Impact of Survey Mode on Measuring Water, Sanitation, and Hygiene (WaSH) Outcomes: Insights from a Randomized Control Trial in Ethiopia

by

Dewan Abdullah Al Rafi

(Under the Direction of Ellen McCullough & Gregory Colson)

Abstract

This study examines how survey mode influences self-reported Water, Sanitation, and Hygiene (WaSH) outcomes, focusing on social desirability bias as a key mechanism. Using a randomized controlled trial (RCT) in Ethiopia's North Wollo zone, we compare phone-based and face-to-face survey responses. Our findings reveal that Phone survey group has reported significantly higher rates of improved water access by 9.3 percentage points, improved sanitation facilities by 27.5 percentage points, and increased hand washing frequency by 0.96 times higher, compared to In-person respondents. These differences are consistent with increased social desirability bias in telephone surveys. Our findings remain robust to individual enumerator effects. We investigated several alternative mechanisms including survey fatigue, enumerator-specific effects, and enumerator learning; finding no empirical support for these explanations. While we cannot exclude the possibility that effects stem from in-person enumerators' ability to verify respondent claims, the evidence suggests social desirability bias as the primary driver. Our findings highlight important methodological implications for development research and policy evaluation, emphasizing the need for adjustments in survey design to mitigate mode-induced measurement biases in WaSH-related studies.

Index words: Survey Methodology, WaSH Outcomes, Data Quality, Social Desirability Bias

Impact of Survey Mode on Measuring Water, Sanitation, and Hygiene (WaSH) Outcomes: Insights from a Randomized Control Trial in Ethiopia

by

Dewan Abdullah Al Rafi

M.Sc., The University of Arizona, 2023

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Master of Science

Athens, Georgia

2025

Impact of Survey Mode on Measuring Water, Sanitation, and Hygiene (WaSH) Outcomes: Insights from a Randomized Control Trial in Ethiopia

by

Dewan Abdullah Al Rafi

Major Professors:	Ellen McCullough Gregory Colson
Committee:	Susana Ferreira

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2025

Dedication

To My Parents,
Dewan Abdul Mazid and Rebeka Sultana
Whose love and sacrifice have brought me here.

Acknowledgments

I would like to express my heartfelt gratitude to my beloved partner, *Radita Hossain* and my sister, *Sanzida Taurin*. Thank you for your unwavering love and support.

I am deeply grateful to *Dr. Ellen McCullough* for her invaluable guidance, support, and patience throughout this journey. Her mentorship has not only enriched my knowledge and opened up new opportunities but has also instilled in me a profound appreciation for her expertise. I am forever indebted to her for her unwavering dedication to my growth and success.

Lastly, I would like to extend my sincere thanks to my friends — especially *Imtiaj, Anika, Rezwana, Kaies, and Medha* — for their constant help and support along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data collection forms the backbone of scientific research, underpinning hypothesis testing, empirical validation, and policy formulation. The rigor of data collection methods—experiments, observations, archival research, and surveys—determines the reliability, validity, and generalizability of findings. Among these, surveys are widely employed for gathering behavioral and demographic data, especially in development contexts. In-person surveys allow richer interaction and clarification but are resource-intensive; phone surveys offer cost-effectiveness and broader geographic reach but suffer from lower response rates, shorter interviews, and potential measurement bias (Dillman et al., 2016; Lavrakas, 2008).

The growing use of phone surveys in low-income countries, including Ethiopia—where mobile penetration remains lower than in regional peers like Kenya and Malawi (Statista, 2025) has introduced new challenges related to representativeness, attrition, and data quality. Notably, survey mode can affect how respondents report outcomes, raising concerns over accuracy and comparability. Due to limited engagement and contextual distractions, phone surveys often elicit higher rates of satisficing and fatigue, particularly among less educated respondents. (Anderson et al., 2023; Holbrook et al., 2003; O'Leary et al., 2024; Pariyo et al., 2019). These issues are further exacerbated by contextual factors like hunger or fasting (Abate et al., 2022; Orquin & Kurzban, 2015). Sector-specific studies confirm survey mode effects across domains such as labor, health, and sensitive topics like partner intimacy (Abate et al., 2022; Arthi et al., 2018; Beland & St-Pierre, 2007; Gaddis et al., 2019; Kelly et al., 2013), though some indicators (e.g., food security) show minimal variation depending on trust and surveyor credibility (Nord & Hopwood, 2007). However, no study has systematically examined these mode effects in the WaSH (Water, Sanitation, and Hygiene) sector.

Despite extensive research on water, sanitation, and hygiene (WaSH) in multiple dimensions, a critical gap remains in understanding how different data collection methods influence the measurement of WaSH outcomes. This gap is particularly important because accurate measurement is essential for designing effective interventions and tracking progress toward public health goals. Inaccurate data especially on self-reported behaviors such as hand washing frequency or toilet usage can lead to misinformed policy decisions and misallocation of resources. Given that improved WaSH conditions are directly linked to reductions in waterborne diseases and child mortality, ensuring the reliability of survey data is not merely a methodological concern but a public health imperative. Understanding how data collection methods shape what we observe is thus crucial to getting the diagnosis right and saving lives through more effective WaSH interventions. The challenge is compounded by the nature of WaSH assessments, which rely both on observable infrastructure (e.g., water sources, sanitation facilities) and on self-reported behaviors prone to recall and social desirability biases. Understanding how different modes of data collection affect these measurements is key to improving the accuracy and impact of WaSH monitoring and interventions.

Social desirability bias is especially concerning in phone surveys, where lack of personal accountability and increased respondent anonymity may inflate positive responses (Beland & St-Pierre, 2007; Contzen et al., 2015; Holbrook et al., 2003). Demographic factors such as gender, education, and socio-economic status moderate this bias (de Leeuw, 2005; Huhtanen et al., 2015). Without appropriate mitigation strategies—like interviewer training, indirect questioning, or adjusting survey mode—data quality may be compromised, misleading policymakers (Krumpal, 2013).

While some studies explore mode effects in other domains, none employ experimental methods to assess how survey mode shapes self-reported WaSH outcomes. Given the behavioral complexity and social framing of WaSH practices, understanding how mode influences responses is crucial for improving data reliability and quality.

Our study examines the effect of survey mode on self-reported WaSH outcomes through a randomized controlled trial in Ethiopia's North Wollo zone. Comparing phone and in-person surveys among participants in graduation-from-poverty programs, we assess mode effects on reported access to improved water sources, sanitation use, and handwashing practices, controlling for confounders. The study makes two key contributions. First, it provides the first experimental evidence on how survey mode affects reporting in the WaSH sector, ad-

dressing a major gap in development data quality. Second, it identifies social desirability bias as a core mechanism driving inflated reporting in phone surveys—extending its application from sensitive domains to everyday hygiene behaviors.

This research has significant real-world implications. By showing how phone surveys can systematically inflate WaSH indicators, it raises concerns about overestimated progress toward Sustainable Development Goal 6[1]. Misleading data may lead to premature policy conclusions, misallocated resources, and underserved populations. Our findings underscore the need for improved survey design and methodological rigor to ensure accurate tracking of development outcomes and better-targeted interventions.

[1] *Sustainable Development Goal 6 (SDG 6), or "Clean Water and Sanitation," aims to ensure the availability and sustainable management of water and sanitation for all by 2030, encompassing access to safe drinking water, sanitation, and hygiene, as well as water quality and ecosystem protection*

Chapter 2

Methodology

2.0.1 Study Area

Our study examines a population of low-income women in Ethiopia who are enrolled in multi-faceted graduation-from-poverty programs. We focus on the North Wollo zone, where our two partner NGOs - CARE Ethiopia and World Vision Ethiopia—overlap in program implementation. Specifically, we study participants from CARE Ethiopia's Livelihoods for Resilience (LfR) program and World Vision Ethiopia's Strengthen PSNP4 Institutions and Resilience (SPIR) program, both of which are implemented in collaboration with the Ethiopian NGO Organization for Rehabilitation and Development in Amhara (ORDA). Our sample is drawn randomly from eligible households that opted into this safety net program, rather than from the broader population of eligible households in the region. Using program enrollment records, we randomly select VESA[2] groups from the Meket and Wadla woredas of North Wollo[3].

[2] *VESA stands for Village Economic and Social Association. These are informal associations of targeted Productive Safety Net Program (PSNP) clients established for internal savings and lending, and accessing services from the Livelihoods for Resilience Activity team.*

[3] *Meket and Wadla are woredas, or districts, located in North Wollo Zone of the Amhara Region in northern Ethiopia. These are the third-level administrative division in Ethiopia, after regions (Amhara) and zones (North Wollo).*

2.0.2 Survey Mode and Data Collection

We recruited women from each household within the selected VESA, regardless of her participation status in the LfR group. Once recruited, participants were randomly assigned to one of the survye mode arms described below. Each survey mode arm involved a distinct set of pre-tested survey questions and a predetermined data collection method. To assess the impact of survey mode on responses, we employed both traditional in-person interviews and phone surveys. Data collection for both methods took place within a 7-day period following the baseline interviews in each VESA. Additionally, each household was randomly assigned a specific interview time slot for a validation visit within the 7 days. We have provided mobile phones and charging support to the respon-

dents who were assigned to phone call data collection. Below are the description about the two survey groups:

- Phone Survey Group: In our study, the phone survey group comprises respondents who were provided with a mobile phone connection and received daily calls—one per day—focused on household WaSH-related questions, totaling seven calls over the survey period. In addition to WaSH content, these respondents also participated in phone-based modules on diet and financial behavior, although the later fall outside the scope of the present analysis. All WaSH-related questions used in this study are described in detail in Section 2.0.4. The phone survey group represents one-third of the total study sample. The overall project design included two phone survey arms —phone time use and phone diet, as well as one true control group that was surveyed entirely in person. Each group was assigned an equal number of respondents. For the purposes of our analysis, we designate the phone diet group as the "phone survey group", as it is the group from which all WaSH data were collected via phone calls. The remaining two-thirds of the sample, including the true control and phone time use groups, are collectively treated as the in-person survey group.

