Institute for Fiscal Studies

Britta Augsburg
Antonella Bancalari
Zara Durrani
Madhav Vaidyanathan
Zach White

Working paper
21/46

When nature calls back: sustaining behavioural change in rural Pakistan

Economic and Social Research Council

When Nature Calls Back: Sustaining Behavioural Change in Rural Pakistan*

Britta Augsburg, Antonella Bancalari, Zara Durrani, Madhav Vaidyanathan and Zach White

Abstract

We implement a mixed method approach that combines a randomized controlled trial and qualitative data collection to assess whether, and if so how, behavioural 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, reduced reversal to unsafe sanitation, but only where initial conditions are unfavourable —i.e. poor public infrastructure and low-quality sanitation facilities. Promotion efforts are hence best targeted towards those that face larger difficulties in constructing and maintaining high quality sanitation. (*JEL* C93, I12, I15, I18, O18, Q53)

Keywords: behaviour change, sustainability, basic services, water and sanitation, health, infrastructure, maintenance.

*__Augsburg__: Institute for Fiscal Studies (e-mail: britta_a@ifs.org.uk); **Bancalari**: School of Economics and Finance, University of St. Andrews, and Institute for Fiscal Studies (e-mail: antonella.bancalari@st-andrews.ac.uk); **Durrani**: Oxford Policy Management, Pakistan (e-mail: zara.durrani@opml.co.uk); **Vaidyanathan**: Oxford Policy Management, UK (e-mail: madhav.vaidyanathan@opml.co.uk); **White**: GSMA (e-mail: zwhite@gsma.com). The pre-analysis research design (documented in 'WASH Results Programme: 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. We thank Molly Lipscomb, Raymond Guiteras and participants at the IFS Development Seminar and the ADBI Sanitation and Development Conference for their useful comments and suggestions.

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

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

Although preventive behaviours 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, both in terms of initial adoption as well as maintenance of facilities. Safe sanitation – the adequate disposal of human waste – has been recognised as an indispensable element of disease prevention and primary healthcare programmes (e.g. the Declaration of Alma-Ata, 1978), and is included under SDG 6. Safe sanitation is regarded as one of the most cost-effective health interventions (OECD, 2011; Guy Hutton and Mili Varughese, 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 improved sanitation behaviours is only part of the solution. Reversion, or 'slippage', back to unsafe sanitation behaviour 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 behaviour 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 practices 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

[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)

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 continued follow-up activities on sustaining behaviour 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 delivering health promotion and hygiene messaging through community meetings and other participatory activities. The aim is to 'trigger' collective behavioural change and encouraging villages to construct and use toilets. The original proponents of CLTS placed a heavy emphasis on using psychosocial levers (particularly disgust and shame) to trigger behaviour (Kar and Chambers, 2008). Other central tenants of the approach are that changes should be community-led and that the construction of latrines should not be subsidised. The focus of CLTS activities centre on promoting initial uptake rather than sustained change, and the approach has been criticised for a lack of attention on slippage back to OD. Gertler et al. (2015) finds that a CLTS campaign with intensive follow-up activities was effective at improving sanitation behaviour in Mali, but the authors are not able to disentangle the effect of the initial triggering meetings from the follow-up activities. The aim of this study is to disentangle those effects.

We use a sequential explanatory mixed methods approach where the design was mixed at the methodological (conceptual) stage of the study and integrated at the data interpretation (inferential) stage (Ivankova et al., 2006; Teddlie and Tashakkori, 2006; Leech and Onwuegbuzie, 2009). The sequencing consisted of two distinct phases following Tashakkori and Teddlie (1998) and Creswell et al. (2003). In the first, quantitative, phase, 123 villages that had earlier undergone CLTS mobilisation and triggering activities were randomly allocated to receive follow-up activities, or not. CLTS monitoring data were used to select a random subset of 1,191 beneficiaries for inclusion in the field experiment. Beneficiary households are those that constructed a household toilet during the CLTS implementation phase. Study participants were surveyed twice: at baseline in 2016 (before the implementation of follow-up activities) and two years later in 2018. Data was collected and main impacts were estimated before the start of the second, qualitative, phase.

This second phase, which used a multiple case study design following Stake (1995) andYin (2003), was implemented in four purposefully selected treatment and two control villages, and consisted of 21 focus group discussions (FGDs) and 32 individual interviews. It was designed to help explain and elaborate on the quantitative results obtained in the first phase, focusing on knowledge, attitudes, practices, and determinants of sustained sanitation behaviour change. The qualitative and quantitative component hence answered different aspects of the same question. The purpose of mixing qualitative and quantitative methods was to provide robustness and depth to the research findings; increase understanding of how

changes occurred (or failed to occur) as expected; and allow each method to enquire into areas that could be investigated with the other method.

We find that two years after the baseline, 30% of CLTS beneficiary households in control communities have at least one member over 5 who goes for OD as his/her main defecation practice. Such a significant non-use of toilets adopted earlier has been observed in other contexts, such Odisha, India (Barnard et al., 2013; Cairncross et al., 2005). The latter study reports that almost 20% of CLTS beneficiary households declared OD as their main practice four and 10 years after intervention implementation.[2]

Over-time decrease in adoption is not a sanitation-specific phenomenon. Banerjee et al. (2010) finds that while 30% of children have at least one vaccination injection, only 1% have being fully immunised because most children drop-out after the initial three injections. A series of studies indicate that long-term adoption happens only in the presence of positive social learning. Dupas (2014) finds that adoption of anti-malarial bed nets increased by $\approx$ 20 percentage points (ppts) one year after the initial campaign due to 'learning by doing'. On the contrary, Kremer and Miguel (2007) finds that adoption of deworming tablets dropped by 3.1 ppts for children with social ties to the cohort exposed to deworming during the two previous years, mostly due to unmet initial high expectations of effectiveness. Hence, follow-up activities that mobilise entire communities to learn about the quality of the health product can be effective at sustaining long-term behavioural change, which is what we find.

Follow-up activities were effective at sustaining behavioural change, but only in the province in which reversal to OD at baseline was the greatest.[3][4] In this province, Sindh, beneficiary households in villages allocated to follow-up activities were, on average, 14 percentage points (ppts) less likely to revert to OD, representing a 25% lower OD relative to control beneficiary households at endline. We find a comparable increase (13 ppts) in the probability of using functional toilets, indicating that the lower OD in treated villages was primarily due to sustained operability of the facilities. The effects cannot be ascribed to pre-treatment differences in sanitation behaviour nor other observables across treatment arms, and we find that they are robust to different specifications and adjustments in inference due to multiple-hypothesis testing.

We also provide suggestive evidence that the effect was sustained over time. We leverage variation in the time gap between households last exposure to follow-up activities and the endline survey to stratify the treatment sample. This analysis suggests that the effect on OD practices was almost twice as large one year after the activities finished compared to the average ITT effect. Qualitative evidence suggests that one driving factor is the volunteer staff within communities, who continued motivational engagement with households on their own accord.

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

[3] Reversal to unsafe sanitation was significantly greater in the province of Sindh than in Punjab. At endline, 57% of control households in Sindh reported OD practice, up from 26% at baseline, compared to only 13% in Punjab, up from 8%.

[4] Because the experiment was not designed to causally identify the effect of the initial CLTS campaign, reversal is not defined with respect to the OD level achieved due to this campaign.

These estimates hide meaningful heterogeneity. We find that promotion efforts are best targeted towards those that are most likely to slip back into unsafe behaviour. We analyse as heterogeneity dimensions factors that are likely to influence toilet functionality, both at the village and household levels. We demonstrate that the material circumstances in which people live in — including lack of access to public infrastructure and services, unfavourable environmental conditions of neighbourhoods, asset poverty, and poor quality of sanitation facilities — are all factors that enhance the effectiveness of follow-up activities. The baseline public infrastructure that a household has access to, as well as their sanitation quality, are the key drivers of the observed impacts. A 'joint analysis', where we include in the same regression the estimation of all heterogenous 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 quality of latrines 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.

Our results suggest that follow-up activities can sustain over a two-year period more than half the initial improvement in OD and toilet usage attained by CLTS campaigns, as estimated in the literature. Experimental evidence of the effectiveness of CLTS campaigns in sanitation behaviour yields mixed results (Abramovsky et al., 2018), 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).[5]

1 Context and intervention

Pakistan, the world's sixth 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 practised 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 to 57% in Punjab (Bureau of Statistics, Planning and Development Department, Government of Punjab, and UNICEF, 2014; Sindh Bureau of Statistics and UNICEF, 2014). Sindh is generally poorer and less developed than Punjab also on a number of other dimensions, and study districts are below their respective district averages. 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 for example with a score of 31.1 significantly lower in Badin than compared to

[5] Guiteras et al. (2015) find no effect of CLTS on sanitation behaviour in rural Bangladesh, Briceño et al. (2017) find a reduction of OD by 0.12 ppts and increase in toilet usage by 0.15 in rural Tanzania and Cameron et al. (2019) find a reduction of -0.09 ppts in OD and an increase in toilet construction by 2.4 ppts in rural Indonesia.

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 with 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, 2016). 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, 2016). A map of the location of the study is shown in the Online Appendix A1.

CLTS is a widely used community-level information and mobilisation intervention across low- and middle-income countries.

CLTS 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 programme. Next, leaders arrange a community meeting, which focuses on 'triggering' the intended beneficiaries to change their defecation behaviours. Activities conducted use the psychosocial levers of shame and disgust (at OD) and collective peer pressure. They also highlight the relationship between OD and poor health. It is expected that attendees of the triggering meeting draw up a community action plan to achieve 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 (CRP), are identified. The main task of the CRPs it to follow up on the commitments made. In our study, these CRPs received support from an NGO through trained Social Organisers (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 faeces enter the environment), the intensity and types of follow-up activities tend to vary by implementing agency and context.

The intervention we analyse 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 2014–2016. The third step of CLTS campaign was treated as a separate phase (Phase 2), implemented between 2016 and 2018 The distinction between Phases 1 and 2 was a result of funding modalities. The NGO received funding from an international NGO, which in turn received funding for implementation using a Payment by Results (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 NGO faced the potential of payment disallowances in case targets were not met.[7]

A programme 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 programmes implemented by

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

[7]Household surveys were implemented by an external agency to check on toilet construction, toilet functionality, and hygiene behaviour. These surveys were not conducted in control communities and need to therefore be seen as part of the intervention following the advice of Zwane et al. (2011).

NGOs:

"Most programme 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 programme end." (White, Zach and Colin, Jeremy, 2019)

The aim of this paper is to evaluate the effectiveness of this second phase, which we call 'follow-up activities', in terms of sustaining improvements in sanitation practices.

'Follow-up activities' centred around the community — targeting all community members in treatment communities, as well as specific stakeholders (e.g. the village committee). Two types of community gatherings were implemented. Firstly, broad-based community meetings (BBCMs), which assessed 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 assigned tasks designed during the CLTS campaign (Phase 1). Secondly, Health and Hygiene Sessions (HHS), aimed at maintaining awareness regarding personal, domestic, and environmental hygiene conditions. These sessions 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 the 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) remains the same.

The programme's monitoring data confirms 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

A sequential explanatory mixed methods approach was used to assess the effectiveness of the follow-up activities, with two distinct components, implemented sequentially (Tashakkori and Teddlie, 1998; Creswell et al., 2003; Onwuegbuzie and Johnson, 2006; Leech and Onwuegbuzie, 2009). The first, quantitative, component ('field experiment') was designed using a clustered randomised controlled trial

approach in order to assess the impact of the follow-up activities on outcomes of interest: the sustainability of sanitation behaviour—i.e. the practice of OD of at least one household member above the age of 5 years, and use of functional latrines. To gain a deeper understanding of impacts and their mechanisms, the quantitative component was coupled with a complementary qualitative research study, the data collection for which took place after completion of the field experiment with input of initial, top-level findings.

2.1 Quantitative methods

Our quantitative analysis was conducted on 123 villages randomly chosen from the complete list of 307 villages covered during the CLTS campaign in the provinces of Sindh and Punjab.[8] Of these villages, 61 were randomly assigned to receive follow-up activities ('treatment villages'). The remaining 62 villages were assigned to a control arm, where no follow-up activities took place.[9] Online Appendix Table A1 gives a break-down of sample size (both villages and households) by experimental arm and province. We use the term "beneficiary" households and households interchangeably, given that the whole sample of the study includes only households that were beneficiaries of the CLTS campaign in Phase 1.

A lot of effort went into ensuring adherence to the randomisation. 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 an MoU was signed between all implementing agencies and the research team outlining responsibilities on both sides. It was also agreed with the programme 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 NGOs programme 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 NGO, follow-up activities were indeed formally implemented in treated areas only.[10] Interviews with field staff confirm this.

We sampled study participants from the pool of households that constructed a latrine during Phase 1 of the CLTS campaign (henceforth 'beneficiary households'), which make up on average 22% of the study villages' population (31% in Sindh and 12% in Punjab). Programme 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

[8] 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.

[9] One of these programme villages had less 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 the main sample was drawn.

[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 Health Department conducted standard visits delivering health messages. These were however not specifically targeted at control communities only, but implemented routinely in villages.

(10-212 in Sindh with median of 44; 8-1337 in Punjab, median of 54); see Online Appendix A2 for the distribution of sampled households per village).

