Easy, Informative, and Cheap?
On the Effectiveness of Interactive Voice Response Calls

November 10, 2023

Alexandra Avdeenko[c,d,e], Jakob Gaertner[a], Marc Gillaizeau[a,b], Ghida Karbala[a], Laura Montenbruck[b], Giulia Montresor[a,b] & Atika Pasha[a,b] [1]

Abstract

The expansion of modern communication is creating enormous opportunities to gather survey data remotely. In collaboration with two NGOs across three provinces in rural Pakistan, we assess the efficiency of telephonic interviews conducted by enumerators versus interactive voice recording (IVR, or robocalls). Our results show that interviews led by enumerators largely outperform robocalls in survey quality. In a panel survey with 12,017 NGO beneficiaries, robocalls had lower call pick-up (by 57%), consent (by 92%), and interview completion rates (by 91%). Mistrust in automatized calls, respondent unavailability, and not wanting to lose phone credit were self-reported reasons to drop IVR calls. Testing robocall framing, we find that female voice recordings can partly mitigate low IVR consent rates. Similarly, medical and religious contextualization improve outcomes such as overall item response. However, with 88 times the price of a completed enumerator-led interview, robocalls are substantially less cost-effective in our setting.

Keywords: Survey Methods; Data Collection; RCT; Response Rates; IVR; CATI
JEL Codes: C81; D8; I1; I18

[1] [a] Center for Evaluation and Development, [b] University of Mannheim, [c] The World Bank Group, [d] Heidelberg University, [e] CEPR. Corresponding author: pasha@c4ed.org. The team thanks the Federal Ministry for Cooperation and Development for funding this project. Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged. Moreover, the implementation of this project would have been more challenging without our local branch in Pakistan, C4DE. We thank our colleagues for the local lead in trainings, communication, technical implementation, reporting, translations, and contracting. In particular, we thank Arooba Khurram, Ghulam Murtaza, Muhammad Usama, Shafahat Hussain, Sharafat Hussain, Usama Waheed, and Waheed U Zaman. We would also like to thank our implementing partners, ACTED Pakistan and NRSP, who organized and conducted the phone calls and the interventions. At NRSP, we would like to especially thank Adnan Hussain, Ayub Zafar, Dr. Nabeela Shahid, Ghulam Mustafa Haider, Inam Bari, Khaliq Qureshi, Muhammad Saad Iqbal, Muhammad Tahir Waqar, Pervaiz Ahmed, and Tazeem Khan. We would also extend our thanks to the participants of the European Society for Population Economics 2021 conference and the Methods and Measurement Conference 2021, for their invaluable input and comments on the paper. AEA RCT Registry: November 24, AEARCTR-0006517. Disclaimer: The findings, interpretations, and conclusions expressed in this event do not necessarily reflect the views of the World Bank.

1 Introduction

The COVID-19 pandemic has revealed the limits of real-time surveillance via in-person surveys. In this context, the possibility to acquire data through safer (i.e., remote) and faster technologies has become crucial, and empirical evidence on their performance is of utmost practical interest. Among various remote survey technologies, enumerator-led computer assisted telephone interviews (CATI) are, by now, a well-established technique to monitor the condition and needs of vulnerable populations, especially for routine crisis monitoring in low- and middle-income countries (LMICs) (Zezza et al. 2021, Henderson and Rosenbaum 2020, Glazerman et al. 2020, Greenleaf et al. 2021, Phadnis et al. 2021). Another technology, namely Interactive Voice Response (IVR) a.k.a. robocalls, is gaining traction as an automated survey mode that can be implemented cheaply, and could achieve greater and faster coverage than enumerator-led CATI (Corkrey and Parkinson 2002, Tsoli et al. 2017, Sacks et al. 2015, Aicken et al. 2016). Yet, in spite of the growing interest in cheap and effective ways of collecting data - especially for rural, hard-to-reach areas and low-income settings - evidence in this regard remains scarce and inconclusive.

Robocalls have been mostly tested in high-income countries. In the context of health-related interventions, they have been found effective in fostering healthy behavior through frequent interactions with the patients, though with modest effects for response rates in surveys (see Corkrey and Parkinson [2002] for an overview; Tsoli et al. 2017, Coombes and Gregory 2019). Evidence from Germany (Kreuter et al. 2008) and the USA (Metzger et al. 2000) shows that self-interviewing methods (and robocalls in particular) outperform enumerator-led interviews when the goal is to elicit sensitive information.[2] The latter finding seems to be corroborated in a low-income setting by a study in Burkina Faso showing that female respondents are more likely to report using modern contraceptives in telephonic interviews than in face-to-face surveys (Greenleaf et al. 2020).[3] However, Maffioli [2020] shows that enumerator-led surveys are more successful than robocalls in terms of response, cooperation, refusal, and contact rates when collecting data on experiences with Ebola in Liberia. These differences could, among other things, be driven by differences in the target population's experience with, or access to, digital technologies, but also by differences in trust in the source of the call. While the performance of robocalls in high-income settings seems promising, research in lower-income countries provides less clear-cut conclusions.

