Journal of Development Economics 158 (2022) 102933

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Regular article

When nature calls back: Sustaining behavioral change in rural Pakistan ☆

Britta Augsburg [a], Antonella Bancalari [a,b,*], Zara Durrani [c], Madhav Vaidyanathan [c], Zach White [d]

[a] Institute for Fiscal Studies, UK
[b] School of Economics and Finance, University of St Andrews, UK
[c] Oxford Policy Management, UK
[d] GSMA, UK

ARTICLE INFO

JEL classification:
C93
I12
I15
I18
O18
Q53

Keywords:
Behavior
Sustainability
Basic services
Sanitation
Health
Maintenance

ABSTRACT

We implement a randomized controlled trial and a qualitative study to assess whether, and if so how, behavioral change can be sustained. We do so in the context of Pakistan's national sanitation strategy to combat open defecation, Community-Led Total Sanitation. Our findings demonstrate that continued follow-up activities that build on the original intervention lead to only modest reductions in reversal to unsafe sanitation on average, but gain in importance where initial conditions are unfavorable, i.e. poor public infrastructure and sanitation facilities. Promotion efforts are hence best targeted towards those who face larger difficulties in constructing and maintaining high-quality sanitation. The effects were sustained at least one year after the implementation of activities.

Long-term continued adoption of welfare-improving goods and practices is fundamental to achieving the Sustainable Development Goals (SDGs). A drop in adoption of, and reversal to, unsafe behavior is, however, common after achieving initial improvements (Dupas and Miguel, 2017). Promotion campaigns have for example proven effective at increasing learning and triggering behavioral change, but how this change can be sustained over time is less well understood (Madajewicz et al., 2007; Duflo et al., 2015; Banerjee et al., 2015; Tarozzi et al., 2021; Chong et al., 2020; Hussam et al., 2021). Studies on the effect of one-off community mobilization campaigns and commitment devices with no follow-up on sustained healthy behaviors yield mixed evidence (Cairncross et al., 2005; DellaVigna and Malmendier, 2006; Kremer and Miguel, 2007; Dupas, 2009; Banerjee et al., 2010; Bjorkman Nyqvist et al., 2017) and little is known about drivers of sustainability (Chirgwin et al., 2021).

In this paper we seek to help fill this knowledge gap. Using a complementary mixed-methods approach, combining a randomized controlled trial (RCT) with qualitative research, we study the effectiveness of continued community-based mobilization after initial behavioral change was achieved. We do so in the context of sustained adoption of health infrastructure, i.e. household sanitation, in rural Pakistan.

Although preventive behaviors and technologies are highly cost-effective, there is systematic underinvestment in preventive healthcare in low- and middle-income countries (LMICs) (Dupas, 2011). Sanitation is a particularly important example of underinvestment, in terms of both initial adoption and maintenance of facilities. Safe sanitation – the adequate disposal of human waste – has been recognized as an indispensable element of disease prevention and primary healthcare programs (e.g. the Declaration of Alma-Ata, 1978), and is included under SDG 6. Safe sanitation is also regarded as one of the most cost-effective health interventions (OECD, 2011; Hutton and Varughese,

☆ We gratefully acknowledge financial support from FCDO (at that time DfID), Contract: DFID 6507 "Monitoring, Verification and Evaluation Service Provider for the WASH Results Programme" and from the ESRC Centre for the Microeconomic Analysis of Public Policy (ES/T014334/1). The pre-analysis research design (documented in 'WASH Results Program: MVE Component – RCT Research Study Design Document: Understanding the role of follow-up (outcome phase) activities in achieving sustainable WASH outcomes') was approved by SEQAS, the UK Department of International Development (DFID)'s Independent Quality Assurance Service for Evaluation Outputs.

* Corresponding author at: Institute for Fiscal Studies, UK.
E-mail addresses: britta_a@ifs.org.uk (B. Augsburg), antonella.bancalari@ifs.org.uk (A. Bancalari), zara.durrani@opml.co.uk (Z. Durrani), madhav.vaidyanathan@opml.co.uk (M. Vaidyanathan), zwhite@gsma.com (Z. White).

https://doi.org/10.1016/j.jdeveco.2022.102933
Received 30 November 2021; Received in revised form 28 June 2022; Accepted 4 July 2022
Available online 12 July 2022

2016), even when the exact benefit–cost ratios are still the subject of debate (Bancalari, 2020; Whittington et al., 2020).

Despite its cost-effectiveness, 2.3 billion people lacked access to basic sanitation in 2015, when our study was conceptualized, a number that decreased but remains significant today (1.9 billion in 2020) (Joint Monitoring Program WHO-UNICEF, 2021).[1] This low coverage has been identified as one of the leading causes of malnutrition and early-life mortality (Prüss-Ustün et al., 2014, 2019). There is also growing recognition that initial uptake of safe sanitation behaviors is only part of the solution. Reversion, or 'slippage', back to unsafe sanitation behavior once access has been achieved can be substantial (Tyndale-Biscoe et al., 2013; UNICEF, 2014). Therefore, there is a call for research into how toilets can be maintained and rehabilitated to sustain long-term behavior change (Orgill-Meyer et al., 2019; Chirgwin et al., 2021).

Pakistan is a particularly relevant context in which to study sustained underinvestment in sanitation. With one-fifth of the rural population practicing open defecation (OD), the country is one of the largest contributors to OD worldwide. Pakistan is estimated to have lost 3.9% of GDP in 2017 alone due to premature deaths, healthcare treatment, and lost time and productivity induced by inadequate water and sanitation (World Bank, 2018).

Within this context, we leverage the recent completion of a Community-Led Total Sanitation (CLTS) campaign to study the effectiveness of follow-up activities in sustaining behavior change. Implemented in more than 60 Latin American, Asian and African countries, CLTS is widely used to promote first-time adoption of private toilets (also referred to as latrines). The approach forms part of at least 25 countries' national strategies to combat OD (Zuin et al., 2019), including Pakistan's 2006 National Sanitation Policy (Government of the Islamic Republic of Pakistan Ministry of Environment, 2006) and the later published Pakistan Approach to Total Sanitation (Government of the Islamic Republic of Pakistan Ministry of Environment, 2011).

CLTS entails the delivery of health promotion and hygiene messaging through community meetings and other participatory activities. The aim is to 'trigger' collective behavioral change and encourage households to construct and use toilets. The original proponents of CLTS placed a heavy emphasis on using psychosocial levers, particularly disgust and shame, in this process (Chambers and Kar, 2008). Other central tenets of the approach are that changes should be community-led and that the construction of household latrines should not be subsidized. CLTS activities center on promoting initial uptake rather than sustained change, and the approach has been criticized for a lack of attention on slippage back to OD. Based on pooling data from four RCTs evaluating CLTS in different contexts, Gertler et al. (2015) postulate that CLTS indeed increases the use of such household facilities. The authors are, however, not able to disentangle the effect of the initial triggering meetings from the follow-up activities.

In this study, we use an RCT to robustly delineate the impact of follow-up activities. 123 villages that had earlier undergone CLTS mobilization and triggering activities were randomly allocated to receive follow-up activities, or not. The study sample within these villages is what we refer to as CLTS 'beneficiary households', i.e. households that constructed a household toilet *after* CLTS triggering took place and *before* follow-up activities were randomized. These households were identified from CLTS implementers' monitoring data.

To provide context within which the intervention and the main impacts are embedded, and to understand mechanisms through which the intervention impacts emerge, we complemented quantitative with qualitative methods. To this end, our data collection consisted of two distinct stages following Tashakkori and Teddlie (1998) and Creswell et al. (2003).

In the first, quantitative, stage, study participants were surveyed twice: at baseline in 2016 (after CLTS triggering and before the implementation of follow-up activities) and two years later in 2018. Data were collected and main impacts were estimated before the start of the second, qualitative, stage. This second stage, which used a multiple case study design following Stake (1995) and Yin (2003), was implemented in four purposefully selected treatment and two control villages, and consisted of 21 focus group discussions (FGDs) and 32 individual interviews with village leaders, members of both beneficiary and non-beneficiary households, as well as program staff.

Survey instruments for both the quantitative and the qualitative study were designed to capture information on our primary outcomes – i.e. sanitation behavior – and potential drivers of and barriers to this behavior.

In the quantitative analysis, we first estimate the prespecified specifications, namely the intention to treat (ITT) average effects on OD and the use of a functional latrine (also using ANCOVA models). At this point, we also estimate heterogeneous ITT effects by geographic location. This heterogeneity dimension was perceived as a relevant one during the study's inception because, while CLTS was implemented by one local non-governmental organization (NGO), the implementation of follow-up activities was overseen by two different international NGOs. Randomization had therefore been stratified by the two provinces in which these international NGOs participated, Sindh and Punjab. These initial results further informed the qualitative survey instruments to allow for a more focused approach.

We next estimate heterogeneous ITT effects following steps closely in line with those outlined in Ferraro and Miranda (2013). We first test for heterogeneity with Crump et al. (2008)'s non-parametric approach, we then select a set of heterogeneity dimensions based on previous literature and policy relevance, and we finally identify the nature of the heterogeneity in a regression form using interaction terms. The purpose was to identify subgroups that are most responsive to the intervention and to provide guidance for policymakers to target the intervention in a more cost-effective manner (Heckman et al., 1997; Djebbari and Smith, 2008).

The qualitative data were analyzed following a thematic analysis approach (Ivankova et al., 2006), focusing on interpretation of recurring stories and experiences shared by participants to identify themes. The qualitative analysis and estimation of heterogeneous ITT effects were two independently conducted efforts. We bring the two methodologies together by using the qualitative evidence to support the findings from the heterogeneous analysis and gain a deeper understanding of the individuals' motivations.

Between the quantitative baseline and endline surveys, the share of CLTS beneficiary households in control communities that have at least one member over 5 years of age who goes for OD doubled up to 34% and that use functional latrines dropped from 86% to less than 70%. Such a significant drop in toilet usage has been observed in other contexts, such as India (Cairncross et al., 2005; Barnard et al., 2013; Orgill-Meyer et al., 2019). The last study reports that almost 20% of CLTS beneficiary households declared OD as their main practice both 4 and 10 years after intervention implementation.[2]

Over-time decrease in adoption of technologies is not a sanitation-specific phenomenon. For example, Banerjee et al. (2010) find that while 30% of children have at least one vaccination injection, only 1% received a full immunization course as most drop out after the initial

[1] Basic sanitation services are defined as those that hygienically separate excreta from human contact, but where excreta are not safely managed (Joint Monitoring Program WHO-UNICEF, 2017).

[2] On the other hand, Crocker et al. (2017a), who assess sustainability of CLTS (which included a component of follow-up activities) in Ethiopia and Ghana, find sustained impacts on OD in three out of four study cities. However, as acknowledged by the authors, these were assessed after one year and impacts could look different after a longer period.

three injections. Another example comes from Hanna et al. (2016), who find that adoption of eco-friendly cooking stoves declined markedly over time as households failed to make the necessary maintenance investments to keep them functional.

Households randomly allocated to follow-up activities, which were attended by household members and generated lasting awareness, were 6 percentage points (ppts) more likely to continue adopting safe sanitation behaviors. This effect is, however, only marginally significant, and how precisely it is estimated depends on the choice of specification.[3]

We find that the effectiveness of the follow-up activities differed by geographic location. Beneficiary households living in villages allocated to follow-up activities in the province of Sindh were, on average, 14 ppts less likely to revert to OD, representing a 25% lower OD relative to control beneficiary households at endline. We find a comparably greater (13 ppts) use of functional toilets in treated villages, indicating that the lower OD was primarily due to sustained operability of the facilities. The effects cannot be ascribed to pre-treatment differences in sanitation behavior nor to other observables across treatment arms, and we find that they are robust to different specifications. In contrast, effects for the province of Punjab are close to zero and insignificant.

Since provinces differ not only in terms of who is involved in implementation but also in intervention-relevant characteristics. Sindh is a poorer province and the one with higher initial levels of OD and in which reversal to OD was the greatest among control households. As such, we investigate the sustainability of ITT effects and heterogeneity in observed impacts focusing on the province of Sindh only. Doing so allows for a clean separation between program effectiveness and initial slippage to OD, geographical features and implementation specifics. However, our main conclusions are robust to conducting the analysis on the full sample. We find that the effect was sustained at least one year after the follow-up activities were implemented. We leverage variation in the time gap between households' last exposure to follow-up activities and the endline survey to stratify the treatment sample. We provide suggestive evidence that the effect on OD practices was almost twice as large one year after the activities finished as the average ITT effect. Qualitative evidence reveals that one driving factor is the volunteer staff within communities, who continued motivational engagement with households of their own accord.

We also demonstrate that follow-up promotion efforts are best targeted towards those who are most likely to slip back into unsafe behavior. We find that the material circumstances in which people initially live (including lack of access to public infrastructure and services, unfavorable environmental conditions of neighborhoods, asset poverty and poor status of sanitation facilities) are all factors that enhance the effectiveness of follow-up activities. A 'joint analysis', where we include in the same regression the estimation of all heterogeneous effects, shows that follow-up activities are particularly effective in villages with poor infrastructure: beneficiary households are, on average, 28 ppts less likely to revert to OD than treated households in villages with better initial public infrastructure. Likewise, beneficiary households allocated to follow-up activities with poorer status of the sanitation and hygiene facilities at baseline were, on average, 16 ppts less likely to practice OD than treated households with better sanitation facilities. We also provide suggestive evidence that follow-up activities incentivized households living in these poor conditions to conduct regular maintenance of their latrines, as well as leveraged better latrine technologies and improved social norms to keep individuals away from harmful sanitation practices.