We used programme monitoring data from the time of the baseline survey to randomly select a subset of 1,191 beneficiary households. 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. Similarly, the proportion of beneficiary households within a community cannot be taken as a proxy of latrine coverage in those villages. As other households within the villages may have constructed a latrine prior to, or after, Phase 1. This sampling strategy did not seek to be representative of village-level characteristics, but was rather designed to specifically evaluate sustained behavioural change after an initial intervention, 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 programme switched from Phase 1 to Phase 2 activities. An endline survey was administered two years later, coinciding with the end of the programme in 2018. We were able to re-interview 95% of the original households. As we show in the Online Appendix, Table A6, larger households and poorer (measured as those with a higher BISP –Benazir Income Support Programme– Poverty score) were less likely to attrit. The attrition rate (5%) is balanced across treatment arms and several other baseline household characteristics.

Figure 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: (1) is the start of Phase 1 (CLTS mobilisation and triggering); (2) indicates the start of Phase 2 (follow-up activities), which ends at point (3). Point (2), which was midway through the intervention, was the start of the RCT Research Study and data collection. Our data hence do not allow us to make a judgement about the impact of the CLTS campaign in this context. The fundamental focus of this research design is to assess the additional benefit of the follow-up activities, therefore speaking solely to the right-hand side of Figure 1. However, as outlined above, all households sampled were recorded by the implementing NGO as having constructed a latrine during Phase 1 of the programme. As such the research team have confidence they were beneficiaries of the programme. What is not known for Phase 1 is the counterfactual.

['Figure 1: Research design here']

At baseline, at least one household member practices OD (self-reported) in 17% of study households (way higher in the province Sindh: 30%) while at least one uses a functioning latrine (self-reported complemented with latrine observation) in 83% of households. In line, 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 that there was either not complete behaviour change among beneficiary households at the start of the study, or that slippage back to open defecation

had already taken place.[11]

Study households typically consist of seven members, equally split between males and females, and on average one child under five, as shown in Appendix Table A4. 90% of household heads are male, on average in their early-forties, and with an average 3 years of schooling. 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 than Punjab, as shown in Appendix Table A5. Only 16% are asset rich.[12] 26% of study households have a sanitation facility of high quality and 12% share the facility. 69% have a latrine of improved technology with water seal.[13]

Furthermore, an average village in Sindh is poorer compared to the whole sample of study. 43% of villages have high availability of public infrastructure and 31% a good supply of sanitation goods and services.[14] 85% have had a flood/drought during the last two years. 87% of villages report that they have officially been declared ODF. See Online Appendix, Table A3 for more information on how each variable is computed. 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 follow-up activities of a CLTS campaign on sanitation practices, using a cluster randomised assignment to treatment. We focus on the intention-to-treat (ITT) estimates.[15] Because randomization was successful in creating observationally-equivalent groups in the experiment, and attrition was random, we can estimate treatment effects by restricting the sample to post-baseline observations. We estimate treatment effects Tj on the sanitation practices Y_{ij}, which we also refer to as 'sanitation behaviour', in household or individual i living in village j using the following ANCOVA specification:

$$Y_{ij} = \beta_0 + \beta_1 T_j + \beta_2 Y_{ij}^0 + \epsilon_{ij} \tag{1}$$

[11] Another possible explanations for this finding is non-sample error in the survey implementation, including the misidentification of beneficiary households, and errors in the underlying data set of programme 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, that included working with social organisers and CRPs to confirm households, and that the programme monitoring data were independently verified under the PbR modality.

[12] 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.

[13] High quality sanitation defined as above the median of the baseline distribution of the computed sanitation quality index that includes having a permanent superstructure, a roof, a form of closure, requires repairs, and being clean.

[14] 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.

[15] Though conservative, an ITT analysis ensures that estimates are not subject to bias arising from selection, i.e. those that choose not to join activities may be different from other beneficiary households in the village. Further, ITT is the more 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 programme than a measure of impacts on those reached.

where T_j is an indicator variable equal to one if the 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 α_1 captures the average ITT estimate.

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

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

where d_j is an indicator for the heterogeneity dimension measured at the village-level (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. In this analysis, we adjust p-values for multiple-hypothesis testing using the bootstrap-based procedure proposed by List et al. (2019).[16] 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 quantitative study was complemented by a qualitative research study intentionally sequenced *after* the completion of the endline data collection to avoid confounding the effects of the main treatment. It was designed at the village and household levels to contextualise and establish a catalogue of possible responses to the quantitative findings, and to establish and explain motivations and mechanisms.

The qualitative research was conducted in the second quarter of 2018 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 lays 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 with prevalence or population size. As such, a small number of cases (villages) were selected that maximised the range of variation on dimensions of interest to the RCT study while answering the 'whys' and 'hows' of the endline quantitative survey findings. The two primary indicators of interest used to shortlist programme villages for inclusion as qualitative sample case studies were the (1) 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 geographical close to the chosen treatment communities, in order to ensure efficiencies in time and logistics and to allow better comparison

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

across potentially similarly contextualised communities.

Semi-structured interviews were conducted by a local team fluent in the local languages. A total of 20 community 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 programme staff prior to the village-level fieldwork.

Figure A2 in the Online Appendix summarises the qualitative sample reached as part of the research. Respondents for each instrument type (FGDs, IDIs, and KIIs) were purposively 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 in the study villages. As such, the qualitative results report knowledge, attitudes, practices, and determinants of sanitation behaviour change in both CLTS beneficiaries (in Phase 1, as described above) 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 analysing 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. Thematic analysis rejects a quantitative approach to analysing qualitative data (such as frequency or cluster analysis) and instead focuses on interpretation of the stories and experiences shared by participants in order to identify and examine themes in as rigorous way as possible. As such, the first stage of 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 organised 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 (2016) notes, 'all coding is a judgement call' and team members had to be cognisant of potential own biases and of the fact that this coding process is subjective to a degree.

Following the analysis of the qualitative data, the quantitative and qualitative research teams met to discuss and compare findings and piece together the emerging story. This gave the team a chance to validate findings and discuss areas of (dis)agreement and to go back to the data for further analysis. This process of triangulating and bringing together the quantitative and qualitative findings was iterative throughout the analysis and writing process.

Throughout this paper we make use of illustrative quotes that were identified as helpful 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 Main results

We focus on two key outcome variables: whether at least one household member above five years of age practised OD during the previous two weeks, and whether the dwelling has a functional toilet being used by household members. Both outcomes are self-reported, but the second outcome was also corroborated by observations.

We first find that among beneficiaries of the Phase 1 CLTS campaign, reversal to OD was common. Concentrating first on the whole sample, Figure 2, Panel A, 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, 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 behaviour], 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).[17]

['Figure 2: OD and functional latrine levels at baseline and endline, by treatment arm' here]

Panel A also indicates that OD rates are lower in treated than in control communities, however only marginally so (30% in treatment compared to 34% in control). 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 to control households (see Online Appendix, Table A9). 35% of households in the control villages report having heard about, and 25% to have 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. This could be either due to recall error or other promotion efforts.[18] Or because some CRPs report that they continued

[17] 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)). The fact that this quote comes from a treatment community underlines the extreme case sampling applied for the qualitative part of the study.

[18] 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 were 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.

to work on their own accord after the initial CLTS campaign had ended, as highlighted during qualitative interviews:

'The programme 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 emphasised 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 on our own accord. The [NGO] people have not contacted us since.' (Male CRP KII, C).

Turning back to Figure 2, Panels B 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 to only 13% in Punjab, up from 8%), differences between experimental arms at endline are also much more noteworthy in this province at 13ppts (44% OD for treated and 57% for control households). In Punjab, the difference stands at only 3ppts (16% versus 13%).

We next confirm these differences more formally in Table 1. This table presents the estimated ITT effect of being allocated to receive follow-up activities on our two pre-specified outcomes. We show estimates with and without accounting for the baseline value of the dependent variable (ANCOVA model). We present the results in the full sample (Panel A), and splitting the sample by province: Sindh in Panel B and Punjab in Panel C. We are able to split the sample as such because the randomization was stratified by province. Randomisation was successful in both provinces, as shown in Appendix, Balance Table A4.[19]

['Table 1: Effect of follow-up activities on prevalence of OD and using a functional toilet' here]

While coefficients have the expected signs, suggesting an average reduction in reversal to OD by 5ppts (and increase in use of functioning latrine by 6ppts), these impacts are not statistically significant. We find, however that the heterogeneity by province we noted in Figure 2 is confirmed: estimates for Punjab (Panel B) suggest null impacts, while follow-up activities decreased the probability of OD significantly in Sindh (Panel B), and did so by 14 ppts among treated households, which translates into 25% lower OD compared to control villages at endline. The effect size is consistent across specifications, but more precisely estimated using ANCOVA, significant at the 5% level. The reduction in OD is accompanied by a comparable increase (of 13 ppts) in the probability of using a functioning toilet, significant at the 10% level. 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 household at endline that report

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

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, T2)

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 8% of households reporting to openly defecate, compared to more than 25% in Sindh (as Figure 2 shows). Arguably, follow-up activities in Punjab were thus not needed on average, while in Sindh follow-up activities 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 unfavourable. 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 treatment effects and heterogeneity in observed impacts, will focus on Sindh in order to allow for a clean separation between initial slippage to OD, geographical features and implementation modality/partner.[20]. 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 treatment effects

Follow-up activities are designed to address the sustainability of sanitation infrastructure and behaviour. 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. 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 week, within the first quarter of 2018, the variation comes purely from differences in intervention implementation[21]

The average gap is four and a half months, with a minimum of three months and a maximum of 12 months. The time gap was three months for 65% of treated households, six months for 18% of treated households, and more than nine months for 17% of households.

Ideally, this time gap was random across communities. However, the implementation of activities was

[20]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 sanitation quality above the median of the baseline distribution is not present anymore in the Sindh sample. See Appendix Table A5

[21]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. 3.53% households the intervention ended in 2017-Q1; in 13.33% households in 2017-Q2; in 18.43% households in 2017-Q3; and, in 64% households in 2017-Q4.

conducted in blocks of 10 villages and social organizers visiting the less remote ones first. In line, when analysing selection into the time-gap, we find that the those with three- and six-month time gaps had access to better public infrastructure at baseline than those who experienced a larger time gap. Where the time gap was 9+ months, sanitation facilities were more likely to be of high quality at baseline. The joint significance of baseline characteristics is balanced overall for the sample of analysis for the time gap of three months and 9+ months (82% of the sample) (see the Online Appendix, Table A10). Nevertheless, when estimating the treatment effects stratified by time gap, we control for the corresponding imbalances.

Figure 3 shows the estimated treatment effect when stratifying the treatment sample by time gap between follow-up activities and endline survey. We find evidence that follow-up activities generate a sustained behavioural change. Although the stratification makes the estimates predictably noisy, the results suggest that the reduction in OD due to follow-up activities is sustained over time and even greater 9+ months after the intervention was finalised compared to 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. We cannot reject, however, the null hypothesis that the point estimates after nine months are statistically similar to the point estimates of the other time-gap samples (three and six months after implementation). The positive effect on the use of a functional latrine is also sustained over time.

['Figure 3: Effects of follow-up activities on behaviour, by time gap between the activities and outcome measurement' here]

4 Heterogeneity

We are interested in what makes follow-up activities an effective strategy to induce sustained changes in a village's sanitation environment. Because we find that reversal to unsafe sanitation behaviour 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 heterogenous effects by characteristics collected at baseline, after Phase 1 activities were completed and before Phase 2 activities started.[22]

4.1 Heterogeneity by village-level characteristics

We find that follow-up activities are mostly effective in villages with lower availability and quality of public infrastructure, limited supply of sanitation services and with worse environmental conditions. The first two factors are measured by building indexes using item response theory. The 'public infrastructure index' includes information on the quality of the village's roads, bridges, and access to piped water, as well as availability of schools, hospital, police station, bank, market, shops, and community centre. The

[22]We show in Table A5 that the baseline heterogeneity dimensions considered are balanced between treatment and control in the Sindh sample .

'sanitation supply index' captures the availability of services needed to build latrines, including whether the village has a mason, a plumber, and a cement and brick supplier, as well as access to a sanitation hardware store. See Table A3 in the Online Appendix for more details about the construction of the indexes and their components.

The study communities are usually farming and labour communities and the majority of the households with latrines have just simple pit latrine systems. Material availability is a challenge both in terms of access and affordability. The closest sanitation supplies are usually in the nearest urban centre, and transportation can be problem. People in some communities reported travelling 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 rikshaw or Suzuki, so we used a rikshaw 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).

The third factor is a village's vulnerability to shocks that can deteriorate infrastructure directly, 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 (i.e. latrines overflow due to heavy rain, which creates negative health spillovers (Bancalari and Martinez, 2018)) and indirectly affect the ability of households to maintain latrines (i.e. temporary income shocks that reduce the capacity to invest in housing).

Latrine collapse due to rain was a key reason for latrines becoming non-functional. This was most often the case for basic, mud, or *katchi* latrines, which were more susceptible to weather conditions.[23]

'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'. (SOs 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

[23]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.

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 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] behaviour 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)

Our proxy for weather vulnerability is an indicator on whether the village was affected by a flood/drought in the two years prior to the baseline survey.

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

Follow-up activities were of relevance in villages with low availability and access to public infrastructure and high vulnerability to weather shocks. 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 (26 ppts) if allocated to receive follow-up activities. Only the interaction term "Follow-up activities x high-quality village infrastructure" is statistically significant. [24]

We see a similar pattern when looking at village sanitation supply and village 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 5.

[24] As highlighted by Crocker et al. (2017), 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 behaviour. It is therefore possible, that these village-level heterogenous 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 Appendix Table A12). We also find that the village-level heterogenous effects survive when including the interaction "Follow-up activities x Eligible OD rate (Low)". This interaction is not statistically significant (see Appendix Table A13).