[2]Kreuter et al. [2008] consider the effects of two self-administrating methods of conducting digital questionnaires, namely robocalls and web surveys, as compared to enumerator-led interviews, on the likelihood of reporting potentially sensitive information among students in Germany. The authors find that self-administration works better than enumerator-led data collection when it comes to the reporting of sensitive information, with web surveys performing best and robocalls second best. Similarly, in a study in the US, Metzger et al. [2000] find that sexually active homosexual men interviewed via audio computer-assisted self-interviewing methods are more open to sharing self-reports of HIV risk behavior than individuals assigned to interviewer-administered assessments.

[3]Relatedly, in a qualitative study in Guinea, researchers show that due to the sensitive nature of the information, many individuals feel embarrassed to visit a hospital in person. Instead, a discrete smartphone-based solution with integrated self testing was found to have a high degree of acceptance, particularly for younger people, who constitute the most vulnerable group for sexually transmitted infectionss (STIs) (Aicken et al. 2016).

With the original goal to assess the COVID-19 pandemic situation and related needs in remote areas, we designed a study that contributes to the literature on the effectiveness of robocalls in low-income countries. In close cooperation with two large Non-Governmental Organizations (NGOs) in Pakistan, we repeatedly conducted phone interviews with 12,071 of their beneficiaries between August 2020 and January 2021. Set in rural Pakistani villages,[4] our study seeks to investigate two main research questions. First, are IVR calls able to generate equal or higher (a) consent rates, (b) interview completion rates, and (c) response rates (including response to sensitive questions) than enumerator-led calls? To answer this question, we measure the impact of randomly allocating individuals to various interview modes, namely enumerator-led calls, IVR calls, or an alternation of both modes over several waves of data collection. The latter experimental variation seeks to test whether, over time, a change in the mode of interview encourages a person to stay engaged and thereby decreases attrition. Second, we want to understand whether collecting data via different survey modes can decrease the cost of data collections as compared to purely enumerator-led data collections.

The results of our study indicate that, in a setting with high illiteracy and high levels of poverty, collecting data through robocalls is not optimal with respect to response rates and data quality. Only a third of the study participants called via IVR picked up the call, whereas the response rate to enumerator-led calls is close to 80% (a statistically significant difference of nearly 45 percentage points). Similarly, consent rates (in cases where the call was picked up) are 89 percentage points lower for IVR calls than for enumerator-led calls (the latter with a mean of 96.5%). In terms of data quality for calls where consent was acquired, item non-response is also more frequent for robocalls than for enumerator-led calls (nearly a 93% lower response). Interview completion rates for IVR calls (conditional on consent) are 86.7 percentage points lower than those of enumerator-led calls, which perform much better at a completion rate of approximately 96%. At endline, all respondents were contacted through enumerator-led calls. They reported three main reasons for not having stayed on during a robocall: being busy, fear of losing phone credit while participating, and a general mistrust in robocalls.

Our paper also shows that IVR is unsuccessful in collecting sensitive information on household health and respondents' social and labor market engagement. At a time when social distancing was heavily emphasized and even mandated in certain areas, and divulging one's (non-)compliance can be considered sensitive, IVR does not encourage higher responses to questions on the latter than enumerator-led calls. On the contrary, we find that individuals receiving robocalls are less likely to respond to sensitive questions. Interview completion rates for sensitive questions are 91.9 percentage points lower for robocalls, compared to the enumerator-led interview mean of 99%. This is at odds with existing research emphasizing a lack of data quality when collecting sensitive information in direct interaction with survey respondents.[5]

[4]To describe the context, only every fourth person living in rural areas of Pakistan can read, and over half of the rural population of the country falls into the two lowest wealth quintiles (NIPS Pakistan and ICF [2019]).

[5]This includes randomized response methods and list experiments. See, for instance, Warner [1965], Böckenholt et al. [2009], Blair et al. [2015], Lensvelt-Mulders et al. [2005] for randomized response methods and Asadullah et al. [2021], Porter et al. [2021], Aronow et al. [2015] for list experiments. For instance, Rosenfeld et al. [2016] compare different methods to capture information on voting and find that direct questioning

Our findings have direct implications for practitioners, especially when considering the cost of the two data collection methods under scrutiny. Inevitably, high quality responses and potential for credible outreach are closely related to the costs involved in data collections and information studies. Robocalls have been praised for being a cheap alternative to enumerator-led interviews. This is also the case in our study, where the fixed costs required to set up and conduct data collections via robocalls are roughly 16% that of the costs required to set up and conduct an enumerator-led data collection. Yet, while being cheap, robocalls are significantly less effective in our study, with an important cost-effectiveness trade-off.[6] Contrasting the lower costs of making a robocall with the lower response rates they achieve, we find that, due to the decreasing marginal impact of fixed costs, enumerator-led calls outperform robocalls in cost-effectiveness for large data collections. Robocalls are more cost-effective than enumerator-led phone interviews up to a threshold of 481 completed interviews. However, given the low completion and pick-up rates to robocalls, the completition of 481 interviews requires a sample frame with more than 200,000 possible respondents. Beyond small samples, enumerator-led interviews are advantageous and more cost-effective. Similarly, our findings don't support the hypothesis that an iteration between both modes can enhance cost-effectiveness by exploiting the low cost of robocall technology as well as the higher gross completion rates of enumerator-led calls.

