Public Service Delivery, Exclusion and Externalities: Theory and Experimental Evidence from India*

Alex Armand Britta Augsburg Antonella Bancalari Maitreesh Ghatak

November 11, 2023

Abstract

This study explores the interaction between the quality of public services, the implementation of user fees, and the resulting potential for exclusion, that can lead to negative externalities. Our theoretical framework takes account of the possible externalities that result from excluded users accessing alternative options in the context of sanitation, i.e., open defecation, and challenges the conventional wisdom that higher quality unequivocally leads to increased use. Instead, it highlights the ambiguity that results from a simultaneous increase in usage due to improved services (quality effect) and a decrease caused by the fees (price-elasticity effect). We then provide empirical evidence from a randomized controlled trial, where we incentivized the quality of water and sanitation services in the two largest cities of Uttar Pradesh, India. We show that higher service quality increases fee compliance but excludes some users, leading to unintended negative health externalities. Our detailed data provides evidence that results are driven by changes in caretaker behaviour. This finding highlights the need to be cautious regarding user fees, especially for public services involving significant externalities, and in settings where the users are very poor. (*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, 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, Fernanda Estevan, 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, 6$^{\text{th}}$ Zurich Conference on Public Finance in Developing Countries, CEPR WE_ARE, and LISER for helpful comments. We would like to thank Tom Glinnan, Kerry Neitzel and Garima Rastogi for outstanding research assistance. This project would not have been possible without the insightful support of Abhijit Banerji, Sneha Santwana, Bhartendu Trivedi, and the teams at FINISH Society and Morsel Research and Development. We are also indebted to study participants for their willingness to contribute to the study. We gratefully acknowledge financial support from 3ie – International Initiative for Impact Evaluation (grant no. DPW1-1105), and from the ESRC's Centre for the Microeconomic Analysis of Public Policy at the IFS (ES/T014334/1). Bancalari's work was supported by the 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 do not follow the order presented in the plan, all results are present either in the main text or in the appendix.

1 Introduction

Public service delivery is of profound importance to development, with governments and international organizations consistently emphasizing the need to improve their quality to boost use. This objective is in fact included in eight out of the 17 Sustainable Development Goals (United Nations, 2020). However, delivering higher quality services comes at a cost. Given the reality of budgetary constraints, user fees are therefore frequently advocated as a key element for funding public services, particularly in low- and middle-income countries (L&MICs) (World Bank, 2004).[1] This endorsement persists despite a limited understanding of the trade-offs between cost-effectiveness, allocative efficiency, equity, and quality of the service when user fees are used (Gertler et al., 1987; Dupas, 2014a). Further, the commonly held assumption that improved services automatically lead to increased usage may not be valid in the context of user fees, especially when improving quality leads to greater monitoring of payment. In this case, individuals who cannot or choose not to pay may try to use the service without paying or may face exclusion. Other than being undesirable from an equity point of view, exclusion from essential services can result in overall negative externalities from the alternative options excluded users use (e.g., open defecation in the context of sanitation). This may not only undermine the positive effects of enhanced quality among those who use the services, but also dampen the ability of entire communities to escape poverty (Stavins, 2011; Greenstone and Jack, 2015).[2] Such unintended consequences are often overlooked in the literature, leaving the implications of improving public services ambiguous (Dupas, 2014a).

In this paper, we offer new theoretical and empirical insights into the interplay between the quality of public services, the implementation of user fees, the potential for exclusion, and the resulting externalities. Our theoretical framework of public service provision, while simple, introduces several novel elements to provide insight into this interplay, and, crucially, takes account of the externalities that can arise from exclusion.[3] Empirically, we provide evidence from a large-scale randomized controlled trial (RCT) to assess the impact of an improvement in public service quality on usage of the service, fee compliance, and the overall externalities involved. The evidence we present challenges the conventional belief that higher service quality unequivocally leads to increased utilization, underscoring instead how enhancements in service quality in fee-funded models can lead to service exclusion and negative externalities. Our detailed data on both providers

[1] In L&MICs, user fees commonly finance essential services such as health and education (Bird, 2010).

[2] Examples of negative externalities include open defecation instead of sewerage systems, driving a car instead of using public transport, burning fuel instead of connecting to electrification networks, or disposing of garbage in the environment rather than utilizing collection services. In these examples, there may be other outside options that do not involve negative externalities, such as cycling.

[3] In the context sanitation, Gautam (2023) focuses on a household's private sanitation choice and the resulting externalities, but does not consider issues relating to service providers and the interaction between suppliers and users.

and users allows us to shed light on the underlying mechanisms, namely fee enforcement.

Our theoretical framework considers a public good that is excludable. These services encompass a wide range of public services, where individuals are heterogeneous in the valuation of the public service, the ability and the willingness to pay user fees, and can choose between accessing the public service for a fee or opt for an outside option that generates negative externalities. Our focus is not the optimal mechanism of provision, which is the main focus of the theoretical literature (see, for example, Norman, 2004; Hellwig, 2005; Gravel and Poitevin, 2019). Rather, we assume the implementation of user fees, and characterize the conditions that determine the demand response to enhanced service quality. Our analysis points to four key effects. First, the quality effect, where improved services naturally attract a larger user base. Second, the price-elasticity effect, whereby greater monitoring of fee-payment leads to reduced demand regardless of users' income levels. Third, the income effect, impacting those financially unable to pay for the service. Finally, the externality effect, which raises the relative cost of the outside option as more individuals opt for it.

We show that, holding income and externality effects constant, the impact of improving public services on the user base depends on the relative sizes of the price-elasticity effect and the quality effect. In particular, the user base decreases if the price-elasticity effect dominates the quality effect, and increases otherwise. The income effect can reinforce a negative price-elasticity effect, and it is possible for the user base to decrease even if the quality effect dominates the price-elasticity effect. In our framework, the externality effect reduces the attractiveness of the outside option, mitigating a negative price-elasticity effect. Its presence requires considering outcome measures beyond the user base to capture the effects that arise from user exclusion.

Empirical evidence on these effects is based on an experiment that exogenously shifts the quality of a basic service in the two major cities of Uttar Pradesh (UP), India's largest state (Government of India, 2011). Our experiment was implemented in partnership with governmental and non-governmental organizations (NGOs) and revolves around community toilets (CTs). Present in many L&MICs, these public services offer essential access to hygiene and sanitation through communal facilities. They primarily cater to informal settlements (or *slums*), where overcrowding, limited space and inadequate housing constrain access to safely managed private toilets. In these areas, and unlike public toilets, CTs primarily serve a specific group of residents lacking access to private toilets. In India, these services are prevalent nationwide and operate with user fees. Facilities are poorly maintained and dirty. Moreover, non-payment among users is common and there is a general low willingness to pay (WTP) for the service (National Geographic, 2017). Nevertheless, this public service frequently represents the sole viable alternative to the outside option of using unimproved facilities or resorting to open defecation (OD). These practices impose significant health externalities (Coffey et al., 2018; Geruso and Spears, 2018), often hindering human and economic development (Miguel and Kremer, 2004; Bleakley, 2007; Adukia, 2017; Lipscomb and Mobarak, 2017; Augsburg and Rodríguez-Lesmes, 2018; Orgill-Meyer and Pattanayak, 2020; Spears, 2020).

The design of our RCT is as follows. Subsequent to an extensive mapping of the universe of CTs and the slums they serve in the two cities, we randomly allocated 70 of the 110 CTs to a treatment group, while the remaining 40 served as the control group. The intervention in the treatment group, labelled as *maintenance* treatment. It targeted the individuals in charge of service delivery and fee collection in the facility, referred to as the *caretaker*. It stimulated service quality by providing a one-off grant to rehabilitate the facilities in the first two months, and a large bimonthly financial reward (roughly 40% of the caretaker's monthly salary) to keep the facilities clean in the following 10 months. The primary objective of our intervention was to enhance service quality. It is important to clarify that 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 spanning a period of 18 months, we collected objective measures of the service's quality, the number of users, and the prevalence of payment among them. To map intervention effects into behavioral responses, we complemented objective measures with panel survey data and behavioral measures for both caretakers and slum residents.

We provide evidence demonstrating that the maintenance treatment leads to sustained improvements in service quality. However, caretakers undergoing this treatment not only improve their maintenance efforts, but they also devote a larger share of their time to monitoring activities that include fee collection (8.5% larger than the control group). As a result, the maintenance treatment increases by 16.7% the proportion of users who pay the fee compared to the control group, accompanied by a notable reduction in the frequency of service usage by residents.

This pattern is not driven by an increased willingness to pay for the service among residents, or a shift their ability to pay. These results are therefore consistent with a mechanism in which, net of the externality effect, the price-elasticity effect dominates the quality effect, leading to the exclusion of some residents from using the service. Residents' adaptation to the outside option aligns with this mechanism. At the time of the endline survey, the average share of respondents who practiced OD the day before the interview was 39.1% in the maintenance treatment group, compared to 21.0% in the control group. The negative externalities stemming from increased OD are confirmed by an increase in reported health issues in treated areas. Residing near a facility in the maintenance treatment group increases the self-reported incidence of having curative expenditure. The lack of an effect on the intensive margin of these expenditures points towards an increase in the incidence of diseases that are not costly to treat, such as infectious diseases linked to high prevalence of OD. This significant finding underscores the potential unintended consequences of program design, operating through externalities, which may not be adequately captured by a simple cost–benefit analysis.

During the maintenance intervention, we also introduced a sensitization campaign to test whether these mechanisms are amplified or reduced by shifting the valuation of the outside option. This campaign targeted slum residents living in a randomly selected half of the catchment areas assigned to the maintenance treatment. The campaign was designed to raise awareness about the importance of adopting safe sanitation practices. Despite the presence of various governmental and non-governmental initiatives ongoing at the time of the intervention, notably the Government of India's flagship *Swachh Bharat Mission (SBM)*, the sensitization campaign further raised awareness of the negative health externalities of not using the service.[4] However, we do not observe any effect of sensitization on the behavior of residents or caretakers.

Our results provide a fresh perspective to the literature on public service delivery in general, and specifically in the context of public health and sanitation, whose importance and policy relevance cannot be exaggerated. Improving access to water, sanitation, and hygiene (WASH) services has large benefits in terms of reduced health externalities. In 2020, an estimated 3.6 billion people worldwide lacked access to safely managed sanitation services, with approximately half residing in urban areas. In India, only 37% of the 0.48 billion people living in urban areas had such access (WHO, 2021). Enhancing these services in urban areas is crucial, as the rapid urbanization is causing a surge in demand for essential services, surpassing existing capacities (Bryan et al., 2020). Notably, slums present one of the most challenging environments because non-payment is expected to be pervasive and free-to-use services are in dire conditions (Marx et al., 2013). In such contexts, fees play a critical role in sustaining the viability of service delivery.

We contribute to several literatures spanning development, public service delivery, health and sanitation, and also, state capacity.

Firstly, we shed light on the mechanisms driving quality of public services, challenging the belief that enhanced service quality leads to greater use. In a model funded by user fees, incentives to boost service quality can indirectly stimulate user exclusion, and thus have similar consequences to raising user fees. Higher user fees have been shown to limit access in a variety of settings, including education (Fafchamps and Minten, 2007; Kremer and Holla, 2009; Lucas and Mbiti, 2012; Andrabi et al., 2020; Romero et al., 2020), health (Ito and Tanaka, 2018; Beuermann and

[4] The SBM is a government-led initiative aimed at promoting toilet construction and hygiene practices, and ending OD, through a mix of awareness creation and subsidy provision. Recent evidence suggests that SBM led to a substantial increase in toilet coverage in rural India (Chatterjee et al., 2023).

Pecha, 2020), and water and sanitation services (Szabo, 2015). We highlight the potential unintended consequences of this exclusion, as well as underlying mechanisms, leading to negative externalities.

Secondly, we complement existing evidence on the role of state capacity and bureaucracy for service delivery (Burgess et al., 2017; Rasul and Rogger, 2018; Bandiera et al., 2021; Akhtari et al., 2022; Fenizia, 2022; Best et al., 2023). Top-down incentives enhancing the performance of local providers are effective at raising the quality of service delivery, in line with the evidence on incentives for pro-socially motivated jobs (Besley and Ghatak, 2018).[5] Our findings underscore the significance of local providers not only in shaping the quality of public services, but also in guaranteeing universal access. Furthermore, our contribution extends to the understanding of the constraints providers face when maintaining public infrastructure, a topic often overlooked in the literature (Duflo et al., 2012).[6]

Thirdly, we contribute to the literature studying the causes for the underprovision of basic services in L&MICs. As Burgess et al. (2020) posit, funding electricity services with user fees in poor institutional settings is associated with a high prevalence of non-payment and, consequently, service rationing. 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 tools like disconnection threats (Jack and Smith, 2015; Coville et al., 2020; Jack and Smith, 2020; Rockenbach et al., 2023). Adding to this stream of literature, our findings indicate that top-down incentives for quality improvements increase fee payment, with limited additional effects observed from a sensitization campaign.

Finally, our study complements the literature on tax collection in L&MICs since user fee collection has some similarities with tax collection and the "exit" option in the case of taxes is the informal economy. An emerging body of research has shed light on the significant challenges associated with expanding tax bases in these countries (Besley and Persson, 2013; Pomeranz and Vila-Belda, 2019), with a particular focus on property taxes (Weigel, 2020; Balan et al., 2022). Notably, while user fees often constitute the largest portion of the overall tax burden in L&MICs (see, for example, Paler et al., 2017), there is relatively little knowledge about the mechanisms of fee collection. Our results provide novel evidence on the incentives of fee collectors, which closely relates to tax collection. While Khan et al. (2016) show that performance-pay among tax collectors can raise tax revenues, our study reveals that incentives aimed at raising the quality of

[5] Banerjee et al. (2008) provides the theoretical foundations for top-down incentives for the management of public goods in poor institutional settings, where coordination among citizens fails.

[6] A vast literature focuses primarily on the effects of expanding the infrastructure, rather than maintaining it. For L&MICs and specific to the water and sanitation infrastructure, refer to Devoto et al. (2012), Meeks (2017), Alsan and Goldin (2019), and Bancalari (2020).

service delivery can also indirectly stimulate collection.

2 Theoretical framework

We are interested in the decision-making process of a community of residents as to whether or not to access a service. Their payoff for using the service depends on its quality, y, which in turn is determined by the provider's maintenance effort, x.[7] For simplicity, we set x to be a discrete binary choice $x \in \{0, 1\}$, with the relationship between quality and maintenance effort given by $y = g(x)$, with $y_1 = g(1) > y_0 = g(0)$.

Accessing the service costs a fixed user fee equal to c. The service provider collects these fees using effort e, which we again assume to be a binary variable $e \in \{0, 1\}$, where 0 indicates no effort. Because we are interested in comparing alternative combinations of x and e, we do not model the incentives of the provider in allocating these two types of efforts, and the possible inter-dependence among them, and effectively assume that x and e are exogenous.[8]

We now turn to users. Users who are able to pay the fee can decide whether to pay, a decision denoted by $\delta \in \{0, 1\}$, where 0 indicates the decision not to pay the fee. We assume that δ depends on a user's type, and on the fee-collection effort e. We consider three types of users indicated by $\tau \in \{1, 2, 3\}$ which, following the terminology of Tirole (1996), we call: (1) the honest type, who always pays the fee; (2) the dishonest type, who never pays the fee; and (3) the opportunist type, who pays the fee if the service provider chooses $e = 1$ and does not pay the fee otherwise. We set the fraction of type 1 (honest) users to be α, the fraction of type 2 (dishonest) users to be β, and the fraction of type 3 users (opportunists) to be $1 - \alpha - \beta$.

We include the possibility that some users are not able to pay the fee even when they are willing. We model this possibility in the form of an income shock. Let $\phi \in \{0, 1\}$ denote whether someone has not or has been subject to an income shock. With probability p, $\phi = 0$ and users value the cost of using the service at the actual fee, c, and, with probability $1 - p$, $\phi = 1$, meaning they experience a budget tightening shock that turns the cost of having to pay the fee to λc, where $\lambda > 1$ is sufficiently large to make the outside option always preferable. We assume that this shock is independent of θ, type τ, or any other aspects of the environment (e.g., the quality or fee-collection efforts) and happens at the beginning of the period. This does not have to be interpreted literally as a shock, but as another fixed characteristic of users and as the form in which we introduce an income effect to the model that allows us to distinguish between willingness to pay and ability

[7] We do not consider infrastructural investments because our focus is on maintenance and quality-enhancement of existing infrastructure.

[8] In the framework of a multitasking model, x and e are not necessarily substitutes, but can move in the same direction. For instance, it is likely that providers, who need to collect revenues to cover operation and maintenance of the service, impose monitoring to exclude users not willing to pay since they increase costs at no economic benefit.

to pay. The fraction of honest types who pays is therefore αp, while the fraction of opportunist types when induced to pay is $(1 - \alpha - \beta)p$. Dishonest types never pay the user fee, and therefore their fraction is β.

We refer to the alternative to using the service as the outside option. In the empirical part of our study, the outside option to sanitation services is represented by OD. We denote the share of the community's residents using the service by r, and the share using the outside option by $1 - r$. Using the outside option generates a fixed payoff of θv and a disutility $\gamma(1 - r)$, which captures the negative externalities associated with it. The parameter θ captures residents' heterogeneous valuation of the outside option, such as tastes, social norms, or socioeconomic characteristics of users, and v is a scaling factor. We assume that θ is distributed with a density function $f(\theta)$ and cumulative distribution function $F(\theta)$, and, without loss of generality, we normalize v to one.

The parameter γ captures the magnitude of the negative externality, which depends on the share of residents using the outside option. Note that using the service too could involve some costs due to externalities and so we can interpret γ as the *net* marginal externality cost to an individual of having more users using a particular method. To the extent that the externality from users of a particular method (say, the outside option) affects all users through overall externalities, then that would be an additive term for either method and would cancel out when figuring out the individual's decision-making. However, that would matter for welfare and individual impacts.[9]

2.1 Individual decision-making

Heterogeneity among residents originates from their valuation of the outside option, their propensity to pay the fee for the public service, and their ability to pay. For simplicity, we assume that these characteristics are orthogonal. The choice of an individual of type (τ, θ, ϕ) is then to decide whether to use the public service or not, and if they choose to use it, whether to pay the fee. An individual will use the public service if

$$u(y) - \delta(\tau, e)\,[p\,c + (1 - p)\lambda\,c] \geq \theta - \gamma(1 - r). \tag{1}$$

The marginal service user (i.e., the user who is indifferent between using the service and the outside option) is defined by their type (τ, θ). Type 1 (honest) users will use and pay for the

[9] All users use services more than once a day, although in our model we formulate it as a choice that a given user type makes once. A simple way of allowing for the same user to make different choices would be to allow for a stochastic cost parameter ϵ to the payoff from using the service on a given day. This would be in addition to the negative income shock, which is best interpreted as an individual-specific shock that is not contingent on any particular use on a particular day. Person-use is then be the new unit of aggregation, and each individual can mix service use and the outside option within a day. To the extent that this is a zero-mean shock, our framework can be used directly, using the expressions as the expected payoff for a given user.

service when θ is equal or smaller than the indifference threshold $\theta_1(y, r)$ that guarantees that $u(y) - pc = \theta - \gamma(1 - r)$. Similarly, type 2 (dishonest) users will use the service when θ is equal or smaller than the indifference threshold $\theta_2(y, r)$ that guarantees that $u(y) = \theta - \gamma(1 - r)$, but will stop using the service, independent of θ, when the service provider monitors fee-payment. Finally, the indifference threshold for type 3 (opportunist) users, $\theta_3(y, r)$, depends on whether the service provider monitors fee-payment, and coincides with θ_1 when $e = 1$, and with θ_2 when $e = 0$.