['Table 2: Effect of follow-up activities on prevalence of OD and using a functional toilet, by village characteristics' here]

We next examine in which part of the distribution of village-level public infrastructure quality follow-up activities are more effective at sustaining behavioural change. Figure 4 shows local polynomial smooth plots (with confidence intervals at the 90% level) for OD (Panel A) and the 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.[25] Figure A3 in the Appendix shows that the heterogenous effects by the distribution of the sanitation supply index are not as pronounced as by the public infrastructure index. Online Appendix, Figure A4, Panel A, shows that the distributions of the village-level heterogeneity dimensions are very similar across treatment and control groups.

['Figure 4: Heterogeneous effects of follow-up activities on behaviour, by public infrastructure' here]

4.2 Heterogeneity by household-level characteristics

The previous section suggests that follow-up activities were most effective at sustaining behavioural change in communities with poorer public infrastructure. However, this may proxy overall poverty, which bundles together several constraints affecting decision making and behaviour across households. 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 caved 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, T2)

'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 labour 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, T2)

To analyse how the effectiveness of follow-up visits interacts with poverty, we use as a proxy a standard baseline household asset index (e.g. TVs, radios, vehicles, productive assets). Columns (1) and (2) of Table 3, which has the same set-up as Table 2, 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 ($\approx 27\%$ lower OD). This effect is accompanied by an increase of the same magnitude in the probability of using a functional latrine.

Beyond poverty, we disentangle this heterogeneity further by looking at factors likely influencing the

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

ability to keep a toilet functional. These factors are: quality of the sanitation facility; whether the facility is shared; and the latrine technology.[26] Although these factors are correlated with asset poverty, they are not perfectly correlated, allowing us to analyse a more specific heterogeneity dimension.[27] To examine these factors, we consider three indexes.

The first index captures sanitation quality and is created using variables that capture if the sanitation facility has a roof, curtains or door, and a functional handwashing facility, its observed cleanliness status (faeces, flies, odour), as well as whether it needs repairs.[28]

The second index summarises 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.[29]

The third index focuses on whether the latrine is of an improved technology (following the (Joint Monitoring Program WHO-UNICEF, 2017) classification) and has a water seal.[30]

Table 3 shows that follow-up activities are more effective at sustaining behaviour change when the sanitation facility is badly maintained and shared, but also when the underlying technology is an improved latrine. 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 it is shared compared to control households with those baseline characteristics, respectively ($\approx$ 26% lower OD). These heterogeneous effects are driven by an increase of the same magnitude in the probability of using a functional latrine. Notably, 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) when the initial latrine is of improved technology compared to control households with improved latrines.

The interaction term is only statistically significant for the heterogeneity along the latrine technology dimension. When adjusting p-values for multiple-hypothesis testing, the interaction is not statistically significant anymore. Only the effect of follow-up activities on the behaviour of households with low sanitation quality remains statistically significant at the 10% level.

[‘Table 3: Effect of follow-up activities on prevalence of OD and using a functional toilet, by household

[26] 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 in the latter group 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, shown in Columns (3) and (4) of Appendix Table A12)

[27] See Table A11 in the Online Appendix for correlations of these indexes with our poverty proxy.

[28] See Table A3 in the Online Appendix for more details about the construction of the indices and their components.

[29] The main reservation and concern regarding sharing outside of the household was related to overuse (i.e. pits getting filled very quickly) and cleanliness.

[30] Qualitative findings suggest that the intervention had some influence in getting households to upgrade their latrine structure. One female IDI respondent (T1) expressed it as follows: *‘We built our latrine five years ago... Before this [programme] 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.’* .

characteristics' here]

We next examine in which part of the distribution of sanitation quality index follow-up activities are more effective at sustaining behavioural change. Figure 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 quality 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 quality index.[31] Online Appendix, Figure A3 shows that heterogenous effects by the distribution of the household asset index are not as pronounced as by the sanitation supply index.

['Figure 5: Heterogeneous effects of follow-up activities on behaviour, by sanitation quality' here]

4.3 Predominant heterogeneity

We show in this section that households living in villages with worse public infrastructure and initially using a sanitation facility of low quality need follow-up activities the most to prevent slippage back into OD. We provide suggestive evidence that follow-up activities incentivized households living in these poor conditions to conduct regular maintenance of their latrines.

Given that the dimensions of heterogeneity we consider are correlated (see Table A11 in the Appendix for pairwise correlations of heterogeneity dimensions at baseline), 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 equations (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 analyse when are follow-up activities most effective. The estimate of the average effect of follow-up activities on OD and functional latrine remains similar in magnitude, but it is not precisely estimated anymore. Interestingly, we find evidence that neither household asset poverty nor whether the toilet is shared are the main mechanisms explaining the results. What survives the joint analysis are the heterogeneous effects by village-level public infrastructure and the household-level sanitation quality. The drop in OD (increase in functional latrines) is higher in villages with worse public infrastructure, and in households with poorer sanitation quality.

['Table 4: Joint analysis' here]

To understand well the underlying mechanism that makes follow-up activities effective, we conduct one final exercise. Namely, we analyse if follow-up activities have an effect on the probability of a household

[31] Figure A4 in the Appendix shows that the distribution of each index is similar across treatment arms

having maintained their latrine in the previous two years, and we look at heterogeneous effects along the dimensions that survived the joint analysis exercise. We find that follow-up activities increased the probability of maintaining latrines by 7 ppts if the baseline sanitation quality is low compared to control households with low sanitation maintenance (87% higher maintenance; see Table A15 in the Appendix). However, the interaction term is not statistically significant, and nor are the average effects.

5 Robustness checks

We conduct the following robustness checks on our analysis.

To start with, in order to check that our results are not dependent on our chosen set of controls, we estimate our average and heterogeneous effects using the Post-Double Selection LASSO procedure for selection of control variables (Tibshirani, 1996; Belloni et al., 2014). Table A16 in the Appendix shows that the average effects, both with and without ANCOVA, remain robust and are highly precisely estimated. Online Appendix, Table A17 shows that the heterogenous effects by village public infrastructure remain robust and the heterogeneous effect along weather vulnerability is now statistically significant. Treated households that suffer from weather shocks (i.e. floods or draught) are 30 ppts less likely to practice OD and 41 ppts more likely to use a functional latrine, compared to control households that also suffer from these shocks. Heterogeneity by household assets and sanitation facility characteristics, however, are not statistically significant anymore (see Table A18 in the Online Appendix). We find that follow-up activities are 19 ppts more effective at increasing the use of functional latrines when the latrine technology is improved.[32]

A second robustness check is conducted by exploiting the full sample of study households and villages, pooling together Sindh and Punjab provinces. Estimating heterogenous effects in the full sample increases the statistical power to detect heterogenous effects. We re-estimate our heterogeneity analysis on the pooled sample, adding province stratification dummies. Our heterogeneous results remain robust when conducting the analysis including Punjab. The heterogenous 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 15 ppts larger (and statistically significant) for households with the asset and sanitation quality indexes below the median at baseline compared to 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 quality below the median at baseline compared to similar control households (see Tables A19 and A20 in the Online Appendix). 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 quality index (see Table A21 in

[32] We run an additional robustness check where we exclude two outlier villages in terms of the number of beneficiaries from the analysis. See Online Appendix **??** for the distribution. Results become, if anything, stronger, and are available upon request.

the Online Appendix). Treated households with an improved latrine with 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 to control households with an improved latrine.

6 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 three main contributions toward this end.

First, we show that follow-up activities can be an effective tool to achieve sustained sanitation behaviour change and reduce fall-back to OD significantly – by 14 ppts in the context studied. This is a considerable impact in terms of magnitude. Rigorous studies suggest that the average impact of sanitation interventions on OD ranges between 4 and 19 ppts (Patil et al., 2014; Guiteras et al., 2015; Augsburg et al., 2020), and that impacts are significant in poor contexts only (Abramovsky et al., 2018) (such as in very poor areas of Mali, where CLTS led to a decrease by more than 23ppts in OD; see Pickering et al. (2015)). The importance of follow-up activities has long been stressed by practitioners, yet has often been overlooked in programme implementation (Hailemariam et al., 2019). Our findings fill a critical evidence gap by providing robust evidence to support this view, and thereby respond to the call of more research on CLTS implementation, as stressed in a recent reviews (Venkataramanan et al., 2018; Chirgwin et al., 2021).

Second, our heterogeneity analysis provides guidance on how implementers can most effectively target limited resources. Follow-up activities are likely to be effective when targeted to where the initial latrine quality is poor and in communities with inferior public infrastructure and where latrines are more likely to collapse. Either strategy is feasible to implement within the CLTS framework. Implementing agencies can use secondary data to target follow-up activities across communities. Additionally, within communities, CRPs, who typically know their communities well, can be trained to target their support to specific households.

Third, follow-up activities implemented as part of CLTS achieve convergence: households living in more difficult environments are supported in achieving sustained OD behaviour. This results in more equitable public good delivery, and ultimately a more equitable society.

However, follow-up activities are not a silver bullet. For one, while they reduce slippage back to unhealthy sanitation behaviour, they do not eliminate it. Notwithstanding the environment, we show that such activities enable households to sustain healthy sanitation behaviour that are initially better placed to do so (i.e. that already owned an improved latrine). For those that do not, study households perceived lack of finance and lack of access to toilet construction materials to be important constraints, as shown in other contexts (Ben Yishay et al., 2017; Lipscomb and Schechter, 2018; Augsburg and Rodríguez-Lesmes, 2020). Our results demonstrate that follow-up activities are an essential and effective component

of sanitation promotion programmes, but there remains a need to further develop strategies for sustaining behaviour change.

References

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

Augsburg, B., B. Caeyers, S. Giunti, and S. Smets (2020). Labelled Loans and Human Capital Investments. *Institute for Fiscal Studies Working Paper W20/20*.

Augsburg, B. and P. Rodríguez-Lesmes (2020, dec). Sanitation dynamics: Toilet acquisition and its economic and social implications in rural and urban contexts. *Journal of Water Sanitation and Hygiene for Development 10*(4), 628–641.

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

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

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

Banerjee, A. V., S. M. Barnhardt, and E. Duflo (2015). Movies, Margins, and Marketing: Encouraging the Adoption of Iron-Fortified Salt. In *Insights in the Economics of Aging*, pp. 285–306. National Bureau of Economic Research, Inc.

Barnard, S., P. Routray, F. Majorin, R. Peletz, S. Boisson, A. Sinha, and T. Clasen (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).

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives 28*(2), 29–50.

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

Bjorkman Nyqvist, M., D. De Walque, and J. Svensson (2017). Experimental Evidence on the Long-Run Impact of Community-Based Monitoring. *American Economic Journal: Applied Economics 9*(1), 33–69.

Briceño, B., A. Coville, P. Gertler, and S. Martinez (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., K. Shordt, S. Zacharia, and B. K. Govindan (2005). What Causes Sustainable changes in Hygienic Behaviour? A Cross-Sectional Study from Kerala, India. *Social Science and Medicine 61*(10), 2212–20.

Cameron, L., S. Olivia, and M. Shah (2019, may). Scaling up sanitation: Evidence from an RCT in Indonesia. *Journal of Development Economics 138*, 1–16.

Chirgwin, H., S. Cairncross, D. Zehra, and H. Waddington (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 Systematic Reviews 17*.

Chong, A., M. Gonzalez-Navarro, D. Karlan, and M. Valdivia (2020, jun). Do Information Technologies Improve Teenagers' Sexual Education? Evidence from a Randomized Evaluation in Colombia. *World Bank Economic Review 34*(2), 371–392.

Creswell, J. W., V. L. Piano Clark, M. Gutmann, and W. Hanson (2003). Advanced mixed methods research designs.

Crocker, J., D. Saywell, and J. Bartram (2017). Sustainability of community-led total sanitation outcomes: Evidence from Ethiopia and Ghana. *International Journal of Hygiene and Environmental Health 220*, 551–557.

DellaVigna, S. and U. Malmendier (2006). Paying not to go to the gym. *American Economic Review 96*(3), 694–719.

Duflo, E., P. Dupas, and M. Kremer (2015, sep). Education, HIV, and early fertility: Experimental evidence from Kenya. *American Economic Review 105*(9), 2757–2797.

Dupas, P. (2009, may). What matters (and What Does Not) in households' decision to invest in malaria prevention? *American Economic Review: Papers and Proceedings 99*(2), 224–230.

Dupas, P. (2011). Health Behavior in Developing Countries. *Annual Review of Economics 3*, 425–449.

Dupas, P. (2014). Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment. *Econometrica 82*(1), 197–228.

Dupas, P. and E. Miguel (2017). Impacts and Determinants of Health Levels in Low-Income Countries. In A. Banerjee and E. Duflo (Eds.), *Handbook of Economic Field Experiments*, Volume 2. Elsevier B.V.

Gertler, P., M. Shah, M. L. Alzua, L. Cameron, S. Martinez, and S. Patil (2015). How does Health Promotion Work? Evidence from the dirty Business of Eliminating Open Defecation.

Government of the Islamic Republic of Pakistan Ministry of Environment (2006, September). National Sanitation Policy.

Government of the Islamic Republic of Pakistan Ministry of Environment (2011, March). Pakistan Approach to Total Sanitation (PATS).

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

Guy Hutton and Mili Varughese (2016). The Costs of Meeting the 2030 Sustainable Development Goal Targets on Drinking Water, Sanitation, and Hygiene. Technical report, World Bank, Washington D.C.

Hailemariam, M., T. Bustos, B. Montgomery, R. Barajas, L. B. Evans, and A. Drahota (2019, jun). Evidence-based intervention sustainability strategies: A systematic review. *Implementation Science 14*(1), 57.

Hussam, R., A. Rabbani, G. Reggiani, and N. Rigol (2021). Rational Habit Formation: Experimental Evidence from Handwashing in India. *American Economic Journal: Applied Economics Forthcomin*.

