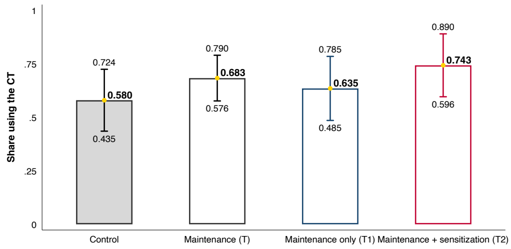
WTP for service use is elicited to the respondent of the household survey and the spouse (up to two respondents per household), and is measured 4 times during the study in conjunction with the household survey. WTP is elicited using a standard incentivized version of the multiple price list (or take-it-or-leave-it) methodology. Participants were prompted to choose between different amounts of cash (ranging from INR 0 to 60 with increases of INR 5) and a bundle of 10 tickets to use the CT in the catchment area where they live. During follow-up 5, following the introduction at the end of the study of monthly passes for families in a limited number of CTs, we also elicited the WTP for a monthly family pass for up to five members. In total, participants face 13 combinations.

Figure S.4.5: Treatment effects on list randomization outcomes

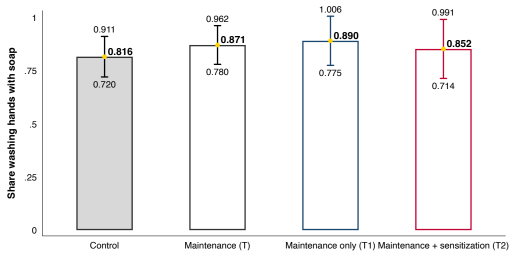
A. OD



B. CT use



C. Hand-washing with soap



Note. Confidence intervals are built using statistical significance at the 10 percent level and assuming errors are clustered at the level of the catchment area. Randomization of lists was performed at individual level, and data were collected during follow-up 5 only. Supplementary Material S.4 provides details about measurement.

After all choices are made, one of the options is then randomly selected by drawing a numbered ball from a bag, and the decisions are realized. Following the realization of the game, in the case of the bundle of tickets being assigned, the respondent could allocate the 10 tickets or some of them to either male or female use. Before participating in the game, the participant was introduced to a practice round of the game using a bar of soap to facilitate familiarity with the rules. The exact explanation of the game read by the enumerator to the participant was as follows:

> *Now let us do the prize draw for 10 tickets to use the [CT name]. These tickets are being officially provided by [CT name] as a promotion to encourage people to use the CT. They can be used at any time in the next 2 months. You will be given the choice later to decide how many of the 10 tickets you would like to be for men and boys, and how many you would like to be for women and girls. We are going to ask you to make a series of choices between either receiving these 10 tickets or instead receiving amounts of cash. At the end of all of the choices, you will draw a ball from a bag to determine which one of these choices will be randomly selected for your lucky draw – you will get the tickets or the money, depending on what you chose. This means that any one of the choices that you make could be selected at the end. Therefore it is in your best interest just to answer your honest opinion about which option you would prefer in every single choice.*

In conjunction with the incentivized version of the WTP elicitation, we also collected information from participants about the price that female and male residents face to use the CT, and we asked about WTP for the use of a hypothetical higher-quality CT in a non-incentivized setting. For the latter, participants were asked directly whether they would purchase a ticket for different amounts of money, ranging from INR 0 to 10, with increases of INR 1. The exact question reads as follows:

> *Now I want to tell you an imaginary story. Imagine that starting from tomorrow, the owners of the nearest CT decided to change the price for using the defecation cubicles. At the same time, they would improve the quality of the CT to the highest standard, ensuring it was very clean, had good hand-washing facilities, and that all the cubicles had a light and a lock. Would you be willing to buy a ticket to use the defecation cubicles of the community toilet, if the price was...*

S.4.6 Lab-in-the-field experiment: adapted dictator game

To measure preference for maintenance among potential users, we played an adapted dictator game in which participants are endowed with INR 50 and are given the option to donate all or part of it to a fund to purchase cleaning products for the CT. This component was administered to the respondent of the household survey and the spouse (up to two respondents per household), and measured in conjunction with each household survey. Having collected all the contributions to the cleanliness of the CT within each slum, the total amount was used to purchase cleaning products, which were then delivered to the caretaker. The exact setting reads as follows:

> *I would like to inform you that as an additional thank-you for participating in this study, you will receive an extra INR 50 in cash. We are asking all participants to choose between keeping some or all of this INR 50 for themselves, and donating some or all of this INR 50 for a special fund for cleaning products that we will deliver to the CT. How would you like to split the INR 50 between cash for yourself, and donation to the cleaning product fund for your CT?*

Similarly, to measure pro-social motivation for the cause among caretakers, we implemented an adapted dictator game in which the caretaker is endowed with INR 50 and is given the option to donate all or part of it to fund a sanitation project implemented by our partner, FINISH Society. Pro-social motivation among caretakers was measured during each CT survey. Having collected the contributions from all caretakers, the total amount was donated to the FINISH Society project. The exact setting faced by the caretakers reads as follows:

> *I would like to inform you that as a thank-you for participating in this study, you will receive INR 100 in cash. You can keep the full amount for yourself or you have the opportunity to donate some or all of it to FINISH Society to help with improving water access, sanitation and hygiene in disadvantaged areas of India. How would you like to split the INR 100 between cash for yourself and donation to charity?*

S.4.7 Lab-in-the-field experiment: public goods game (PGG)

We implemented a standard public goods game with the experiment participants. The game is based on the voluntary contribution mechanism, in which participants receive an endowment of INR 100, and they have to decide whether to keep the endowment or to invest part or all of it in a public pot. The contributions in the group are increased by a multiplier and shared equally among participants. The multiplier is randomly varied at catchment-area level to either double or triple the contributions. The game is designed so that while the total return to the investment in the pots is higher than the return from keeping the endowment, there is no incentive to invest in the former because of the higher individual pay-off that can be obtained from keeping the endowment. The dominant strategy is therefore to not contribute at all, while the social optimum is to invest in the pot. We played simultaneously with three groups of equal size (ranging from four to six participants) in each community. Participants also received INR 20 as show-up fee. The instructions given to groups of six participants are the following (note that x is either 2 or 3):

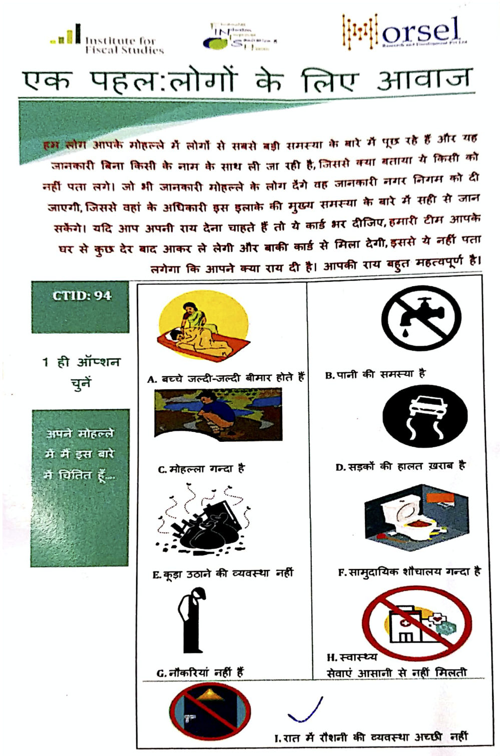
> *In this game, each player receives an endowment of INR 100 and you can choose to contribute (C) to the shared pot or keep (K) it. Out of the INR 100, you can decide how much to contribute and how much to keep. Secretly, you will put your donation amount in the pink envelope and the amount you want to keep in the blue envelope. All contributions will be summed and we will increase the total contribution by [x]. The final pot will be split equally among players. Let's look at some examples. If all 6 players contribute the INR 100, their individual payoffs would be equal to INR [$600 \cdot x/6$]; if one player contributes and other players keep the endowment, then the payoff of each player contributing is equal to INR [pot $\cdot\, x/6$], and the payoff of the player keeping is equal to INR [$100 +$ pot $\cdot\, x/6$]; if all players keep, then their individual payoffs are INR 100.*