Individual decision-making highlights three important results. First, offering poor-quality services for free does not eliminate the outside option unless the upper bound of the distribution of θ is extremely low for all types (e.g., strong aversion to the outside option due to social norms or awareness). In other words, if the utility obtained from using the service is low, we will always observe individuals reverting to the outside option, thus generating negative externalities. Second, because users value higher-quality services more, we know that $\theta_1(y_0, r) < \theta_1(y_1, r)$ and $\theta_2(y_0, r) < \theta_2(y_1, r)$. In other words, holding everything else constant, a higher quality of service provision will attract more users. Third, holding everything else constant, the valuation threshold θ is higher for those not paying the fee, i.e., $\theta_1(y, r) < \theta_2(y, r)$. That is, when there is no monitoring of fee-payment, we will always have larger shares of service users among dishonest and opportunistic users as compared to honest users because of non-payment.

2.2 Equilibrium

Proposition 1. *We define π as the fraction of fee-payers, with $\pi = \alpha p$ when $e = 0$ and $\pi = \alpha < 1 - \beta$ when $e = 1$. For given levels of y and $\pi \in (0, 1)$, there is a unique r^* that satisfies:*

$$r = \pi F(\theta_1(y, r)) + (1 - \pi) F(\theta_2(y, r)) \equiv g(r). \tag{2}$$

Proof. Because $\theta_1(y, r)$ and $\theta_2(y, r)$ are decreasing in r, and $F(\theta)$ is decreasing in θ, $g(r)$ is strictly decreasing in r and the end-points for $r = 0$ and $r = 1$ satisfy $g(0) > 0$ and $g(1) < 1$. By continuity, there is a unique r^* that satisfies $r = g(r)$. We present the argument graphically in Figure 1, in which the thick curve in the middle cuts the 45-degree line. □

Equation (2) helps us to highlight four forces driving service use in presence of user fees. First, a **quality effect**. When y increases, both $\theta_1(y, r)$ and $\theta_2(y, r)$ increase, and therefore, $g(r)$ goes up for all values of r. If there is a quality improvement of the service, then more users of all types use the service. This effect would tend to increase the equilibrium value of r.

Second, a **price-elasticity effect**. When the fee-collection effort increases, for the same level of y, π increases from αp to $(1 - \beta) p$. This change would decrease $g(r)$ for all values of r as

$\theta_1(y, r) < \theta_2(y, r)$. Intuitively, if all the opportunist types have to pay fees, some of them will drop out of the service and use the outside option instead. A higher fee-collection effort would therefore tend to reduce the equilibrium value of r.

Third, an **income effect**. Since those who cannot pay the fee will not use the service if they have to pay the fee, a decrease in p would reduce the use of the service. Negative income shocks would therefore have a tendency to reduce the equilibrium value of r. Empirically, if we were to give people a cash transfer that on average compensates them for the user fee when they use the service, then this effect would not matter.

Fourth, an **externality effect**. If r rises, the outside option becomes more attractive, leading to downward shifts in $\theta_1(y_1, r)$ and $\theta_2(y_1, r)$. However, if r falls, the outside option becomes less attractive and $\theta_1(y_1, r)$ and $\theta_2(y_1, r)$ shift upwards. In both cases, there is a negative feedback loop. In general, if the price-elasticity and income effects dominate in the neighborhood of the initial equilibrium, then the externality effect would temper the initial negative effect on service use by making the outside option less attractive. However, if the quality effect dominates around the initial equilibrium and r goes up, then the externality effect would dampen this effect by making the outside option more attractive.[10]

Because the above four effects are at work, $g(r)$ may not shift uniformly up or down when the quality of service provision increases. Therefore, the equilibrium value of r may be higher or lower depending on which effects dominate in the middle. To understand this ambiguous effect, we proceed to characterize and compare the equilibrium across two extreme scenarios: a scenario C, in which the service provider exerts no maintenance and fee-collection efforts, i.e., $x = 0$ and $e = 0$, and a scenario T, in which the service provider exerts full effort, i.e., $x = 1$ and $e = 1$.

There are different reasons why we want to consider a scenario in which both maintenance and fee-collection efforts increase. First, fees are often needed to fund maintenance, and thus the service provider who wants to improve service quality might also want to increase revenue collection. Second, the service provider may want to avoid congestion or actively exclude users who disrespect the public good, because they make maintenance costlier. Finally, in the long run, a better-maintained infrastructure can generate more time for other activities, such as fee collection.

The proposition below characterizes our main result in terms of the outcome from moving from

[10]We assume a linear externality term. However, if this term is nonlinear, it could introduce complexity. For instance, if the disutility of the outside option increases in $(1 - r)$ at a faster rate as more people use it, even slight quality improvements could lead to a significant increase in adoption. Conversely, if the disutility initially rises slowly but then sharply as a threshold is crossed, substantial quality enhancements may be necessary to raise adoption rates. In our empirical context, evidence from Cameron et al. (2022) suggests that health benefits from sanitation become noticeable only beyond a certain latrine coverage threshold. If health externalities affect various sanitation methods similarly, they might offset each other in individual decision-making. In our model, externalities apply exclusively to alternative service users, and the shape of the externality cost function dictates the extent to which service usage replaces the outside option.

scenario C to scenario T:

Proposition 2. *Assume first that there is no income effect, i.e., $p = 1$. Transitioning from scenario C to scenario T leads to an increase in r if the utility gain from the improved service, $u(y_1) - u(y_0) \equiv \Delta u$, is larger than the fee c. In this case, the quality effect outweighs the price-elasticity effect. If Δu is smaller than the fee c, transitioning from scenario C to scenario T leads to a reduction in r. If there is a income effect, then there exists a critical value $\underline{p} > 0$ such that, for $p \leq \underline{p}$, the number of users of the public service in scenario T could go down as compared to the scenario C even if $\Delta u > c$. In this case, the sum of the price-elasticity and the income effects outweighs the quality effect.*

Proof. Assume first that all users are able to pay the fee, i.e., $p = 1$. In scenario C, the relevant indifference thresholds of θ are equal to $\theta_1^C = u(y_0) + \gamma(1-r) - c$ for type 1 users and to $\theta_2^C = u(y_0) + \gamma(1-r)$ for type 2 users. In scenario T, these indifference thresholds are instead equal to $\theta_1^T = u(y_1) + \gamma(1-r) - c$ and $\theta_2^T = u(y_1) + \gamma(1-r)$. For a given value of r, the difference in indifference thresholds for type 1 users in both scenarios, i.e., $\theta_1^T - \theta_1^C$ is equal to the same difference for type 2 users, i.e., $\theta_2^T - \theta_2^C$, and corresponds to the utility gain resulting from an improved service, i.e., $u(y_1) - u(y_0) \equiv \Delta u$. Similarly, the difference in indifference thresholds across types in a specific scenario is the same across scenarios, i.e., $\theta_2^C - \theta_1^C = \theta_2^T - \theta_1^T$, and equal to c.

The total demand for the service in scenario C is therefore equal to

$$r_C = \alpha F\left(\theta_1^C\right) + (1-\alpha) F\left(\theta_1^C + c\right), \tag{1}$$

and the total demand for the service in the scenario T is instead equal to

$$r_T = (1-\beta) F\left(\theta_1^C + \Delta u\right) + \beta F\left(\theta_1^C + \Delta u + c\right). \tag{2}$$

If the increase in utility derived by an improved quality outweighs the cost of using the public service in the payoff of users, i.e., $\Delta u > pc$, we therefore obtain the following ranking:

$$F\left(\theta_1^C\right) < F\left(\theta_1^C + c\right) < F\left(\theta_1^C + \Delta u\right) < F\left(\theta_1^C + \Delta u + c\right).$$

Because any weighted average of $F\left(\theta_1^C\right)$ and $F\left(\theta_1^C + c\right)$ is strictly smaller than any weighted average of $F\left(\theta_1^C + \Delta u\right)$ and $F\left(\theta_1^C + \Delta u + c\right)$, it follows that r_T is strictly larger than r_C. That is, the solid curve in Figure 1 shifts up uniformly (indicated by the dashed curve), and transitioning from scenario C to scenario T leads to an increase in r. The opposite is true when the cost of using the public service in the payoff of users outweighs the increase in utility derived by an improved quality.

Next, we show that if we take into account the income effect, the above result has to be modified. Take the extreme case of $p = 0$. Because $(1-\alpha) > \beta$, it is possible that

$$(1-\alpha) F\left(\theta_1^C\right) > \beta F\left(\theta_1^C + \Delta u\right) \quad (3)$$

even when $\Delta u > c$. If (3) holds, then there would be a critical value $\underline{p} > 0$ defined by

$$\alpha \underline{p} F\left(\theta_1^C\right) + (1-\alpha) F\left(\theta_1^C + c\right) = (1-\beta)\, \underline{p} F\left(\theta_1^C + \Delta u\right) + \beta F\left(\theta_1^C + \Delta u + c\right) \quad (4)$$

such that, for $p \leq \underline{p}$, the number of users of the public service in scenario T could go down. That is, the solid curve in Figure 1 shifts down (indicated by the dotted curve). □

This result highlights that whether improving the quality of public services increases use depends critically on what is happening with fees – their level, the payment monitoring system, and the distribution of the willingness and ability to pay in the user population. We provide here the intuition behind this result in the absence of income shocks. In scenario C, all honest types with $\theta \leq \theta_1(y_0, r)$ and all opportunist and dishonest types with $\theta \leq \theta_2(y_0, r)$ will use the service. Of all users, only the fraction of honest types who are not subject to the negative income shock will pay for the service.

Under scenario scenario T, the quality and the price-elasticity effects operate in opposite directions. When only service quality increases, keeping all else constant, all honest and opportunist types with $\theta \leq \theta_1(y_1, r)$ and all dishonest types with $\theta \leq \theta_2(y_1, r)$ will use the service. On the other hand, when only monitoring increases, keeping all else constant, all honest types, independent of their valuation, will continue to use the service, and pay for it. For opportunist types, usage will also remain constant, but a fraction $(1-\alpha-\beta)p$ will now pay. Since dishonest types never pay, now faced with $e = 1$, they will discontinue their use of the service. Overall, the fraction paying will increase from αp to $(1-\beta)p$, and the total demand will drop.

When both quality and monitoring increase, the overall results are ambiguous. For the honest and dishonest types, service use goes up since quality is better. However, among opportunist types, there is a possible drop. If $\Delta u > c$ then for opportunist types who were already using the service (the inframarginals), the payoff from using the service is higher even after paying fees because of the quality improvement. However, if $\Delta u < c$, then for some of the opportunist types who were using the service before any effort was exerted, and who are not subject to a negative income shock, the payoff from using the service is lower after paying fees, despite the quality improvement, and thus some of them will switch to the outside option.

We now turn to the empirical part of the paper. The theoretical framework developed above will serve as a guide for interpreting our main empirical results and will also direct us in conducting further analysis to verify its plausibility. These analyses will encompass aspects related to service provision, use and payment, and the broader health implications arising from externalities.

3 Background and interventions

Background. Our experiment is implemented in the slums of Lucknow and Kanpur, the capital and the second largest city, respectively, of the Indian state of UP. In 2015, Lucknow was the 129th largest city worldwide with 3.2 million inhabitants, and Kanpur was the 141st with 3.0 million inhabitants. Their populations are expected to grow by 2035 by 59% and 37%, respectively (United Nations, 2018).[11]

Our study cities exemplify the rapid pace of urbanization experienced by cities in L&MICs, often resulting in the proliferation of slums. In Lucknow and Kanpur, slum residents are 13% and 15% of the population, respectively, comparable to India's capital, Delhi (Government of India, 2011). Slums represent an extreme case of both lack of access to safe sanitation services and 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 (WHO, 2021). These unsanitary conditions are common in our study area, where more than 40% of slum residents lack access to private toilets (Government of India, 2011). Shared sanitation facilities remain a common solution for the foreseeable future in the slums of L&MICs.[12]

When providing services to residents, these facilities are known as CTs. Arranged in gender-specific areas, they offer sanitation, hand-washing and bathing facilities. Services are generally rendered by a long-term public–private partnership funded by user fees, with each access costing a standard fee of 5 Indian rupees (INR, corresponding to US$ 0.07).[13] At market prices, an average household of four members living in a slum could spend up to 8% of their average household income on CT user fees if all adult members were to use the service.

Service delivery is performed by caretakers who are in charge of the daily operation and management (O&M) of the CT. These activites include maintenance (i.e., by cleaning the facility or supervising cleaners), and collecting user fees. Because CTs lack physical-access control technologies, the only way that fee-payment is monitored is by the presence of the caretaker at the

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

[12]Under the urban component of the SBM, toilets are envisioned for 80% of urban households engaging in OD. The remaining share is assumed to be catered by CTs due to space constraints (Government of India, 2017).

[13]There is almost no variation of the fee across the two cities. Nominal INR are converted to nominal US$ using the 2019 average exchange rate of US$1 = INR 70.42 (International Monetary Fund, 2020).

entrance. Higher-level managers regularly gather the fees collected by caretakers and distribute cleaning agents and tools. Caretakers are hired centrally and receive a fixed salary. The salary is, on average, equal to INR 5,000 (US$71) per month and do not include any performance-based financial reward (Armand et al., 2020).

The quality of the service rendered is substandard. CTs are poorly maintained, as reflected by the low quality of the facility, and the limited availability of functioning hand-washing facilities. Free-to-use CTs are not as common, but also exist in the study area, though their quality is much worse (Appendix A).

Low quality is associated with low payment of user fees (panel A in Figure 2). On average, only 65% of users pay the fee in CTs (median 68%), and all users pay the fee only in 20% of CTs. Payment is only partially enforced by caretakers, with only 8% of residents reporting that they had been prevented from using the facility because they were unwilling to pay the fee. The distribution of payment measured at the CT level closely mirrors the distribution of residents who elicit a positive WTP to use the service (panels A and B, Figure 2). Using an incentivized measure of WTP (explained in Section 5.2), only 65% of residents indicated a non-zero WTP for a single CT use at baseline. WTP on the intensive margin remains very low (panel C, Figure 2). On average, WTP amounts to INR 1.40, which is a mere 28% of the official fee.

Maintenance intervention. With the purpose of boosting the quality of the service delivered by the CTs, we introduced a positive shock to the maintenance effort of caretakers. We initiated the intervention by pushing rehabilitation through a one-off grant. This grant was disbursed in the initial two months of the intervention, and it was offered directly to the caretaker. Depending on the specific requirements of the facility, 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).

We then introduced an incentive scheme to motivate caretakers in their cleaning efforts. Based on prior research findings, we chose an output-based absolute payment system with discrete rewards, emphasizing individual performance to mitigate the impact of social comparisons (e.g., Ashraf et al., 2014b). From the second to the twelfth month of the intervention, a bimonthly 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. Firstly, caretakers received INR 500 (US$7.10) for maintaining visible cleanliness of latrines.

Secondly, they received INR 500 (US$7.10) for ensuring the availability of soap in the hand-washing facilities. Lastly, caretakers who kept bacteria counts below a specified standard were rewarded INR 1,000 (US$14.20). We allocated a higher incentive for reducing pathogen exposure due to its significant health externality and its widespread presence (Appendix A). The specified standard is the baseline median value of *E. coli* bacteria count in the study CTs. Importantly, caretaker incentives did not relate to the amount of revenue they collected through user fees.

Caretakers received feedback on their past performance to gauge the effort required to meet the criteria (Bandiera et al., 2015), but payments were tied only to current performance to deter gaming over time (Bénabou and Tirole, 2003). In each round, the total potential incentive was thus INR 2,000 (US$28.40), roughly 40% of caretakers' 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 as compared to other interventions that showed effects on exerted effort.[14] We set high expected payments because, in the context of pro-social tasks, these have been shown to be more effective (Ashraf et al., 2014a).

Every two months and for a total of four times during the study, enumerators verified the conditions for the reward during random visits and delivered payments accordingly. Caretakers received on average INR 779 (US$11.06) in the first round of incentives, INR 1,036 (US$14.71) in the second round, INR 1,058 (US$ 15.02) in the third round, and INR 972 (US$13.80) in the last round. These amounts correspond to 39%, 52%, 53% and 49% of the potential reward, respectively.

Sensitization campaign. Separately from the maintenance intervention, we additionally implemented a sensitization campaign among residents. This intervention targeted only at a random sub-sample of the catchment areas of CTs in the maintenance treatment group. The campaign aimed to increase awareness of negative externalities resulting from OD, highlighting the importance of paying the fee to fund the services offered by CTs. The campaign was executed through different means. Firstly, door-to-door visits were conducted three times in April–June 2018, July–September 2018, and January–March 2019. Secondly, leaflets were distributed posters were put up summarizing the main messages. Lastly, voice message reminders were sent to study households' mobile phones on a monthly basis.

Implementation. We implemented interventions in partnership with Lucknow and Kanpur Municipal Corporations, Sulabh International, and the zone and city managers of the CTs. The interventions were implemented by FINISH Society, a Lucknow-based NGO. Appendix C provides the location and the timeline of activities, while Appendix E describes operational details, including the cost of interventions.

[14] 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.

4 Research design: sampling and randomization

The research design is an RCT with the treatment unit being a CT. Because a listing of CTs operating in the two cities was not available at the time of the experiment, we conducted in 2017 a census of all CTs in Lucknow and Kanpur. We gathered data on the facilities' location, their physical characteristics, their management, maintenance practices, and their users. These data formed the basis for selecting CTs operating with user fees and used mostly by residents, i.e., the most common model of public service delivery in slums. A total of 110 facilities were identified, 52 in Lucknow and 58 in Kanpur. Further details about sampling are provided in Appendix C.

To create exogenous variation in the quality of the service, each CT was randomly allocated 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. Using the rich census information, we built blocks using Mahalanobis-distance relative proximity, and randomly allocated each CT within a block to a treatment arm using a lottery with equal probability of assignment. To further minimize the risk of treatment contamination, CTs that are within 400 meters of each other were allocated to the same treatment arm. As a result, 40 CTs were allocated to the control group and 70 CTs were allocated to the maintenance treatment. In addition, to avoid contamination, while caretakers work only in one facility, we limited their rotation to different facilities in agreement with the service managers of each city.[15] We further cross-randomized the provision of the sensitization campaign to residents living in the proximity of CTs in the maintenance treatment group. Because this intervention was not implemented through the main unit of treatment, i.e., the CT, we discuss it in detail in Section 6.2.