Ivankova, N. V., J. W. Creswell, and S. L. Stick (2006). *Using Mixed-Methods Sequential Explanatory Design:From Theory to Practice*, Volume 18. Sage Publications.

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.

Kar, K. and R. Chambers (2008). *Handbook on Community-Led Total Sanitation*, Volume 44. Institute of Development Studies.

Kremer, M. and E. Miguel (2007). The illusion of sustainability. *Quarterly Journal of Economics 112*(3), 1007 – 1065.

Kresche, E., M. Lipscomb, and L. Schechter (2020). Externalities and Spillovers from Sanitation and Waste Management in Urban and Rural Neighborhoods. *Applied Economic Perspectives and Policy 42*, 395–420.

Leech, N. and A. Onwuegbuzie (2009). A Typology of Mixed Methods Research Designs.

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

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

Madajewicz, M., A. Pfaff, A. van Geen, J. Graziano, I. Hussein, H. Momotaj, R. Sylvi, and H. Ahsan (2007, nov). Can information alone change behavior? Response to arsenic contamination of groundwater in Bangladesh. *Journal of Development Economics 84*(2), 731–754.

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

OECD (2011). Benefits of Investing in Water and Sanitation: An OECD Perspective. Technical report, OECD, Paris.

Onwuegbuzie, A. J. and R. B. Johnson (2006). The Validity Issues in Mixed Research.

Orgill-Meyer, J., S. K. Pattanayak, N. Chindarkar, K. L. Dickinson, U. Panda, S. Rai, B. Sahoo, A. Singha, and M. Jeuland (2019, aug). Long-term impact of a community-led sanitation campaign in India, 2005–2016. *Bulletin of the World Health Organization 97*(8), 523–533A.

Patil, S. R., B. F. Arnold, A. L. Salvatore, B. Briceno, S. Ganguly, J. M. Colford, and P. J. Gertler (2014, aug). The Effect of India's Total Sanitation Campaign on Defecation Behaviors and Child Health in Rural Madhya Pradesh: A Cluster Randomized Controlled Trial. *PLoS Medicine 11*(8), e1001709.

Pickering, A. J., H. Djebbari, C. Lopez, M. Coulibaly, and M. L. Alzua (2015). Effect of a community-led sanitation intervention on child diarrhoea and child growth in rural Mali: a cluster-randomised controlled trial. *The Lancet Global Health 3*(11), e701–e711.

Prüss-Ustün, A., J. Bartram, T. Clasen, J. M. Colford, O. Cumming, V. Curtis, S. Bonjour, A. D. Dangour, J. De France, L. Fewtrell, M. C. Freeman, B. Gordon, P. R. Hunter, R. B. Johnston, C. Mathers, D. Mäusezahl, K. Medlicott, M. Neira, M. Stocks, J. Wolf, and S. Cairncross (2014, aug). Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: a retrospective analysis of data from 145 countries. *Tropical medicine and international health 19*(8), 894–905.

Prüss-Ustün, A., J. Wolf, J. Bartram, T. Clasen, O. Cumming, M. C. Freeman, B. Gordon, P. R. Hunter, K. Medlicott, and R. Johnston (2019, June). *International Journal of Hygiene and Environmental Health*, 765–777.

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., R. Maertens, K. Matin Ahmed, and A. van Geen (2020). Demand for Information on Environmental Health Risk, Mode of Delivery, and Behavioral Change: Evidence from Sonargaon, Bangladesh. *The World Bank Economic Review 35*, 1–29.

Tashakkori, A. and C. Teddlie (1998). Mixed methodology: Combining qualitative and quantitative approaches.

Teddlie, C. and A. Tashakkori (2006). A General Typology of Research Designs Featuring Mixed Methods.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*(Methodological), 267–288.

Tyndale-Biscoe, P., M. Bond, and R. Kidd (2013). ODF Sustainability Study - Sanitation Learning HubSanitation Learning Hub. Technical report, Plan International.

UNDP (2016). Human Development Report 2016 Human Development for Everyone. Technical report, United National Development Program, New York.

UNICEF (2014). EVALUATION OF THE WASH SECTOR STRATEGY "COMMUNITY APPROACHES TO TOTAL SANITATION" (CATS) Executive Summary EVALUATION REPORT EVALUATION OFFICE. Technical report, UNICEF2014.

Venkataramanan, V., J. Crocker, A. Karon, and J. Bartram (2018). Community-led total sanitation: A mixed-methods systematic review of evidence and its quality. *Environmental Health Perspectives 126*.

White, Zach and Colin, Jeremy (2019). Monitoring, Evaluation, and Verification Component of the WASH Results Programme.

Whittington, D., M. Radin, and M. Jeuland (2020, jan). Evidence-based policy analysis? The strange case of the randomized controlled trials of community-led total sanitation. *Oxford Review of Economic Policy 36*(1), 191–221.

World Bank (2018). Sanitation Overview.

Yin, R. (2003). *Case study research: Design and methods*. Number 3. Sage Publications.

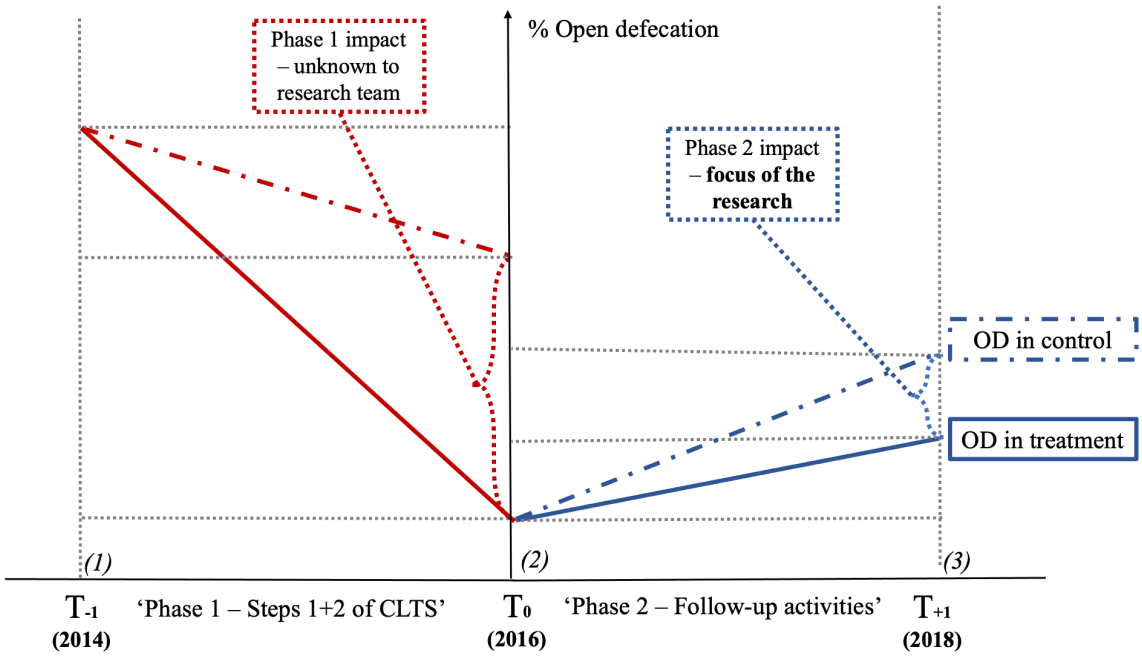
Zuin, V., C. Delaire, R. Peletz, A. Cock-Esteb, R. Khush, and J. Albert (2019, dec). Policy Diffusion in the Rural Sanitation Sector: Lessons from Community-Led Total Sanitation (CLTS). *World Development 124*, 104643.

Zwane, A. P., J. Zinman, E. Van Dusen, W. Pariente, C. Null, E. Miguel, M. Kremer, D. S. Karlan, R. Hornbeck, X. Giné, E. Duflo, F. Devoto, B. Crepon, and A. Banerjee (2011, feb). Being surveyed

can change later behavior and related parameter estimates. *Proceedings of the National Academy of Sciences of the United States of America 108*(5), 1821–1826.

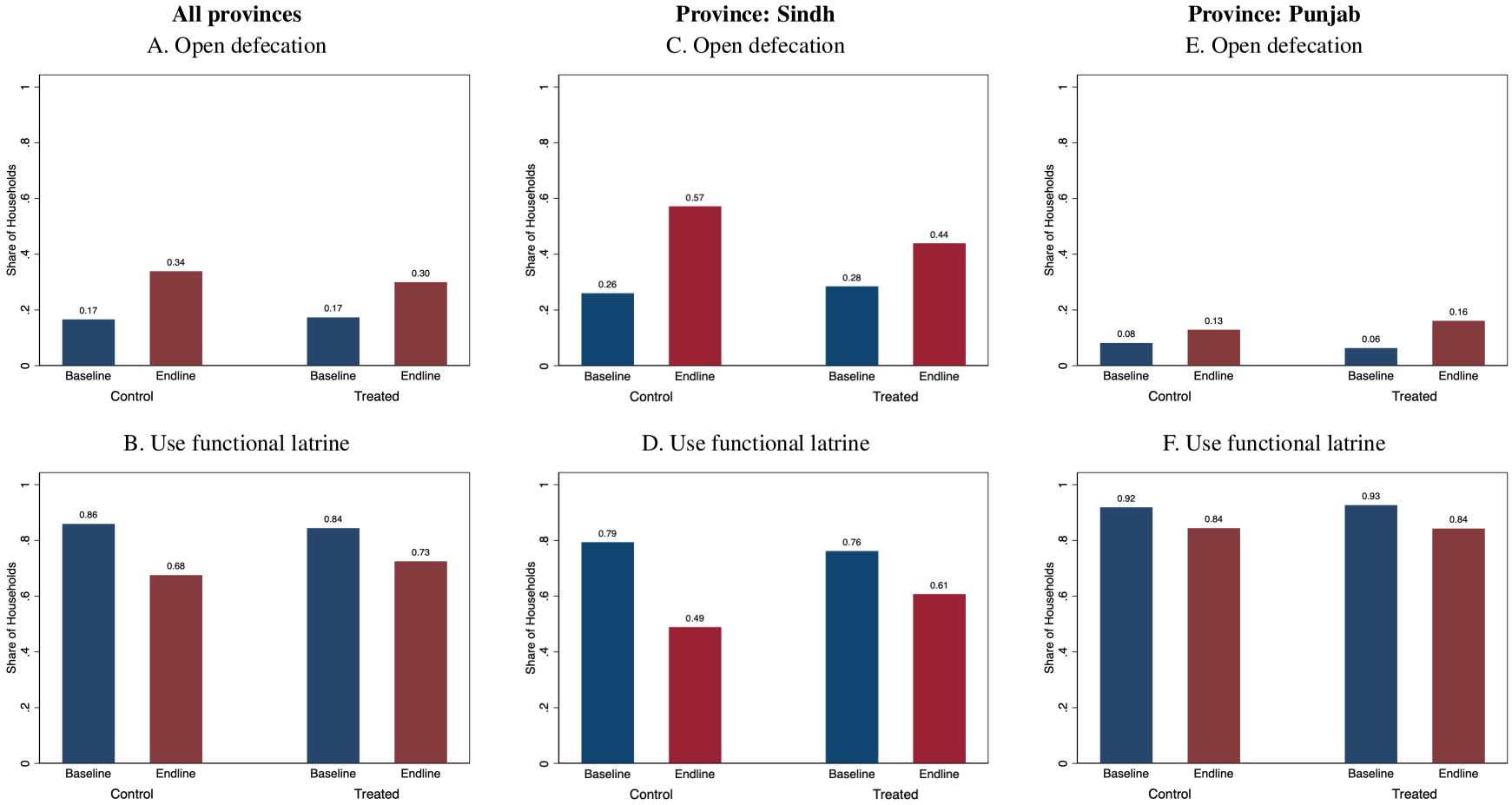
Figures

Figure 1: Research design



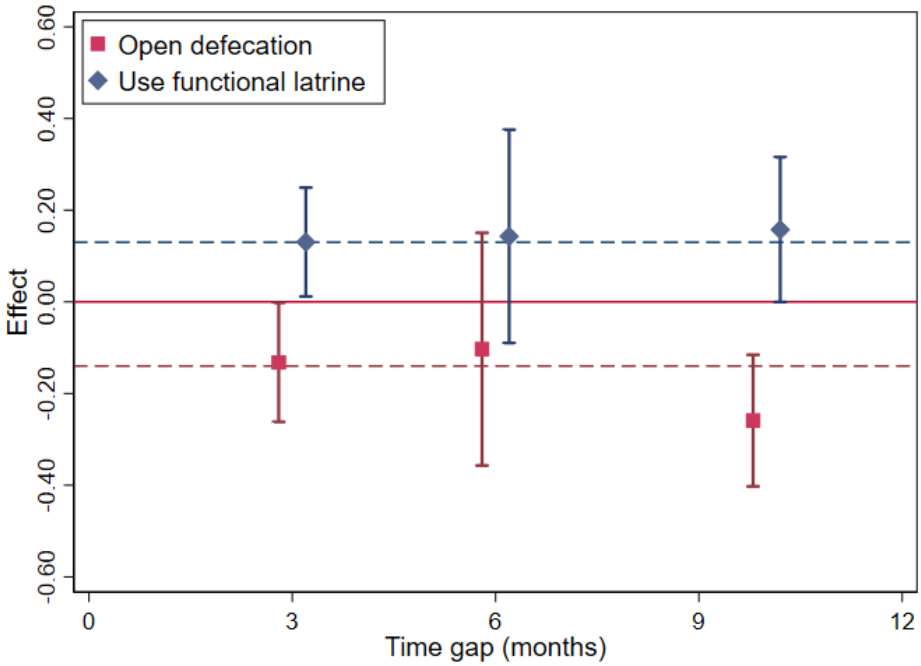
Notes. This figure presents a graphical representation of the research design. The y-axis represents the percentage of beneficiary households defecating in the open. The y-axis represents time: (1) start of Phase 1 (CLTS mobilisation and triggering); (2) start of the Phase 1 (follow-up activities), which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

Figure 2: Mean behaviour at baseline and endline, by treatment arm



Notes. 'Open defecation' (Panels A, C and E) is an indicator of whether a household reports at least 1 member older than 5 years old that openly defecates; and 'Use functional latrine' (Panels B, D and F) is an indicator of whether the household has a functional latrine in the dwelling and members use it.

