Public Service Delivery, Exclusion, and Externalities[*]

Alex Armand Britta Augsburg Antonella Bancalari Maitreesh Ghatak

March 30, 2025

Abstract

When public services are funded through user fees, incentivizing providers to improve quality while holding nominal fees constant can have ambiguous welfare effects. We identify a novel channel, both theoretically and empirically, where incentives to enhance quality also intensify payment enforcement, leading to user exclusion. Using a randomized controlled trial in the context of water and sanitation services in the slums of two large Indian cities, we find that service quality and fee compliance improve as providers increase efforts in both maintenance and monitoring. However, heightened monitoring excludes users, generating negative externalities. These findings underscore the need for financing models that better align provider incentives with equitable access. (*JEL* C93, H40, I15, Q53)

Keywords: public service, exclusion, externality, maintenance, user fees, payment, water and sanitation, health.

[*] **Armand**: Nova School of Business and Economics – Universidade Nova de Lisboa, CEPR, IZA, NOVAFRICA, Institute for Fiscal Studies (e-mail: alex.armand@novasbe.pt); **Augsburg**: Institute for Fiscal Studies (e-mail: britta_a@ifs.org.uk); **Bancalari**: Institute for Fiscal Studies, IZA (e-mail: antonella.bancalari@ifs.org.uk); **Ghatak**: London School of Economics, BREAD, CEPR (e-mail: m.ghatak@lse.ac.uk). We are grateful to Nava Ashraf, Orazio Attanasio, Oriana Bandiera, Rachel Cassidy, Joan Costa-Font, Joshua Deutschmann, Pascaline Dupas, Greg Fisher, Sebastian Galiani, Paul Gertler, Anders Jensen, Michael Kremer, David Lagakos, Eliana La Ferrara, Meera Mahadevan, A. Mushfiq Mobarak, 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, 2023 BSE Summer Forum, Econometric Society Winter School, Yale, LSE, IFS, UCL, CSAE Oxford, University of St. Andrews, University of Essex, 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, Valentin Post, 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). Bancalari's work was supported by an LSE Doctoral Scholarship. 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). The presentation of the results does not follow the order presented in the plan; all results are present either in the main text or in the appendix.

1 Introduction

Essential public services and the role of state capacity in their provision are fundamental to economic development and social welfare (Besley and Ghatak, 2006). However, with limited resources at their disposal, governments face a trade-off between guaranteeing an adequate quality of basic services and expanding their accessibility. Improving the quality of services requires additional resources, which are often collected through user fees. Charging fees for essential public services, predominant in low and middle-income countries (LMICs), can exclude individuals who are unwilling or unable to pay (e.g. Gertler et al., 1987; Dupas, 2014a). Excluded individuals are therefore forced to resort to inferior outside options that often generate negative externalities, such as unsafe healthcare, open defecation, illegal waste disposal, school dropout, or reliance on polluting fuels instead of electricity networks. Since regulating such externalities is a key role of government (Besley and Persson, 2009), policymakers should aim at expanding access by reducing constraints on users, such as lowering fees. However, this option comes at the potential cost of lower quality due to congestion externalities and reduced revenue.

While the literature often frames the quality-accessibility trade-off solely as a function of government budget constraints, service delivery ultimately depends on the behavior of providers. Understanding how their incentives shape this trade-off is therefore crucial. Over the past two decades, a large literature has opened the black box of state capacity by studying the personnel economics of the public sector—from bureaucrats to front-line service providers (Bandiera et al., 2024)—focusing on how selection, incentives, and monitoring shape performance and service quality. In particular, a growing body of evidence shows that incentives can improve outcomes in various public services, although they may also have unintended effects due to multitasking considerations (see Finan et al., 2017 for a review). However, we know very little about this core issue of the trade-off between quality and accessibility in public service delivery when we take into account incentives of service providers, especially when they have no direct control over user access (i.e., user fees are fixed).

In this paper, we investigate how incentivizing service quality among providers affects the quality of public services, the risk of user exclusion, and the potential externalities that may arise as unintended consequences. We combine a theoretical framework with empirical evidence from a large-scale randomized controlled trial (RCT) that improved the quality of an important basic service by offering providers incentives based on measured quality, as well as direct grants. We go beyond the well-understood quantity-quality trade-off frontier, as the nominal user fee is held constant. We show that efforts to increase quality through incentives may result in user exclusion and negative externalities. Combining theoretical insights with rich data to investigate the dynamics between providers and users, we uncover the mechanisms behind this somewhat surprising result.

We model the delivery of a fee-funded public service from the perspective of a policymaker whose objective function accounts for both the social benefits of increased usage and the mitigation of negative externalities associated with the outside option. To achieve this goal, the policymaker contracts a provider to operate the service by exerting two types of effort: monitoring user fee payments, which determines the probability of fee collection, and conducting service maintenance, which determines the level of quality. Demand depends positively on quality and negatively on user fees, which are assumed to be exogenously given, and payment monitoring.

One might expect that incentivizing higher quality in a public service while holding nominal fees constant would have positive effects on demand. However, in the presence of multidimensional agency problems, this is not necessarily the case. Our model is similar in spirit to the classic multitasking model by Holmstrom and Milgrom (1991), which demonstrates that when an agent's role involves multiple dimensions subject to incentive problems, incentivizing one dimension may cause the agent to shift effort away from another—potentially more valuable but less accurately measured—task, resulting in efficiency loss. In contrast, we do not assume that the provider's tasks vary in measurability and allow the efforts between tasks to be independent in the agent's cost function. In particular, we uncover a complementarity across the two efforts arising from the demand for the service. Improvements in quality (through maintenance effort) increase demand, which increases the returns to monitoring fee payment. In contrast, a greater effort to collect fees enhances the resources available to maintain quality. This complementarity explains why incentivizing higher quality can increase user exclusion despite quality improvements, potentially resulting in negative welfare effects when negative externalities are present.

We provide empirical evidence on these mechanisms through an RCT that exogenously changes the quality of water and sanitation services in the two main cities of Uttar Pradesh (UP), India's largest state (Government of India, 2011). The experiment revolves around community toilets (CTs). Common in many LMICs, these public services operate with user fees and provide essential hygiene and sanitation through communal facilities, mainly serving specific groups of residents in informal settlements (or slums). In this challenging context, where overcrowding and inadequate housing limit private toilets, CTs remain the only alternative to unimproved facilities or open defecation (OD), which impose severe negative health, human capital, and environmental externalities.[1] Therefore, the social benefit of CTs is extremely high. Consider that in 2020,

[1] An extensive literature highlights the direct and indirect consequences of lacking access to sanitation. Examples from the economics literature include Miguel and Kremer (2004), Bleakley (2007), Adukia (2017), Augsburg and Rodríguez-Lesmes (2018), Coffey et al. (2018), Geruso and Spears (2018), and Spears (2020). The literature on strategies to reduce OD focuses mainly on private investments rather than public services (Guiteras et al., 2015; Lipscomb

half of the estimated 3.6 billion people worldwide who lacked access to safe sanitation resided in urban areas, India being one of the main contributors to this statistic (World Health Organization, 2021).

CTs are ideal for testing our theoretical predictions, as they are widespread in India and typically charge uniform fixed fees per use. The policymaker–provider interaction is simplified by the fact that the provision of services is often carried out by a single person, the caretaker, who maintains the facility and collects fees.[2] Furthermore, improvements in CT quality are salient, as facilities are often in extremely poor conditions, resulting in a low willingness to pay (WTP) and non-payment of user fees.

The design of the experiment is as follows. Subsequent to an extensive mapping of the universe of CTs and the slums they serve, we randomly assigned 70 of the 110 CTs present in the two cities to a treatment group, while the remaining 40 served as the control group. The intervention in the treatment group, called maintenance treatment, targeted the caretaker and aimed to exogenously change the effort put into maintaining the quality of the service. We boosted this effort by providing a one-off grant to renovate the facilities in the first two months of the intervention and a large bimonthly financial reward (approximately 40% of the caretaker's monthly salary) to incentivize cleanliness in the following 10 months. The primary objective of the intervention was to improve the quality of the service, while none of the intervention components were designed with the explicit aim of enforcing fee compliance.

We assess treatment impacts using a unique set of measurements, combining observations, survey responses, and incentivized behavioral measurements. Starting in April 2018, and over a period of 18 months, we collected observations of the quality of the service, the number of users during the peak hours, and the payment of user fees. To map intervention effects into behavioral responses, we complemented these objective measures with panel survey data and behavioral measures for both caretakers and slum residents living in the catchment areas of CTs.

We find that the maintenance treatment leads to sustained improvements in service quality (10.4% higher than the control group mean). Caretakers who undergo this treatment not only improve their maintenance efforts (11.9% better cleaning), but also dedicate a greater share of their time to monitoring activities that enable fee collection (9.5% greater). These results are consistent

and Schechter, 2018; Cameron et al., 2021; Gautam, 2023; Gautam et al., 2025), and information campaigns (Augsburg et al., 2022; Cameron et al., 2022; Abramovsky et al., 2023). The externalities of OD have been found to be more detrimental to child health in areas where people live more closely together (Hathi et al., 2017). Furthermore, recent evidence from public service delivery highlights how decentralization can generate water pollution externalities (Lipscomb and Mobarak, 2017).

[2] Especially in LMICs, where available technology limits alternative ways of collecting fees, both tasks are performed by the same provider. Examples include managers of health clinics, early childhood development centers and schools, caretakers of community water stations, and microgrid electricity hubs, and operators of minibuses for public transportation and pay-per-use garbage collection.

with a complementarity between providers' efforts to monitor fee payments and maintain quality, as proposed by our model.

Ex ante, the effect of increasing service quality on demand is theoretically ambiguous. We find a decrease in the observed number of users, primarily driven by residents rather than passersby (8.2% fewer resident users than the control group mean). This finding is accompanied by a notable reduction in the self-reported number of daily service uses by residents (9.7% lower). At the same time, the maintenance treatment increases the proportion of users who pay the fee by 17.9% compared with the control group, reducing free-riding in the use of the service. This pattern is not driven by an increased WTP for the service among residents.

Our findings suggest a mechanism whereby the higher costs to users, resulting from increased monitoring efforts by the caretaker and a higher likelihood of fee payment, outweigh the benefits of greater maintenance efforts. Consequently, more users are excluded from the service than are attracted by improved quality. This finding relies on the assumption that the fee collection system is imperfect and subject to agency problems by the provider. In a scenario where entry is perfectly regulated through mechanized means—which, in practice, still imperfectly control access and require supervision (e.g., users jumping barriers in public transport)–improving quality under fixed fees may lead to congestion, causing some users to forgo the service. Although this may limit the welfare gains from the intervention, it does not increase exclusion relative to the pre-intervention baseline.

We rule out alternative mechanisms that might otherwise explain changes in use and quality, including the possibility that users are discouraged from using the service due to congestion resulting from improved quality. Furthermore, we rule out the possibility that the results are driven by changes in the timing of use that could affect our observations, caretaker absenteeism, opening hours, or temporary facility closures.

User exclusion is consistent with data on residents' reliance on the outside option, OD. Due to the sensitive nature of this behavior, we use the list randomization technique to measure the prevalence of OD, reducing social desirability bias and increasing the reliability of responses (see, e.g., Karlan and Zinman, 2012). At the time of the endline survey, the average proportion of respondents who had practiced OD the day before the interview was 21.0% in the control group. The maintenance treatment more than doubled this prevalence, with a mean prevalence of 43.8%. The larger effect on the prevalence of OD compared to the effect on observed service use is due to the fact that the latter focuses only on use during the peak hours, whereas the former captures patterns of use throughout the day. The well-documented negative externalities of increased OD are confirmed in our setting by a rise in self-reported health problems. Residing near a facility in the maintenance treatment group increases the likelihood of incurring medical expenses for illness (7.4% higher than the control mean).

To understand how incentivizing service quality interacts with exogenous increases in demand, we complemented the maintenance intervention with a sensitization campaign to raise awareness of the negative externalities of OD. We randomly selected half of the facilities in the maintenance treatment group and implemented the campaign among residents living in the slum they serve. In the other half, we implemented only the maintenance intervention. Despite the existence of various initiatives underway at the time of the intervention, including the Indian Government's initiative to end OD through awareness creation and subsidy provision—the Swachh Bharat Mission—we find that the campaign raises awareness of the negative health consequences of not using CTs. However, it does not alter the behavior of residents or caretakers.

Our results offer a fresh perspective to the literature on public service delivery. First, we complement the existing evidence on the role of incentives in service delivery. Our finding that financial rewards for service providers incentivize performance aligns with evidence on the importance of extrinsic rewards in prosocially motivated jobs (Besley and Ghatak, 2018), in particular of financial rewards to improve the performance of front-line workers providing basic services (see, for example, Glewwe et al., 2010; Lazear, 2000; Duflo, 2012; Ashraf et al., 2014; Behrman et al., 2015; Caria et al., 2025), and complements evidence on bureaucrat incentives (Burgess et al., 2017; Rasul and Rogger, 2018; Bandiera et al., 2021; Akhtari et al., 2022; Besley et al., 2022; Fenizia, 2022; Best et al., 2023). However, we also find that incentives lead to user exclusion and negative externalities, underscoring how they can undermine universal access and highlighting the often-overlooked trade-offs that providers face when maintaining public infrastructure (Duflo et al., 2012).[3] This mechanism has not been explored in the literature studying the ability of the state to manage externalities (Bandiera et al., 2024).

Second, our study contributes to the literature on funding models. Prior research has extensively documented that user fees are regressive and exclusionary. Numerous studies show that higher fees, often in the context of free versus privatized services, reduce access to essential services such as education (Fafchamps and Minten, 2007; Lucas and Mbiti, 2012; Andrabi et al., 2020; Romero et al., 2020), health (Ito and Tanaka, 2018; Beuermann and Pecha, 2020; Rubli, 2023; Dupas and Jain, 2024; Bronsoler et al., 2025), garbage collection (Fullerton and Kinnaman, 1996), and clean water (Szabo, 2015). Theoretical work has emphasized these effects (Gertler et al., 1987; Norman, 2004; Hellwig, 2005; Gravel and Poitevin, 2019). Although most studies focus on fee pricing, less attention has been paid to the mechanisms underlying fee collection itself. Our study addresses this gap by treating user fees as exogenously determined and demon-

[3] The literature on public infrastructure focuses primarily on the effects of new constructions or expanding infrastructure, rather than maintenance. For LMICs and specific to the water and sanitation infrastructure, see, e.g., Alsan and Goldin (2019), and Bancalari (2024).

strates that incentives to improve service quality can inadvertently drive user exclusion, similar to the effects of increasing fees. By highlighting this overlooked channel, we provide key insights into the trade-offs of funding public services through user fees, emphasizing the need to balance service quality and equitable access, for instance, through free-to-use services, as discussed in Section 6. In addition, since fees often constitute the largest share of the overall tax burden in LMICs, our findings also complement the recent literature on broadening the tax base in settings with low state capacity (see, e.g., Pomeranz and Vila-Belda, 2019).

Finally, we contribute to the literature studying the causes for the under-provision of basic services in LMICs. As Burgess et al. (2020) posit, when services are funded by user fees in poor institutional settings, there is a high prevalence of non-payment because services are often perceived as rights, and, as a result, services are rationed. A growing body of literature explores various solutions to address non-payment of electricity and water bills, including pre-paid meters, information campaigns, commitment devices, and heavy-handed approaches such as threats to disconnect users (Jack and Smith, 2015, 2020; Coville et al., 2023; Rockenbach et al., 2023). We provide novel evidence on these mechanisms, highlighting the role of the providers' incentives, the consequences of user exclusion and the limited role of demand sensitization.

2 Theoretical framework

We analyze the service delivery problem from the perspective of a policymaker who contracts a provider to deliver a public service in exchange for the payment of a user fee p. Because our focus is on understanding the consequences of increasing quality while keeping the fees fixed, we assume that the fee is exogenously given and that fee payment might not be perfectly monitored. A certain degree of imperfect enforcement is present in fee-funded services, which is particularly limited in LMICs (Burgess et al., 2020). We discuss the implications of eliminating fees from our theoretical model in Section 6.

To deliver the service, the provider undertakes two types of effort: monitoring the payment of fees ($e_1 \in [0, 1]$) and maintaining the quality of the service ($e_2 \in [0, 1]$). We assume that if an effort of e_1 is supplied, then the probability of collecting a user fee from a given user is e_1 and the expected fee for the user is $\tilde{p} \equiv pe_1$. Similarly, if e_2 is the maintenance effort, then the quality that results is $q = e_2$.

The cost of exerting the efforts (e_1, e_2) is given by the following cost function:

$$c\left(e_1, e_2\right) = \frac{1}{2}e_1^2 + \frac{1}{2}e_2^2, \tag{1}$$

which assumes efforts to be independent. Although we do not explicitly consider the possibility that the two types of effort are complements or substitutes in the cost function, as discussed in the remainder of Section 2, there exists a natural complementarity between the two efforts via their effect on demand, which motivates our choice to keep the cost function simpler. Nevertheless, it is straightforward to extend our framework to allow for complementarity or substitutability.[4]

Users decide whether or not to use the public service and, conditional on using it, whether or not to pay the fee depends on the provider's monitoring effort. We follow the conventional assumption that the demand for the service is increasing in quality and decreasing in fees. We assume that the demand function is a reduced-form expression aggregating the individual choice by the users, and for tractability we impose a linear functional form:

$$D = \alpha q - \frac{1}{2}\beta\tilde{p} + \varphi. \tag{2}$$

Here, $\alpha > 0$ captures the effect of an increase in quality on demand (quality sensitivity), $\beta > 0$ captures the negative effect of fees on demand (price sensitivity), and $\varphi > 0$ is a constant that reflects the exogenous part of demand that is not responsive to changes in either the effort to monitor fee payment or the effort to perform improvements in quality, such as social norms. In principle, the demand function includes both extensive and intensive margins.[5]

Those who do not demand this service rely on an outside option that is subject to negative externalities, generating a social cost if the demand decreases. We capture this by considering a (net) social gain $s \geq 0$ obtained for each use of the service, while the total gain is given by sD.[6]

2.1 The first-best scenario

In the first-best scenario, we assume that there are no agency problems and that the provider will carry out the levels of e_1 and e_2 stipulated by the policymaker. The policymaker takes into account

[4] For instance, the cost function can be expanded with the additive term $\eta e_1 e_2$, where $\eta > 0$ captures substitutability (e.g., due to time constraints) and $\eta < 0$ captures complementarity (e.g., both tasks can be done simultaneously).

[5] We can provide micro-foundations to the demand by assuming that an individual's value of using the service is $\theta u(q)$, where $u(q)$ is increasing in q and is subject to diminishing marginal utility. The (expected) payment of the service leads to a disutility $\mu\tilde{p}$, where μ captures the marginal utility of money. With an outside option valued at v, an individual uses the service if:

$$\theta u(q) - \mu\tilde{p} \geq v.$$

With users differentiated in terms of θ, μ, and v, the demand function can be derived given the joint density $f(\theta, \mu, v)$. We obtain an intensive margin by allowing users to choose the number of times they use the service (x). If the per-use payoff from using the service does not depend on x, then the analysis is unchanged, since the payoffs from using the service and the outside option are $(\theta u(q) - \mu\tilde{p})\,x$ and vx, respectively. If the total payoff from using the service is instead $\theta u(q,x) - \mu\tilde{p}x$, with $u(q,x)$ increasing in x but subject to diminishing marginal utility, then the payoff from using the service would be decreasing in x. We could therefore derive the total demand by maximizing the payoff with respect to x, and use the optimal x together with the decision on whether to use the service.

[6] Let the private cost of the service per user be γ and σ the social gain, such that $s = \sigma - \gamma$ denotes the social gain net of cost from use of the facility. To rule out uninteresting cases, we assume $\sigma - \gamma \geq 0$.

the positive social gain from increasing the use of the service, and its expected payoff is given by

$$\hat{\pi}\left(e_1, e_2\right) = \left(s + pe_1\right)\left(\alpha e_2 - \frac{1}{2}\beta e_1 p + \varphi\right) - \frac{1}{2}e_1^2 - \frac{1}{2}e_2^2, \tag{3}$$

where the first component represents the total benefits derived from individuals making use of the service, while the second is the cost associated with running the service. The optimal effort levels can be solved by maximizing $\hat{\pi}\left(e_1, e_2\right)$ with respect to e_1 and e_2. The first-order conditions can be written as

$$\begin{aligned} e_1 &= \max\left\{\frac{p\left(\alpha e_2 - s\frac{1}{2}\beta + \varphi\right)}{1 + \beta p^2}, 0\right\} \\ e_2 &= \alpha\left(s + pe_1\right). \end{aligned} \tag{4}$$

There are three interesting observations on the relationship between the provider's efforts that come out of these expressions and that are worth mentioning before solving explicitly for e_1 and e_2. The first observation is that even though the two types of effort are additively separable in the cost function, and hence are neither substitutes nor complements, they are complements through their effect on the demand for the service. The higher e_2, the higher the quality of the service, and the higher the demand. With higher demand, the returns of the monitoring of the payment of user fees increase, leading to a higher value of e_1. In contrast, the higher the collection of user fees due to a higher level of e_1, the more resources are available for maintenance and the more effort it is worth putting into attracting more users, and so e_2 will be higher. The second observation is that the monitoring of fee payment e_1 is decreasing in social gain s. At the same time, the maintenance effort e_2 is increasing in s, and because e_2 increases demand, this incentive to increase quality partly mitigates the direct negative effect of a higher s on the monitoring of fee payments. The third observation is that any exogenous increase in demand through φ would increase both the monitoring effort e_1 and the maintenance effort e_2 as more users provide more fees.

We derive the optimal effort levels of the provider, e_1^* and e_2^*, by solving the first-order conditions:

$$\begin{aligned} e_1^* &= \max\left\{p\frac{\varphi - \left(\frac{1}{2}\beta - \alpha^2\right)s}{1 + p^2\left(\beta - \alpha^2\right)}, 0\right\} \\ e_2^* &= \alpha\,\frac{s(1 + \frac{1}{2}\beta p^2) + p^2\varphi}{1 + p^2\left(\beta - \alpha^2\right)}. \end{aligned} \tag{5}$$

The second-order conditions of the maximization problem imply $1 + p^2\left(\beta - \alpha^2\right) > 0$, and therefore the denominator is positive in both expressions. If we explicitly carry out comparative statics with respect to the parameters, then the effects we outlined intuitively are confirmed.