- In-Person Group: WaSH-related information for the remaining two-thirds of the sample was collected through in-person visits. This segment of the sample is referred to as the "in-person survey group". Notably, half of this group had been provided with mobile phones as part of a separate component of the broader project; however, those phone connections were not used for WaSH data collection and are therefore irrelevant to our analysis. For the purpose of this study, we aggregate two subgroups —the true control group and the phone time use group into a single In-person survey group, since both were administered WaSH questions during In-person validation visits.

Validation visit during our data collection process is an important part. We have used two different data integration process where validation visit date works as the reference day. In our main data set we have single observation for each of our In-person and ture control group respondents which we have collected through traditional in-person interview method. But we have 7 phone call observations and one validation visit which falls within the phone interview date range. To compare the phone call data with in-person data we generate a single response by collapsing the 7 phone calls to one. In this study we have done this using (1) the validation day phone call, which is basically selecting the phone call that

occurred on the randomly assigned validation visit day and (2) averaging all seven responses for daily calls into one single observation.

2.0.3 Experimental Design

In this sub-section we are going to explain the entire experimental design in simple words. The questionnaire employed in this study is essential to the experimental design. Both sample groups were administered the same set of questions, with the only variation being the mode of data collection. In Figure 2.1 we have outlined the experimental design in one single framework. The process began with a baseline survey at **Day -1**, conducted across the entire sample, which primarily focused on household-level information. These included household size, family characteristics, decision-making processes within the household, as well as agricultural-related information such as the number of plots owned by the household, parcel size, and the farm decision-making process. Additionally, GPS coordinates were recorded for each household to enable spatial analysis. Following the baseline survey, a mobile phone connection was provided to the phone survey group, and daily phone calls were initiated over the next seven days.

In Figure 2.1, green-shaded areas represent components relevant to this study, while red-shaded areas fall outside its scope. The in-person group received mobile phones as part of a separate experiment focused on daily time use, where they were called to report their activities for "yesterday" only. These data are unrelated to the current study. However, all in-person households were randomly assigned a validation visit between **Day 1 and Day 7**[4]. During these visits, enumerators collected data on WaSH behaviors, household and individual diet, and household financial conditions, using both "usual" and "yesterday" recall periods. The use of two recall frames allows for comparison of short- and long-term behavioral patterns. This approach is a well-established method in various development tracking surveys, including the Living Standards Measurement Study (LSMS), Joint Monitoring Program (JMP) to capture a broader spectrum of behaviors.

[4] *Day 4 is used illustratively in the figure; the actual validation day was randomly assigned within the range.*

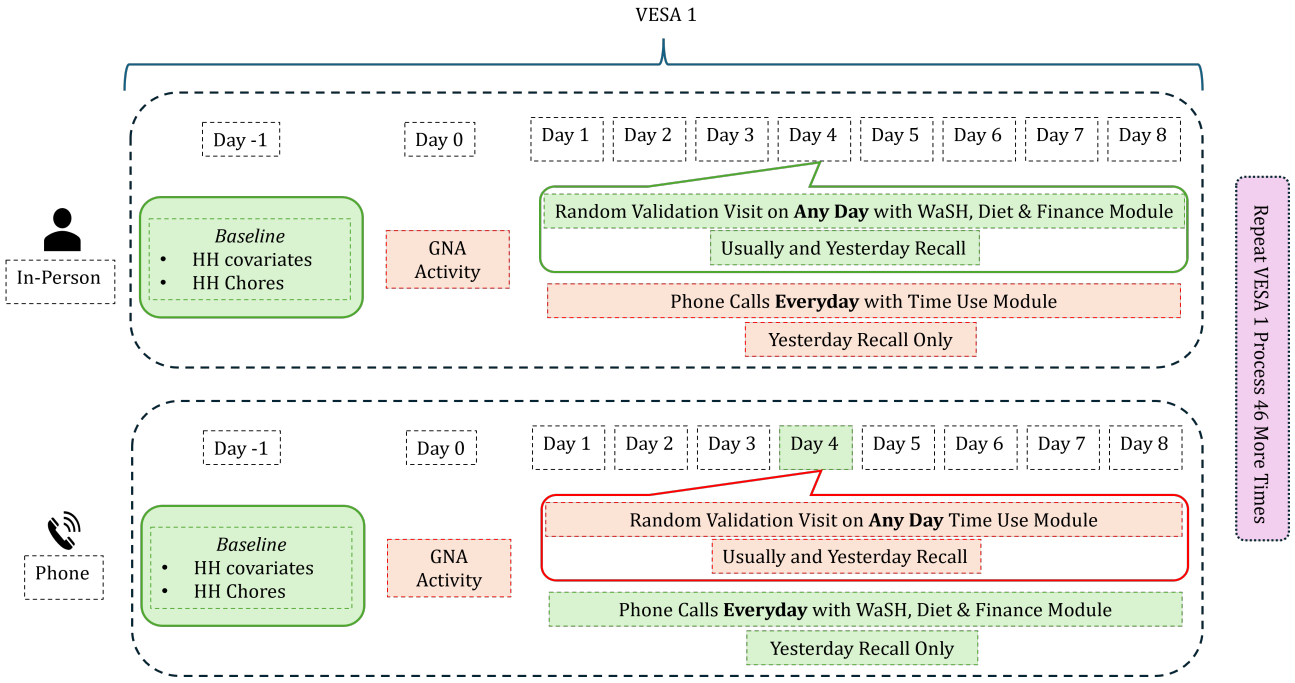

Figure 2.1: Experimental Design of Our Study

The phone survey group received mobile phones and was contacted daily to report WaSH behaviors, household diet, and financial outcomes, using only a "yesterday" recall frame. Similar to the in-person group, phone survey respondents also received a randomly assigned validation visit (illustratively shown on Day 4). Although the content of this visit is not directly relevant to our analysis, we use the visit date as a reference point for aligning data. As detailed in Section 2.0.2, we extract the phone survey data corresponding to the validation visit date for comparison—this dataset is referred to as the *validation day comparison data*. In addition, we construct a second dataset, referred to as the *7-Day Average* data, by averaging all seven daily responses into a single observation per respondent. After completing the data collection process using the above mentioned design for one VESA, we replicated the same experiment across the remaining 46 VESAs, covering all units in the study sample.

2.0.4 Questionnaire and WaSH Outcome

Tables A.2 - A.4 present the sanitation module questions included in our survey, designed to assess key aspects of water access, water treatment, storage, sanitation facility usage, hand washing behavior, and waste disposal. These questions

align with the WHO-UNICEF Joint Monitoring Programe (JMP) for Water Supply, Sanitation, and Hygiene (WaSH), which provides a standardized framework for classifying water and sanitation conditions based on their ability to reduce contamination and improve public health outcomes (WHO & UNICEF, 2005). The questionnaire captures both objective infrastructure availability (e.g., type of water source and toilet facility) and behavioral hygiene practices (e.g., hand washing and water treatment habits). These distinctions allow us to measure sanitation conditions beyond mere facility access and assess whether households actively engage in practices that mitigate health risks.

The survey first identifies the main source of drinking water (Q1), distinguishing between piped water, wells, springs, and surface water. Households may rely on multiple water sources for different purposes, which is why Q17 asks whether the same water source is used for drinking, cooking, bathing, and other household activities. Since different water sources have varying contamination risks, this distinction helps assess overall exposure to unsafe water. Additionally, we classify water sources as improved or unimproved (Q15) following the JMP definitions, where improved sources such as piped water and protected wells—offer better safeguards against contamination compared to unimproved sources like unprotected springs and surface water (WHO & UNICEF, 2005).

To assess water safety practices, respondents were asked whether they treated their drinking water the previous day (Q2) and whether they generally treat their water before consumption (Q16). The distinction between daily and habitual treatment is important, as water treatment behaviors can vary due to seasonality, economic constraints, or risk perception (Anthonj et al., 2022). Households that inconsistently treat their water may still be exposed to waterborne diseases despite having access to an improved source. We also asked about water storage practices (Q4), recognizing that even improved water sources can be contaminated if stored improperly (Shaheed et al., 2014). Households using covered plastic containers with lids are less likely to experience secondary contamination compared to those using open containers.

We assessed sanitation facility access through multiple questions to capture both stable and temporary facility usage. Respondents reported the type of toilet facility they used the previous day (Q5), while a separate variable (Q14) classifies their yesterday's sanitation facility into a binary outcome variable. The classification follows WHO-JMP definitions, where improved sanitation includes flush toilets and ventilated pit latrines, while unimproved sanitation includes open pits and shared latrines. The reason for adding separate question is, some households may rely on different facilities at different times and situations, high-

lighting the importance of considering both short-term access and long-term infrastructure availability.

Another key aspect of the sanitation module is the disposal of child feces (Q6), which is a critical indicator of sanitation practices. Unsafe disposal—such as discarding feces in open areas or rinsing them into drainage systems—can contribute to environmental contamination and increase disease transmission risks. The question differentiates between yesterday's practice vs. usual behavior, recognizing that occasional unsafe disposal may still pose significant health risks. The response categories allow us to classify feces disposal methods into safe (toilet/garbage) vs. unsafe (open disposal, drainage, or composting), helping assess the effectiveness of sanitation interventions targeted at young children.

Hand washing access and behavior were measured at multiple locations. Q7 asks whether there is a hand washing station with water and soap near the toilet, while Q9 and Q10 assess the availability of hand washing stations inside the household compound and near food preparation areas, respectively. Since access to a hand washing facility does not always imply consistent usage, Q8 specifically asks whether respondents washed their hands the previous day, providing a behavioral measure of hygiene compliance. These questions help differentiate between infrastructure availability and actual hand washing practices, which is crucial for evaluating hygiene interventions.

Following the WHO and UNICEF joint WaSH monitoring guidelines, we classified improved water sources and sanitation facilities as those that provide better safeguards against contamination, which is crucial for preventing waterborne diseases. Improved drinking water sources include options such as piped water, protected wells, and rainwater collection, all of which reduce the risk of exposure to harmful pathogens. Similarly, improved sanitation facilities are designed to manage human waste safely and include systems such as flush toilets connected to sewer systems or septic tanks, as well as ventilated improved pit latrines.