5 Data

To obtain information on both the service provision and residents, we gathered a substantial amount of primary data. Appendix B provides definitions of the variables used in this study, and replicates the list of pre-registered outcomes, indicating the exhibits in which we present the corresponding results. Appendix F provides the scripts of behavioural measurements.

5.1 Data from CTs

Through unannounced visits, we collected information about service delivery using objective measurements. Independent observers collected information about the quality of facilities, including maintenance and cleanliness. Observers also collected samples from randomly selected

[15] During the study period, we did not observe rotation across the study CTs. Whenever a caretaker was replaced, the implementing team paid regular visits to inform the new caretakers about the intervention.

spots on the floor of facilities. These were then analyzed in a laboratory to identify the presence and counts of bacteria. On average, more than three types of hazardous bacteria, including *E. coli* and salmonella, were found in each facility in each round. Finally, observers documented use and payment at the facility entrance by recording the number of users and the number of those paying the fee. This count was performed for 1 hour between 5 and 7a.m., when most residents of the community use the facility (henceforth called *rush hour*). Observers relied on the expertise of caretakers to identify use and payment behavior separately for residents of the slums and for by-pass users.

These measurements were collected at baseline in April–June 2018, and in four waves of bi-monthly 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).[16]

Objective measurements of service delivery were supplemented with survey data, collected with the same timing. Surveys were administered with the caretakers. The questionnaire, implemented consistently across survey waves, covered management and maintenance practices in the provision of the service, as well as fee-collection efforts. Appendix Table D1 presents descriptive statistics of facilities and their caretakers at baseline. In 80% of facilities, the CT is operated by a single caretaker; caretakers are generally male (82%), have roughly 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 meeting managers.

Attrition was kept to a minimum between baseline and follow-up surveys. The average number of observations per facility equaled 3.9 and per caretaker equaled 3.8, with no differential attrition across treatment groups (Appendix D.1).[17]

5.2 Data from residents

We supplement data about service provision with information from the individuals using, or in need of using, the CT. Because CTs are serving slum residents, a volatile population by definition, we first had to build a standard sampling frame for this population of interest. Following the census of CTs described in Section 5.1, during the second half of 2017, we performed a geographical mapping of the slums surrounding each facility, followed by a census of all slum residents.

[16]To monitor the intervention's progress, we further implemented a mid-intervention measurement 2 months after the baseline, in July–September 2018, right after providing the one-off grant and before incentivizing the caretakers.

[17]For two CTs of the maintenance treatment group, another CT opened after baseline in their proximity. We allocated these CTs to the same treatment arm, implemented the interventions and surveyed them during follow-up rounds.

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. Based on this information, we narrowed the population of interest down to include slum residents who encounter the decision-making scenarios outlined in our theoretical framework (Section 2). This group, which we label as *residents* throughout the paper, comprises both current users and potential users of each CT. They are households living in the slum that expressed no intention of relocating from the census and where at least one member reported not using a private latrine for defecation. We further restricted this population to those residing in the catchment area of a CT.[18] 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 the sample of residents closely mirror those of the broader population of slum residents in India (Appendix C).

In conjunction with the baseline CT survey (Section 5.1), a baseline survey was administered to the sampled residents. The targeted respondent was the household's main decision-maker, in most cases the household head. The questionnaire covered the household's sociodemographic characteristics, the respondent's sanitation behavior, and family members' health status. Appendix Table D2 presents descriptive baseline statistics. On average, household heads are 45 years old, men, with primary education or less. The baseline health status of slum residents is poor and slums are a high-disease environment: almost 30% of households had a sick member and 60% faced out-of-pocket (OOP) expenditures in curative care (i.e., doctors' visits due to illness).

The sample of residents was revisited twice in the follow-up period. These surveys happened in conjunction with the follow-ups 2 and 4 of the CT survey.[19] Measurements from the household survey are analyzed at the household level (i.e., one observation per household).

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 the behavioral instruments are analyzed at the respondent level, using up to two observations per household.

The first measurement is the elicitation of residents' WTP for CT use. 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 market prices are well known (Andersen et al., 2006;

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

[19] 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 Appendix D.2.

Berry et al., 2020). We prompted 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, for example, 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. While the market value of 10 tickets is INR 50 (US$ 0.71), we offered different amounts of cash, ranging from INR 0 to 60 (US$ 0.85, above the current market price to deal with truncation) 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.[20]

The second measurement aims to capture accurately the prevalence of OD among residents. Because this is a sensitive behavior subject to recall bias, we used a list randomization technique (see, for example, 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 hold true, rater than which ones. We randomly allocated 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 proportion 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 individuals' types. Specifically, we use adapted dictator games to measure caretakers' pro-social motivation for the cause, and citizens' willingness to contribute to the cleanliness of the CT. Because we observe no treatment impacts along these dimensions, we discuss them in Appendix D.9 and as heterogeneity dimension (see Section 6).

On average, each study household was reinterviewed 1.65 times out of a possible 2 interviews, with 7.9% of baseline households not participating in any of the follow-up assessments. To minimize sample loss during revisits, we interviewed additional households that were randomly chosen from the baseline sampling frame. We observe no differential attrition across treatment groups and being a replacement household is orthogonal to treatment allocation (Appendix D.1).

[20]This measure is conditional on the quality of the closest CT. On average, residents are willing to pay above the market price of INR 5 when asked about a hypothetical higher-quality CT (see Appendix F).

6 Results

We use the data collected during the follow-up rounds to study behavioral responses of both service providers and residents to the maintenance intervention. Exploiting 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 CT/household/individual i in catchment area j at time t using the following specification:

$$Y_{ij} = \beta\, T_j + \alpha\, \mathbf{X}_{ij} + \epsilon_{ij}. \tag{5}$$

Here T_j is an indicator variable equal to 1 if the catchment area j received the *maintenance* intervention, and 0 otherwise. $\mathbf{X}_{ij}$ is a set of indicator variables capturing randomization strata. The error term ϵ_{ij} is assumed to be clustered by catchment area when the analysis is performed at the household or individual level. Results are robust to alternative assumptions about clustering, such as catchment area and data collection round when the outcome of analysis is at individual or household level, and by catchment area when the outcome of analysis is at catchment area level. Equation (5) estimates the impact of the *maintenance* treatment T throughout the study, following the pre-registered specification, but focusing on the overall impact of the *maintenance* treatment. Results using this specification are presented in Section 6.1. In Section 6.2, we present results using the pre-registered specification, which estimates separately the impact of providing the *maintenance* treatment in the CT 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 the multiple follow-up measurements to average out noise and increase power. In Appendix D.2, we show the treatment effects estimated for each survey round separately.

We provide evidence that supports the interpretation of β as the causal effect of the maintenance treatment. Randomization was successful in creating observationally equivalent groups in the experiment for both household and catchment area characteristics (Appendix Tables D1–D2), and for outcomes measured at baseline (Appendix D.2). We support main estimates with estimates using alternative specifications, including ANCOVA to control for potential imbalances at baseline, ordinary least squares (OLS) with inverse probability weights to correct for attrition (Appendix D.3), and using machine-learning approaches for the selection of control variables (Appendix D.4). Moreover, in addition to the features introduced in the design of the experiment and described in Section 4, we alleviate contamination concerns by showing evidence against spillover effects across treatment arms (Appendix D.5).

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 List et al. (2019) 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 indicated at the bottom of each table. Sections 6.1 and 6.2 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 Appendices D.4 and D.8, we discuss heterogeneous effects for all outcome variables using the causal forest procedure of Athey et al. (2019) and using sub-sample analysis by pre-specified dimensions. The estimated effects tend to be homogeneous across different heterogeneity dimensions.

6.1 Boosting quality of public services

The successful implementation of the *maintenance* intervention is reflected in significant differences in measures of exposure to the intervention across experimental arms (Appendix D.6). We use two indicators: the transfer to a CT and the transfer to a caretaker. The transfer to a CT includes the value of the initial grant (non-zero only during the mid-intervention survey) received by treated CTs 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, played in all treatment arms. The transfer to a caretaker comprises financial rewards provided in treated CTs and amounts retained by caretakers in each round of the adapted dictator game capturing caretaker' pro-social motivation for the cause, played in all treatment arms. Over the study period, the average transfer to a treated CT amounted to INR 25,270 (US$ 358.84), 16 times larger than the average transfer to a control CT. Similarly, caretakers in the control group received on average INR 373 (US$ 5.30), while caretakers in the treatment groups received an additional INR 4,179 (US$ 59.34).

Quality of the service

Table 1 presents estimates of treatment effects on CT quality (column 1) and inputs into service delivery (columns 2–4), focusing on indicators that proxy the main variables of our theoretical framework. Quality (y) is 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. Appendix D.7 details the construction of the index. Service delivery inputs are divided into the caretakers' maintenance (x) and fee-collection effort (e). Maintenance efforts include cleaning and rehabilitation. Cleaning is measured as an index that includes the number of tools, equipment 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. Rehabilitation is measured as an indicator variable equal to 1 if the facility received repairs and/or deep cleaning in the month previous to the visit, and 0 otherwise. We assess fee-collection effort with monitoring effort, measured by the caretaker's 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 conducted at the CT level.

We find that the maintenance intervention consistently improved the quality of service delivery. On average, the maintenance treatment leads to an increase of 6.4 percentage points in the quality index, 10.1% higher than the control mean, and this effect remains robust to multiple hypothesis testing with a p-value of 0.03. The maintenance treatment shifts the distribution of the quality of service delivery index, which is detectable mostly at higher levels of quality (Panel A of Figure 3). A Kolmogorov–Smirnov test of the equality of the distributions of the index 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 for the structural quality of the facility and the presence of harmful bacteria (Appendix D.7).

To gain insight into *how* quality of service delivery increases we estimate intervention impacts on inputs. We first show that maintenance effort increased. Cleaning performance improved significantly by 5.7 percentage points, 11.1% greater than the control mean, and this effect remains robust to multiple hypothesis testing with a p-value of 0.01. This effect is driven by improved inputs, the use of cleaners, and a correct implementation of cleaning procedures (Appendix D.7). Despite the significant transfers made to the CT at the start of the intervention, we observe an insignificant effect on rehabilitation during the follow-up period.[21]

We observe an increase in caretakers' effort allocated to monitoring activities that enable fee collection. Importantly, fee collection was not incentivized by the intervention. As a result of the maintenance treatment, caretakers spent a significantly larger share of their time on monitoring activities, which increased by 6 percentage points, 8.5% higher than the control mean. The p-value of this effect is 0.07, which adjusts to 0.15 when accounting for multiple hypothesis testing. This finding indicates that the caretaker's presence at the facility's payment point serves as an effective incentive for payment. On average, the treatment did not prompt caretakers to implement stricter payment enforcement, consistent with only 7.6% of study households in the control group reporting being denied entry due to non-payment of the service fee (Table D15). However, we do estimate a positive effect on enforcement in facilities where payment rates were initially low

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

(Appendix D.10).

We know from our model that, when both maintenance and fee-collection efforts shift, the effect on service utilization is theoretically ambiguous, especially because neither the intervention nor the measurements shifted the ability to pay the fee among residents, nor did they affect the net marginal externality cost.[22] As we expect both the quality and the price-elasticity effects to be at play, the following section analyzes how changes in quality manifest in terms of payment and use of the service.

Payment and use of the service

Table 2 turns to outcomes related to payment (π) and use (r) of the service. For both parameters of the theoretical model, we present effects on alternative measures. Column (1) shows the impact on the share of users paying the fee, again relying on data collected by independent observers during rush hour in the CT. Column (2) focuses on the share of residents surveyed who are willing to pay a positive amount, while column (3) focuses on the amount (in INR) residents are willing to pay (both incentivized measures as explained in Section 5.2).

Column (4) documents the effect on the total number of users, relying on data collected by independent observers during rush hour in the CT (Section 5.1), while columns (5) and (6) document impacts on the number of uses among residents, as reported by household heads when surveyed, distinguishing households that regularly use the CT and other residents. The remaining columns are about payment. The specifications in columns (1), (2) and (4) are at the CT level, columns (5) and (6) at the household level and column (3) at the respondent level.

The maintenance treatment leads to a significant increase in the share of users paying the user fee, with an average increase of 9.3 percentage points (representing a 16.7% increase over the control group mean). This estimate is robust to multiple hypothesis testing (the corrected p-value is 0.08). The maintenance 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 3). A Kolmogorov–Smirnov test of the equality of the distributions of the index in the control and maintenance treatment groups is rejected at the 1% confidence level. The positive effect on payment is greater among residents rather than by-passers (Appendix D.11).[23]

These effects are observed despite finding no change in WTP among residents. We find no significant effect on the share of residents who are willing to pay a positive amount for using the

[22] Appendix D.12 provides further evidence against the presence of an income effect induced by the WTP games.

[23] The reductions in users and increases in payment translate into a small positive effect on revenues, but the estimate is not statistically significant (Appendix D.10). 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.

CT, even though our baseline observations indicated that the distribution of this outcome closely mirrored the distribution of the share of total users paying the fee (Section 3). Furthermore, we find no significant effect on the WTP for using the service. 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 market price of INR 5. These results suggest that the rise in payment is primarily linked to the enhanced monitoring efforts of caretakers rather than a surge in WTP among residents. In addition, due to the small amounts, it is unlikely that the WTP measurement generated income or price-elasticity effects. Having received free tickets to use the CT as compared to cash does not impact use in the following survey round (Appendix D.12).

At the same time, we observe a reduction in usage. The maintenance treatment reduces the total number of CT users by on average 2 out of 34 users during rush hour, relative to the control group. This corresponds to 5.7% lower use as compared to the control group's average. While the estimate is not statistically significant across rounds, the negative effect on total use is precisely estimated in follow-ups 1 and 4 (Appendix D.2).

We find precise estimates when focusing on usage at the intensive margin as reported by residents. The maintenance intervention decreases the number of reported uses by 0.11 among residents who are regular CT users (8.0% higher than control mean) and by 0.19 percentage points among non-regular users (25.3% higher than control mean), both effects robust to multiple hypothesis testing (the corrected p-values are 0.06 and 0.10, respectively). These results are supported by data collected by observers at the CT level, which highlight that the effect on total users is concentrated among residents (Appendix D.11).

The results indicate that the benefits of improving quality do not outweigh the higher costs for users associated with greater monitoring efforts. The overall negative effect in usage suggests that, net of the externality effect, the price-elasticity effect was strong enough to offset the quality effect.

Outside option and externalities

Because we find that the the maintenance treatment decreased usage (Section 6.1), we study how it impacts the share of residents relying on the outside option in Table 3. In column (1) we show impacts on having practiced OD the day before the interview (i.e., the main outside option for the sample of residents). Because OD is a sensitive behavior and there is high awareness of negative externalities from OD among residents (66.0% of the control group and 69.1% of the treatment group are aware of this), we rely on the measurement captured by the list randomization technique detailed in Section 5.2. 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 respondents who claimed to practice OD the previous day by 17.2 percentage points, compared to a share of 21.0% in the control group. The p-value of this effect is 0.03, adjusted to be 0.22 when considering multiple hypothesis testing.[24] The coefficient on the number of occasions when other sanitation practices were used, using self-reported data, is also positive, but not statistically significant (Appendix D.11).

The observed increase in the use of the outside option serves as additional evidence in support of the price-elasticity effect offsetting the quality effect. Characteristics that proxy poverty (i.e. female-headed households and those with fewer assets) correlate significantly with stopping using the service at follow-up in response to the maintenance treatment (Appendix D.13). These results are in line with evidence of user fees being regressive (Gertler et al., 1987).

To understand the externality effects associated with the outside option, we study the health status of residents in columns (2)–(6) of Table 3. The dependent variable in column (2) is the (self-reported) morbidity, measured by an indicator variable equal to 1 if any household member had fever, diarrhea or cough during the two weeks prior to the interview, and 0 otherwise. Columns (2)–(5) focus on OOP health expenditures, distinguishing between curative and preventive expenditures, and analyzing changes both at the extensive and intensive margins. Curative expenditures include costs associated with doctor visits during illnesses, medicines and diagnostics, and hospitalization. Preventive expenditures include all costs associated with paying CT fees to use the sanitation facilities, access to drinking water and hygiene, scheduled medical checks and preventive goods like vaccines, bednets, and anti-worm tablets. The specifications in columns (2)–(6) are at the household level.

We document results consistent with an increase in water-borne diseases from greater OD. While, on average, we find no treatment effect on morbidity, the maintenance treatment increases the probability of spending a positive amount on curative healthcare by 4.9 percentage points (7.7% higher than the control mean and the p-value is adjusted from 0.05 to 0.26). The effect is larger and more precisely estimated during the first follow-up survey, and in the same follow-up we also find that the maintenance treatment increased self-reported morbidity by 7.6 percentage points (Appendix Figure D3). We find no effects on the intensive margin of curative expenditures, in line with water-borne diseases treated with low-cost therapies, nor on preventive expenditures.

[24] Using the same technique, we find that, in the control group, 58.4% of respondents used the CT the previous day and 82.0% of respondents washed hands with soap. We do not find any significant effect of the maintenance treatment for these variables (Appendix D.9). 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: due to the quality effect, more individuals in the maintenance treatment were likely to use the CT the previous day of the interview, but with a lower frequency.

6.2 Adding demand-side sensitization

We investigate the additional effects of shifting the valuation of the outside option (θ in our theoretical model) through a sensitization campaign aiming at raising awareness about the negative consequences of OD among residents. See Section 3 for details about the campaign.

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

$$Y_{ij} = \beta_1 \, T1_j + \beta_2 \, T2_j + \alpha \, \mathbf{X}_{ij} + \epsilon_{ij} \tag{6}$$

where $T1_j$ is an indicator variable for whether CT j received the maintenance treatment but the residents in its catchment area did not receive any sensitization campaign (*maintenance only* group), and 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). The randomization created observationally equivalent groups across all the treatment arms for household, catchment area, and CT characteristics measured at baseline (Appendix Tables D1–D2 and Section D.2).

The exposure to WASH campaigns was already relatively high among study households, including the awareness creation efforts of the government of India's Swachh Bharat Mission. This exposure reflected in high baseline awareness of the negative consequences of OD. Despite these ongoing campaigns, our sensitization campaign was effective at reaching the targeted population and improving awareness (Appendix D.6). 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 (8.2 percentage points higher than the average share in the control group, which was 0.65%) and the recall of posters (27.5 percentage points higher than the control group). The sensitization treatment group also raised awareness of externalities from OD (4.7 percentage points higher than the control group mean 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 maintenance only treatment and of the maintenance plus sensitization treatment, estimated with equation (6) and pooling together 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 Appendix D.2.

We find no differential effects between providing 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 mostly driven by top-down incentives and that raising awareness of externalities is not enough to offset the positive effects on OD practice.

7 Conclusion

Our research advances our understanding of the complexities of decentralized public service delivery. We demonstrate that an external quality boost in 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 the resulting externalities have an impact on overallpublic health. These findings highlight the importance of considering both user perspectives and provider incentives when shaping policies related to the provision of basic services.