When comparing our results with the literature on the impact of CLTS campaigns, our average ITT effect estimated in Sindh is comparable to approximately half the most optimistic impact that CLTS programs have been found to achieve, suggesting this is a worthwhile addition. Experimental evidence of the effectiveness of CLTS campaigns in sanitation behavior, however, yields mixed results, with the most optimistic point estimate being a decrease in OD and increase in toilet ownership by 25 ppts in rural villages of Mali (Pickering et al., 2015).[4] The effectiveness CLTS campaigns with limited follow-up support and monitoring has also been found to be greatest where initial conditions are unfavorable (Abramovsky et al., 2018).

1. Context and intervention

Pakistan, the world's fifth most populous country, has made significant progress in improving access to clean water, decent toilets and good hygiene. However, despite the improvements made, the challenge remains substantial. At the time of this study's baseline survey, in 2016, over 24 million people in the country practiced OD (Joint Monitoring Program WHO-UNICEF, 2017).

The study is situated in the provinces Sindh (district Badin) and Punjab (districts Bahawalpur and Rahimyar Khan). In 2014, before the onset of our study, 38% of rural households used improved sanitation facilities in Sindh, compared with 57% in Punjab (Sindh Bureau of Statistics and UNICEF, 2014; Bureau of Statistics, Planning and Development Department, Government of Punjab, and UNICEF, 2014).[5]

Sindh is generally poorer and less developed than Punjab also on a number of other dimensions, and study districts are below their respective province averages. For example, living standards as measured on an index scale from 0 to 100 (capturing access to clean water, clean fuel, electricity, adequate sanitation, roof quality and basic household assets) were significantly lower in Badin, with a score of 31.1, than the province and country averages of 67.6 and 74.5 respectively. Punjab is generally better off with a score of 83.0, but Bahawalpur and Rahimyar Khan are in line with the country average at 77.5 and 75.2 respectively. Similarly, the Human Development Index (HDI) was estimated at 0.41 in Badin (0.64 in Sindh), 0.645 in Bahawalpur and 0.625 in Rahimyar Khan (0.732 in Punjab) and 0.68 in Pakistan as a whole (UNDP, 2017). This puts Badin below the sub-Saharan Africa HDI average (0.52, 2015 data) and below the average of the least developed countries (0.51) (UNDP, 2017). A map of the location of the study is shown in Figure A1 in the online appendices.

CLTS is a widely used community-level information and mobilization intervention across low- and middle-income countries. It is typically implemented in three steps. As a first step, the implementing agency enters the targeted village or community to build rapport and introduce the program. Next, leaders arrange a community meeting, which focuses on 'triggering' the intended beneficiaries to change their defecation behaviors. Activities conducted use the psychosocial levers of shame and disgust and collective peer pressure. They also highlight the relationship between OD and poor health. It is expected that attendees of the triggering meeting will draw up a community action plan to achieve open defecation free (ODF) status, typically posted in a public spot. The third step of CLTS, which is the focus of this study, draws on this action plan. So-called 'natural leaders', or Community Resource Persons (CRPs) as they were named in the program we study, are identified. The main task of the CRPs is to follow up on the

[3] Our study is underpowered to estimate statistical significance in such small effects, as both ex-ante and ex-post power calculations were based on a magnitude of 8 ppts.

[4] Guiteras et al. (2015) find no effect of CLTS on sanitation behavior in rural Bangladesh, Briceño et al. (2017) find a reduction of OD by 12 ppts and increase in toilet usage by 15 ppts in rural Tanzania, and Cameron et al. (2019) find an increase in toilet construction by 2.4 ppts in rural Indonesia.

[5] Throughout the paper, we follow the (Joint Monitoring Program WHO-UNICEF, 2017) classification that improved sanitation facilities are those designed to hygienically separate excreta from human contact, meaning that the excreta produced should either be: treated and disposed of in situ, stored temporarily and then emptied and treated off-site, or transported through a sewer with wastewater and then treated off-site.

commitments made. In our study, these CRPs received support from an NGO through trained Social Organizers (SOs).[6]

While the main steps of CLTS follow a specific pattern and include a commonly used set of activities (e.g. transect walks, mapping of defecation areas, calculation of how much feces enter the environment), the intensity and types of follow-up activities tend to vary by implementing agency and context.

The intervention we analyze in this paper was implemented by an NGO operating at national scale. Steps one and two of the CLTS campaign were classified as the first phase (Phase 1), implemented during 2014–15. The third step of the CLTS campaign was treated as a separate phase (Phase 2), implemented between 2016 and 2018.

The intervention studied was also unusual in that it was implemented within the context of one of the first large-scale payment-by-results (PbR) programs in sanitation. The local implementing NGO received funding from two international NGOs, which in turn received funding from the donor through the PbR modality. During Phase 1, payments were predominantly contingent on achievement of toilet construction targets. During Phase 2, payments were contingent on the degree to which constructed latrines remained functional and in use. The international NGOs faced the possibility of payments being disallowed if targets were not met. However, it is important to note that this PbR modality was not passed on to the local implementing NGO in their contract. As such, the implementing NGO did not face the possibility of payments being disallowed if certain targets were not met, though the international NGO managing the implementing NGO did have strong incentives to ensure these were met. A further important clarification on incentives is that the beneficiary households themselves did not receive any financial transfers, either for the initial construction of the latrine or for subsequently maintaining it as functional.

A program evaluation, commissioned by the funder, noted that the PbR modality had focused implementing agencies' attention on completing the follow-up activities and ensuring outcomes. Additionally, it was noted by the evaluators that this differed from many other sanitation programs implemented by NGOs:

'Most program log frames or results frameworks anticipate the delivery of both outputs and outcomes, but having a dedicated outcome phase ["Phase 2"] is rare and its inclusion was felt by many respondents to be a significant and positive feature [...] Furthermore, being held accountable for delivering against the outcome indicators helped to ensure that suppliers remained active in community support and engagement to facilitate the transition from outputs to outcomes up to the program end.' (White and Colin, 2019)

The aim of this paper is to evaluate the effectiveness of this second phase (the implementation of follow-up activities) in terms of sustaining improvements in sanitation practices.

Follow-up activities centered on the community, targeting all community members in treatment communities, as well as specific stakeholders (e.g. the village committee). Two types of community activities were implemented. First, broad-based community meetings (BBCMs) were held to assess general water, sanitation and hygiene (WASH) status within the community based on discussions promoted by local leaders and community walks. The aim of these gatherings was to review the progress on keeping the community open defecation free as per the routine workplans and to assign tasks designed during the CLTS campaign (Phase 1). Second, Health and Hygiene Sessions (HHS), aimed at maintaining awareness regarding personal, domestic and environmental hygiene conditions, were conducted door-to-door, as well as in community meetings. Besides delivering health and hygiene messages, the sessions provided hand-washing demonstrations and training on operation and maintenance of latrines. It should be stressed that there was a degree of continuity between Phase 1 and Phase 2 activities: though different in focus (continued use, rather than initial uptake), the core modality of delivery (community-level engagement by SOs and CRPs) remained the same (see Table A7 in the Appendix for more details).

The program's monitoring data confirm that these meetings were implemented as planned. That is, BBCMs were arranged quarterly and HHSs were conducted every 2–3 months (nine times over the project period), with average village attendance per session of 216 and 147 people respectively (76% and 52% of the village population). CRPs met SOs on a monthly basis, reviewing their progress as per their routine workplans and assigned tasks. Some additional capacity-building activities were implemented at the village level, focusing on disaster risk reduction, equity, inclusion, and sanitation operation and maintenance.

2. Methodology

Combining a randomized controlled trial (RCT) with qualitative research, we study the effectiveness of continued community-based mobilization after initial behavioral change was achieved.

We implemented a complementary mixed-methods approach to answer our research questions, with two distinct data collection stages (Tashakkori and Teddlie, 1998; Creswell et al., 2003; Onwuegbuzie and Johnson, 2006; Leech and Onwuegbuzie, 2009). The first stage was the implementation of the clustered RCT for which data collection happened in the first quarter of both 2016 and 2018. The second stage was the implementation of the multiple-case qualitative study, which went to the field in the second quarter of 2018 (see Table A8 in the Appendix for more details of the study timeline).

2.1. Quantitative methods

The quantitative data collection was conducted on 123 villages randomly chosen from the complete list of 307 villages in the provinces of Sindh and Punjab in which Phase 1 CLTS activities had been implemented.[7] Of these villages, 61 were randomly assigned to receive follow-up activities ('treatment villages').[8] The remaining 62 villages were assigned to a control arm, where no follow-up activities took place.[9] Table A1 in the online appendices gives a breakdown of sample size (both villages and households) by experimental arm and province.

A lot of effort went into ensuring adherence to the randomization. This included discussions with implementing staff at several levels and more detailed record-keeping on implementation as part of the regular monitoring activities, ensuring continued awareness about the treatment/control distinction. At study inception, a Memorandum of Understanding was signed between all implementing agencies and the research team outlining responsibilities on both sides. It was also agreed with the program funder that results in control villages would not be included in the assessment of supplier results and payments. As such, these villages were not included in the sample frame for the monitoring surveys of the implementing agency. The endline survey was timed to end three months prior to the end of the NGO's program to allow the NGO to carry out some follow-up activities in control communities. According to the records of the implementing agency and collaborating international NGOs, follow-up activities were indeed

[6] The Social Organizers were full-time NGO staff typically responsible for 25–30 communities.

[7] By villages, we refer to the 'database village', as defined by the implementing institution. This is in most cases what is referred to as a 'sub-village' within a 'revenue' village, which was the unit of intervention for the CLTS implementation and is also the lowest level geographical unit used in official statistics in Pakistan.

[8] The Phase 1 villages that were not selected to be included in the evaluation study also received follow-up activities.

[9] One of these program villages had fewer than 10 beneficiaries, so in order to reach our targeted beneficiary sample size we added an additional village to the sample. The village was taken from a back-up list that had been created at the same time as the main sample was drawn.

formally implemented in treated areas only.[10] Interviews with field staff confirm this.

We sampled study participants from the pool of 10,261 beneficiaries —i.e. households that were recorded to have constructed a household latrine during Phase 1 of the CLTS campaign, which make up on average 22% of the study villages' populations (31% in Sindh and 12% in Punjab). Program villages had on average 83 Phase 1 beneficiary households (45 in Sindh and 119 in Punjab) ranging from 8 to 1,337, with a median of 45 (10 to 212 in Sindh, median 44; 8 to 1,337 in Punjab, median 54); (see Figure A2 for the distribution of sampled households per village).

We used program monitoring data from the time of the baseline survey to randomly select a subset of 1,191 beneficiary households.[11,12] An important caveat of this sampling strategy is that the sample is only representative of CLTS *beneficiary* households that constructed latrines within the study areas, not of the general population living in these villages. We therefore cannot say anything about non-beneficiary households, such as potential spillover effects, nor can we take the proportion of beneficiary households within a community as a proxy of latrine coverage in study villages. This sampling strategy did not seek to be representative of village-level characteristics, but rather was designed to specifically evaluate sustained behavioral change after an initial uptake of latrines, and within the constraints the field experiment allowed for.

We conducted two survey rounds as part of the quantitative field experiment. A baseline survey was implemented in early 2016 — at the point at which the program switched from Phase 1 to Phase 2 activities. An endline survey was administered two years later, coinciding with the end of the program in 2018. We were able to re-interview 95% of the original households. As we show in Table A6, larger households and poorer ones (measured as those with a higher BISP – Benazir Income Support Program – Poverty score) were less likely to attrit. The attrition rate of 5% is balanced across treatment arms and several other baseline household characteristics.

Fig. 1 provides a graphical representation of the research design, highlighting some of the intervention features and key assumptions made. The *y*-axis represents the hypothetical percentage of beneficiary households defecating in the open. The *x*-axis represents time: point (1) is the start of Phase 1 (CLTS mobilization and triggering); point (2) indicates the start of Phase 2 (follow-up activities), which ends at point (3).

Point (2), which is midway through the intervention, is the start of the RCT Research Study and data collection. As such, our data hence do not allow us to provide quantitative estimates of Phase 1's impact as the counterfactual is unknown. However, as described above, our sample frame was the universe of households that constructed a toilet during Phase 1 of the program as reported by the implementing NGOs and verified by an independent agency. We do not know whether these households built a latrine because of the Phase 1 CLTS intervention, or whether they would have done so anyway absent the intervention.