2.2 The second-best scenario

In the second-best scenario, there are agency problems between the policymaker and the provider, with e_1 and e_2 no longer contractable. Unlike the policymaker, we assume that the provider does not directly value the net social gain of usage (s), but benefits from keeping the service running and collecting fees. To capture the latter, for simplicity, we assume that the provider keeps a fraction λ of user fees. It is reasonable to assume that in any fee-funded service, the provider has a certain degree of gain from fee collection (e.g., salaries might be paid out of user fee revenues), meaning that the following analysis applies to a wide range of services.

Because the policymaker cannot directly contract provider effort but still wants to increase the quality of the service to reduce negative externalities of the outside option, we assume that the policymaker incentivizes the provider with a bonus b paid for higher values of e_2.[7] The resulting expected payoff of the provider $\pi\left(e_1,e_2\right)$ is given by

$$\pi\left(e_1,e_2\right)=\lambda pe_1\left(\alpha e_2-\frac{1}{2}\beta e_1 p+\varphi\right)+be_2-\frac{1}{2}e_1^2-\frac{1}{2}e_2^2. \tag{6}$$

For comparison, the policymaker payoff $\hat{\pi}\left(e_1,e_2\right)$ in the first-best scenario had $\lambda=1$ and $b=0$, and allowed for $s>0$. To determine the optimal values of e_1 and e_2 chosen by the provider, we maximize $\pi\left(e_1,e_2\right)$ with respect to e_1 and e_2, and derive the following first-order conditions:

$$\begin{aligned} e_1 &= \frac{\lambda p\left(\alpha e_2+\varphi\right)}{1+\lambda\beta p^2} \\ e_2 &= \alpha\lambda pe_1+b. \end{aligned} \tag{7}$$

Notice that e_1 and e_2 are still complements via a mechanism similar to the one in the first-best scenario. Explicitly solving the two first-order conditions for e_1 and e_2, we obtain the optimal effort levels of the provider e_1^* and e_2^*:

$$\begin{aligned} e_1^* &= \lambda p\frac{\varphi+\alpha b}{1+\lambda p^2\left(\beta-\lambda\alpha^2\right)} \\ e_2^* &= \alpha\lambda^2p^2\frac{\varphi+\alpha b}{1+\lambda p^2\left(\beta-\lambda\alpha^2\right)}+b. \end{aligned} \tag{8}$$

Here, the second-order conditions imply that $1+\lambda p^2\left(\beta-\lambda\alpha^2\right)>0$, ensuring that both efforts

[7] We could consider an alternative approach that rewards usage. The expected payoff for the provider would be $\left(b+\lambda pe_1\right)\left(\alpha e_2-\frac{1}{2}\beta e_1 p+\varphi\right)-c\left(e_1,e_2\right)$, where b is the bonus. Setting $b=s$ and $\lambda=1$ (meaning that the provider is the residual claimant) allows us to achieve the first-best scenario. However, measuring usage with accuracy and separately from fee collection and quantifying b based on social returns might be challenging. In addition, when λ is high, providers may have incentives to gain control of fees and increase them, thus excluding poorer households.

exerted by the provider are strictly positive.

The optimal effort levels of the provider e_1^* and e_2^* in the presence of agency problems have important implications for understanding the consequences of incentivizing service quality. Whenever the provider has an incentive to collect fees ($\lambda > 0$) incentivizing improvements in quality ($b > 0$) increases both e_1 and e_2 due to complementarity in efforts. Given the social benefits in the use of the service that are not taken into account by the provider, incentivizing quality may not be optimal. Quality incentives may increase the monitoring effort e_2, raising the quality of the service. However, these incentives can also push the monitoring effort e_1 higher than the policymaker would prefer in the first-best scenario, leading to the exclusion of users.

Comparing the first-best scenario to the second-best scenario, we can summarize the main implications of our model in three main lessons.

1. The two types of effort that characterize service delivery in fee-funded services are complements through their effect on the demand for the service.

2. The effect of increasing quality on the demand for the service is theoretically ambiguous and depends on the sensitivity of the demand to the quality and the price of the service.

3. In settings with significant social externalities, to the extent that social benefits are not taken into account by the service provider, the net social welfare effect of incentivizing quality is ambiguous and hinges on how monitoring efforts exclude users, who then turn to the outside option.

3 The experiment

We now turn to the empirical part of the paper. The theoretical framework developed in Section 2 will guide our interpretation of the main empirical results and inform further analysis to assess its plausibility.

We study a large-scale RCT that focuses on the provision of hygiene and sanitation services through shared public complexes known as CTs. Often arranged in gender-specific areas, they offer sanitation, hand-washing, and bathing facilities. Unlike public toilets, which are common throughout the world, CTs serve primarily residents in their vicinity. This service is widespread in the urban centers of LMICs and, in particular, in slums, where access to private sanitation is constrained. Slums represent an extreme case of both the lack of access to safe sanitation services and the high prevalence of OD. Of the estimated half a billion people practicing OD worldwide, about 10% live in urban areas, with India being the most affected (World Health Organization, 2021).

The context of the experiment is the informal settlements in rapidly growing cities in LMICs, where public services are pushed beyond capacity (Bryan et al., 2020). The experiment is implemented in the slums of Lucknow and Kanpur, the capital and the second largest city, respectively, of the Indian state of Uttar Pradesh. In 2015, Lucknow was the 129th largest city worldwide with 3.2 million inhabitants, and Kanpur was the 141st with 3.0 million inhabitants. These are fast-growing cities, with their populations expected to grow by 59% and 37%, respectively, between 2015 and 2035 (United Nations, 2018).[8] The rapid pace of urbanization has resulted in limited access to basic services in large areas of these cities. In Lucknow and Kanpur, the residents of the slums are 13% and 15% of the population, respectively, comparable to India's capital, Delhi (Government of India, 2011). Unsanitary conditions are common in these areas, where more than 40% of the residents do not have access to private toilets (Government of India, 2011).

In India, CTs remain a common solution for the foreseeable future and are argued to be the most appropriate alternative for slums.[9] Services are generally rendered by a long-term public–private partnership funded by user fees, with each use priced at a fixed standard fee of 5 Indian rupees (INR, corresponding to US$ 0.07).[10] While local municipal authorities are in charge of choosing the appropriate fee (Government of India, 2017), in practice, fees are fixed—that is, they do not vary frequently over time, they do not adjust to changes in demand or costs, and they are almost uniform across India. In our setting, this is confirmed by the absence of fee variation between the two cities. An average household of four members living in a slum could spend up to 8% of their average household income on fees if all adult members were to use and pay for every service use.

The provider of the service is the *caretaker*, who is in charge of the daily operation and management of the CT. These activities include monitoring the payment of user fees and maintaining service quality, such as cleaning the facility or supervising cleaners. In line with our assumption of exogenous fees, the caretaker cannot alter the user fee, which is taken as given. Because facilities lack physical-access control technologies, the only way to monitor fee payment is through the presence of the caretaker at the entrance.

In our setting, caretakers are hired centrally and receive a fixed salary to manage a single CT. The salary is on average equal to INR 5,000 (US$71) per month. Their contract does not include any performance-based financial rewards for quality improvement; however, to retain their job,

[8] These prospects are similar to growing cities such as Accra (Ghana), Amman (Jordan), and Hyderabad (Pakistan), and of metropolises such as Karachi (Pakistan), Cairo (Egypt), and Manila (The Philippines).

[9] Under the urban component of the Swachh Bharat Mission, toilets are envisioned for 80% of urban households engaging in OD. The remaining 20%, which represented more than 86 million inhabitants by 2015, is assumed to be covered by shared sanitation facilities due to space constraints (Government of India, 2017).

[10] Nominal INR are converted to nominal US$ using the 2019 average exchange rate of US$1 = INR 70.42 (International Monetary Fund, 2020).

they are aware that fee collection is essential. As their salary is funded through fee revenues, caretakers need to monitor fee payments to avoid being relocated or fired. Furthermore, the resources available to the caretakers for operating the facility are directly related to fee revenues, which are used to fund cleaning agents, tools, and cleaners, based on the amount collected.

In India, the quality of the service rendered by CTs is substandard (see, for example, anecdotal evidence from National Geographic, 2017). The facilities are poorly maintained, as shown by their low quality and the limited availability of functioning hand-washing facilities. Online Appendix A presents an analysis of CTs in our study area, which confirms the poor state of the facilities. We also highlight how low quality is associated with low payment of user fees in panel A of Figure 1. On average, only 65% of the users pay the fee (median 68%), and all users pay the fee only in 20% of the facilities. Payment is only partially enforced by caretakers, and only 8% of residents reported that they had been prevented from using the facility because they were unwilling to pay the fee. These statistics closely mirror the distribution of residents with positive WTP to use the service (panel B, Figure 1), captured using an incentivized measure described in Section 4.2. WTP on the intensive margin remains very low (panel C). On average, WTP amounts to INR 1.40—a mere 28% of the official fee—driven by a large share of respondents who are not willing to pay any amount. Free-to-use CTs are not as common, but they also exist in the study area, although their quality is much worse.

Access to sanitation in LMICs and the solution provided by CTs offer a unique setting to test the theoretical predictions discussed in Section 2. CTs operate just like a general service funded by user fees, with use depending on the fee and the extent to which payment is monitored, the quality of the facility, and an exogenous component to demand, which in our case is often tied to social norms. In addition, like many other services in LMICs, the monitoring of payment and quality maintenance are performed by an individual person and for a single service. The misalignment between the returns of the caretaker and those of a policymaker caring about social welfare is salient. The main alternative to using CTs is to practice OD, which means that if use drops, OD and the well-known negative health externalities to the community increase. Hence, the social gains from increasing users are crucial in this setting.

3.1 The interventions

The experiment was motivated by the goal of encouraging the use of CTs to reduce the social cost associated with OD. In the context of CTs, because the quality of the service also depends on the quality of the facility and of the infrastructure, which are beyond the influence of the caretaker, we can augment the model to allow for an exogenous component of quality. We let the quality of the service $q(e_2)$ be equal to $e_2 + a$, where $a > 0$ is the exogenous component. Hence, the demand for the service that would determine CT use would be equal to $\alpha(a + e_2) - \frac{1}{2}\beta e_1 p + \varphi$.

Because a policy of reducing fees (directly or indirectly through monitoring) would not be feasible given the need for user fees to fund the service, the equation suggests two alternative ways of increasing demand: first, by improving quality (by increasing a or by incentivizing the caretaker to increase effort e_2 through b); second, by increasing exogenous demand through φ, for example, by introducing an information campaign to increase awareness. However, as long as the caretaker benefits from a greater fee collection ($\lambda > 0$)—as in our setting, where caretaker's salary and maintenance inputs are funded by fee revenues—the theoretical framework in Section 2 suggests that the net welfare effect remains ambiguous between both interventions due to the complementarity between e_2, which increases the demand for CTs, and e_1, which excludes users. Even in the absence of incentives to increase quality ($b = 0$), direct increases in a or φ would lead to higher e_2, as it becomes worthwhile to attract users, which in turn increases e_1.

We implemented two interventions that capture these two policy options in partnership with the Lucknow and Kanpur Municipal Corporations, Sulabh International, and the zone and city managers of the CTs. The locations of the cities and CTs within the cities are shown in Online Appendix C. The activities were implemented by FINISH Society, a non-governmental organization based in Lucknow.[11] The following paragraphs describe each intervention, with operational details, including intervention costs, provided in Online Appendix E. Online Appendix C provides details of the intervention timeline.

Maintenance intervention

The first intervention focuses on boosting the quality of the service. We introduce two components. We started the intervention by targeting a through a one-time grant. The grant was paid in the initial two months of the intervention and was offered directly to the caretaker. The caretaker had the flexibility to allocate the grant into one of three packages of equal value: repairs and/or refurbishments (chosen by 41% of caretakers), deep cleaning of the facility and the sanitation system (chosen by 41%), or the provision of tools and agents along with training in maintenance best practices (chosen by 18%). The average value of each package was INR 25,000 (US$355), roughly the total cost of running the service for 2.5 months (Online Appendix E).

We then introduced an incentive scheme to motivate caretakers in their maintenance efforts, aiming to increase e_2 by setting $b > 0$ in the payoff of the provider (6). Based on previous research findings, we chose an output-based absolute payment system with discrete rewards, emphasizing individual performance to mitigate the impact of social comparisons (see the review in Besley and Ghatak, 2018). From the second to the twelfth month of the intervention, a bimonthly

[11] See more details about FINISH Society at www.finishsociety.org.

financial reward system was implemented for caretakers based on their compliance with various indicators, ensuring a clean and healthy facility, and, in line with Holmstrom (2017), taking into account the full portfolio of activities related to CT quality that the caretaker can engage in. These indicators were identified as the primary drivers of inadequate service delivery during baseline assessments. First, the caretakers received INR 500 (US$7.10) to maintain visible cleanliness of the latrines. Secondly, they received INR 500 (US$7.10) to ensure the availability of soap in the handwashing facilities. Lastly, caretakers who kept the bacteria counts below a specified standard were rewarded INR 1,000 (US$14.20). We allocated a higher incentive to reduce pathogen exposure due to its significant health consequences and its widespread presence (Online Appendix A). The specified standard is the baseline median value of *E. coli* bacteria count in the facilities that are part of the study. Importantly, incentives did not relate to the amount of user fees collected.

Caretakers received feedback on their past performance to gauge the effort required to meet the criteria, but the payments were tied only to current performance to deter gaming over time—based on lessons from Bandiera et al. (2015) and Bénabou and Tirole (2003). In each round, the total potential incentive was therefore INR 2,000 (US$28.40), roughly 40% of the caretaker's average monthly salary. In all rounds combined, this amount adds up to INR 8,000 (US$113.60), or 13% of the annual salary. These expected payments are large compared to other interventions that showed effects on exerted effort.[12] We set high expected payments because, in the context of prosocial tasks, financial incentives have been found to be effective when their relative value is high (Ashraf et al., 2014).

Every two months and for a total of four times during the study, the enumerators verified the conditions for assigning the reward during random visits and delivered the payments accordingly. The 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. These amounts correspond to 39%, 52%, 53%, and 49% of the potential reward, respectively.

Sensitization campaign

The second intervention focuses on increasing the part of the demand that is exogenous to service quality (φ) through a sensitization campaign among residents living in catchment areas of the study CTs. This intervention was implemented in a randomly selected half of CTs in the maintenance treatment group. The campaign aimed to increase awareness of the negative externalities resulting from OD, highlighting the importance of using the CT. The campaign was carried out

[12] For India and in the context of education, Duflo et al. (2012) and Muralidharan and Sundararaman (2011) offer a reward equivalent to 1% and 3% of a typical teacher's annual salary, respectively.

through different means. First, door-to-door visits were conducted three times in April–June 2018, July–September 2018, and January–March 2019. Secondly, leaflets were distributed among residents and posters were placed to summarize the main messages. Lastly, voice message reminders were sent monthly to the mobile phones of the households included in the study.

3.2 Research design: sampling and randomization

The research design is an RCT with the treatment unit being a CT. Because a list of CTs operating in the two cities was not available at the time of the experiment, in 2017 we conducted a census of all the facilities in Lucknow and Kanpur. We gather data on location, physical characteristics, management and maintenance practices, and types of user. These data formed the basis for selecting CTs that operate with user fees and are used primarily by residents (i.e., the most common model of service delivery in slums). We excluded facilities that were permanently closed or abandoned (i.e., unused or used by residents, but without a caretaker). A total of 110 facilities were identified that met these criteria, 52 in Lucknow and 58 in Kanpur. More details on sampling are provided in Online Appendix C.

To create exogenous variation in the quality of the service, each CT was randomly assigned to one of two groups: the treatment group received the maintenance intervention and the control group did not receive any intervention. For randomization, we first stratified CTs according to the main organization managing the facility and the city it was located in. Using the rich census information, we built blocks of three CTs using the Mahalanobis distance relative proximity, and randomly allocated CTs within a block to a treatment arm using a lottery with equal probability of assignment. Therefore, block identifiers represent the randomization strata. To minimize the risk of contamination by treatment, CTs within 400 m of each other were assigned to the same treatment arm. As a result, 40 CTs were assigned to the control group and 70 CTs were assigned to the maintenance treatment. In addition, while caretakers work only in one facility, we limited their rotation to different facilities in agreement with the service managers of each city. During the study period, we do not observe rotation of caretakers from one study CT to another. Whenever a caretaker was replaced, the implementing team made regular visits to inform the new caretakers about the intervention.

In addition, we cross-randomized the provision of the sensitization campaign to residents living near CTs in the maintenance treatment group. We discuss this in detail in Section 5.2.

4 Data

To obtain information on both service provision and residents, we gathered a substantial amount of primary data. Online Appendix B provides definitions of the variables used in the analysis, and provides the list of pre-registered outcomes. Online Appendix F provides the wording used for behavioral measurements.

4.1 Data from providers

We collected objective measures of service delivery during unannounced visits to study CTs. Independent observers collected information about the quality of facilities, including maintenance and cleanliness. Observers also collected samples from randomly selected spots on the floor of facilities. The samples were then analyzed in a laboratory to identify the presence and counts of bacteria. On average, more than three types of hazardous bacteria, including *E. coli* and salmonella, were found in each facility in each round. Finally, observers documented use and payment at the entrance of the facility by recording the number of users and the number of those who paid the fee. This count was performed for one hour during peak hours (between 5 a.m. and 7 a.m.), when most residents of the community use the facility. The observers relied on the knowledge of the caretakers to identify the usage and payment behavior of the slum residents and passersby who use the facility separately.

Measurements were collected at baseline in April–June 2018, and in four waves of bimonthly follow-up data collection, starting four months after the baseline: in October–November 2018 (follow-up 1), January–March 2019 (follow-up 2), April–May 2019 (follow-up 3), and July–September 2019 (follow-up 4).[13]

Objective measures of service delivery were complemented by primary survey data collected at the same time. The surveys were administered to the caretakers. The questionnaire, implemented consistently across the survey waves, covered maintenance efforts in the provision of the service, as well as efforts to monitor fee payment.

Table D1 in the Online Appendix presents descriptive statistics of the facilities and their caretakers at baseline. In 80% of the facilities, the CT is operated by a single caretaker; caretakers are generally male (82%), have approximately 10 years of experience in their job, and 44% live in the local community. Caretakers allocate 68% of their time to monitoring activities (i.e., collecting fees and supervising cleaners), while the remainder is allocated to activities that keep them away from the fee collection point, such as cleaning the facility themselves, conducting repairs, and

[13] To monitor the intervention's progress, we further implemented a mid-intervention measurement two months after the baseline, in July–September 2018, right after providing the one-off grant and before incentivizing the caretakers.

meeting managers.

Attrition was kept to a minimum between the baseline and follow-up surveys. The average number of observations per facility and per caretaker is equal to 3.9 and 3.8 out of 4, respectively, with no differential attrition between treatment groups (Online Appendix D.1).

4.2 Data from residents

We supplement data on service provision with information from residents of the slums that each CT serves. Because slum communities are a volatile population, we first had to build a standard sampling frame for them. Following the CT census described in Section 4.1, during the second half of 2017, we performed a geographical mapping of the slums surrounding each facility, followed by a census of all residents of the slums. In total, we collected information on more than 30,000 households in both cities, covering their demographics, dwelling characteristics (including geolocation), and their access to basic services.

The population of interest, which we call residents throughout the paper, comprises current users and potential users of each CT. These are households living in the slum that did not intend to relocate and where at least one member reported not using a private latrine for defecation. We also restricted this population to those who reside in the catchment area of a CT.[14] Applying these criteria, we established a sampling frame comprising 5,553 households, from which we randomly sampled 1,573 households, a study sample size aligned with our power calculations (Armand et al., 2018). The average characteristics of this sample closely mirror those of the broader population of Indian slum residents (Online Appendix C).

In conjunction with the CT baseline survey (Section 4.1), a baseline survey was administered to the residents sampled. The targeted respondent was the household's main decision-maker, in most cases the household head. The questionnaire covered the sociodemographic characteristics of the household, the sanitation behavior of the respondent, and the health status of the family members. Table D2 presents descriptive baseline statistics. On average, household heads are 45 years old, male, with primary education or less. The baseline health status of the residents of the slums is poor: almost 30% of the households had a sick member and 60% faced out-of-pocket expenditures on curative care.

The sample of residents was revisited twice during the follow-up period. These surveys were carried out in conjunction with follow-ups 2 and 4 of the CT survey, as shown in Online Ap-

[14] The catchment area is defined as the space inside the slum borders and within a radius of not more than 250 m from the facility. We fix this parameter after studying how service use is affected by the distance (computed using geolocation) between their residences and the closest facility. Proximity is crucial: beyond 250 m, service use decreases rapidly (Online Appendix A).

pendix C.[15] Measurements from the household survey are analyzed at the individual or household level.

Survey data were supplemented with behavioral measurements taken from the most senior male and female decision-makers in the household, who are commonly spouses. We collected these behavioral measurements with each participant alone, without other senior members present. Measurements from behavioral instruments are analyzed at the respondent level, using up to two observations per household.

The first measurement is the elicitation of residents' WTP to use the service. Following extensive piloting, we opted for the incentivized version of the multiple price list (or take-it-or-leave-it) methodology, which performs well in settings where prices are well known (Andersen et al., 2006; Berry et al., 2020). We asked participants to choose between different amounts of cash or a bundle of 10 tickets, allowing them to use the CT in their catchment area. We do not focus on ability to pay because a single use of the CT is relatively cheap and highly recurrent, while ability to pay is a more binding constraint for products with high value relative to household income (see, e.g., Kremer and Miguel, 2007; Ashraf et al., 2010; Dupas, 2014b). One of the options is then randomly drawn and the decisions are realized. We informed the participant that each option has the same probability of being drawn. Although the fee value for 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 fee value to handle truncation) in steps 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.[16]

The second measurement aims to accurately capture the prevalence of OD among residents. Because this behavior is a sensitive issue, we use a list randomization technique to elicit prevalence (see, e.g., Karlan and Zinman, 2012). This technique addresses potential stigma by reading a list of statements to the respondent and only asking how many of these are true, rather than which ones. We randomly assigned each respondent to either a list of 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 share practicing the sensitive behavior, in our case, OD. This measurement was collected at the end of the study, in follow-up 4.

