Nudges for Rural Sanitation:

Evaluating low-touch methods to promote latrine use in rural Bihar

Impact Evaluation Document

20 May 2018

Project Team:
Divya Nair (Director); Pulkit Agarwal (Project Manager); Ritika Rastogi (Associate); Sebastian Lucek (Associate); Karan Nagpal (Economist); Steven Brownstone (Junior Economist)

Table of Contents

Note: This pre-analysis plan (PAP) is being submitted to the Registry for International Development Impact Evaluations (RIDIE) after the baseline survey and the implementation of our intervention are complete, but before the endline survey and final analysis take place. We did not register the PAP before baseline because we were uncertain about the latrine coverage rates in our target geography, about the extent to which the latrines in our target geography were the type we are focussing on in this study, and about the expected rate of compliance with our intervention. This uncertainty necessitated a flexible approach to our baseline survey and sampling strategy that allowed us to make changes to our study design, such as varying the way we sampled villages for the study and how we offered our intervention to the sampled households. The insights gained from these activities have informed this document.

Intervention Overview

I. Intervention Description

In the past five years, an unprecedented number of individual household latrines have been constructed in rural India as part of the Swachh Bharat Mission – Gramin (SBM-G), the Government of India's flagship rural sanitation program. While this has reduced rates of open defecation, a practice widely regarded as severely detrimental to health[1], a substantial proportion of latrine owners continue to engage in open defecation. An important reason for continued open defecation by latrine owners is the presence of behavioural barriers to latrine use, which we outline below in detail. Mitigating these behavioural barriers will be essential for promoting and sustaining exclusive latrine use in rural India.

In this project, we will evaluate whether a bundle of behavioural interventions (nudges) applied in latrines can increase latrine use and decrease open defecation. If the evaluation is successful, the nudge ideas will be shared with government and other stakeholders so that they can be implemented at scale in order to promote sustained, exclusive latrine use in rural India.

To design the nudges, IDinsight's project team, together with ideas42, first identified important behavioural barriers to latrine use through an extensive review of the existing evidence. Then, we designed an initial series of nudges through an iterative design process. Promising nudges were shortlisted by IDinsight with extensive input from ideas42. Each of the nudges targets one of the three behavioural barriers to sustained, exclusive latrine use that we identified as the most salient:

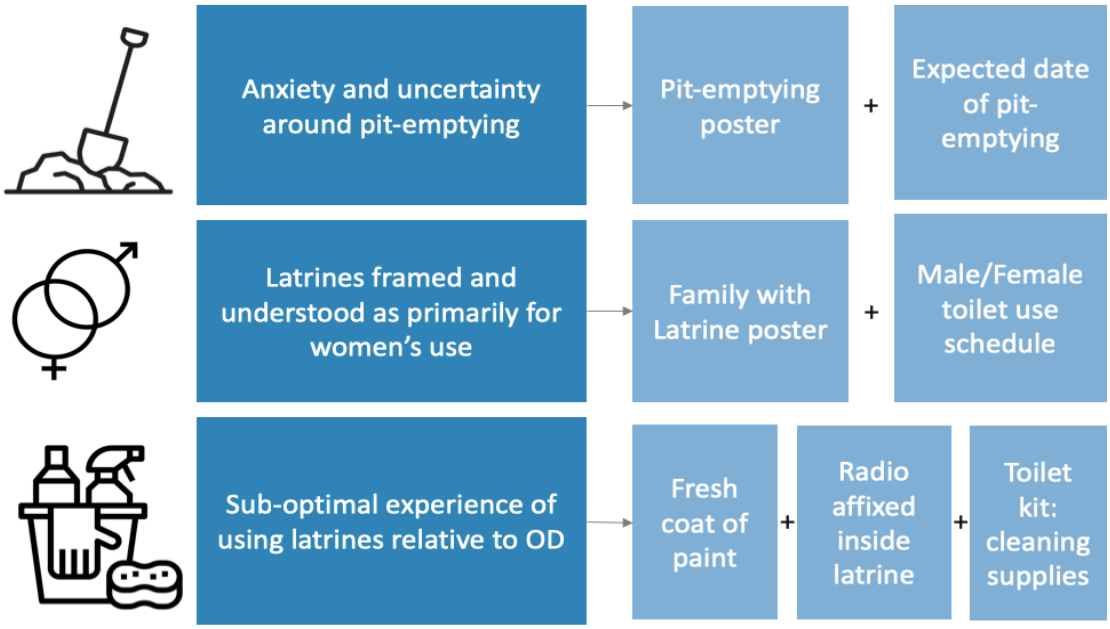
1. **Pit-emptying**: Individuals are anxious that their pits will fill, but they underestimate how long this will take. In addition, people of all castes wish to avoid emptying their pits since this activity is considered impure and was historically performed by the lowest castes, often against their will. Together, these factors lead to a faulty mental model of latrines as a limited resource. Therefore, latrines are a source of anxiety and uncertainty, and latrine owners have low motivation to use their latrines.
2. **Gender norms**: Men and women understand latrines as primarily intended for use by women in the home. For example, in our baseline survey, 94% of respondents say women's use of latrines is more important than men's, and only half of respondents believe that it is appropriate for women and men to share the same latrine. In addition, 99% of respondents in our baseline survey agree with the statement that protecting the dignity of women is a primary reason for owning a latrine. Thus, many men continue practicing open defecation instead of exclusively using latrines.
3. **Convenience and experience**: Many people find latrine use unpleasant and consider it an inconvenience or hassle. In addition, households struggle with finding the right materials and time

[1] Gupta, A., Khalid, N., Deshpande, D., Hathi, P., Kapur, A., Srivastav, N., Vyas, S., Spears, D., & Coffey, D. (2019). *Changes in open defecation in rural North India: 2014-2018.* Working Paper.

required to maintain adequately clean, usable latrines. Thus, many individuals continue practicing open defecation despite having access to a latrine.

Figure 1 below provides the final list of nudges, grouped by the behavioural barrier that is targeted by the nudge. The nudge designs were developed and finalized through an extensive prototyping exercise in our target geography consisting of in-depth interviews, focus group discussions, co-creation sessions, feasibility tests, and meetings with local sanitation stakeholders.

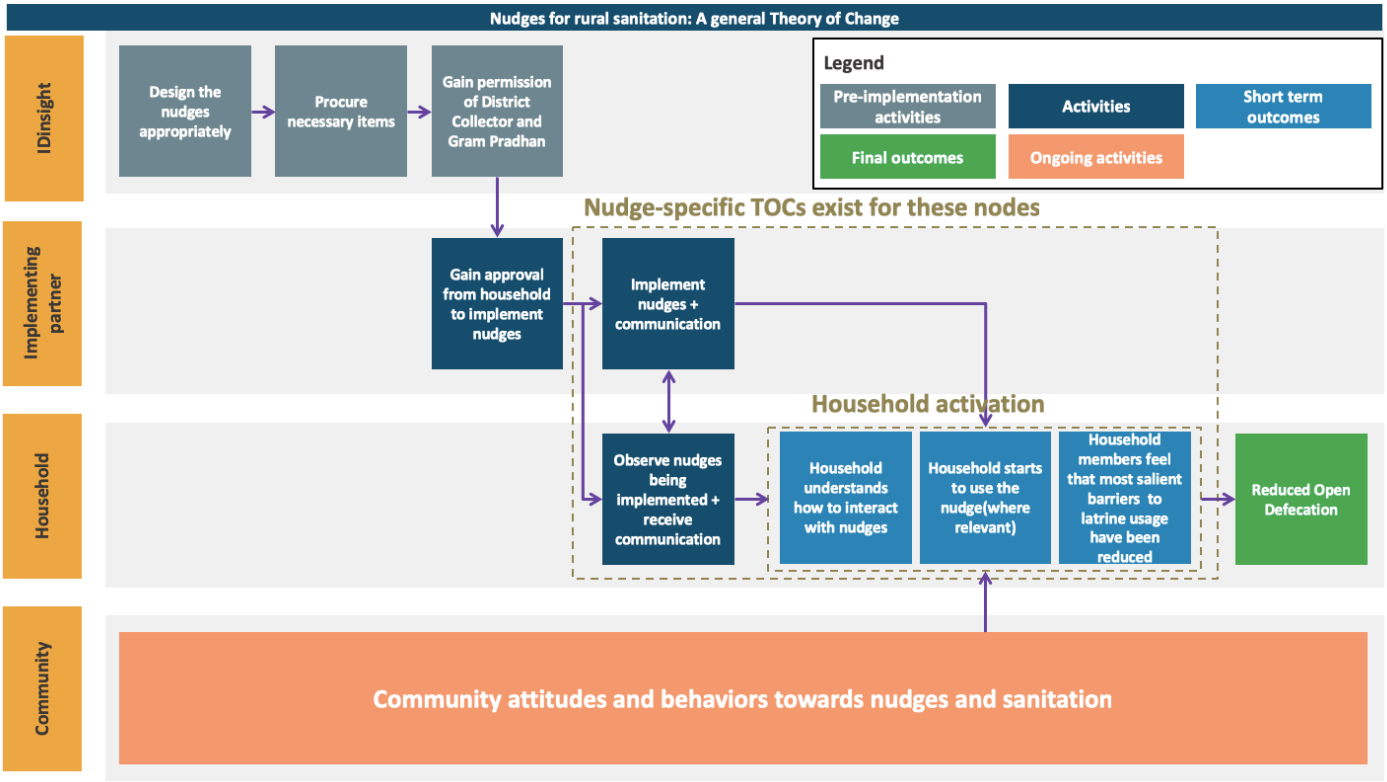
Figure 1: Behavioural barriers and related nudges