For analytical purposes, we created binary and categorical outcome variables to evaluate WaSH conditions. The classification of water sources was used to generate a binary indicator of improved (coded as 1) vs. unimproved (coded as 0) drinking water access. Sanitation facility classification was converted into an improved (coded as 1) vs. unimproved (coded as 0) sanitation variable, reflecting whether respondents have access to a toilet that hygienically separates human waste from human contact. To better understand the scenario we have classified both the yesterday and usually recall data to generate binary variables for

both yesterday and usual recall, which we have presented in our main regression analysis in Table 3.2.

2.0.5 Empirical Strategy

We analyze the effect of the survey method on each of the WaSH outcomes described in the previous parts. For each WaSH outcome, we explore differences in each WaSH outcome across the survey modes using the following model specification:

$$Y_i = \alpha + \beta T_i + \gamma_v + \omega_w + \epsilon_i \tag{2.1}$$

Here, Y_i is the dependent variable representing the WaSH outcome for each individual. T_i is an indicator for survey mode (where $T_i = 1$ for individuals in the Phone survey group meaning, data has been collected using phone calls everyday, and $T_i = 0$ otherwise which refers to the In-person survey group), while γ_v is the fixed effects which by the experimental design controls for village (i.e. VESA group) and ω_w is the survey wave indicator. In addition to binary outcomes for other WaSH indicators, we also analyze a count-dependent variable representing the frequency of hand washing. Since this outcome is a count variable, we use simple OLS to model the relationship between survey method and the frequency of hand washing. While OLS is typically used for continuous outcomes, it is often applied to count data under certain conditions, especially when the counts are not excessively large or over-dispersed (see the data distribution for the frequency of hand washing in Figure A.1 in appendix section). In this case, the assumption is that the count of hand washing events per individual (which is T_i for freq. of hand washing) can be treated as a continuous outcome for the purposes of estimation. We also have estimated logit model for our two binary outcome variables (water source type and toilet type) and Poisson regression for the count outcome variable —frequency of hand washing which is explained in Section A.2 in appendix.

To control for variations due to the study's random assignment design, we include fixed effects at the village economic and social association group level (VESA). These fixed effects help account for VESA-specific characteristics, ensuring that any effects attributed to the survey mode effect are not merely the result of inherent VESA characteristics. Additionally, since the data collection is structured in waves, VESA fixed effects also control for seasonality in responses, as interviews conducted in different seasons might yield different outcomes due to seasonal factors.

2.0.6 Hypothesis

To formalize the research question, we have tested the following hypothesis for all three outcome variables

$$H_0 : \beta = 0 \tag{2.2}$$

This null hypothesis asserts that the coefficients for the survey mode is zero. In other words, it suggests that there is no statistically significant difference between the two groups (Phone and In-person) in terms of the outcome variables under investigation.

By testing this hypothesis, we are essentially evaluating the accuracy and reliability of the data collection methods—specifically, whether there is consistency between responses collected through phone surveys and those gathered through in-person surveys. If the null hypothesis holds, it would imply that the mode of survey (phone or in-person) does not introduce bias or significantly affect the results. On the other hand, if we reject the null hypothesis, it would suggest that the method of survey delivery impacts the outcomes, indicating a potential difference in data quality or response patterns between the two survey modes. This would be critical in determining the comparability and validity of results from these different methods.

Chapter 3

Results

3.0.1 Descriptive Evidence

Figure 3.1 presents descriptive statistics for the three primary outcome variables: water source type, toilet facilities, and frequency of hand washing. These are shown across two comparison strategies: (1) the 7-Day Average dataset, which aggregates daily responses from the phone survey group, and (2) the Validation Day Comparison dataset, which aligns phone and in-person responses based on the randomly assigned validation visit day.

In terms of water source type, the proportion of households reporting access to improved water sources is consistently higher in the phone survey group across both comparison strategies. In the 7-Day Average data, 88% respondents have reported the use of improved water source over phone which is 77% for In-person group . A similar pattern holds in the Validation Day Comparison, where 87% of the phone group shows the use of improved water access. This discrepancy suggests that phone respondents may be more likely to over-report improved water usage, potentially due to social desirability biases.

A comparable pattern emerges for the use of improved toilet facilities. The phone survey group reports higher usage of improved toilets across both the 7-Day Average (80%) and Validation Day datasets(74%). In contrast, the in-person group consistently shows a lower proportion of improved toilet use, which is 45% in both data sets. This pattern underscores the possibility of inflated reporting in phone-based data collection. This divergence is further supporting the hypothesis that survey mode drives the observed differences.

The third panel shows the mean frequency of hand washing per day, measured using continuous responses. In both datasets, the phone survey group reports higher hand washing frequency compared to the in-person group. On aver-

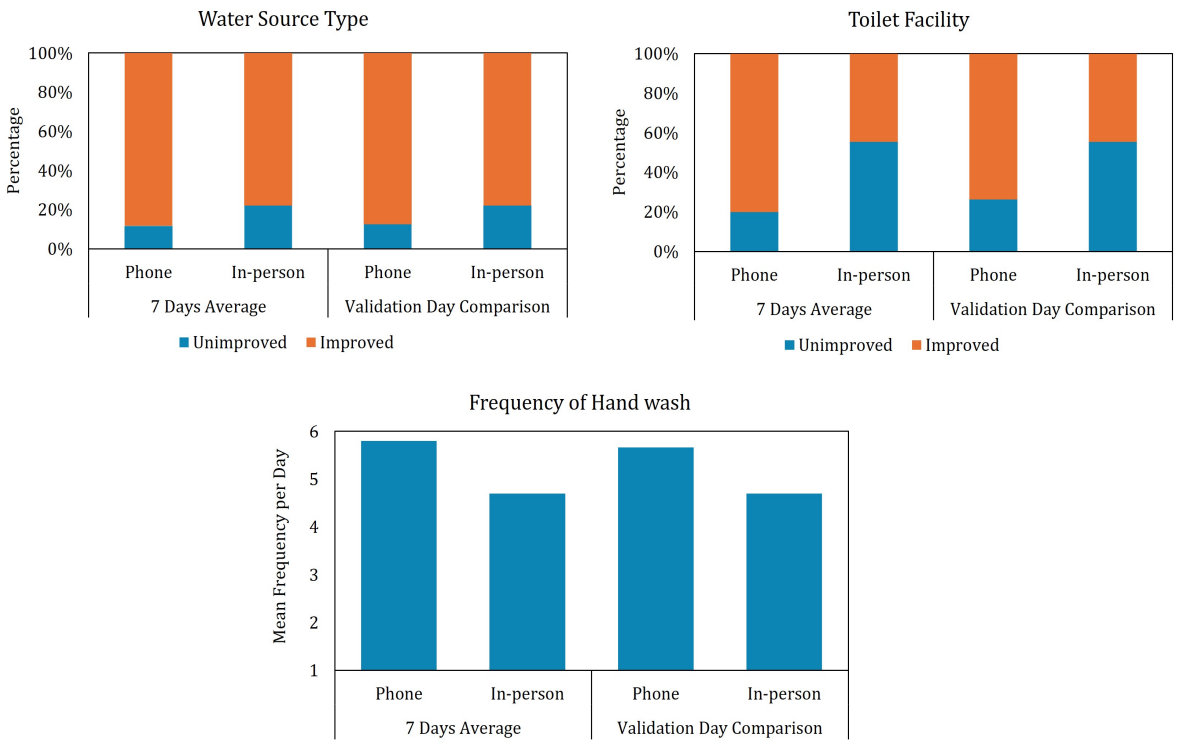

Figure 3.1: Descriptive Statistics Of WASH Outcome Variable by Survey Mode

age, the phone group reports nearly six hand washing instances per day in the 7-Day Average data, while the in-person group reports fewer than five. The Validation Day Comparison reveals a similar trend, reinforcing the concern that self-reported hygiene behaviors may be overstated in phone surveys.

3.0.2 Balance Test

Continuing with the analysis, Table 3.1 provides a balance test comparing household characteristics across the two survey modes: phone and in-person. The objective of this test is to assess whether the two groups are comparable in terms of key demographic and socioeconomic variables, ensuring that any observed differences in outcomes are not driven by pre-existing imbalances. The results demonstrate no significant differences between the two survey modes across the variables tested. For instance, the average age of respondents is 38.48 years in the phone survey group and 37.99 years in the in-person group, with a mean difference of 0.485 years and a p-value of 0.526. Similarly, the age of household

heads is consistent between the two groups, with a negligible mean difference of -0.188 years (p = 0.824).

Table 3.1: Household Characteristics by Survey Mode

Variable	Mean		Difference	t-test
	Phone	In-person		p-value
Age of Respondent	38.080	38.043	0.037	0.964
Age of Household Head	44.971	45.559	-0.588	0.527
Respondent's Education (year)	1.735	1.790	-0.055	0.113
Household Head's Education (year)	5.169	4.676	0.494	0.196
Household Decision Making	1.802	2.551	-0.748	0.151
Agricultural Land Size (ha)	0.296	0.304	-0.007	0.825
Number of HHs	211	434		
Number of Clusters	47	47		

Note: The null hypothesis for the two tailed mean difference test is $H_0 : \mu_{treatment} - \mu_{control} = 0$. *These statistics are from baseline survey data.*

Educational attainment, both for respondents and household heads, also shows minimal variation. Respondents in the phone group report an average of 2.618 years of education compared to 2.485 years in the in-person group (mean difference = 0.133, p = 0.224). Household heads show a similar trend, with 2.325 years of education in the phone group and 2.171 years in the in-person group (mean difference = 0.153, p = 0.150).

The sample size consists of 211 observations for the phone survey and 434 for the in-person survey, distributed across 47 clusters. The absence of statistically significant differences across all variables indicates that the two groups are well-balanced. This comparability reduces the likelihood of confounding effects introduced by differences in household characteristics between survey modes, ensuring the validity of subsequent analyses and strengthening the robustness of the study's findings.

3.0.3 Empirical Evidence

Impact of Survey Mode

Using the specified Ordinary Least Squares (OLS) estimation model described in Equation 2.1, we examine the impact of the use of different survey mode on measuring three key Water, Sanitation, and Hygiene (WaSH) outcomes: water source type, toilet type, and frequency of hand washing. Table 3.2 presents the regression results using the Validation Day Comparison dataset.