We supplement these measurements with additional behavioral measurements to understand caretakers' and residents' responses. Specifically, we use adapted dictator games to measure care-

[15] Similar to the CT survey, a rapid assessment survey was carried out in conjunction with the CT mid-intervention measurement. Impacts on mid-intervention outcomes are presented in Online Appendix D.2.

[16] This measure is conditional on the quality of the closest facility. On average, residents are willing to pay more than the fee when asked about a hypothetical higher-quality CT (see Online Appendix F).

takers' prosocial motivation for the cause and citizens' willingness to contribute to the cleanliness of the CT. We discuss the impacts of treatment along these dimensions in Online Appendix D.8.

On average, each study household was re-interviewed 1.65 times out of a possible two interviews, with 7.9% of baseline households not participating in any of the follow-up surveys. To minimize sample loss during return visits, we interviewed additional households that were randomly chosen from the baseline sampling frame. We do not observe differential attrition between treatment groups, and being a replacement household is orthogonal to treatment allocation (Online Appendix D.1).

5 Results

We use data collected during follow-up rounds to study behavioral responses of both service providers and residents to the maintenance intervention. Using the random allocation to the intervention, we estimate treatment effects by restricting the sample to follow-up observations. We begin by estimating the impact of the maintenance treatment on the outcome $Y_{ij,t}$ of observation i (a CT, a household, or an individual) in catchment area j at survey wave t using the following specification:

$$Y_{ijt} = \beta\, T_j + \alpha\, \boldsymbol{X}_{ijt} + \epsilon_{ijt}. \tag{9}$$

Here, T_j is an indicator variable equal to 1 if the catchment area j received the maintenance intervention, and 0 otherwise. $\boldsymbol{X}_{ijt}$ is a set of indicator variables that capture randomization strata and survey wave fixed effects. The error term ϵ_{ijt} is assumed to be clustered by catchment area when the analysis is performed at the household or individual level.

Equation (9) estimates the impact of the maintenance treatment T throughout the study. The results using this specification are presented in Section 5.1. In Section 5.2, we present results that estimate separately the impact of providing the maintenance treatment with or without a sensitization campaign in the catchment area. Given the low serial correlation found in our outcome variables, we follow McKenzie (2012) and pool multiple follow-up measurements to average out noise and increase power. In Online Appendix D.2, we show the treatment effects estimated for each survey round separately.

Randomization was successful in creating observationally equivalent groups in the experiment for the characteristics of the household and the CT (see Tables D1 and D2), and for the outcome variables measured at baseline (Online Appendix D.2). We support main estimates with estimates using alternative specifications, including ANCOVA to increase precision, ordinary least squares (OLS) with inverse probability weights to correct for attrition (Online Appendix D.3), and machine learning approaches for the selection of control variables (Online Appendix D.4). All results are robust to these alternative specifications. Furthermore, in addition to the features introduced in the design of the experiment (see Section 3.2), Online Appendix D.5 shows evidence against spillover effects between treatment arms.

For inference, we supplement standard p-values with those adjusted for 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 Romano and Wolf (2005, 2016) bootstrap-based procedure. The latter considers all hypotheses tested within a table, separately for outcomes at the CT level and at the household/respondent level. The level of analysis is indicated at the bottom of each table.

Sections 5.1 and 5.2 present estimates of treatment effects focusing on the following groups of outcomes: quality of service delivery, use and payment for the service, outside option, and health consequences. In Online Appendices D.4 and D.7, we discuss heterogeneous effects along pre-specified dimensions for all outcome variables using the causal forest procedure of Athey et al. (2019) and using interaction terms in equation (9), respectively. The estimated effects tend to be homogeneous across different heterogeneity dimensions.

5.1 Increasing the quality of service delivery

We begin by verifying the correct implementation of the maintenance intervention. We focus on two measures of exposure to the intervention. First, the transfer to a CT includes the value of the initial grant received and the subsidized use of tickets from the WTP game, along with products donated by study participants as part of the adapted dictator game measuring citizens' willingness to contribute to the cleanliness of the CT. Second, the transfer to a caretaker comprises the financial rewards provided as part of the treatment and amounts retained by caretakers in each round of the adapted dictator game. See Section 4.2 for details about the games.

The successful implementation of the maintenance intervention is reflected in significant differences in exposure between the experimental arms (Online Appendix D.5). Treated CTs received INR 4,741 (US\$ 55.20) more than control CTs per round and caretakers in treated CTs received INR 757 (US\$ 8.81) more than caretakers in control CTs per round. During the study period, the total transfer to a treated facility was INR 25,270 (US\$ 358.84), 16 times higher than the average transfer to a control CT. Similarly, the 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 on service quality (column 1) and caretaker effort (columns 2–4). Service quality is measured as an index that combines objective measurements of service delivery, including the status of the facility as observed by interviewers, and the lack of harmful bacteria collected with laboratory tests. Online Appendix D.6 details the construction of the index.

The caretaker effort focuses on maintenance (e_2 in our model) and monitoring of fee payment (e_1). Maintenance efforts include cleaning and renovation. Cleaning is measured as an index that includes the number of tools, materials, and cleaning staff employed during the last routine cleaning of the facility, and the caretaker's correct implementation of this process, normalized to be between 0 and 1. Renovation is measured as an indicator variable equal to 1 if the facility underwent repairs and/or deep cleaning in the month prior to the visit, and 0 otherwise. The monitoring effort is measured by the reported allocation of time to fee collection and supervising cleaners, in contrast to activities that take them away from the fee-payment point. The analysis in this table is performed at the CT level.

We find that the intervention consistently improves the quality of service delivery. On average, the maintenance treatment leads to an increase of 6.6 percentage points in the quality index, 10.4% higher than the control mean. This effect remains robust to multiple hypothesis testing with a p-value of 0.01. The treatment shifts the distribution of the index of quality of service delivery, which is detectable mainly at higher levels of quality (panel A of Figure 2). A Kolmogorov–Smirnov test for the equality of index distributions in the control and maintenance treatment groups is rejected at the 1% confidence level. The underlying drivers are improvements in perceived cleanliness, while no significant effect is observed on the structural quality of the facility and the presence of harmful bacteria (Online Appendix D.6).

To gain insight into how quality of service delivery increases, we estimate the impact of the intervention on the caretaker's effort. We find that the maintenance effort increased. Cleaning performance improved significantly by 6.1 percentage points, 11.9% greater than the control mean. This effect remains robust to multiple hypothesis testing with a p-value smaller than 0.01. The effect is driven by improved inputs, the use of cleaners, and a correct implementation of cleaning procedures (Online Appendix D.6). Despite significant transfers to the CT at the start of the intervention, we observe no effect on renovation during the follow-up period.[17]

[17] We find significant improvements in renovation only in the mid-intervention survey, right after the implementation of the grant scheme (Online Appendix D.2). The effect is equal to an increase of 32.6 percentage points in the likelihood of having implemented a renovation compared to the control group. Improvements in quality are not achieved by an increase in labor supply, as caretakers continue to work on average 12 hours a day in all treatment arms. Changes in this dimension might be limited by the fact that the labor supply is closely aligned with the opening times of the

In addition to a greater effort to maintain quality, we also observe an increase in the effort allocated to monitoring activities that enable fee collection. Because fee collection was not incentivized by the intervention, this increase is consistent with our model's finding of complementarity between the efforts made to maintain quality and to collect fees (see Section 2). We find that caretakers spend a significantly larger share of their time on monitoring activities, which increased by 6.7 percentage points, 9.5% higher than the control mean. The p-value of this effect is 0.02, which is adjusted to 0.05 when accounting for multiple hypothesis testing. The maintenance treatment shifts the distribution of the monitoring effort, which is detectable throughout the whole distribution (panel B of Appendix Figure D6). A Kolmogorov–Smirnov test for the equality of the distributions of monitoring effort in the control and maintenance treatment groups is rejected at the 1% confidence level. These findings confirm that the presence of the caretaker at the facility's payment point serves as an effective way to monitor payment of user fees. On average, the treatment did not prompt caretakers to implement stricter payment enforcement, while a significant increase in enforcement is observed in facilities where payment rates were initially low (Online Appendix D.9). In addition, we find no effect on labor supply, measured as the total number of hours worked (panel A of Appendix Figure D6).

Our model shows that, when both the effort to maintain quality and the effort to monitor fee payments shift, the effect on demand for the service is theoretically ambiguous and ultimately a function of quality and price sensitivities; that is, α and β in equation (2). Recall that the price sensitivity in the demand depends on the expected fee and thus on monitoring of fee-payments. In the following subsection, we analyze how changes in quality manifest in terms of demand for the service.

Demand for the service

Table 2 turns to the demand for the service (D), focusing on use and payment by presenting alternative measures. Columns 1 and 2 focus on all users, while columns 3–6 focus on residents. Columns 1 and 2 show the impact on the total number of users and on the share of users paying the fee, respectively. Both indicators rely on data collected by independent observers during the peak hours in the CT (see Section 4.1 for details about these measures). Relying on the same measurement, column 3 documents the effect on the total number of users, but restricted to residents. Focusing instead on survey responses, and therefore limited to residents, column 4 shows impacts on the number of times a person used the service the day prior to the interview. The last two columns focus on payment among residents. Column 5, again relying on data collected by independent observers during the peak hours, presents the effect on the share of residents who

facilities.

used the service and paid the fee. Column 6 focuses on the resident's WTP for a single CT use (in rupees), an incentivized measure explained in Section 4.2. The specifications in columns 1–3 and 5 are at the CT level, column 4 is at the household level, and column 6 at the respondent level.

We begin by focusing on the total number of users during peak hours. Considering all types of users (residents and non-residents), we observe an insignificant decrease in the number of users. However, when focusing on residents, the effect is slightly larger and is precisely estimated. The maintenance treatment decreases the number of resident users during peak hours by 2.3 users (corresponding to a decrease of 8.2% as compared to the average in the control group). The p-value is equal to 0.07 and 0.10 when adjusted for multiple hypothesis testing.

This result is supported by self-reported use among residents in the intensive margin (column 4). The maintenance treatment reduces by 0.12 the number of daily uses, a 9.7% drop over the control group mean. This estimate is robust to multiple hypothesis testing with an adjusted p-value of 0.08. This effect is observed among residents who are regular users of the CT, as well as those who are not (Online Appendix D.10).

At the same time, we observe an increase in payment as a larger share of users are observed to be paying the fee before using the service. Focusing on all types of users (column 2), the maintenance treatment leads to a significant increase in the share of those paying the user fee by 10.0 percentage points, a 17.9% increase over the share in the control group. This estimate is robust to multiple hypothesis testing with an adjusted p-value of 0.01. The treatment shifts the entire payment distribution, with noticeable effects even at lower payment levels and more pronounced impacts at full payment (panel B of Figure 2). A Kolmogorov–Smirnov test of the equality of payment distributions in the control and maintenance treatment groups is rejected at the 1% confidence level. Focusing only on residents, we observe a slightly larger increase in the share of users paying the fee (column 5). The estimate is equal to 11.1 percentage points, or 22.7% over the share in the control group.

The reductions in users and increases in payment translate into a small positive effect on revenues during peak hours (Online Appendix D.9).[18] When considering all types of users, the maintenance treatment generates an increase in revenues by 11.8% over the average revenues in the control group. Among residents, the effect is instead equal to 13.7%.

The effects in use and payment are observed even though there is no effect of the maintenance treatment on WTP among residents (column 6). The average WTP for a single use is equal to INR 1.20 in both the control and the maintenance treatment groups, as compared with the fee of INR 5. Furthermore, we find no significant effect on the likelihood that residents have a positive

[18] As finance reports at the facility level are not available, we estimate revenues using observers' data on users and payments, but restricted to the times during which these data were collected by observers.

WTP or a WTP equal to or greater than the user fee (Online Appendix D.8). Due to the small amounts at stake, it is unlikely that the measurement instrument generated income effects. Having received free tickets to use the CT as compared to cash does not affect the use in the following survey round (Online Appendix D.10).

These results suggest that the increase in payment is mainly related to the increase in the probability of collecting a user fee, which originates from the greater monitoring efforts associated with stimulating the quality of service delivery (Section 5.1). Therefore, the benefits of improving quality do not outweigh the higher costs for users associated with greater monitoring efforts by the caretaker. The overall negative effect in usage described in this section suggests that the price sensitivity of demand was strong enough to offset the quality sensitivity.

Outside option and externalities

Because Section 5.1 highlighted a decrease in users after the maintenance treatment, we study how the intervention affects the share of residents who rely on the outside option. As discussed in Section 3, the main outside option for residents living near the CT is to practice OD. Table 3 presents the results. In column 1, we show impacts on having practiced OD the day before the interview, our measure of OD prevalence measured in the last follow-up survey. As discussed in Section 4.2, because OD is a sensitive behavior and there is a high awareness of the negative externalities it imposes on residents (66.0% of the control group and 69.1% of the treatment group are aware of this), we rely on a list randomization technique to build prevalence. The analysis in column 1 is thus conducted at the respondent level, but can be interpreted only in aggregate terms.

We find that the maintenance treatment increases the share of residents who claimed to have practiced OD the previous day by 22.8 percentage points, compared to a share of 21.0% in the control group. The p-value of this effect is smaller than 0.01, and equal to 0.01 when considering multiple hypothesis testing.[19] The effect of the maintenance treatment on the self-reported number of times a resident used sanitation practices other than the CT during the day before the interview is also positive, but not statistically significant (Online Appendix D.10). The magnitude of the effect on the prevalence of OD captured by list randomization is larger than the effect observed in the corresponding period (i.e., the last follow-up survey) in self-reported data on CT use. Due to the different nature of the two options, this difference is likely due to the attenuation bias caused by the measurement error in self-reported usage. The magnitude is also greater than

[19] Using the same technique, we find that 58.4% of respondents in the control group used the CT the previous day and 82.0% washed their hands with soap. We do not find any significant effect for these variables (Online Appendix D.8). The coefficients on both practicing OD and using the CT the previous day are positive, though the latter is not statistically significant. This could be explained by an increase in mixing methods in the intensive margin: due to improved quality, more residents use the CT, but with a lower frequency due to payment monitoring (see column 4 in Table 2).

the effect on the number of users captured during the observations at the facility, which could be explained by the fact that the observations refer to peak hours only, whereas the list randomization captures behavior throughout the day.

The observed increase in the use of the outside option serves as additional evidence in support of the price sensitivity of demand that offsets the quality sensitivity. Characteristics that proxy poverty, such as female-headed households, larger families, and households with fewer assets, correlate significantly with stopping use of the service in response to the maintenance treatment (Online Appendix D.11).

We address and rule out several alternative explanations linking a decrease in use and an increase in the outside option. First, user exclusion is not driven by overcrowding caused by a surge in demand. One plausible concern is that an initial improvement in quality could lead to increased usage, resulting in long queues that discourage future use. However, the observed use during the peak hours at the beginning of the intervention was not significantly higher in the treated facilities compared with the control facilities (see Figure D2). Second, we do not find any evidence that the users adjusted the timing of their visits to the CT. To verify this channel, in the follow-up surveys, in addition to observed use during peak hours at dawn, we asked observers to repeat the measurement in the afternoon, when the number of users is much lower. Data were collected using the same methodology. Online Appendix D.9 shows that the effect of the maintenance treatment is again negative, but not statistically significant, thus excluding changes in the timing of use. Finally, we dismiss the possibility of a mechanism driven by effects on caretaker absenteeism, changes to the facilities' opening hours, or closure of the facility. Using data from unannounced visits, we show no systematic differences between treatment arms in whether the facility was open or whether the caretaker was present (Figure D3).

To understand the extent to which greater reliance on the outside option generates negative externalities, in columns 2–5 of Table 3, we study the health status of residents. The dependent variable in column 2 is morbidity (self-reported), measured by an indicator variable equal to 1 if any member of the household had a fever, diarrhea, or a cough during the two weeks prior to the interview, and 0 otherwise. Columns 3–5 focus on out-of-pocket healthcare expenditures, focusing on total expenditures, as well as positive expenditures distinguishing between curative and preventive healthcare. Curative expenditures include costs associated with doctor visits for illnesses, medicines and diagnostics, and hospitalization. Preventive expenditures include all costs associated with the use of sanitation facilities, access to drinking water and hygiene, scheduled medical checks, and preventive goods such as vaccines, bed nets, and anti-worm tablets. The specifications in columns 2–5 are at the household level.

We document results consistent with an increase in infectious diseases due to increased OD.

Although, on average, we find no effect on morbidity or total healthcare expenditures, the maintenance treatment increases the probability of spending on curative healthcare by 4.7 percentage points (7.4% higher than the control mean and the p-value is adjusted from 0.03 to 0.12 when considering multiple hypothesis testing). The effect is larger and significantly different from zero during the first follow-up survey (Figure D3). In the same follow-up survey, we also find an increase in self-reported morbidity by 7.6 percentage points. An effect on the probability of spending on curative healthcare without any effect on total expenditures is in line with infectious diseases being treated with low-cost therapies.

5.2 Adding demand-side sensitization

We investigate the additional effects of a sensitization campaign aimed at raising awareness about the negative consequences of OD. Recall from Section 3.1 that this intervention was designed to stimulate the exogenous part of the demand (φ), and to increase users through the demand in combination with an improved service.

We dedicate a separate section to this effect because the sensitization is implemented not at the facility level, but among residents living near the facility, and it is incremental with respect to the main intervention. Within the maintenance treatment arm, we cross-randomized the allocation to this sensitization campaign. Of the 70 CTs that were assigned to the maintenance treatment, we randomly selected 35 CTs and implemented the campaign among residents living in their catchment area. To estimate the differential effects of additionally implementing the sensitization campaign, we estimate the following specification:

$$Y_{ijt} = \beta_1 T1_j + \beta_2 T2_j + \alpha \boldsymbol{X}_{ijt} + \epsilon_{ijt}. \tag{10}$$

Here, $T1_j$ is an indicator variable for whether CT j received the maintenance treatment but the residents in its catchment area were not targeted by the sensitization campaign (maintenance-only group), and $T2_j$ is an indicator variable for whether CT j received the maintenance treatment and in addition the residents in its catchment area were targeted by the sensitization campaign (maintenance plus sensitization group). Randomization created observationally equivalent groups across all treatment arms for household, catchment area, and CT characteristics measured at baseline (Tables D1 and D2, and Online Appendix D.2).

Exposure to water, sanitation, and hygiene (WASH) campaigns was already relatively high among study households, including awareness-creation efforts by the Swachh Bharat Mission. This exposure was reflected in high baseline awareness of the negative consequences of OD. Despite these ongoing campaigns, our sensitization campaign was effective in reaching the targeted population and further improving awareness (Online Appendix D.5). The maintenance plus sensitization treatment group saw significant improvements in various aspects, including the share of study households reporting exposure to a WASH campaign using interactive activities (8.4 percentage points higher than the average share in the control group, which was 0.65%), the recall of posters (16.2 percentage points higher than the control group), and the awareness of externalities of OD (5.6 percentage points higher than the mean of the control group of 66%). These effects are statistically different from those observed in the maintenance-only group.

For all outcomes presented in Tables 1–3, Table 4 reports estimates of the effect of the sensitization campaign estimated with equation (10) combining all follow-up rounds. 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 impacts of the two treatment arms do not differ. Estimates by survey round are presented in Online Appendix D.2.

We find no differential effects between the maintenance intervention with or without the additional sensitization campaign among residents. This suggests that the quality of service delivery and the means to achieve it are primarily driven by top-down incentives, and that raising awareness of externalities is not enough to boost demand.

6 Quality–pricing trade off

We study an intervention that aims to improve the quality of public services while leaving fees unchanged. However, we gain insight into pricing from the fact that, in our framework, the effort to collect fees influences the expected fee paid by users, which can range from 0 to the official fee p. In the first-best scenario, when the policymaker can contract the effort of the service provider, the first-order conditions (4) show that in the absence of net social gains from the service ($s = 0$), the optimal fee collection effort is strictly positive ($e_1 > 0$). However, large net social gains cause this effort to be equal to zero ($e_1 = 0$). Exerting no fee collection effort is equivalent to setting a fee at 0 ($p = 0$). It follows that access to the service should be free whenever the net social gain is sufficiently high.[20]

In the second-best scenario, where the provider receives a fraction λ of the user fees, setting the fees to 0 directly eliminates any effort to collect fees, thus achieving the same objective as in the first-best scenario. In this scenario, incentivizing the provider to improve the quality of service would not lead to fee-based user exclusion and would have the potential to increase social welfare, which is particularly important when potential net social gains are large. As soon as

[20] In the first-best scenario, the optimal fee collection effort e_1^* is decreasing in s when $\beta > 2\alpha^2$, that is, when the negative price sensitivity of demand (β) is large relative to the quality sensitivity (α). It follows that for $s > \frac{\varphi}{\frac{1}{2}\beta - \alpha^2}$, e_1^* should be set to 0.

$p > 0$, the potential for user exclusion makes the welfare effects ambiguous.

This result has important implications in a world where policymakers have incentives to raise the quality of public services while maximizing social welfare. Making the service free to use requires that operations and maintenance be fully subsidized through alternative means, such as taxation. For example, some communities in the United States use property taxes to fund garbage collection while allowing residents to claim deductions against their income tax (Fullerton and Kinnaman, 1996). In LMICs, this result is somewhat of a dilemma because the prevalence of fee-funded services is based on the significant challenges of expanding compliance with taxes that could fund local services (Weigel, 2020; Dzansi et al., 2022). In addition, existing evidence indicates mixed effects of providing free services on user behavior (see, e.g., Szabo, 2015). These limitations are highlighted in our context. Due to the lack of alternative funding schemes, free-to-use CTs provide a much lower quality and are often abandoned (Section 3). Although evidence from Brazil suggests that access to private sanitation improves property tax compliance (Kresch et al., 2023), financing CTs by increasing property taxes may not be feasible as they serve slum areas where property rights are more limited.