II. Theory of Change

The following theory of change (**Figure 2**) presents an overview of the steps required for our intervention to achieve the change we seek–a reduction in open defecation relative to latrine use–by mapping intervention inputs to outputs and behavioural outcomes.

Figure 2: Theory of Change



Impact Evaluation Design

I. Research Questions

1. Primary Impact Evaluation Research Questions

Q1.1: What is the effect of the intervention on rates of open defecation reported by those who are offered the intervention, compared to those who are not offered the intervention?

Q1.2: What is the effect of the nudge intervention on the salience of the three behavioural barriers to latrine use that we have identified as especially salient?

These three behavioural barriers are:

1. Pit-emptying:
    a. Correct estimation of pit-emptying timeline
    b. Reported aversion to pit-emptying
    c. Anxiety associated with latrine upkeep
2. Gender-norm barriers to latrine use
    a. Association of latrines with women's use
    b. Coordination problems between household members of opposite genders
    c. Preference for not sharing latrines between household members of opposite genders
3. Associations of hassle and unpleasantness with latrine use

2. Secondary Impact Evaluation Research Questions

Q2.1: Does treatment affect other latrine-oriented behaviours and perceptions? These include:

1. Number of times latrine(s) is/are cleaned per week
2. Self-reported cleanliness of latrine(s)

Q2.2.: How do treatment effects on all outcomes of interest (OD rates, latrine use rates, barrier salience scores) vary by individual and household characteristics? These include:

- Gender
- Age
- Baseline OD rates
- Religion
- Caste
- Education levels
- Socioeconomic Status (as measured by the Poverty Probability Index (PPI))[2]

[2] Further information about the Poverty Probability Index may be found at www.povertyindex.org/about-ppi

- Household size

II. Evaluation Methodology

We will conduct a household-level randomised control trial (RCT). The main benefit of an RCT lies in the construction of a convincing counterfactual group to simulate what the open defecation rate would have been among households receiving the nudges had they not received them. RCTs are acknowledged as the most rigorous way to estimate the causal effect of an intervention that is attributable to the intervention alone. In addition, because low-touch nudges work largely through targeting sub-conscious mental processes and mental models, those who receive the intervention may not be able to easily identify the effects of the nudges on their own behaviour. Thus, an RCT is especially helpful in understanding the effects of nudges on latrine use behaviours.

We are conducting this study in Manigachhi block in Darbhanga district, Bihar. We began by listing all villages in this block from the 2011 Population Census. We then sorted this list of villages in a random order. Our power calculations suggested a total sample size of 1,777. In each village, our sample includes every household that owns a functional pit latrine, but does not own any septic latrine. Since we did not know ex ante how many such households we would find in any given village, we proceeded down the list of villages (sorted randomly) with our administration of the baseline survey until we reached our target sample size. We reached our target while conducting baseline in the 24th village in our randomly ordered list. Since we wanted to saturate this final village, we exceeded our target sample size for a final sample of 1,872 households. We randomised these households into a treatment group (n=804) that received the nudge intervention, and a control group (n=1,068) that received a placebo intervention (a study lamp) of similar market value as the intervention (for further discussion of the purpose of the placebo intervention, see Section V: Technical Risks below).

Randomisation took place in multiple phases. After baseline was completed in a set of villages, the households in these villages were randomly assigned to treatment and control groups, while baseline took place in the next set of villages. This allowed implementation to happen in villages where the baseline survey had already been administered, while baseline surveying proceeded in the remaining villages.

1. Unit of Analysis and Unit of Treatment

We will have **three** primary units of analysis in our study:

1. **Main respondents**: (n=1,872) those who directly respond to our survey and provide information about themselves and other members of their household. Barrier salience questions will only be asked of main respondents.
    a. Main respondents were chosen during baseline based on availability. If possible, enumerators would survey the primary male decision-maker of the household. Most often, the primary female decision maker was available and would be administered the survey. If neither the primary male or female decision makers were available, the surveyor

would identify a mature member of the household who was able to provide information on other household members.

2. **Full sample of individuals over the age of 18**: (n=5,585) includes main respondents and other members of the household, for whom main respondents will provide demographic information and details on open defecation and latrine use practices.[3]
3. **Households:** (n=1,872) information from each individual will be aggregated at the household level. For example, we will calculate an indicator for whether anyone in the household practices open defecation (see section 3: Outcomes of Interest below).

The primary unit of **treatment** will be the household, since latrines are most commonly owned at the household level and the nudges will be applied to a single household's latrine.

2. Treatment Arms

Our evaluation will have one treatment arm. This choice will ensure that we are able to conduct a rapid evaluation that generates rigorous results to inform the Government of India's rural sanitation policies, as part of the Swachh Bharat Mission. Adding treatment arms to the study rapidly increases sample size, and thus time and resource costs. Importantly, we would expect bundled nudges to have a stronger effect than smaller combinations of nudges. Further, we do not expect any negative complementarities between the nudges. Findings from our evaluation will inform whether more detailed research on individual nudge interventions and related behavioural barriers is warranted. If we find no effects or very small effects of our intervention, further research is unlikely to provide useful insights. If we find sizeable effects of an intervention consisting of all the nudges, then large-scale research across India on low-touch nudge methods to promote sanitation behaviours may be justified.

Each household in the treatment arm will receive all of the finalised nudges. Because surveying extra households in the control group and providing our placebo treatment to control households is logistically simpler than the application of our intervention to latrines in the treatment group, there will be fewer households in the treatment group than in the control group (see *Sample Size Calculations* below for more details on the treatment/control ratio). During prototyping, we estimated the cost of the nudge intervention and used this cost to determine the most appropriate ratio of treatment to control households (0.75).

With only one treatment arm, this impact evaluation will not provide evidence on the relative effects of each nudge. However, the barrier salience modules of our survey are designed to detect changes in the salience of individual behavioural barriers to latrine use that are targeted by a unique set of nudges. For example, should we find that our intervention reduces open defecation rates and the salience of one set of behavioural barriers, while the other behavioural barriers remain just as salient as at baseline, this

[3] An additional 3,864 individuals in our sample are under the age of 18.

would suggest that the nudges targeting that barrier were the primary channel through which the intervention changed open defecation behaviour.