At baseline, at least one household member practices OD (self-reported) in 17% of study households (30% in Sindh) while at least one uses a functioning latrine (self-reported complemented with latrine observation) in 83% of households. In line with this, 12% of beneficiary households indicated that they did not own a functional latrine. Assuming that these households had indeed all constructed a toilet during Phase 1, these figures indicate either that there was not complete behavior change among beneficiary households at the start of the study or that slippage back to open defecation had already taken place.[13]

Study households typically consist of seven members, equally split between males and females, and including on average one child under 5, as shown in Table A4. (See Table A3 for more information on how each variable is computed.) On average, household heads are in their early 40s and have 3 years of schooling; 90% of them are male. 90% of households own the dwelling in which they live. 18% of households have participated in a community project. Households in the province of Sindh are poorer on average than those in Punjab, as shown in Table A5: only 16% are asset rich compared with 46% respectively.[14] Among study households in Sindh, 26% have a sanitation facility observed in good status, 13% share the facility and 69% have a latrine of improved technology with a water seal.[15]

Furthermore, an average village in Sindh is poorer than the average for the whole study sample. 43% of villages have high availability of public infrastructure and 31% a good supply of sanitation goods and services.[16] 85% are vulnerable to weather shocks, defined as having had a flood/drought during the last two years. 87% of villages report that they have officially been declared ODF. In Sindh, with the exception of the age of the household head, there are no statistically significant differences between control and treatment groups.

We estimate the impacts of a CLTS campaign's follow-up activities on sanitation practices, using a cluster randomized assignment to treatment. We focus on the intention-to-treat (ITT) estimates.[17] Because

[10] The implementing agency also kept records about other agencies conducting activities in study areas and double-checked these with the government. No activities were recorded in Sindh. In Punjab, Lady Health Workers of the Health Department conducted standard visits delivering health messages. However, these were not specifically targeted at control communities only, but implemented routinely in villages.

[11] Ex-post power analysis based on a statistical power of 80 and 90%, and a significance level of 5%, assuming two-sided t-tests, suggest that with a sample of almost 1,200 households in 120 clusters, we are able to detect a minimum detectable effect (MDE) of 0.077 for our main outcome OD, which captures 'whether at least one household member aged 5 or above does not use the toilet the household has access to'. We note that this is closely in line with the power analysis conducted at study conceptualization, where the estimated MDE was 0.080. The ex-post MDEs are 0.078 for Sindh and 0.050 for Punjab.

[12] Baseline response rates were high. Only 6% of sampled households could not be interviewed, the vast majority due to agricultural seasonal migration. Where possible (in 88% of cases), these households were replaced by a randomly selected back-up household living in the same community.

[13] Other possible explanations for this finding are non-sample error in the survey implementation, including the misidentification of beneficiary households, and errors in the underlying data set of program beneficiaries. While these are possible errors, they are not seen as likely, or large enough to explain the degree of reversal. Great care was taken in identifying beneficiary households, which included working with SOs and CRPs to confirm households, and the program monitoring data were independently verified under the PbR modality.

[14] Asset rich is defined as above the median of the baseline distribution of the computed household asset index that includes a number of durable assets owned by the household. This reference is set from the baseline distribution including both provinces.

[15] Good status of sanitation facilities is defined as above the median of the baseline distribution of the computed sanitation status index that includes having a permanent superstructure, having a roof, having a form of closure, requiring repairs and being clean.

[16] High availability of public infrastructure is defined as a household living in a village with the value of the public infrastructure index higher than the median of the baseline distribution in the whole sample. This index includes a number of village-level variables, such as having paved roads, a primary school, a public hospital and an improved water source. Good supply of sanitation goods and services is defined as a household living in a village with the value of the sanitation supply index higher than the median of the baseline distribution. This index includes whether or not the village has access to a mason, access to a plumber, access to a cement block producer, access to a sanitary hardware store and access to a brick producer.

[17] Though conservative, an ITT analysis ensures that estimates are not subject to bias arising from selection, i.e. those who choose not to join activities may be different from other beneficiary households in the village. Further, ITT is an interesting parameter for the purpose of policymaking, since it is reasonable to assume that full compliance will never be achieved and therefore that the ITT is a better measure of the expected benefit of the program than a measure of impacts on those reached.

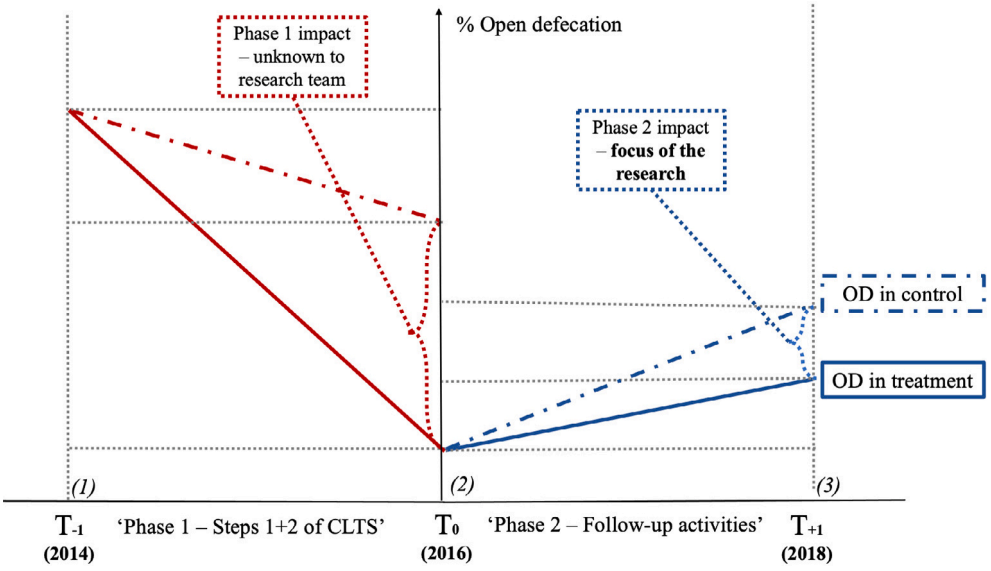

Fig. 1. Research design.
Notes. This figure presents a graphical representation of the research design. The *y*-axis represents the percentage of beneficiary households defecating in the open. The *x*-axis represents time: (1) start of Phase 1 (CLTS mobilization and triggering); (2) start of Phase 2 (follow-up activities), which ends at point (3). Point (2), which is midway through the program, is the start of the RCT Research Study and the start of data collection.

randomization was successful in creating observationally equivalent groups in the experiment, and attrition was random, we can estimate ITT effects by restricting the sample to post-baseline observations. We estimate ITT effects T_j on the sanitation practices Y_{ij}, which we also refer to as 'sanitation behavior', in household or individual i living in village j using the following ANCOVA specification:

$$Y_{ij} = \beta_1 T_j + \beta_2 Y_{ij}^0 + \epsilon_{ij} \quad (1)$$

where T_j is an indicator variable equal to 1 if village j is allocated to receive follow-up activities and 0 otherwise. Y_{ij}^0 is the baseline value of the outcome, added to increase statistical power (McKenzie, 2012). The error term ϵ_{ij} is assumed to be clustered by village. The parameter of interest, $\beta 1$, captures the average ITT estimate.

To investigate heterogeneous impacts, we expand the specification in Eq. (1) to:

$$Y_{ij} = \alpha_1 T_j + \alpha_2 T_j * d_j + \alpha_3 d_j + \alpha_4 Y_{ij}^0 + \epsilon_{ij} \quad (2)$$

where d_j is an indicator for the heterogeneity dimension measured at the village level at baseline (in some specifications, the dimension is measured at the household level). When the heterogeneity dimension is continuous, d_j is an indicator of whether the dimension in each household is above the median of the baseline distribution of the dimension. α_2 captures the heterogeneous effect.

In this analysis, we adjust *p*-values for multiple hypothesis testing using the bootstrap-based procedure proposed by List et al. (2019).[18] This procedure has been proven to asymptotically control the family-wise error rate (i.e. the probability of one or more false rejections) and to be asymptotically balanced (i.e. the marginal probability of rejecting any true null hypothesis is approximately equal in large samples).

2.2. Qualitative methods

The qualitative research study took place in the second quarter of 2018, intentionally sequenced *after* the completion of the endline data collection to avoid any risk of confounding the effects of the main treatment. It was designed at the village and household levels to provide context within which the intervention and the main impacts are embedded, and to better understand any underlying drivers focusing on knowledge, attitudes, practices, and determinants of sustained sanitation behavior change.

Data collection was conducted in a purposefully selected subset of four treatment and two control villages from the quantitative survey sample. Sampling for the qualitative component of the mixed-methods study was embedded in purposive extreme case sampling. Its logic lies in selecting information-rich case studies for in-depth study. The purpose of sampling in qualitative research is generally to look at variations and comparisons and less so to look at prevalence or population size. As such, a small number of cases (villages) were selected that maximized the range of variation on dimensions of interest to the RCT study while answering the 'whys' and 'hows' of the intervention impacts established through the RCT. The two primary indicators of interest used to shortlist program villages for inclusion as qualitative sample case studies were (1) the percentage of beneficiaries who openly defecate and (2) the percentage of beneficiaries who have a toilet. Based on these two indicators, communities at both extremes – high and low percentages – were selected to ensure a mix of responses, and breadth of themes to be covered. Control communities were selected to be geographically close to the chosen treatment communities, in order to ensure efficiencies in time and logistics and to allow better comparison across potentially similarly contextualized communities.

Semi-structured interviews were conducted by a local team fluent in the local languages. A total of 20 community focus group discussions (FGDs) across age and gender subgroups, 20 household-level in-depth interviews (IDIs) and 12 key informant interviews (KIIs) were completed with village leaders and CRPs. In addition, one FGD was conducted with district-level program staff prior to the village-level fieldwork.

Table A2 summarizes the qualitative sample reached as part of the research. Respondents for each instrument type (FGDs, IDIs and KIIs) were purposely selected once the team arrived in the sample villages, keeping gender, age group, preferred defecation practices, and knowledge of the local community in mind.

It is important to note that, while the quantitative survey focused only on the beneficiary households within study villages, the qualitative research was conducted with a wider group of respondents/households

[18] We do not adjust for multiple hypothesis testing in our main analysis as we are only considering functioning toilet ownership and OD behavior as outcomes.

in the study villages. As such, the qualitative data cover knowledge, attitudes, practices, and determinants of sanitation behavior change in both CLTS beneficiaries and non-beneficiary households. The advantage of this approach is that it situates the quantitative findings in the wider village context and helps us to draw conclusions (with care, given the sample size) that go beyond the beneficiary population in study villages.

The approach to analyzing the qualitative data was based on thematic analysis (Ivankova et al., 2006). Thematic analysis is an inductive approach to research that requires more involvement and interpretation from the researcher. Instead of a quantitative approach to analyzing qualitative data (such as frequency or cluster analysis), thematic analysis focuses on interpretation of the stories and experiences shared by participants in order to identify and examine themes as rigorously as possible. As such, the analysis began in the field during the daily debriefs.

Once fieldwork was completed, audio-recordings were transcribed and translated into English. Using the evaluation questions, as well as initial themes that emerged during the daily debriefs and an analysis workshop, a coding framework was developed to guide the initial stages of analysis. The coding framework, or 'node tree', comprises descriptive codes known as nodes and sub-nodes against which data from the KIIs, FGDs and observations could be organized according to emergent themes. Before starting to code, the qualitative research team discussed the nodes to ensure common understanding and to try to ensure consistency between each village. This joint process was key given that, as (Saldana, 2015) notes, 'all coding is a judgment call' and team members had to be cognizant of potential own biases and of the fact that this coding process is subjective to a degree.

While the qualitative analysis and the estimation of heterogeneous ITT effects were two independently conducted efforts, the two research teams brought the two methodologies together by using the qualitative evidence to support the estimated heterogeneous effects and gain a deeper understanding of the individuals' motivations.

Throughout this paper, we make use of illustrative quotes that were identified as helpful ones to explain the quantitative field experiment findings. Such quotes are used where a respondent, or set of respondents, clearly articulated a prominent point arising from the thematic analysis.

3. Average ITT effects

We focus on two key outcome variables. The first is whether at least one household member above 5 years of age practiced OD during the previous two weeks. Since this measure was self-reported and can suffer from reporting bias (Coffey et al., 2014), our second outcome is whether household members use a functional toilet located in the dwelling, which enumerators were able to corroborate. In every household surveyed, enumerators collected observations of the existence and status of the latrine in the dwelling.[19]

We first find that among beneficiaries of the Phase 1 CLTS campaign, reversal to OD was common. Concentrating first on the whole sample, Panel A of Fig. 2 shows that at endline, 34% of beneficiary households that were not allocated to receive follow-up activities ('control') reported OD practice (up from 17% at baseline). In line with this, Panel B shows that only 68% of control households reported using a functioning toilet at endline, down from 86% at baseline.