The remainder of this paper is structured as follows: In the next section, we explain the experimental setup of our study. We describe the data and key outcomes of the study in Section 3 and the estimation strategy in Section 4. Section 5 presents the findings, and the cost-effectiveness of the two survey modes is presented in Section 6. Finally, Section 7 concludes.

2 Experimental setup and design

The experiment (using two survey modes) was implemented across three waves of data collection. The preceding baseline and succeeding endline data collections were conducted purely via enumerator-led calls. In total there were five waves of data collection, three follow-up waves (FUs) and the baseline and endline data collection waves.

Experimental design. For enumerator-led calls, an enumerator interviewed individuals on the phone, using scripted interview questions, and entered the provided responses in a CATI software (hereafter referred to as enumerator-led calls). During robocalls, individuals were asked pre-recorded interview questions. Respondents could then dial in their desired answer on their phone's keypad.

leads to significant under-estimation of sensitive votes. Similarly, Blattman et al. [2016] validate face-to-face responses to sensitive questions via in-depth qualitative interviews and observations, and find under-reporting of expenditure data, which also correlated with being assigned to a cash program.

[6] Phone interviews have been found to be cheaper than face-to-face surveys (Garlick et al. 2020), and robocalls techniques seem to bear even more potential. Similarly, in a comparison with Demographic and Health Surveys, Maffioli [2020] finds that robocalls bear potential for collecting data at a low cost ($24 per completed interview) in Liberia.

Individuals were randomly assigned to three groups: (1) the control group (C) (35% of the total sample) was always contacted through enumerator-led calls in all waves of data collection; (2) the second group (another 15% of the total sample), namely treatment arm 1 (T1), was contacted only via robocalls in all three waves; (3) The third group (50% of the total sample), or treatment arm 2 (T2), was randomly split in every wave, with part of the group receiving enumerator-led calls and the other part receiving robocalls. Consequently, each individual in T2 received an alternation of enumerator-led calls and IVR calls across the three FUs of data collection. Each individual in the T2 sample received exactly one IVR call over the three waves of data collection. Table 1 presents the three possible alternation sequences for individuals in T2.

Table 1: Random alternations within T2

	Individual 1	*Individual 2*	*Individual 3*
Wave # 1	IVR call	Enumerator-led call	Enumerator-led call
Wave # 2	Enumerator-led call	IVR call	Enumerator-led call
Wave # 3	Enumerator-led call	Enumerator-led call	IVR call

Note: Table 1 displays the division of the T2 sample into three groups of 33%, and the type of call they received across the waves of data collection. The individuals here represent the call order within the three groups which the T2 sample was divided into.

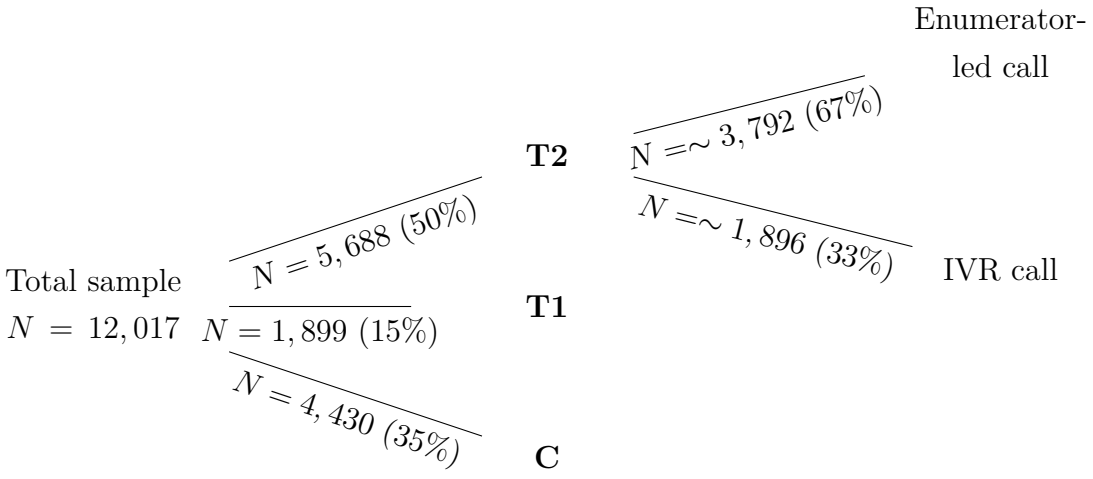
The overall sample was randomly assigned such that, per wave, 33% of the sample received a robocall ($N =\sim 1,896$), while the remaining 67% received enumerator-led calls ($N =\sim 3,792$). The 33% of the sample that received the robocall per wave was never overlapping with the enumerator-led call sample within any of the waves, i.e. no individual received both modes of calls in the same wave.[7]

Figure 1 summarizes the experimental design of the survey-mode variation based on the estimation sample of 12,017.[8]

[7]In preparation for each wave, assigned IVR schedules were compared with the planned enumerator-led calls in the same wave, to avoid an overlap of assigned calls in a single wave.