3. Outcomes of Interest

Open defecation: We will collect the following three indicators for the main respondent:

1. Whether the respondent practiced open defecation and/or used a latrine at all in the previous day.
2. Whether the respondent practiced open defecation and/or used a latrine at all in the morning during the past week.
3. Whether the respondent practiced open defecation and/or used a latrine at all at other times of day besides the morning.

Our primary regression specification for the main-respondent-level analysis will be an indicator variable that takes a value of 1 if the main respondent reports defecating in the open in response to any of the three questions above (See Appendix A: Impact Evaluation Pre-Analysis Plan for further details).

For all other individuals in each household, we will ask the main respondent to report whether the individual practiced open defecation at all within the past week.

We will calculate average treatment effects across three different measures of open defecation corresponding to different samples:

1. Main respondents only sample: whether the main respondent practiced OD at any point within the past week.
2. Full sample over the age of 18: whether the individual practiced OD at any point within the past week. Importantly we will restrict this sample to individuals over and including the age of 18 as we designed the nudges for adults and believe they are most likely to work for adults.
3. Household-level: whether any individual in the household practiced OD at any point within the past week.

As a robustness check, we installed door clickers on a random subset of latrines in the treatment and control groups. We will measure latrine use from the door clicker, and this will serve as an additional outcome for our analysis, controlling for the number of latrine users (proxied by household size).

Barrier Salience: To measure the salience of each behavioural barrier, we will ask a series of approximately eight questions per behavioural barrier. Each question will be coded into a binary variable, with a value of 1 indicating that the respondent provided an answer consistent with the barrier being salient and 0 indicating the opposite. For example, if a respondent agrees with the statement “I enjoy using my latrine,” this response would be coded as 0, since agreement with this statement is consistent with the experience and usage barrier not being salient for this respondent. We will aggregate the individual questions into three indices (corresponding to the three target behavioural barriers) following the methodology

developed by Anderson (2008).[4] This index will serve as the overall measure of barrier salience for comparison between the treatment and control groups. The specific questions that will be used to compute aggregate barrier salience scores are reported in Appendix D below.

III. Sample Size

We conducted a series of sample-size calculations aimed at informing our research design. These calculations are based on power calculation conventions and the ongoing OPM/3ie RCT evaluating latrine-use interventions in Bihar.[5] Unlike these studies, we targeted a minimum detectable effect size of 5 percentage points due to the low-touch nature of the intervention. The conservative sample size parameters used for this calculation are listed below. (Further details about the assumptions and sources used to conduct these power calculations may be found in *Appendix C* of this document, *Sample Size Calculation Details*).

Parameter	Pre-baseline estimates	Baseline Value	Source
α	0.05	0.05	Statistical convention
Power (1-β)	0.8	0.8	Statistical convention
Baseline-endline correlation	0.7	0.7	OPM/3ie Bihar study
Proportion of eligible households (with a functional latrine) where at least one member defecates in the open.	0.525	0.377	OPM/3ie Baseline study. While this study comes from a similar geography in rural Bihar, this value is higher than most other studies we have seen, including our own baseline value. We selected this value to ensure we are powered conservatively. (Post-baseline: observed value in our baseline data)

[4] Anderson, Michael L. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association,* 103.484 (2008): 1481-1495.

[5] Saith, R., Viswanathan, S., Lamba, S., Joshi, S., Purty, N., Datta, S. & Harris, J. (2018). *Improving H.A.B.I.T: Households' Attitudes and Behaviours to Increase Toilet Use.* Baseline Report. Oxford Policy Management & Ideas42.

Minimum detectable effect size (reduction in OD)	5%	4.53%	Target minimum detectable effect size (Post-baseline: output of power calculation in Stata)
Treatment/control Ratio	0.75	0.75	Chosen based on the optimal cost of the intervention in treatment and control households.
Total sample size	1,777	1,872	Pre-baseline: Stata Power output, with an additional 10% added as a safeguard against attrition Baseline: actual sample size
Target treatment group sample size	762	804	As above
Target control group sample size	1,015	1,068	As above

1. Parameters for Power Calculations

We exceeded our target sample size for a final sample size of 1,872 (804 treatment and 1,068 control, for a treatment/control ration of 75.28%). This was due to the size of the villages we were surveying and the reliance of our sampling strategy on strictly following the randomised order of villages and saturating each village (for an explanation of our sampling strategy, see section IV below). Using the actual sample size and baseline OD values to calculate the minimum detectable effect size, we find that the minimum effect that can be detected with a 95% confidence interval and power of 0.8 is 4.53 percentage points. This means that if we cannot reject the null that the nudges had no effect on open defecation rates, we can conclude that the true effect of the nudges is unlikely to be greater than 4.53 percentage points.

IV. Sampling and Randomisation

We seek to estimate the causal impact of the nudges on open defecation rates by comparing those who receive the nudges with a counterfactual control group. We will estimate the intent-to-treat (ITT) and not the treatment-on-the-treated (ToT) effect of our nudge intervention on open defecation rates (the ITT will be our primary specification). Importantly, we do not expect these two treatment effects to be very different from each other, as we encountered low rates of refusal of our intervention during implementation (3.78% of households). In addition, the policy, if implemented at scale, would likely encounter some refusal, and thus the ITT estimate will provide a more policy-relevant estimate of changes in open defecation due to our intervention.

1. Population

Our population of interest is latrine-owning households in Darbhanga District, Bihar. We chose Darbhanga district since it is a district in which our NGO implementing partner, FINISH Society, was able to work and had contact with local officials but did not have any active programming in place.

2. Sampling Criteria

We are interested in targeting the households that are the primary targets of the Swachh Bharat Mission: households that are induced to build their latrines by SBM-G programming and the Rs. 12,000 subsidy allocated by the SBM-G. The mission shares our goal of promoting sustained, exclusive latrine use throughout rural India, and we hope that our intervention, if successful, can help inform future SBM-G programming. Determining which households fall into this category is difficult since households may not provide accurate information about their eligibility for SBM subsidies and, especially, whether their decision to construct a latrine was motivated primarily by the subsidy. To target these households, we turn to a useful proxy: latrine type. Households with septic tank latrines[6], the cost of which far exceeds the SBM-G subsidy, are clearly motivated above and beyond the SBM-G and are likely to engage in higher latrine use given their higher revealed willingness-to-pay for latrines. Thus, we target households that own functional latrines, but do not own any septic latrines.

A wide variety of latrine types are found in rural Bihar. The following criteria were used to determine whether a household falls into our sampling frame through brief observation and self-reported data:

1. Household has a fully-functional latrine
2. The household does not own any septic latrines

3. Sampling Strategy and Randomisation Procedure

The first step of our sampling procedure was **block selection**.[7] We selected Manigachhi block based on the following criteria:

1. *Water table depth:* Out of all of the available blocks within Darbhanga district, we selected Manigachhi block since it had the lowest water-table level in official data.[8] Due to concerns about the proximity of leach pits to wells, we wanted to avoid geographies in which the use of latrines could plausibly contribute to water-source contamination. On average, our sample villages are not statistically significantly different from Manigachhi block and Darbhanga district on a number

[6] Septic latrines are defined as latrines connected to a tank, often rectangular in shape, that does not have any leaching mechanism.
[7] Block selection took place through observation of data and through qualitative observation and discussion in consultation with our implementing partner, other IDinsight project teams, and relevant local stakeholders.
[8] Source: http://cgwb.gov.in/GW-Scenario.html

of key demographic variables (caste, religion, age, gender ratio, SES, education). In addition, we find little theoretical justification for correlations between water table depth and any other variables that may affect our outcome variables or the effects of our intervention.