In challenging the conventional belief that higher service quality invariably attracts more users, our study underscores the need for a balanced approach that carefully weighs quality enhancements against the risk of user exclusion and potential externalities. One crucial implication is that in cases where citizens perceive access to basic services as a right, fee-funded models of public service delivery may not be effective. As the international community endeavors to achieve universal access to safe and affordable basic services, particularly within the United Nations' Sustainable Development Goals, it is imperative to deepen our understanding of how to sustainably provide these services in economically disadvantaged settings. This extends to exploring service delivery mechanisms that ensure equitable access without compromising quality.

While providing free basic services in these areas may seem like an intuitive solution, it presents formidable challenges. Existing evidence indicates mixed effects of providing free services on user behavior (see, for example, Szabo, 2015). Expanding basic services when potential users are not willing to pay for its use has also proven ineffective (Lee et al., 2020). Additionally, subsidizing services may inadvertently discourage investments in infrastructure maintenance, potentially perpetuating poor service quality (McRae, S., 2015).

Therefore, it is crucial to enhance our knowledge regarding the design of effective mechanisms for tax collection and redistribution that can fund service delivery in the poorest settings. While our results highlight the significance of top-down incentives, further evidence is needed to design effective mechanisms that stimulate bottom-up incentives. Understanding the constraints to collective action in areas characterized by prevalent coordination failures and resistant social norms is a vital research objective. For instance, investigating the effectiveness of monitoring technologies in these environments to create and reinforce new local norms of respect for the public good is an area that warrants further exploration.

References

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.

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

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

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

Ashraf, N., Berry, J., and Shapiro, J. M. (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *American Economic Review*, 100(5):2383–2413.

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

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

Balan, P., Bergeron, A., Tourek, G., and Weigel, J. L. (2022). Local elites as state capacity: How city chiefs use local information to increase tax compliance in the Democratic Republic of the Congo. *American Economic Review*, 112(3):762–797.

Bancalari, A. (2020). The unintended consequences of infrastructure development. IFS working paper W20/32, Institute for Fiscal Studies, London, United Kingdom.

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

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

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

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

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

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

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

Best, M. C., Hjort, J., and Szakonyi, D. (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.

Bird, R. (2010). Taxation and Decentralization. Technical report, The World Bank, Washington D.C.

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

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

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

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

Chatterjee, U., Mobarak, A. M., Munoz, J., and Sen, I. K. (2023). Progress on sanitation in rural India: Reconciling diverse evidence. Technical report, World Bank Policy Research Working Paper 185062, Washington D.C.

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

Coville, A., Galiani, S., Gertler, P. J., and Yoshida, S. (2020). Pipe dreams: Enforcing payment for water and sanitation services in Nairobi's slums. NBER Working Paper 27569, National Bureau of Economic Research.

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

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.

Dupas, P. (2014a). Pricing and user fees. *Encyclopedia of Health Economics*, 3:136–141.

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

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.

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

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.

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.

Greenstone, M. and Jack, B. K. (2015). Envirodevonomics: A research agenda for an emerging field. *Journal of Economic Literature*, 53(1):5–42.

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.

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.

Khan, A. Q., Khwaja, A. I., and Olken, B. A. (2016). Tax farming redux: Experimental evidence on performance pay for tax collectors. *The Quarterly Journal of Economics*, 131(1):219–271.

Kremer, M. and Holla, A. (2009). Improving education in the developing world: What have we learned from randomized evaluations? *Annu. Rev. Econ.*, 1(1):513–542.

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

Lee, K., Miguel, E., and Wolfram, C. (2020). Does household electrification supercharge economic development? *Journal of Economic Perspectives*, 34(1):122–44.

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.

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

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.

Marx, B., Stoker, T. M., and Suri, T. (2013). The economics of slums in the developing world. *Journal of Economic Perspectives*, 27(4):187–210.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Stavins, R. N. (2011). The problem of the commons: Still unsettled after 100 years. *American Economic Review*, 101(1):81–108.

Szabo, A. (2015). The value of free water: Analyzing South Africa's free basic water policy. *Econometrica*, 83(5):1913–1961.

Tirole, J. (1996). A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *The Review of Economic Studies*, 63(1):1–22.

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

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

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

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

World Bank (2004). World development report. Technical report, World Bank.

Table 1: Service delivery

Dep. variable:	**Service delivery**	**Inputs to service delivery**		
	Quality	Maintenance		Monitoring
		Cleaning	Rehabilitation	
	(1)	(2)	(3)	(4)
Maintenance (T)	0.064	0.057	-0.027	0.060
	(0.024)	(0.016)	(0.053)	(0.032)
	[0.01, 0.03]	[0.00, 0.01]	[0.62, 0.62]	[0.07, 0.15]
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 (5). Standard errors clustered by catchment area are reported in parentheses. The p-values 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 6 for details). Dependent variables by column: (1) *Quality*, 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*, index including the number of tools, 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 (see Appendix Table D10 for individual components); (3) *Rehabilitation*, 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*, share of worked hours allocated by the caretaker to collecting fees and supervising cleaners, rather than conducting activities away from the entrance or off-site. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix B.

Table 2: Use and payment for the service

Dep. variable:	**Payment for the service**			**Use of the service**		
	Share of users paying the fee	Share of residents willing to pay a positive amount	WTP among residents	Users	Number of uses among residents	
					Regular users	*Other residents*
	(1)	(2)	(3)	(4)	(5)	(6)
Maintenance (T)	0.093	-0.003	0.009	-1.941	-0.110	-0.193
	(0.042)	(0.022)	(0.087)	(1.626)	(0.047)	(0.094)
	[0.03, 0.08]	[0.90, 0.90]	[0.92, 0.92]	[0.24, 0.41]	[0.02, 0.06]	[0.04, 0.10]
Mean (control group)	0.556	0.648	1.205	33.903	1.383	0.763
Observations	434	222	6001	434	2417	883
Catchment areas	110	109	109	110	109	102
Observation rounds	4	2	2	4	2	2
Level of analysis	CT	CT	Respondent	CT	Household	Household
Measurement	Observed	Incentivized	Incentivized	Observed	Self-reported	Self-reported

Note. Estimates based on CT-level OLS regressions using equation (5). Standard errors clustered by catchment area are reported in parentheses. The p-values 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 6 for details). Dependent variables by column: (1) *Share of users paying the fee*, observed share of users who pay the entry fee; (2) *Share of residents willing to pay a positive amount*, share of residents with a positive WTP in 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; (3) *WTP among residents*, 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; (4) *Users*, total number of users observed; (5)–(6) *Number of uses among residents*, 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 strata indicators for the city and the provider of the CT. Columns (5) and (6) are estimated on relevant sub-samples. Additional details about the variables are presented in Appendix B.

Table 3: Outside option and health consequences

Dep. variable:	Practiced OD	Morbidity	Health expenditure			
			Curative		Preventive	
			Extensive	Intensive	Extensive	Intensive
	(1)	(2)	(3)	(4)	(5)	(6)
Maintenance (T)	0.172	0.029	0.049	-35.277	-0.003	4.542
	(0.080)	(0.027)	(0.025)	(195.308)	(0.003)	(56.857)
	[0.03, 0.22]	[0.28, 0.73]	[0.05, 0.26]	[0.86, 0.97]	[0.44, 0.88]	[0.92, 0.92]
Mean (control group)	0.210	0.451	0.636	1700.010	0.992	741.053
Observations	817	3323	3298	3298	3323	3322
Catchment areas	107	109	109	109	109	109
Observation rounds	1	2	2	2	2	2
Level of analysis	Respondent	Household	Household	Household	Household	Household
Measurement	List randomization	Self-reported	Self-reported	Self-reported	Self-reported	Self-reported

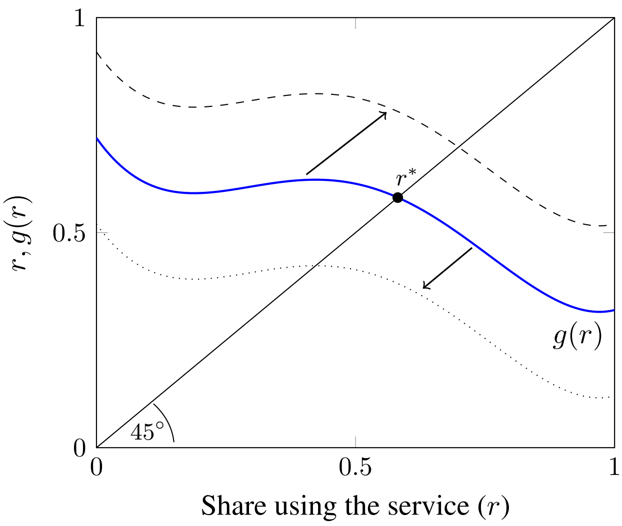
Note. Estimates based on household-level OLS regressions using equation (5). Standard errors clustered by catchment area are reported in parentheses. The *p*-values presented in brackets, the first from individual testing, the second adjusting for testing that each treatment is jointly different from zero for all outcomes presented in the table (see Section 6 for details). Dependent variables by column: (1) *Practiced OD*, 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) *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; (3) *Curative expenditure - extensive*, indicator variable equal to 1 if the respondent had positive curative expenditures, and 0 otherwise; (4) *Curative expenditure - intensive*, level of curative healthcare expenditures (in rupees); (5) *Preventive expenditure - extensive*, indicator variable equal to 1 if the respondent had positive preventive expenditures, and 0 otherwise; (6) *Preventive expenditure - intensive*, level of preventive healthcare expenditures (in rupees). 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)–(6) include only 109 catchment areas in the sample because the dependent variables were measured only in rounds 3 and 5, after a catchment area was displaced. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix B.

Table 4: The effect of sensitization

	Maintenance only			Maintenance + sensitization			$T1 = T2$
	β	se	p-value	β	se	p-value	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quality	0.07	0.03	0.01	0.05	0.03	0.09	0.58
Maintenance: cleaning	0.06	0.02	0.00	0.06	0.02	0.00	0.85
Maintenance: rehabilitation	-0.04	0.06	0.47	-0.01	0.06	0.85	0.60
Monitoring	0.05	0.04	0.22	0.07	0.04	0.04	0.35
Share of users paying	0.08	0.05	0.09	0.11	0.05	0.03	0.54
Share of residents with positive WTP	0.01	0.03	0.74	-0.03	0.02	0.28	0.22
WTP among residents	0.09	0.11	0.41	-0.07	0.10	0.49	0.16
Users	-2.61	1.85	0.16	-1.25	1.81	0.49	0.42
Number of uses among residents:							
Regular users	-0.06	0.05	0.28	-0.16	0.06	0.01	0.13
Other	-0.23	0.11	0.04	-0.16	0.11	0.16	0.58
Practiced OD	0.19	0.10	0.05	0.16	0.09	0.08	0.71
Morbidity	0.03	0.03	0.36	0.03	0.03	0.34	1.00
Health expenditure:							
Curative (extensive)	0.04	0.03	0.17	0.06	0.03	0.04	0.52
Curative (intensive)	31.25	227.29	0.89	-98.73	226.54	0.66	0.58
Preventive (extensive)	-0.00	0.00	0.41	-0.00	0.00	0.60	0.73
Preventive (intensive)	20.09	64.90	0.76	-10.44	63.43	0.87	0.61

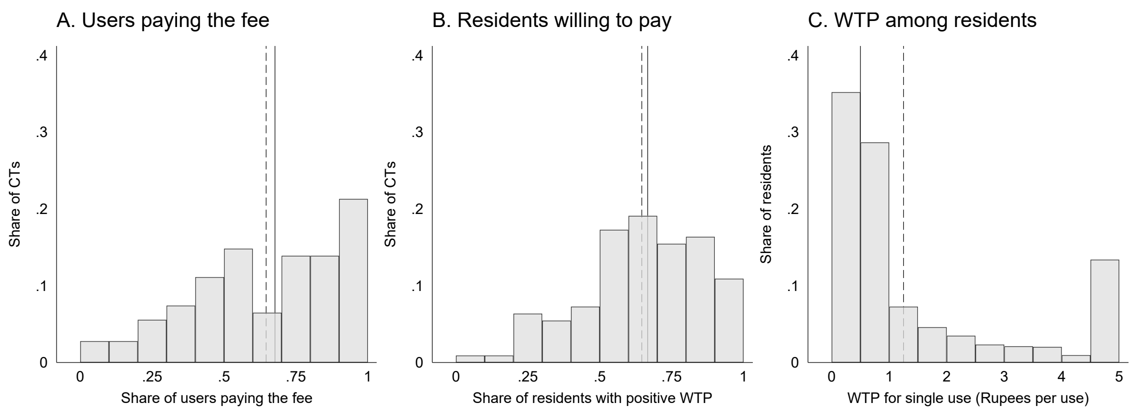
Note. In columns (1)–(6), estimates are based on CT-, respondent- or household-level OLS regressions using equation (6) for each outcome. $p-$values are presented in columns (3) and (6), the first from individual testing, the second adjusting for jointly testing that each treatment is different from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

Figure 1: Equilibrium level of service use



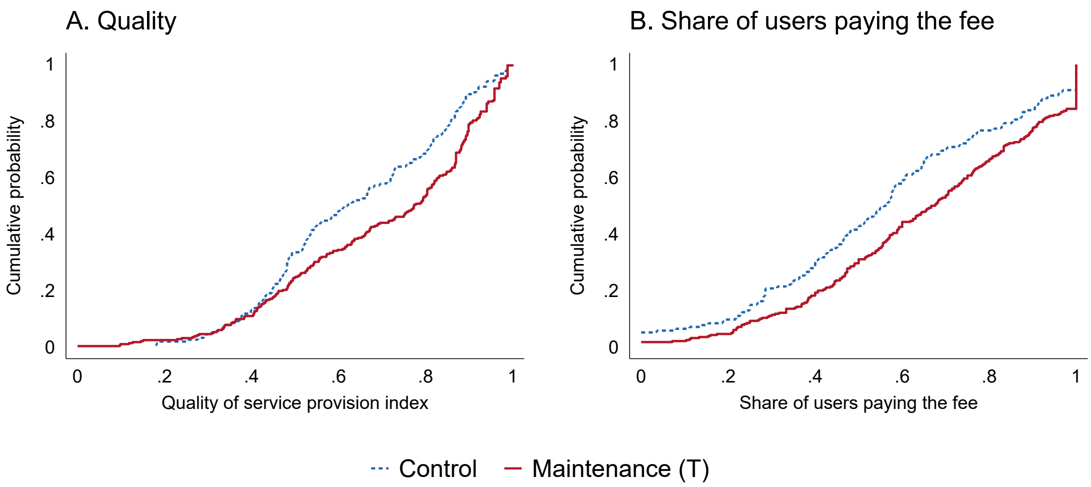
Note. The figure provides a graphical proof of Proposition 1. The function $g(r)$ represents the demand for the service, defined by equation (2), where r is the share of individuals using the service. A share $1 - r$ uses the outside option. The equilibrium point r^* is given by the intersection between $g(r)$ and the 45-degree line. Details of the theoretical framework are provided in Section 2.

Figure 2: Payment and WTP for CT use, at baseline



Note. Data collected at baseline. Panel A reports the (observed) share of users who pays the fee for the use of the CT during 1 hour at dawn. 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 market price for a single CT use. The solid vertical lines represent the sample median, and the dashed vertical lines represent the sample mean. Additional details about the variables are presented in Appendix B.

Figure 3: 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 group. 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 Appendix B.

ONLINE APPENDIX

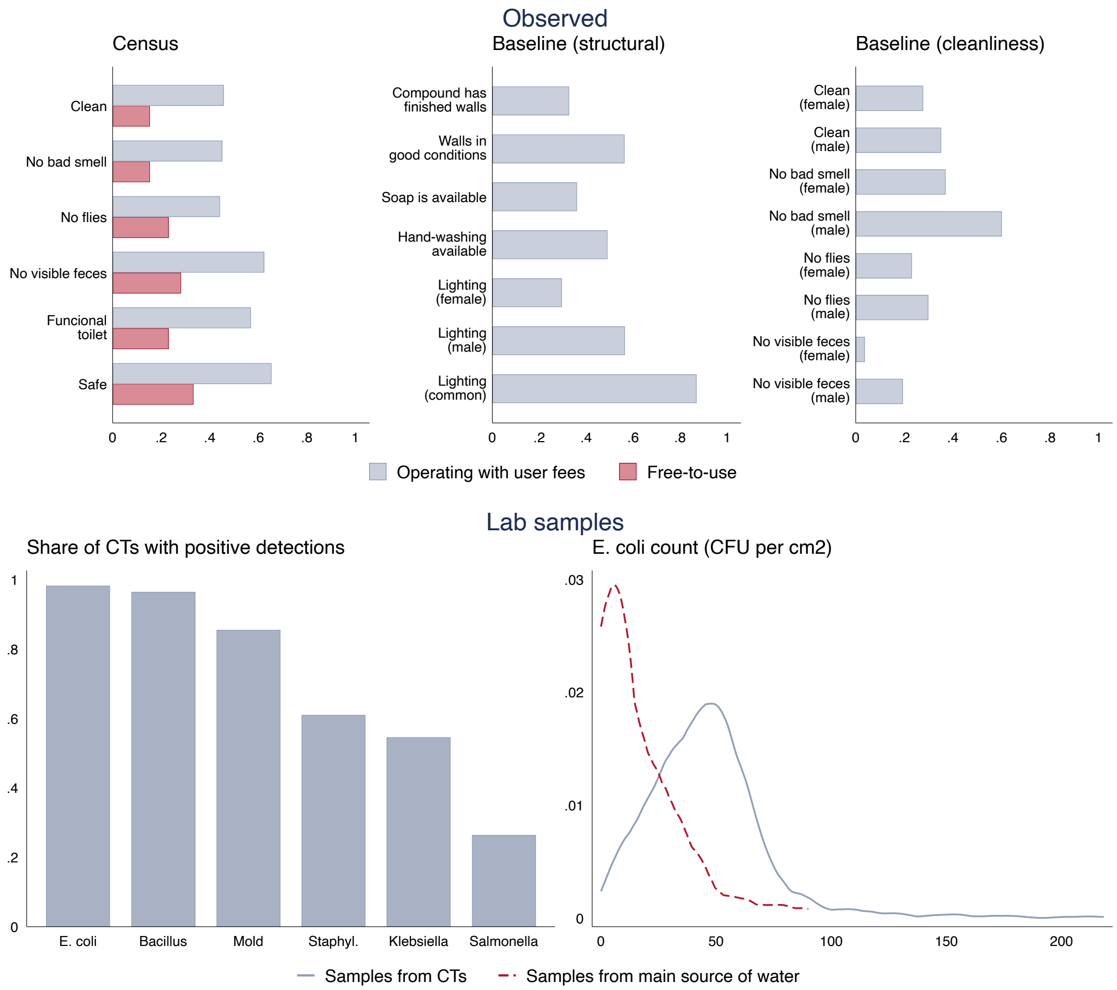
**Public Service Delivery, Exclusion and Externalities:
Theory and Experimental Evidence from India**

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

A Status quo of service delivery in study area

Figure A1 summarizes the status of service delivery in the study area. The top panel presents statistics on observed dimensions of delivery (see Section 5.1), comparing free- and pay-to-use facilities and summarizing the average status of pay-to-use facilities included in the study as collected by observers at baseline. The bottom panel presents instead summary statistics for the prevalence of bacteria and mold in the surfaces of CTs and in water samples collected in their proximity.