[8]The baseline sample size was 12,652. However, due to delayed implementation of the robocalls, one part of T2 was not treated through IVR calls in any FU. Therefore, this group did not receive a "mixed" treatment and was removed from the final estimation sample. See Section B.1 in the Online Appendix for more details on randomization, and how the estimation sample was attained from the randomization sample. Figure O.1 in the Online Appendix shows the respective allocations based on the baseline sample.

Figure 1: Survey-mode variation - Impact evaluation design (estimation sample)



Note: Figure 1 summarizes the impact evaluation design of the survey-mode variation with the estimation sample. Related figure: Appendix Figure O.1.

Balance. Table A.1 in the Appendix presents the balance tests across the survey-mode variation.[9] In order to account for the minor imbalances, and to reduce the size of standard errors, we control for baseline covariates in all treatment effect estimations in the analysis. [10]

3 Data

Target population. The target population in our study was vulnerable households residing in the rural areas of Sindh, Punjab, and Khyber Pakhtunkhwa (KP). The study surveyed individuals selected from the beneficiary databases of two large Pakistani NGOs– ACTED Pakistan, and the National Rural Support Programme (NRSP).[11] Given that all survey data was to be collected remotely, the sampling frames were restricted to beneficiaries for whom phone numbers were available.[12]

Data collection. We first conducted a baseline survey between August and October 2020. At baseline, all sampled individuals were called and invited to participate in the study by enumerators. The respondents who consented to participate in this study at baseline were then repeatedly interviewed over a period of five months: Following the baseline survey, three FUs waves of interviews were conducted with the same respondent (Figure 2). Each FU wave was

[9]The respective balance test on the baseline sample is displayed in the Online Appendix Table O.1. The results do not differ in the baseline sample either.

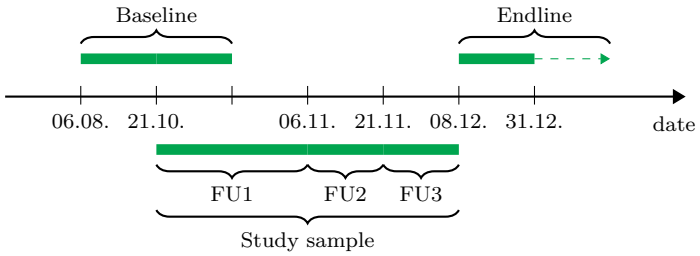
[10]Power calculations were performed at baseline (and prior to the randomization into treatment arms) on the sample used for randomization for a given consent rate. For the number of robocalls that were picked up, our study is underpowered, given the realized consent rate. These calculations are available upon request.

[11]This study is part of a larger project on COVID-19 awareness campaigns, the details within the pre-analysis plan can be found under the AEA registry number AEARCTR-0006555 (further details can be found **here**).

[12]We cleaned over 100,000 phone numbers and merged the beneficiary lists to the 2017 Pakistan census data whenever possible. Through this exercise, phone numbers of beneficiaries belonging to the same villages were identified. A village and the corresponding households were kept in the sample only if the village counted at least 20 different phone numbers.

scheduled to last two weeks.[13] Finally, the endline data collection took place between December 2020 and January 2021 and was conducted via enumerator-led calls only.

Figure 2: Timeline



Note: Figure 2 displays timeline of data collection waves for this study. The analysis sample for our experiment comes from individuals interviewed during FU waves 1, 2, and 3. Descriptive statistics are presented for the baseline and endline samples.

In the first FU wave, when IVR calls where implemented for the first time, it was not possible for individuals to unmask an IVR call based on the number, i.e., individuals could not distinguish whether the call was initiated by an enumerator or by a machine before picking up the call. However, as phone numbers remained identical for both survey modes in all waves, respondents could decide not to pick up the phone upon recognizing the number or even blocking the number if they were not interested in participating any longer. Table A.2 in the Appendix shows the per wave numbers of attempts for the completed interview for the enumerator-led sample. For the robocalls the maximum number of attempts was consistently limited to three.