We found that, the Phone survey group demonstrated a significant positive effect on water source type. Respondents in this group are 9.3 percentage points ("Yesterday" recall) more likely ($p < 0.05$) to report using improved water sources compared to the In-person group. The effect of collecting data via Phone survey on toilet facilities is even more substantial. In the "Yesterday" recall model of toilet type, respondents in the Phone group are 28.3 percentage points more likely ($p < 0.001$) to report access to improved sanitation facilities compared to the In-person group. The most pronounced impact is observed in the frequency of hand washing. The Phone survey group reports a 0.96 times higher ($p < 0.001$) frequency of hand washing everyday compared to the In-person group.

Table 3.2: Effect of Phone Survey Mode Treatment on WaSH Outcome Variables

	Water Source Type		Toilet Type		Freq. of Hand Washing
	Model 1	Model 2	Model 1	Model 2	Model 1
Phone Group	0.079*	0.093*	0.281***	0.283***	0.961***
	[0.01,0.15]	[0.01,0.17]	[0.19,0.37]	[0.19,0.37]	[0.51,1.41]
N	617	617	617	617	614
No. of Cluster	47	47	47	47	47
IP Mean	0.793	0.779	0.449	0.447	4.692
Recall	Usually	Yesterday	Usually	Yesterday	Yesterday

Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$). *This table displays regression results from simple linear regression, specified in equation 2.1. Values in parenthesis represents 95% CI. All the models include VESA level clustering of standard error and constant term. Since we do not have Usual recall question for freq. of hand washing, so we are presenting only the yesterday recall coefficient. IP Mean represents the mean of the WaSH variables for In-person group . Results from 7 Days Average data is in the Appendix section.*

In this regression analysis, we use both “Usually" recall and “Yesterday" recall to compare the difference between more stable, long-term habitual behavior and short-term, potentially variable behavior. The result of both Table 3.2 and Table A.1 helps us to conclude that, we can reach the same conclusion regardless of how we construct the outcome variable and what ever recall period we consider. The results of the OLS estimates in this section appear to be meaningful and align with expectations, suggesting that the phone survey group reports significantly higher positive outcomes in WaSH indicators. However, several factors could explain these findings. One possibility is that respondents in the phone survey group are accurately reporting their true behavior, reflecting genuine improvements in WaSH outcomes.

On the contrary, a potential external factor influencing the results is verifiability, which refers to whether respondents perceive their answers as being subject to external validation. If respondents believe that their responses can be cross-checked or verified by an external party, they may be more likely to report their behavior accurately. Conversely, if they believe their responses are not verifiable, they may be more inclined to provide socially desirable but potentially inaccurate answers. In this study, verifiability could differ between the In-person and Phone survey groups due to differences in data collection methods. For instance, in the In-person survey group, respondents may have been more aware that their responses could be corroborated with face to face observations, leading to more truthful reporting. In contrast, the Phone survey group, may have felt less accountable for the accuracy of their responses, potentially leading to over-reporting of positive behaviors.

Although verifiability could play a role in explaining differences in responses between the two groups, we are unable to directly analyze its effect due to data limitations. The current dataset does not include information on whether respondents perceived their responses as verifiable, nor do we have external measures to objectively validate their reported behaviors. As a result, while verifiability remains a plausible factor influencing the differences in reported WaSH outcomes, its precise impact cannot be isolated in this study, which remains as the limitation.

Robustness to Enumerator Effect

Figure 3.2 presents a series of coefficient plots illustrating the robustness of our main results to enumerator-specific effects. Each panel corresponds to one of the three WaSH outcome variables: access to improved water sources (top left), access to improved toilet facilities (top right), and frequency of hand washing

(bottom). The blue dot labeled "All" indicates the estimated survey mode effect —i.e., the effect of phone surveys relative to in-person surveys—using the full sample. The remaining points represent estimates from a leave-one-out approach, where we systematically exclude each enumerator (labeled 1 to 11) from the sample and re-estimate the survey mode effect using the same regression specification described in Equation 2.1.

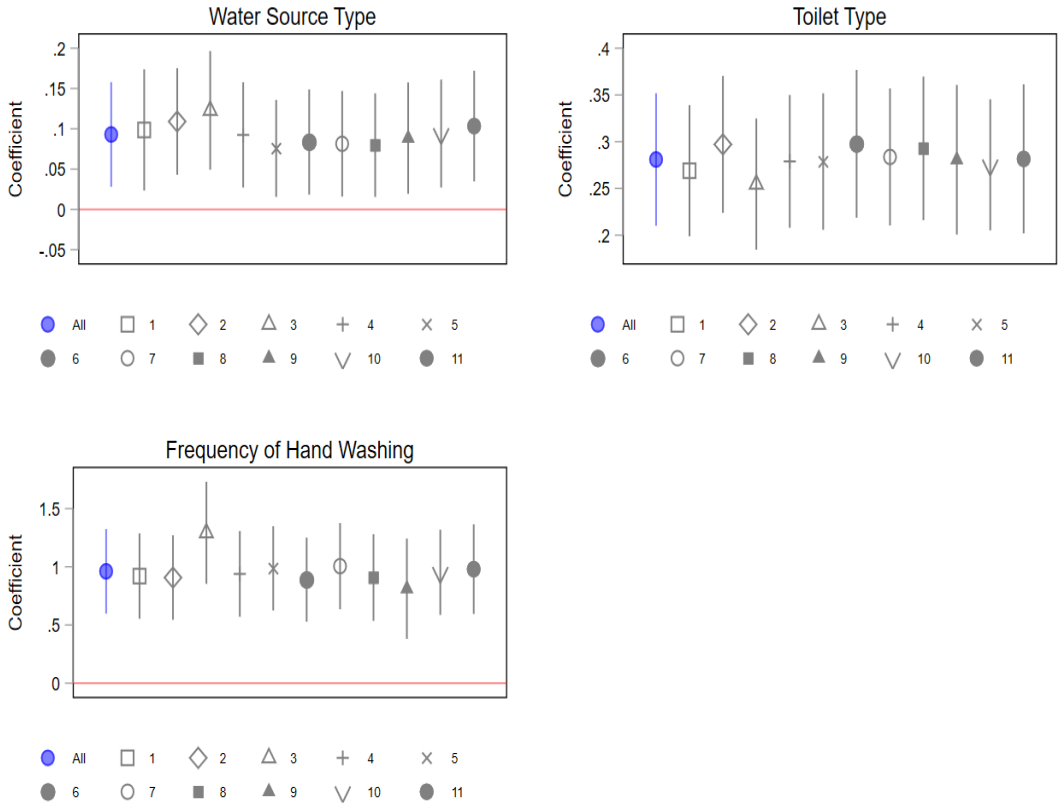

Figure 3.2: Robustness of the Impact of Survey Mode to Enumerator Effect

Note: We have used all the WaSH outcome variables as the dependent variable where we regress that with the same specification of equation 1. Here each omitted enumerator is presented by each marker number 1 to 11. One enumerator is omitted due to no validation visit on his name. Additionally, ***All*** *means, we have used all the observations to regress WaSH outcomes with respect to survey modes. The coefficient on the y-axis represents the coefficient for the phone call group from each regression where the reference group is in-person data group.*

Across all outcome variables, the survey mode effect remains stable and statistically consistent when any single enumerator is dropped from the analysis. The coefficients exhibit minimal variation, and confidence intervals overlap substan-

tially with the full-sample estimate. This pattern suggests that the observed differences between phone and in-person survey responses are not driven by any specific enumerator's behavior, style, or influence on respondent reporting.

There could be another potential factor that might have some impact of the estimated survey model impact that we have found in Section 3.0.3. This factor is referred as the social desirability bias, where respondents overstate positive behaviors or outcomes to align with perceived social norms or expectations. This bias could artificially inflate the estimated survey mode effects by introducing a systematic upward shift in reported behaviors, making the phone survey group appear to have better WaSH conditions than they actually do. To address this concern, we conduct further analysis in a subsequent section to examine the presence and extent of social desirability bias in the responses.

3.0.4 Social Desirability Effect

Previous studies related to measuring the impact of different survey modes have identified a new dimension in measurement-related study, called social desirability effect, which refers to the intentional misreport by the respondent during the interview. Some studies have argued that respondents are more willing to report the truth when there is a chance to access the information by the enumerator in a different way or there is an assurance of high confidentiality (Evans et al., 1977; Himmelfarb & Lickteig, 1982). Holbrook et al., 2003 argued that due to the lack of rapport and interpersonal trust between enumerator and respondent in phone survey, leads to a bias in the responses. In our study we hypothesized that there is no social desirability in different survey mode.

We have calculated a composite score of social desirability based on 7 question related to WaSH outcomes. These questions are not directly the outcomes that we have used in our analysis but are directional to either a socially desired response or not for WaSH related outcomes, which we have presented in Table 3.3. We have coded all the socially desired responses as 1 and not the desired answer as 0 and summed everything up to get an index. A higher social desirability index value indicates a higher presence of desirability in the response. Further, we have used the desirability index as a proxy of social desirability.

In Figure 3.3 we are representing the comparative performance of the Phone and In-person survey groups across multiple WaSH-related indicators included in the social desirability score. Each axis corresponds to a specific WaSH-related

Table 3.3: Social Desirability Score Calculation

Questions and Categories	*Response classification*	
	1 = Desired	***0 = Not Desired***
1. Is there any hand washing station (water only or water and soap/soapy matertoiletial/soapy water) inside the house compound ?	Yes	No, Not Applicable
2. Is there any hand washing station maintained at the place of food preparation or child feeding area?	Yes	No, Not Applicable
3. Did you treat your water in any way to make it good to drink YESTERDAY?	Yes	No, Not Applicable
4. Is there any hand washing station maintained at the place of food preparation?	Yes	No, Not Applicable
5. YESTERDAY, what kind of container did you use to store water	Container with a plastic handle and a lid that can be used to close the top	Clay container, plastic container, aluminium container, Not Applicable
6. Is there any hand washing station (water only or water and soap/soapy mater toiletial/soapy water) next to the toilet ?	Yes	No, Not Applicable
7. Did you wash your hand yesterday ?	Yes	No

question outlined in Table 3.3, such as the presence of hand washing stations, water treatment, or appropriate water storage. The Phone survey group consistently scores higher across all indicators, as evidenced by the larger area enclosed by the solid blue line compared to the dashed orange line representing the In-person survey group. This suggests that respondents in the Phone survey group are more likely to provide socially desirable answers. The observed differences align with the hypothesis that social desirability bias may influence the response of Phone survey group, as they display a stronger tendency toward reporting behaviors or conditions perceived as favorable. This highlights the potential role of social desirability effect in shaping the outcomes, which is further explored using the composite desirability index in the analysis. From the mean differ-

Figure 3.3: Comparison of WaSH-Related Social Desirability Indicators Between Phone and In-person Survey Groups

ence test we found a significant difference in social desirability score between the Phone and In-person survey group (see Figure 3.4).