These statistics mirror the sentiment expressed in some of the qualitative interviews:

'There was a change [in defecation behavior], but I feel all that effort is now going to waste. Lots of people have stopped using latrines [...] Those whose latrines caved in with the rains or because they were not improved – all these people have reverted to open defecation. Yes, children, women, the elderly – everyone is included in this.' (Male Community Leader, KII, T)[20]

'Out of 70–80 households you can say that there only 10 or so in which there has been a change and they have started using latrines. The rest of them are using the same old method I have told you about (open defecation) [...] People say the programme was good, but that it should have run for longer in our village.' (Male, Community Leader, KII, C)

'Everyone has a latrine in their house, but some of these have caved in now [...] Only two or three houses in the community had installed improved latrines. (For the others,) Those whose latrines have collapsed may make them again, but for now they have reverted to open defecation in bushes outside their houses.' (Female, Community Leader, KII, C)

Panel A also indicates that OD rates at endline are lower in treated than in control communities, although only marginally so (30% for treated compared with 34% for control households). This is despite evidence of treatment fidelity. Follow-up activities significantly increased the likelihood of reported exposure: treated households are 20 ppts more likely to report having heard about and 17 ppts more likely to report having attended activities, compared with control households (see Table A9). 35% of households in the control villages report having heard about, and 25% having attended, any sanitation activities and promotion campaigns. Since these questions refer to *any* sanitation activity, not specifically the follow-up activities, it is expected that some activities would also be reported by control households as well. This could be due to either recall error or other promotion efforts.[21] Otherwise, it could be because CRPs continued to work of their own accord after the initial CLTS campaign had ended, as highlighted during qualitative interviews:

'The program started in 2014. Within six months we had been successful in constructing a latrine in every household. The SO worked here even after this [latrine construction] for a year, i.e. till 2016. We emphasized that people should keep their latrines and their homes clean and made them understand. The SOs stopped coming in 2016. But we [the CRPs] continued the work of our own accord. The [NGO] people have not contacted us since.' (Male CRP, KII, C)

Turning back to Fig. 2, Panels C to F indicate significant differences between provinces. While reversal to unsafe sanitation was significantly greater in Sindh than in Punjab (at endline, 57% of control households in Sindh reported OD practice, up from 26% at baseline, compared with only 13% in Punjab, up from 8%), differences between experimental arms at endline are also much more noteworthy in the former province, at 13 ppts (44% OD for treated and 57% for control households). In Punjab, the difference stands at only 3 ppts (16% versus 13%).

We next confirm these differences more formally in Table 1, which presents the estimated ITT effect of being allocated to receive follow-up activities on our two prespecified outcomes. We show estimates with and without accounting for the baseline value of the dependent variable (ANCOVA model). We also present estimates using the Post-Double Selection LASSO procedure (LASSO) for selection of baseline control variables (Tibshirani, 1996; Belloni et al., 2014).

We present the results for the full sample (Panel A), and when splitting the sample by province: Sindh in Panel B and Punjab in Panel

[19] Ideally, in the analysis of self-reported behavior, we would control for a proxy of social desirability. However, since these data are not available, we rely on the assumption that the follow-up activities did not introduce more response bias than the Phase 1 CLTS implementation might have already done. If this assumption is violated, our impacts will be upward biased. This bias, however, is not present when collecting observations of toilet ownership. Social desirability might affect the household's willingness to let the enumerator collect observations of their toilets, but such corroboration was possible in 99% of the cases (always possible in Sindh) and it is balanced across treatment and control.

[20] Information is given here on who provided the quote (male/female, CRP, Community Leader, individual respondent), during which type of interview (FGDs, IDIs, KIIs) and in which village (control (C) or treatment (T)).

[21] We aimed to reduce recall error by asking about activities in the last 12 months, rather than the whole two years that the Phase 2 intervention lasted. We note that none of the follow-up activities was more intense in the first than in the second year of implementation. Of course, attendance might have reduced over time. Unfortunately, we do not have this detailed information.

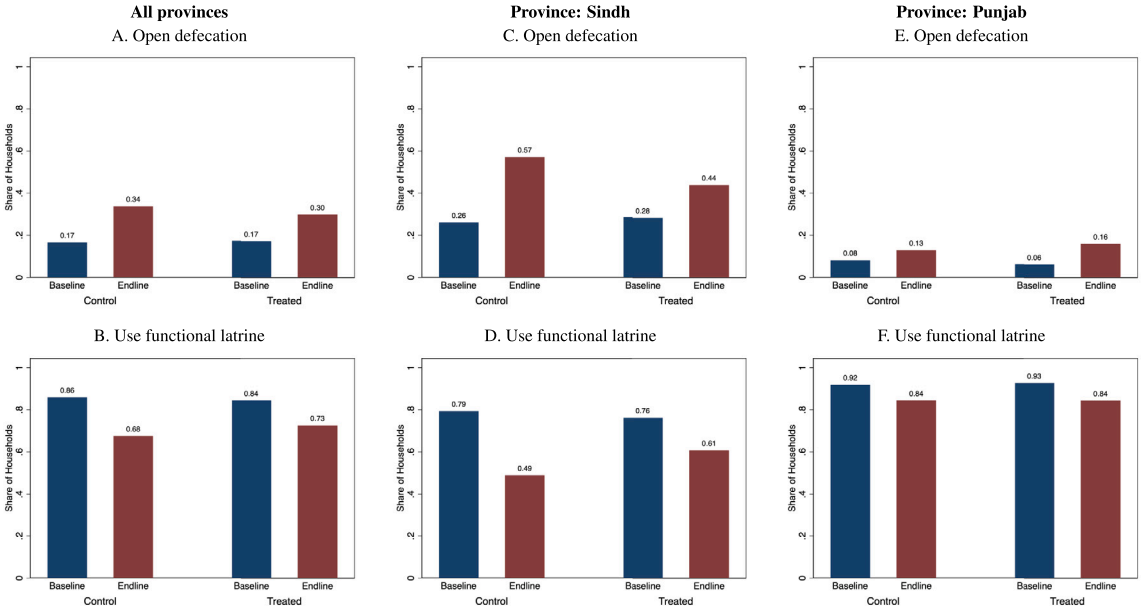

Fig. 2. Mean OD and functional latrine use levels at baseline and endline, by treatment arm.
Notes. 'Open defecation' (Panels A, C and E) is an indicator for whether a household reports at least one member older than 5 years who openly defecates. 'Use functional latrine' (Panels B, D and F) is an indicator for whether the household has a functional latrine in the dwelling and members use it.

Table 1
Effect of follow-up activities on prevalence of OD and use of a functional latrine.

Outcome	Open defecation			Use functional latrine		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Full sample						
Follow-up activities	−0.05	−0.05	−0.05	0.06	0.06	0.06*
	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
ANCOVA	No	Yes	Yes	No	Yes	Yes
LASSO	No	No	Yes	No	No	Yes
Control mean (EL)	0.34	0.34	0.34	0.68	0.68	0.68
Villages	123	123	123	123	123	123
Households	1,132	1,132	1,132	1,132	1,132	1,132
Panel B: Sindh sample						
Follow-up activities	−0.13	−0.14**	−0.16***	0.12	0.13*	0.16**
	(0.08)	(0.07)	(0.06)	(0.08)	(0.07)	(0.07)
ANCOVA	No	Yes	Yes	No	Yes	Yes
LASSO	No	No	Yes	No	No	Yes
Control mean (EL)	0.57	0.57	0.57	0.49	0.49	0.49
Villages	61	61	61	61	61	61
Households	551	551	551	551	551	551
Panel C: Punjab sample						
Follow-up activities	0.02	0.03	0.02	0.01	0.00	0.01
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
ANCOVA	No	Yes	Yes	No	Yes	Yes
LASSO	No	No	Yes	No	No	Yes
Control mean (EL)	0.13	0.13	0.13	0.84	0.84	0.84
Villages	62	62	62	62	62	62
Households	581	581	581	581	581	581

Notes. Primary data, household level. 'Open defecation' (columns (1), (2) and (3)) is an indicator for whether a household reports at least one member older than 5 years who openly defecates. 'Use functional latrine' (columns (4), (5) and (6)) is an indicator for whether the household has a functional latrine in the dwelling and members use it. 'Follow-up activities' is the β_1 parameter from Eq. (1). Panel A corresponds to the full sample of analysis. We additionally include province fixed effects in this panel. Panels B and C restrict the sample to the provinces of Sindh and Punjab, respectively. Standard errors (in parentheses) are clustered at the unit of randomization (villages). Statistical significance denoted by $^*p < 0.1$, $^{**}p < 0.05$ and $^{***}p < 0.01$.

C. We are able to do this split since the randomization was stratified by province. Randomization was successful in both provinces, as shown in Table A4.[22]

The estimated coefficients suggest a modest average reduction in reversal to OD of 5 ppts and an increase in the use of functional latrines by 6 ppts. How precisely the effect is estimated depends on the choice of estimation specification. The only significant positive impact we find is on the use of a functional latrine, which is statistically significant at the 10% level (Panel A, column (6), using both ANCOVA and LASSO).

Next, we examine heterogeneous ITT effects by our prespecified dimension, province. The table confirms the heterogeneity we noted in Fig. 2: estimates for Punjab (Panel C) suggest null impacts, while follow-up activities decreased the probability of OD significantly in Sindh (Panel B), and did so by 16 ppts among treated households, which translates into 28% lower OD compared with control villages at endline. The effect size is consistent across specifications, but more precisely estimated using ANCOVA and LASSO. The reduction in OD is accompanied by a comparable increase (of 16 ppts) in the probability of using a functioning toilet. This result suggests that control households were more likely to revert to OD, while treated households were more likely to keep their toilets functional. We find a big overlap between OD and the use of functional latrines, as 95% of households at baseline and 93% of households at endline that report not practicing OD were using a functional latrine. One Community Leader in a treated village stressed that as long as households have a functioning toilet, they stay away from OD:

'People who have functional latrines use them. Those whose latrines have fallen in either use other people's latrines or they go [i.e. open defecate] in the fields.' (Male Community Leader, KII, T)

There are three potential reasons for the null effects in Punjab and the significant effects in Sindh. First, the slippage to OD from the CLTS campaign in Phase 1 was very low in Punjab, with only 6%–8% of households reporting open defecation at baseline, compared with more

[22] There is an imbalance in the share of households with sanitation status above the median of the baseline distribution, but the difference in all baseline characteristics is not jointly significant.

than 25% in Sindh (as Fig. 2 shows). Arguably, follow-up activities in Punjab were thus less needed on average, while follow-up activities in Sindh acted as a much required 'booster' after the CLTS campaign in Phase 1. Second, Sindh is a poorer province than Punjab, as discussed in Section 2.1. Follow-up activities may be more effective where initial conditions are unfavorable. Third, the implementing agency collaborated with different international NGOs in the two provinces, which led to slightly different approaches used in the follow-up activities. While the core set of activities was the same, activities in Sindh included a 'capacity' component, consisting of training on disaster risk reduction and operations and maintenance.

The rest of the paper, which investigates the sustainability of ITT effects and heterogeneity in observed impacts, will focus on Sindh in order to allow for a clean separation between program effectiveness and initial slippage to OD, geographical features and implementation modality/partner.[23] We will, however, present results for the full sample in the Appendix for completeness and note that they confirm our main conclusions.

3.1. Sustainability of ITT effects

Follow-up activities are designed to address the sustainability of sanitation infrastructure and behavior. A natural question is for how long such activities are needed and /or whether they can be phased out. We can speak to these questions given significant variation in the time gap between completion of intervention activities and when the endline survey took place in the province of Sindh. Because the horizon of the study is short, however, we can only speak about sustainability of ITT effects in the short run (up to a year after the follow-up activities were implemented).

We measure the time gap as the distance between the last quarter of intervention and the time of the endline survey. Since the endline survey was conducted over four weeks, within the first quarter of 2018, the variation comes purely from differences in intervention implementation.[24]

The average gap is four and a half months, with a minimum of three months and a maximum of twelve months. The time gap was three months for 70% of treated households, six months for 16% of them and nine months or more for 14%.

Ideally, the time gap would be random across treatment communities. However, the implementation of activities was conducted in blocks of 10 villages and SOs visited the less remote ones first. In line with this, when analyzing selection into the time gap, we find that those with a three-month time gap had a dwelling of stronger material and access to better public infrastructure at baseline than those who experienced a longer time gap. Those with a six-month time gap were more likely to have a latrine of improved technology with a water seal than those with shorter or longer time gaps. The joint significance of baseline characteristics is balanced overall (see Table A11).

Because stratification was not conducted based on the time gap, we further check whether treatment and control villages are balanced within the sample of each time gap. We find that households surveyed six months after receiving the treatment were larger, had a more fragile dwelling and had fewer assets than control households. Those surveyed nine months or more after being treated were more likely to own their house, but it is more likely to be a fragile dwelling and have access to worse public infrastructure at baseline than control households. The joint significance of baseline characteristics, however, is balanced between treatment and control for every sample of analysis (see Table A10). Nevertheless, when estimating the ITT effects stratified by time gap, we control for the corresponding imbalances.

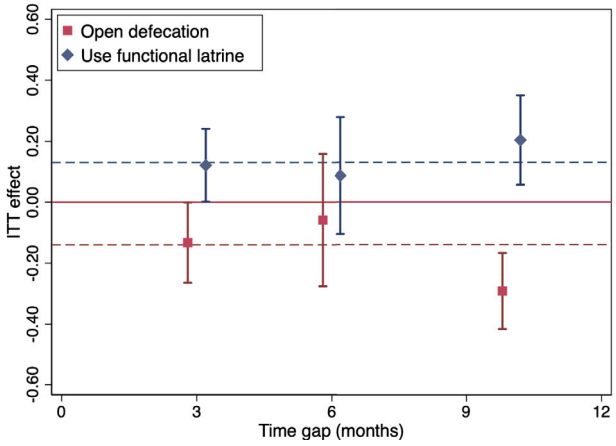

Fig. 3. Effects of follow-up activities on behavior, by time gap between the activities and outcome measurement.
Notes. 'Open defecation' is an indicator for whether a household has at least one member older than 5 years who openly defecates. 'Use functional latrine' is an indicator for whether the household has a functional latrine in the dwelling used by its members. The dotted lines show the average treatment effect (linear probabilities) for the whole sample. The plots indicate the point estimate of the effects of follow-up activities in each time gap and the 90% confidence interval. Time gap, indicated on the *x*-axis, is measured as the months elapsed between the end of follow-up activities and the measurement of outcomes. All estimations control for the baseline characteristics presented in Table A10. Sample includes households interviewed at endline in Sindh.