7 Conclusion

Our research advances our understanding of the complexities of public service delivery. We show that, in the presence of user fees, top-down incentives to increase the quality of service delivery can lead to sustained improvements and increased fee compliance. However, this improvement comes at the expense of excluding some users from the service and increasing negative externalities resulting from the outside option. We provide evidence on the mechanisms that lead to these unintended consequences, through the complementarity in provider's maintenance and fee-payment monitoring efforts.

These findings highlight the need for financing models that better align provider incentives with equitable access when shaping public service policies. It is crucial to enhance our knowledge of how to design effective mechanisms for the delivery and financing of public services, especially in the poorest settings. For example, while our results highlight the significance of top-down incentives, more evidence is needed to design effective bottom-up incentives. In our context, we show that increasing demand has no effect compared to top-down incentives. Understanding the constraints on collective action in areas characterized by prevalent coordination failures and resistant social norms is therefore an important research objective.

References

Abramovsky, L., Augsburg, B., Lührmann, M., Oteiza, F., and Rud, J. P. (2023). Community matters: Heterogeneous impacts of a sanitation intervention. *World Development*, (165):106197.

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

Akhtari, M., Moreira, D., and Trucco, L. (2022). Political turnover, bureaucratic turnover, and the quality of public services. *American Economic Review*, 112(2):442–493.

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.

Andrabi, T., Das, J., Khwaja, A. I., Ozyurt, S., and Singh, N. (2020). Upping the ante: The equilibrium effects of unconditional grants to private schools. *American Economic Review*, 110(10):3315–3349.

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.

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

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.

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

Augsburg, B., Bancalari, A., Durrani, Z., Vaidyanathan, M., and White, Z. (2022). When nature calls back: Sustaining behavioral change in rural Pakistan. *Journal of Development Economics*, 158:102933.

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

Bancalari, A. (2024). The unintended consequences of infrastructure development. *Review of Economics and Statistics*, forthcoming.

Bandiera, O., Best, M. C., Khan, A. Q., and Prat, A. (2021). The allocation of authority in organizations: A field experiment with bureaucrats. *Quarterly Journal of Economics*, 136(4):2195–2242.

Bandiera, O., Callen, M., Casey, K., La, E., Ferrara, C. L., and Teachout, M. (2024). State effectiveness. *International Growth Centre Evidence Paper*.

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.

Behrman, J. R., Parker, S. W., Todd, P. E., and Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in mexican high schools. *Journal of Political Economy*, 123(2):325–364.

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., Burgess, R., Khan, A., and Xu, G. (2022). Bureaucracy and development. *Annual Review of Economics*, 14(1):397–424.

Besley, T. and Ghatak, M. (2006). Public goods and economic development. In Banerjee, A., Benabou, R., and Mookherjee, D., editors, *Understanding Poverty*, pages 285–302. Oxford University Press, Oxford.

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

Besley, T. and Persson, T. (2009). The origins of state capacity: Property rights, taxation, and politics. *American economic review*, 99(4):1218–1244.

Best, M. C., Hjort, J., and Szakonyi, D. (2023). Individuals and organizations as sources of state effectiveness. *American Economic Review*, 113(8):2121–2167.

Beuermann, D. W. and Pecha, C. J. (2020). The effect of eliminating health user fees on adult health and labor supply in Jamaica. *Journal of Health Economics*, 73:102355.

Bleakley, H. (2007). Disease and development: Evidence from hookworm eradication in the American South. *Quarterly Journal of Economics*, 122(1):73–117.

Bronsoler, A., Gruber, J., and Seira, E. (2025). Private sector provision as an "escape valve": The Mexico diabetes experiment. *Review of Economic Studies*, 92(1):129–161.

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. *The Economic Journal*, 127(605):F117–F141.

Cameron, L., Gertler, P., Shah, M., Alzua, M. L., Martinez, S., and Patil, S. (2022). The dirty business of eliminating open defecation: The effect of village sanitation on child height from field experiments in four countries. *Journal of Development Economics*, 159:102990.

Cameron, L., Santos, P., Thomas, M., and Albert, J. (2021). Sanitation, financial incentives and health spillovers: A cluster randomised trial. *Journal of Health Economics*, 77:1024556.

Caria, S., Deserranno, E., Kastrau, P., and León-Ciliotta, G. (2025). The allocation of incentives in multi-layered organizations. *Journal of Political Economy*.

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., and Yoshida, S. (2023). Financing municipal water and sanitation services in Nairobi's informal settlements. *Review of Economics and Statistics*, forthcoming.

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. Technical Report 4, J-PAL Urban Services Review Paper.

Duflo, Esther, H. R. R. S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*.

Dupas, P. (2014a). Global health systems: Pricing and user fees. In Culyer, A. J., editor, *Encyclopedia of Health Economics*, volume 2, pages 136–141. Elsevier, San Diego.

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

Dupas, P. and Jain, R. (2024). Women left behind: gender disparities in utilization of government health insurance in India. *American Economic Review*, 114(10):3345–3383.

Dzansi, J., Jensen, A., Lagakos, D., and Telli, H. (2022). Technology and tax capacity: Evidence from local governments in Ghana. Working Paper 29923, National Bureau of Economic Research. Revised June 2024.

Fafchamps, M. and Minten, B. (2007). Public service provision, user fees and political turmoil. *Journal of African Economies*, 16(3):485–518.

Fenizia, A. (2022). Managers and productivity in the public sector. *Econometrica*, 90(3):1063–1084.

Finan, F., Olken, B. A., and Pande, R. (2017). The personnel economics of the developing state. *Handbook of economic field experiments*, 2:467–514.

Fullerton, D. and Kinnaman, T. C. (1996). Household responses to pricing garbage by the bag. *American Economic Review*, 86(4):971–984.

Gautam, S. (2023). Quantifying welfare effects in the presence of externalities: An ex-ante evaluation of sanitation interventions. *Journal of Development Economics*, 164:103083.

Gautam, S., Gechter, M., Guiteras, R. P., and Mobarak, A. M. (2025). To use financial incentives or not? Insights from experiments in encouraging sanitation investments in four countries. *World Development*, 187:106791.

Gertler, P., Locay, L., and Sanderson, W. (1987). Are user fees regressive?: The welfare implications of health care financing proposals in Peru. *Journal of Econometrics*, 36(1):67–88.

Geruso, M. and Spears, D. (2018). Neighborhood sanitation and infant mortality. *American*

Economic Journal: Applied Economics, 10(2):125–162.

Glewwe, P., Ilias, N., and Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, 2(3):205–227.

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.

Government of India (2017). Guidelines for Swachh Bharat Mission – Urban. Swachh Bharat Mission, Ministry of Housing and Urban Affairs, Government of India.

Gravel, N. and Poitevin, M. (2019). Optimal provision of a public good with costly exclusion. *Games and Economic Behavior*, 117:451–460.

Guiteras, R. P., Levinsohn, J., and Mobarak, A. M. (2015). Encouraging sanitation investment in the developing world: A cluster-randomized trial. *Science*, 348(6237):903–906.

Hathi, P., Haque, S., Pant, L., Coffey, D., and Spears, D. (2017). Place and Child Health: The Interaction of Population Density and Sanitation in Developing Countries. *Demography*, 54(1):337–360.

Hellwig, M. F. (2005). A utilitarian approach to the provision and pricing of excludable public goods. *Journal of Public Economics*, 89:1981–2003.

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

Holmstrom, B. and Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7(special issue):24–52.

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

Ito, T. and Tanaka, S. (2018). Abolishing user fees, fertility choice, and educational attainment. *Journal of Development Economics*, 130:33–44.

Jack, B. K. and Smith, G. (2015). Pay as you go: Prepaid metering and electricity expenditures in South Africa. *American Economic Review: Papers & Proceedings*, 105(5):237–241.

Jack, B. K. and Smith, G. (2020). Charging ahead: Prepaid metering, electricity use, and utility revenue. *American Economic Journal: Applied Economics*, 12(2):237–241.

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.

Kresch, E. P., Walker, M., Best, M. C., Gerard, F., and Naritomi, J. (2023). Sanitation and property tax compliance: Analyzing the social contract in Brazil. *Journal of Development Economics*, 160:102954.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5):1346–1361.

Lipscomb, M. and Mobarak, A. M. (2017). Decentralization and pollution spillovers: Evidence from the re-drawing of county borders in Brazil. *Review of Economic Studies*, 84(1):464–502.

Lipscomb, M. and Schechter, L. (2018). Subsidies versus mental accounting nudges: Harnessing mobile payment systems to improve sanitation. *Journal of Development Economics*, 135:235–254.

Lucas, A. M. and Mbiti, I. M. (2012). Access, sorting, and achievement: The short-run effects of free primary education in Kenya. *American Economic Journal: Applied Economics*, 4(4):226–253.

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

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.

Norman, P. (2004). Efficient mechanisms for public goods with use exclusions. *Review of Economic Studies*, 71(4):1163–1188.

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.

Rockenbach, B., Tonke, S., and Weiss, A. R. (2023). A large-scale field experiment to reduce non-payments for water: From diagnosis to treatment. *Review of Economics and Statistics*, forthcoming:1–45.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Romano, J. P. and Wolf, M. (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, 113:38–40.

Romero, M., Sandefur, J., and Sandholtz, W. A. (2020). Outsourcing education: Experimental evidence from Liberia. *American Economic Review*, 110(2):364–400.

Rubli, A. (2023). Trade-offs between access and quality in healthcare: Evidence from retail clinics in Mexico. *Journal of Public Economics*, 224:104938.

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.

United Nations (2018). World urbanization prospects: The 2018 revision. New York, US: United Nations Department of Economic and Social Affairs, Population Division.

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.

World Health Organization (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: Service delivery

Dep. variable:	**Quality**	**Efforts in service delivery**		
		Maintenance		**Monitoring**
		Cleaning	Renovation	
	(1)	(2)	(3)	(4)
Maintenance (T)	0.066	0.061	-0.012	0.067
	(0.022)	(0.014)	(0.048)	(0.028)
	[0.00, 0.01]	[0.00, 0.00]	[0.80, 0.80]	[0.02, 0.05]
Mean (control group)	0.636	0.513	0.625	0.707
Observations	434	434	434	434
Catchment areas	110	110	110	110
Observation rounds	4	4	4	4
Level of analysis	CT	CT	CT	CT
Measurement	Observed	Self-reported	Self-reported	Self-reported

Note. Estimates based on CT-level OLS regressions using equation (9). Standard errors clustered by catchment area are reported in parentheses. The 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) quality, an index computed by aggregating indicator variables about the structural quality of the facility, its cleanliness, and the lack of bacteria, and re-scaled to be between 0 (lowest in-sample quality) and 1 (highest in-sample quality); (2) cleaning, an index including the number of tools, materials, and cleaners used during the last cleaning of the facility and the caretaker's knowledge about this process, normalized to be between 0 and 1 (see Table D10 for individual components); (3) renovation, an 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; (4) monitoring, the share of worked hours allocated by the caretaker to collect fees and supervise cleaners, rather than conducting activities away from the entrance or off-site. All specifications include indicator variables for data collection rounds and randomization strata. Additional details about the variables are presented in Online Appendix B.

Table 2: Demand for the service

Dep. variable:	**All users (residents and passersby)**		**Residents**			
[0.5ex]	Users during the peak hours	Share paying the fee	Users during the peak hours	Number of daily uses	Share paying the fee	WTP (amount)
	(1)	(2)	(3)	(4)	(5)	(6)
Maintenance (T)	−1.781	0.100	−2.258	-0.117	0.111	0.008
	(1.518)	(0.032)	(1.250)	(0.059)	(0.034)	(0.080)
	[0.24, 0.23]	[0.00, 0.01]	[0.07, 0.10]	[0.05, 0.08]	[0.00, 0.09]	[0.92, 0.90]
Mean (control group)	33.903	0.556	27.519	1.198	0.489	1.205
Observations	434	434	434	3303	434	6001
Catchment areas	110	110	110	109	110	109
Observation rounds	4	4	4	2	4	2
Level of analysis	CT	CT	CT	Household	CT	Respondent
Measurement	Observed	Observed	Observed	Self-reported	Observed	Incentivized

Note. 'All users' refers to anyone using the service, while 'residents' focuses only on people who live in the slum and do not have access to private sanitation (see Section 4.2). Estimates based on CT-level OLS regressions using equation (9). Standard errors clustered by catchment area are reported in parentheses. The 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) total number of users observed during the peak hours; (2) observed share of users who pay the entry fee during the peak hours; (3) total number of users living in the slum observed during the peak hours; (4) number of times the respondent used the CT for defecation in the day previous to the interview; (5) observed share of users living in the slum who pay the entry fee during the peak hours; (6) incentivized willingness to pay for a single CT use (in rupees), elicited for a bundle of 10 tickets and divided by 10 to get a single-use WTP. All specifications include indicator variables for data collection rounds and randomization strata. Additional details about the variables are presented in Online Appendix B.

Table 3: Outside option and health consequences among residents

Dep. variable:	Outside option	Health consequences			
	Practiced OD	Morbidity	Health expenditure	Preventive healthcare	Curative healthcare
	(1)	(2)	(3)	(4)	(5)
Maintenance (T)	0.228	0.035	0.040	−0.003	0.047
	(0.071)	(0.023)	(0.070)	(0.003)	(0.021)
	[0.00, 0.01]	[0.14, 0.35]	[0.57, 0.48]	[0.31, 0.48]	[0.03, 0.12]
Mean (control group)	0.210	0.451	6.937	0.992	0.636
Observations	817	3323	3306	3322	3298
Catchment areas	107	109	109	109	109
Observation rounds	1	2	2	2	2
Level of analysis	Respondent	Household	Household	Household	Household
Measurement	List randomization	Self-reported	Self-reported	Self-reported	Self-reported

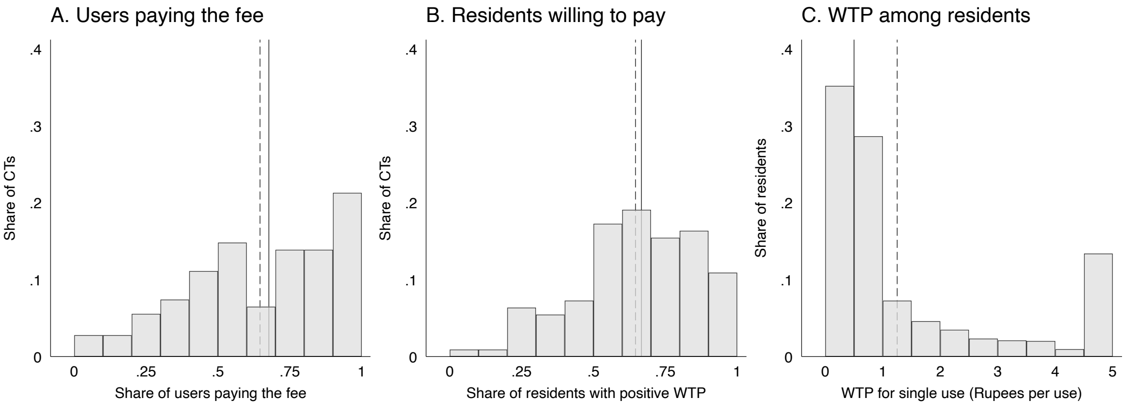
Note. Estimates refer to residents, that is, people living in the slum and not having access to private sanitation (see Section 4.2). Estimates based on household-level OLS regressions using equation (9). Standard errors clustered by catchment area are reported in parentheses. The p-values are presented 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) share of study participants who practiced OD the day before the interview, obtained using the list randomization technique applied to the most senior male and female household member in follow-up 4; (2) an indicator variable equal to 1 if any household member had a fever, diarrhea, or a cough during the two weeks previous to the interview, and 0 otherwise; (3) total health expenditure during the month previous to the interview, reported in logarithms; (4) an indicator variable equal to 1 if the respondent spent on preventive healthcare, and 0 otherwise; and, (5) indicator variable equal to 1 if the respondent spent on curative healthcare, and 0 otherwise. Column 1 includes only 107 catchment areas because, due to the randomization of lists to respondents, a number of areas do not have respondents with the list of items including OD. Columns 2–5 include only 109 catchment areas in the sample because the dependent variables were measured only in rounds 3 and 5, after one slum was displaced. All specifications include indicator variables for data collection rounds and randomization strata. Additional details about the variables are presented in Online Appendix B.

Table 4: The effect of demand sensitization

	Maintenance-only			**Maintenance plus sensitization**			$T1 = T2$
	β	se	p-value	β	se	p-value	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quality	0.08	0.03	0.00	0.05	0.03	0.08	0.44
Maintenance: cleaning	0.06	0.02	0.00	0.06	0.02	0.00	0.91
Maintenance: rehabilitation	−0.03	0.05	0.50	0.01	0.06	0.82	0.34
Monitoring	0.05	0.03	0.08	0.08	0.03	0.01	0.33
Users during peak hour	−2.49	1.75	0.16	−0.99	1.66	0.55	0.34
Share of users paying	0.09	0.04	0.02	0.11	0.04	0.01	0.52
Users during peak hour (residents)	−2.60	1.38	0.06	-1.88	1.45	0.20	0.58
Share of users paying (residents)	0.09	0.04	0.02	0.13	0.04	0.00	0.37
Number of daily uses	−0.10	0.07	0.17	−0.13	0.07	0.05	0.63
Resident's WTP (amount)	0.09	0.10	0.32	−0.08	0.10	0.38	0.10
Practiced OD	0.21	0.08	0.01	0.24	0.08	0.00	0.71
Morbidity	0.03	0.03	0.26	0.04	0.03	0.14	0.74
Health expenditure	0.02	0.09	0.80	0.06	0.07	0.41	0.64
Preventive healthcare	−0.00	0.00	0.43	−0.00	0.00	0.34	0.90
Curative healthcare	0.04	0.03	0.16	0.06	0.02	0.01	0.37

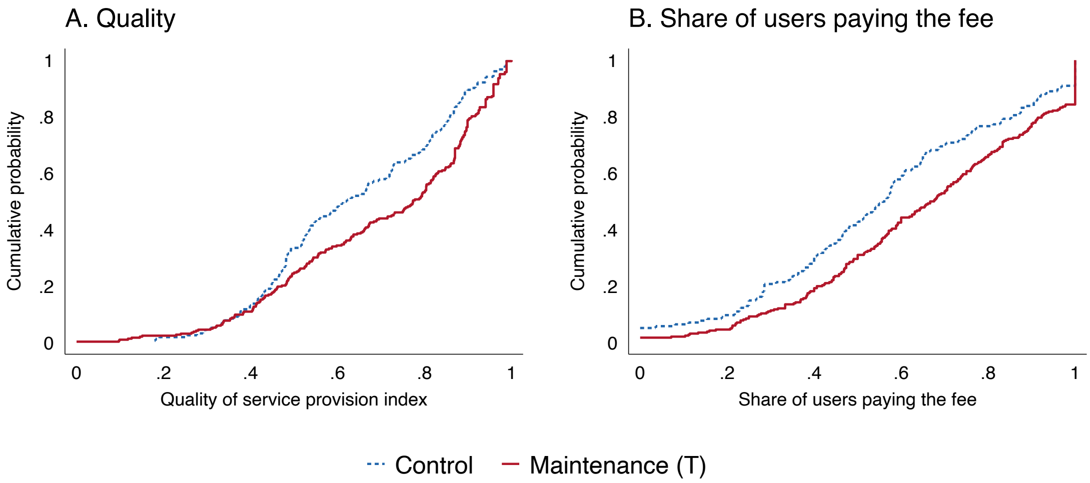
Note. In columns 1–6, estimates are based on CT-, respondent-, or household-level OLS regressions using equation (10) for each outcome. The p-values are presented in columns 3 and 6, the first from individual testing, the second adjusting for jointly testing that each treatment is different from zero 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 Online Appendix B. All specifications include indicator variables for data collection rounds and randomization strata.

Figure 1: Payment and WTP for the service



Note. Data collected at baseline. Panel A reports the (observed) share of users who pays the fee for the use of the CT during the peak hours. Panel B shows the share of residents in the catchment area of a CT who are willing to pay a positive amount for using the CT, estimated using the incentivized elicitation of WTP. Panel C 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 fee 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 Online Appendix B.

Figure 2: Quality of and payment for service delivery at follow-up, by treatment group



Note. The figure shows the empirical cumulative distribution functions of the quality of service delivery index (Panel A) and of payment (Panel B) distinguishing between control and treatment groups. The sample includes all follow-up measurements. The p-value of a Kolmogorov–Smirnov test of equality of distributions is equal to 0.003 for Panel A, and 0.002 for Panel B. Additional details about the variables are presented in Online Appendix B.

ONLINE APPENDIX

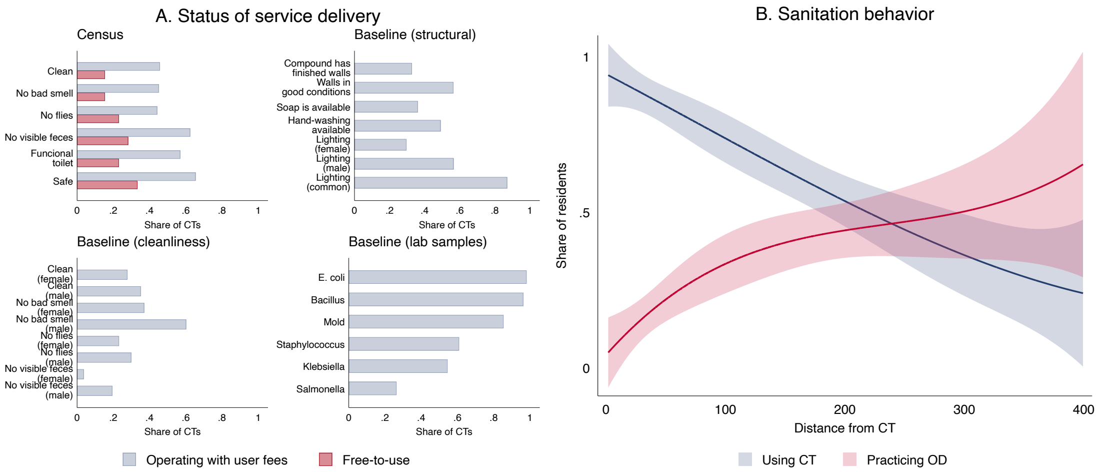
Public Service Delivery, Exclusion and Externalities

Alex Armand, Britta Augsburg, Antonella Bancalari, and Maitreesh Ghatak

A Status quo of service delivery in study area

Figure A1 summarizes service delivery in the study area. Panel A presents statistics on observed delivery dimensions, comparing free and pay-to-use facilities. In panel B, using self-reported data from the census of residents, we present cubic fits for the relationships between the distance from the facility and the use of the service or OD.