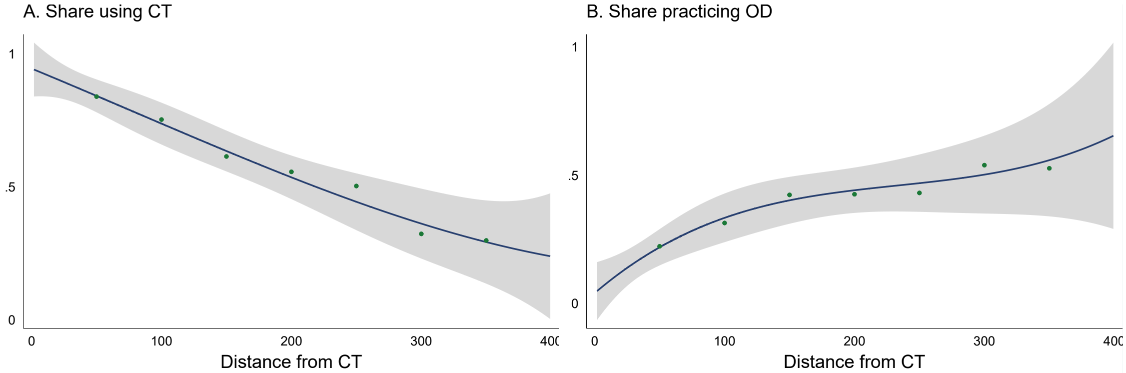
Figure A1: Status of service delivery in study area



Note. Share of CTs that have or have access to each characteristic. For the top panel, the left figure refers to information collected during the CT census and includes all CTs in Lucknow and Kanpur, while the middle and right figures refer to information collected at baseline for the sample of facilities included in the study. *Female*, *male* and *common* indicate the area of the CT where the information is observed. For the bottom panel, the left figure presents the share of CTs where each bacteria type or mold was detected in at least one of the three samples, while the right figure shows the distribution of the E. coli count from CT and water samples. The distribution fits are estimated non-parametrically using kernel density estimation assuming an Epanechnikov kernel function. Bandwidths are estimated by Silverman's rule of thumb. The collection of samples is described in Appendix F. For details about these data sources see Section 5.1.

To understand service use, we study how distance from a facility affects use using self-reported data from the census of residents (see Section 5.2). Figure A2 presents cubic fits for the relationships between the distance from the facility and self-reported use of the service (panel A), or OD (panel B).

Figure A2: Sanitation behavior, by distance from a facility



Note. Data source is the slum resident census (see Section 5.2). The figures present cubic fits of the share of residents using the CT (panel A) and of the share practicing OD (panel B) on the distance between the resident's dwelling and the closest CT. Dots show averages for equally spaced intervals. The shaded area presents the 90% confidence intervals, assuming standard errors are clustered at the slum level. The sample includes all households considered eligible for the study (see Section 5).

B Definition of variables

B.1 Definition of outcome variables

Variable [model parameter]	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 expenditures [$\gamma(1-r)$]	Expenditures incurred during the month previous to the interview: *curative*, OOP 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*, OOP expenses associated with regular doctor checks, vaccines, anti-worm tablets, bednets, and prenatal tests, and travel costs associated with these expenses. *Extensive* (margin) is an indicator variable equal to 1 if the respondent had positive expenditures, and 0 otherwise. *Intensive* (margin) is the level of expenditures (in rupees). The variable is self-reported by the household head during all residents' survey rounds, except for the mid-intervention survey.
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 [x]	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: rehabilitation [x]	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 [e]	Share of worked hours allocated by the caretaker to collecting fees and supervising cleaners. Alternative activities are those away from the fee-collection point, such as conducting repairs, cleaning the facility, and meeting the manager. The variable is self-reported by the caretaker.
Morbidity [$\gamma(1-r)$]	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 [$r \cdot \pi$]	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 INR 5). Information is collected using observation during the rush hour.
Number of uses among residents [r]	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 [$(1-r)$]	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 [y]	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.

(continued on next page)

Variable [model parameter]	Description
Refused entry for not paying	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 [π]	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 [π]	Share of users entering the CT and paying the entry fee. The variable uses data from observers collected at the entrance of the CT in every round.
Transfer to the CT [-]	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. Information based on project's administrative data.
Transfer to the caretaker [-]	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.
Users [r]	Total number of users entering the CT (reported in logarithms). The variable uses data from observers collected at the entrance of the CT in every round.
WTP among residents [π]	WTP 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.
Additional sources	
Basemaps [-]	Basemaps were created using ArcGIS® software by Esri®, and used in line with the Esri Master License Agreement, specifically for the inclusion of screen captures in academic publications.

Note. Model parameter is the parameter in the theoretical framework that relates to the empirical measurement (see Section 2). Appendix F provides detailed descriptions and scripts of each measurement.

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, D17, D14
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).	D14
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.	D8
Hygiene	Hygienic practices of respondents and family members (see *Note*).	D14

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 5.2).

C Study location, timeline, and sampling

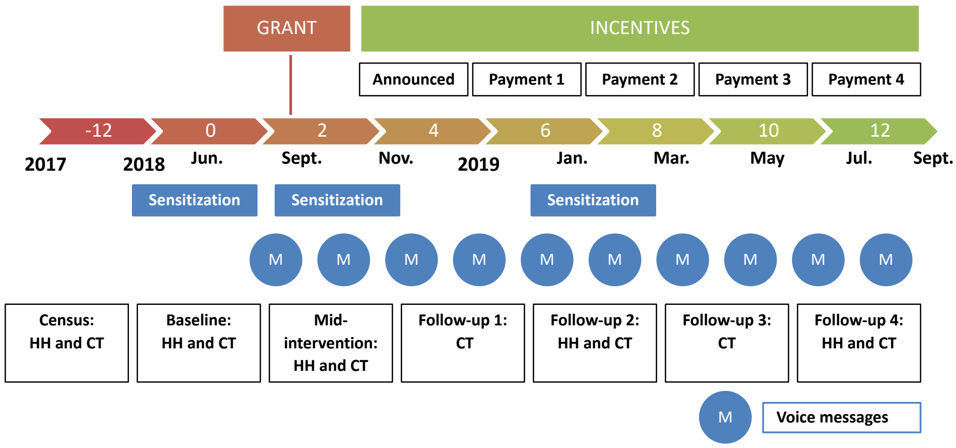
Figure C1 describes the timeline of interventions and data collection efforts and a summary of the sampling procedure, while Figure C2 provides the spatial distribution of CTs in the study. To obtain the sampling frame, during the first half of 2017, we performed a mapping of slums and a census of all 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.

To identify residents, during the second half of 2017, we performed a census of all households living within slum borders and within 400 meters of the selected CTs. The questionnaire was administered to household heads and gathered information for more than 30,000 households. Among censused households, we selected residents according to the conditions described in Section 5.2. We sampled up to 17 eligible households per catchment area. For areas with fewer than 10 eligible households available within 150 meters from a CT, we selected all households within this bound, and randomly selected the remaining ones from the households living 150–250 meters from a CT.

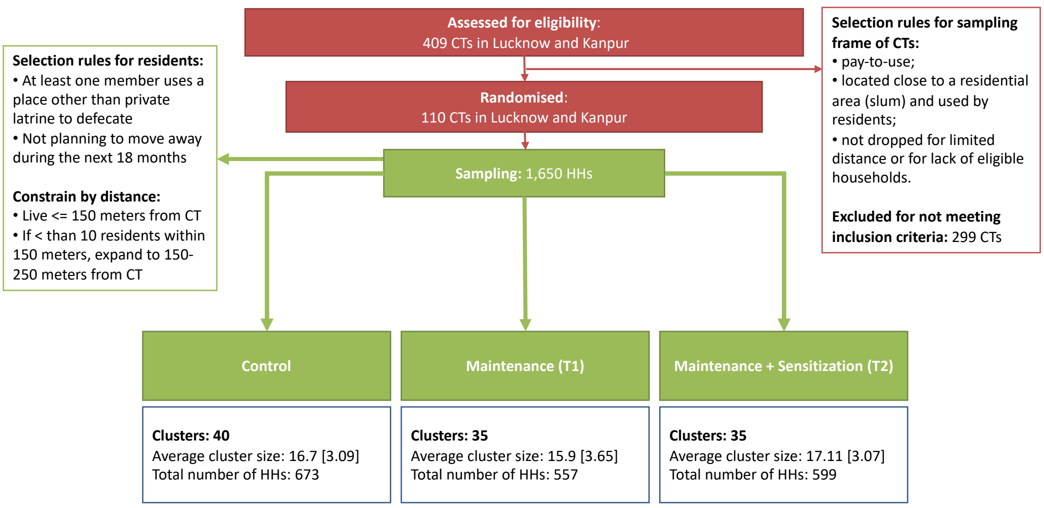
In total, we obtained a sample of 1,650 households. This sample presents average characteristics aligned with the slum residents across all states of India and in UP, as detailed in the 2011 Indian Slum Population Census. The share of male individuals is 52% and 53% in Indian’s and UP’s slums, and 53% in our study sample. Similarly, the share of children is 12%, 14% and 9%, respectively. In terms of caste composition, the study sample is for 45% represented by individuals being part of a scheduled caste, versus only 20% and 22% in India’s and UP’s slums, respectively. Literacy rate tends to be instead lower in the study sample as compared to the rest of the country (46% versus 78% and 69%, respectively).

Figure C1: Timeline and sampling strategy

A. Timeline of interventions and data collection efforts

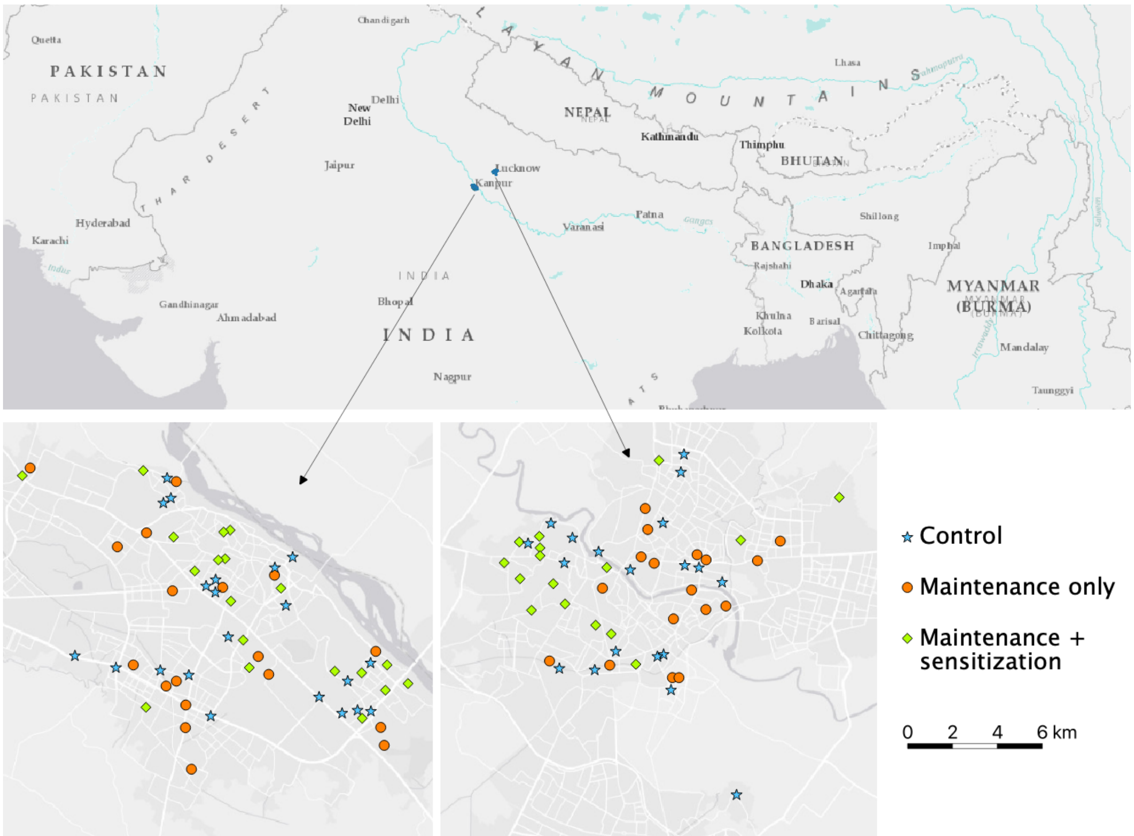


B. Sampling design



Note. Panel A shows the timeline of interventions and data collection efforts. 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 5. Panel B 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 5.

Figure C2: 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 present the balance test for characteristics at baseline. The null hypothesis of equal means across the treatment arms cannot be rejected for any of the characteristics other than for the household head's education, though the difference in means of all characteristics are 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	2.78	2.34	3.23	0.32
	[8.85]	[9.29]	(1.88)	(2.11)	(2.19)	
Distance to closest CT	0.54	0.58	-0.06	-0.04	-0.07	0.76
	[0.44]	[0.66]	(0.11)	(0.11)	(0.11)	
Surrounding market	0.33	0.35	-0.04	-0.01	-0.06	0.82
	[0.47]	[0.48]	(0.10)	(0.11)	(0.11)	
Surrounding road	0.84	0.88	-0.06	-0.05	-0.08	0.67
	[0.37]	[0.33]	(0.07)	(0.09)	(0.09)	
Surrounding government office	0.25	0.20	0.07	0.08	0.06	0.69
	[0.43]	[0.41]	(0.08)	(0.10)	(0.10)	
Only residents use CT	0.12	0.07	0.07	0.07	0.07	0.53
	[0.32]	[0.27]	(0.06)	(0.07)	(0.07)	
Single caretaker	0.80	0.82	-0.04	0.03	-0.11	0.28
	[0.40]	[0.39]	(0.07)	(0.08)	(0.09)	
Share of female caretakers	0.18	0.22	-0.06	-0.02	-0.10	0.42
	[0.37]	[0.39]	(0.07)	(0.08)	(0.08)	
Caretaker is also cleaner	0.27	0.28	-0.02	-0.02	-0.03	0.96
	[0.45]	[0.46]	(0.09)	(0.10)	(0.10)	
Caretaker is from local community	0.44	0.49	-0.07	-0.11	-0.02	0.60
	[0.50]	[0.51]	(0.10)	(0.12)	(0.12)	
Caretaker's experience (months)	125.28	129.91	-5.43	1.37	-11.53	0.86
	[103.45]	[109.34]	(22.81)	(26.60)	(25.96)	
CT is cleaned frequently	0.86	0.87	-0.02	-0.02	-0.02	0.97
	[0.35]	[0.34]	(0.07)	(0.08)	(0.08)	
Time allocated to managing	0.68	0.66	0.03	0.03	0.02	0.58
	[0.14]	[0.11]	(0.03)	(0.03)	(0.03)	
Capacity	13.00	13.21	-0.32	-0.46	-0.17	0.94
	[5.57]	[5.52]	(1.11)	(1.27)	(1.34)	
Daily opening hours	17.76	17.88	-0.19	-0.35	-0.02	0.53
	[1.49]	[1.59]	(0.28)	(0.36)	(0.27)	
Share of functioning toilets	0.90	0.88	0.03	0.05	0.01	0.47
	[0.22]	[0.23]	(0.04)	(0.04)	(0.05)	
WTP (avg. catchment area)	1.41	1.44	-0.05	-0.03	-0.06	0.95
	[0.83]	[0.65]	(0.15)	(0.17)	(0.20)	
Distance from CT (avg. catchment area)	128.71	128.77	-0.01	-2.22	2.21	0.94
	[49.56]	[43.87]	(9.26)	(10.21)	(12.25)	

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

Table D2: Household characteristics at baseline, by treatment group

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

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

Table D3 tests whether attrition created differences across treatment arms. For the CT survey, columns (1)–(2) show estimates of treatment effects on the number of observations and interviews with caretakers in the follow-up rounds.[2] Concerning the survey of residents, columns (2)–(6) in Table D3 estimate the probability of attrition in the sample of residents as a function of the treatment status.[3] In order to maintain

[2]We collected data for all of the 110 selected CTs at the 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, given that some CTs closed temporarily/permanently for refurbishment, and one slum was completely displaced after follow-up 1. In some cases, we were able to collect observations, while not being able to survey caretakers.

[3]We interviewed 1,575 households at baseline (an average of 12 households per cluster), 1,532 households during

a comparable sample size in all follow-up surveys, we handled attrition with replacements at random using the sampling frame used for the baseline sampling. Column (7) tests whether the replacement was introduced differently across treatment arms. For all outcomes, we do not observe any significant difference across treatment groups.