2. *Ability of implementing partner to work in block:* We ensured that FINISH society was able to mobilize their existing resources sufficiently to conduct high-quality implementation.
3. *Absence of other intensive sanitation initiatives,* including active ODF efforts. We ensured that we are not working in an area already receiving other extensive interventions, besides universal Swachh Bharat Mission programming, that may complicate our implementation process and the interpretation of our results.
4. *Non-outlier on key demographic indicators in the 2011 Indian Census and 2013 Economic Census:* In order to ensure the villages in which we work are comparable to others in the region on demographic measures that could affect estimated treatment effects (i.e. that could affect the generalizability of our results), we prioritised selecting a block that was not an outlier on key demographic variables:
    a. Gender ratio
    b. Literacy rates
    c. Caste demographics
    d. Religious demographics
    e. Age
    f. Socioeconomic status

After selecting a block, we constructed our sample, conducted randomisation, and implemented the intervention through the following steps:

1. **Village listing:** Using a list of all villages in Manigachhi block from the 2011 Population Census of India, we ordered the villages in a random order.
2. **Census:** In the first villages on this list, we began a census of households to determine the extent of households that meet our sampling criteria in our target geography.
3. **Saturation:** After determining that we would need to include many villages in our study to achieve the target sample size, we ended the census and shifted to a saturation methodology. Moving through villages following the random order created in step one, we administered the baseline survey to every household that met our sampling criteria in each village. Villages were double-checked by field team leaders after saturation was complete to ensure all households were properly assessed for their fit with the sampling criteria. If this check discovered any households meeting the sampling criteria in which a survey had not been conducted, the survey was conducted in these households.
4. **Stratified Randomisation:** Households were randomised into treatment and control groups, stratified by baseline open defecation rates[9] and week of baseline. After the baseline survey was completed in one set of villages, these villages would be randomised as one group, so that we could move on to surveying the next round of villages, while the implementation team administered the intervention in the first group of villages.

[9] Specifically, an indicator variable for whether the main respondent reports open defecating within the last week.

5. **Offer and administration of intervention:** Households in the treatment group received an offer to have the nudge intervention implemented to their latrine from our implementing partner, FINISH Society. The offer to receive the nudges was framed similarly to every household. Those who consented immediately received the intervention through our implementing partner. Similarly, control households were offered a placebo intervention (a study lamp).

Randomisation took place using the statistical software package Stata. Randomisation occurred in a series of stages, using the following procedure:

1. **Collect Baseline Data in a set of villages:** Prior to randomisation, we conducted baseline data collection in a set of villages, saturating each village.
2. **Create randomisation strata:** We created randomisation strata based on baseline OD rates, as well as the week in which baseline was conducted.
3. **Stratified random assignment within the set of villages:** We conducted stratified random assignment of households to the treatment and control group in Stata.

After completing a group of villages, we would repeat the three steps above in a new set of villages. This methodology was chosen to allow baseline and implementation to proceed simultaneously. After one set of villages was surveyed and randomised, the survey team moved on to the next set while the implementation team implemented in the households of the preceding set.

Our original Treatment/Control (T/C) ratio target was 0.68. During baseline collection, we decided to adjust this target, settling finally on a ratio of 0.75. We accidentally included an additional 100 control households in the original budget and contract with our implementing partner, and we randomised at 0.68 to take these households into account. Due to timing considerations, and the likelihood that we would need to include extra households in the study regardless, due to our saturation sampling approach as well as the large size of the final village in our sample, we later decided to exclude the 100 extra control households from our targets and use the original target fraction of 0.75. The first two strata were randomised at a T/C ratio of 0.68. Succeeding strata were randomised at a higher T/C ratio to achieve our overall T/C ratio target. Our overall T/C ratio for the study was 0.75, as specified in our sample size calculations. To account for stratified randomisation, we include stratum fixed effects in all our regression specifications, as reported in Appendix A.

The following table displays each of the strata in our study and the corresponding T/C ratio:

	Household OD rate	Number of Households	Number of Villages	Treatment/Control Ratio
Village-group 1	1	135	8	0.66 (54/81)
Village-group 1	0	174	8	0.69 (71/103)
Village-group 2	1	109	7	0.60 (41/68)
Village-group 2	0	127	8	0.65 (50/77)
Village-group 3	1	170	3	0.75 (73/97)

Village-group 3	0	310	4	0.73 (131/179)
Village-group 4	1	246	3	0.88 (115/131)
Village-group 4	0	307	3	0.88 (144/163)
Village-group 5	1	110	1	0.75 (47/63)
Village-group 5	0	180	1	0.75 (77/103)
Excess village-group[10]	1	3	2	0.5 (1/2)
Excess village-group	0	1	1	0 (0/1)

4. Balance Tests

The primary reason for conducting an RCT is to create a convincing counterfactual group to estimate the causal effects of the intervention. RCTs accomplish this by ensuring that the treatment and control groups are *ex ante* identical in terms of any observable characteristics that may affect outcomes, besides treatment status. Due to random chance, however, the treatment and control groups may have systematic differences, which may reduce our confidence that any differences in outcomes we observe will be due to the nudge intervention alone. To assess this problem, we conducted t-tests for differences of means on relevant baseline variables to determine whether balance between the treatment and control arms had been successfully achieved.

Our balance tests show no significant differences at conventional levels between the treatment and control groups across any variables of interest. **Tables 1 and 2** below display these results. The final column in these tables reports the difference in means between the treatment and control groups.

V. Data Collection and Validation

1. Types of Data

Our data will be grouped as follows:

- **Individual characteristics** including self-reported open defecation, salience of the behavioural barriers, and other individual-level characteristics such as age and education.
- **Household characteristics** such as caste, religion, and socioeconomic status.
- **Latrine characteristics** such as functionality and latrine-type etc.
- **Barrier salience questions** aimed at understanding the extent of each barrier for the main respondent.
- **Nudge-related question** for endline specifically, aimed at understanding the extent of exposure to the nudges, the duration of the nudges, and perceptions of the intervention.

[10] 4 households in 2 villages were erroneously not surveyed in time for inclusion in the randomisation set. These households were randomised as a separate group at the end.

Table 1: Balance tests for sample of main respondents

	(1) Control		(2) Treatment		(3) Total		T-test Difference
Variable	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)-(2)
% SC	1068	16.479	804	16.542	1872	16.506	-0.063
		(1.136)		(1.311)		(0.858)	
% Brahmin	1068	8.708	804	8.582	1872	8.654	0.126
		(0.863)		(0.988)		(0.650)	
% Muslim	1068	12.640	804	14.552	1872	13.462	-1.912
		(1.017)		(1.244)		(0.789)	
OD Rate (%)	1068	21.629	804	20.771	1872	21.261	0.858
		(1.260)		(1.432)		(0.946)	
% Female	1068	57.865	804	56.343	1872	57.212	1.522
		(1.512)		(1.750)		(1.144)	
Age	1068	40.860	804	41.505	1872	41.137	-0.645
		(0.452)		(0.521)		(0.341)	
Household Size	1068	5.096	804	5.103	1872	5.099	-0.008
		(0.073)		(0.082)		(0.055)	
PPI Score	1059	32.356	798	32.447	1857	32.395	-0.091
		(0.495)		(0.597)		(0.381)	

Notes: The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 2: Balance tests for full sample

	(1) Control		(2) Treatment		(3) Total		T-test Difference
Variable	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)-(2)
OD Rate (%)	5407	18.291	4055	17.509	9462	17.956	0.782
		(0.526)		(0.597)		(0.395)	
Age	5466	27.520	4104	27.576	9570	27.544	-0.056
		(0.278)		(0.318)		(0.209)	
% Literate	4933	62.396	3725	62.228	8658	62.324	0.168
		(0.690)		(0.794)		(0.521)	
% Female	5458	53.830	4104	53.752	9562	53.796	0.077
		(0.675)		(0.778)		(0.510)	

Notes: The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

2. Data Quality and Validation

IDinsight is able to ensure high data quality through the work of our field-management staff, who are experts at training and recruiting well-performing survey teams and administering surveys. The following additional steps will be followed to ensure data is accurate and high-quality:

- **Survey administration** will make use of SurveyCTO, a program that allows for programming surveys into smartphones to ensure data collection can be more easily controlled and monitored by research teams. We will program the survey according to best practices to prevent surveyor mistakes or malpractice, avoid inconsistencies, and ensure high-quality data collection. Some examples of best practices include audio audits of survey sessions, GPS coordinate reports, and the use of photos to validate location.
- **High frequency data quality checks** will be performed every day for both rounds of data collection to identify issues in data quality and inform us as to which surveyors to monitor.
- **Audio audits** triggered by suspicious data patterns will allow us to audit individual surveys and hear the interaction between surveyors and respondents. We will focus especially on surveyors flagged in our high-frequency data quality checks.
- **Clear contractual language and incentives** for surveyors will ensure that they are aware that quality data-collection is a precondition to continued work on the project team and receipt of related benefits. In addition, these expectations will be set during training. Due to IDinsight's extensive data collection efforts in the state of Bihar, surveyors will be incentivized to do their best work in order to have continued employment opportunities with IDinsight.
- **Strong and proven surveyor training program**. Across IDinsight engagements, including in an extensive 27,000-household survey IDinsight recently completed across India, IDinsight has developed and refined best practices for surveyor training. This includes aspects of training such as the type of training manuals that work best, the style and content of training, management by IDinsight's professional Field Managers who run the training, and optimized training schedules.