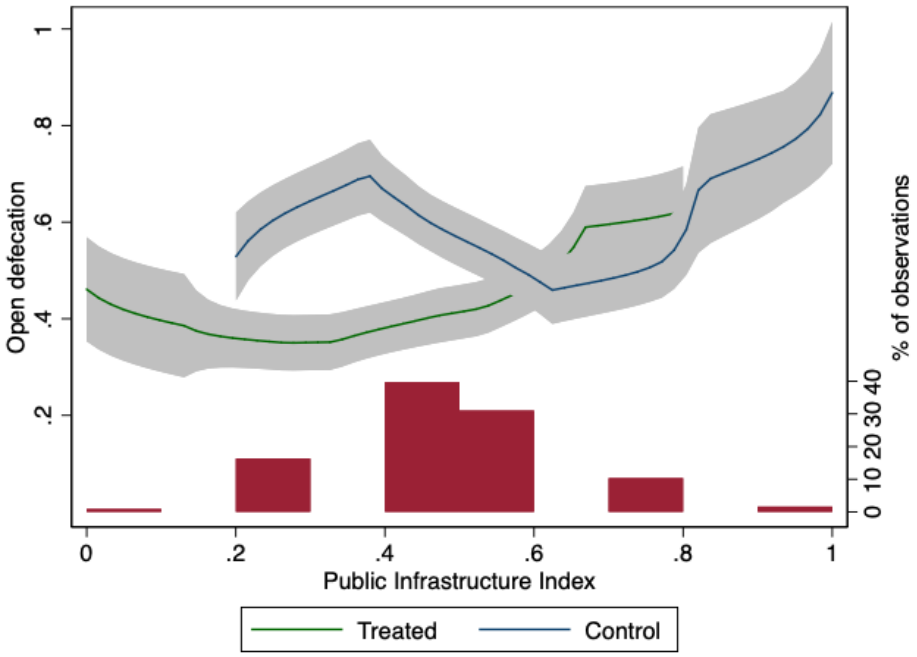
Figure 3: Effects of follow-up activities on behaviour, by time gap between the activities and outcome measurement



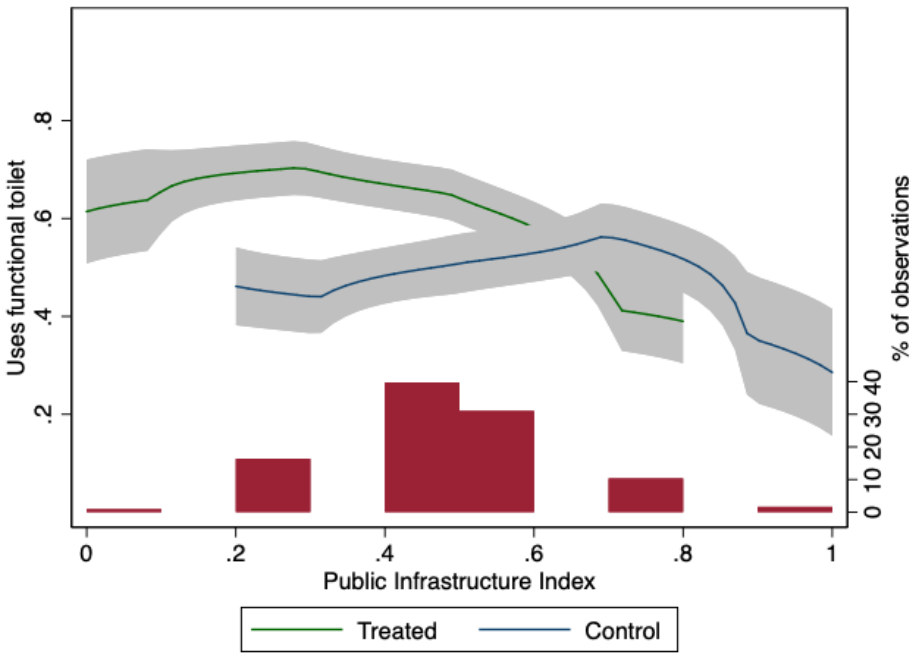
Notes. 'Open defecation' is an indicator of whether a household has at-least 1 member older than 5 years that openly defecates; and 'Functional latrine' of whether the household has a functional latrine in the dwelling used by its members. The dotted lines show the average treatment effect (linear probabilities) for the whole sample. The plots indicate the point estimate of the effects of follow-up activities in each time gap and 95 percent confidence interval. Time gap, indicated in the x-axis, is measured as the months elapsed between the end follow-up activities and the measurement of outcomes. All estimations include the following baseline characteristics correlated with the time gap: number of children in the household, gender and age of the household head, sanitation quality index, whether latrine is improved and the village-level public infrastructure index. Sample includes households interviewed at endline in Sindh.

Figure 4: Heterogeneous effects of follow-up activities on behaviour, by public infrastructure

A. Open defecation



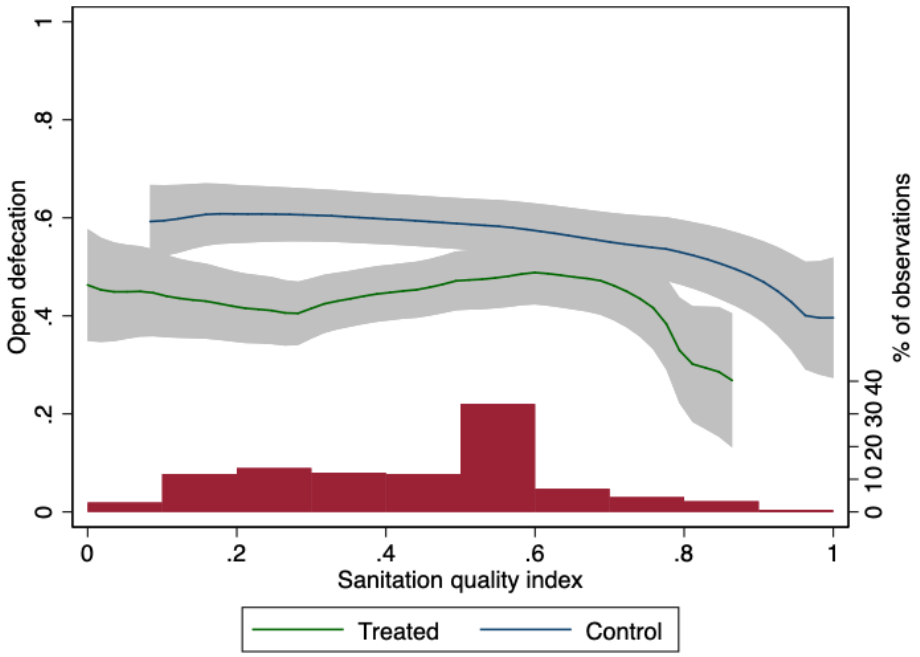
B. Use functional latrine



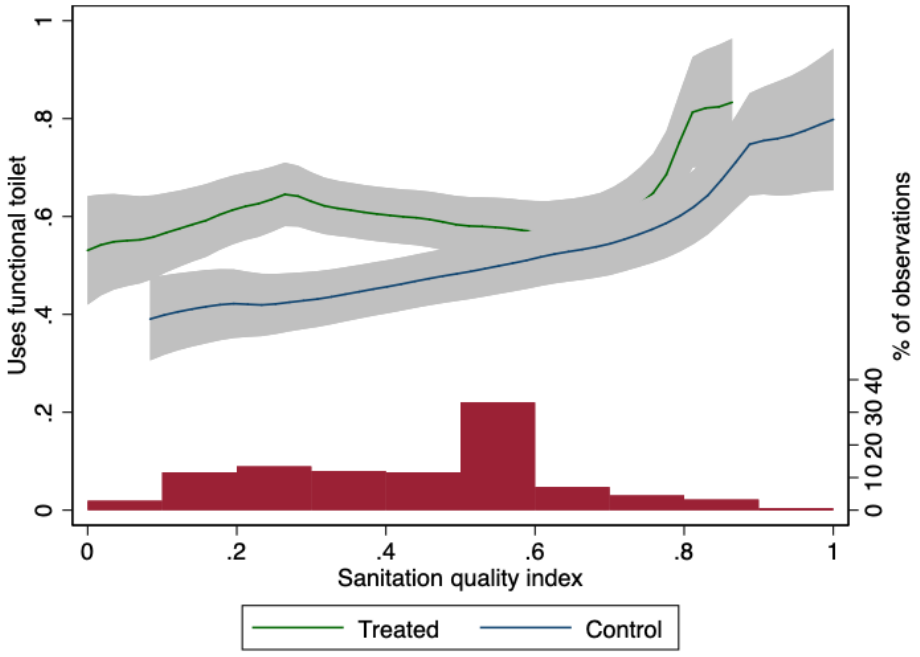
Notes. 'Open defecation' (Figure A) is an indicator of whether a household has at-least 1 member older than 5 years that openly defecates. 'Use functional latrine' (columns (2), (4) and (6)) is an indicator for whether the household has a functional latrine in the dwelling used by its members. 'Public infrastructure' is an index capturing the availability of infrastructure in the village. The histogram located at the bottom of each plot shows the distribution of observations. Sample includes households interviewed at endline in Sindh.

Figure 5: Heterogeneous effects of follow-up activities on behaviour, by sanitation quality

A. Open defecation



B. Use functional latrine



Notes. 'Open defecation' (Figure A) is an indicator of whether a household has at-least 1 member older than 5 years that openly defecates. 'Use functional latrine' (columns (2), (4) and (6)) is an indicator for whether the household has a functional latrine in the dwelling used by its members. 'Sanitation quality' is an index capturing the quality of the bathroom located in the household's dwelling. The histogram located at the bottom of each plot shows the distribution of observations by the sanitation quality index. Sample includes households interviewed at endline in Sindh.

Table 1: Effect of follow-up activities on behaviour

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

Notes. Primary data, household level. 'Open defecation' (columns (1) and (2)) is an indicator of whether a household reports at least 1 member older than 5 years old that openly defecates. 'Use functional latrine' (columns (3) and (4)) is an indicator of whether the household has a functional latrine in the dwelling and members use it. 'Follow-up activities' is the β_1 parameter from equation 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 parenthesis) are clustered at the unit of randomization (villages). Statistical significance denoted by * p¡0.1 , ** p¡0.05, and *** p¡0.01.

Table 2: Heterogeneous effects of follow-up activities on behaviour, by village characteristics

Outcome	Public infrastructure		Sanitation supply		Weather vulnerability	
	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)
	[0.05]	[0.05]	[0.13]	[0.32]	[0.97]	[0.83]
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)
	[0.19]	[0.08]	[0.77]	[0.92]	[0.82]	[0.92]
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)
	[0.98]	[0.68]	[0.65]	[0.59]	[0.11]	[0.26]
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 of whether a household has at-least 1 member older than 5 years that 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 having 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 2 years. 'Follow-up activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (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.

Table 3: Heterogeneous effects of follow-up activities on behaviour, by household characteristics

Outcome	Household assets		Sanitation quality		Shared facility		Latrine technology	
	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)
	[0.13]	[0.23]	[0.09]	[0.16]	[0.26]	[0.25]	[0.43]	[0.64]
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)
	[0.79]	[0.97]	[0.70]	[0.63]	[0.84]	[0.90]	[0.43]	[0.28]
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)
	[0.67]	[0.5]	[0.44]	[0.74]	[0.48]	[0.81]	[0.12]	[0.07]
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) and (5)) is an indicator of whether a household has at-least 1 member older than 5 years that 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. 'Household assets' (columns (1) and (2)) is an indicator for being above the median household asset index index. 'Sanitation quality' (columns (3) and (4)) is an indicator for being above the median of the sanitation quality index. 'Shared facility' (columns (5) and (6)) is an indicator for whether the household bathroom and latrine are shared. '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 activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (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.

Table 4: Joint analysis

Outcome	Open defecation	Use functional latrine
	(1)	(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 quality (Low)	-0.16*	0.18*
	(0.09)	(0.10)
FU x Shared latrine (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 of whether a household has at-least 1 member older than 5 years that 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 *** $p<0.01$.

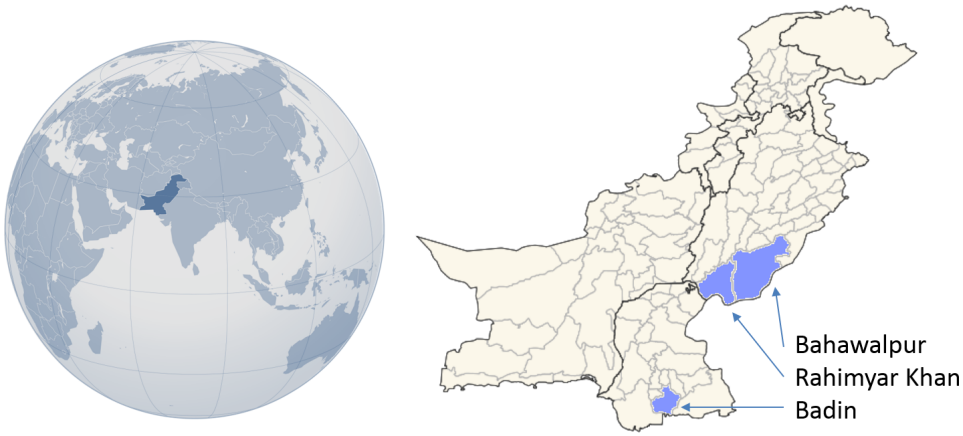
ONLINE APPENDIX

This Online Appendix provides additional information on the data collection and robustness checks.