Fig. 3 shows the estimated ITT effects (using an ANCOVA specification) when stratifying the treatment sample by time gap between follow-up activities and endline survey in Sindh. Although the stratification makes the estimates predictably noisy, the results suggest that the reduction in OD due to follow-up activities is sustained over the first year. We also find suggestive evidence that the average ITT effects are even greater nine to twelve months after the intervention was finalized than they were three and six months after. The point estimate of the effect on OD in the 9+ months sample is twice as large as the average estimated ITT effect (dashed red line) of −0.14. However, we cannot reject the null hypothesis that the point estimates after nine months are statistically similar to the point estimates of the other time-gap samples. The positive ITT effect on the use of a functional latrine is also sustained over the first year.[25]

4. Heterogeneity

We next move beyond average ITT estimates to understand the heterogeneity of responses across villages and households. We are interested in what makes follow-up activities an effective strategy to sustain improvements in a village's sanitation environment. Because we find that reversal to unsafe sanitation behavior is reduced and that more toilets are kept functional in Sindh, our focus is on factors that are likely supporting the upkeep of latrines – both at the village level (Section 4.1) and at the household level (Section 4.2). For this, we estimate heterogeneous ITT effects by characteristics collected at baseline, after Phase 1 activities were completed and before Phase 2

[23] Not surprisingly given the stratification, randomization led to balanced baseline characteristics in the province of Sindh. In fact, the imbalance in the share of households with the sanitation status index above the median of the baseline distribution is not present in the Sindh sample. We still find an imbalance in the age of the household head, but the results remain robust when controlling for this characteristic (see Table A5).

[24] We cannot create a continuous measure of the time gap, since intervention implementation is reported only by quarter of the year. In Sindh, follow-up activities were implemented between the first quarter of 2016 and the last quarter of 2017. See Figure A3 for the full time-gap distribution.

[25] In Appendix Figure A4 we show the results for the full sample, including both Sindh and Punjab. Although the estimates are less precise, we also find that the point estimates are larger in magnitude for the 9+ months sample.

activities were started. Table A5 shows that the baseline heterogeneity dimensions considered are balanced between treatment and control in the Sindh sample. Moreover, Figure A6 shows that the distributions of the heterogeneity dimensions are very similar across treatment and control groups. This balance suggests no systematic bias when drawing inference from comparing outcomes in treatment and control villages within subgroups.

We are interested in exploring heterogeneity in ITT effects beyond the prespecified dimension (i.e. province). To reduce the chances of getting statistically significant differences across subgroups when no true treatment effect heterogeneity exists, we use four complementary steps akin to Ferraro and Miranda (2013).

First, we test for heterogeneity with Crump et al. (2008)'s non-parametric approach, but without yet characterizing the nature of it. We test for two null hypotheses: (i) the average ITT effect for the subgroup with X values of the heterogeneity dimension is equal to zero for *all values* of X; and (ii) the average (ITT) treatment effect (ATE) for the subgroup with X values of the heterogeneity dimension is equal to the ATE for *all values* of X. We test both whether the conditional average treatment effect is equal to zero (Zero CATE) and whether the conditional average treatment effect is constant (Constant CATE). Table A12 summarizes the results. We start by testing the hypotheses with a specification including all baseline characteristics from Tables A4 and A5 and their squares. The null hypotheses of Zero CATE and Constant CATE are rejected at the 1% significance level for the outcome 'Open defecation' and at the 5% significance level for 'Use functional latrine', implying that some subgroups may have average effects different from zero (see Panel A of Table A12).

Second, we select a set of heterogeneity dimensions both at the village level and at the household level, based on theory and policy relevance. We discuss this selection process in Sections 4.1 and 4.2. Using alternative methods of choice including all the selected dimensions, only the ones at village level or only those at household level, we always reject the null hypothesis of Zero CATE. We reject the null hypothesis of Constant CATE for six out of eight specifications (see Panels B to D of Table A12). Altogether, there is strong evidence that the follow-up activities generated heterogeneous ITT effects.

We then turn to identify the nature of the heterogeneity. The results of this analysis are presented in Sections 4.1 and 4.2. Before pursuing a parametric analysis with interaction terms, we plot mean outcomes for treatment and control villages along the distribution of each heterogeneity dimension. Finally, we identify differential effects in outcomes across subgroups through interaction terms between the treatment and dummy variables capturing whether the value of the heterogeneity dimension for a household or village is above the median of the baseline distribution. We estimate this using Eq. (2) and the ANCOVA model. This last step allows us to identify the subgroups that are most responsive to the treatment (Heckman et al., 1997; Djebbari and Smith, 2008). We also estimate marginal effects through interaction terms between the treatment and the continuous heterogeneity dimensions.

To mitigate even further the risk of finding statistically significant differences across subgroups when no true heterogeneous effect exists, we supplement standard inference with multiple hypothesis testing. In each table, we present the p-values for the significance of each individual coefficient and p-values adjusted for multiple hypotheses. The latter consider all hypotheses tested within a table.

4.1. Heterogeneity by village-level characteristics

There are a number of reasons why treatment effects may differ across villages. We focus on characteristics that can affect the ability and willingness to sustain healthy sanitation behaviors. Following Cameron et al. (2019), we first explore heterogeneity by accessibility of the village, such as having a paved road, and general public infrastructure quality. We build a 'public infrastructure index' that includes binary indicators for whether or not the village main road is paved, the village internal roads are paved, the village has access to a primary school, the village has access to a public hospital and the village has access to improved water.

Second, we examine differences by supply-side sanitation market access, following Guiteras et al. (2015). Supply failures such as lack of access to markets where toilet components are sold, or lack of information about repairs and maintenance methods, may impede sustained adoption of latrines. We build a 'sanitation supply index', which captures the availability of services needed to build latrines, including whether the village has access to a mason, a plumber, a cement and brick supplier, and a sanitation hardware store. See Table A3 for more details about the construction of the indexes and their components.

The study communities are usually farming and manual labor communities and the majority of the households with latrines just have simple pit latrine systems. Material availability is a challenge in terms of both access and affordability. The closest sanitation supplies are usually in the nearest urban center, and limited transportation is a pressing issue. People in some communities reported traveling three to four hours to the nearest city to buy sanitary materials. Accessibility of supplies and services is particularly challenging for poorer households without the necessary financial resources:

'We got the materials for the latrines from Badin [city]. Materials from there are usually brought in a rickshaw or Suzuki, so we used a rickshaw as well. It was very difficult. [T]here is no facility [i.e. sanitation hardware store] for getting supplies in our area.' (Male IDI respondent, C)

The qualitative interviews stressed that such poor access often affected both the quality of the initial sanitation infrastructure built (as a result of the CLTS campaign) and the maintenance of the facility:

'X's house has a non-concrete latrine with a WC [water closet] in it. The WC has not been properly installed in the latrine and it is just placed on the top. The women of the house use this latrine to urinate only. They do not use it for defecation because there is no sewerage system and there is no ditch outside. The males of the house defecate under the trees, in the fields or in the bushes outside. Females defecate inside the house in a space surrounded by walls of bushes and sticks.' (Female IDI respondent, T)

Third, we analyze heterogeneity by a village's vulnerability to shocks that can cause infrastructure to deteriorate, considering in particular weather shocks such as droughts and floods. Our study province, Sindh, experiences extreme climatic variations throughout the year, with heavy rains and frequent flooding. Weather shocks can directly affect the functionality of latrines (e.g. latrines can overflow due to heavy rain, which creates negative health spillovers (Bancalari and Martinez, 2018)) and indirectly affect the ability of households to maintain latrines because of temporary income shocks that reduce the capacity to invest in housing. Our proxy for weather vulnerability is an indicator of whether the village was affected by a flood/drought in the two years prior to the baseline survey.

Latrines collapsing due to rain was a key reason for them becoming non-functional. This was most often the case for basic, mud or *katchi* latrines, which were more susceptible to weather conditions.[26]

'The trouble people faced was seasonal rains, which demolished the mud latrines they had constructed with their hard work. These people were not able to purchase the required materials to build them again due to a lack of resources. This was also a reason why they were turning towards open defecation'. (SO, FGD, Badin)

'Every house in our village has a latrine. All the latrines are basic. A lot of latrines got ruined during rain. There are some people who made their own latrines again, and there are some who use their relatives' latrine and say that they will make their own latrine [in the future], and then there are some people like us who resort to open fields and defecate there. Yes, there

[26] The term *katchi* usually refers to something in the early stage of development. In this case, *katchi* latrines were usually described as basic, mud or un-cemented latrines or pit systems.

Table 2
Heterogeneous effects of follow-up activities on prevalence of OD and using a functional latrine, by village characteristics.

	Public infrastructure		Sanitation supply		Weather vulnerability	
Outcome	OD	Latrine	OD	Latrine	OD	Latrine
	(1)	(2)	(3)	(4)	(5)	(6)
Follow-up activities	−0.26***	0.26***	−0.17**	0.14	−0.04	0.11
	(0.09)	(0.08)	(0.08)	(0.09)	(0.19)	(0.15)
Follow-up activities x Yes/High	0.27*	−0.34**	0.07	−0.02	−0.12	0.02
	(0.15)	(0.14)	(0.14)	(0.15)	(0.20)	(0.17)
Follow-up activities (Yes/High)	0.00	−0.08	−0.10	0.13	−0.16**	0.13*
	(0.98)	(0.51)	(0.38)	(0.29)	(0.03)	(0.10)
Control mean (No/Low)	0.64	0.44	0.63	0.46	0.51	0.57
Control mean (Yes/High)	0.51	0.53	0.47	0.54	0.58	0.48
Villages	61	61	61	61	61	61
Households	551	551	551	551	551	551

Notes. 'OD' (columns (1), (3) and (5)) is an indicator for whether a household has at least one member older than 5 years who openly defecates. 'Latrine' (columns (2), (4) and (6)) is an indicator for whether the household has a functional toilet in the dwelling used by its members. 'Public infrastructure' (columns (1) and (2)) is an indicator for being above the median village public infrastructure index. 'Sanitation supply' (columns (3) and (4)) is an indicator for being above the median sanitation supply index. 'Weather vulnerability' (columns (5) and (6)) is an indicator for whether the village had floods or drought in the last two years. 'Follow-up activities' shows the estimates of the α_1 parameter and 'Follow-up activities x Yes/High' presents the estimates of the α_2 parameter from Eq. (2). 'Follow-up activities (Yes/High)' is a post-estimation result of the linear combination $\alpha_1 + \alpha_2$. All regressions control for the corresponding heterogeneity dimension. All coefficients are estimated from an ANCOVA model. Sample includes households interviewed at endline in Sindh. Standard errors clustered at the unit of randomization (villages) in parentheses. *p*-values adjusted for multiple hypothesis testing by heterogeneity dimension in brackets. Statistical significance denoted by $^*p < 0.1$, $^{**}p < 0.05$ and $^{***}p < 0.01$.

are a lot of people who have started going to the fields again to defecate.' (Male IDI respondent, T)

In this vein, some respondents explicitly stressed the importance of follow-up:

'I don't think people will continue their [improved sanitation] behavior because, firstly, those who were there to make them understand have gone. And then on top of that, in the rainy season latrines collapse and poor people like us can't construct them again … [H]alf the village had previously made latrines – even if these were for the use of women only. But then half the village's latrines collapsed in the rain and so all the women have started using open spaces [to defecate] again. Other than myself and a few village people, no one else has contacted the villagers again for follow-up.' (Male CRP, KII, T)

Table 2 shows the heterogeneous ITT effects of follow-up activities on the probability of a household practicing OD (columns (1), (3) and (5)) and of having a functional latrine (columns (2), (4) and (6)) by these three village-level factors. Columns (1) and (2) show the heterogeneous effects by the public infrastructure index, columns (3) and (4) by the sanitation supply index, and columns (5) and (6) by the weather vulnerability indicator. 'Follow-up activities' shows the estimates of the α_1 parameter from Eq. (2) and 'Follow-up activities x Yes/High' of the α_2 parameter. 'Follow-up activities (Yes/High)' is a post-estimation result of the linear combination $\alpha_1 + \alpha_2$. All coefficients are estimated using an ANCOVA specification.

Follow-up activities were of relevance in villages with low availability of and access to public infrastructure. Households living in communities with low-quality public infrastructure were significantly less likely to perform OD (by 26 ppts) and more likely to own a functional latrine (by 26 ppts) if allocated to receive follow-up activities.[27]

We see a similar pattern when looking at village sanitation supply and weather vulnerability, with follow-up activities being more effective in areas where the households' ability to keep their toilets functional is likely hampered. The interaction terms are, while consistent, not as precisely estimated. Only the heterogeneity by public infrastructure quality survives multiple hypothesis testing. However, we are likely underpowered due to a small sample size in terms of detecting significant differences for other dimensions, as we will show through robustness checks in Section 4.4.