Figure A1: Status of service delivery and sanitation behavior in study area



Note. Panel A displays data from the CT census, and from the baseline survey for study facilities. For details see Section 4.1 and Appendix F. Panel B uses data from the slum resident census (Section 4.2). Shaded areas indicate 90% confidence intervals, based on standard errors clustered at the slum level. This analysis includes households eligible for the study (Section 4).

B Definition of variables

B.1 Definition of outcome variables

Variable	Description
Awareness	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 all rounds of the residents' survey.
Caretaker ever refused entry	Indicator variable for whether the respondent reports having observed the caretaker refusing entry in the CT to someone. The variable is self-reported by the household head.
Health expenditure	Expenditures incurred during the month previous to the interview: *curative*, out-of-pocket expenditures for 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; *preventive*, out-of-pocket expenses associated with hygiene, access to clean drinking water, regular doctor checks, vaccines, anti-worm tablets, bednets, prenatal tests, and travel costs associated with these expenses. The amount is reported in rupees and transformed using a logarithmic transformation. *Preventive healthcare* is an indicator variable equal to 1 if the respondent spent money on preventive healthcare, and 0 otherwise. *Curative healthcare* is an indicator variable equal to 1 if the respondent spent money on curative healthcare, and 0 otherwise. These variables are self-reported by the household head during all residents' survey rounds, except for the mid-intervention survey.

(continued on next page)

Variable	Description
Interactive activities	Indicator variable equal to 1 if the respondent is aware of any activity about WASH, and 0 otherwise. The variable is self-reported by the household head during the survey of residents.
Maintenance: cleaning	Index including the number of tools (broom, mop, safety equipment, liquid tools such as water, pressurized water and disinfectants), equipment and cleaners used during the last cleaning of the facility and the caretaker's knowledge about this process, normalized to be between 0 and 1, with 1 indicating the best cleaning input. The variable aggregate survey responses from the CT survey. The baseline survey asks for information only on use of the broom, and disinfectants, while the full list is available for the following rounds.
Maintenance: renovation	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. The variable aggregates responses from the CT survey and project's administrative data collected during all rounds.
Monitoring	Share of worked hours allocated by the caretaker to collecting fees and supervising cleaners.
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 each residents' survey round.
Monthly revenues	Revenues in rupees imputed using information from observers about the number of people using the CT and the share of them who is paying the fee (assuming a standard fee of 5 rupees). Information is collected using observation during the rush hour.
Number of daily uses	Number of times the respondent used the CT for defecation out of the two times previous to the intervention. *Regular users* are respondents that reported using the CT regularly. Data collected in every residents' survey round.
Posters	Indicator variable equal to 1 if the respondent is aware of any WASH poster in the CT, and 0 otherwise. The variable is self-reported by the household head during the survey of residents.
Practiced OD	Aggregate share of study participants who practiced OD the day before the interview. Data are obtained for the most senior male and female household member in follow-up 4 using the list randomization technique (Appendix F). At individual level, the variable is equal to the number of items reported by the respondents assigned to the group including the practice of OD minus the average number of items reported by respondents in the group without sensitive items.
Quality	Index computed aggregating indicator variables about the status of the facility, its cleanliness and the lack of bacteria, and re-scaled to be between zero (lowest in-sample quality) and one (highest in-sample quality). The variable aggregate survey responses from the CT survey, data from observers, and data from laboratory tests collected in all rounds.
Refused entry	Indicator variable for whether the respondent reports being refused entry in the CT for not having paid the fee. The variable is self-reported by the household head in the survey of residents.
Share of residents willing to pay a positive amount	Share of residents with a positive WTP in the incentivized WTP game for a single CT use, elicited for a bundle of ten tickets and divided by 10 to get at single use WTP. Data collected in every residents' survey round for the most senior male and female household member. The data is aggregated for each CT catchment area.
Share of users paying	Share of users entering the CT during the peak hour and paying the entry fee. The variable uses data from observers collected at the entrance of the CT in every round. The observers were instructed to identify users who reside in the slum, as opposed to other users, which allows computing this statistic for residents only. Information about residents is available only for the follow-up measurements.
Transfers	*Transfer provided to the CT* in the corresponding period as part of the intervention (in thousands of rupees). It includes, for the maintenance treatment group, the value of the initial grant to treated CTs, and, for both treatment and control group, the amount of subsidized tickets from the WTP game and the value of products bought as part of the adapted dictator game among residents. *Transfer provided to the caretaker* in the corresponding period as part of the intervention (in thousands of rupees). It includes the financial incentive for treated CTs and the amounts kept from the adapted dictator game for all CTs. Information based on project's administrative data.

(continued on next page)

Variable	Description
Users during the peak hour	Total number of users entering the CT during the peak hour. The variable uses data from observers collected at the entrance of the CT in every round. The observers were instructed to identify users who reside in the slum, as opposed to other users, which allows computing this statistic for residents only. Information about residents is available only for the follow-up measurements. In the follow-up period, this measurement was supplemented by the total number of users during an afternoon hour, when the number of users is lower.
WTP (amount)	Willingness to pay for a single CT use (in INR). The variable is incentivized and elicited for a bundle of ten tickets, and divided by 10 to get at single use WTP. Data collected in every residents' survey round for the most senior male and female household member.

Note. Appendix F provides detailed descriptions and scripts of measurement. Basemaps are from Esri®, and used in line with the Esri Master License Agreement, specifically for the inclusion of screen captures in academic publications.

B.2 Pre-registered outcomes and reference to the text

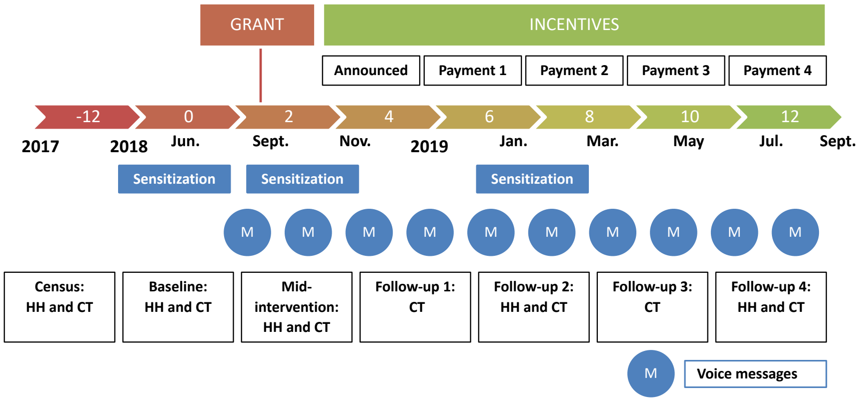
Primary outcomes	Description from pre-analysis plan	Table
Quality	Quality will be proxied using observations and lab results from samples taken at the CT to capture dirtiness, bacteria count, bad infrastructure quality.	1
Sanitation behavior (residents)	Sanitation practices of respondents and family members. In particular, we will focus on survey reports of CT use and open defecation (see *Note*).	2, 3, D15, D13
Sanitation behavior (CT level)	We will measure CT usage through tallies at the CT at specific times of the day: number of users and % users that pay	2
Willingness to pay	Elicited for the two primary decision-makers per household. Incentivized WTP for bundle of 10 tickets to use the nearest CT (using multiple price list).	2
Demand for cleanliness	Eliciting willingness to contribute to cleanliness of the CT through a donation game. Amount donated out of 50Rs (continuous variable 1-50).	D13
Secondary outcomes	**Description from pre-analysis plan**	**Table**
CT management	Management of CTs as reported in CT surveys by caretakers: % time allocated to clean and/or supervise cleaner, collect fee; CT cleaned more than twice per day and adequate cleaning.	1
Health status	Health situation of household members reported in household surveys.	3
Sanitation attitudes, expectations and knowledge	Priors about sanitation practices and the connection with illnesses and safety reported in household surveys.	D7
Hygiene	Hygienic practices of respondents and family members (see *Note*).	D13

Note. Due to concerns that arose during the baseline survey, when we observed high awareness of hygiene and of OD externalities, leading to potential stigma in self-reported sanitation behavior, we also collected sanitation behavior data using a list randomization technique (Section 4.2).

C Study location, timeline, and sampling

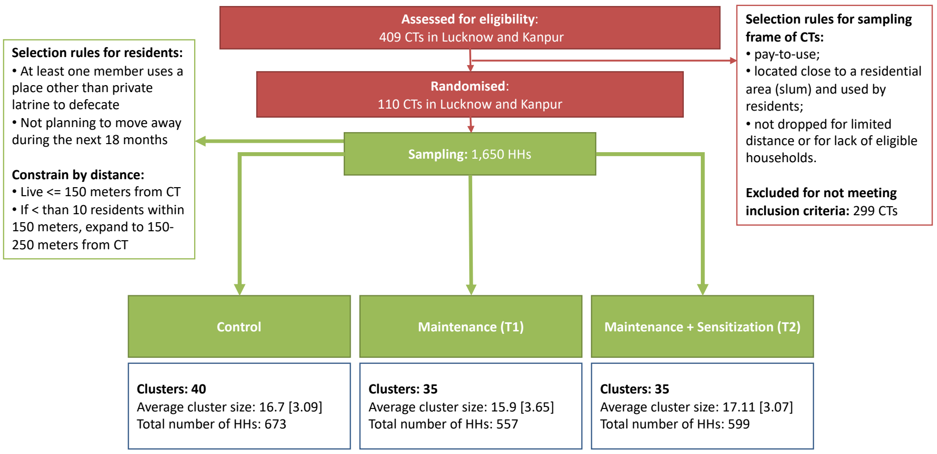
Figures C1 and C2 describes the timeline of interventions and data collection efforts and a summary of the sampling procedure. Figure C3 provides the spatial distribution of CTs in the study. To obtain the sampling frame, during the first half of 2017, we performed a mapping of slums and a census of all CTs in both study cities. *Slums* are defined in accordance to the census of 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. We identified 201 facilities in Lucknow and 208 CTs in Kanpur. Out of these, we dropped free-to-use facilities and/or facilities located outside slum areas (generally near market areas and used primarily by non-residents). To avoid cases in which residents can choose between different CTs, we drop those closer than 300 meters to each other, and those with two other CTs within 350 meters. In addition, we also dropped CTs in whose catchment areas are living fewer than 8 eligible households. This resulted in a total of 110 CTs.

Figure C1: Timeline of interventions and data collection efforts



Note. M indicates the delivery of voice messages. HH and CT indicates the collection of the household and CT surveys, respectively. Details about data collection activities are reported in Section 4. Further details of the procedure are discussed in Section 4.

Figure C2: Sampling strategy

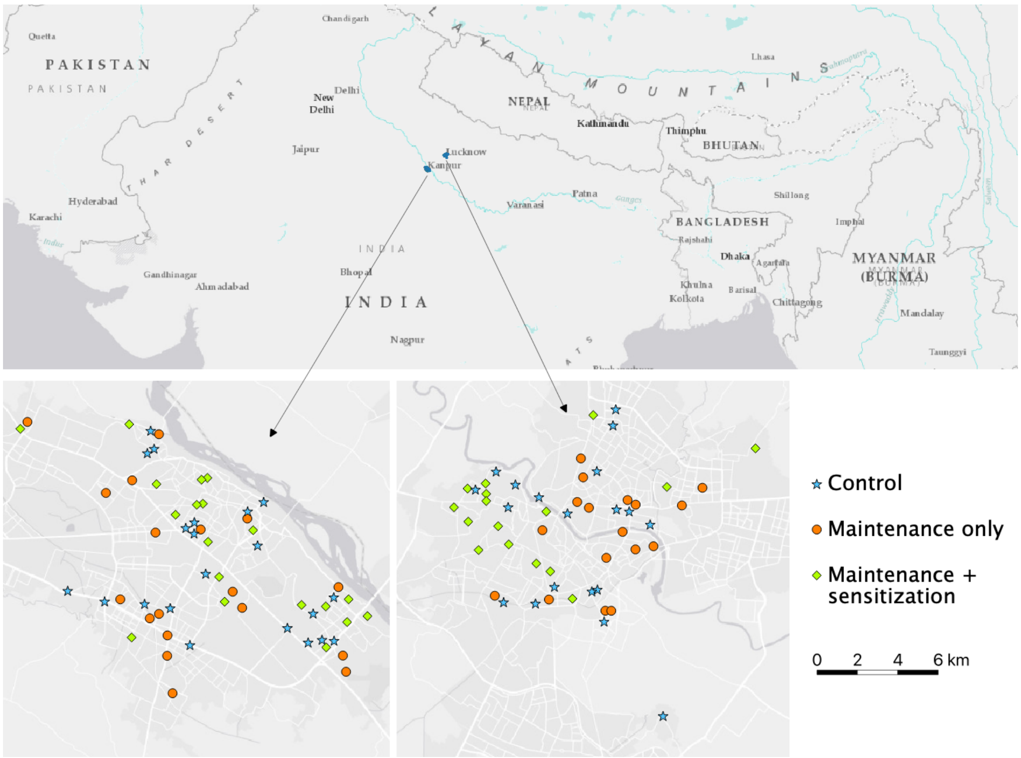

Assessed for eligibility:
409 CTs in Lucknow and Kanpur
Randomised:
110 CTs in Lucknow and Kanpur
Sampling: 1,650 HHs
Selection rules for residents:
• At least one member uses a place other than private latrine to defecate
• Not planning to move away during the next 18 months
Constrain by distance:
• Live <= 150 meters from CT
• If < than 10 residents within 150 meters, expand to 150-250 meters from CT
Selection rules for sampling frame of CTs:
• pay-to-use;
• located close to a residential area (slum) and used by residents;
• not dropped for limited distance or for lack of eligible households.
Excluded for not meeting inclusion criteria: 299 CTs
Control
Maintenance (T1)
Maintenance + Sensitization (T2)
Clusters: 40
Average cluster size: 16.7 [3.09]
Total number of HHs: 673
Clusters: 35
Average cluster size: 15.9 [3.65]
Total number of HHs: 557
Clusters: 35
Average cluster size: 17.11 [3.07]
Total number of HHs: 599

Note. The figure summarizes the procedure followed for the selection of CTs and the sampling of households within their catchment areas. Further details of the procedure are discussed in Section 4.

To identify residents we conducted a census of all households living within the slum boundaries and within 400 meters of the selected CTs in the second half of 2017. The questionnaire was administered to household heads, and information was collected from more than 30,000 households. From the enumerated households, we select residents according to the conditions described in Section 4.2. We sampled up to 17 eligible households per catchment area. For areas with fewer than 10 eligible households within 150 meters from a CT, we selected all households within this limit and randomly selected the remainder from households living within 150–250 meters from a CT.

In total, we obtained a sample of 1,650 households. This sample has average characteristics consistent with slum dwellers in all states of India and in UP, as reported in the 2011 Indian Slum Population Census. The proportion of males is 52% and 53% in the slums of India and UP, respectively, and 53% in our study sample. Similarly, the proportion of children is 12%, 14% and 9% respectively. In terms of caste composition, 45% of the study sample belongs to a scheduled caste, compared to 20% and 22% in India's and UP's slums, respectively. Literacy rates tend to be lower in the study sample than in the rest of the country (46% versus 78% and 69%, respectively).

Figure C3: Geographical distribution of CTs



Note. The figure shows the location of the Lucknow and Kanpur and the distribution of the CTs selected for the study. Details about the procedure to select CTs is provided in Appendix C. Basemap source: Esri (see Appendix B for attributions).

D Additional analysis

D.1 Balance in observable characteristics and attrition

Tables D1 and D2 show the balance test for characteristics at baseline. The null hypothesis of equal means across treatment arms cannot be rejected for any of the characteristics except for education of the household head, although the difference in means of all characteristics is not jointly significant.

Table D1: CT characteristics at baseline, by treatment group

	Descriptive statistics		**Differences from control group, by treatment group**			
	All	Control	Maintenance	Maintenance only	Maintenance + sensitization	p-value joint test (4)-(5)
	(1)	(2)	(3)	(4)	(5)	(6)
Year of construction	1996.98	1995.26	3.08	2.89	3.29	0.28
	[8.85]	[9.29]	(1.90)	(2.09)	(2.35)	
Distance to closest CT	0.54	0.58	-0.01	-0.02	-0.01	0.99
	[0.44]	[0.66]	(0.08)	(0.10)	(0.07)	
Surrounding market	0.33	0.35	-0.02	-0.01	-0.03	0.96
	[0.47]	[0.48]	(0.10)	(0.12)	(0.12)	
Surrounding road	0.84	0.88	-0.05	-0.03	-0.08	0.71
	[0.37]	[0.33]	(0.08)	(0.09)	(0.09)	
Surrounding government office	0.25	0.20	0.04	0.07	0.00	0.76
	[0.43]	[0.41]	(0.09)	(0.11)	(0.11)	
Only residents use CT	0.12	0.07	0.07	0.06	0.07	0.57
	[0.32]	[0.27]	(0.06)	(0.08)	(0.08)	
Single caretaker	0.80	0.82	-0.04	0.01	-0.10	0.42
	[0.40]	[0.39]	(0.08)	(0.09)	(0.09)	
Share of female caretakers	0.18	0.22	-0.06	-0.03	-0.08	0.59
	[0.37]	[0.39]	(0.07)	(0.08)	(0.08)	
Caretaker is also cleaner	0.27	0.28	-0.02	-0.01	-0.02	0.98
	[0.45]	[0.46]	(0.09)	(0.11)	(0.11)	
Caretaker is from local community	0.44	0.49	-0.09	-0.14	-0.04	0.41
	[0.50]	[0.51]	(0.10)	(0.11)	(0.12)	
Caretaker's experience (months)	125.28	129.91	-11.49	-3.54	-19.25	0.73
	[103.45]	[109.34]	(22.82)	(24.23)	(26.91)	
CT is cleaned frequently	0.86	0.87	-0.04	-0.04	-0.04	0.86
	[0.35]	[0.34]	(0.07)	(0.08)	(0.09)	
Monitoring	0.68	0.66	0.01	0.02	0.00	0.79
	[0.14]	[0.11]	(0.03)	(0.04)	(0.03)	
Capacity	13.00	13.21	-0.47	-0.51	-0.43	0.91
	[5.57]	[5.52]	(1.12)	(1.29)	(1.32)	
Daily opening hours	17.76	17.88	-0.24	-0.37	-0.11	0.52
	[1.49]	[1.59]	(0.27)	(0.33)	(0.30)	
Share of functioning toilets	0.90	0.88	0.03	0.05	0.01	0.59
	[0.22]	[0.23]	(0.04)	(0.05)	(0.05)	
WTP (avg. catchment area)	1.41	1.44	-0.04	-0.04	-0.04	0.97
	[0.83]	[0.65]	(0.17)	(0.19)	(0.21)	
Distance from CT (avg. catchment area)	128.71	128.77	1.91	-0.23	4.24	0.93
	[49.56]	[43.87]	(10.39)	(11.17)	(13.05)	

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 the maintenance 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 randomization strata. 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.

Table D2: Household characteristics at baseline, by treatment group

	Descriptive statistics		**Differences from control group, by treatment group**			
	All	Control	Maintenance	Maintenance only	Maintenance + sensitization	p-value joint test (4)-(5)
	(1)	(2)	(3)	(4)	(5)	(6)
Household head is male	0.75	0.73	0.04	0.05**	0.02	0.10
	[0.43]	[0.44]	(0.02)	(0.03)	(0.03)	
Household head is married	0.77	0.76	0.01	0.01	0.01	0.87
	[0.42]	[0.43]	(0.02)	(0.02)	(0.03)	
Age of household head	45.44	46.04	-0.99	-0.94	-1.05	0.34
	[12.82]	[13.42]	(0.67)	(0.77)	(0.79)	
Age of spouse	39.14	39.39	-0.41	-0.68	-0.12	0.61
	[11.39]	[12.00]	(0.59)	(0.73)	(0.63)	
Household head has no education	0.54	0.55	-0.03	-0.08**	0.02	0.03
	[0.50]	[0.50]	(0.03)	(0.04)	(0.04)	
Spouse has no education	0.45	0.45	-0.01	-0.01	-0.02	0.86
	[0.50]	[0.50]	(0.03)	(0.03)	(0.03)	
Household members	4.94	4.94	-0.00	0.01	-0.02	0.98
	[1.99]	[2.08]	(0.13)	(0.15)	(0.14)	
Household members (0-5 y.o.)	0.47	0.50	-0.04	-0.04	-0.04	0.71
	[0.77]	[0.82]	(0.05)	(0.06)	(0.05)	
Household members (older than 5 y.o.)	4.47	4.44	0.03	0.04	0.02	0.96
	[1.83]	[1.92]	(0.11)	(0.13)	(0.12)	
Muslim	0.17	0.12	0.06	0.09*	0.03	0.20
	[0.37]	[0.32]	(0.04)	(0.05)	(0.05)	
Spent on religious items	0.94	0.94	-0.01	-0.01	-0.01	0.73
	[0.25]	[0.24]	(0.01)	(0.01)	(0.01)	
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	0.01	0.01	0.00	0.70
	[0.15]	[0.16]	(0.01)	(0.02)	(0.01)	
Household members per room	3.99	3.90	0.13	0.05	0.20*	0.22
	[1.86]	[1.94]	(0.11)	(0.13)	(0.12)	
Access to piped water	0.71	0.70	0.00	-0.01	0.02	0.80
	[0.45]	[0.46]	(0.05)	(0.06)	(0.05)	
Access to private toilet	0.08	0.07	0.01	0.00	0.02	0.74
	[0.27]	[0.26]	(0.01)	(0.02)	(0.02)	
Expenditure on CT use (INR)	180.75	173.72	-1.58	-10.56	7.79	0.73
	[244.60]	[221.49]	(16.06)	(17.74)	(22.74)	
Distance to CT (meters)	126.22	126.42	2.22	3.28	1.10	0.93
	[79.90]	[80.42]	(7.61)	(8.41)	(9.50)	

Note. Columns (1)–(2) report sample mean with standard deviation in brackets for the whole sample and for the control group, respectively. Column (3)–(5) report differences from the control group of treatment groups, estimated using OLS and controlling for randomization strata. 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.