Questionnaires. The questionnaire consisted of four main modules relating to COVID-19 and the mobility restrictions that were imposed as a result of the pandemic. The questionnaire module addressed the following topics: 1) awareness of COVID-19 symptoms, 2) labour market engagement in past 7 days, 3) social interactions in the past 7 days and 4) a module on health and COVID-19 like symptoms. In contrast to the baseline and endline surveys, the FU surveys were intentionally short. During each FU, enumerator-led calls lasted between five to ten minutes, while IVR calls were, upon the advise of the mobile service provider, limited to a maximum duration of five minutes. IVR interviews were restricted to eleven questions, nine of which were identical to those asked in enumerator-led interviews. To keep FU interviews short, and receive information on as many outcomes as possible, each individual was randomly assigned one of the modules mentioned above, along with the health module, which was asked in every interview. The remaining three modules related to the awareness of COVID-19 symptoms, labour market engagement in past 7 days, and social interactions in the past 7 days, were iterated at random across participants, making sure that all modules were covered once by the end of the data collection. Appendix Table A.3 presents the full questionnaire for IVR interviews. The nine questions that were similarly asked during enumerator-led interviews are highlighted.

[13] In each wave, a small percentage (around 17%) of enumerator-led calls were not carried out in order to not delay the two-weeks time schedule. Calls that were not made during a given wave were then allocated to the next wave.

Summary statistics. Table 2 presents characteristics of our estimation sample at baseline. Over 60% of respondents are men, and the average respondent is around 38 years old. The average household (HH) size is 9, implying that we generally have very large households in our sample.[14]

Almost two thirds of responding households own either land or livestock, as can be expected for a population residing in rural areas. The average individual income earned in the last 7 days is 890 Pakistani Rupees (PKR), amounting to an average weekly income of around 5 United States Dollar (USD).[15] On average, 43% of respondents reported having worked outside the household in the last 7 days. About 15% of the households reported having a member falling sick in the last two weeks. Of the reported sick individuals in the household, at least one (or 0.9) had symptoms similar to those seen in COVID-19 patients– these symptoms are also self reported. On average, the responding household member (HHm) traveled 1.12 days in the last 7 days for visiting (social or religious), and worked around 2 days outside their home.

Table 2: Summary statistics of baseline characteristics

	(1) Mean	(2) SD	(3) Min	(4) Max	(5) N
Female	0.38	0.48	0.00	1.00	12,017
Age	38.13	11.23	18.00	85.00	12,017
HH size	9.00	4.67	1.00	30.00	12,017
HH owns either land or livestock	0.65	0.47	0.00	1.00	12,017
Income (last 7 days)	888.92	1253.76	0.00	3500.00	12,017
Worked (last 7 days)	0.43	0.49	0.00	1.00	12,017
Some HHm fell sick (last 14 days)	0.15	0.35	0.00	1.00	12,017
# of HHm sick with common COVID-19 symptoms (last 14 days)	0.10	0.40	0.00	9.00	12,017
# of days HHm traveled for visit (last 7 days)	1.12	2.17	0.00	7.00	12,017
# of days HHm worked outside home (last 7 days)	2.09	2.80	0.00	7.00	12,017

Note: Table 2 displays statistics on the estimation sample at baseline. Column (1) contains the mean values, column (2) standard deviation, column (3) minimum values, column (4) maximum values and column (5) number of observations.

Outcomes. Table 3 describes the primary and secondary outcomes used in the analysis. Primary outcomes comprise a set of binary variables that capture the overall rates of interview response, consent, completion and item response. Secondary outcomes are binary variables measuring response rates to sensitive questions within the COVID-19 pandemic context.

Pick up the call captures interview overall response rates, taking the value of 1 if an individual picks up the call. *Consent to interview* indicates whether individuals who pick up the call also agree to be interviewed. Therefore, it is defined for individuals who pick up the call and continue until the consent question is presented. The two other primary outcomes are measured for the population of individuals that consents to participate in the interview. *Complete interview* indicates whether an individual stays on the call until the end of the survey. Additionally, *Respond to all questions*, indicates whether an individual responds to all nine questions that

[14]PSLM 2015-16 reports an average household size of 6.47 members for rural households. The 2017 census, on the other hand gives and average household size of 6.45 overall, but 7.9 for rural KP, 6.46 for Punjab and 5.47 for Sindh. However, the instance of rural outmigration is considered high in Pakistan, and may have reduced during COVID, leading to a larger household size in our sample.

[15]The conversion rate is 1 USD = 177.025 PKR, extracted on 8th of January, 2022 from this link.

Table 3: Primary and secondary outcomes

#	(1) Outcome	(2) Description
Panel A - Primary outcomes		
1	Pick-up the call	Fraction of full sample picking up the call
2	Consent to interview	Fraction of sample, that picks up the call, consenting to participate in the survey
3	Complete interview	Fraction of sample, consenting to the interview, and completing the interview (not ending the interview before all questions are presented)
4	Respond to all questions	Fraction of sample that completes the interview, responding to all 9 questions asked
Panel B - Secondary outcomes		
5	Respond to all sensitive questions	Fraction of sample, that completes the interview, responding to all 5 sensitive health and non-health questions asked
6	Respond to all sensitive health questions	Fraction of sample, that completes the interview, responding to all 2 sensitive health questions asked
7	Respond to all sensitive non-health questions	Fraction of sample, that completes the interview, responding to all 3 sensitive non-health questions asked

Note: Table 3 displays the primary and secondary outcomes of interest presented in the main analysis in section 5. All outcomes are dummy variables. Additional outcomes can be found in Table A.4.

the IVR calls and enumerator-led interviews had in common. This outcome depends on the individual completing the interview in the first place.