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)	0.074	0.110	0.033	-0.011	-0.020	-0.013	0.007
	(0.074)	(0.114)	(0.058)	(0.028)	(0.036)	(0.031)	(0.028)
	[0.32]	[0.33]	[0.57]	[0.68]	[0.57]	[0.67]	[0.81]
Panel B							
Maintenance only (T1)	0.073	0.044	0.042	-0.013	-0.026	-0.016	0.011
	(0.074)	(0.142)	(0.063)	(0.030)	(0.037)	(0.035)	(0.033)
	[0.32]	[0.76]	[0.50]	[0.66]	[0.48]	[0.64]	[0.75]
Maintenance + sensitization (T2)	0.075	0.173	0.025	-0.010	-0.015	-0.010	0.003
	(0.076)	(0.107)	(0.066)	(0.030)	(0.041)	(0.034)	(0.033)
	[0.32]	[0.11]	[0.71]	[0.75]	[0.72]	[0.78]	[0.92]
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

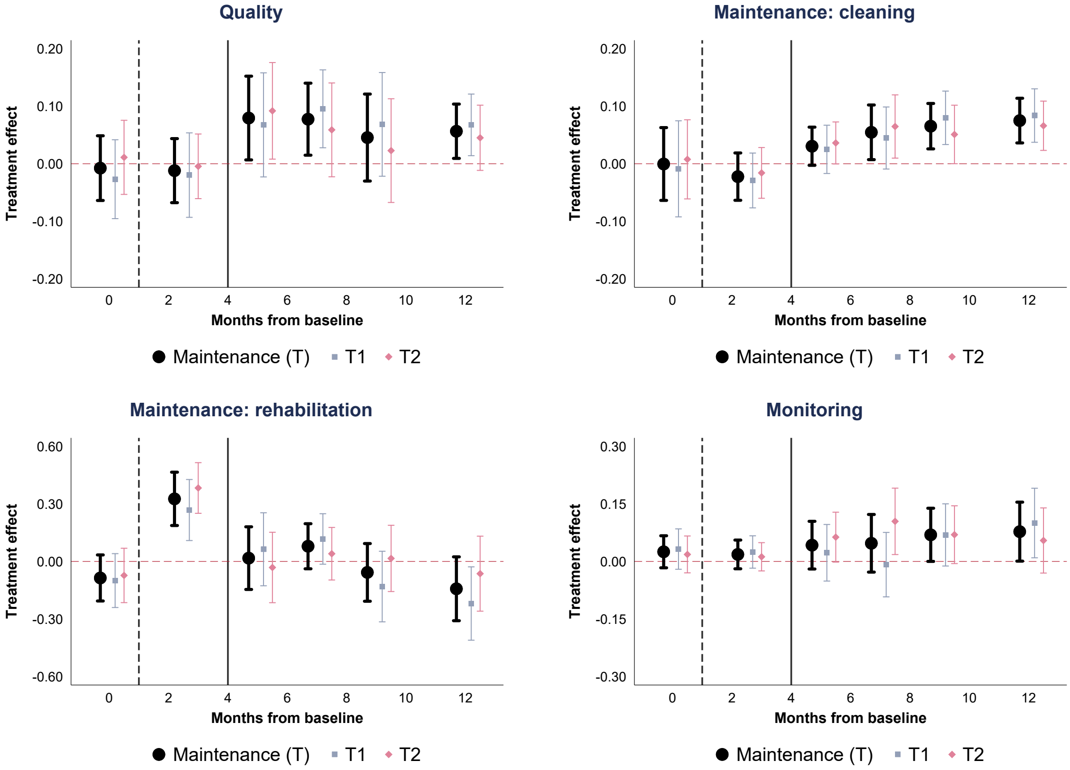
Note. Estimates based on OLS regressions using equation (5) in panel A, and equation (6) in panel B. Robust standard errors are presented in parenthesis in columns (1)–(2). Standard errors clustered by catchment area are presented in parenthesis in columns (3)–(7) The p-values are presented in brackets. 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 (it was interviewed in any of the follow-ups, but it was not interviewed at baseline), 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 include strata indicators for the city and the provider of the CT. Figure C1 provides the timing of each follow-up survey.

mid-intervention survey, 1,578 households at follow-up 2, and 1,772 households in the follow-up 4. On average, each interview, including all modules, had an average duration of one hour.

D.2 Estimates of treatment effects by survey round

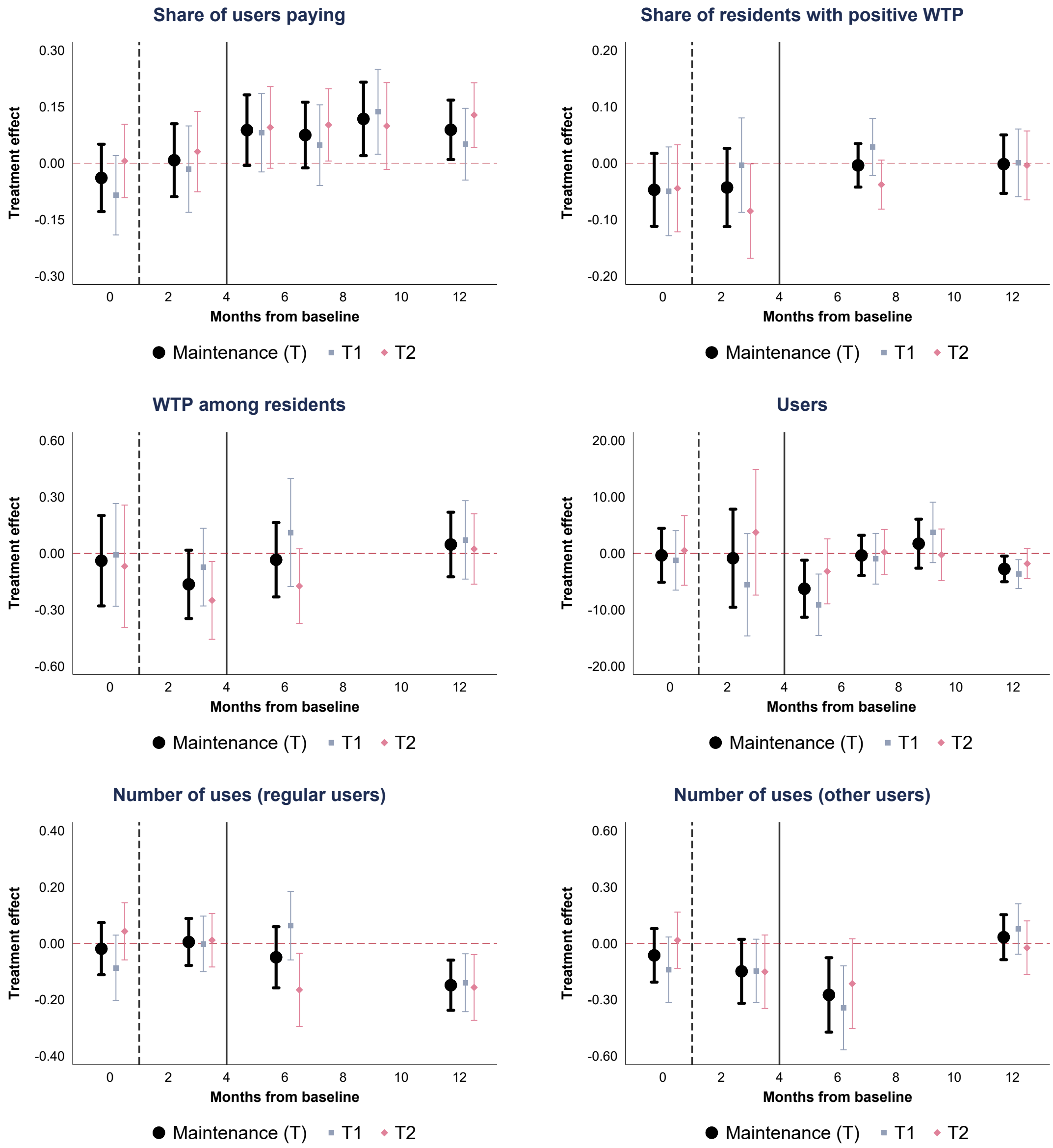
For the outcomes presented in Tables 1–3, this section presents estimates of equation (5) and equation (6) separately for each survey. Estimates are presented in Figures D1–D3. The upper part of each panel presents estimates of treatment effects on the corresponding variable, while the lower part reports the evolution over time of the average of the corresponding variable in the control group.

Figure D1: Timing of effects for outcomes in Table 1



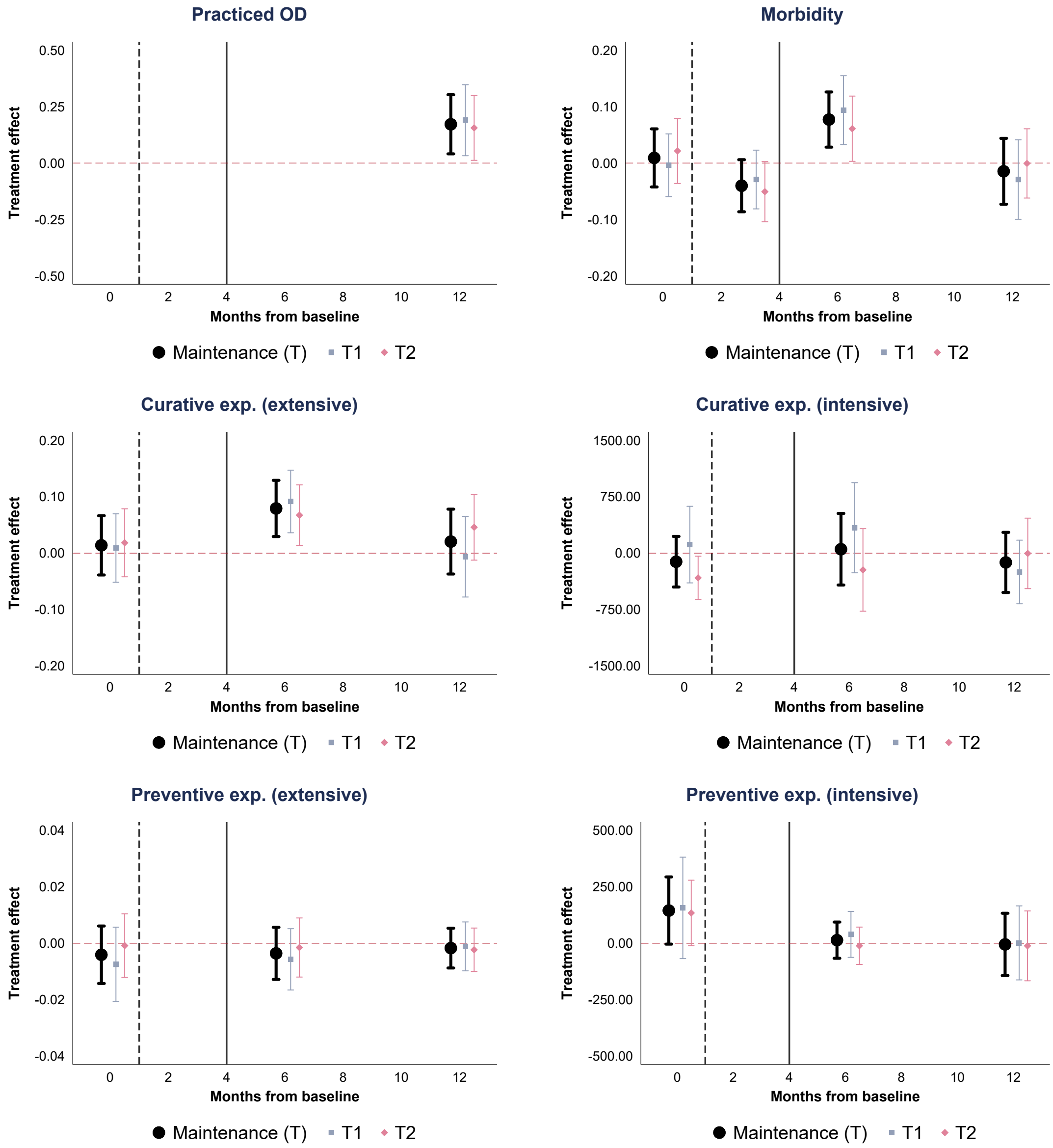
Notes. Estimates based on CT-level OLS regressions using equation (5) and equation (6) separately for each data collection period. Period 0 indicates the *baseline* measurement. The measurement in between the two vertical lines is the *mid-intervention* measurement. All subsequent periods (to the right 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 include strata indicators for the city and the provider of the CT.

Figure D2: Timing of effects for outcomes in Table 2



Notes. Estimates based on CT-level OLS regressions using equation (5) and equation (6) 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 include strata indicators for the city and the provider of the CT.

Figure D3: Timing of effects for outcomes in Table 3



Notes. Estimates based on household-level OLS regressions using equation (5) and equation (6) 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 include strata indicators for the city and the provider of the CT. 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 (5) and (6) adding the value at baseline of the dependent variable as a control variable (ANCOVA specification). Table D4 also present estimates of treatment effects using equations (5) and (6) 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: rehabilitation	-0.01	0.05	0.79	1			
Monitoring	0.05	0.03	0.09	1			
Share of users paying	0.10	0.04	0.01	1			
Share of residents with positive WTP	-0.01	0.02	0.56	1			
Users	-1.94	1.62	0.23	1			
WTP among residents	0.01	0.08	0.90	1	0.02	0.09	0.82
Practiced OD	0.17	0.08	0.03	0	0.16	0.08	0.05
Number of uses (regular users)	-0.11	0.04	0.02	1	-0.11	0.05	0.02
Number of uses (other users)	-0.11	0.09	0.24	1	-0.19	0.10	0.05
Morbidity	0.03	0.03	0.29	1	0.03	0.03	0.28
Curative exp. (extensive)	0.05	0.02	0.05	1	0.05	0.03	0.06
Curative exp. (intensive)	-34.25	194.71	0.86	1	-61.42	203.51	0.76
Preventive exp. (extensive)	-0.00	0.00	0.43	1	-0.00	0.00	0.39
Preventive exp. (intensive)	0.20	56.36	1.00	1	1.82	54.75	0.97

Note. Estimates based on respondent- and household-level OLS regressions using equation (5), controlling for the baseline value of the dependent variable if available (see *ANCOVA specification* column, 1 = Yes) 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 strata indicators for the city and the provider of the CT. Specifications where the level of analysis is the respondent also include gender. Additional details about the variables are presented in Appendix B.

D.4 Robustness to the inclusion of control variables

We follow two approaches. First, Table D5 presents estimates of the effect of the maintenance treatment (T) using equation (5) 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).

Second, we show robustness to estimation of treatment effects via causal forest. 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 summa-

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	0.06	0.02	0.01	0.05	0.02	0.03	434
Maintenance: cleaning	0.06	0.02	0.00	0.06	0.01	0.00	434
Maintenance: rehabilitation	-0.03	0.05	0.62	-0.04	0.05	0.47	434
Monitoring	0.06	0.03	0.07	0.06	0.03	0.04	434
Share of users paying	0.09	0.04	0.03	0.08	0.04	0.04	434
Share of residents with positive WTP	-0.00	0.02	0.90	-0.01	0.02	0.78	222
Users	-1.94	1.63	0.24	-2.01	1.63	0.22	434
WTP among residents	0.01	0.09	0.92	-0.01	0.08	0.94	6001
Practiced OD	0.17	0.08	0.03	0.16	0.08	0.04	817
Number of uses (regular users)	-0.11	0.05	0.02	-0.10	0.05	0.07	2417
Number of uses (other users)	-0.19	0.09	0.04	-0.12	0.09	0.20	883
Morbidity	0.03	0.03	0.28	0.01	0.03	0.65	3323
Curative exp. (extensive)	0.05	0.02	0.05	0.05	0.03	0.07	3298
Curative exp. (intensive)	-35.28	195.31	0.86	-154.06	245.31	0.53	3298
Preventive exp. (extensive)	-0.00	0.00	0.44	0.00	0.00	0.74	3323
Preventive exp. (intensive)	4.54	56.86	0.94	14.09	73.68	0.85	3322

Note. Columns (1)–(3) show estimates using equation (5), while columns (4)–(6) show estimates using the PDSL procedure (Tibshirani, 1996; Belloni et al., 2013), with selection over a large number of baseline-level control variables. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. N indicates the sample size. In order to have the same sample size of estimates as in the main tables, missing values are replaced by the value 0 and an indicator variable equal to 1 if the observation had a missing value is introduced for all variables. Additional information about outcome variables is provided in Appendix B.

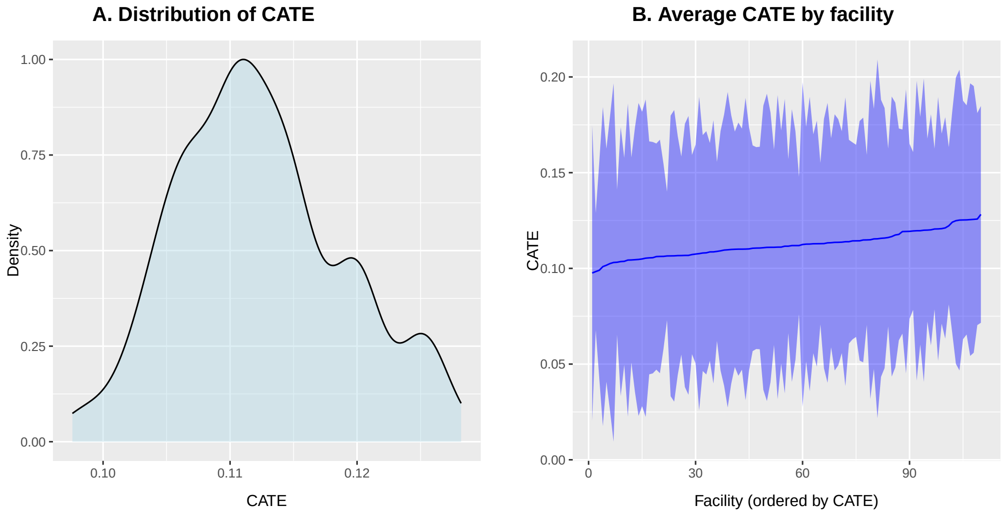
rizes the causal forest results on heterogeneity of the effect on payment. Panel A shows the distribution of the Conditional ATE (CATE), while panel B averages the CATE at CT level and includes the 90% confidence interval.

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	0.002	1.000
Maintenance: cleaning	0.061	0.016	0.000	0.000	1.000
Maintenance: rehabilitation	-0.032	0.054	0.549	0.182	1.000
Monitoring	0.055	0.036	0.122	0.064	1.000
Users	-1.623	1.678	0.333	0.159	0.576
Share of users paying	0.113	0.044	0.010	0.000	1.000
Share of residents with positive WTP	-0.004	0.023	0.870	0.695	1.000
WTP among residents	-0.002	0.098	0.985	0.576	0.992
Practiced OD	0.236	0.092	0.011	0.000	1.000
Number of uses (regular users)	-0.131	0.074	0.076	0.013	0.999
Morbidity	-0.007	0.032	0.817	0.573	0.997
Curative exp. (extensive)	0.043	0.034	0.203	0.049	0.999
Curative exp. (intensive)	-130.364	250.225	0.602	0.264	1.000
Preventive exp. (extensive)	0.002	0.005	0.739	0.214	1.000
Preventive exp. (intensive)	6.024	81.13	0.941	0.551	1.000

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 payment



Note. Panel A shows 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). Panel B shows the average CATE at CT level with the 90% confidence interval. Additional information about the variables is provided in Appendix B.

D.5 Spillover analysis

Table D7 shows a test for contagion or spillover effects by estimating heterogeneous treatment effects according to the minimum distance to a CT in the treatment group. Among all outcome variables, we do not observe any heterogeneous effect, suggesting the absence of spillover effects.

Table D7: Contagion and spillover effects

	Effect of Maintenance (T), by distance to another treatment unit						**Het. test**
	Close to another unit			Far from another unit			
	β	se	N	β	se	N	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quality	0.05	0.03	208	0.07*	0.04	226	0.61
Maintenance: cleaning	0.06***	0.02	208	0.05*	0.02	226	0.68
Maintenance: rehabilitation	-0.00	0.07	208	-0.04	0.09	226	0.72
Monitoring	0.07*	0.04	208	0.05	0.06	226	0.82
Share of users paying	0.08	0.05	208	0.07	0.07	226	0.96
Share of residents with positive WTP	-0.00	0.03	108	-0.00	0.03	114	0.98
WTP among residents	-0.14	0.14	2329	0.06	0.12	3672	0.33
Users	-0.84	2.27	208	-3.68	2.48	226	0.39
Number of uses (regular users)	-0.11	0.07	970	-0.11*	0.06	1447	0.96
Number of uses (other users)	-0.31	0.19	321	-0.05	0.10	562	0.16
Practiced OD	0.26*	0.14	334	0.09	0.10	483	0.28
Morbidity	0.03	0.04	1299	0.03	0.04	2024	0.94
Curative exp. (extensive)	0.04	0.04	1289	0.05	0.03	2009	0.90
Curative exp. (intensive)	126.82	285.46	1289	-197.58	264.06	2009	0.36
Preventive exp. (extensive)	-0.01**	0.00	1299	0.00	0.00	2024	0.15
Preventive exp. (intensive)	28.04	100.45	1298	-35.09	71.52	2024	0.51

Note. Close to (*far from*) indicates whether the minimum distance from a CT in the treatment group is below or equal to (above) the sample median. Variables referring to catchment areas are averages of the corresponding variable within the catchment area. In columns (1)–(6), estimates are based on CT-, respondent- or household-level OLS regressions using equation (5) separately for each category. Column (7) presents a heterogeneity test based on CT-level OLS regressions using equation (5) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. Standard errors are clustered by catchment area for CT-level outcomes and by catchment-area–round for respondent- and household-level outcomes. The dependent variables are indicated in the rows and are defined in Appendix B. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1.