Table D3 tests whether attrition led to differences between treatment arms. In the *CT survey*, we collected data for all 110 CTs selected at baseline, but only for 108 in the mid-intervention survey, 109 in follow-up 1, 107 in follow-up 2, 105 in follow-up 3, and 106 in follow-up 4, as some CTs were closed for rehabilitation, and one slum was completely displaced after follow-up 1. In some cases, we were able to collect observations, but were unable to interview caretakers. In two different CTs in the *maintenance* treatment group, another CT opened in their vicinity after baseline. We assigned these CTs to the same treatment arm, performed the interventions, and interviewed them during the follow-up rounds. In the *resident survey*, we interviewed 1,575 households at baseline (an average of 12 households per cluster), 1,532 households at the mid-intervention survey, 1,578 households at follow-up 2, and 1,772 households at follow-up 4. On average, each interview lasted one hour. In order to maintain a comparable sample size across surveys, we handled attrition by randomly replacing respondents using the sampling frame used for the baseline survey. Column (7) tests whether replacement was introduced differently across treatments.

Table D3: Attrition in CT and residents' measurements

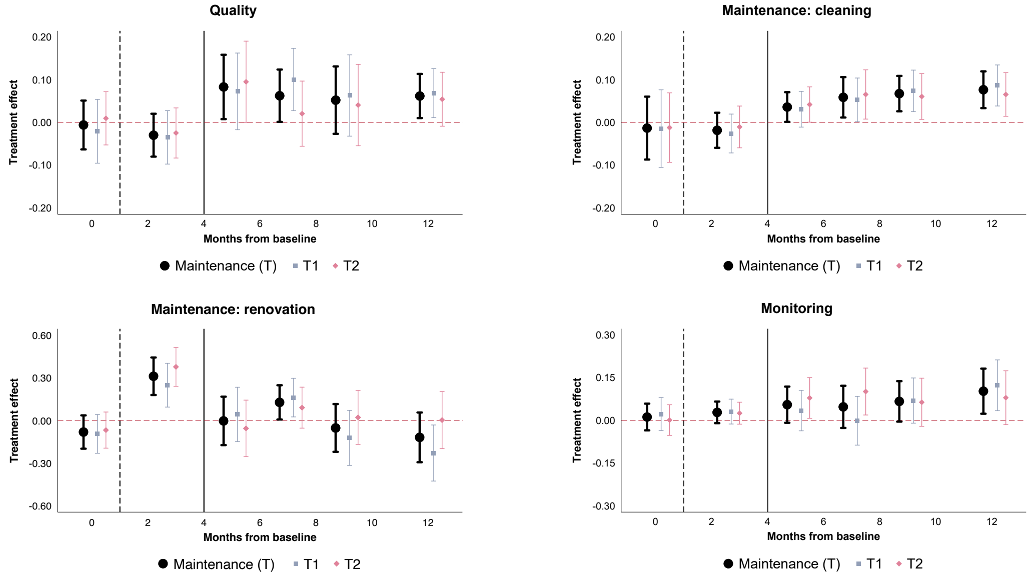
	CT measurements		Residents' measurements				
	Observations	Interviews	Interviews	Interviewed at BL and not in ...			Replacement
				Any	FU 2	FU 4	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A							
Maintenance (T)	0.084	0.140	0.032	-0.012	-0.021	-0.010	0.012
	(0.073)	(0.119)	(0.049)	(0.024)	(0.028)	(0.027)	(0.023)
	[0.25]	[0.24]	[0.52]	[0.60]	[0.44]	[0.70]	[0.60]
Panel B							
Maintenance only (T1)	0.084	0.077	0.045	-0.015	-0.033	-0.012	0.011
	(0.079)	(0.145)	(0.053)	(0.025)	(0.031)	(0.030)	(0.026)
	[0.29]	[0.60]	[0.40]	[0.55]	[0.29]	[0.68]	[0.68]
Maintenance + sensitization (T2)	0.084	0.202	0.018	-0.010	-0.010	-0.008	0.013
	(0.078)	(0.125)	(0.058)	(0.028)	(0.033)	(0.032)	(0.030)
	[0.29]	[0.11]	[0.75]	[0.72]	[0.77]	[0.80]	[0.66]
Mean (dep. var.)	3.973	3.873	1.645	0.079	0.200	0.154	0.221
Observations	110	110	1573	1573	1573	1573	3323

Note. Estimates based on OLS regressions using equation (9) in panel A, and equation (10) in panel B. Robust standard errors are presented in parenthesis in columns (1)–(2). Standard errors clustered by catchment area are presented in parenthesis in columns (3)–(7) The p-values are presented in brackets. Dependent variables by column: (1) *Observations*, number of follow-up surveys where CT observation were collected; and (2) *Interviews*, number of post-intervention surveys with the caretaker; (3) *Interviews*, number of post-intervention surveys for households interviewed at baseline; (4)–(6) *Interviewed at BL and not in ...*, indicator variable equal to 1 if the household was interviewed at baseline, but was not re-interviewed after, and zero if re-interviewed; (6) *Replacement*, indicator variable equal to 1 if the household is part of the replacement sample, and 0 otherwise. In columns (3)–(6), the sample is restricted to baseline observations, while in column (7) the sample is restricted to follow-up observations. All specifications control for randomization strata. Figure C1 provides the timing of each follow-up survey.

D.2 Estimates of treatment effects by survey round

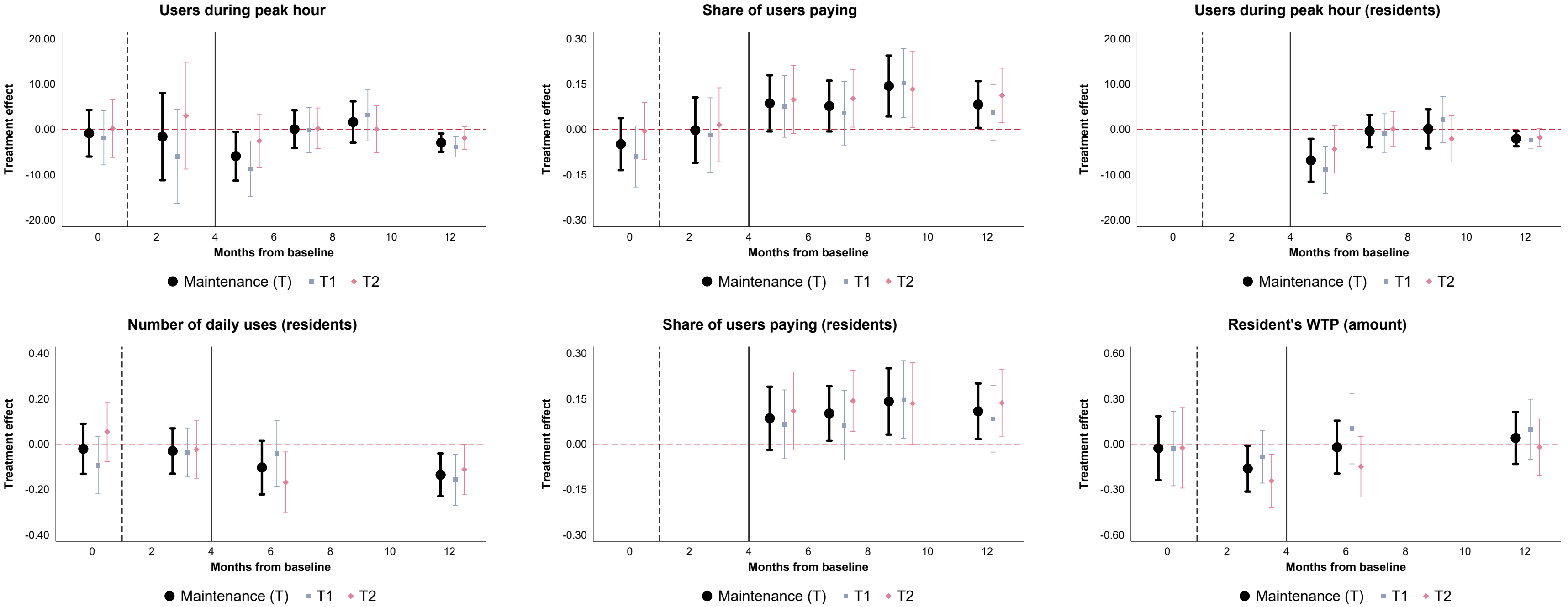
Figures D1–D3 present estimates of equation (9) and equation (10) separately for each survey.

Figure D1: Timing of effects for outcomes in Table 1



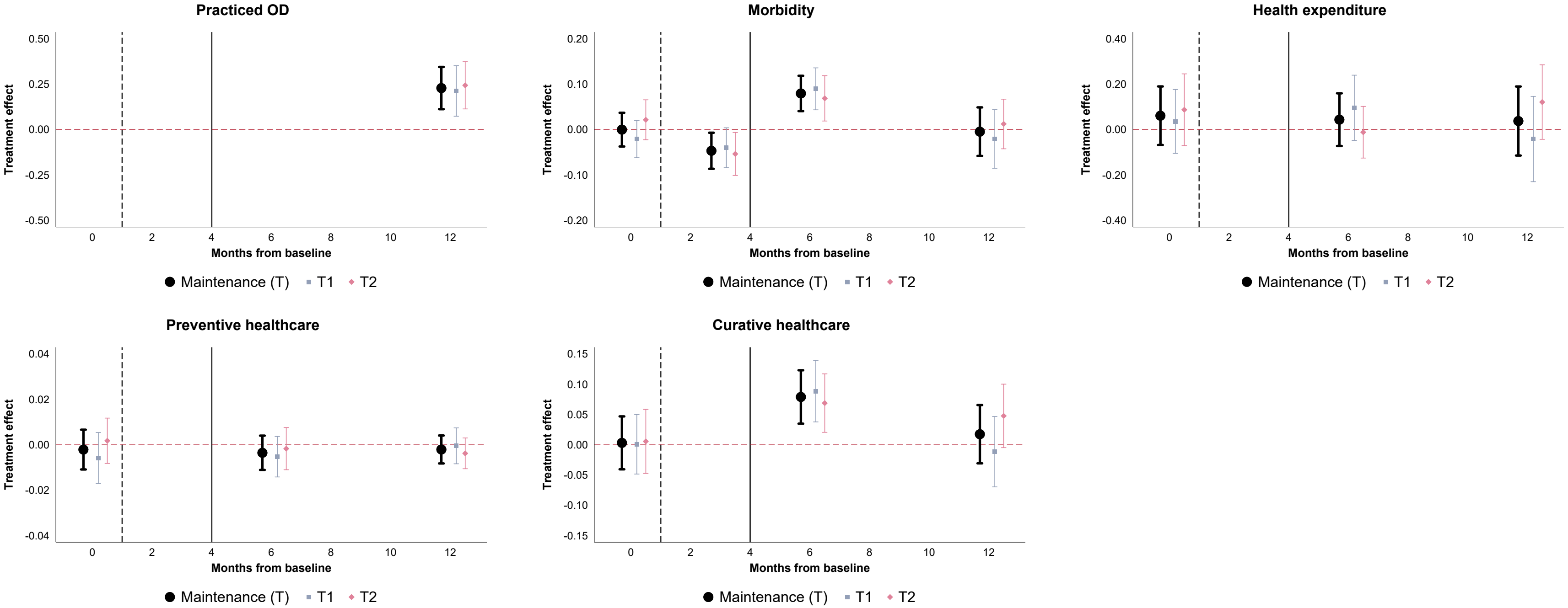
Notes. Estimates based on CT-level OLS regressions using equation (9) and equation (10) separately for each data collection period. Period 0 indicates the *baseline* measurement. The measurement in between the two vertical lines is the *mid-intervention* measurement. All subsequent periods (to the right of the vertical solid line) are the *follow-up* measurements. Confidence intervals are computed at the 90% level of confidence using robust standard errors. Outcome variables are defined in Appendix B. All specifications control for randomization strata.

Figure D2: Timing of effects for outcomes in Table 2



Notes. Estimates based on CT-level OLS regressions using equation (9) and equation (10) separately for each data collection period. Period 0 indicates the *baseline* measurement. The measurement in between the two vertical lines is the *mid-intervention* measurement. All subsequent periods (to the right of the vertical solid line) are the *follow-up* measurements. Confidence intervals are computed at the 90% level of confidence using robust standard errors. Outcome variables are defined in Appendix B. All specifications control for randomization strata.

Figure D3: Timing of effects for outcomes in Table 3



Notes. Estimates based on household-level OLS regressions using equation (9) and equation (10) separately for each data collection period. Period 0 indicates the *baseline* measurement. The measurement in between the two vertical lines is the *mid-intervention* measurement. All subsequent periods (to the right of the vertical solid line) are the *follow-up* measurements. Confidence intervals are computed at the 90% level of confidence using standard errors clustered at the catchment area. Outcome variables are defined in Appendix B. All specifications control for randomization strata. Respondent-level regressions include a control for the gender of the respondent.

D.3 Robustness using ANCOVA and IPW specifications

Table D4 presents estimates of treatment effects using equations (9) and (10) adding the value at baseline of the dependent variable as a control variable (ANCOVA specification). Table D4 also presents estimates of treatment effects using equations (9) and (10) weighting observations by inverse probability weights (IPW) to account for attrition (Wooldridge, 2002).

Table D4: Estimates with ANCOVA and IPW specifications

	ANCOVA				**IPW**		
	β	se	p-value	ANCOVA	β	se	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quality	0.06	0.02	0.01	1			
Maintenance: cleaning	0.06	0.01	0.00	1			
Maintenance: renovation	0.00	0.05	0.97	1			
Monitoring	0.06	0.03	0.02	1			
Users during peak hour	-1.62	1.47	0.27	1			
Share of users paying	0.11	0.03	0.00	1			
Users during peak hour (residents)	-2.26	1.25	0.07	0			
Share of users paying (residents)	0.11	0.03	0.00	0			
Resident's WTP (amount)	0.00	0.07	0.96	1	0.02	0.08	0.78
Practiced OD	0.23	0.07	0.00	0	0.22	0.07	0.00
Number of daily uses	-0.12	0.05	0.02	1	-0.11	0.06	0.07
Morbidity	0.04	0.02	0.13	1	0.03	0.02	0.14
Health expenditure	0.03	0.07	0.63	1	0.04	0.07	0.61
Preventive healthcare	-0.00	0.00	0.30	1	-0.00	0.00	0.32
Curative healthcare	0.05	0.02	0.03	1	0.05	0.02	0.03

Note. Estimates based on respondent- and household-level OLS regressions using equation (9), controlling for the baseline value of the dependent variable if available (see *ANCOVA specification* column (4), 1 = Yes ANCOVA, 0 = ANCOVA not possible) in columns (1)–(3), and weighting observations by inverse probability weights in columns (5) to (7). Column (4) indicates whether the baseline value is available. 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 randomization strata. Specifications where the level of analysis is the respondent also include gender. Additional details about the variables are presented in Appendix B.

D.4 Robustness to the inclusion of control variables

Table D5 presents estimates of the effect of the maintenance treatment (T) using equation (9) in columns (1)–(3), and the post-double selection LASSO (PDSL) procedure in columns (4)–(6). The PDSL procedure provides a method for model selection in the presence of a large number of control 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 D1); *caretaker characteristics* (variables related to caretakers at baseline included in Table D1); *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 D2); *outcome variables* (when available, we include the baseline value of outcomes presented in Tables 1–3). Table D6 presents estimates of ATE of the maintenance treatment on all outcome variables using the causal forest procedure of Athey et al. (2019). In the procedure, we use the set of variables from Appendix D.4. Figure D4 summarizes the causal forest results on heterogeneity of the effect on payment, showing the distribution of the Conditional ATE (CATE) and the average CATE at CT level with the 90% confidence interval.

Table D5: Effect of the maintenance 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)
Quality	0.07	0.02	0.00	0.05	0.02	0.02	434
Maintenance: cleaning	0.06	0.01	0.00	0.06	0.01	0.00	434
Maintenance: renovation	-0.01	0.05	0.80	-0.03	0.05	0.58	434
Monitoring	0.07	0.03	0.02	0.06	0.02	0.00	434
Total users during peak hour	-1.78	1.52	0.24	-2.20	1.32	0.10	434
Share of users paying	0.10	0.03	0.00	0.10	0.03	0.00	434
Users during peak hour (residents)	-2.26	1.25	0.07	-2.71	1.03	0.01	434
Share of users paying (residents)	0.11	0.03	0.00	0.11	0.03	0.00	434
Number of daily uses	-0.12	0.06	0.05	-0.12	0.05	0.03	3303
Resident's WTP (amount)	0.01	0.08	0.92	-0.01	0.08	0.85	6001
Practiced OD	0.23	0.07	0.00	0.17	0.06	0.01	817
Morbidity	0.03	0.02	0.14	0.04	0.02	0.07	3323
Health expenditure	0.04	0.07	0.57	0.03	0.06	0.61	3306
Preventive healthcare	-0.00	0.00	0.31	-0.00	0.00	0.40	3322
Curative healthcare	0.05	0.02	0.03	0.05	0.02	0.02	3298

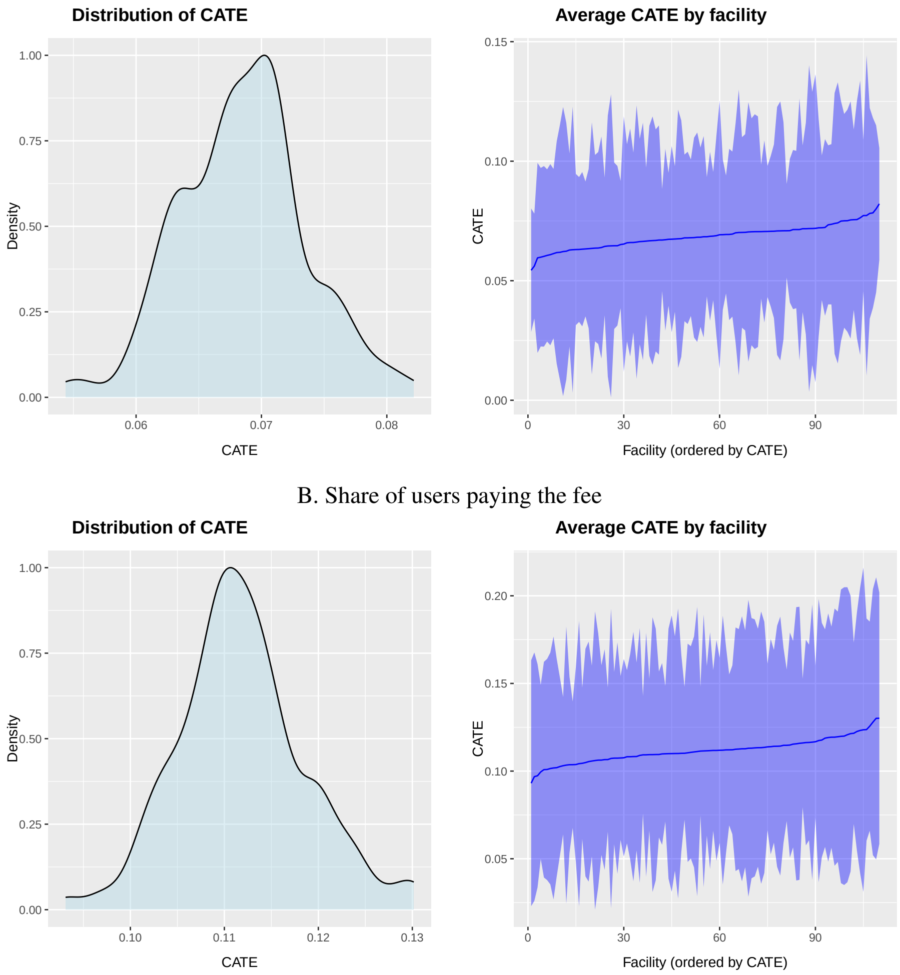
Note. Columns (1)–(3) show estimates using equation (9), while columns (4)–(6) show estimates using the PDSL procedure (Tibshirani, 1996; Belloni et al., 2013), with selection over a large number of baseline-level control variables. All specifications include indicator variables for data collection rounds and randomization strata. 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. Additional information about outcome variables is provided in Appendix B.

Table D6: 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)
Quality	0.069	0.027	0.012	0.000	1.000
Maintenance: cleaning	0.06	0.016	0.000	0.000	1.000
Maintenance: renovation	-0.031	0.055	0.578	0.261	1.000
Monitoring	0.056	0.036	0.116	0.097	1.000
Users during peak hour	-1.578	1.682	0.348	0.160	0.604
Share of users paying	0.115	0.044	0.008	0.000	1.000
Users during peak hour (residents)	-1.927	1.388	0.165	0.050	0.998
Number of daily uses	-0.085	0.068	0.212	0.09	0.372
Share of users paying (residents)	0.123	0.046	0.007	0.000	1.000
WTP among residents	-0.002	0.100	0.984	0.549	0.901
Practiced OD	0.229	0.092	0.013	0.000	1.000
Morbidity	0.018	0.028	0.516	0.23	0.954
Health expenditure	0.026	0.078	0.743	0.363	0.993
Curative healthcare	-0.002	0.003	0.458	0.238	1.000
Preventive healthcare	0.04	0.027	0.141	0.049	0.986

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 D.4, 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 (heterogeneity present) of the coefficient on the quality of the estimates of treatment heterogeneity. Additional information about outcome variables is provided in Appendix B.

Figure D4: Conditional ATE of the maintenance treatment on quality and payment


A. Quality

B. Share of users paying the fee

Note. Distribution of CATE is the distribution of the Conditional ATE (CATE) of the maintenance treatment on payment computed using the cluster-robust causal forest procedure of Basu et al. (2018) and Athey and Wager (2019). *Average CATE by facility* is the average CATE at CT level with the 90% confidence interval. Additional information about the variables is provided in Appendix B.