The box plot illustrates the distribution of Social Desirability Scores across Phone and In-person groups differentiated by survey method and data integration type. The In-person (7 days average) group consistently shows a lower median score and a narrower inter-quartile range, suggesting respondents in face-to-face surveys were less likely to provide socially desirable answers. In contrast, both Phone survey groups — whether aggregated over a 7-day average or using validation-day comparison data — show higher medians and wider distributions, indicating a greater tendency toward socially desirable responses in phone-based interviews.

Interestingly, the two phone survey groups are nearly identical in distribution, suggesting that the method of data integration (7-day average vs. validation-day comparison) has no effect on the observed pattern of social desirability score. This reinforces that survey mode, rather than how the data are aggregated, is likely the primary driver of differences in social desirability scores.

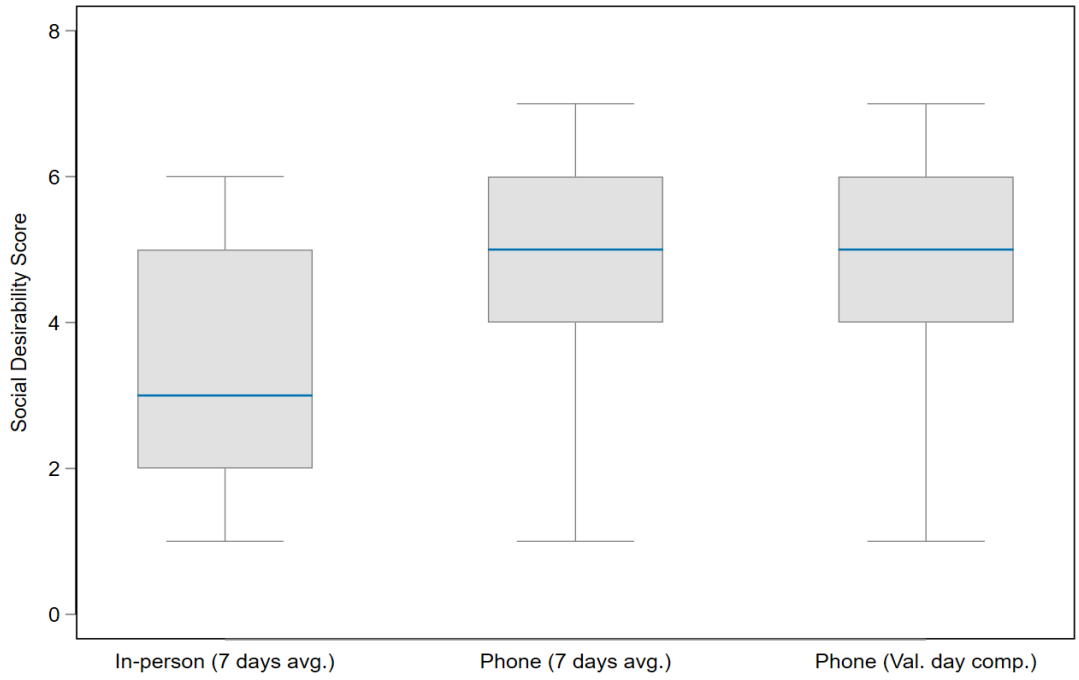

Figure 3.4: Inspecting Social Desirability Score

Note: The in-person survey group is identical across both data sets, so it is displayed only once in the box plot. The primary purpose of presenting the distribution of social desirability scores is to highlight differences across survey methods and data integration types.

One possible explanation for this pattern is the role of social presence and perceived accountability in shaping response behavior. In phone interviews, respondents may experience a greater sense of anonymity and reduced social pressure, making it easier for them to modify their responses to appear more favorable. In contrast, face-to-face interviews create a stronger social presence, where non-verbal cues and direct interaction with an interviewer may encourage more honest reporting. This aligns with Holbrook et al., 2003, who found that social desirability bias is heightened in phone surveys due to reduced social monitoring. However, this differs from Heerwegh, 2009, who suggests that face-to-face interactions could also induce social desirability effects due to the interviewer's immediate presence.

Enumerator Effect

A team of 12 enumerators conducted the survey after undergoing identical training and working under a single survey coordinator. To streamline logistics, we assigned each enumerator exclusively to either phone or in-person interviews. This fixed assignment raised concerns that the estimated survey mode effects—such as those shown in Table 3.2, might reflect enumerator-specific influences rather than true mode effects. To test this possibility, we performed a robustness check. We subset the dataset by enumerator and ran separate regressions, each using the social desirability score as the dependent variable and survey mode as the independent variable. To account for unobserved heterogeneity, we included VESA-level fixed effects. Following the method used by Abate et al., 2022, we aimed to determine whether any single enumerator disproportionately influenced the estimated survey mode effect.

Figure 3.5 shows the results of this analysis. The blue marker represents the survey mode effect estimated from the full dataset, while the other markers show estimates from regressions that exclude one enumerator at a time. All coefficients remain positive and statistically significant, and they closely align with the full-sample estimate. This consistency suggests that no individual enumerator drives the survey mode effect.

These results indicate that the phone survey group consistently reports more socially desirable responses than the in-person group. The robustness of these findings supports the conclusion that survey mode—rather than enumerator characteristics—drives the observed differences in response behavior. This aligns with prior studies (e.g., Kreuter et al., 2008), which emphasize the role of structural factors like anonymity and perceived judgment in shaping social desirability bias. Our findings contribute to the survey methodology literature by highlighting the value of accounting for interviewer-related biases in experimental survey designs (Blair et al., 2020).

Enumerator Experience

While our earlier robustness checks account for enumerator identity, we need to examine how enumerator experience over time might influence responses. As Abate et al., 2022 notes, enumerator experience can bias social desirability measurements and potentially affect survey mode effect. We conducted a formal non-linear analysis to identify such patterns in our data. Without direct information on enumerators' prior fieldwork experience, we use the order of survey administration as a proxy for experience—capturing how interviewer behavior might evolve during data collection. The reasoning is that as enumerators be-

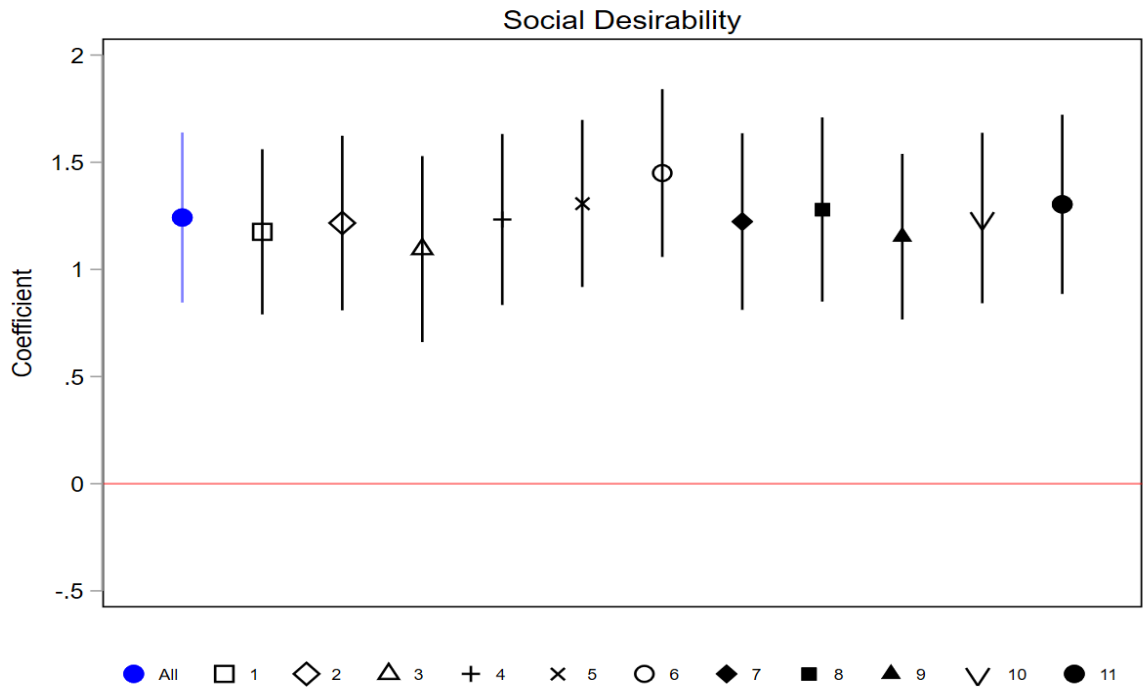

Figure 3.5: Robustness of Social Desirability Mechanism to Individual Enumerator Effect

Note: We have used social desirability score as the dependent variable where we regress that with the same specification of equation 2.1. Here each omitted enumerator is presented by each marker number 1 to 11. One enumerator is omitted due to no validation visit on his name. Additionally, all means, we have used all the observations to regress social desirability with respect to survey modes. The coefficient on the y-axis represents the coefficient for the survey mode from each regression.

come more familiar with the survey instrument and gain confidence, they may unconsciously adjust their question delivery or respondent interactions. These subtle changes can affect how respondents perceive the interview, potentially influencing their tendency to provide socially desirable responses, especially on sensitive or value-laden topics.