Secondary outcomes are *response rates to questions referring to sensitive information.* Specifically, the surveys included health-sensitive questions relating to COVID-19 infection, which is particularly stigmatized in the study area. Non-health sensitive questions were also asked. These inquired about engagement in activities considered to increase the risk of infection (travel, work, social gathering).

Appendix Table A.4 describes the tertiary outcomes. These comprise binary variables capturing *response rates to each sensitive question* asked in the survey, and *count variables for the number of questions* answered within the health and non-health categories. All primary, secondary and tertiary outcomes were defined in a Pre-analysis Plan (PAP) prior to the endline analysis.

4 Estimation

We conduct an Intention-to-Treat Effect (ITT) analysis, including every subject who was randomized into the study and received calls as scheduled.[16] All causal effects are estimated using

[16]Note that part of the sample was randomized to T2 but never received any IVR call, because of delays in implementation and technical issues with the mobile company. This sample was dropped from the estimation. However, since the number of questions within the IVR interviews were limited, we cannot be certain that the same individual responded to the robocoll in each wave, over the same contact number. For contacts receiving

Ordinary Least Square (OLS) regressions, pooling the data across the three FU waves Standard errors are clustered at the individual level to account for repeated measurements. We complement the regression analysis with descriptive statistics from the enumerator-led endline survey to provide insights into individual experiences with the IVR survey technology. In the following, we describe the main empirical specifications used for the analysis of each experimental variation.

To estimate the impact of the different mobile technologies used on the indicators of interest, we use the following regression equation:

$$Y_i = \beta_0 + \beta_1 IVR_i + \beta_2 mixed_i + \eta \mathbf{R}_{iv} + \mu_i + \zeta + \epsilon_i \quad (1)$$

The dependent variable Y_i refers to indicators described in Table 3 for individual i. IVR_i and $mixed_i$ are binary treatment arm indicators. IVR_i equals one if the individual was assigned to T1, and zero otherwise. There term $mixed_i$ equals one if the individual was assigned to the mixed treatment arm, T2, and zero otherwise. The omitted category is the control group C, i.e., individuals assigned to receiving enumerator-led calls only. The term $\mathbf{R}_{iv}$ corresponds to a matrix of covariates used in randomization, including both individual and village-level characteristics as described in Table A.1 in the Appendix.

The terms μ and ζ capture NGO and interview wave fixed effects, respectively.

5 Results

After examining the performance of each of the experiments, we find that enumerator-led calls outperform IVR calls in all outcomes we test for. The combination of both technologies considerably reduces the negative effect associated with IVR calls, as evident in results for the T2 arm.

5.1 Improving response behavior through survey mode variation

Table 4 shows the differential effects for groups T1 and T2 as compared to group C (enumerator-led calls only) for the main outcomes. In column 1, we observe that individuals in group T1 are 45.1 percentage points less likely to pick up the call (significant at 1% level). This amounts to a 57% drop in pick-up probability for IVR calls compared to enumerator-led calls in group C where the mean pick-up rate is 79%.The assignment to the mixed treatment group (T2) has a statistically significant negative effect on pick-up rates (i.e., 9.7 percentage points, or about 12% lower compared to the control group).

However, this negative effect is much lower than that observed for group T1 that received only IVR calls, and the coefficients differ significantly across the two treatment arms (as shown

enumerator-led calls, we implemented checks to make sure we were contacting the same individual over time.

by the significant test for equal means). Therefore, alternating IVR and enumerator-led calls considerably mitigates the adverse outcomes of IVR.[17]

Table 4: Survey-mode variation - Response (overall and item), consent and completion

	(1) Pick-up the call, full sample	(2) Consent to interview, conditional on pick-up	(3) Complete interview, conditional on consent and pick-up	(4) Respond to all questions, conditional on consent and pick-up
T1	-0.451***	-0.888***	-0.867***	-0.848***
	(0.009)	(0.007)	(0.025)	(0.022)
T2	-0.097***	-0.108***	-0.006*	-0.013***
	(0.006)	(0.004)	(0.003)	(0.005)
Statistical tests (p-values)				
Equal	0.000	0.000	0.000	0.000
Jointly zero	0.000	0.000	0.000	0.000
Mean	0.790	0.965	0.956	0.912
SD	0.407	0.183	0.204	0.283
Obs.	30,306	19,956	17,096	17,096
R^2	0.128	0.441	0.126	0.069

Note: Sample description for indicator 1- full sample, indicator 2- only sample that picked up the call, indicator 3- only sample that picked up the call and consented to interview, indicator 4- only sample that picked up the call, consented to interview and were asked all modules named r, i, sd and k. The control group are pure enumerator-led calls. Control variables: randomization variables, wave and IP dummies. Significance levels are indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are shown for each estimation below the coefficient. Statistical tests refer to the difference in the coefficients of the two treatment arms.