Appendix 1 Location and sampling

Figure A1 shows the study location, the provinde Sindh and district Badin.

Figure A1: Study location



Notes. Source: ESRI.

Table A1 provides a breakdown of sample sizes (villages and households) by treatment arm.

Table A1: Treatment allocation per treatment arm

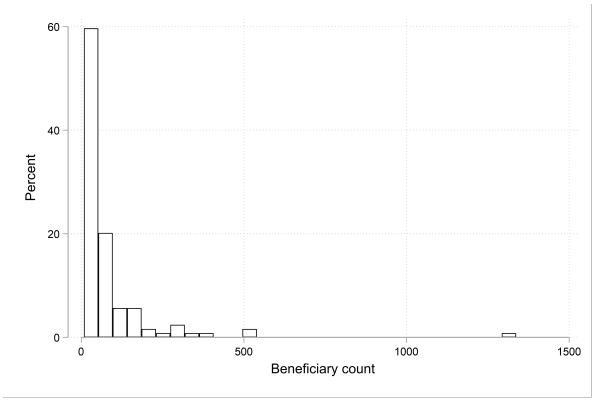
	Treatment		**Control**		**Total**	
	Number	%	Number	%	Number	%
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Villages						
Sindh	30	49.18	31	50	61	49.59
Punjab	31	50.82	31	50	62	50.41
Total	61	100	62	100	123	100
Panel B. Households						
Sindh	284	47.81	298	49.92	582	48.87
Punjab	310	52.19	299	50.08	609	51.13
Total	594	100	597	100	1191	100

Notes. Panel (A) shows the distribution of villages by treatment arm and Panel (B) of households per treatment arm.

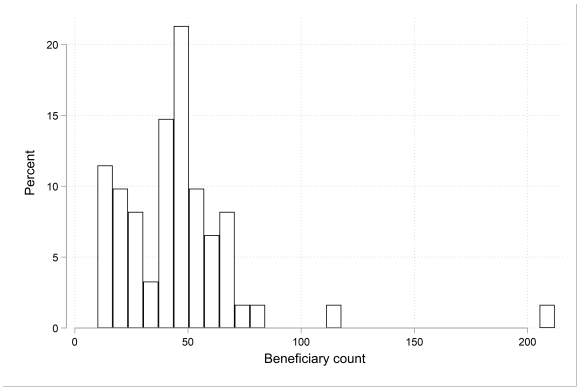
Figure A2 shows the distribution number of beneficiaries households per study village.

Figure A2: Distribution of number of beneficiaries households per study village

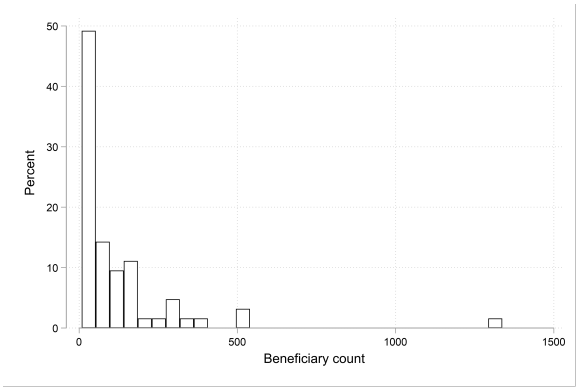
A. Full sample



B. Sindh



C. Punjab



Notes.'Beneficiary count' is the total number of beneficiaries sampled in a single village.

Table A2: Qualitative research design sample sizes

Method	Respondent	T	C	Total
(1)	(2)	(3)	(4)	(5)
KIIs	Village leaders and resource persons	2	2	12 KIIs (8 T and 4 C)
IDIs	Household representatives	4	2	20 IDIs (16 T and 4 C)
FGDs	Youth adult (18–39 years)	2	2	20 FGDs (16 T and 4 C)
	Middle-aged (40–65 years)	2	2	
	(NGO) Social Organisers	1		1 FGD
Total				53

Notes. 'T' denotes treated and 'C' control villages. 'KIIs' denote key informant interviews, 'IDIs' in-depth interviews , and 'FGDs' focus group discussions. 4 KIIs were conducted with resource persons (3 treatment; 1 control village); 8 KIIs were conducted with village leaders (5 treatment; 3 control). Village leaders include lady Health Workers, local counsellors, land owners, and community elders. Household representatives were any adult man or woman residing in the house; equal division by gender in each village. Sample for Sindh only included 4 resource persons and 8 village leaders.

Appendix 2 Variable definition and additional details about measurements

Table A3: Variable definition and additional details about measurements

Variable	Description
Awareness	Binary variable that takes a value of 1 if the main respondent in the household reported to have heard about village sanitation promotion campaigns in the previous 12 months, and 0 otherwise.
Attendance	Binary variable that takes a value of 1 if any member of the household reported to have attended village sanitation promotion campaigns in the previous 12 months, and 0 otherwise.
Open defecation	Binary variable that takes a value of 1 if at-least one member of the household above the age of 5 is reported to openly defecate, and 0 otherwise.
Functional latrine	Binary variable that takes a value of 1 if the household was reported as having a functional latrine in the dwelling and its members are reported to be using it, and 0 otherwise.
Latrine maintenance	Binary variable that takes a value of 1 if the household latrine was reported to have undergone repairs for maintenance in the last 2 years, and 0 otherwise.
Heterogeneity dimensions	— Village level
Public infrastructure index	Standardized index that ranges from 0 to 1, constructed using an item-response theory one-parameter logistic model on a number of village level variables. These include 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. The following sources of water are considered to be of improved technology – piped system, pubic tap/standpipe, tube well/bore well, handpump, protected well, protected spring or rainwater.
Sanitation supply index	Standardized index that ranges from 0 to 1, constructed using an item-response theory one-parameter logistic model on a number of village level variables. These include binary indicators for 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.
Weather vulnerability	Binary variable that takes a value of 1 if the village is reported to have had a flood or drought in the last two years, and 0 otherwise.
Heterogeneity dimensions	— Household level
Household asset index	Standardized index that ranges from 0 to 1, constructed using an item-response theory one-parameter logistic model on a number of household level variables. These include binary indicators for whether or not the household is reported to own a fridge, freezer, washing machine, cooking stove, cooking range, microwave, sewing machine, sofa, air-conditioner, air cooler, fan, geyser, heater, television, DVD player, computer, landline phone, mobile phone, tractor, scooter, rickshaw, car, bicycle, other vehicles, at-least one livestock.
Sanitation quality index	Standardized index that ranges from 0 to 1, constructed using an item-response theory one-parameter logistic model on a number of household level variables. These include binary indicators for whether or not the household latrine is observed to have a permanent superstructure, have a roof, have a curtain/door/form of closure, require no major repairs, require no minor repairs, have no faeces in the pan, have no flies and have no odour. In addition the index also considers binary indicators for whether or not a handwashing facility is observed in the household dwelling and whether or not water and soap were observed as present and available at the handwashing facility.
Shared facility	Binary variable that takes a value of 1 if the household sanitation facility is shared with people from other households, and 0 otherwise.
Latrine technology	Binary variable that takes a value of 1 if the household latrine is of improved technology and is observed to have a water seal, and 0 otherwise. The household latrine is considered to be of improved technology if it has a pour flush system, is a ventilated improved pit latrine, is a pit latrine with a slab or a composting toilet.

Appendix 3 Balance and attrition

Table A4: Balance at baseline

	All	**Control**	**Treatment - Control**
	Mean	Mean	Difference
	(SE)	(SE)	[p-value]
Outcomes			
Open defecation	0.17	0.17	0.01
	(0.01)	(0.03)	[0.84]
Use functional latrine	0.83	0.84	-0.01
	(0.03)	(0.05)	[0.84]
Household characteristics			
Household size	6.56	6.39	0.33
	(0.26)	(0.31)	[0.52]
Number of males	3.31	3.34	-0.07
	(0.15)	(0.19)	[0.81]
Number of children under 5	0.89	0.84	0.12
	(0.10)	(0.13)	[0.56]
Head, male	0.89	0.90	-0.03
	(0.03)	(0.04)	[0.60]
Head, age	41.72	43.66	-3.85
	(1.19)	(1.73)	[0.10]
Head, years of schooling	3.15	2.85	0.60
	(0.36)	(0.52)	[0.41]
House is owned	0.91	0.92	-0.01
	(0.03)	(0.04)	[0.78]
Dwelling of strong material	26.83	29.51	-5.31
	(4.01)	(5.86)	[0.51]
BISP score	21.97	22.87	-1.79
	(0.89)	(1.23)	[0.32]
Worked on community project	0.18	0.18	-0.00
	(0.03)	(0.05)	[0.97]
Household asset ($>$ median)	0.46	0.44	0.04
	(0.05)	(0.06)	[0.65]
Sanitation quality ($>$ median)	0.51	0.61	-0.19
	(0.05)	(0.06)	[0.04]
Shared bathroom	0.25	0.20	0.10
	(0.04)	(0.05)	[0.17]
Improved latrine	0.72	0.75	-0.07
	(0.04)	(0.05)	[0.41]
Village characteristics			
Public infrastructure ($>$ median)	0.48	0.51	-0.06
	(0.05)	(0.06)	[0.53]
Sanitation supply ($>$ median)	0.49	0.49	-0.01
	(0.05)	(0.06)	[0.93]
Flood/Drought in the last 2 years	0.48	0.49	-0.02
	(0.05)	(0.07)	[0.80]
Declared OD free	0.61	0.66	-0.10
	(0.05)	(0.06)	[0.26]

Notes. Sample includes households interviewed at endline, both in the provinces of Sindh and Punjab. Standard errors in parenthesis and p-values in brackets.

Table A5: Balance at baseline, Sindh province

	All	**Control**	**Treatment - Control**
	Mean	Mean	Difference
	(SE)	(SE)	[p-value]
Outcomes			
Open defecation	0.31	0.27	0.09
	(0.06)	(0.08)	[0.46]
Use functional latrine	0.80	0.83	-0.06
	(0.05)	(0.07)	[0.57]
Household characteristics			
Household size	7.15	6.77	0.75
	(0.40)	(0.39)	[0.35]
Number of males	3.66	3.67	-0.02
	(0.24)	(0.27)	[0.96]
Number of children under 5	1.08	0.93	0.29
	(0.15)	(0.19)	[0.34]
Head, male	0.90	0.93	-0.06
	(0.04)	(0.05)	[0.42]
Head, age	41.70	45.43	-7.34
	(1.83)	(2.62)	[0.04]
Head, years of schooling	2.72	2.30	0.83
	(0.45)	(0.67)	[0.36]
House is owned	0.90	0.90	0.00
	(0.04)	(0.06)	[0.97]
Worked on community project	0.31	0.30	0.02
	(0.06)	(0.08)	[0.85]
Household asset ($>$ median)	0.16	0.13	0.06
	(0.05)	(0.06)	[0.53]
Sanitation quality ($>$ median)	0.26	0.33	-0.14
	(0.06)	(0.09)	[0.22]
Shared facility	0.13	0.15	-0.04
	(0.04)	(0.06)	[0.66]
Improved latrine	0.69	0.72	-0.06
	(0.06)	(0.08)	[0.61]
Village characteristics			
Public infrastructure ($>$ median)	0.43	0.53	-0.21
	(0.06)	(0.09)	[0.10]
Sanitation supply ($>$ median)	0.31	0.33	-0.04
	(0.06)	(0.09)	[0.72]
Weather vulnerability	0.85	0.87	-0.03
	(0.05)	(0.06)	[0.76]
Declared OD free	0.87	0.90	-0.06
	(0.04)	(0.06)	[0.49]

Notes. Sample includes households interviewed at endline in the provinces of Sindh. Standard errors in parenthesis and p-values in brackets.

Table A6: Attrition at endline

	(1)	(2)	(3)	(4)
Follow-up	-0.02	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.05)	(0.05)
OD		0.00	0.00	0.00
		(0.00)	(0.00)	(0.00)
Functional toilet		-0.00	-0.00	-0.00
		(0.00)	(0.00)	(0.00)
Follow-up x OD			0.00	0.00
			(0.00)	(0.00)
Follow-up x Functional toilet			-0.00	-0.00
			(0.00)	(0.00)
Household size				-0.01***
				(0.00)
Number of males				0.01
				(0.00)
Number of children under 5				0.00
				(0.01)
Head, male				-0.00
				(0.00)
Head, age				-0.00
				(0.00)
Head, years of schooling				-0.00
				(0.00)
Household owns house				0.00
				(0.00)
Strong dwelling				0.00
				(0.00)
BISP Poverty score				-0.00*
				(0.00)
Worked on community project				-0.00
				(0.00)
Attrition rate	0.05	0.04	0.04	0.04
P-value for F-test on covariates	.	0.07	0.22	0.01
No. of observations	1,186	1,185	1,185	1,185

Notes. Sample includes households interviewed at baseline, both in the provinces of Sindh and Punjab. The dependent variable is a dummy variable equal to one if the household was not interviewed at endline and zero otherwise. Standard errors clustered at the unit of randomization (villages) in parenthesis. Statistical significance denoted by * p<0.1 , ** p<0.05, and *** p<0.01.