Figure 3.6 presents a quadratic fit plot illustrating the non-linear relationship between social desirability scores and survey number, which serves as a proxy for enumerator experience, separately for the Phone survey and In-person survey groups. The observed patterns suggest that enumerator experience systematically influences response tendencies, but its effects vary depending on the mode of data collection.

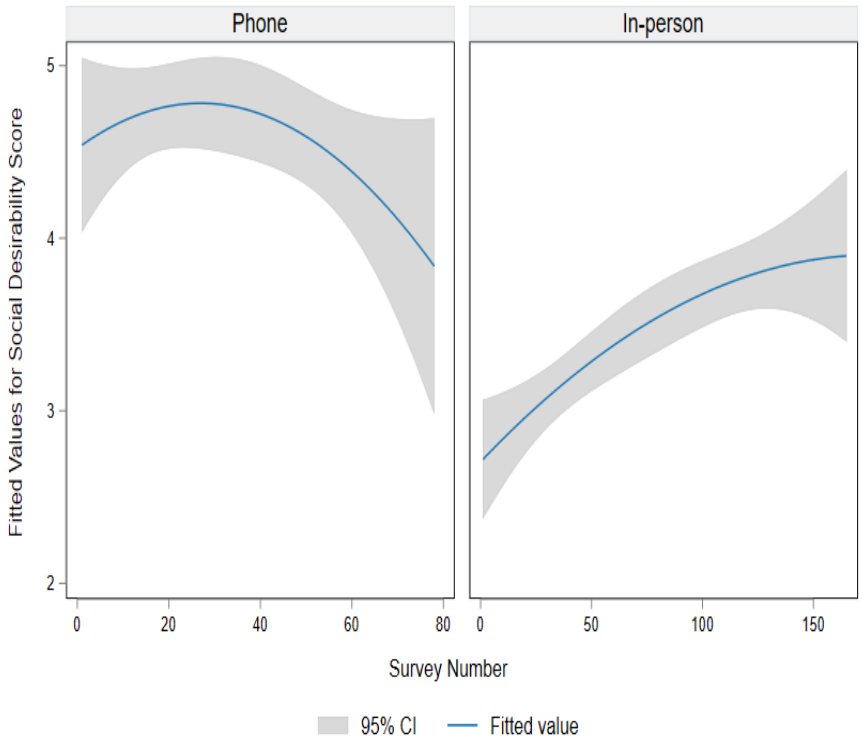

Figure 3.6: Non-Linear Relationship Between Fitted Values and Enumerator Experience

Note: In this figure, survey number means the n^{th} number of survey for each enumerator. First of all, we have sorted the entire data set by enumerator name and date of data collection. Then we assigned a serial number to each observation. Since, we do not have any other variable related to enumerators prior experience in data collection, then it is the only possible way to isolate the impact of enumerator experience. Theoretically, as the survey number increases, the enumerator is gaining more and more experience in our survey.

In the Phone survey group (left panel), social desirability scores initially increase with enumerator experience, peak at an intermediate stage, and then decline as enumerators conduct more surveys. This inverted U-shaped pattern suggests that early-stage enumerators may unintentionally yield more socially desirable responses, possibly due to limited familiarity with standardized probing techniques or inconsistencies in question delivery. However, as enumerators gain experience, they may develop more neutral interviewing techniques, leading to a reduction in social desirability bias. This pattern aligns with findings from Holbrook et al., 2003, who argue that interviewer effects can be more pronounced

in the early stages of survey administration but tend to stabilize over time as enumerators become more proficient in maintaining a neutral stance.

In contrast, the In-person group (right panel), exhibits an upward trend, where social desirability scores start at a lower level and gradually increase with enumerator experience. This suggests that early-stage In-person interactions may produce more candid responses, potentially due to a lack of established rapport between the enumerator and respondent. However, as enumerators gain experience, they may develop subtle social cues—whether through nonverbal communication, tone of voice, or familiarity—that encourage respondents to provide more socially desirable answers. This is consistent with Heerwegh, 2009, who found that interviewer effects in face-to-face settings can become more pronounced as familiarity increases, leading to a greater tendency for respondents to adjust their answers based on perceived social expectations.

Survey Fatigue - Fasting

Another potential reason for getting a socially desired response from the Phone survey group could be survey fatigue on the respondents' part due to long questionnaire or impatience during the survey (Abate et al., 2022; Anderson et al., 2023). Survey fatigue—where respondents become less engaged, more inattentive, or adjust their answering behavior due to cognitive or physical exhaustion—is a critical concern in survey-based research. One way to proxy fatigue is through fasting status, as fasting has been shown to influence cognitive performance, self-regulation, and decision-making (O'Leary et al., 2024). Research in psychology and behavioral economics suggests that hunger and calorie restriction can lead to cognitive depletion, reduced attentional control, and increased reliance on heuristic decision-making (Orquin & Kurzban, 2015). In survey settings, these effects may manifest as greater response fatigue, lower effort in answering questions, or increased reliance on socially desirable responding. Given these cognitive and behavioral effects, fasting serves as a useful proxy for respondent fatigue, allowing us to examine whether survey fatigue moderates the social desirability mechanism underlying survey mode effects.

In addition to current fasting status, we include lagged fasting, whether the respondent fasted on the day prior to the survey as a key explanatory variable. Given that our outcome variable relies on a "yesterday" recall (i.e., self-reported behavior from the previous day), lagged fasting is particularly relevant in isolating how the respondent's physical or cognitive condition on the day being reported may affect their responses. To assess whether the effect of survey mode

varies with fasting-related fatigue, we interact both current fasting and lagged fasting with the survey mode indicator.

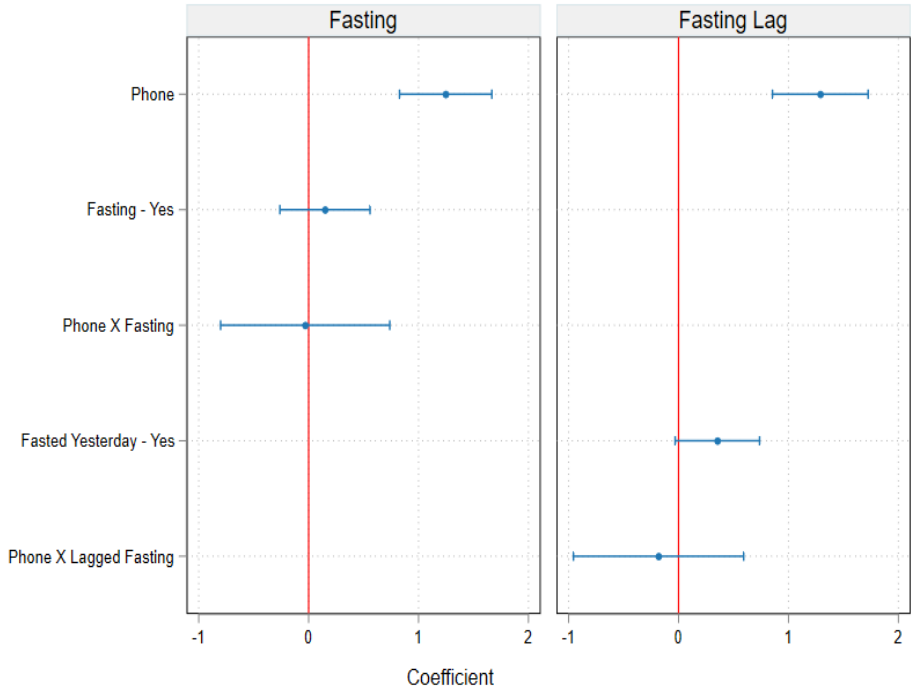

Figure 3.7: Effect of Fasting and Lagged Fasting on Social Desirability Score

In this analysis, we utilize validation day comparison data to examine the effects of fasting on survey responses. Both models incorporate wave fixed effects to account for temporal variations and cluster standard errors at the VESA level to control for unobserved heterogeneity. For the classification of lagged fasting, we define it as whether the day prior to the survey date was a fasting day. The fasting calendar for Ethiopia provides a structured framework for this classification. Specifically, the 40 days preceding May 2, 2021, correspond to Lent, during which all days are designated as fasting days. However, May 2 itself (Easter/Fasika) is not considered a fasting day, marking the beginning of a festive period. In the 40 days following May 2, no days are classified as fasting days, as this period is one of celebration. Beyond these 80 days, fasting days are systematically observed on Wednesdays and Fridays, which we use for our classification in this study.

Figure 3.7 presents coefficient plots from two regression models examining the relationship between fasting, lagged fasting, and response behavior, with a particular focus on the interaction terms (Phone × Fasting and Phone × Lagged Fasting). These interaction terms test whether fasting-induced fatigue alters the extent to which respondents engage in socially desirable responding, thereby affecting the estimated survey mode effects.

The results indicate that fasting and lagged fasting influence responses overall but do not systematically moderate the survey mode effect, suggesting that the

social desirability mechanism remains intact. In both models, the Phone survey coefficient is positive and statistically significant, meaning that respondents in the Phone survey group exhibit higher outcome values on average, independent of fasting status. The left panel shows that fasting itself has an overall effect on responses, as indicated by the positive coefficient for Fasting, which is consistent with research suggesting that hunger and cognitive depletion can shape decision-making processes. However, the Phone × Fasting interaction term is not statistically significant, implying that fasting does not systematically alter the survey mode effect. A similar pattern emerges when considering lagged fasting in the right panel. The positive coefficient for *Fasted Yesterday*, suggests that prior fasting may have residual cognitive effects on response behavior. However, the Phone × Lagged Fasting interaction term remains non-significant, indicating that any carryover effects from fasting do not differentially affect responses in the Phone survey group.