D.5 Implementation of interventions and spillover analysis

Table D7 shows the effect of the treatments on indicators of exposure to interventions. We focus on transfers as part of the maintenance intervention and on indicators of the sensitization campaign. In columns (1) and (2), transfers are per period, and therefore total transfers can be obtained by multiplying the estimate by the number of rounds of observation. Table D8 presents a test for contagion or spillover effects by estimating heterogeneous treatment effects according to the minimum distance to a CT in the treatment group. The estimates are based on the equation (11) (see Appendix D.7). We do not observe any heterogeneous effect for any of the outcome variables, suggesting the absence of spillover effects.

Table D7: Exposure to the interventions, by component

	Maintenance		**Sensitization campaign**		
	Transfer to the ...		Recall of WASH campaigns		Awareness
	CT	Caretaker	Interactive activities	Posters	
	(1)	(2)	(3)	(4)	(5)
Panel A					
Maintenance (T)	4.741	0.757	0.057	0.096	0.035
	(0.054)	(0.033)	(0.016)	(0.021)	(0.017)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.04]
Panel B					
Maintenance only (T1)	4.646	0.745	0.032	0.033	0.014
	(0.070)	(0.040)	(0.018)	(0.022)	(0.021)
	[0.00]	[0.00]	[0.09]	[0.13]	[0.51]
Maintenance + sensitization (T2)	4.850	0.771	0.084	0.162	0.056
	(0.071)	(0.045)	(0.019)	(0.025)	(0.018)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]
T1 = T2 (p-value)	0.026	0.638	0.015	0.000	0.060
Mean (control group)	0.315	0.063	0.646	0.327	0.659
Std. dev. (control group)	0.358	0.025	0.478	0.469	0.474
Observations	560	560	4844	3323	4844
Catchment areas	110	110	110	109	110
Observation rounds	5	5	3	2	3

Note. In columns (1) and (2), estimates are based on CT-level OLS regressions using equation (9) in panel A, and equation (10) 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 (9) in panel A, and equation (10) in panel B. Standard errors clustered by catchment area–round are reported in parentheses. The p-values 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 defined in Appendix B. All specifications include indicator variables for data collection rounds and randomization strata.

Table D8: Contagion and spillover effects

	Effect of maintenance treatment from equation (11)				
	Main effect		Heterogeneity by distance to another treatment unit		
	β	se	δ	se	N
Outcome variable	(1)	(2)	(3)	(4)	(5)
Quality	0.07***	0.02	-0.02	0.04	434
Maintenance: cleaning	0.06***	0.01	-0.03	0.02	434
Maintenance: renovation	-0.01	0.05	-0.02	0.04	434
Monitoring	0.06**	0.03	-0.05	0.04	434
Users during peak hour	-1.78	1.51	0.09	1.54	434
Share of users paying	0.10***	0.03	-0.08	0.05	434
Resident's WTP (amount)	-0.04	0.08	-0.14	0.09	4717
Users during peak hour (residents)	-2.19*	1.23	1.55	1.46	434
Share of users paying (residents)	0.11***	0.03	-0.08	0.05	434
Number of daily uses	-0.11*	0.06	-0.04	0.07	2601
Practiced OD	0.15**	0.08	0.04	0.09	633
Morbidity	0.05*	0.03	0.02	0.03	2617
Health expenditure	0.07	0.07	0.06	0.08	2603
Preventive healthcare	-0.01	0.00	0.01	0.00	2616
Curative healthcare	0.06**	0.02	0.03	0.02	2600

Note. The heterogeneity dimension is the distance between the CT in the slum and the closest CT outside the slum and selected for the study. Estimates are based on service-, caretaker-, respondent- and household-level OLS regressions using equation (11). Columns (1)–(2) refer to the main effect, while columns (3)–(4) refer to the interaction between the between the treatment indicator T and the distance to another treatment unit (reported in standardized units). Column (5) reports the number of observations. Standard errors clustered by catchment area. The dependent variables are indicated in the rows and are defined in Appendix B. All specifications include indicator variables for data collection rounds and randomization strata. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1.

D.6 Quality and inputs of service delivery: construction and effects

We measure the overall **quality of service delivery** using all the observed indicators related to the facility's structural quality and cleanliness of the facility and the absence of harmful bacteria. We use item response theory (IRT), a technique used to describe the relationship between individual responses to questionnaire items and an unobserved latent characteristic (Gordon et al., 2012; Kline, 2014). We use a two-parameter IRT model with an ability score, which represents the weights in the index, and a discrimination score, which measures how well the indicator discriminates between low and high quality. The index is rescaled to be between 0 (lowest quality) and 1 (highest quality).[2] Table D9 gives the list of all indicators included. In addition, we construct three separate indices using IRT to measure structural quality, visible cleanliness and the absence of bacteria. Figure D5 shows the effect of the maintenance treatment on each component. Regarding caretaker inputs and routine maintenance, Figure D6 shows the empirical cumulative distribution functions of the total number of hours worked by the caretaker (panel A) and the proportion of time spent on monitoring activities (panel B), distinguishing between the control and treatment groups, while Table D10 shows the estimates of treatment effects on the individual indicators used to construct the routine maintenance indicator used in the main text.

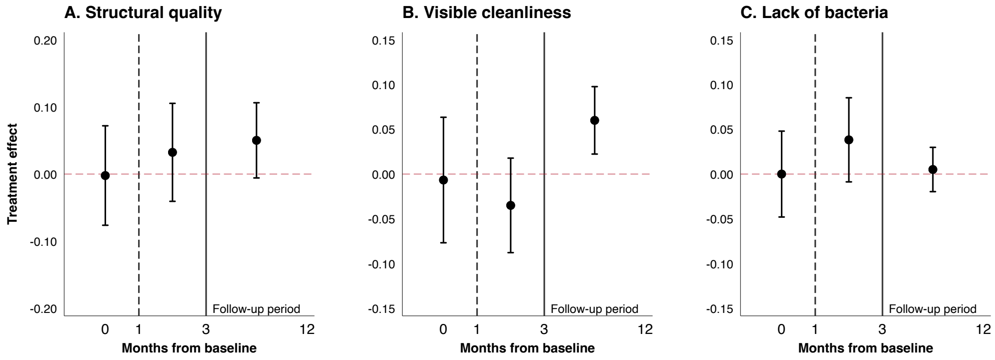
[2] We compute the index separately for the baseline and subsequent surveys because the baseline survey includes a smaller number of indicators. At baseline, we use a one-parameter IRT model due to convergence.

Table D9: Indicators used for the construction of the quality index

Category	Indicator variables	Ability score	Discrimination
Structural quality	All cubicle doors are functioning	1.971	0.247
Structural quality	All locks are functioning	-0.603	0.435
Structural quality	Compound has finished walls	2.259	0.412
Structural quality	Internal walls are in good condition	3.156	0.294
Structural quality	Soap is available and visible for both genders	1.731	0.572
Structural quality	Hand-washing facility available for both genders	1.667	0.811
Structural quality	Female area has lighting	1.842	1.002
Structural quality	Male area has lighting	1.751	1.059
Structural quality	Common area has lighting	2.960	0.762
Visible cleanliness	Toilets in female area are not dirty	0.699	3.705
Visible cleanliness	Toilets in female area do not stink	0.640	4.121
Visible cleanliness	Flies not present in the female area	0.837	3.904
Visible cleanliness	Toilets in male area are not dirty	0.570	4.843
Visible cleanliness	Toilets in male area do not stink	0.771	3.431
Visible cleanliness	Flies not present in the male area	0.525	5.990
Visible cleanliness	Feces not visible inside the latrine in the female area	1.009	5.186
Visible cleanliness	Feces not visible outside the latrine in the female area	1.200	4.523
Visible cleanliness	Feces not visible inside the latrine in the male area	0.987	3.699
Visible cleanliness	Feces not visible outside the latrine in the male area	1.192	3.134
Visible cleanliness	Common area is not dirty	1.276	2.924
Visible cleanliness	Common area does not stink	1.254	3.254
Visible cleanliness	not present in the common area	1.272	2.764
Visible cleanliness	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
Lack of bacteria	Bacteria of bacillus are not detected	2.148	-3.145
Lack of bacteria	Bacteria of staphylococcus are not detected	-25.405	-0.097
Lack of bacteria	Bacteria of salmonella are not detected	38.091	0.025
Lack of bacteria	Bacteria of klebsiella are not detected	10.820	-0.123
Lack of bacteria	Mold is not detected	3.537	-0.455

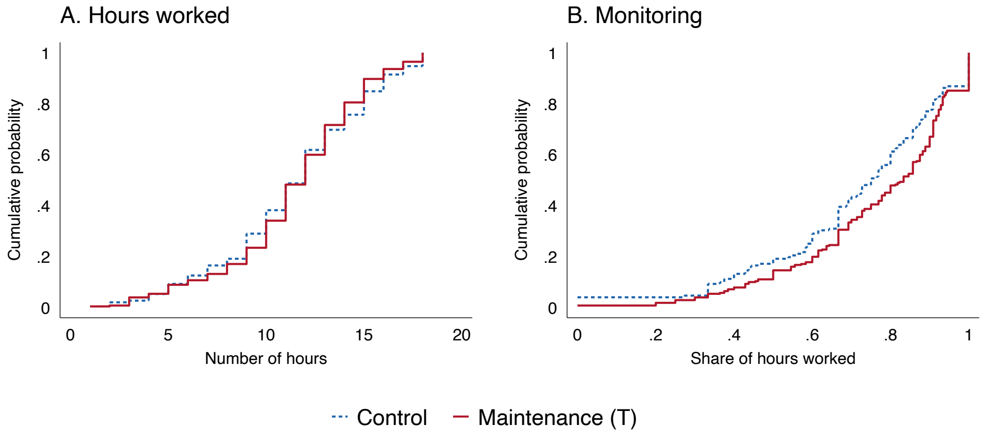
Note. All indicator variables are equal to 1 if the condition is true, and 0 otherwise. The table reports the main parameters in the index build using IRT: the ability score and the discrimination. Observations are restricted to the mid-intervention survey and all follow-ups surveys. The manual for observers defines the rules for the visual evaluation of CTs. *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 contaminated black waters leaking from a septic tank, pit/cesspool or pipes.

Figure D5: Effect on CT quality by component of the index



Note. Each panel presents estimates of treatment effects based on OLS regressions using equation (9) at the CT level. Confidence intervals are built using statistical confidence at the 90% level. Period 0 indicates the *baseline* measurement. The measurement in between period 1 and 3 is the *mid-intervention* measurement. All subsequent periods (to the right of the vertical solid line) are the *follow-up* measurements and are pooled together. See Section 3 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 randomization strata.

Figure D6: Caretaker's labour supply and effort



Note. The figure shows the empirical cumulative distribution functions of the total number of hours worked by the caretaker (Panel A) and of the share of time allocated to monitoring activities (Panel B), distinguishing between control and treatment group. The sample include all follow-up measurements. The p-value of a Kolmogorov–Smirnov test of equality of distributions is equal to 0.900 for Panel A, and 0.020 for Panel B. Additional details about the variables are presented in Appendix B.

Table D10: Inputs in routine maintenance

Dep. variable:	**Tools used during routine maintenance**						**Other inputs**	
	Broom or brush	Mop	Bucket of water	Disin-fectants	Pressurized water	Safety equip-ment	Employed cleaners	Correct imple-menta-tion
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Maintenance (T)	-0.001	0.073	0.040	0.006	0.042	0.033	0.146	0.118
	(0.012)	(0.032)	(0.043)	(0.024)	(0.027)	(0.024)	(0.080)	(0.039)
	[0.90]	[0.02]	[0.35]	[0.81]	[0.12]	[0.17]	[0.07]	[0.00]
Mean (control group)	0.987	0.717	0.711	0.947	0.066	0.039	0.579	0.059
Observations	434	434	434	434	434	434	434	434
Catchment areas	110	110	110	110	110	110	110	110
Observation rounds	4	4	4	4	4	4	4	4

Note. Estimates based on CT-level OLS regressions using equation (9). Standard errors clustered by catchment area are reported in parentheses. The p-values presented in brackets. Dependent variables are indicator variables for whether the tools were used in the last routine maintenance, whether cleaners were employed, and whether the caretakers applies correct cleaning procedures. *Correct implementation* is an 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. All specifications include indicator variables for data collection rounds and randomization strata. Additional details about the variables are presented in Appendix B.

D.7 Treatment heterogeneity by pre-specified dimensions

We estimate heterogeneous effects with respect to characteristic H (measured at baseline) on the outcome Y_{ij} of CT/household/individual i in catchment area j using the following specification:

$$Y_{ij} = \beta\, T_j + \delta\, T_j \cdot H_{ij} + \gamma\, H_{ij} + \alpha\, \mathbf{X}_{ij} + \epsilon_{ij}. \tag{11}$$

Presence of heterogeneous treatment effects is captured by the parameter δ being statistically different from zero. Tables D11 and D12 present estimates of β and δ for heterogeneity by service delivery characteristics, and by other catchment area and individual characteristics. For ease of interpretation of the coefficients on the interaction terms, we standardize H_{ij} in equation (11) to interpret heterogeneous effects per one-standard-deviation change in H_{ij}.

Table D11: Effect of maintenance treatment, heterogeneous effects 1/2

	Main effect		**Heterogeneity**		
Heterogeneity dimension	β	se	δ	se	N
Outcome variables	(1)	(2)	(3)	(4)	(5)
A. Quality of the service					
Quality	0.07***	0.02	0.03	0.03	434
Maintenance: cleaning	0.06***	0.01	0.01	0.01	434
Maintenance: renovation	-0.01	0.05	-0.09**	0.04	434
Monitoring	0.07**	0.03	0.06***	0.02	434
Users during peak hour	-1.76	1.49	-2.42	1.64	434
Share of users paying	0.10***	0.03	-0.03	0.03	434
Users during peak hour (residents)	-2.24*	1.22	-2.29	1.41	434
Share of users paying (residents)	0.11***	0.03	-0.05	0.04	434
Number of daily uses	-0.11*	0.06	-0.17***	0.06	2601
Resident's WTP (amount)	-0.03	0.08	0.00	0.11	4717
Practiced OD	0.18**	0.07	-0.05	0.09	633
Morbidity	0.05*	0.02	-0.02	0.03	2617
Health expenditure	0.07	0.07	0.16	0.09	2603
Preventive healthcare	-0.01*	0.00	-0.00	0.00	2616
Curative healthcare	0.05**	0.02	0.03	0.02	2600
B. WTP for service use[1]					
Quality	0.07***	0.02	0.01	0.03	434
Maintenance: cleaning	0.06***	0.01	-0.04**	0.02	434
Maintenance: renovation	-0.01	0.05	-0.06	0.07	434
Monitoring	0.07**	0.03	-0.00	0.05	434
Users during peak hour	-1.61	1.46	-3.42	2.22	434
Share of users paying	0.10***	0.03	-0.03	0.05	434
Users during peak hour (residents)	-2.12*	1.22	-1.77	1.73	434
Share of users paying (residents)	0.11***	0.03	-0.01	0.05	434
Number of daily uses	-0.11*	0.06	0.02	0.05	2601
Resident's WTP (amount)	-0.03	0.08	-0.05	0.09	4717
Practiced OD	0.16**	0.07	-0.01	0.09	633
Morbidity	0.05*	0.03	0.01	0.02	2617
Health expenditure	0.07	0.07	0.04	0.05	2603
Preventive healthcare	-0.01*	0.00	0.01*	0.00	2616
Curative healthcare	0.05**	0.02	0.04*	0.02	2600
C. Environmental contamination[1]					
Quality	0.07***	0.02	0.05*	0.03	434
Maintenance: cleaning	0.06***	0.01	0.04**	0.02	434
Maintenance: renovation	-0.03	0.04	-0.25***	0.04	434
Monitoring	0.07**	0.03	-0.02	0.04	434
Users during peak hour	-1.97	1.50	-2.79	1.70	434
Share of users paying	0.10***	0.03	-0.07	0.05	434
Users during peak hour (residents)	-2.44**	1.21	-2.58**	1.27	434
Share of users paying (residents)	0.11***	0.03	-0.08	0.05	434
Number of daily uses	-0.12**	0.06	-0.13*	0.08	2601
Resident's WTP (amount)	-0.03	0.08	-0.09	0.09	4717
Practiced OD	0.16**	0.07	-0.03	0.08	633
Morbidity	0.05*	0.03	-0.04	0.03	2617
Health expenditure	0.07	0.07	-0.02	0.08	2603
Preventive healthcare	-0.01*	0.00	0.00	0.00	2616
Curative healthcare	0.05**	0.02	0.01	0.03	2600

Note. Heterogeneity dimensions are reported in each panel's title and measured at baseline. When these variables are missing at baseline, we impute them using the average within the catchment area, or within the randomization strata (if the variable remains missing). Estimates are based on service- or caretaker-level OLS regressions using equation (11). Columns (1)–(2) refer to the main effect, while columns (3)–(4) refer to the interaction between the between the treatment indicator T and the heterogeneity variable H (reported in standardized units). Column (5) reports the number of observations. Standard errors clustered by catchment area. The dependent variables are indicated in the rows and are defined in Appendix B. All specifications include indicator variables for data collection rounds and randomization strata. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1. [1] For outcome variables measured at the CT-level, the heterogeneity dimension is averaged within the catchment area. *Environmental contamination* is captured by the average E.coli count in water samples collected in the slum (see Appendix F).

Table D12: Effect of maintenance treatment, heterogeneous effects 2/2

	Main effect		Heterogeneity		
Heterogeneity dimension	β	se	δ	se	N
Outcome variables	(1)	(2)	(3)	(4)	(5)
A. Knowledge of hygiene practices[1]					
Quality	0.07***	0.02	-0.01	0.02	434
Maintenance: cleaning	0.06***	0.01	0.01	0.02	434
Maintenance: renovation	-0.01	0.05	0.07	0.05	434
Monitoring	0.07**	0.03	0.04	0.03	434
Users during peak hour	-1.74	1.49	0.46	1.94	434
Share of users paying	0.10***	0.03	-0.02	0.05	434
Users during peak hour (residents)	-2.28*	1.21	-0.74	1.59	434
Share of users paying (residents)	0.11***	0.03	-0.04	0.05	434
Number of daily uses	-0.11*	0.06	-0.04	0.05	2601
Resident’s WTP (amount)	-0.02	0.08	-0.04	0.07	4717
Practiced OD	0.16**	0.08	0.07	0.09	633
Morbidity	0.05*	0.03	-0.03	0.02	2617
Health expenditure	0.07	0.07	-0.15**	0.06	2603
Preventive healthcare	-0.01*	0.00	-0.01**	0.00	2616
Curative healthcare	0.05**	0.02	-0.05*	0.03	2600
B. Social capital[1]					
Quality	0.07***	0.02	0.01	0.03	434
Maintenance: cleaning	0.06***	0.01	-0.02	0.02	434
Maintenance: renovation	-0.01	0.05	0.02	0.05	434
Monitoring	0.07**	0.03	0.05	0.03	434
Users during peak hour	-1.92	1.53	0.26	1.35	434
Share of users paying	0.10***	0.03	-0.02	0.04	434
Users during peak hour (residents)	-2.39*	1.25	-0.45	1.18	434
Share of users paying (residents)	0.12***	0.03	-0.01	0.04	434
Number of daily uses	-0.11*	0.06	-0.04	0.04	2601
Resident’s WTP (amount)	-0.03	0.08	-0.04	0.07	4717
Practiced OD	0.16**	0.08	0.07	0.08	633
Morbidity	0.05*	0.03	0.00	0.02	2617
Health expenditure	0.07	0.07	-0.01	0.06	2603
Preventive healthcare	-0.01*	0.00	-0.00	0.00	2616
Curative healthcare	0.05**	0.02	-0.04*	0.02	2600
C. Caretaker’s intrinsic motivation					
Quality	0.07***	0.02	0.04	0.03	434
Maintenance: cleaning	0.06***	0.01	-0.00	0.02	434
Maintenance: renovation	-0.01	0.05	-0.01	0.06	434
Monitoring	0.07**	0.03	0.02	0.04	434
Users during peak hour	-1.92	1.47	-2.96	1.94	434
Share of users paying	0.10***	0.03	0.03	0.04	434
Users during peak hour (residents)	-2.34*	1.22	-1.96	1.65	434
Share of users paying (residents)	0.11***	0.03	0.05	0.04	434
Number of daily uses	-0.11*	0.06	-0.02	0.08	2601
Resident’s WTP (amount)	-0.03	0.09	-0.06	0.12	4717
Practiced OD	0.18**	0.07	-0.20**	0.08	633
Morbidity	0.05*	0.03	0.02	0.04	2617
Health expenditure	0.07	0.07	-0.02	0.07	2603
Preventive healthcare	-0.01*	0.00	-0.00	0.00	2616
Curative healthcare	0.05**	0.02	0.01	0.03	2600

Note. Heterogeneity dimensions are reported in each panel’s title and measured at baseline. When these variables are missing at baseline, we impute them using the average within the catchment area, or within the randomization strata (if the variable remains missing). Estimates are based on respondent- and household-level OLS regressions using equation (11). Columns (1)–(2) refer to the main effect, while columns (3)–(4) refer to the interaction between the between the treatment indicator T and the heterogeneity variable H (reported in standardized units). Column (5) reports the number of observations. Standard errors clustered by catchment area. The dependent variables are indicated in the rows and are defined in Appendix B. All specifications include indicator variables for data collection rounds and randomization strata. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1.. [1]For outcome variables measured at the CT-level, the heterogeneity dimension is averaged within the catchment area. *Knowledge of hygiene practices* is captured by awareness, an indicator variable equal to 1 if the respondent reports that OD generates a health externality for their family, and 0 otherwise. *Social capital* is captured by trust in the community, an indicator variable equal to 1 if the respondent trust the community to keep the CT clean, and 0 otherwise. *Caretaker’s intrinsic motivation* is captured by the caretaker’s prosocial motivation, measured by the share of the endowment that is donated by the caretaker in the adapted dictator game (see Appendix F).