VI. Technical Risks

We are confident in the ability of this study design to produce internally valid estimates of the effects of the nudge intervention on latrine use behaviour. Nonetheless, we have identified various areas where the study design is at risk of producing biased results. The following sections discuss these risks our project may face, each of which poses challenges to our ability to recover unbiased estimates of the effects of our intervention. In addition, we highlight the steps we will take to minimize these risks and ensure that our analysis produces rigorous causal impact estimates.

1. Attrition

Households may drop out from the study between baseline and endline. There are three reasons why we worry about attrition. First, attrition could be random, in which case there are no systematic differences between those we can and cannot find at endline. This does not affect internal validity, but reduces the

statistical power of the study. Second, attrition may be non-random, in which case the attriters may differ significantly from the non-attriters, changing the composition of our sample. This still preserves internal validity, but may reduce the generalizability of our results. Finally, most damaging for internal validity, we may have differential attrition between the treatment and control groups.

We have taken or will take the following steps to reduce attrition in general:

- We will inform households that there are two survey rounds, and ask for consent for both survey rounds at baseline. Only those who consent to participating in both surveys will be included in the study.
- Identifying information and contact information for each household and their neighbours will be collected during the baseline survey, along with best times to contact household members.
- Data on key outcome variables will be collected before other data to guard against attrition while the survey is being administered. Should respondents become tired in the midst of the survey, they will already have answered the key questions related to our outcomes and can be a part of our analysis, rather than counted as attrited. This may be especially important as our survey will be slightly different between the treatment and control groups.
- We will make multiple attempts to reach each respondent from our baseline survey at endline.
- We powered our study conservatively (adding an additional 10% to our calculated sample size) to guard against loss of statistical power due to attrition.

To test whether the attriters vary from the non-attriters in a statistically significant way, we will compare baseline characteristics between the two groups. These characteristics include the outcome variables and all control variables used in our analysis. In case attrition is non-random, we will employ multiple imputation and/or inverse probability weighting to address this, using all available information on the attriters from baseline.

Finally, we will test whether attrition is differential between the treatment and control groups, a major risk to the internal validity of our estimates. We will conduct this analysis by regressing a binary outcome variable equal to 1 if the household is surveyed during both baseline and endline, and 0 otherwise, on treatment status, strata fixed effects, and baseline values of our covariates. We will additionally test for differential attrition through regressing attrition on a dummy variable for treatment, all baseline covariates, and the treatment dummy interacted with baseline covariates. We will conduct a joint F-test for all of the interaction terms.

In case attrition is significantly different across treatment and control groups, we will bound our treatment effect estimates using Manski and/or Lee bounds:

1. **Manski bounds** are a highly conservative bound on Average Treatment Effect estimates. The lower bound estimate assumes all attriters are not affected by the intervention, and the upper bound estimate assumes all attriters are affected by the intervention.
2. **Lee bounds** are another way to bound the Average Treatment Effects that generate tighter bounds on the ATE than Manski bounds. But this method requires additional monotonicity

assumptions–that treatment can affect attrition only in one direction (both treatment and the absence of treatment cannot be causing attrition).

In some cases, Manski and Lee bounds may be too wide to be informative. If warranted, we may make use of inverse probability weighting to account for differential attrition in our sample.

2. Spillovers

Neighbours of households who receive the nudge intervention may be aware of, discuss, or even share aspects of the intervention, leading to spillovers. For example, neighbours may hear about some of the additions made to others' latrines and imitate them on their own. If the neighbours are in the control group, then any aspects of the intervention that affect neighbours' behaviour, either positively or negatively, would prevent us from recovering unbiased estimates of the treatment effect. Such a control group whose behaviour is affected by the intervention given to treatment households is no longer truly a control group – it no longer provides a good comparison of what would have happened had the treatment group not received the nudges.

Spillovers could occur through one of two channels: **The first channel is vision and discussion**. Individuals in the control group may see the nudges in their neighbours' latrines and/or discuss the nudges with other members of their community. Relatedly, if the nudges have a large effect on defecation practices in a village, control households' understanding of the norms of defecation in their community could change. This change in perceived norms–what individuals think others in their community do or think–could potentially alter control households' behaviour. **The second channel through which spillovers could occur is actual usage**. This would occur if a member of the control group uses a latrine belonging to someone in the treatment group, such as a neighbour or family member. This use could be sporadic or regular depending on the relationships between individuals in the treatment and control groups. Spillovers through this channel could also occur if households borrow concepts from the nudge intervention or the nudges themselves directly from the treatment group and apply these to their own latrines.

After conducting extensive qualitative work in rural Bihar during the design of our nudge intervention, we believe spillovers are unlikely to pose major issues to our study for the following two reasons:

1. **The nature of the nudges is such that they are tailored to the users of a single latrine**. The two most publicly visible components of the intervention–the pit-emptying poster and gendered-usage poster–include crucial components relevant only to one household. The pit-emptying poster contains an expected date of pit-emptying that is specific to each household, and the gendered usage poster contains a toilet use schedule created for one family. In addition, the cleaning supplies and the radio will be inside one latrine and most likely used for that latrine alone (though households could potentially sell or gift these items to other households). Overall, we believe the nudges are sufficiently individualized and private that control households will not be able to engage with the nudges to the extent required to change their behaviour, and thus spillovers are not a major concern.

2. **The nudges are "low-touch," and we are targeting a small minimum detectable effect size.** We expect an effect size that is no greater than 10% (our theoretical prior grounded in our theory of change and our assessment of recent studies on open defecation and similar behavioural interventions). Additionally, our intervention targets only a subset of all households in a given village. Of all households in a village, we focus on pit-latrine owners who do not own septic latrines. This excludes septic latrine owners and those who do not own latrines. Of the households meeting our sampling criteria, only a fraction (43% on average) will receive our intervention. Based on all of the above considerations, we do not expect our intervention to appreciably change control households' expectations of their neighbour's beliefs and behaviours to a degree large enough to influence the control group's behaviour. One caveat to this, however, would be if social norms and expectations are only formed within groups of individuals who own the same types of latrine. Even still, the low-touch nature of the intervention suggests changes in social-norms around open defecation and latrine use are unlikely. We expect our intervention to work through changing individual and household behaviour alone, a crucial aspect of what makes the intervention "low-touch."

Nonetheless, spillovers are a threat to our ability to estimate an unbiased treatment effect of our intervention. Despite our expectations, households may borrow, copy, or be influenced by the intervention, and these actions may impact control households' defecation behaviour.

The following steps have been taken, or will be taken, to prevent and address spillovers:

- During prototyping, nudge designs were evaluated on their potential for generating spillovers and the likelihood of sharing between neighbours.
    - During prototyping, we asked a sample of villagers who did not receive the intervention whether they were aware of the intervention in their neighbours' homes. Most respondents were aware that some intervention had been given to their neighbours, but did not know the details of what the intervention consisted of. This suggests that spillovers are not likely to be important in this particular context.
- During the process evaluation and at endline, recipients and non-recipients of the intervention will be surveyed about their awareness of the nudges, and whether they have discussed or shared the nudges with their neighbours, and whether and how frequently they have used neighbours' latrines. In addition, our process evaluation, baseline, and endline survey will include a short social network analysis to better understand with whom households engage in sanitation-related discussions and assess the extent to which control households may have been exposed to the treatment in nearby treatment households. If relevant, these variables may also play a role as control variables in our analysis. These questions will come after key outcome questions to avoid biasing the results.
- All control households received a placebo treatment of similar monetary value. This placebo treatment is a study-lamp. The placebo will help avoid situations where control households resent not receiving the intervention and will diminish jealousy and comparison between households.