S.4.8 SCA: voice-to-the-people initiative

We distribute a card (Figure S.4.6) to participants asking about the most pressing issue in their community. Participants were invited to circle only one option among the following: children are frequently ill, water availability is limited, the community is dirty, the quality of roads is poor, there is no waste collection, the CT is dirty, jobs are missing, access to healthcare is limited, and lighting at night is poor. We used visual representations to facilitate selection of the issue among illiterate participants. Individual anonymized codes allowed the returned cards to be matched with the households in the sample. Two letters with the results for the corresponding cities were sent to the heads of sanitation and environment at the municipal corporations of Lucknow and Kanpur in October 2019. The specific instruction reads as follows:

> *We are collecting anonymous reports about the most pressing issue in your community. We will communicate this to the district municipal corporation to raise awareness among administrators. If you want to tell us your opinion, fill in the card and return to one of our team members. Your voice is important! In my community, I am most concerned about: [circle only one option]*

Figure S.4.6: Card distributed for the voice-to-the-people initiative



Note. The figure shows the card distributed to participants as part of the voice-to-the-people initiative. The options read in Hindi: children frequently ill (A); limited water availability (B); community is dirty (C); poor quality of roads (D); no trash collection (E); CT is dirty (F); no jobs (G); limited access to healthcare (H); poor lighting at night (I). In each household, up to two participants had the option to mark one of the issues.

S.4.9 Additional data sources

Basemaps throughout the paper were created using ArcGIS® software by Esri®. Basemaps are used in line with the Esri Master License Agreement, specifically for the inclusion of screen captures in academic publications. We use the *World Light Gray Base* (sources: Esri, HERE, Garmin, ®OpenStreetMap contributors, and the GIS User Community).

Appendix and Supplementary Material Bibliography

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, 9(4):383–405.

Armand, A., Augsburg, B., Bancalari, A., and Trivedi, B. (2020). Community toilet use in slums: Willingness-to-pay and the role of informational and supply side constraints. *3ie Impact Evaluation Report 113*.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1179–1203.

Athey, S. and Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. *Observational Studies*, 5.

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 115(8):1943–1948.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2017). Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. Technical report, National Bureau of Economic Research.

Cortes, K. E., Fricke, H., Loeb, S., Song, D. S., and York, B. N. (2021). Too little or too much? Actionable advice in an early-childhood text messaging experiment. *Education Finance and Policy*, 16(2):209–232.

Flores, G. E., Bates, S. T., Knights, D., Lauber, C. L., Stombaugh, J., Knight, R., and Fierer, N. (2011). Microbial biogeography of public restroom surfaces. *PLoS ONE*, 6(11):1–7.

Gent, J. F., Ren, P., Belanger, K., Triche, E., Bracken, M. B., Holford, T. R., and Leaderer, B. P. (2002). Levels of household mold associated with respiratory symptoms in the first year of life in a cohort at risk for asthma. *Environmental Health Perspectives*, 110(12):A781–A786.

Gordon, D., Howe, L. D., Galobardes, B., Matijasevich, A., Johnston, D., Onwujekwe, O., Patel, R., Webb, E. A., Lawlor, D. A., and Hargreaves, J. R. (2012). Authors' response to: Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low- and middle-income countries: The case for multiple correspondence analysis. *International Journal of Epidemiology*, 41(4):1209–1210.

Government of India (2011). Census of India 2011. Office of the Registrar General and Census Commissioner, Ministry of Home Affairs, Government of India, New Delhi, India.

Jenkins, R. and Maddocks, S. (2019). *Bacteriology methods for the study of infectious diseases*. Academic Press.

Kline, T. (2014). *Psychological Testing: A Practical Approach to Design and Evaluation*. Sage Publications.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

WHO (2017). Guidelines for drinking-water quality, 4th edition: 1st addendum. Guidelines Review Committee – Water, Sanitation, Hygiene and Health, World Health Organization April 2017, World Health Organisation.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.