The results indicate that while fasting-induced fatigue may affect overall response behavior, it does not systematically alter survey mode effects, meaning that social desirability bias remains the key mechanism explaining differences between Phone and In-person survey groups. If cognitive depletion from fasting significantly interfered with social desirability, we would expect fasting respondents to show either a weaker or stronger tendency toward socially desirable responses. However, the non-significant interaction terms suggest that fasting does not meaningfully disrupt this mechanism. Instead, survey mode, rather than temporary cognitive strain, drives social desirability bias. These findings have important methodological implications. They suggest that fasting-induced survey fatigue does not compromise the validity of survey mode effects, reinforcing the argument that survey mode is the primary driver of social desirability bias rather than short-term cognitive depletion. This highlights the importance of survey design considerations in fasting-prone contexts, ensuring that fasting status does not introduce systematic bias in experimental estimates

Chapter 4

Conclusion

This study provides compelling evidence that survey mode significantly influences the reporting of Water, Sanitation, and Hygiene (WaSH) outcomes in development contexts. Through a randomized controlled trial in Ethiopia's North Wollo zone, we demonstrate that phone-based surveys yield systematically different response patterns compared to face-to-face interviews across multiple WaSH indicators. Our findings indicate that respondents surveyed via phone are significantly more likely to provide socially desirable responses, leading to inflated reports of improved water access, sanitation, and hygiene behaviors.

The robustness of these effects across different outcomes specifications, data integration approaches, and enumerator assignments suggests that these differences are not artifacts of methodological choices, but rather systematic variations in response behavior driven by survey mode. Our analysis of social desirability bias reveals that phone respondents exhibit a higher tendency to provide socially desirable answers, reinforcing concerns about data validity in phone-based surveys. Additionally, our investigation into enumerator effects and survey fatigue (proxied by fasting status) confirms that these biases persist independent of enumerator characteristics or temporary cognitive depletion, further strengthening the argument that social desirability bias is the primary mechanism driving survey mode effects.

These findings have important implications for both research methodology and policy implementation in low-resource settings. First, while phone-based surveys offer cost and logistical advantages, they may introduce systematic biases in WaSH-related data collection, affecting the reliability of policy-relevant insights. Given the increasing reliance on phone surveys for development research, program monitoring, and impact evaluations, researchers must carefully consider potential distortions in self-reported behavioral data. Second, the substantial

differences observed between survey modes suggest that data collected through different methods may not be directly comparable, particularly in longitudinal studies or cross-sectional analyses where survey mode transitions occur. Adjustments for social desirability bias—such as mode calibration techniques, indirect questioning, or survey design modifications—are necessary to improve data accuracy and comparability.

Beyond WaSH outcomes, this study contributes to the broader survey methodology literature by demonstrating that survey mode effects extend beyond previously documented domains (e.g., economic measures, health indicators) and are particularly pronounced in self-reported behavioral outcomes. Future research should explore alternative strategies to mitigate social desirability bias in phone surveys and examine whether similar response patterns emerge in different geographical regions and development contexts. Additionally, a deeper investigation into the psychological and social mechanisms underlying survey mode effects could yield valuable insights for improving remote data collection methods.

In conclusion, while phone surveys provide valuable flexibility in data collection, their impact on data quality and response patterns cannot be ignored. As remote survey methodologies become increasingly prevalent, it is essential for researchers and policymakers to recognize, account for, and mitigate survey mode biases to ensure the validity and reliability of research findings and subsequent policy recommendations.

Appendix A

A.1 Appendix - A

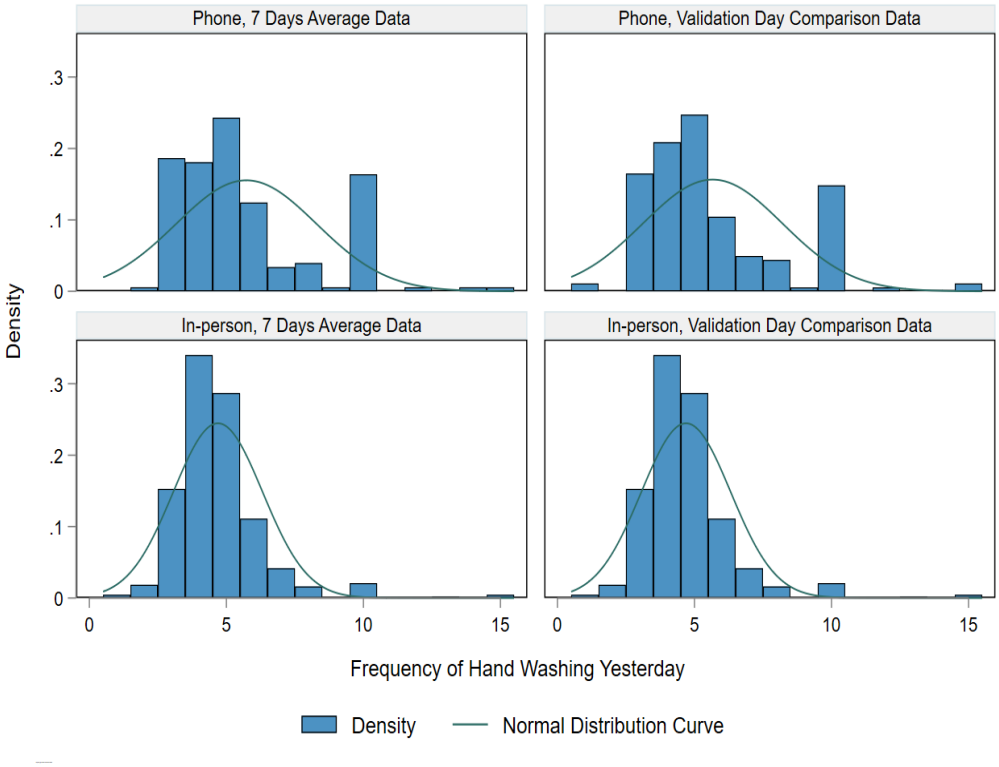

Figure A.1: Frequency Distribution of Hand Washing Events/day

Table A.1: Effect of Phone Survey Mode Treatment on WaSH Outcome Variables

	Water Source Type		Toilet Type		Freq. of Hand Washing
	Model 1	Model 2	Model 1	Model 2	Model 1
Phone Group	0.089*	0.102*	0.348***	0.350***	1.015***
	[0.01,0.17]	[0.02,0.18]	[0.27,0.43]	[0.27,0.43]	[0.57,1.46]
N	640	640	641	641	609
No. of Cluster	47	47	47	47	47
IP Mean	0.793	0.779	0.449	0.447	4.692
Recall	Usually	Yesterday	Usually	Yesterday	Yesterday

Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$). *IP Mean represents the mean for In-person group.This table displays regression results from simple linear regression, specified in equation 2.1. In this table we have used 7-Days average data. Values in parenthesis represents 95% CI. All the models include VESA level clustering of standard error and constant term. Since we do not have Usual recall question for freq. of hand washing, so we are presenting only the yesterday recall coefficient.*

Table A.2: **Survey Questionnaire and Responses**

Questions and Categories	*Phone*			*In-Person*		
1. What was your main source of drinking water?	***yesterday***			***yesterday***	***rainy***	***dry***
Piped into dwelling	0.81			0.20	0.60	0.60
Piped into compound/yard	8.80			59.48	59.48	60.48
Public tap outside compound	55.94			59.48	59.48	60.48
Protected/covered well	1.84			7.58	6.99	6.99
Protected spring	12.63			5.19	4.79	4.79
Open/unprotected spring	4.91			14.77	13.17	12.77
River/lake/pond/stream/dam	6.75			6.19	6.59	6.19
Others	0.22					
Not Applicable	8.10					
	n/a	***yes***	***no***	***n/a***	***yes***	***no***
2. Did you treat your water in any way to make it good to drink YESTERDAY?	7.13	17.93	74.95	0.00	0.80	99.20
3. Did you use the same source of drinking water for all other purposes such as cooking, bathing, and washing clothes and household items YESTERDAY?	6.75	74.84	18.41	0.00	85.43	14.57
4. What kind of container did you use to store water YESTERDAY?	***percentage***			***percentage***		
Clay container	0.32			0.60		
Aluminum/metal/steel container	0.43			0.00		
Plastic container	18.74			21.56		
Container with a plastic handle and a lid on top	75.92			77.842		
Others	0.05			0.00		
Not Applicable	4.54			0.00		
5. What kind of toilet facility did you use YESTERDAY?						
Pit latrine/traditional pit toilet	80.18			54.69		
Ventilated improved pit latrine (VIP)	0.92			0.20		
Flush toilet	0.05			0.00		
Communal/shared latrine	0.49			0.00		
Pan/bucket	0.00			9.58		
No facility/bush/field	18.36			35.53		
6. What did you do to dispose of young child's stools	***yesterday***			***yesterday***	***usually***	
Drop in the toilet/garbage can	40.17			18.09	10.78	
Rinse/wash away in open area	14.96			3.41	2.09	

Table A.3: **Survey Questionnaire and Responses**

Questions and Categories	*Phone*			*In-Person*		
7. Is there any hand washing station (water only or water and soap/soapy material/soapy water) next to the toilet?	***percentage***			***percentage***		
Not Applicable	0.40			6.39		
Refuse to response	0.34			0.00		
Yes	76.06			13.17		
No	23.20			80.44		
8. Did you wash your hand YESTERDAY?						
Yes	100.00			99.6		
No	0.00			0.40		
	yes	***no***	***n/a***	***yes***	***no***	***n/a***
9. Is there any hand washing station in the house/the compound?	48.42	51.31	0.27	54.29	45.71	0.20
10. Is there any hand washing station maintained at the place of food preparation?	78.02	21.98	0.00	52.10	47.90	0.00
11. Did you have a hand washing station maintained at the place of food preparation for the child or child feeding area YESTERDAY?	91.79	6.80	1.40	49.50	48.30	2.20
12. Where did you prepare food for the youngest child in this household YESTERDAY?	***percentage***			***percentage***		
Inside the house	32.29			53.49		
On the veranda/outside the house	64.20			0.60		
Others	0.16			0.20		
Food is not mashed for youngest child	1.62			28.34		
Not Applicable	1.62			16.97		
Don't Know	0.11			0.00		
Refuse to respond	0.00			0.40		
13. Where did you feed the child YESTERDAY?						
Inside the house	94.17			66.27		
On the veranda/outside the house	2.54			0.00		
Others	0.05			0.00		
Food is not mashed for youngest child	1.51			16.37		
Not Applicable	1.67			17.37		
Don't Know	0.05			0.05		

Table A.4: **Survey Questionnaire and Responses**

Questions and Categories	*Phone*		*In-Person*	
	Improved	***Unimproved***	***Improved***	***Unimproved***
14. Type of toilet facility *(Calculated)*	81.16	18.84	36.81	63.19
15. Water source type *(Calculated)*	88.12	11.88	53.82	46.18
16. Do you treat your water in any way to make it good to drink?	***yesterday***		***yesterday***	***usually***
Yes	18.55		0.92	1.38
No	74.33		99.08	98.62
Not Applicable	7.13		0.00	0.00
17. Do you use the same source of drinking water for all other purposes such as cooking, bathing, and washing clothes and household items?				
Yes	74.12		72.81	0.00
No	19.10		27.19	0.00
Not Applicable	6.78		0.00	0.00

A.2 Appendix - B

We also have used the Logistic regression for our binary outcome variable to identify the survey mode effect. Here we have followed the regression specification below:

$$Pr(Y_i = 1) = \lambda(\beta X_i + \gamma_v + \omega_w) \tag{A.1}$$

Besides, due to the count nature of the frequency of hand washing variable, we have also used the Poisson regression with the following specification:

$$E[Y_i|X_i, \gamma_v, \omega_w] = exp(\beta X_i + \gamma_v + \omega_w) \tag{A.2}$$

In both equation above, Y_i is the dependent variable, X_i is the survey mode, where in-person survey is the reference group; γ_v and ω_w is the VESA and survey wave level fixed effect to control for unobserved variation and β is the coefficient of interest. Additionally, λ is the logistic cumulative distribution function[5], and exp is the exponential function ensuring the expected value is non-negative.