3. Social Desirability Bias

Open defecation practices may be a sensitive topic for survey respondents. Additionally, the politics of ODF (Open Defecation Free) status declaration as part of the SBM-G in contexts where open defecation is still practiced may lead survey respondents to underreport open defecation and overreport latrine use. Due to the large-scale nature of the Swachh Bharat Mission and the extensive nature of latrine-use communications throughout rural India, participants in the study may feel that there is a "right answer" they should provide when reporting open defecation.

Importantly, studies continue to report non-zero rates of open defecation in our target geography, suggesting social desirability bias is not an overwhelming problem.[11] In addition, if answers are manipulated in the same way in treatment and control groups, any differences found in our analysis will still be attributable to the intervention. Still, social desirability bias is a cause for concern. For example, households that receive the latrine-related intervention may feel an obligation to give more socially desirable answers due to the fact that the interventions they received visibly target open defecation behaviour.

The following steps will be taken to minimize the risk of social desirability bias:

- Questions about OD and latrine use will be included among other demographic questions so as not to place too much emphasis on latrine use.
- OD questions will be asked as select-multiple questions: allowing individuals the opportunity to report both using a latrine and practicing open defecation so as not to unduly highlight open defecation behaviour.
- Questions about OD will be prefaced by a value-neutral statement (e.g. "We have spoken with many people in this area. Some people prefer to defecate in the open and some prefer to use latrines.")
- The purpose of the nudge designs and research (eliminating OD) will not be directly revealed.
- Our research team will take care to establish neutrality and non-judgement before administering surveys. Our surveyor training will highlight this important task.
- The survey and intervention implementation teams will be kept separate, so that those who receive the intervention do not feel compelled to inflate their reported latrine use out of a desire to please the implementers. The groups will not be explicitly affiliated in communications over the course of the project.
    - The implementing partner will make offers to receive the nudges, apply the nudges to households, and provide any other implementation support needed.

[11] Gupta, A., Khalid, N., Deshpande, D., Hathi, P., Kapur, A., Srivastav, N., Vyas, S., Spears, D., & Coffey, D. (2019). *Changes in open defecation in rural North India: 2014-2018.* Working Paper.

Saith, R., Viswanathan, S., Lamba, S., Joshi, S., Purty, N., Datta, S. & Harris, J. (2018). *Improving H.A.B.I.T: Households' Attitudes and Behaviours to Increase Toilet Use.* Baseline Report. Oxford Policy Management & Ideas42.

- The survey team will collect evaluation data at separate times from implementing partner activities.

4. Evaluation-Driven Effects

Evaluation-driven effects refer to effects on the outcome variable driven by the evaluation itself. These effects are related to evaluation activities, and are independent of the actual intervention. For example, if the implementation team simply comes to the household's door and discusses the nudges, this may have an effect on defecation behaviour, regardless of whether the nudges are ever implemented. If there are evaluation-driven effects in our evaluation, then we would be worried that the treatment effects we estimate are biased due to the inclusion of these effects in the estimate.

A number of possible evaluation-driven effects could pose problems for our project. Households may change their behaviour to impress evaluators, or may temporarily change their behaviour since they are participating in the study and may intuit the study's purpose. In addition, households in the control group may change their behaviour in order to attempt to receive the intervention if it is rolled out to more households. We will take the following steps to prevent bias to our results due to evaluation-driven effects:

- Study protocols will intentionally minimize the potential for evaluation-driven effects by specifying clear actions and protocols for all steps of the evaluation.
- We will ensure the evaluation team and implementation team work separately and have clearly defined roles. Teams will have strict instructions not to engage in unnecessary interactions with households on topics related to defecation that may increase the probability of evaluation-driven effects.
- We will ensure similar staff-interaction in treatment and control groups. The same surveys, survey enumerators, and protocols will be used in interactions with both groups.

5. Concerns Related to External Validity

While many of IDinsight's engagements are decision-focused and emphasize internal over external validity concerns, this project contains a notable knowledge-focused component. Thus, we would be concerned if the results from our evaluation do not generalize to other similar contexts.

Importantly, we do not expect the results from our impact evaluation to generalize easily to contexts beyond rural Bihar. The reasons for open defecation and behavioural barriers to latrine use vary across India. In addition, the extent and nature of spillovers, as well as compliance rates, may vary widely between villages, even those that are near to one another.

Prior to conducting baseline, we conducted a series of statistical tests using 2011 Indian census data to ensure that Darbhanga district is not significantly different from the average Bihar district on a set of key demographic variables. We likewise used the same tests to analyse Manigachhi block relative to the

average of all blocks in Darbhanga district, and we concluded on the basis of these tests that the two are not significantly different on key observable demographic characteristics. Thus, we believe **our results will be generalisable to the context of rural Bihar**. The population of our study reflects the kind of population that would be likely to receive this intervention should it be scaled up.

Additionally, a number of techniques exist to estimate bounds on the population average treatment effect for experiments where inclusion in the randomisation set is opt-in or geographic locations are not randomly sampled.[12] If desired by stakeholders, we may seek to employ such methods to make claims about the external validity of our study.

Also, a number of other studies are currently being conducted on latrine ownership and defecation practices throughout Bihar and much of India. As feasible, we will seek to relate our own data to the data in other studies to better understand the generalizability of our results.

[12] See, for example Andrews, I., & Oster, E. (2017). *Weighting for External Validity* (No. w23826). National Bureau of Economic Research.

Appendix A: Impact Evaluation Pre-Analysis Plan

Primary Specifications

The following sections provide detailed analysis plans for each research question. The first questions below are grouped together into one section since they will be analysed identically, but with different outcome variables.

Q1.1: What is the effect of the intervention on rates of latrine use and open defecation among those who are offered the intervention and those who are not offered the intervention?
+
Q1.2: What is the effect of the nudge intervention on the salience of the three behavioural barriers to latrine use that we have identified as most relevant?
+
Q2.1: Does treatment affect other latrine-oriented behaviours, such as upkeep, repairs, and cleaning of toilets?

Primary Specification (linear regression model) for analyses at the level of all household members older than 18:

$$(1) \qquad Y_{ihs} = \beta T_h + X_h'\theta_1 + X_i'\theta_2 + \lambda_s + \epsilon_{ihs}$$

Primary Specification (linear regression model) for analyses at the level of main respondents and households:

$$(2) \qquad Y_{hs} = \beta T_h + X_h'\theta_1 + \lambda_s + \epsilon_{hs}$$

For individual *i* living in household *h* in stratum *s*.