D.8 Effects on other behavioral measurements

We played two adapted dictator games to measure caregivers' prosocial motivation for the cause and residents' willingness to contribute to the quality of the CT service. Columns (1)–(4) in Table D13 show estimates of the treatment effects on these outcomes. In addition, using information from the list of randomization questions, columns (5)–(6) show estimates of treatment effects on the proportion of study participants who used the CT and washed their hands with soap the day before the interview. The measures are described in Appendix F.

Table D13: Other behavioral measurements

Dep. var.:	Caretaker	Residents				
	Prosocial motivation for the cause	WTP is positive	WTP $\geq$ fee	Willingness to contribute to quality	Used CT	Washed hands with soap
	(1)	(2)	(3)	(4)	(5)	(6)
Maintenance (T)	-0.030	-0.009	-0.001	-0.003	0.108	0.010
	(0.025)	(0.019)	(0.012)	(0.007)	(0.089)	(0.072)
	[0.23]	[0.64]	[0.92]	[0.66]	[0.23]	[0.90]
Mean (control group)	0.343	0.641	0.112	0.212	0.584	0.820
Observations	434	6001	6001	6001	810	839
Catchment areas	110	109	109	109	109	106
Observation rounds	4	2	2	2	1	1

Note. Estimates based on household-level OLS regressions using equation (9). Standard errors clustered by catchment area are reported in parentheses and p-values in brackets. Dependent variables by column: (1) *Prosocial motivation for the cause*, share of the endowment that is donated by the caretaker in the adapted dictator game; (2) *WTP is positive*, indicator variable equal to 1 if the incentivized WTP for a single CT use (in rupees), elicited for a bundle of ten tickets and divided by 10 to get at single use WTP, is positive, and 0 otherwise; (3) *WTP $\geq$ fee*, indicator variable equal to 1 if the incentivized WTP for a single CT use (in rupees), elicited for a bundle of ten tickets and divided by 10 to get at single use WTP, is equal or larger than INR 5, and 0 otherwise; (4) *Willingness to contribute to quality*, share of the endowment that is donated by the respondent in the adapted dictator game; (5) *Used CT*, number of items reported by the respondents assigned to the group including the use of the CT minus the average number of items reported by respondents in the group without sensitive items; (6) *Washed hands with soap*, number of items reported by the respondents assigned to the group including hand-washing with soap minus the average number of items reported by respondents in the group without sensitive items. In columns (5)–(6), the sample is restricted to respondents assigned to the list with the corresponding sensitive item. All specifications include indicator variables for data collection rounds and randomization strata. Additional details about the variables are presented in Appendix B.

D.9 Payment enforcement, timing of use, and revenues

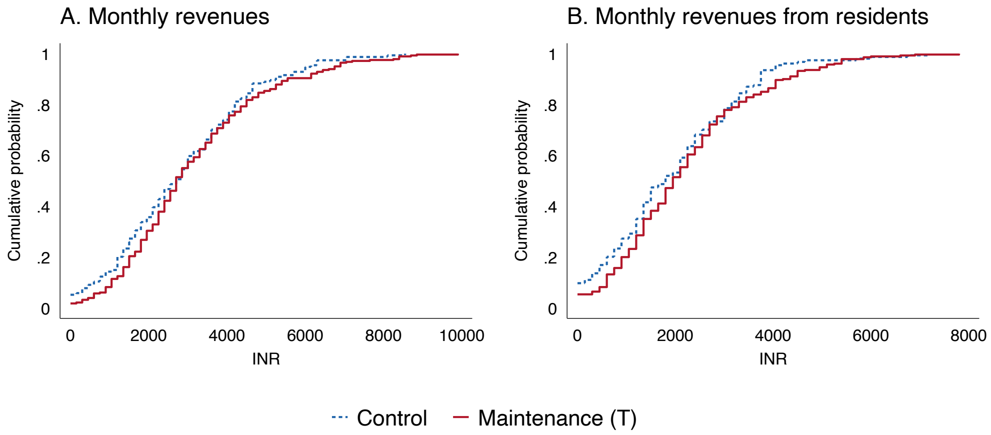
Columns (1)–(2) in Table D14 provide estimates of treatment effects on monthly revenues estimated using observation during the peak hour. Revenues are imputed using information from the observers on the number of people using the CT and the proportion of them paying the fee (assuming a standard fee of INR 5). Figure D7 shows the cumulative distribution functions of these measures of service revenues, distinguishing by treatment group. Column (3) shows the effects on the number of users observed in the afternoon, when the number of users is lower. Columns (4)–(7) show treatment effects for different indicators of payment enforcement reported by the sample of residents.

Table D14: Service revenues, alternative timing of use, and payment enforcement

Dep. variable:	**Monthly revenues during rush hour**		**Alternative timing**	**Caretaker ever refused entry**		**Refused entry for not paying**	
	Total	From residents	Users during afternoon	All CTs	Low payment	All CTs	Low payment
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Maintenance (T)	335.722	268.163	-0.383	0.019	0.069	0.009	0.051
	(195.127)	(143.555)	(0.470)	(0.020)	(0.028)	(0.020)	(0.022)
	[0.09]	[0.06]	[0.42]	[0.34]	[0.02]	[0.64]	[0.03]
Mean (control group)	2840.260	1954.870	12.955	0.076	0.044	0.074	0.041
Observations	434	434	434	1641	812	1641	812
Catchment areas	110	110	110	109	53	109	53
Observation rounds	4	4	4	1	1	1	1

Note. Columns (1)–(3) show estimates based on CT-level OLS regressions using equation (9), while columns (4)–(7) show estimates based on household-level OLS regressions using equation (9). Standard errors clustered by catchment area are reported in parentheses and p-values in brackets. Dependent variables are defined in Appendix B. *Low payment* restricts the sample to CTs that at baseline presented a share of users paying the fee below the sample median. All specifications include indicator variables for data collection rounds and randomization strata. Details about the variables are presented in Appendix B.

Figure D7: Distribution of service revenues during rush hour, by treatment



Note. The figure shows the empirical cumulative distribution functions of total service revenues per month including all users (Panel A) and only regular users (Panel B), and distinguishing between control and treatment group. The sample include all follow-up measurements. The p-value of a Kolmogorov–Smirnov test of equality of distributions is equal to 0.459 for Panel A, and 0.346 for Panel B. Additional details about the variables are presented in Appendix B.

D.10 Use of the service and self-reported use of the outside option

Table D15 shows estimates of treatment effects on self-reported use of the CT and the outside option. In addition, to test whether the distribution of free tickets influenced use, we exploit variation stemming from the distribution of tickets for free CT use as part of the incentivized WTP measurement (see Section 4.2), and estimate a reduced form regression at the household level on the daily uses of the CT at time t on $\tilde{c}_{i,t}$ using the following specification:

$$use_{ij,t} = \lambda_0 + \lambda_1\, tickets_{ij,t-1} + \lambda_2\, WTP_{ij,t-1} + \Omega_t + \epsilon_{ij,t} \qquad (12)$$

where $tickets_{ij,t}$ is an indicator variable equal to 1 if the household received free tickets instead of cash during the previous visit, $WTP_{ij,t}$ is the WTP for CT use elicited in conjunction with the distribution of tickets, Ω_t capture time fixed effects, and $\epsilon_{ij,t}$ captures idiosyncratic unobserved determinants of service use and is assumed to be clustered at the CT level. Because the distribution of tickets versus cash depends on the WTP of the respondent and on a random number extracted as part of the WTP game, we assume that conditional on WTP, receiving the tickets is exogenous to unobserved determinants of service use. Table D16 presents the results.

Table D15: Self-reported use of the CT and of the outside option among residents

	Daily uses of the CT		**Daily uses of other practices**	
	Regular users	Other residents	Regular users	Other residents
	(1)	(2)	(3)	(4)
Maintenance (T)	-0.127	-0.191	0.055	0.068
	(0.042)	(0.080)	(0.044)	(0.091)
	[0.00]	[0.02]	[0.21]	[0.46]
Mean (control group)	1.383	0.763	0.221	0.800
Observations	2417	883	2417	883
Catchment areas	109	102	109	102
Observation rounds	2	2	2	2

Note. Estimates based on household-level OLS regressions using equation (9). Standard errors clustered by catchment area are reported in parentheses and p-values in brackets. Dependent variables by column: (1)–(2) is the number of times a the respondent used the CT for defecation in the day previous to the interview; (3)–(4) is the number of times the person defecated not using the CT the day before the interview. *Regular users* are respondents that reported using the CT regularly. All specifications include indicator variables for data collection rounds and randomization strata. Additional details about the variables are presented in Appendix B.

Table D16: Effect of receiving free tickets versus cash

Dependent variable:	**Number of uses among residents**			
Sub-sample:	Regular users		Other residents	
	(1)	(2)	(3)	(4)
Received free tickets (previous visit)	0.057	0.046	0.157	0.166
	(0.064)	(0.064)	(0.129)	(0.133)
	[0.37]	[0.47]	[0.23]	[0.21]
WTP (previous visit)	0.007	0.009	0.024	0.018
	(0.011)	(0.012)	(0.024)	(0.024)
	[0.54]	[0.45]	[0.32]	[0.45]
Maintenance (T)		-0.120		-0.307
		(0.046)		(0.093)
		[0.01]		[0.00]
Mean (control group)	1.401	1.401	0.765	0.765
Observations	1830	1830	593	593
Catchment areas	109	109	93	93
Observation rounds	2	2	2	2
Level of analysis	Household	Household	Household	Household

Note. Estimates based on household-level OLS regressions using equation (12). Standard errors clustered by catchment area are reported in parentheses. The p-values presented in brackets. The dependent variable is the number of times the respondent used the CT for defecation in the day previous to the interview (*regular users* are respondents that reported using the CT regularly). All specifications include indicator variables for data collection rounds and randomization strata. Additional details about the variables are presented in Appendix B. The measurement of WTP is described in Section 4.2.

D.11 Selection in sanitation behavior

Table D17 shows the correlates of changes in sanitation behavior in the maintenance treatment group.

Table D17: Selection in sanitation behavior between baseline and follow-up 4

Dep. variable:	Stopped using CT	Reduced CT uses
	(1)	(2)
Household head is male	-0.133***	0.016
	(0.041)	(0.044)
Age of household head	0.000	-0.001
	(0.001)	(0.001)
Household members	-0.034***	-0.004
	(0.012)	(0.009)
Muslim	0.028	0.033
	(0.055)	(0.070)
General caste	0.027	0.055
	(0.084)	(0.078)
Asset index	-0.313*	0.078
	(0.168)	(0.130)
Access to private toilet	0.207***	0.035
	(0.070)	(0.067)
Distance to CT (meters)	0.001***	-0.001***
	(0.000)	(0.000)
Awareness of externalities	0.008	-0.006
	(0.041)	(0.040)
Observations	836	663

Note. Sample restricted to the residents in the maintenance treatment group. Dependent variables by column: (1) *Stopped using CT*, indicator variable equal to 1 if used the CT at baseline and stopped using it at follow-up 4, and equal to 0 if continued using CT at follow-up 4; (2) *Reduced CT uses*, indicator variable equal to 1 if used the CT at baseline more frequently than at follow-up 4, and equal to 0 if continued using CT at same frequency at follow-up 4. Dependent variables are reported by respondents of the household survey. Sample restricted to catchment areas allocated to the maintenance treatment. Standard errors clustered at slum level are reported in parentheses. Statistical significance denoted by *** p<0.01, ** p<0.05, * p<0.1.

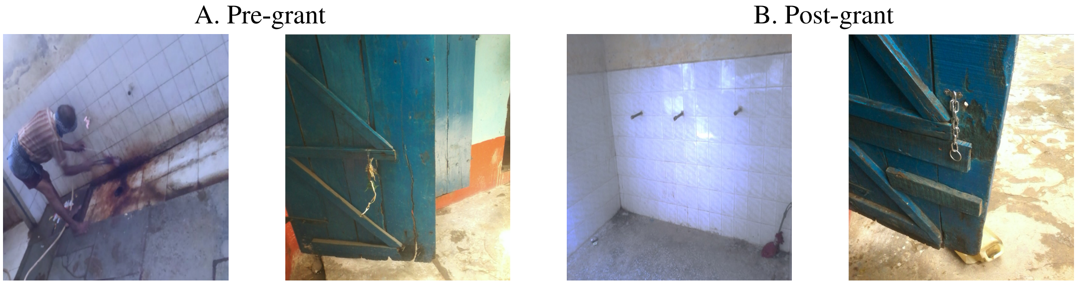
E Details about the interventions

E.1 Design

Maintenance intervention. The caretaker(s) could choose one of three *grant* packages of similar monetary value: *Deep cleaning* (septic tank sewage removal, unblocking latrines and sewerage pipes, and cleaning walls, floors and inside toilets); *Repairs* (sanitation/water connection repairs and/or infrastructure refurbishment); or *Cleaning tools and agents* (four pairs of gloves, five floor cleaners, four toilet disinfectants, five liquid soaps, four toilet brushes, two cloths, four nose masks, two brooms, two bucket and mop sets, three detergents, two hand wash dispensers, two dustpans and two dustbins, and training in their use). Photographs of the CT area to be improved were taken before the work was carried out. Our partner FINISH arranged and supervised the work with an external contractor, who was used in all facilities, and conducted the training through both theoretical and practical sessions.

In the *financial reward*, caretakers could receive: INR 500 conditional on the availability of soap in hand-washing facilities for both sexes; INR 500 conditional on the visible cleanliness of the latrines (whether the cubicles were free of visible feces both inside and outside the latrines); INR 1,000 conditional on the bacterial count being maintained at a minimum standard (i.e. 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 the measures taken. In CTs with more than one caretaker (20% of the sample), the financial reward was shared. After two months and every two months, the conditions were checked by observers.

Figure E1: Examples of grant use



A. Pre-grant

B. Post-grant

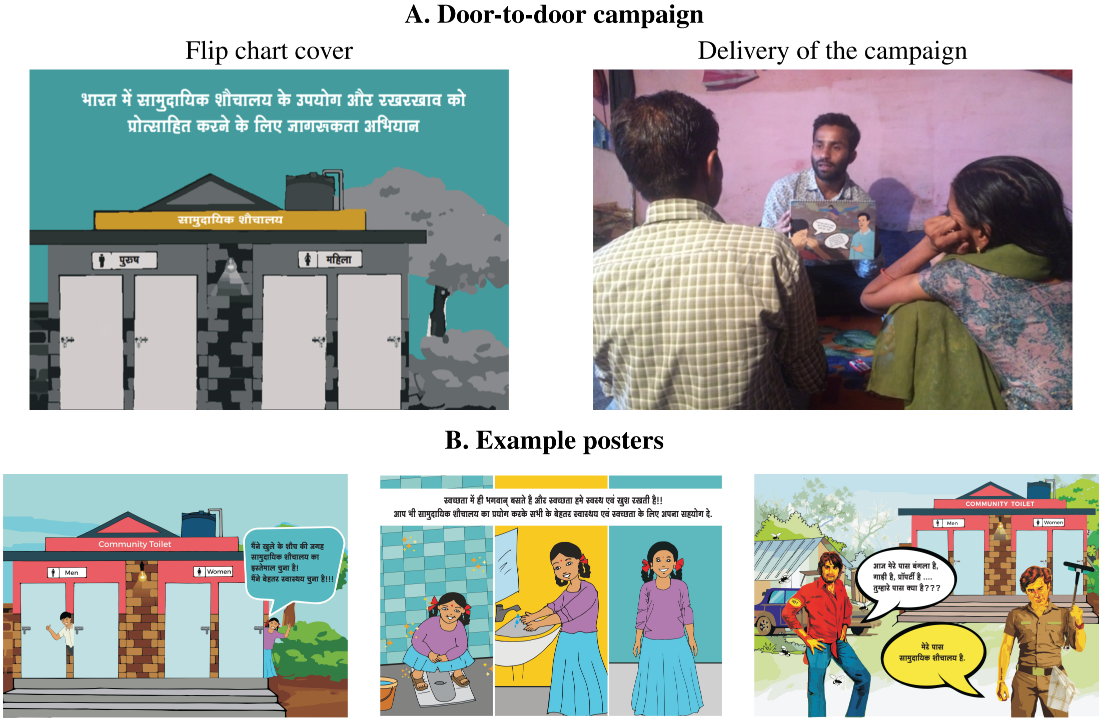
Note. Example of deep cleaning of walls and repair of locks in a CT in Lucknow. Panel A shows the status before the intervention, while panel B shows the status after the deep cleaning. Source: Antonella Bancalari.

Sensitization campaign. The campaign was designed in conjunction with our partner FINISH and a local graphic designer. All members of study households were targeted. We provided key messages regarding the risks of unsafe sanitation behavior through different means: *door-to-door visits* (Panel A, Figure E2), using a flip chart; a four-page summary *leaflet* , and a series of *posters* (Panel B, Figure E2) placed outside and inside the CTs; and ten *voice messages* sent to study households between study months 1 and 11. Households listened to an average of 7 messages. All study participants received a message about opening hours; those in the maintenance treatment group were informed about grant-funded improvements; and those in the maintenance plus sensitization group received information provided during the sensitization campaign, such as the negative health externalities of OD.

E.2 Cost of interventions and quality scenarios

Table E1 summarizes the total project costs under the maintenance (Panel A) and sensitization (Panel B) interventions. Drawing on input from our implementing partner, FINISH Society, and the Lucknow Municipal Corporation, Table E2 presents estimated operation and maintenance (O&M) costs for the median community toilet (CT) in our study sample (built 20 years ago, with four female WCs, six male WCs, two urinals, and approximately 150 daily users). A household can generate a potential monthly expenditure of at least INR 600 if all members over the age of five use the CT once daily and pay the INR 5 fee. The monthly maintenance cost for the 'status quo' scenario is INR 10,200 (US$144.85). Salaries account for 78% of the total budget and cover the cost of a full-time caretaker and a cleaner for daily routine cleaning. We consider an alternative scenario that supports an 'improved' level of maintenance, assuming that the number of users remains constant. This scenario includes a higher salary for a more experienced cleaner, higher input costs and an annual investment in cleaning equipment, such as a pressure washer, which costs approximately INR 20,000 (US$ 284.01). This scenario results in a total of INR 28,800 (US$ 408.97) per month, with salaries accounting for 63% of the total. We are not suggesting that this scenario is optimal. It can certainly be improved. The table also shows the cost per eligible household (see Appendix C for eligibility and proximity criteria), of which there are 34 in the median CT. In panel B of table E2, we convert the total costs of the maintenance intervention (table E1) into monthly costs. Adding these costs gives a total monthly cost of INR 13,544 (US$ 192.33) per CT.

Figure E2: Sensitization campaign materials



A. Door-to-door campaign

Flip chart cover

Delivery of the campaign

B. Example posters

भारत में सामुदायिक शौचालय के उपयोग और रखरखाव को प्रोत्साहित करने के लिए जागरूकता अभियान
सामुदायिक शौचालय
पुरुष
महिला
Community Toilet
Men
Women
मैंने खुले के शौच की जगह सामुदायिक शौचालय का इस्तेमाल चुना है! मैंने बेहतर स्वास्थय चुना है!!!
स्वच्छता में ही भगवान् बसते है और स्वच्छता हमे स्वस्थ एवं खुश रखती है!!
आप भी सामुदायिक शौचालय का प्रयोग करके सभी के बेहतर स्वास्थय एवं स्वच्छता के लिए अपना सहयोग दे.
COMMUNITY TOILET
Men
Women
आज मेरे पास बंगला है, गाड़ी है, प्रॉपर्टी है .... तुम्हारे पास क्या है???
मेरे पास सामुदायिक शौचालय है.

Note. The left picture in 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"; the right picture shows a moment of the sensitization campaign (source: Morsel). Panel B shows three of five posters placed on the walls of CTs read in Hindi: 'I choose to always defecate in CTs, I choose better health' (left); 'Health is happiness and cleanliness is godliness. Do your bit by using CTs' (middle); the right one 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'.

Table E1: 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.

Table E2: 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% of total management cost to the maintenance grant implementation and 50% 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.

F Measurements

Bacteria presence. We gathered bacteria data analysed in a laboratory: 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); genus *Bacillus*; genus *Staphylococcus*, genus *Klebsiella*; and genus *Salmonella*. 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 whether swabs were collected in male or female areasDuring each visit, the enumerator then collected three samples: one from the floor to enter the cubicle hallway, and two 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. We further collected up to two water samples per catchment area from randomly selected water sources frequently used as reported by residents during the baseline survey.

List randomization. The questionnaire for follow-up 4 was supplemented with a list randomization technique. Respondents were randomly allocated to one of four groups and received a list of statements (Table F3), and were asked to report how many of them were true. Group A received a list of statements related to general behavior. Groups B–D received the same list and one extra statement capturing sensitive behavior.[3]

Table F3: Statements used for list randomization

Group A	Group B	Group C	Group D
- I cooked yesterday - I bought milk yesterday - I watched TV yesterday	- I cooked yesterday - I bought milk yesterday - I watched TV yesterday - I defecated in the open yesterday	- I cooked yesterday - I bought milk yesterday - I watched TV yesterday - I used the CT to defecate yesterday	- I cooked yesterday - I bought milk yesterday - I watched TV 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).

WTP for service use. WTP for service use is elicited from the respondent of the household survey and his/her spouse 4 times during the study (in conjunction with the household survey) 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. In total, participants face 13 combinations. After making all the choices, one of the options was randomly selected by drawing a numbered ball from a bag, and the decisions are realized. Before participating in the game, the participant was introduced to a practice round using a bar of soap to facilitate familiarity with the game. The wording was:

> *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*

[3] Self-reported sanitation behavior was measured by asking survey respondents where each demographic group defecated the last two times. To prevent under-reporting of OD 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 CTs defecate in the open.”

> *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.*

Incentivized WTP was supplemented by an hypothetical question about a higher-quality CT. We capture this alternative measure by asking: “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 that it was very clean, had good hand-washing facilities, and that all cubicles had a light and a lock. Would you be willing to buy a ticket, if the price was [. . .] INR?”

Adapted dictator games. To measure the preference for maintenance among residents, we played an adapted dictator game in which participants are endowed with INR 50 and 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 wording was:

> *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 prosocial 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. Prosocial 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 wording was:

> *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?*
>
> *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.*

Appendix Bibliography

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.

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.

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.

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

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, WHO.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.