[5] We can denote the logistic cumulative distribution function as

$$\lambda(z) = \frac{1}{1 + e^{-z}}$$

For both Logistic and Poisson regression we hae used the marginal effect to identify the impact of survey mode on WaSH outcome variables. In this case we have used the following derivative:

- For Logistic model

$$\frac{\partial \Pr(Y_i = 1)}{\partial X_i} = \lambda' (\beta X_i + \omega_w + \gamma_v) \cdot \beta$$

  Here $\lambda'(z)$ is the derivative of the Logistic CDF which is $\lambda'(z) = \lambda(z)(1-\lambda(z))$

- For Poisson regression:

$$\frac{\partial E[Y_i \mid X_i]}{\partial X_i} = E[Y_i \mid X_i] \cdot \beta$$

Table A.5: Effect of Phone Survey Mode Treatment on WaSH Outcome Variables

	Water Source Type		Toilet Type		Freq. of Hand Washing
	Model 1	Model 2	Model 1	Model 2	Model 1
Phone Group	0.590*	0.679**	1.252***	1.264***	0.186***
	[0.11,1.07]	[0.17,1.19]	[0.85,1.65]	[0.86,1.67]	[0.11,0.27]
N	617	617	617	617	614
No. of Cluster	47	47	47	47	47
IP Mean	0.793	0.779	0.449	0.447	4.692
Recall	Usually	Yesterday	Usually	Yesterday	Yesterday

*Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$). IP Mean represents the mean for In-person group.This table displays regression results from logistic regression for water source type , and toilet type. For frequency of hand washing we are using poisson regression. In this table we have used Validation Day Comparison data. The above mentioned coefficients are marginal effect whereas values in parenthesis represents 95% CI. All the models include VESA level clustering of standard error and constant term. Since we do not have Usual recall question for freq. of hand washing, so we are presenting only the yesterday recall coefficient.*

Table A.6: Effect of Phone Survey Mode Treatment on WaSH Outcome Variables

	Water Source Type		Toilet Type		Freq. of Hand Washing
	Model 1	Model 2	Model 1	Model 2	Model 1
Phone Group	0.680*	0.764**	1.633***	1.646***	0.195***
	[0.16,1.20]	[0.23,1.30]	[1.20,2.06]	[1.21,2.09]	[0.12,0.28]
N	640	640	641	641	609
No. of Cluster	47	47	47	47	47
IP Mean	0.793	0.779	0.449	0.447	4.692
Recall	Usually	Yesterday	Usually	Yesterday	Yesterday

*Note: ***($p < 0.001$), **($p < 0.01$), *($p < 0.05$). IP Mean represents the mean for In-person group.This table displays regression results from logistic regression for water source type , and toilet type. For frequency of hand washing we are using poisson regression. In this table we have used 7 Days Average data. The above mentioned coefficients are marginal effect whereas values in parenthesis represents 95% CI. All the models include VESA level clustering of standard error and constant term. Since we do not have Usual recall question for freq. of hand washing, so we are presenting only the yesterday recall coefficient.*

Bibliography

Abate, G. T., De Brauw, A., Hirvonen, K., & Wolle, A. (2022). Measuring consumption over the phone: Evidence from a survey experiment in urban Ethiopia. *Journal of Development Economics*, *161*, 103026. https://doi.org/10.1016/j.jdeveco.2022.103026

Anderson, E., Lybbert, T. J., Shenoy, A., Singh, R., & Stein, D. (2023). Does survey mode matter? Comparing in-person and phone agricultural surveys in India. *Journal of Development Economics*, *166*, 103199. https://doi.org/10.1016/j.jdeveco.2023.103199

Anthonj, C., Setty, K. E., Ferrero, G., Yaya, A.-M. A., Poague, K. I. H. M., Marsh, A. J., & Augustijn, E.-W. (2022). Do health risk perceptions motivate water - and health-related behaviour? A systematic literature review. *The Science of The Total Environment*, *819*, 152902. https://doi.org/10.1016/j.scitotenv.2021.152902

Arthi, V., Beegle, K., De Weerdt, J., & Palacios-López, A. (2018). Not your average job: Measuring farm labor in Tanzania. *Journal of Development Economics*, *130*, 160–172. https://doi.org/10.1016/j.jdeveco.2017.10.005

Beland, Y., & St-Pierre, M. (2007, March). *Mode effects in the Canadian Community Health Survey: A comparison of CATI and CAPI*. Wiley Online Library. https://doi.org/10.1002/9780470173404

Blair, G., Coppock, A., & Moor, M. (2020). When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments. *American Political Science Review*, *114*(4), 1297–1315. https://doi.org/10.1017/s0003055420000374

Contzen, N., De Pasquale, S., & Mosler, H.-J. (2015). Over-Reporting in Handwashing Self-Reports: potential explanatory factors and alternative measurements. *PLoS ONE*, *10*(8), e0136445. https://doi.org/10.1371/journal.pone.0136445

de Leeuw, E. D. (2005). *To mix or not to mix data collection modes in surveys* (tech. rep. No. 2). Bureau of Economic and Business Research.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2016, January). *Internet, Phone, Mail and Mixed-Mode Surveys: the Tailored design method* (4th). John Wiley & Sons Inc.

Evans, R. I., Hansen, W. B., & Mittelmark, M. B. (1977). Increasing the validity of self-reports of behavior in a smoking in children investigation. *Journal of Applied Psychology*, *62*(4), 521–523. https://doi.org/10.1037/0021-9010.62.4.521

Gaddis, I., Oseni, G., Palacios-Lopez, A., & Pieters, J. (2019). Measuring Farm Labor: Survey Experimental Evidence from Ghana. *The World Bank Economic Review*, *35*(3), 604–634. https://doi.org/10.1093/wber/lhaa012

Heerwegh, D. (2009). Mode Differences between Face-to-Face and Web Surveys: An Experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, *21*(1), 111–121. https://doi.org/10.1093/ijpor/edn054

Himmelfarb, S., & Lickteig, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology*, *43*(4), 710–717. https://doi.org/10.1037/0022-3514.43.4.710

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires. *Public Opinion Quarterly*, *67*(1), 79–125. https://doi.org/10.1086/346010

Huhtanen, P., Mustonen, H., & Mäkelä, P. (2015). Effects of telephone versus face-to-face survey modes on reports of alcohol-related attitudes, harms and alcohol consumption. *Journal of Substance Use*, *21*(4), 407–413. https://doi.org/10.3109/14659891.2015.1040091

Kelly, C. A., Soler-Hampejsek, E., Mensch, B. S., & Hewett, P. C. (2013). Social desirability bias in sexual behavior reporting: Evidence from an interview mode experiment in rural malawi. *International Perspectives on Sexual and Reproductive Health*, *39*(01), 014–021. https://doi.org/10.1363/3901413

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and web Surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. https://doi.org/10.1093/poq/nfn063

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Lavrakas, P. (2008, January). *Encyclopedia of Survey Research Methods*. SAGE Publications, Inc. https://doi.org/10.4135/9781412963947

Nord, M., & Hopwood, H. (2007). Does interview mode matter for food security measurement? Telephone versus in-person interviews in the Current Population Survey Food Security Supplement. *Public Health Nutrition*, *10*(12), 1474–1480. https://doi.org/10.1017/s1368980007000857

O'Leary, J., Georgeaux-Healy, C., & Serpell, L. (2024). The impact of continuous calorie restriction and fasting on cognition in adults without eating disorders. *Nutrition Reviews*, *83*(1), 146–159. https://doi.org/10.1093/nutrit/nuad170

Orquin, J. L., & Kurzban, R. (2015). A meta-analysis of blood glucose effects on human decision making. *Psychological Bulletin*, *142*(5), 546–567. https://doi.org/10.1037/bul0000035

Pariyo, G. W., Greenleaf, A. R., Gibson, D. G., Ali, J., Selig, H., Labrique, A. B., Al Kibria, G. M., Khan, I. A., Masanja, H., Flora, M. S., Ahmed, S., & Hyder, A. A. (2019). Does mobile phone survey method matter? reliability of computer-assisted telephone interviews and interactive voice response non-communicable diseases risk factor surveys in low and middle income countries. *PLOS ONE*, *14*(4), 1–25. https://doi.org/10.1371/journal.pone.0214450

Shaheed, A., Orgill, J., Montgomery, M. A., Jeuland, M. A., & Brown, J. (2014). Why "improved" water sources are not always safe. *Bulletin of the World Health Organization*, *92*(4), 283–289. https://doi.org/10.2471/blt.13.119594

Statista. (2025). Digital & Connectivity Indicators - Telecommunication.

WHO & UNICEF. (2005). *Water for life : making it happen*. World Health Organization; UNICEF.