Appendix 4 Intervention details and timeline

Table A7: Program description

Activity	Description	Details		
		Activities	Times	Attendance
(1)	(2)	(3)	(4)	(5)
Hygiene Session	Community level promotion events held by NGO social organisers. During these events the social organisers hold sessions on hand washing practices with soap on critical timings and personal hygiene and demonstrations. Additionally, the social organisers discuss issues related to safe sanitation practices. The NGO aimed that at least one participants form every HH take part in the hhygiene session.	29	9	216
Broad Based Community Meeting (BBCM)	BBCM were community level events organised by NGO social organisers. These events included reviewing the ODF targets within the community and planning around what needed to be done to meet these (in the case of non-ODF villages). These meetings included eliciting family and community commitments and pledges and publically recognising achievements. The aim of these activities was to create social pressure and promote interest behaviours as social norms.	29	3	147

Notes. 'Activities' (column 3) is the number of communities activity implemented; 'Times' (column 4) is the average number of times activity took place in phase 2 (per village); and 'Attendance' is the average number of people attending (per activity).

Table A8: Study timeline

Phase	Date	Events
Phase 1	May 2014	Contracts signed between finder and INGO partners
	June 2015	Phase 1 implementation begins. Activities predominantly CLTS. Payments to INGO partner based on three key output indicators: The number of poor people having access to safe and reliable drinking water sources. The number of poor people having access to basic or improved household latrines. The number of poor men, women, and school children that have been reached by handwashing promotion in villages and schools
	Dec 2015	Deadline for the achievement of Phase 1 results
Phase 2	Jan/Feb 2016	Baseline quantitative data collection for research
	Feb/Mar 2016	Phase 2 implementation begins. Programme activities focused on sustaining behaviour change. Payments to INGO partner based on three key outcome indicators: 90% of poor people across the project districts continue to use reliable, safe drinking water sources, 75% of poor people across project districts continue to use basic or improved latrines, and A composite hygiene indicator composed of: improvements in knowledge, observation of handwashing facilities, and change in self-reported behaviours.
	Mar 2018	Phase 2 results assessed, end of programme.
	Mar 2018	Endline quantitative data collection for research
	Apr/May 2018	Endline qualitative data collection

Appendix 5 Implementation fidelity

Table A9: Awareness and attendance of follow-up activities

Outcome	Awareness		Attendance	
	(1)	(2)	(3)	(4)
Follow-up activities	0.20***	0.19***	0.17***	0.16***
	(0.05)	(0.05)	(0.04)	(0.04)
ANCOVA	No	Yes	No	Yes
Control mean (EL)	0.35	0.35	0.25	0.25
Villages	123	123	123	123
Households	1,132	1,132	1,132	1,132

Notes. Primary data, household level. 'Awareness' (columns (1) and (2)) is an indicator of whether the main respondent heard about village sanitation promotion campaigns in the previous 12 months. 'Attendance' (columns (3) and (4)) is an indicator of whether any household member attended village sanitation promotion campaigns in the previous 12 months. Sample include households surveyed at endline, both in the provinces of Sindh and Punjab. Standard errors (in parenthesis) are clustered at the unit of randomization (villages). Statistical significance denoted by * p¡0.1 , ** p¡0.05, and *** p¡0.01.

Appendix 6 Time-gap analysis

Table A10: Sample selection for time - gap analysis

Outcome	3 months	6 months	9 months
	(1)	(2)	(3)
Household size	-0.01	0.00	-0.01
	(0.01)	(0.01)	(0.01)
Number of males	0.00	0.00	0.01
	(0.01)	(0.02)	(0.02)
Number of children under 5	0.03*	0.02	0.02
	(0.02)	(0.02)	(0.02)
Head, male	0.00**	0.00	-0.00
	(0.00)	(0.00)	(0.00)
Head, age	0.00	0.00**	0.00
	(0.00)	(0.00)	(0.00)
Head, years of schooling	-0.00	-0.00	0.00
	(0.01)	(0.01)	(0.01)
House is owned	-0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
Dwelling of strong material	0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)
Household asset index	0.11	-0.00	0.26
	(0.20)	(0.23)	(0.22)
Sanitation quality	-0.06	0.18	0.23*
	(0.11)	(0.13)	(0.13)
Shared bathroom	-0.09	-0.13	-0.15
	(0.15)	(0.12)	(0.12)
Latrine technology	0.03	0.12*	0.06
	(0.05)	(0.07)	(0.07)
Public infrastructure	0.54**	0.56*	0.18
	(0.25)	(0.31)	(0.34)
Sanitation supply	-0.40	0.06	0.28
	(0.38)	(0.48)	(0.48)
Weather vulnerability	0.22	-0.19	0.01
	(0.19)	(0.17)	(0.18)
Joint F-test	0.19	0.06	0.77
Villages	61	61	61
Households	551	551	551

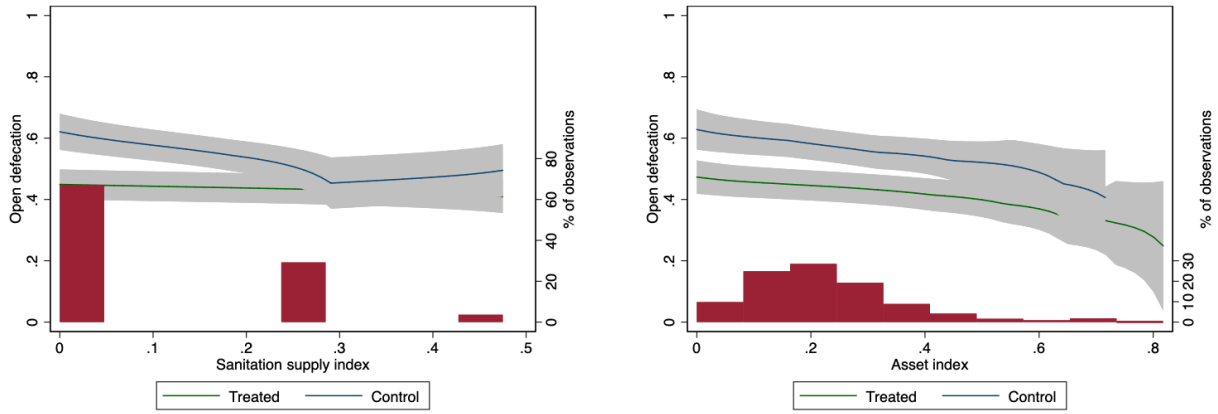
Notes. Primary data, household-level. '3 months' is an indicator $= 1$ if interviewed at endline 3 months after follow-up activities or control, $= 0$ otherwise; '6 months' $= 1$ if interviewed at endline 6 months after follow-up activities or control, and $= 0$ otherwise; and '9 months' $= 1$ if interviewed at endline 9 months after follow-up activities or control, and $= 0$ otherwise. 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 parenthesis. Estimates are reported based on baseline characteristics. Statistical significance denoted by * $p<0.1$, ** $p<0.05$, and *** $p<0.01$.

Appendix 7 Additional information for heterogeneity analysis

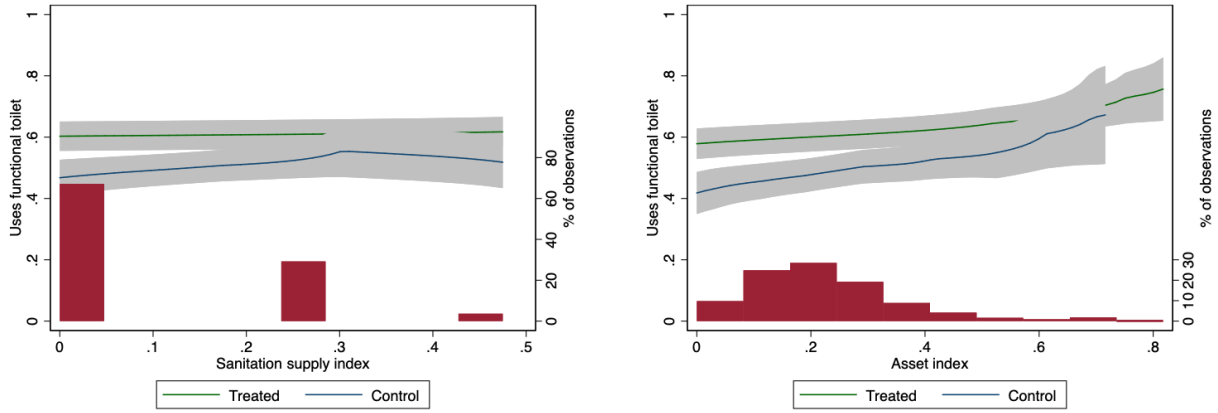
Appendix 7.1 Heterogeneity by initial OD

Figure A3: Heterogeneous effects of follow-up activities on behaviour, by village sanitation supply and household asset

A. Open defecation



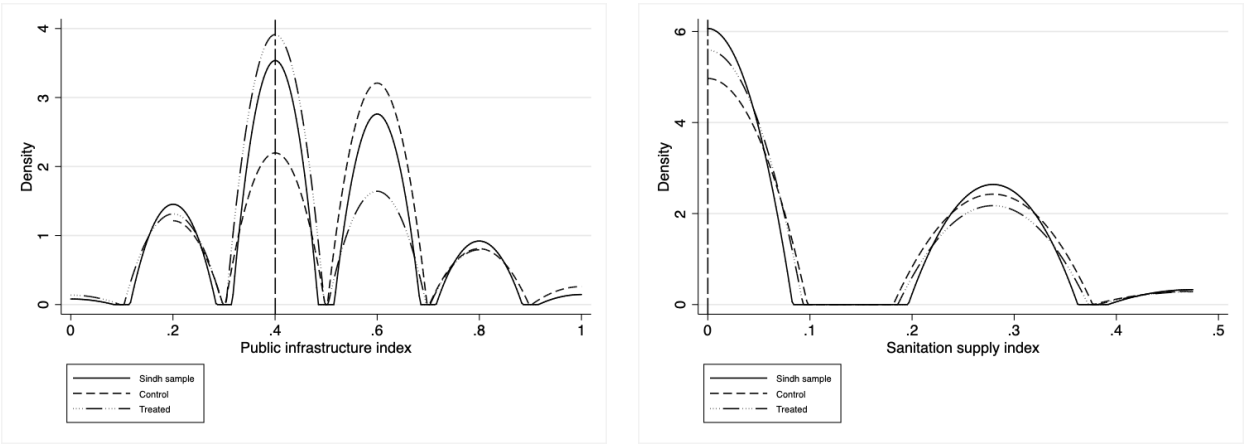
B. Functional latrine



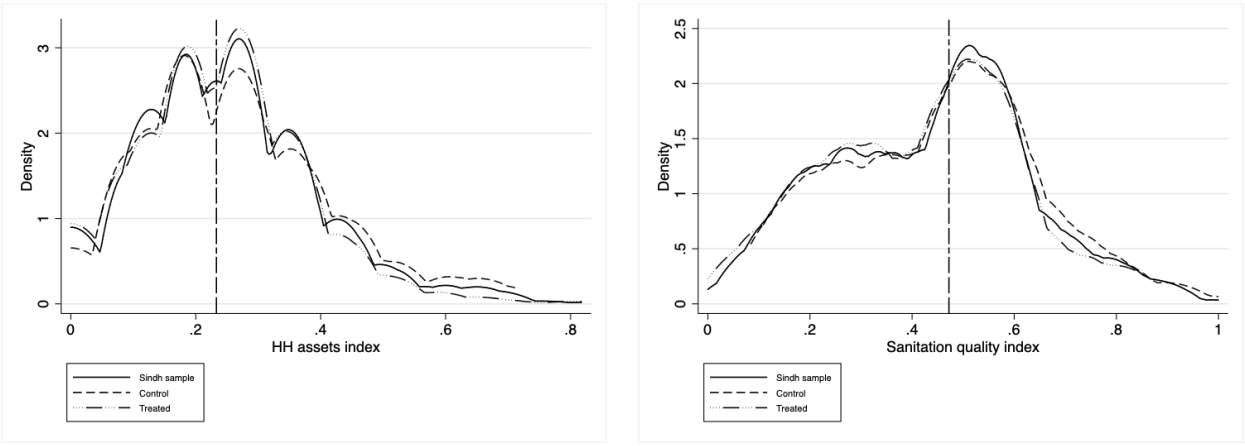
Notes. 'Open defecation' (Figure A) is an indicator of whether a household has at-least 1 member older than 5 years that openly defecates. 'Use functional latrine' (columns (2), (4) and (6)) is an indicator for whether the household has a functional latrine in the dwelling used by its members. 'Sanitation quality' is an index capturing the quality of the bathroom located in the household's dwelling. The histogram located at the bottom of each plot shows the distribution of observations by the sanitation quality index. Sample includes households interviewed at endline in Sindh.

Figure A4: Distribution of heterogeneity dimensions

A. Village level



B. Household level



Notes. 'Public infrastructure' is an index capturing the availability of public infrastructure in the village. 'Sanitation supply' is an index capturing the availability of sanitation-related goods and services in the village. 'HH asset' is an index capturing the household's ownership of assets. 'Sanitation quality' is an index capturing the quality of the bathroom located in the household's dwelling. Sample includes households interviewed at endline in Sindh.