D.6 Implementation of interventions across treatment groups

Table D8 shows the effect of treatments on indicators of exposure to the interventions. We focus on transfers as part of the maintenance intervention, and of indicators of the sensitization campaign. In columns (1) and (2), transfers are per period, and thus, total transfers can be obtained by multiplying the estimate with the number of observation rounds.

Table D8: 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)	4.739	0.761	0.053	0.090	0.031
	(0.060)	(0.034)	(0.020)	(0.028)	(0.018)
	[0.00]	[0.00]	[0.01]	[0.00]	[0.10]
Panel B					
Maintenance only (T1)	4.645	0.746	0.023	0.019	0.008
	(0.081)	(0.045)	(0.025)	(0.031)	(0.022)
	[0.00]	[0.00]	[0.35]	[0.54]	[0.71]
Maintenance + sensitization (T2)	4.839	0.776	0.083	0.160	0.053
	(0.074)	(0.047)	(0.021)	(0.029)	(0.020)
	[0.00]	[0.00]	[0.00]	[0.00]	[0.01]
T1 = T2 (p-value)	0.063	0.636	0.009	0.000	0.042
Mean (control group)	0.315	0.063	0.646	0.327	0.660
Std. dev. (control group)	0.358	0.025	0.478	0.469	0.474
Observations	560	560	4844	3323	4793
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 (5) in panel A, and equation (6) 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 (5) in panel A, and equation (6) 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 6 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 strata indicators for the city and the provider of the CT.

D.7 Quality and inputs of service delivery: construction and effects

To construct a measure capturing the overall **quality of service delivery**, we use all the observed indicators related to the facility's structural quality and cleanliness, and to the lack of harmful bacteria. We build the index using item response theory (IRT), a technique used to describe the relationship between individual responses to questionnaire items and an unobserved latent trait (Gordon et al., 2012; Kline, 2014). We build the index using a two parameter IRT model with the two parameters being an ability score, which could be used as a weight in constructing the index, and a discrimination score, which measures how well the indicator differentiates between low- and high-quality. The index is re-scaled to be between 0 (lowest quality) and 1 (highest).[4] Table D9 provides the list of all indicators included. In addition, we build three separate indices using IRT to measure structural quality, visible cleanliness, and the lack of bacteria. Figure D5 shows the effect of the maintenance treatment on each component.

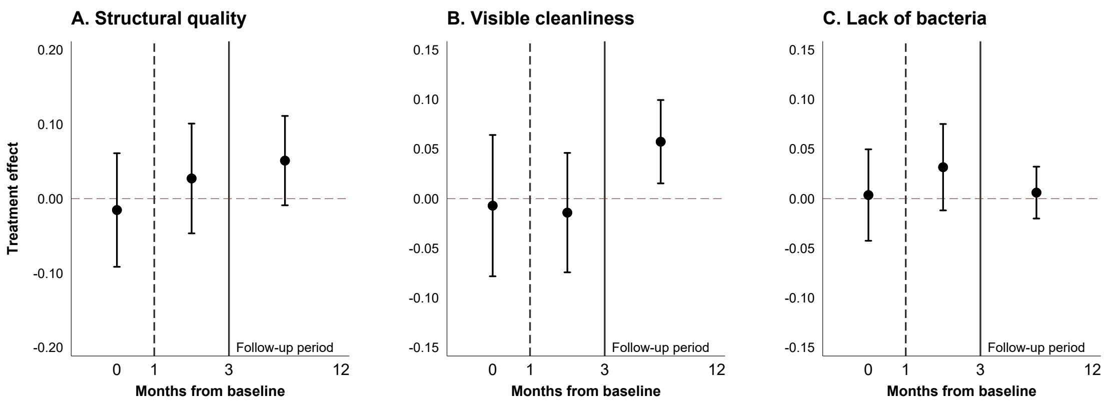
[4] We compute the index separately for baseline and for the subsequent surveys due to the fact that the baseline survey includes a lower number of indicators. At baseline, due to convergence, we adopt a one parameter IRT model.

Table 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 for computing the index. 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.

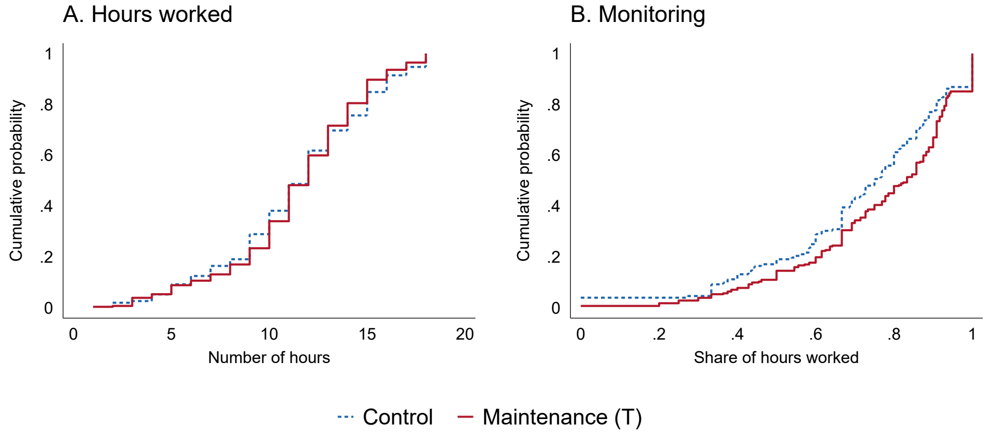
Figure D5: Effect on CT quality by component of the index



Note. Each panel presents estimates of treatment effects based on OLS regressions using equation (5) 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 strata indicators for the city and the provider of the CT.

Concerning instead caretaker's 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 of the share of time allocated to monitoring activities (Panel B), distinguishing between control and treatment group, while Table D10 shows estimates of treatment effects on the individuals indicators used to build the routine maintenance indicator used in the main text.

Figure D6: Caretaker's labour supply and time use



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	Disinfectants	Pressurized water	Safety equip-ment	Employed cleaners	Correct imple-menta-tion
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Maintenance (T)	-0.001	0.072	0.040	0.006	0.041	0.033	0.149	0.115
	(0.011)	(0.032)	(0.043)	(0.025)	(0.026)	(0.024)	(0.077)	(0.038)
	[0.92]	[0.03]	[0.35]	[0.80]	[0.12]	[0.17]	[0.05]	[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 (5). 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 strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix B.

D.8 Treatment heterogeneity by pre-specified dimensions

This section presents estimates of heterogeneous effects by a series of pre-registered variables. Table D11 presents an analysis of heterogeneity for CT- and caretaker-level outcomes. Tables D12 and D13 refer instead to respondent- and household-level outcomes. For continuous heterogeneity dimensions, we build two categories: (i) Lower (or shorter for distance), when the variable is smaller than or equal to the sample median; (ii) Higher (or longer for distance), otherwise.

Table D11: Heterogeneity by catchment area or CT characteristics

	Effect of maintenance treatment, by category						Het. test
	β	se	N	β	se	N	p-value
Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. WTP in catchment area		**Lower**			**Higher**		
Quality	0.08**	0.04	219	0.07**	0.03	215	0.79
Maintenance: cleaning	0.10***	0.03	219	0.02	0.02	215	0.01
Maintenance: rehabilitation	0.01	0.06	219	-0.07	0.07	215	0.39
Monitoring	0.09**	0.04	219	0.04	0.04	215	0.37
Share of users paying	0.10	0.07	220	0.08	0.05	214	0.83
Share of residents with positive WTP	-0.03	0.03	114	0.02	0.03	108	0.26
Users	0.93	2.24	220	-3.86*	2.29	214	0.13
B. Quality of the service		**Lower**			**Higher**		
Quality	0.07**	0.03	236	0.04	0.03	198	0.58
Maintenance: cleaning	0.05**	0.02	236	0.05**	0.02	198	0.94
Maintenance: rehabilitation	0.07	0.07	236	-0.15**	0.07	198	0.03
Monitoring	0.05	0.05	236	0.07*	0.04	198	0.70
Share of users paying	0.09	0.06	236	0.08	0.06	198	0.89
Share of residents with positive WTP	-0.02	0.03	120	0.00	0.03	102	0.67
Users	-0.61	1.96	236	-3.57	2.65	198	0.37
C. Users		**Lower**			**Higher**		
Quality	0.01	0.04	164	0.09***	0.03	270	0.09
Maintenance: cleaning	0.04	0.03	164	0.06***	0.02	270	0.61
Maintenance: rehabilitation	-0.03	0.08	164	-0.01	0.07	270	0.85
Monitoring	0.03	0.04	164	0.07	0.04	270	0.66
Share of users paying	0.09	0.06	163	0.10*	0.05	271	0.85
Share of residents with positive WTP	0.02	0.04	82	-0.02	0.03	140	0.43
Users	-5.67*	3.27	163	0.11	1.79	271	0.15
D. Payment		**Lower**			**Higher**		
Quality	0.07**	0.03	213	0.06*	0.03	221	0.91
Maintenance: cleaning	0.07***	0.02	213	0.05**	0.02	221	0.57
Maintenance: rehabilitation	0.04	0.08	213	-0.08	0.07	221	0.29
Monitoring	0.09	0.06	213	0.03	0.03	221	0.34
Share of users paying	0.12**	0.06	213	0.08	0.05	221	0.70
Share of residents with positive WTP	0.02	0.03	108	-0.02	0.03	114	0.42
Users	-1.13	2.38	213	-2.93	2.19	221	0.58
E. Caretaker's pro-social motivation		**Lower**			**Higher**		
Quality	0.06*	0.03	211	0.05	0.04	223	0.87
Maintenance: cleaning	0.04	0.02	211	0.07***	0.02	223	0.35
Maintenance: rehabilitation	0.10	0.08	211	-0.17**	0.06	223	0.01
Monitoring	0.05	0.06	211	0.06*	0.03	223	0.92
Share of users paying	0.10	0.07	212	0.07	0.05	222	0.63
Share of residents with positive WTP	0.04	0.03	108	-0.05	0.03	114	0.05
Users	1.41	2.62	212	-4.92**	2.08	222	0.06

Note. Categories for heterogeneity analysis are defined at baseline, with *lower* (*higher*) indicating whether the variable is smaller than or equal to (larger than) the sample median. Variables referring to catchment areas are averages of the corresponding variable within the catchment area. In columns (1)–(6), estimates are based on CT- or caretaker-level OLS regressions using equation (5) separately for each category. Column (7) presents a heterogeneity test based on CT- or caretaker-level OLS regressions using equation (5) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. The p-value is relative to the significance of the coefficient on the interaction term. Standard errors clustered by catchment area. The dependent variables are indicated in the rows and are defined in Appendix B. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1. *Caretaker's pro-social motivation* is the share of the endowment that is donated by the caretaker in the adapted dictator game (see Appendix F).

Table D12: Heterogeneity by individual characteristics: household-level outcomes

	Effect of maintenance treatment, by category						**Het. test**
	β	se	N	β	se	N	p-value
Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. WTP for service use		**Lower**			**Higher**		
Number of uses (regular users)	-0.01	0.07	1103	-0.09	0.06	1314	0.32
Number of uses (other users)	-0.41***	0.13	396	-0.10	0.12	487	0.06
Morbidity	0.02	0.04	1507	0.04	0.04	1816	0.70
Curative exp. (extensive)	0.05	0.04	1495	0.03	0.04	1803	0.72
Curative exp. (intensive)	320.84	337.49	1495	-221.94	287.70	1803	0.20
Preventive exp. (extensive)	-0.01*	0.01	1507	0.00	0.00	1816	0.04
Preventive exp. (intensive)	-29.04	81.06	1507	58.55	77.28	1815	0.33
WTP among residents	-0.00	0.11	2627	0.02	0.10	3374	0.90
Practiced OD	0.13	0.10	351	0.21*	0.11	466	0.63
B. Awareness of externality		**Lower**			**Higher**		
Number of uses (regular users)	-0.09	0.09	665	-0.04	0.06	1752	0.68
Number of uses (other users)	-0.29**	0.12	222	-0.22**	0.11	661	0.97
Morbidity	0.08*	0.04	893	0.01	0.04	2430	0.27
Curative exp. (extensive)	0.05	0.04	886	0.03	0.03	2412	0.60
Curative exp. (intensive)	250.15	440.75	886	-63.57	230.21	2412	0.49
Preventive exp. (extensive)	-0.01	0.01	893	-0.00	0.00	2430	0.66
Preventive exp. (intensive)	22.15	82.76	892	15.79	70.98	2430	0.97
WTP among residents	0.03	0.15	1622	-0.00	0.09	4379	0.80
Practiced OD	0.22*	0.12	200	0.15*	0.09	617	0.79
C. Trust in the community		**Lower**			**Higher**		
Number of uses (regular users)	-0.03	0.06	1899	-0.17*	0.10	518	0.17
Number of uses (other users)	-0.34***	0.11	628	-0.01	0.16	255	0.13
Morbidity	0.03	0.03	2545	0.04	0.04	778	0.93
Curative exp. (extensive)	0.04	0.03	2529	0.05	0.05	769	0.92
Curative exp. (intensive)	-17.13	256.49	2529	230.17	445.08	769	0.67
Preventive exp. (extensive)	-0.01	0.00	2545	0.01	0.01	778	0.07
Preventive exp. (intensive)	0.42	67.44	2544	83.04	123.05	778	0.49
WTP among residents	-0.01	0.10	4593	0.04	0.15	1408	0.67
Practiced OD	0.16*	0.08	609	0.21	0.18	208	0.79
D. Distance to CT		**Shorter**			**Longer**		
Number of uses (regular users)	-0.12**	0.05	1242	0.01	0.08	1175	0.20
Number of uses (other users)	-0.08	0.13	385	-0.31**	0.12	498	0.18
Morbidity	0.05	0.04	1639	0.01	0.04	1684	0.36
Curative exp. (extensive)	0.04	0.04	1628	0.04	0.03	1670	0.99
Curative exp. (intensive)	-402.77	350.54	1628	440.03	266.99	1670	0.05
Preventive exp. (extensive)	-0.01	0.01	1639	-0.00	0.00	1684	0.63
Preventive exp. (intensive)	-6.97	83.78	1638	41.44	80.02	1684	0.66
WTP among residents	-0.13	0.12	2948	0.16	0.10	3053	0.04
Practiced OD	0.22**	0.10	441	0.10	0.12	376	0.43

Note. Categories for heterogeneity analysis are defined at baseline, with *lower* (*higher*) indicating whether the variable is smaller than or equal to (larger than) the sample median. In columns (1)–(6), estimates are based on respondent- and household-level OLS regressions using equation (5) separately for each category. Column (7) presents a heterogeneity test based on CT-level OLS regressions using equation (5) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. The p-value is relative to the significance of the coefficient on the interaction term. Standard errors are clustered by catchment-area–round of observation. The dependent variables are indicated in the rows and are defined in Appendix B. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Specifications where the level of analysis is the respondent also include gender. Statistical significance denoted by *** p<0.01, ** p<0.05, * p<0.1. *Trust in the community* refers to the trust in the community to keep the CT clean.

Table D13: Heterogeneity by catchment area characteristics: household-level outcomes

	Effect of maintenance treatment, by category						Het. test
	β	se	N	β	se	N	p-value
Outcome variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Water quality		**Lower**			**Higher**		
Number of uses (regular users)	-0.03	0.08	1168	-0.08	0.07	1249	0.73
Number of uses (other users)	-0.17	0.14	443	-0.21	0.16	440	0.81
Morbidity	0.07	0.05	1620	-0.00	0.04	1703	0.24
Curative exp. (extensive)	0.06	0.04	1608	0.02	0.04	1690	0.40
Curative exp. (intensive)	608.29**	294.36	1608	-464.78	297.68	1690	0.02
Preventive exp. (extensive)	-0.00	0.01	1620	-0.00	0.01	1703	0.98
Preventive exp. (intensive)	145.51*	80.40	1620	-99.13	93.05	1702	0.05
WTP among residents	0.03	0.11	2931	-0.01	0.13	3070	0.86
Practiced OD	0.19*	0.11	409	0.14	0.12	408	0.66
B. Quality of the service		**Lower**			**Higher**		
Number of uses (regular users)	-0.06	0.07	1277	-0.07	0.08	1140	0.78
Number of uses (other users)	-0.03	0.13	437	-0.38***	0.13	446	0.10
Morbidity	0.03	0.04	1728	0.02	0.05	1595	0.93
Curative exp. (extensive)	0.01	0.03	1716	0.08	0.05	1582	0.23
Curative exp. (intensive)	-434.02	295.00	1716	489.64	329.67	1582	0.04
Preventive exp. (extensive)	-0.01	0.01	1728	-0.00	0.01	1595	0.60
Preventive exp. (intensive)	-35.71	89.05	1727	73.35	94.95	1595	0.43
WTP among residents	-0.11	0.11	3148	0.08	0.14	2853	0.30
Practiced OD	0.16*	0.08	413	0.22	0.14	404	0.66
C. Payment		**Lower**			**Higher**		
Number of uses (regular users)	-0.15*	0.08	1226	0.03	0.07	1191	0.13
Number of uses (other users)	-0.17	0.15	403	-0.25*	0.13	480	0.72
Morbidity	-0.02	0.04	1643	0.08*	0.04	1680	0.11
Curative exp. (extensive)	0.05	0.04	1627	0.03	0.05	1671	0.74
Curative exp. (intensive)	254.04	322.26	1627	-151.50	317.52	1671	0.36
Preventive exp. (extensive)	-0.00	0.01	1643	-0.00	0.01	1680	0.72
Preventive exp. (intensive)	-21.23	94.45	1642	62.99	96.21	1680	0.57
WTP among residents	-0.04	0.14	2975	0.06	0.11	3026	0.57
Practiced OD	0.14	0.09	405	0.20	0.12	412	0.66
D. Caretaker's pro-social motivation		**Lower**			**Higher**		
Number of uses (regular users)	-0.05	0.08	1181	-0.04	0.07	1236	0.99
Number of uses (other users)	-0.29*	0.17	388	-0.16	0.13	495	0.45
Morbidity	0.02	0.05	1582	0.05	0.04	1741	0.71
Curative exp. (extensive)	0.03	0.05	1570	0.07*	0.04	1728	0.54
Curative exp. (intensive)	146.72	347.02	1570	-122.59	313.02	1728	0.56
Preventive exp. (extensive)	0.00	0.01	1582	-0.01	0.00	1741	0.34
Preventive exp. (intensive)	36.67	93.90	1582	17.65	89.99	1740	0.88
WTP among residents	0.16	0.11	2844	-0.14	0.13	3157	0.07
Practiced OD	0.22**	0.11	411	0.13	0.12	406	0.55

Note. Categories for heterogeneity analysis are defined at baseline, with *lower* (*higher*) indicating whether the variable is smaller than or equal to (larger than) the sample median. In columns (1)–(6), estimates are based on respondent- and household-level OLS regressions using equation (5) separately for each category. Column (7) presents a heterogeneity test based on CT-level OLS regressions using equation (5) and adding an interaction term between the treatment indicator T and an indicator variable for the first category. The p-value is relative to the significance of the coefficient on the interaction term. Standard errors are clustered by catchment area–round of observation. The dependent variables are indicated in the rows and are defined in Appendix B. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Specifications where the level of analysis is the respondent also include gender. Statistical significance is denoted by *** p<0.01, ** p<0.05, * p<0.1. *Caretaker's pro-social motivation* is the share of the endowment that is donated by the caretaker in the adapted dictator game (see Appendix F).