We next examine in which part of the distribution of village-level public infrastructure quality follow-up activities are more effective at sustaining behavioral change. Fig. 4 shows local polynomial smooth plots (with confidence intervals at the 90% level) for OD (Panel A) and for use of a functional latrine (Panel B) by the public infrastructure index. It shows that follow-up activities are only effective when the public infrastructure index is below 0.6.[28] We find similar results when analyzing heterogeneity based on continuous measures of village characteristics. Panel A of Figure A7 in the Appendix confirms the result by showing the marginal heterogeneous ITT effect by quantile of the continuous heterogeneity dimension. Figure A5 and A7 in the Appendix show that the heterogeneous ITT effects by the distribution of the sanitation supply index are not as pronounced as by the public infrastructure index.

4.2. Heterogeneity by household-level characteristics

The previous section suggests that follow-up activities were most effective at sustaining behavioral change in communities with poorer public infrastructure. However, this may proxy overall poverty, which bundles together several constraints affecting decision-making and behavior across households.[29]

Previous literature suggests that triggering initial sanitation adoption is only effective when there are no resource constraints. Cameron

[27] As highlighted by Crocker et al. (2017a) and in a review by Kresche et al. (2020), higher sanitation adoption can make it more worthwhile, or easier, for other households to follow improved sanitation behavior. It is therefore possible that these village-level heterogeneous ITT effects might be driven by initial OD rates. Because we lack a baseline census with data on the practice of OD for every household in the village, we are not able to proxy village-level OD. As a second-best and highly incomplete measure, we show impacts by village-level OD prevalence among beneficiary households that were surveyed at baseline. We find that follow-up activities are not more/less effective depending on whether the household is located in a village above/below the median OD rate of eligible households (see columns (1) and (2) of Tables A14 and A16). We also find that the village-level heterogeneous ITT effects survive when including the interaction 'Follow-up activities x Eligible OD rate (High)'. This interaction is not statistically significant (see Table A15).

[28] We have only a few observations at the extremes of the distribution, making inference difficult.

[29] The adoption of sanitation behavior has also been shown to differ by religion in India (Geruso and Spears, 2018), but our setting is more homogeneous along this dimension than the neighboring country.

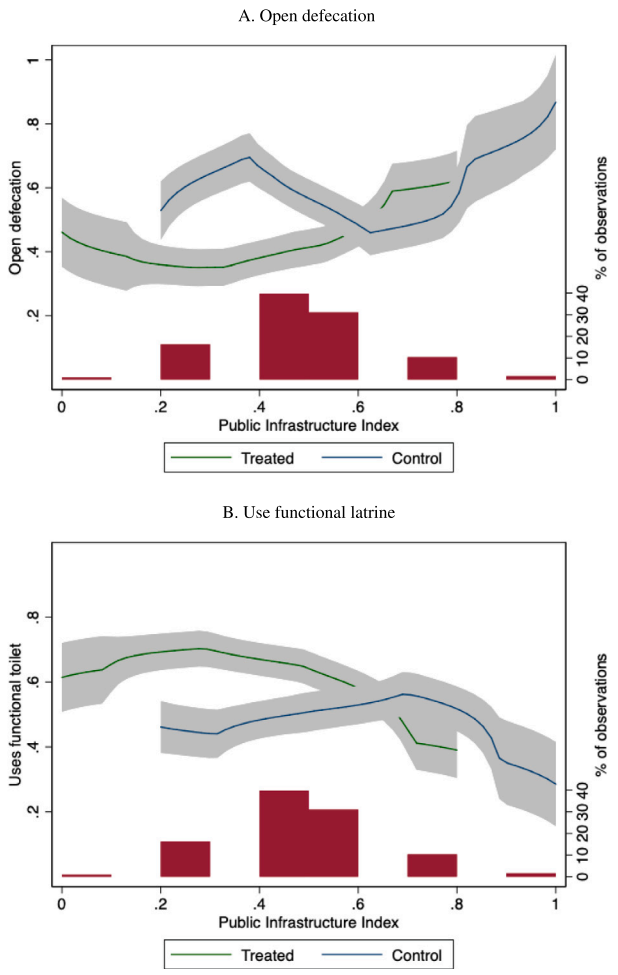

Fig. 4. Heterogeneous effects of follow-up activities on prevalence of OD and use of a functional latrine, by public infrastructure.
Notes. Shows local polynomial smooth plots (with confidence intervals at the 90% level). 'Open defecation' (Panel A) is an indicator for whether a household has at least one member older than 5 years who openly defecates. 'Use functional latrine' (Panel B) is an indicator for whether the household has a functional latrine in the dwelling used by its members. 'Public infrastructure' is an index capturing the availability of infrastructure in the village (see Table A3 for more details). The histogram located at the bottom of each plot shows the distribution of observations. Sample includes households interviewed at endline in Sindh.

et al. (2019) find that latrine promotion is only effective among relatively richer households and Guiteras et al. (2015) find that it is only effective when accompanied by subsidies. Subsidies to take up sanitation facilities have also been found to be more effective for those with stable incomes (Lipscomb and Schechter, 2018).

To analyze how the effectiveness of follow-up activities interacts with poverty, we use as a proxy a standard baseline household asset index (indicating ownership of things such as TVs, radios, vehicles and productive assets). In the words of one CRP and a Community Leader:

'Some people's latrines have fallen [...] and they haven't made them again [Probe: Were households whose latrines contacted again? Why did they not repair them?] I [CRP] visited these households and told them to repair their latrines but they said we don't have the money to keep making latrines again and again. They said we are poor, and we don't have money – ask [the NGO] to make them for us.' (Female CRP, KII, T)

'The biggest factor is poverty. Even the most basic katchi latrines cost PKR 10,000 to make. Buying the materials, the cost of transporting them here, then paying for the labor and the mason for construction – this is very hard for the poor. People who use the fields to defecate these days get very embarrassed but they don't have another choice.' (Male Community Leader, KII, T)

Columns (1) and (2) of Table 3, which has the same set-up as Table 2 and also presents estimates from ANCOVA models, show that asset-poorer households allocated to follow-up activities were 16 ppts less likely to practice OD than asset-poorer households in control villages (≈27% lower OD). This effect is accompanied by an increase of a similar magnitude in the probability of using a functional latrine. The interaction term, however, is not statistically significant. This finding is in line with Ben Yishay et al. (2017) and Devoto et al. (2012)'s findings that the effectiveness of releasing financial constraints for water and sanitation adoption does not differ by initial poverty. Two offsetting forces may be at play: while richer households are better able to invest in keeping latrines functional, they may already be doing so. Worse-off households might be in greater need of the nudge provided by follow-up activities. Also, given that we sampled from households who had constructed a toilet prior to the baseline survey, the poorer households might be those with a relatively higher interest in sanitation in spite of their resource constraints.

We unbundle poverty status by looking at differential effects by the initial quality of the household's sanitation facility. Watson (2006) finds that public sanitation investments are less effective at improving the disease environment when there are better pre-existing sanitation facilities. When considering private investments, however, Lipscomb and Schechter (2018) find that the effectiveness of subsidies to adopt sanitary facilities is not greater for individuals who have already invested in better technologies in the past. They argue that these individuals would have been more likely to continue investing in the future absent the intervention.

We explore heterogeneous effects along the following factors: status of the sanitation facility; whether the facility is shared or not; and the latrine technology. While the status of the sanitation facility and whether it is shared are more closely related to continuous maintenance of the whole facility, the technology is related to the quality of the latrine initially built.[30] Although these factors are correlated with asset poverty, they are not perfectly correlated, allowing us to analyze a more specific heterogeneity dimension.[31] To examine these factors, we consider three indicators.

The first is an index that captures sanitation status and is created using observations of whether the sanitation facility has a permanent superstructure, a roof, curtains or door, and a functional handwashing facility, its observed cleanliness status (feces, flies, odor) and whether it needs repairs.[32]

The second indicator measures whether or not the sanitation facility is shared with other households, which captures potential coordination problems in the maintenance of the facility. Some respondents in the qualitative research expressed openness in allowing relatives (and at times non-relatives) to share a latrine under certain circumstances.[33]

[30] Given the relatively high OD baseline prevalence, one might wonder whether owning, or using, a toilet to begin with interacts with follow-up activities. One might imagine a situation where a household already owning a toilet experienced a latrine collapse due to rain, and that follow-up activities, implemented shortly after, motivated reconstruction. We find some support for this hypothesis, but with insignificant coefficients on the interaction between OD at baseline and our treatment indicator, as shown in columns (3) and (4) of Tables A14 and A16 in the Appendix).

[31] See Table A13 for correlations of these indexes with our poverty proxy.

[32] See Table A3 for more details about the construction of the indices and their components.

[33] The main reservations and concerns regarding sharing outside of the household were related to overuse (e.g. pits getting filled very quickly) and cleanliness.

Table 3
Heterogeneous effects of follow-up activities on prevalence of OD and using a functional toilet, by household characteristics.

	Household assets		Sanitation status		Shared facility		Latrine technology	
Outcome	OD	Latrine	OD	Latrine	OD	Latrine	OD	Latrine
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Follow-up activities	−0.16**	0.14*	−0.17**	0.16**	−0.14*	0.13*	−0.07	0.04
	(0.07)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.09)
Follow-up activities x Yes/High	0.06	−0.00	0.09	−0.08	−0.08	−0.02	−0.13	0.16*
	(0.12)	(0.12)	(0.10)	(0.09)	(0.16)	(0.16)	(0.10)	(0.10)
Follow-up activities (Yes/High)	−0.10	0.13	−0.08	0.08	−0.21	0.11	−0.20**	0.21***
	(0.39)	(0.23)	(0.41)	(0.36)	(0.16)	(0.48)	(0.02)	(0.01)
Control mean (No/Low)	0.60	0.46	0.63	0.41	0.55	0.50	0.78	0.24
Control mean (Yes/High)	0.47	0.60	0.42	0.72	0.58	0.50	0.48	0.60
Villages	61	61	61	61	61	61	61	61
Households	551	551	551	551	551	551	551	551

Notes. 'OD' (columns (1), (3), (5) and (7)) is an indicator for whether a household has at least one member older than 5 years who openly defecates. 'Latrine' (columns (2), (4), (6) and (8)) is an indicator for whether the household has a functional toilet in the dwelling used by its members. 'Household assets' (columns (1) and (2)) is an indicator for being above the median household asset index. 'Sanitation status' (columns (3) and (4)) is an indicator for being above the median of the sanitation status index. 'Shared facility' (columns (5) and (6)) is an indicator for whether the household's bathroom and latrine are shared with other households. 'Latrine technology' (columns (7) and (8)) is an indicator for whether the household latrine is of improved technology and has a water seal. 'Follow-up activities' shows the estimates of the α_1 parameter and 'Follow-up activities x Yes/High' presents the estimates of the α_2 parameter from Eq. (2). 'Follow-up activities (Yes/High)' is a post-estimation result of the linear combination $\alpha_1 + \alpha_2$. All regressions control for the corresponding heterogeneity dimension. All coefficients are estimated from an ANCOVA model. Sample includes households interviewed at endline in Sindh. Standard errors clustered at the unit of randomization (villages) in parentheses. *p*-values adjusted for multiple hypothesis testing by heterogeneity dimension in brackets. Statistical significance denoted by $^*p < 0.1$, $^{**}p < 0.05$ and $^{***}p < 0.01$.

The third indicator focuses on whether the latrine is of an improved technology and has a water seal.[34]

Table 3 shows that follow-up activities are more effective at sustaining behavior change when the sanitation facility has poor status and is shared. Columns (3) and (5) show that households allocated to follow-up activities were 17 ppts and 14 ppts less likely to practice OD if they have a bad-quality sanitation facility and if it is shared than control households with those baseline characteristics, respectively (≈26% lower OD). These heterogeneous effects are accompanied by an increase of a similar magnitude in the probability of using a functional latrine (see columns (4) and (6)). Only the effect of follow-up activities on the behavior of households with poor sanitation status remains statistically significant at the 10% level.

Furthermore, column (7) shows that households allocated to follow-up activities are 20 ppts less likely to practice OD (and 21 ppts more likely to use latrines in column (8)) when the initially built latrine is of improved technology with a water seal than control households with similar latrines. The interaction term is only statistically significant for heterogeneity along the latrine technology dimension. When adjusting *p*-values for multiple hypothesis testing, however, the interaction is no longer statistically significant.

We next examine in which part of the distribution of the sanitation status index follow-up activities are more effective at sustaining behavioral change. Fig. 5 shows that the prevalence of OD is lower (Panel A) and the use of functional latrines higher (Panel B) in treatment than control groups when the sanitation status index lies between 0.1 and 0.5, and the difference disappears thereafter. It is worth noting that very few observations are located at the extremes of the sanitation status index.[35] Figure A5 in the Appendix shows that heterogeneous effects by the distribution of the household asset index are not as pronounced as by the sanitation status index. The results shown in Panels C and D of Figure A7 in the Appendix based on continuous measures of these household characteristics confirm that the heterogeneous effects are not statistically significant.

4.3. Predominant heterogeneity

In this section, we present suggestive evidence that households living in villages with worse public infrastructure and using a sanitation facility of low quality need follow-up activities the most to prevent slippage back into OD. We also provide evidence in support of follow-up activities incentivizing households to conduct regular maintenance of their latrines and changing social norms.