Table A11: Pairways correlation for heterogeneity dimensions

	Sanitat quality	HH asset	Shared latrine	Latrine technology	Public inf	Sanit supply	Weather vulnerab
Sanitation quality	1.00	0.52	0.07	0.27	0.12	0.14	-0.34
HH asset	0.52	1.00	0.13	0.23	0.13	0.27	-0.53
Shared latrine	0.08	0.13	1.00	0.02	0.11	0.16	-0.19
Latrine technology	0.27	0.23	0.02	1.00	-0.04	0.03	-0.09
Public infrastructure	0.12	0.13	0.11	-0.04	1.00	0.41	-0.05
Sanitation supply	0.14	0.28	0.16	0.03	0.41	1.00	-0.35
Weather vulnerability	-0.34	-0.53	-0.20	-0.09	-0.06	-0.35	1.00

Table A12: Heterogeneous effects of follow-up activities on behaviour, by OD

	Beneficiaries OD rate		HH practiced OD	
Outcome	OD	Latrine	OD	Latrine
	(1)	(2)	(3)	(4)
Follow-up activities	-0.04	0.07	-0.15*	0.18**
	(0.11)	(0.08)	(0.08)	(0.08)
Follow-up x Yes/High	-0.19	0.13	0.01	-0.13
	(0.13)	(0.11)	(0.09)	(0.09)
Follow-up (Yes/High)	-0.23***	0.20**	-0.14*	0.04
	(0.00)	(0.01)	(0.05)	(0.58)
Control mean (No/Low)	0.31	0.75	0.46	0.58
Control mean (Yes/High)	0.75	0.31	0.81	0.31
Villages	61	61	61	61
Households	551	551	551	551

Notes. 'OD' (columns (1) and (3)) is an indicator of whether a household has at-least 1 member older than 5 years that openly defecates. 'Latrine' (columns (2) and (4)) is an indicator for whether the household has a functional toilet in the dwelling used by its members. 'Beneficiaries OD rate' is an indicator of whether the HH is located in a village that is above the median proportion of eligible HHs that have at-least 1 member older than 5 years that openly defecates at baseline. 'HH practiced OD' is an indicator of whether the HH practices OD at baseline. 'Follow-up activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (Yes/High)' is a post-estimation result of the linear combination $\alpha_1 + \alpha_2$. All regressions control for the corresponding heterogeneity dimension. 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.

Table A13: Joint analysis, including village OD rate (eligible households)

Outcome	Open defecation	Use functional latrine
	(1)	(2)
Follow-up activities (FU)	-0.02	0.04
	(0.22)	(0.21)
Village level		
FU x Public infrastructure (Low)	-0.26**	0.28**
	(0.12)	(0.12)
FU x Sanitation supply (Low)	-0.04	-0.01
	(0.11)	(0.11)
FU x Vulnerability (Low)	0.11	-0.09
	(0.17)	(0.12)
FU x Eligible OD rate (Low)	0.20	-0.17
	(0.13)	(0.11)
Household level		
FU x Asset (Low)	-0.07	0.06
	(0.11)	(0.10)
FU x Sanitation quality (Low)	-0.13	0.15
	(0.10)	(0.10)
FU x Shared latrine (No)	0.03	-0.04
	(0.13)	(0.11)
FU x Latrine technology (Low)	0.15	-0.21**
	(0.09)	(0.10)
Villages	61	61
Households	551	551

Notes. 'Open defecation' (column (1)) is an indicator of whether a household has at-least 1 member older than 5 years that 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 *** $p<0.01$.

Table A14: Heterogeneous effects of follow-up activities on behaviour, by OD, all provinces

	Beneficiaries OD rate		HH practiced OD	
Outcome	OD	Latrine	OD	Latrine
	(1)	(2)	(3)	(4)
Follow-up activities	0.00	0.01	-0.04	0.06
	(0.04)	(0.04)	(0.04)	(0.04)
Follow-up x Yes/High	-0.13*	0.12*	-0.05	0.02
	(0.08)	(0.07)	(0.06)	(0.07)
Follow-up (Yes/High)	-0.13**	0.14**	-0.09	0.08
	(0.04)	(0.03)	(0.15)	(0.22)
Control mean (No/Low)	0.16	0.84	0.24	0.77
Control mean (Yes/High)	0.59	0.43	0.66	0.38
Villages	123	123	123	123
Households	1,132	1,132	1,132	1,132

Notes. 'OD' (columns (1) and (3)) is an indicator of whether a household has at-least 1 member older than 5 years that openly defecates. 'Latrine' (columns (2) and (4)) is an indicator for whether the household has a functional toilet in the dwelling used by its members. 'Beneficiaries OD rate' is an indicator of whether the HH is located in a village that is above the median proportion of eligible HHs that have at-least 1 member older than 5 years that openly defecates at baseline. 'HH practiced OD' is an indicator of whether the HH practices OD at baseline. 'Follow-up activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (Yes/High)' is a post-estimation result of the linear combination $\alpha_1 + \alpha_2$. All regressions control for the corresponding heterogeneity dimension. Sample includes households interviewed at endline in both provinces, Sindh and Punjab. 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.

Table A15: Heterogeneous effects of follow-up activities on reported maintenance

	Public inf	Sanit supply	Sanit quality
	(1)	(2)	(3)
Follow-up activities	0.04	0.04	0.07*
	(0.04)	(0.05)	(0.04)
Follow-up activities x Yes/High	0.04	0.03	-0.08
	(0.07)	(0.06)	(0.05)
Follow-up activities (Yes/High)	0.08	0.07	-0.01
	(0.22)	(0.12)	(0.85)
Control mean (No/Low)	0.10	0.10	0.08
Control mean (Yes/High)	0.08	0.05	0.10
Villages	61	61	61
Households	551	551	551

Notes. The outcome in each column is an indicator of whether the household latrine was maintained in the last 2 years. 'Public inf' is an indicator for being above the median of village-level public infrastructure index. 'Sanit supply' is an indicator for being above the median of village-level sanitation supply index. 'Sanit quality' is an indicator for being above the median of the household-level sanitation quality index. 'Follow-up activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (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). Statistical significance denoted by * p<0.1 , ** p<0.05, and *** p<0.01.

Appendix 8 Additional information for robustness checks

Appendix 8.1 Robustness - LASSO

Table A16: Effect of follow-up activities on behaviour, LASSO

Outcome	Open defecation		Use functional latrine	
	(1)	(2)	(3)	(4)
Follow-up activities	-0.15**	-0.16***	0.15**	0.16**
	(0.07)	(0.06)	(0.06)	(0.07)
ANCOVA	No	Yes	No	Yes
Control mean (EL)	0.57	0.57	0.49	0.49
Villages	61	61	61	61
Households	551	551	551	551

Notes. Primary data, household level. 'Open defecation' (columns (1) and (2)) is an indicator of whether a household reports at least 1 member older than 5 years old that openly defecates. 'Use functional latrine' (columns (3) and (4)) is an indicator of whether the household has a functional latrine in the dwelling and members use it. Sample includes households interviewed at endline in Sindh. Standard errors (in parenthesis) are clustered at the unit of randomization (villages). Statistical significance denoted by * p¡0.1 , ** p¡0.05, and *** p¡0.01.

Appendix 8.2 Robustness - all provinces

Table A17: Heterogeneous effects of follow-up activities on behaviour, by village characteristics, LASSO

Outcome	Public infrastructure		Sanitation supply		Weather vulnerability	
	OD	Latrine	OD	Latrine	OD	Latrine
	(1)	(2)	(3)	(4)	(5)	(6)
Follow-up activities	-0.32***	0.35***	-0.16**	0.13	-1.50***	1.53***
	(0.08)	(0.07)	(0.08)	(0.08)	(0.03)	(0.01)
Follow-up activities x Yes/High	0.25*	-0.37***	-0.08	0.20	1.20***	-1.12***
	(0.13)	(0.14)	(0.15)	(0.13)	(0.04)	(0.03)
Follow-up activities (Yes/High)	-0.07	-0.01	-0.24*	0.33***	-0.30***	0.41***
	(0.52)	(0.91)	(0.05)	(0.00)	(0.00)	(0.00)
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 of whether a household has at-least 1 member older than 5 years that 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 having 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 2 years. 'Follow-up activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (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.

Table A18: Heterogeneous effects of follow-up activities on behaviour, by households characteristics, LASSO

Outcome	Household assets		Sanitation quality		Shared facility		Latrine technology	
	OD	Latrine	OD	Latrine	OD	Latrine	OD	Latrine
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Follow-up activities	-0.16**	0.18**	-0.19***	0.17**	-0.16**	0.17**	-0.09	0.05
	(0.07)	(0.07)	(0.06)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)
Follow-up activities x Yes/High	0.04	-0.03	0.09	-0.07	0.01	-0.06	-0.10	0.19**
	(0.12)	(0.10)	(0.10)	(0.09)	(0.12)	(0.15)	(0.09)	(0.09)
Follow-up activities (Yes/High)	-0.13	0.14	-0.09	0.11	-0.15	0.11	-0.19**	0.24***
	(0.25)	(0.15)	(0.33)	(0.20)	(0.16)	(0.46)	(0.01)	(0.00)
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) and (5)) is an indicator of whether a household has at-least 1 member older than 5 years that 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. 'Household assets' (columns (1) and (2)) is an indicator for being above the median household asset index index. 'Sanitation quality' (columns (3) and (4)) is an indicator for being above the median of the sanitation quality index. 'Shared facility' (columns (5) and (6)) is an indicator for whether the household bathroom and latrine are shared. '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 activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (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.

Table A19: Heterogeneous effects of follow-up activities on behaviour, by village characteristics, all provinces

Outcome	Public infrastructure		Sanitation supply		Weather vulnerability	
	OD	Latrine	OD	Latrine	OD	Latrine
	(1)	(2)	(3)	(4)	(5)	(6)
Follow-up activities	-0.13**	0.16***	-0.11*	0.11*	-0.00	0.04
	(0.06)	(0.05)	(0.06)	(0.06)	(0.04)	(0.04)
Follow-up activities x Yes/High	0.15*	-0.20**	0.11	-0.10	-0.12	0.05
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Follow-up activities (Yes/High)	0.03	-0.04	0.00	0.02	-0.13*	0.10
	(0.64)	(0.51)	(1.00)	(0.73)	(0.07)	(0.18)
Control mean (No/Low)	0.37	0.64	0.46	0.58	0.20	0.79
Control mean (Yes/High)	0.31	0.71	0.22	0.77	0.54	0.51
Villages	123	123	123	123	119	119
Households	1,132	1,132	1,132	1,132	1,095	1,095

Notes. 'OD' (columns (1), (3) and (5)) is an indicator of whether a household has at-least 1 member older than 5 years that 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 having 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 2 years. 'Follow-up activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (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 restricted to households interviewed at endline, both in the provinces of Sindh and Punjab. 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.

Table A20: Heterogeneous effects of follow-up activities on behaviour, by households characteristics, all provinces

Outcome	Household assets		Sanitation quality		Shared facility		Latrine technology	
	OD	Latrine	OD	Latrine	OD	Latrine	OD	Latrine
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Follow-up activities	-0.13**	0.12*	-0.12**	0.14***	-0.07	0.08*	0.01	0.00
	(0.06)	(0.06)	(0.05)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)
Follow-up activities x Yes/High	0.15**	-0.10	0.14**	-0.16***	0.07	-0.03	-0.09	0.09
	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)
Follow-up activities (Yes/High)	0.02	0.02	0.02	-0.02	0.00	0.05	-0.09*	0.10**
	(0.54)	(0.59)	(0.61)	(0.56)	(0.99)	(0.43)	(0.06)	(0.02)
Control mean (No/Low)	0.53	0.51	0.49	0.51	0.35	0.69	0.53	0.45
Control mean (Yes/High)	0.16	0.83	0.18	0.86	0.21	0.73	0.27	0.76
Villages	123	123	123	123	123	123	123	123
Households	1,132	1,132	1,132	1,132	1,132	1,132	1,132	1,132

Notes. 'OD' (columns (1), (3) and (5)) is an indicator of whether a household has at-least 1 member older than 5 years that 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. 'Household assets' (columns (1) and (2)) is an indicator for being above the median household asset index index. 'Sanitation quality' (columns (3) and (4)) is an indicator for being above the median of the sanitation quality index. 'Shared facility' (columns (5) and (6)) is an indicator for whether the household bathroom and latrine are shared. '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 activties' shows the estimates of the α_1 parameter and 'Follow-up activties x Yes/High' presents the estimates of the α_2 parameter from equation 2. 'Follow-up activties (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, both in the provinces of Sindh and Punjab. 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.

Table A21: Joint analysis, all provinces

Outcome	Open defecation	Use functional latrine
	(1)	(2)
Follow-up activities	0.10	-0.14
	(0.11)	(0.11)
Village level		
FU x Public infrastructure (Low)	-0.14*	0.17**
	(0.08)	(0.08)
FU x Sanitation supply (Low)	0.01	-0.02
	(0.08)	(0.08)
FU x Vulnerability (Low)	0.02	0.06
	(0.09)	(0.09)
Household level		
FU x Asset (Low)	-0.10	0.08
	(0.07)	(0.06)
FU x Sanitation quality (Low)	-0.13**	0.18***
	(0.06)	(0.06)
FU x Shared latrine (No)	-0.06	0.04
	(0.07)	(0.06)
FU x Latrine technology (Low)	0.15**	-0.16**
	(0.06)	(0.07)
Villages	119	119
Households	1,095	1,095

Notes. 'OD' (column (1)) is an indicator of whether a household has at-least 1 member older than 5 years that openly defecates. '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. The sample is lower because the values of the 'Weather vulnerability' indicator is missing for four villages. Sample includes households interviewed at endline, both in the provinces of Sindh and Punjab. Standard errors clustered at the unit of randomization (villages) in parentheses. Statistical significance denoted by * $p<0.1$, ** $p<0.05$, and *** $p<0.01$.