The variables are defined as follows:

- Y_{ihs} refers to the outcome variable of interest
- T_h is an indicator variable that takes the value of 1 for households in the treatment group and 0 for those in the control group
- X'_h is a vector of household-level control variables measured at baseline:
    - PPI score (a common measure of socioeconomic status)
    - Household size
    - Caste
    - Religion
    - Number of latrines
    - Cleanliness of latrines
- X'_i is a vector of individual-level control variables measured at baseline:
    - Gender
    - Age
    - Literacy
    - Educational attainment
    - Baseline OD rate

    - Baseline responses to questions measuring the salience of behavioural barriers
- λ_s represents stratum-level fixed effects for a total of 12 strata.
- ϵ_{ihs} is the error term for an individual observation
- ϵ_{hs} is the error term for a household observation

Given that our treatment is assigned at the household level, we will not cluster standard errors for analyses at the level of households and main respondents (since there is one main respondent per household). For analyses at the level of all household members, we will cluster standard errors at the household level.[13]

Definitions of outcome variables:

Var Name	Code	Definition
Open Defecation (main respondents)	od_main_resp_any	Coded as 1 if the main respondent reports engaging in open defecation within the previous week, 0 otherwise. Maximum of three variables: 1. Indicator for whether the respondent reports engaging in open defecation the previous day 2. Indicator for whether the respondent reports engaging in open defecation in the morning during the previous week 3. Indicator for whether the respondent reports engaging in open defecation at other times of the day during the previous week.
Open Defecation (all individuals)	od_i	OD_i is coded as 1 if a respondent over the age of 18 reports that household member i has engaged in open defecation during the previous week, and 0 otherwise.
Open Defecation (households)	od_hh_any	Coded as 1 if anyone in the household reports engaging in open defecation during the previous week, and 0 otherwise.
Barrier Salience	bar_#_salience	Anderson index constructed from likert-scale and yes/no statements geared around barrier salience For these statements (with the exception of yes/no and multiple-choice questions) respondents will be given three options for responding: one that indicates strong agreement, one that indicates medium agreement, and one that indicates disagreement. When relevant, the questions will

[13] Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (No. w24003). National Bureau of Economic Research.

		always be worded so that the two agreement statements will apply to the more socially undesirable choice so that respondents can signal agreement with a socially undesirable statement without indicated full agreement. For the purposes of analysis, we will collapse both levels of agreement into one, and create from each question an indicator variable that takes on a value of 1 if the respondent agrees with the statement. The precise questions used to address each barrier are reported in Appendix D below.
Observed latrine use	clicker_count	Defined as the count of number of times a latrine door has been opened as counted by clickers installed on latrine doors (the final count-the initial count on the clicker when installed)
Latrine upkeep	lat_upkeep / lat_clean	Number of times in week that latrine is cleaned and overall condition (clean or not clean, self-reported and reported by enumerator)

Hypotheses:

- (Null hypothesis) H_0: $\beta = 0$
- (Alternative hypothesis) H_a: $\beta \neq 0$

Unit of Analysis: Individual (i), clustered by household (h); household (h)

Outline of Stata Code:

```
local controls "..."
reg <outcome> treat `controls' i.strata_id, vce(cluster hh_code)
```

Secondary Research Question 2.2: How do treatment effects vary by individual and household characteristics?

Primary Specification (linear regression model):

$$Y_{ihs} = \gamma_1 T_h + \gamma_2 Z_{i,h,s} + \beta\left(T_h * Z_{i,h,s}\right) + X_h'\theta_1 + X_i'\theta_2 + \lambda_s + \epsilon_{ihs} \quad (3)$$

All variables are as defined above, and $Z_{i,h,s}$ refers to the variables along which heterogeneous treatment effects are calculated, which may be individual- or household- level variables including:

- Gender

- Age
- Literacy
- Educational attainment
- Baseline OD rates
- Religion (indicator variables across categories)
- Caste (indicator variables across categories)
- PPI scores
- Household size
- Percentage nudges retained
- Reported experience of coercion under SBM-G

These subgroup tests are exploratory, rather than part of our principle research question. Nevertheless, to account for the fact that we are testing multiple hypotheses, and therefore the rate at which we do not reject null hypotheses that should be rejected (i.e. make a false discovery), we will check whether the subgroup results are robust to adjusting the p-values using the Benjamini and Hochberg procedure.

Hypotheses:

- (Null hypothesis) H_0: $\beta = 0$
- (Alternative hypothesis) H_a: $\beta \neq 0$

Unit of Analysis: Individual (i), clustered by household (h)

Sample: Under high take-up, the sample will consist of all households.

Outline of Stata Code:

```
local controls "..."
reg <outcome> treat##<z-variable>`controls' i.strata_id, vce(cluster
hh_code)
```

Robustness Checks

1. Door clicker

We will estimate the same equation as the primary household specification above (equation 1) above but with the door clicker reading as the outcome variable.

2. Sample including children

We will estimate equation (1) including all individuals under 18 to determine whether our intervention also had similar effects in a sample including children and young adults.

3. Difference-in-differences specification

We will test whether our treatment effects are robust to using a difference-in-differences specification, in which we compare the *change* in reported OD rates between the treatment and control groups. Similar to before, our regression specification would be:

$$\Delta Y_{hvs} = \beta T_h + X_h^{'}\theta_1 + \lambda_s + \epsilon_{hvs} \quad (4)$$

and

$$\Delta Y_{ihvs} = \beta T_h + X_h'\theta_1 + X_i'\theta_2 + \lambda_s + \epsilon_{ihvs} \quad (5)$$

For individual *i* living in household *h* in stratum *s*.

ΔY_{hvs} is the change in the outcome variable between baseline and endline. The other variables are the same as defined above.

4. Village Fixed Effects

We will estimate the following model with village fixed effects to account for unobservables at the village level. If our randomisation process is correct, we should expect to find that the inclusion of village fixed effects does not qualitatively change the results.

Individual-level outcomes:

$$Y_{ihvs} = \beta T_h + X_h'\theta_1 + X_i'\theta_2 + \lambda_s + \alpha_v + \epsilon_{ihvs} \quad (6)$$

Household-level outcomes:

$$Y_{hvs} = \beta T_h + X_h'\theta_1 + \lambda_s + \alpha_v + \epsilon_{hvs} \quad (7)$$

Appendix B: Descriptive Statistics Pre-Analysis Plan

We will generate descriptive statistics (mean, median, standard deviation, maximum, minimum etc) for the following variables:

Survey:

- Survey durations
    - By enumerator
    - By gender
    - Others present
- Follow-up rate
- Consent given
- Attrition
- Presence of others
- Responsiveness
- Comprehension
- Number of incomplete surveys
- Number of villages
- Households per village
- T/C ratio per village and per randomisation stratum

Household level:

- Number of latrines
- Type of latrines
- Latrine dates of construction
- Age distribution within households
- Amount paid for latrines
- Latrines by SES quartile
- Latrines with doors
- Who built the latrine(s)
- Number of times latrines are cleaned in week
- Who does the cleaning within the family
- OD (at all sample levels) and barrier salience by:
    - Religion
    - Caste
    - Gender
    - Literacy
    - Education
    - PPI Quartile
    - Enumerator gender

Individual level:

- HH role
- Marriage

- Occupation
- Religion
- Caste
- surnames/ jatis
- Age

Process indicators:

- Which individual nudges were given or not given (observed and reported)
- Reasons provided for any nudges not being present
- Extent of water shortages in study period
- Functionality of latrines in study period
- Indicators of whether all nudges are being used as intended
- Questions related to the study lamps
- Social network
- Summary of all likert-style questions:
    - The new paint makes my toilet look better than before
    - The new paint makes using the toilet a more pleasant activity.
    - Did your family use the toilet kit (toilet brush, harpic and odonil) to keep the toilet clean?
    - My toilet is cleaner now
    - Toilet kit makes using the toilet a more pleasant activity
    - Can you read the message written in the pit emptying poster?
    - Can you please tell me what is the main message of the poster?
    - How did you understand the message in the poster?
    - Knowledge of pit emptying time reduces my worry about emptying the pit.
    - Knowledge of pit emptying time allows me to use the toilet with less worry.
    - Do you listen to radio in the toilet?
    - Does anyone else in the family listen to radio in the toilet?
    - Listening to radio in the toilet makes using the toilet a more pleasant experience.
    - Can you read the message written in the family poster?
    - Can you please tell me what is the main message of the poster?
    - How did you understand the message in the poster?
    - In the past week how often did men in your family follow the toilet schedule in the morning or evening?
    - In the past week how often did women in your family follow the toilet schedule in the morning or evening?
    - It is easy to use the toilet schedule
    - Men in the family like to use the toilet schedule
    - Women in the family like to use the toilet schedule
    - Men like to use the toilet schedule more than women in the family

Appendix C: Initial sample size calculation assumptions

1. Research questions and outcomes

Our power calculation assumptions were chosen to power primary research question 1 (effect of treatment on OD and latrine use outcomes). Open defecation is the most important outcome we will analyse in this study and provides justification for the study as a whole.