Robocalls also show a 88.8 percentage points lower likelihood of consent to the interview (C group mean: 96.5%, column (2)), amounting to 92% lower consent in IVR calls. Group T2 performs better than group T1 in terms of consent rates, exhibiting a 10.8 percentage point reduction (11.1%) compared to the control group.

The estimated treatment effect for the IVR technology shows 86.7 percentage points lower interview completion rates compared to enumerator-led interviews (C group mean: 95.6%, column (3)). This large effect is driven by the fact that all enumerator-led calls where the respondent consented to participate were completed, while nearly 91% of respondents assigned to robocalls dropped out of the call before reaching its end.[18] The combination of the two technologies in the T2 arm mitigates this large negative effect of robocalls, showing a decrease of only 0.6 percentage points with respect to the control group.

When looking at response rates to all interview questions in column (4), we again observe a strikingly large negative effect of the robocalls, 84.8 percentage points, amounting to about 93% lower universal response compared to group C (given a control group mean of 91.2%). This implies that, for the same set of questions, the item response rate was considerably different

[17]Note that we capture the effect of only one alternation between the two technologies. This is due to implementation restrictions, as described in Section 3. A larger number of waves would allow more alternations between the two technologies, allowing a deeper analysis on the optimal number and timing of alternations between IVR and enumerator-led calls.

[18]As noted in Table A.5.

between the two technologies. In contrast, group T2 only experienced a decrease of 1.4% with respect to group C for this indicator.[19]

We then test whether IVR calls also underperform on response to sensitive questions. Table 5 reports differential treatment effects on response rates to health and non-health sensitive questions. We detect significant treatment effects on the outcomes response to all sensitive questions, as well as for the response outcomes restricted to only health or non-health sensitive questions. More specifically, the T1 arm shows that robocalls lowers the likelihood of responding to sensitive questions by 91.9 percentage points (group C mean: 98.9%, column 1). Compared to the overall response rate for sensitive questions, the indicator for response to health related questions is also statistically significantly lower for T1 by 55.4 percentage points (column 2). This implies a change of 55.5%, compared to the nearly universal response rate for the control group—99.4%. Response to non-health related questions is also statistically significantly lower for T1 by 92.3 percentage points (column 3), implying a state of nearly no response against the universal response rate for the control group (99.4%). Results for the T2 arm are in line with previous findings, showing mild negative effects for both indicators (6.5 and 1.4 percentage points for health and non-health, respectively).

In our context, it appears that the robocalls are visibly disadvantageous to enumerator-led calls. With the IVR survey mode, the respondents are more likely to drop the call or to not respond to questions. As the call duration increases, the likelihood of the call being dropped increases as well. We plot the frequency of item response for each of the nine questions in the robocalls, per wave, in Figure A.1. There appears to be a clear downward trend, where the number of respondents staying on till the last question declines considerably.

Relatedly, the estimated effects are consistently negative for robocalls for response to each of the sensitive health and non-health questions individually, as illustrated in Table A.7 in the Appendix.[20] We additionally estimate treatment effects on outcomes measuring the length of the interview, as presented in Table A.8 in the Appendix. While the response rates are dummy variables, the length indicators count response to each question included in the indicator. The length indicators—i.e., the total number of questions answered, the total number of health questions answered, and the total number of non-health questions answered—suggest that response rates to the presented questions are higher for interviews conducted by enumerators.

It is important to keep in mind that questionnaire modules other than the health module

[19]While the Table 4 conditions the completion and response to all questions on having acquired consent in the first place, we also run the unconditional estimation for these two indicators, with the set of interviews that did register response in IVR calls. As can be seen in Table A.6 the unrestricted sample shows even worse completion rate and item non-response. However, since these are individuals that did not have a chance to attempt a response of interview completion (since they refused consent), these results are not surprising and intuitive.

[20]The estimates may, however, be affected by the order of questions within the IVR questionnaire. Each outcome is generated on the basis of questions asked consecutively within the robocall, the order in which is displayed in Table A.7. The two health module questions were asked before the three non-health module questions. Due to the larger number of robocalls being dropped prior to the non-health questions being presented, the health questions have a response rate of around 41%—this is only 7% for the non-health questions (Table 2). In fact, when comparing each of the non-health questions, the response rate declines the later the question is asked. From left to right, the increasing size of the estimated coefficients is a reflection of the lower response rate for the non-health modules, over the duration of the interview.