D.9 Effects on other behavioral measurements

Among caretakers and among residents, we played two adapted dictator games to measure pro-social motivation for the cause among caretakers and residents' willingness to contribute to the quality of CT service among residents. These games are described in Appendix F. Columns (1)–(4) in Table D14 shows estimates of treatment effects on the behavior measured by the modified dictator game and additional outcomes obtained from the WTP game described in Section 5.2. In addition, using the information from the list randomization questions, columns (5)–(6) shows estimates of the treatment effects on the share of study participants that, the day before the interview, used the CT and washed hands with soap hand-washing with soap. Appendix F describes how list randomization questions were collected.

Table D14: Other behavioral measurements

Dep. var.:	**Caretaker**		**Residents**			
	Pro-social motivation for the cause	WTP is positive	WTP is equal or larger than market price	Willingness to contribute to quality	Used CT	Washed hands with soap
	(1)	(2)	(3)	(4)	(5)	(6)
Maintenance (T)	-0.021	-0.010	-0.000	-0.004	0.114	0.057
	(0.028)	(0.024)	(0.013)	(0.009)	(0.109)	(0.076)
	[0.46]	[0.68]	[0.99]	[0.67]	[0.30]	[0.45]
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 (5). Standard errors clustered by catchment area are reported in parentheses and p-values in brackets. Dependent variables by column: (1) *Pro-social 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 is equal or larger to market price*, 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 strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix B.

D.10 Payment enforcement and revenues

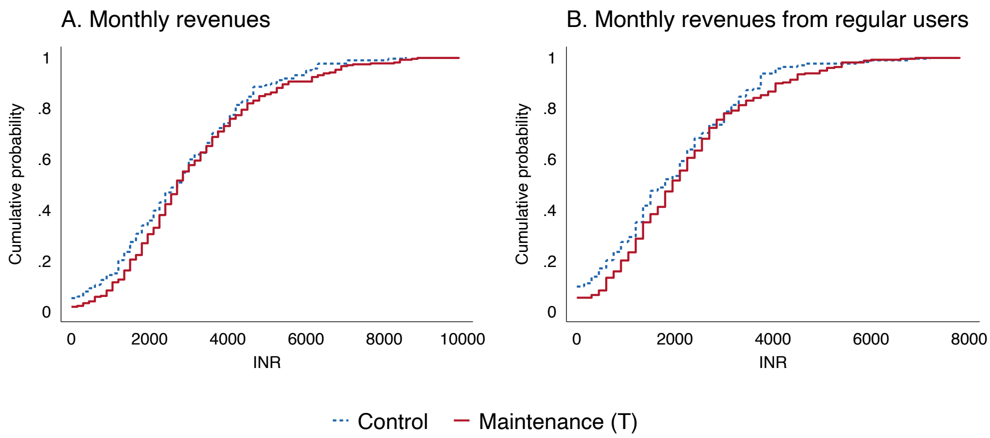
Columns (1)–(4) in Table D15 present treatment effects for different indicators of payment enforcement, reported by the sample of residents. Columns (5)–(6) provides instead estimates of treatment effects on monthly revenues estimated using observation during the rush hour. Revenues are imputed using information from observers about the number of people using the CT and the share of them who is paying the fee (assuming a standard fee of INR 5). Figure D7 shows cumulative distribution functions of these measures of service revenue, distinguishing by treatment group.

Table D15: Payment enforcement and service revenues during rush hour

Dep. variable:	**Caretaker ever refused entry**		**Refused entry for not paying**		**Monthly revenues**	
Sample of CTs:	All	Low payment	All	Low payment	Total	From regular users
	(1)	(2)	(3)	(4)	(5)	(6)
Maintenance (T)	0.015	0.050	0.006	0.047	269.915	246.780
	(0.020)	(0.024)	(0.020)	(0.022)	(253.857)	(184.871)
	[0.45]	[0.04]	[0.78]	[0.04]	[0.29]	[0.18]
Mean (control group)	0.076	0.044	0.074	0.041	2840.260	1954.870
Observations	1641	812	1641	812	434	434
Catchment areas	109	53	109	53	110	110
Observation rounds	1	1	1	1	4	4

Note. In columns (1)–(4), estimates based on household-level OLS regressions using equation (5). In columns (5)–(6), estimates based on CT-level OLS regressions using equation (5). 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 strata indicators for the city and the provider of the CT. Additional 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.11 Use of the service and self-reported use of the outside option

Table D16 presents estimates of treatment effects on payment for and use of the service, distinguishing between users that are residents of the slum and other users. Table D17 shows instead estimates of treatment effects on self-reported use of the outside option at individual level. Information for the spouse is reported by the respondent.

Table D16: Effects on use and payment for the service, by resident status

Dep. variable:	**Share of users paying**		**Users**	
Type of users:	*Residents*	*Other*	*Residents*	*Other*
	(1)	(2)	(3)	(4)
Maintenance (T)	0.103	0.020	-2.132	0.191
	(0.044)	(0.023)	(1.380)	(0.883)
	[0.02]	[0.39]	[0.13]	[0.83]
Mean (control group)	0.489	0.920	27.519	6.383
Observations	434	337	434	434
Catchment areas	110	107	110	110
Observation rounds	4	4	4	4

Note. Estimates based on CT-level OLS regressions using equation (5). Standard errors clustered by catchment area are reported in parentheses. The p-values 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 6 for details). Dependent variables by column: (1)–(2) *Payment*, observed share of users who pay the entry fee; (3)–(4) *Users*, total number of users observed. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix B.

Table D17: Self-reported use of the outside option

	Number of times [person] defecated not in the CT			
	Regular CT users		Other households	
	Respondent	Spouse	Respondent	Spouse
	(1)	(2)	(3)	(4)
Maintenance (T)	0.033	0.044	0.111	0.037
	(0.044)	(0.044)	(0.099)	(0.112)
	[0.45]	[0.32]	[0.27]	[0.74]
Mean (control group)	0.221	0.188	0.800	0.847
Observations	2417	1456	883	569
Catchment areas	109	109	102	96
Observation rounds	2	2	2	2

Note. Estimates based on household-level OLS regressions using equation (5). Standard errors clustered by catchment area are reported in parentheses and p-values in brackets. Dependent variables by column: (1)–(4) is the number of times the person defecated not using the CT the day before the interview (by demographic group). Information for the spouse is reported by the respondent. All specifications include indicator variables for data collection rounds, and strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix B.

D.12 Free tickets versus cash

To test whether obtaining tickets for free stimulates usage, we exploit variation stemming from the distribution of tickets for free CT use as part of the incentivized WTP measurement (see Section 5.2). We estimate a reduced form regression at the household level on the number of times the respondent used the CT at time t on $\tilde{c}_{i,t}$ using the following specification:

$$use_{ij,t} = \lambda_0 + \lambda_1\, tickets_{ij,t-1} + \lambda_2\, WTP_{ij,t-1} + \Omega_t + \epsilon_{ij,t} \tag{7}$$

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 D18 presents the results. Columns (1)–(2) focus on regular users, while columns (3)-(4) on other

residents. Columns (2) and (4) add the treatment indicator. The distribution of tickets rather than cash does not create a significant difference in the use of the service as measured in the following visit.

Table D18: 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)	0.027	0.021	0.163	0.172
	(0.071)	(0.071)	(0.146)	(0.143)
	[0.70]	[0.77]	[0.27]	[0.23]
WTP (previous visit)	0.009	0.010	0.019	0.014
	(0.013)	(0.013)	(0.028)	(0.026)
	[0.47]	[0.43]	[0.50]	[0.60]
Maintenance (T)		-0.107		-0.265
		(0.051)		(0.107)
		[0.04]		[0.01]
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 (7). 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 strata indicators for the city and the provider of the CT. Additional details about the variables are presented in Appendix B. The measurement of WTP is described in Section 5.2.

D.13 Selection in sanitation behavior

Table D19 shows the correlates of changes in sanitation behavior in the maintenance treatment group.

Table D19: Selection in sanitation behavior between baseline and follow-up 4

Dep. variable:	**Stopped using CT**	**Reduced CT uses**
	(1)	(2)
Household head is male	-0.118***	0.019
	(0.039)	(0.044)
Age of household head	0.001	-0.001
	(0.001)	(0.001)
Household members	-0.032***	-0.004
	(0.012)	(0.009)
Muslim	-0.004	0.040
	(0.056)	(0.071)
General caste	0.042	0.032
	(0.082)	(0.080)
Asset index	-0.305**	0.055
	(0.150)	(0.127)
Access to private toilet	0.239***	0.034
	(0.072)	(0.066)
Distance to CT (meters)	0.001***	-0.001***
	(0.000)	(0.000)
Awareness of externalities	0.008	-0.002
	(0.038)	(0.040)
Observations	829	657

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 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 Intervention design

Maintenance intervention. The *grant* offered three packages of similar monetary value from which the caretaker(s) could select one. *Deep cleaning* includes septic tank sewage removal, unclogging latrines and sewerage pipes, and cleaning walls, floors and inside toilets. *Repairs* includes sanitation/water connection repairs and/or infrastructure refurbishment. *Cleaning tools and agents* included four pairs of gloves, five floor cleaners, four toilet disinfectants, five liquid soaps, four toilet-cleaning brushes, two wipes, four nose masks, two brooms, two bucket and mop sets, three detergents, two hand-washing dispensers, two dustpans and two dustbins, and a training on how to use them. Pictures of the CT area to be improved were taken before the work was done. Our partner FINISH arranged and supervised the work with an external contractor, which was used in all facilities, and implemented the trainings as a theoretical session followed by a practical session about cleaning practices.

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.

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

Sensitization campaign. It targeted all members of study households, in particular the heads of participant households and their spouses. The campaign was designed in conjunction with our partner FINISH and a local graphic designer. We provided key messages regarding the risks of unsafe sanitation behavior through different means. First, *door-to-door visits* (Figure E2) used a flip chart with cartoons and messages targeted at household heads and spouses. This session covered how OD affects your community, how OD affects your family, the benefits of using CTs, what can be done to make the CT better, and the rights when you pay the fee for using the CT.

Figure E2: Door-to-door campaign

A. Flip chart cover B. Delivery of the campaign

Note. Panel A shows the cover of the flip chart used to communicate key messages to residents in slums. It translates from Hindi as "Awareness campaign to encourage CT use and maintenance in India". Panel B shows a moment of the sensitization campaign (source: Morsel).

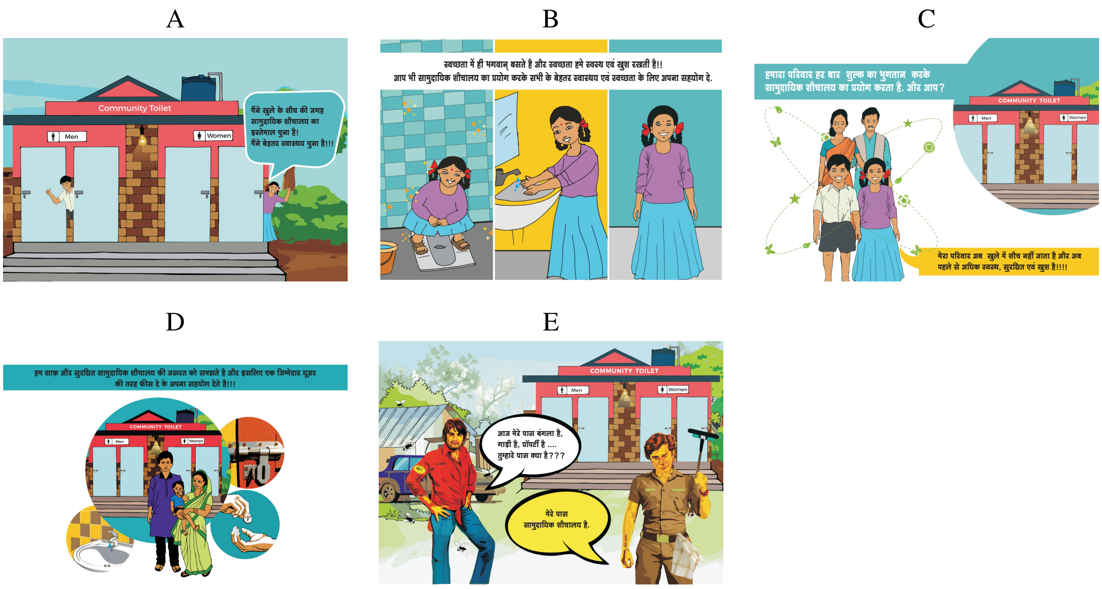
Second, the campaign's message was summarized into a four-page *leaflet* distributed among study households (Figure E3), and in a series of *posters* (Figure E4). We placed three medium-sized and two large posters in the entrance to CTs, in the area close to the hand-washing facilities and in each gender-specific area. Third, using a purposely designed tracking app pre-populated with all mobile phone numbers, we sent a total of 10 *voice messages* between month 1 and 11 of the study. Households listened on average to 7 messages. All study participants received the following voice message: *the CT is open from early morning until late evening*. Study participants in the maintenance treatment group received the message: *your CT has been granted aid to improve its quality. We hope you get to enjoy this better service.* Study participants in the maintenance plus sensitization group received the following messages: *Do you know OD is one of the biggest causes of diarrhea which can even kill your children? Adopting good sanitation behavior will ensure a healthier future for your family. / OD is a big risk for your family's as well as your neighbors' health. Use CTs to defecate instead of polluting and contaminating your community with OD. / Health is wealth! By not defecating in the open you are keeping your health safe and reducing expenses on medicines and treatment. / Cleanliness is godliness! By using CTs, you are contributing towards the cleanliness and health of your community. / Do you know how unsafe it is for women and girls in your family to go for OD? Be the change and adopt the use of CTs. / Using CTs ensures dignity of women in your community. Women should not feel ashamed of going to CTs...it is way better than OD. / Using CTs*

Figure E3: Leaflet circulated during the sensitization campaign

Note. The first page from the left presents the 'benefits of CTs' and includes: (1) improved sanitation facilities; (2) operation and maintenance of infrastructure; (3) safety with doors, locks and lights; (4) hand-washing facilities; and (5) gender-specific areas. The second page presents 'duties of users' and includes: (1) paying the fee to use the CT; (2) not throwing trash into the latrines; (3) flushing after using; (4) not spitting; (5) helping the elderly in the family; (6) accompanying females in the family during darkness; and (7) keeping the facility clean. The third page presents the 'rights of users' and includes: (1) caretakers not allowing free riders; (2) regular cleaning; (3) repairs; (4) respecting opening hours; (5) functional doors, locks and lights; (6) keeping men out of female areas; and (7) respecting and giving priority to females with children and the elderly. The front page of the leaflet, not presented in this figure, is the same as the one of the flip chart (Figure E2).

improves the health of your children and keeps medicines and doctors away.

Figure E4: Posters



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

E.2 Cost of interventions and quality scenarios

Table E1 presents a summary of the cost associated with each activity falling under the maintenance (panel A) and sensitization interventions (panel B). Note that these are total costs throughout the project. Based on input from our implementing partner FINISH Society, as well as Lucknow Municipal Corporation, Table E2 provides information on O&M costs for the median CT in our study sample (built 20 years before, having four female WCs, six male WCs and two urinals, and 150 daily users). A household has the potential to provide monthly revenues of at least INR 600 if all members over 5 years of age use the CT once per day and pay the market fee of INR 5.

Table 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.

The monthly maintenance cost for the current scenario (which we term as 'status quo') is INR 10,200 (US$ 144.85). Under the current scenario, salaries represent 78% of the total budget, and cover the costs for a full-time caretaker and for one cleaner performing a daily routine clean. We consider one alternative cost scenario that was deemed to support an 'improved' maintenance level. Under this scenario, we assume that the number of users remains constant. The scenario introduces a higher salary for the caretaker (which allows hiring a more experienced caretaker), higher input costs, and a yearly investment into cleaning machinery, such as a pressurized water cleaner, which costs about INR 20,000 (US$ 284.01). This scenario leads to a total of INR 28,800 (US$ 408.97) per month, with salaries representing 63% of the total. It is important to note that we do not claim that this scenario is optimal, and it can be improved further. The table also shows cost per eligible household (see 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 intervention expenditures of the maintenance intervention (Table E1) into monthly expenditures. Adding these costs, the total monthly costs become INR 13,544 (US$ 192.33) per CT.

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 and scripts

Bacteria presence. We gathered data about bacteria and mold presence using samples analyzed in the laboratory. For bacteria, we focus on the presence of: 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*.[5]

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

At baseline, we also collected information about residents' access to clean water. During the baseline survey, we asked households about their main source of water, and we then collected water samples from

[5]For information on the effect of bacteria on human health, see Jenkins and Maddocks (2019).

up to two randomly selected sources per catchment area. Figure A1 shows descriptive statistics at baseline for these measurements.

List randomization. The questionnaire for follow-up 4 was supplemented with a list randomization technique. Respondents were randomly allocated to one of four groups. Depending on the group, respondents received a different list of statements, and were asked to report how many of them were true. Table F3 provides the list of statements. Group A received only a list of statements related to general behavior. Groups B–D received the same list and one extra statement capturing sensitive behavior.[6]

Table F3: Statements used for list randomization

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

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

WTP for service use. WTP for service use is elicited to the respondent of the household survey and the spouse (up to two respondents per household) 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 all choices are made, one of the options is randomly selected by drawing a numbered ball from a bag, and the decisions are realized. In the case of the bundle of tickets being assigned, the respondent could allocate the 10 tickets or some of them to either male or female use. Before participating in the game, the participant was introduced to a practice round using a bar of soap to facilitate familiarity with the game. The exact explanation is the following:

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

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 it was very clean, had good hand-washing facilities, and

[6] 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."

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

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

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

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

> *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.

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

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

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

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

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