To determine the key driving mechanisms in the effectiveness of follow-up activities, we run a 'joint analysis' in which we include the treatment variable, all heterogeneity dimensions and their interactions with the treatment variable. Essentially, we estimate a combination of Eqs. (1) and (2) for all heterogeneity dimensions at the same time. The idea is to evaluate which heterogeneity dimension 'survives'.

Table 4 presents the result of this exercise for both of our main outcomes. Note that the values of the heterogeneity dimensions have been inverted from 'High/Yes' to 'Low/No' in order to visualize when follow-up activities are the most effective. The estimate of the average ITT effect of follow-up activities on OD and functional latrine remains similar in magnitude, but it is no longer precisely estimated. Interestingly, we find that neither household asset poverty nor whether the toilet is shared are the predominant heterogeneities. What survive the joint analysis are the heterogeneous ITT effects by village-level public infrastructure and the household-level sanitation status. The drop in OD (increase in functional latrines) is greater in villages with worse public infrastructure and in households with poorer sanitation status.

It is important to note that this 'joint analysis' is not causal, as the dimensions of heterogeneity we consider are correlated among each other (see Table A13 for pairwise correlations of heterogeneity dimensions at baseline). To further understand the mechanisms, we conduct two sets of analyses which suggest that follow-up activities can be effective at sustaining behavioral change through: (i) leveraging an initially better-quality latrine technology to keep toilets functional; (ii) nudging households to invest in maintenance and continuous improvement of their sanitation and water facilities; and (iii) changing social norms related to sanitation.

First, we estimate the ITT effects of follow-up activities on variables capturing the three potential mechanisms using ANCOVA models (see

[34] Qualitative findings suggest that the intervention had some influence in getting households to upgrade their latrine structure. One respondent expressed it as follows:

'We built our latrine five years ago [...] Before this [program] we had a basic latrine without a WC and when [the NGO] arrived they told us about WCs and told us to make an [improved] latrine. Everyone now has a latrine and a WC in their home.' (Female, IDI, T)

[35] Figure A6 shows that the distribution of the sanitation status index is similar across treatment arms.

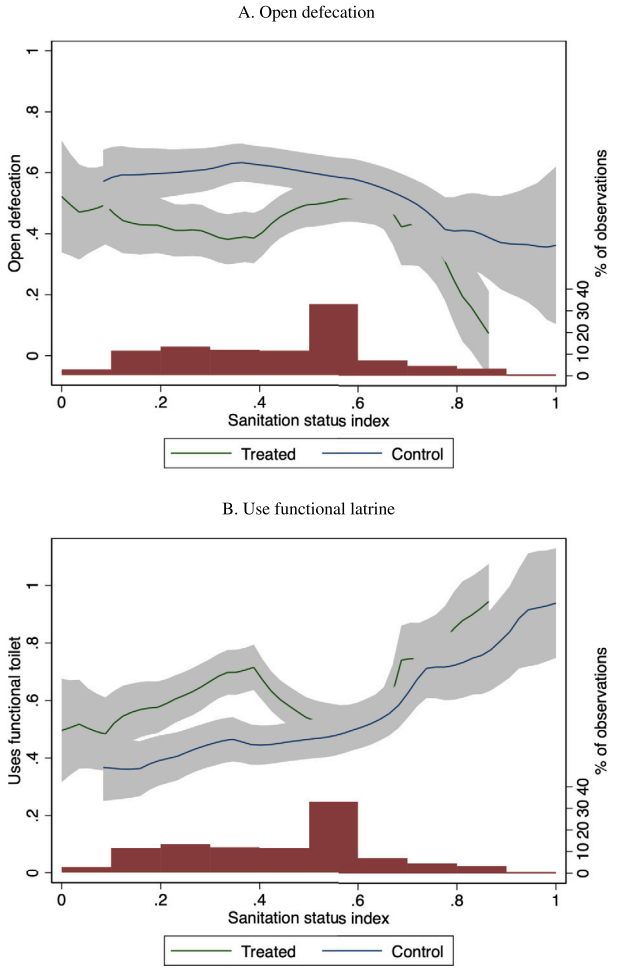

Fig. 5. Heterogeneous effects of follow-up activities on prevalence of OD and using a functional toilet, by sanitation status.
Notes. Shows local polynomial smooth plots (with confidence intervals at the 90% level). 'Open defecation' (Panel A) is an indicator for whether a household has at least one member older than 5 years who openly defecates. 'Use functional latrine' (Panel B) is an indicator for whether the household has a functional latrine in the dwelling used by its members. 'Sanitation status' is an index capturing the status of the sanitation and hygiene facility in the household's dwelling (see Table A3 for more details). The histogram located at the bottom of each plot shows the distribution of observations by the sanitation status index. Sample includes households interviewed at endline in Sindh.

Table A17). In Sindh, we find an increase by 12 ppts in the probability of observing a functional latrine in the dwelling compared with the control. When using the sample with both provinces, we find increases by 3 ppts in having maintained the latrine in the previous two years and by 7 ppts in the probability of observing a hand-washing facility. When using the sample with both provinces, and when concentrating on Sindh only, we also find an increase in the likelihood of observing water available in the bathroom. Interestingly, in Punjab, we estimate a decrease by 8 ppts in the likelihood of the main respondent in the household agreeing with the statement 'Nobody minds OD as it is common'. However, these point estimates are no longer significant when adjusting p-values for multiple hypothesis testing.

In line, participants reported:

'Yes, behaviours and attitudes have changed quite a bit and people care about cleanliness now, whereas first they wouldn't so much. People used to follow what [the NGO] told us in meetings and kept their latrines clean. There is another village where households made latrines, but didn't keep them clean. Now the latrines have become such that they are not worth using' (Female CRP, KII, T)

Table 4
Joint analysis.

Outcome	Open defecation (1)	Use functional latrine (2)
Follow-up activities	0.11	−0.11
	(0.22)	(0.22)
Village level		
FU x Public infrastructure (Low)	−0.28**	0.32**
	(0.14)	(0.14)
FU x Sanitation supply (Low)	−0.04	−0.02
	(0.14)	(0.14)
FU x Vulnerability (Low)	0.06	−0.01
	(0.18)	(0.16)
Household level		
FU x Asset (Low)	−0.06	0.05
	(0.12)	(0.11)
FU x Sanitation status (Low)	−0.16*	0.18*
	(0.09)	(0.10)
FU x Shared facility (No)	0.04	−0.04
	(0.13)	(0.12)
FU x Latrine technology (Low)	0.11	−0.16
	(0.09)	(0.10)
Villages	61	61
Households	551	551

Notes. 'Open defecation' (column (1)) is an indicator for whether a household has at least one member older than 5 years who openly defecates. 'Use functional latrine' (column (2)) is an indicator for whether the household has a functional toilet in the dwelling used by its members. All regressions control for the heterogeneity dimensions. All coefficients are estimated from an ANCOVA model. Sample includes households interviewed at endline in Sindh. Standard errors clustered at the unit of randomization (villages) in parentheses. Statistical significance denoted by $^{*}p < 0.1$, $^{**}p < 0.05$ and $^{***}p < 0.01$.

Second, we analyze whether follow-up activities have an effect on the probability of a household maintaining their latrine, and we look at heterogeneous effects along the dimensions that survived the joint analysis exercise. In Sindh, we find that follow-up activities increased the probability of maintaining latrines by 7 ppts if the baseline sanitation status is low compared with control households with the same initial characteristic (87% higher maintenance; see Table A18). Although the interaction term is not statistically significant, this serves as additional evidence that follow-up activities nudged households that would have otherwise not invested in the upkeep of their sanitation facility, and the resulting functional latrine kept individuals away from OD.

4.4. Robustness checks

We conduct the following robustness checks on our heterogeneity analysis.

To start with, in order to check that our results are not dependent on our chosen set of controls, we estimate our heterogeneous ITT effects using LASSO (in addition to estimating ANCOVA models as in Tables 2 and 3). Table A19 shows that the heterogeneous ITT effects by village public infrastructure remain robust and those by weather vulnerability are now statistically significant. Treated households in villages that suffered from weather shocks (i.e. floods or drought) are 52 ppts less likely to practice OD and 41 ppts more likely to use a functional latrine, compared with control households that also suffered from these shocks. Additionally, Table A20 shows that the heterogeneous ITT effects by latrine technology are more precisely estimated and slightly larger in magnitude.

A second robustness check is conducted by exploiting the full sample of study households and villages, pooling together Sindh and Punjab provinces, and adding province stratification dummies. Estimating heterogeneous ITT effects in the full sample increases the statistical power to detect these effects. Our heterogeneous results remain robust when conducting the analysis including Punjab. The heterogeneous effects by availability of village-level public infrastructure remain consistent, only slightly lower in magnitude. Heterogeneous effects are now more

precisely estimated (at the 5% and 1% significance levels) and larger in magnitude when looking at household-level dimensions. The downward effect on OD is 14–15 ppts larger (and statistically significant) for households with the asset and sanitation status indexes below the median at baseline compared with control households with the same characteristics. Likewise, the positive effect of follow-up activities on using a functional latrine is 16 ppts larger for households with sanitation status below the median at baseline compared with similar control households (see Tables A21 and A22). We also find that in the joint analysis, the heterogeneity along latrine technology survives, in addition to the village-level public infrastructure and household-level sanitation status index (see Table A23). Treated households with better technology – i.e. an improved latrine with a water seal at baseline – are 15 ppts less likely to practice OD and 16 ppts more likely to use a functional latrine at endline, compared with control households with this better technology.

Finally, we estimate the heterogeneous ITT effects in the full sample and using LASSO. The heterogeneous effects by availability of public infrastructure, household assets and sanitation status remain consistent, only slightly lower in magnitude and, for the public infrastructure heterogeneity, less precisely estimated than without LASSO (see Tables A24 and A25).

5. Conclusions

Understanding what factors affect sanitation outcomes in a sustainable manner is important in view of achieving safe sanitation for all, as outlined in SDG 6.2. Our analysis makes several contributions towards this end.

First, our results confirm that initial technology adoption is no guarantee for sustained usage, as previously highlighted by, for example, Hanna et al. (2016) for eco-friendly and health-improving cooking stoves.

Second, our results disclose that a continuation of CLTS at lighter dosage in the form of follow-up activities can lead to significant returns. In the province of Sindh, follow-up activities helped 47% of households that would have abandoned the use of their toilet not to do so, making these households continue safer sanitation practices over the two-year analysis period.[36]

Third, our heterogeneity analysis provides guidance on how implementers can most effectively target follow-up activities. Follow-up activities are likely to be most effective when targeted to where initial latrine quality is poor and where latrines are more likely to collapse, as well as in communities with inferior public infrastructure.[37] In addition, following such a strategy likely achieves convergence: households living in more difficult environments are supported in achieving sustained OD behavior, resulting in more equitable public good delivery, and ultimately a more equitable society. While we are not able to generalize whether follow-up activities would have similar effects on households that constructed a toilet outside a CLTS context, these heterogeneity dimensions suggest that they could be, e.g. that some outside support might be relevant for households living in more difficult environments, independent of whether CLTS triggering was previously implemented or not.

Despite not having the right data to cost follow-up activities, and thereby attempt any meaningful cost-effectiveness analysis, we can draw some policy-relevant conclusions. For one, our results stress that assumptions about post-intervention persistence in any CLTS cost-effectiveness analysis should adjust for slippage in behavior as otherwise the cost-effectiveness would be overstated. Furthermore, follow-up activities will likely improve the cost-effectiveness of CLTS. Previous costing exercises of CLTS in several contexts reveal that the average CLTS program cost is US$38 per *targeted* household (Crocker et al., 2017b). The impact of sanitation uptake, on the other hand, is estimated to lie between 4 and 23 ppts on average (Patil et al., 2014; Guiteras et al., 2015; Pickering et al., 2015; Augsburg et al., 2020), with higher impacts in poorer contexts (Abramovsky et al., 2018). Our estimated impacts of follow-up activities on using functional latrines at 6 ppts on average and 13 ppts in Sindh, do not fall short of these by much. At the same time, their costs are a small portion of overall program costs. Crocker et al. (2017a) highlight that training makes up a large portion of CLTS implementation costs. Local actors' training in particular is estimated to be 56%–61% of program cost. The fact that this training has been completed by the time follow-up activities start makes it conceivable that their implementation is a worthwhile addition to the intervention as a whole.

[36] Such impacts of continued treatment have, for example, been identified in the area of energy conservation by Allcott and Rogers (2014), who show that the effects of continuously sending home energy reports are significant and that households keep responding positively to repeated treatment even after two years.

[37] Whether it would alternatively (or in addition) be more (cost) effective to construct better-quality infrastructure in the first place should be an important consideration in future work.

CRediT authorship contribution statement

Britta Augsburg: Conceptualization, Investigation, Project administration, Methodology, Writing, Supervision. **Antonella Bancalari:** Investigation, Project administration, Methodology, Software, Formal analysis, Writing, Supervision. **Zara Durrani:** Conceptualization, Project administration, Methodology, Data curation, Formal analysis, Writing, Supervision. **Madhav Vaidyanathan:** Software, Formal analysis, Validation, Visualisation. **Zach White:** Conceptualization, Project administration, Investigation, Data curation, Writing, Funding acquisition, Supervision, Formal analysis, Validation.

Data availability

Data will be made available on request.

Acknowledgments

The authors are grateful to Andrew Foster, Raymond Guiteras, Molly Lipscomb, two anonymous referees, and participants at the IFS Development Seminar and the ADBI Sanitation and Development Conference for their useful comments and suggestions.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jdeveco.2022.102933.