The highest level of analysis at which we will analyse open defecation is the household. Thus, we treat households as individual units for the purpose of these power calculations. This provides us with sufficiently conservative estimates to ensure we are powered for all other research questions and aligns with similar studies on open defecation that analyse household-level indicators.

2. Calculation procedure

Sample size was calculated in Stata, a statistical computing program, using Stata's built-in power command along with adjustments to the command created by our team. These adjustments were designed to accommodate more precise information about autocorrelation that is not captured in the basic version of the command. We use a two-proportions test to reflect the nature of our outcome indicator.

3. Components of calculation

Sample Size and Treatment/control Ratio

Sample size is the primary outcome of interest of our power calculations. Because the marginal cost of additional treatment households will be higher than the marginal cost of additional control households, we have specified a treatment/control ratio of 0.75. We made use of the following standard equation for calculating the optimal ratio[14]:

$$\frac{P_t}{P_c} = \sqrt{\frac{C_c}{C_t}}$$

(Our realised costs were INR 890 per control household and INR 1,440 per treatment household, suggesting an optimal T/C ratio of 0.79, close to 0.75.)

Power and Statistical Significance

We have specified statistical significance at the 5% level along with 80% power. This is in line with best practice across the social sciences and within IDinsight's projects.

[14] Equation taken from Glennerster and Takavarasha (2013) *Running Randomized Evaluations: A Practical Guide*

Minimum Detectable Effect Size (MDES)

We have chosen a minimum detectable effect size of 5%. This represents the smallest change in latrine use rates that we are powered to detect given all the above parameters. Notably, this value is smaller than the standard for similar studies. We believe that this is justified due to the low-touch nature of the intervention. Because the intervention is low-touch, our team believes that the effect is likely to be smaller than a high-touch, social-networks-focussed intervention, such as a large-scale community-led total sanitation (CLTS) intervention. In addition, our intervention is inexpensive. Thus, policy-makers may be justified in implementing the intervention even if its effects are small given the simplicity and cost-effectiveness of the intervention. For these reasons, we are confident a 5% MDE is appropriate for this project.

Baseline rates

Our baseline rates are taken from the baseline data of a 3ie/Oxford Policy Management study currently being implemented in rural Bihar. In this study, the baseline OD rate among latrine-owning households is 52.5%. This represents the most recent information from the most similar geography to our study possible. However, this number is significantly higher than in other studies (and also requires a higher sample size to achieve desired power). Thus, we believe this represents a conservative parameter that will help ensure our study is sufficiently powered.

(In our baseline data, we find a household-level OD rate of 37.7%.)

Autocorrelation

In general, for panel surveys on open defecation autocorrelation may range from 0.7-0.9.[15] Lower autocorrelation values require higher sample sizes. Thus, 0.7 represents a conservative estimate of correlation in outcomes between baseline and endline.

Adjustment for Attrition

To adjust for possible attrition, the output of our calculations was inflated by 10% for each arm to ensure power is maintained even if some households drop out of the study between baseline and endline. This value is an estimate of what may be required to address attrition.

4. Sample Size Calculation Description

Using the parameters above, the recommended sample size is n=1,777 with 762 households in the treatment group and 1,015 households in the control group.

[15] Saith, R., Viswanathan, S., Lamba, S., Joshi, S., Purty, N., Datta, S. & Harris, J. (2018). *Improving H.A.B.I.T: Households' Attitudes and Behaviours to Increase Toilet Use.* Baseline Report. Oxford Policy Management & Ideas42.

5. Sensitivity to Changes in Baseline Rates and Autocorrelation

Should baseline rates of OD be different in our target geography, the sample size would be affected. The same holds true for changes in autocorrelation. The following two graphs present how our target sample size would change with changes in these two parameters.

Figure B.1: Total Sample Size

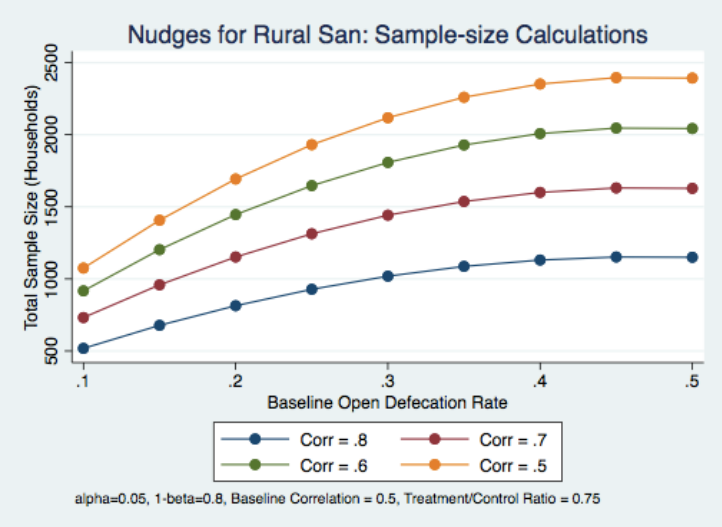
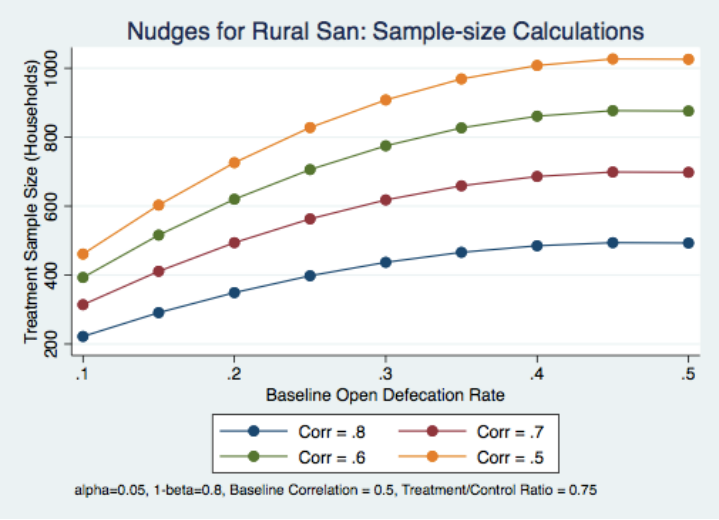


Figure B.2: Treatment arm sample size



Appendix D: Barrier Salience Questions

The following questions will be used to assess the pit-emptying barrier:

1. (Likert) I don't know when my pit will fill. This fact bothers me.
2. (Likert) My pit will fill some day and then I will have to empty the pit. I am worried that I will have to empty the pit
3. (Likert) I am confident that I will find a way to empty my pit once it is full.
4. (Likert) I am willing to self-empty my own pit when it is full.
5. (Likert) My pit will fill one day and I will have to empty it. This fact worries me and prevents me from using the latrine.
6. (multiple choice) Of the following options, which would be your preferred method of emptying your pit? (assessing willingness to self-empty)

The following questions will be used to assess the gender barrier:

1. (yes/no) Do the men of this household use the latrine as frequently as the women of this household do?
2. (Likert) It is more important for women to use the latrine than for men to use the latrine.
3. (Likert) One of the most important reasons to construct a latrine is to protect the dignity of women.
4. (Likert) It is natural for men to defecate in the open.
5. (Likert) It is natural for women to defecate in the open.
6. (Likert) In our household, sometimes when women want to use the latrine they are not able to because it is occupied.
7. (Likert) In our household, sometimes when men want to use the latrine they are not able to because it is occupied
8. (Likert) It is alright for men and women to share the same latrine.

The following questions will be used to assess the experience barrier:

1. (Likert) I prefer defecating in the open to defecating in a latrine.
2. (Likert) Open defecation is less work than using a latrine.
3. (Likert) Defecating in the open is more enjoyable than using a latrine.
4. (Likert) To maintain the latrine it is difficult to decide who will clean the latrine and when.
5. (Likert) I feel that I have to do a lot of work to clean and maintain my latrine.
6. (Likert) I like using my latrine.
7. (Likert) My latrine is dirty.