References

Abramovsky, L., Augsburg, B., Luhrmann, M., Oteiza, F., Rud, J.P., 2018. Community matters: heterogenous impacts of a sanitation intervention. Institute for Fiscal Studies Working Paper W18/28. London, United Kingdom.

Allcott, H., Rogers, T., 2014. The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation. Amer. Econ. Rev. 104 (10), 3003–3037.

Augsburg, B., Caeyers, B., Giunti, S., Smets, S., 2020. Labelled loans and human capital investments. Institute for Fiscal Studies Working Paper W20/20. London.

Bancalari, A., 2020. Can white elephants kill? Unintended consequences of infrastructure development. Institute for Fiscal Studies Working Paper W20/32.

Bancalari, A., Martinez, S., 2018. Exposure to sewage from on-site sanitation and child health: a spatial analysis of linkages and externalities in peri-urban Bolivia. J. Water Sanitation Hyg. Dev. 8 (1), 90–99.

Banerjee, A.V., Barnhardt, S.M., Duflo, E., 2015. In: Wise, D. (Ed.), Insights in the Economics of Aging. National Bureau of Economic Research, Inc.

Banerjee, A., Duflo, E., Glennerster, R., Kothari, D., 2010. Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. BMJ 340 (c2220), 1–9.

Barnard, S., Routray, P., Majorin, F., Peletz, R., Boisson, S., Sinha, A., Clasen, T., 2013. Impact of Indian Total Sanitation campaign on latrine coverage and use: A cross-sectional study in Orissa three years following programme implementation. PLoS One 8 (8), e71438.
Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. 28 (2), 29–50.
Ben Yishay, A., Fraker, A., Guiteras, R., Palloni, G., Buddy Shah, N., Shirrell, S., Wang, P., 2017. Microcredit and willingness to pay for environmental quality: Evidence from a randomized-controlled trial of finance for sanitation in rural Cambodia. J. Environ. Econ. Manage. 86, 121–140.
Bjorkman Nyqvist, M., de Walque, D., Svensson, J., 2017. Experimental evidence on the long-run impact of community-based monitoring. Amer. Econ. J.: Appl. Econ. 9 (1), 33–69.
Briceño, B., Coville, A., Gertler, P., Martinez, S., 2017. Are there synergies from combining hygiene and sanitation promotion campaigns: Evidence from a large-scale cluster-randomized trial in rural Tanzania. PLoS One 12 (11), 1–19.
Bureau of Statistics, Planning and Development Department, Government of Punjab, and UNICEF, 2014. Multiple indicator cluster survey – final report.
Cairncross, S., Shordt, K., Zacharia, S., Govindan, B.K., 2005. What causes sustainable changes in hygienic behaviour? A cross-sectional study from kerala, india. Soc. Sci. Med. 61 (10), 2212–2220.
Cameron, L., Olivia, S., Shah, M., 2019. Scaling up sanitation: Evidence from an RCT in Indonesia. J. Dev. Econ. 138, 1–16.
Chambers, R., Kar, K., 2008. Handbook on Community-Led Total Sanitation. Institute of Development Studies, p. 96.
Chirgwin, H., Cairncross, S., Zehra, D., Waddington, H., 2021. Interventions promoting uptake of water, sanitation and hygiene (WASH) technologies in low- and middle-income countries: An evidence and gap map of effectiveness studies. Campbell Syst. Rev. 17.
Chong, A., Gonzalez-Navarro, M., Karlan, D., Valdivia, M., 2020. Do information technologies improve teenagers' sexual education? Evidence from a randomized evaluation in colombia. World Bank Econ. Rev. 34 (2), 371–392.
Coffey, D., Gupta, A., Hathi, P., Khurana, N., Dean, S., Srivastav, N., Vyas, S., 2014. Revealed preference for open defecation - evidence from a new survey in rural north india. Econ. Polit. Wkly. 38–43.
Creswell, J.W., Piano Clark, V.L., Gutmann, M., Hanson, W., 2003. In: Tashakkori, A., Teddlie, C. (Eds.), Advanced Mixed Methods Research Designs. Sage Publications, pp. 209–240.
Crocker, J., Saywell, D., Bartram, J., 2017a. Sustainability of community-led total sanitation outcomes: Evidence from Ethiopia and Ghana. Int. J. Hyg. Environ. Health 220, 551–557.
Crocker, J., Saywell, D., Shields, K., Kolsky, P., Bartram, J., 2017b. The true costs of participatory sanitation: Evidence from community-led total sanitation studies in Ghana and Ethiopia. Sci. Total Environ. (601–602), 1075–1083.
Crump, R., Hotz, J., Imbens, G., Mitnik, O., 2008. Nonparametric tests for treatment effect heterogeneity. Rev. Econ. Stat. 90, 389–405.
DellaVigna, S., Malmendier, U., 2006. Paying not to go to the gym. Amer. Econ. Rev. 96 (3), 694–719.
Devoto, F., Duflo, E., Dupas, P., Pariente, W., Pons, V., 2012. Happiness on tap: piped water adoption in urban morocco. Amer. Econ. J.: Econ. Policy 4 (4), 68–99.
Djebbari, H., Smith, J., 2008. Heterogeneous impacts in PROGRESA. J. Econometrics 145, 64–80.
Duflo, E., Dupas, P., Kremer, M., 2015. Education, HIV, and early fertility: experimental evidence from kenya. Amer. Econ. Rev. 105 (9), 2757–2797.
Dupas, P., 2009. What matters (and what does not) in households' decision to invest in malaria prevention? Amer. Econ. Rev.: Pap. Proc. 99 (2), 224–230.
Dupas, P., 2011. Health behavior in developing countries. Annu. Rev. Econ. 3, 425–449.
Dupas, P., Miguel, E., 2017. Impacts and determinants of health levels in low-income countries. In: Banerjee, A., Duflo, E. (Eds.), Handbook of Economic Field Experiments, Vol. 2. Elsevier B.V., pp. 3–73.
Ferraro, P., Miranda, J., 2013. Heterogeneous treatment effects and mechanisms in information-based environmental policies: Evidence from a large-scale field experiment. Resour. Energy Econ. 35, 356–379.
Gertler, P., Shah, M., Alzua, M.L., Cameron, L., Martinez, S., Patil, S., 2015. How does Health Promotion Work? Evidence from the Dirty Business of Eliminating Open Defecation. National Bureau of Economic Research.
Geruso, M., Spears, D., 2018. Neighborhood sanitation and infant mortality. Amer. Econ. J.: Appl. Econ. 2 (10), 125–162.
Government of the Islamic Republic of Pakistan Ministry of Environment, 2006. National sanitation policy.
Government of the Islamic Republic of Pakistan Ministry of Environment, 2011. Pakistan Approach to Total Sanitation (PATS).
Guiteras, R.P., Levinsohn, J., Mobarak, A.M., 2015. Encouraging sanitation investment in the developing world: A cluster-randomized trial. Science 348 (6237), 903–906.
Hanna, R., Duflo, E., Greenstone, M., 2016. Up in smoke: the influence of household behavior on the long-run impact of improved cooking stoves. Amer. Econ. J.: Econ. Policy (8), 80–114.
Heckman, J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Rev. Econom. Stud. 64, 487–535.
Hussam, R., Rabbani, A., Reggiani, G., Rigol, N., 2021. Rational habit formation: experimental evidence from handwashing in india. Amer. Econ. J.: Appl. Econ. 14, 1–41.
Hutton, G., Varughese, M., 2016. The costs of meeting the 2030 sustainable development goal targets on drinking water, sanitation, and hygiene. Water and sanitation program: technical paper. Washington D.C.
Ivankova, N.V., Creswell, J.W., Stick, S.L., 2006. Using mixed-methods sequential explanatory design: from theory to practice. Field Methods 18, 3–20.
Joint Monitoring Program WHO-UNICEF, 2017. Sanitation - JMP.
Joint Monitoring Program WHO-UNICEF, 2021. Progress on household drinking water, sanitation and hygiene. 2000–2020 Five years into the SDGs.
Kremer, M., Miguel, E., 2007. The illusion of sustainability. Q. J. Econ. 112 (3), 1007–1065.
Kresche, E., Lipscomb, M., Schechter, L., 2020. Externalities and spillovers from sanitation and waste management in urban and rural neighborhoods. Appl. Econ. Perspect. Policy 42, 395–420.
Leech, N., Onwuegbuzie, A., 2009. A typology of mixed methods research designs. Qual. Quant. 43 (2), 265–275.
Lipscomb, M., Schechter, L., 2018. Subsidies versus mental accounting nudges: Harnessing mobile payment systems to improve sanitation. J. Dev. Econ. 135, 235–254.
List, J.A., Shaikh, A.M., Xu, Y., 2019. Multiple hypothesis testing in experimental economics. Exp. Econ. 22 (4), 773–793.
Madajewicz, M., Pfaff, A., van Geen, A., Graziano, J., Hussein, I., Momotaj, H., Sylvi, R., Ahsan, H., 2007. Can information alone change behavior? Response to arsenic contamination of groundwater in Bangladesh. J. Dev. Econ. 84 (2), 731–754.
McKenzie, D., 2012. Beyond baseline and follow-up: The case for more T in experiments. J. Dev. Econ. 99, 210–221.
OECD, 2011. Benefits of Investing in Water and Sanitation: An OECD Perspective. OECD, Paris.
Onwuegbuzie, A.J., Johnson, R.B., 2006. The validity issues in mixed research. Res. Sch. 13 (1).
Orgill-Meyer, J., Pattanayak, S.K., Chindarkar, N., Dickinson, K.L., Panda, U., Rai, S., Sahoo, B., Singha, A., Jeuland, M., 2019. Long-term impact of a community-led sanitation campaign in India, 2005–2016. Bull. World Health Organ. 97 (8), 523–533A.
Patil, S.R., Arnold, B.F., Salvatore, A.L., Briceno, B., Ganguly, S., Colford, J.M., Gertler, P.J., 2014. The effect of india's total sanitation campaign on defecation behaviors and child health in rural madhya pradesh: a cluster randomized controlled trial. PLoS Med. 11 (8), e1001709.
Pickering, A.J., Djebbari, H., Lopez, C., Coulibaly, M., Alzua, M.L., 2015. Effect of a community-led sanitation intervention on child diarrhoea and child growth in rural Mali: a cluster-randomised controlled trial. Lancet Global Health 3 (11), e701–e711.
Prüss-Ustün, A., Bartram, J., Clasen, T., Colford, J.M., Cumming, O., Curtis, V., Bonjour, S., Dangour, A.D., De France, J., Fewtrell, L., Freeman, M.C., Gordon, B., Hunter, P.R., Johnston, R.B., Mathers, C., Mäusezahl, D., Medlicott, K., Neira, M., Stocks, M., Wolf, J., Cairncross, S., 2014. Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: A retrospective analysis of data from 145 countries. Trop. Med. Int. Health 19, 894–905.
Prüss-Ustün, A., Wolf, J., Bartram, J., Clasen, T., Cumming, O., Freeman, M.C., Gordon, B., Hunter, P.R., Medlicott, K., Johnston, R., 2019. Burden of disease from inadequate water, sanitation and hygiene for selected adverse health outcomes: An updated analysis with a focus on low- and middle-income countries. Int. J. Hyg. Environ. Health 765–777.
Saldana, J., 2015. the Coding Manual for Qualitative Researchers. Sage Publications.
Sindh Bureau of Statistics and UNICEF, 2014. Multiple indicator cluster survey – final report.
Stake, R.E., 1995. the Art of Case Study Research. Sage Publications.
Tarozzi, A., Maertens, R., Matin Ahmed, K., van Geen, A., 2021. Demand for information on environmental health risk, mode of delivery, and behavioral change: evidence from sonargaon, bangladesh. World Bank Econ. Rev. 35, 764–792.
Tashakkori, A., Teddlie, C., 1998. Mixed Methodology: Combining Qualitative and Quantitative Approaches, Vol. 46. Sage Publications.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.
Tyndale-Biscoe, P., Bond, M., Kidd, R., 2013. ODF Sustainability study.
UNDP, 2017. Human Development Report 2016 Human Development for Everyone. United National Development Program, New York.
UNICEF, 2014. Evaluation of the WASH sector strategy "community approaches to total sanitation (CATS)", executive summary. UNICEF Evaluation Office.
Watson, T., 2006. Public health investments and the infant mortality gap: Evidence from federal sanitation interventions on U.S. Indian reservations. J. Publ. Econ. 90 (8), 1537–1560.
White, Z., Colin, J., 2019. Monitoring, Evaluation, and Verification Component of the WASH Results Programme. Oxford Policy Management and ITAD.
Whittington, D., Radin, M., Jeuland, M., 2020. Evidence-based policy analysis? The strange case of the randomized controlled trials of community-led total sanitation. Oxf. Rev. Econ. Policy 36 (1), 191–221.
World Bank, 2018. Sanitation overview.
Yin, R., 2003. Case Study Research: Design and Methods. Sage Publications.
Zuin, V., Delaire, C., Peletz, R., Cock-Esteb, A., Khush, R., Albert, J., 2019. Policy diffusion in the rural sanitation sector: lessons from community-led total sanitation (CLTS). World Dev. 124, 104643.