Table 5: Survey-mode variation - Responses to sensitive questions, **conditional on pick-up, acquiring consent and question being asked**

	(1) Respond to all sensitive questions	(2) Respond to all sensitive health questions	(3) Respond to all sensitive non-health questions
T1	-0.919***	-0.554***	-0.923***
	(0.023)	(0.043)	(0.023)
T2	-0.013***	-0.065***	-0.014***
	(0.003)	(0.011)	(0.002)
Statistical tests (p-values)			
Equal	0.000	0.000	0.000
Jointly zero	0.000	0.000	0.000
Mean	0.989	0.998	0.994
SD	0.102	0.044	0.076
Obs.	8,895	1,214	8,385
R^2	0.430	0.349	0.506

Note: Sample description for indicator 1- only sample that picked up the call, consented to interview and were asked at least one module from r, i and sd, indicator 2- only sample that picked up the call, consented to interview and were asked module r, indicator 3- only sample that picked up the call, consented to interview and were asked at least one module from i and sd. The control group are pure enumerator-led calls. Control variables: randomization variables, wave and IP dummies. Significance levels are indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are shown for each estimation below the coefficient. Statistical tests refer to the difference in the coefficients of the two treatment arms.

were randomized in the enumerator-led calls. Therefore, many individuals were not asked the modules i, k and sd in the C arm. This implies that the number of questions the C respondents could have answered is lower as compared to robocall respondent, since some modules were simply never asked. The mean response count for the length indicators lies between 2.2 and 1.5, even though the total number of questions that could have been answered is 9. For this reason, results for outcomes of the randomized modules (social distancing, labor supply, COVID-19 knowledge) shall be interpreted with caution. Recall that for IVR interviews, all selected questions were asked, given that no module was randomized. Despite the lower probability of being asked the question in enumerator-led interviews, robocalls still appear to be a less effective technology for eliciting item response in our sample. For the sensitive questions indicators, the number of questions answered is always lower for IVR calls.

Overall, the results depict a clear picture for the main research question of this paper– robocalls are not an optimal survey mode in the case of our sample. They consistently perform poorer than enumerator-led interviews, regardless of which outcome is considered. The results are more ambiguous for the mixed treatment group (T2). While this group performs better than T1 when compared to the C arm, it appears that this result is driven by the enumerator-led interviews performing well, while the IVR calls show similar trends as in the T1 arm. Consequently, the automation of interviews, especially if these relate to health monitoring, would be a poor choice in the case of Pakistan, or at least for the type of vulnerable population we consider. This is further discussed under the section on representativeness, where the

different sample characteristics observed across these survey methods shed more light on the performance of robocalls versus enumerator-led interviews.

5.2 Evidence on response behavior between survey modes

Table A.5 presents mean values of our primary and secondary outcomes on the full estimation sample (Part I) and by experimental arm (Part II - IV). On average, a household member picked up the call in 66% of cases. This number is mostly driven by high response rates for enumerator-led interviews (79% compared to 34% in the case of robocalls). Overall, response rates tend to improve over time (as evidenced in Figure A.2 in the Appendix).[21] In the mixed arm (T2), the same increase in response rate is observed for both survey technologies, although the increase is lower in the case of the robocalls–at $\sim 2.5\%$–than for enumerator-led interviews–at $\sim 5\%$.[22]

The difference in pick-up rates between the two survey modes seems strikingly large, and implementation data suggests that there are two main reasons underlying this result, largely stemming from enumerator proactivity and adaptability. Firstly, enumerators could attempt to reach a target several times, when IVR calls would only attempt it three times. Enumerators were paid only for every completed interview, and as such, enumerators attempted up to eight calls to acquire response (and consent) from their contact numbers. Moreover, even if the respondent picked up the call, and factors such as loud background noise, bad connectivity, and other technical issues were observed, the call was concluded and another attempt was made to contact the respondent within the same FU wave. Thus, even actual response was purposely coded as "non-response" by the enumerators if there were any technical barriers faced on either end of the call. This flexibility was not granted to robocalls, where any technical barriers faced by the respondent became actual barriers to the completion of the interview. In addition, in the case of robocalls, any pick-up was considered as a legitimate response by the server–e.g., even a if the respondent picked up the call and stayed online for a second–and hence this contact number was not considered for further attempts. When focusing only on the first attempted call in the group assigned to enumerator-led interviews, the response rate observed was close to 35%, quite similar to the average response rates observed for IVR calls. Consequently, the differences in enumerator reactions might have made enumerator-led interviews more effective than robocalls.

Secondly, we observe a learning effect in enumerator calling strategy, as the data clearly shows that the timing of attempted enumerator-led interviews changed over the FU, and likely adapted to the availability of respondents. Comparing the three waves, the share of individuals called outside the usual 8 AM to 5 PM working hours increased from 14.49% to 18%. Therefore, by the third wave, enumerators made up to a fifth of their designated call load outside the usual working hours, in order to improve response rates, and thereby consent, completion and item response. The robocalls, on the other hand, were all conducted between 8 AM to 5 PM.

[21] The response rate in group T1 increased by around around 3.5 percentage points over the three waves, from 31.96% to 35.5%. For group C, the increase across waves is 75.85% to 82.74%, or around 7 percentage points.

[22] The wave-based participation for the other primary outcomes is shown in Figures A.3, A.4 and A.5.

Individuals that were too busy to respond to the calls during these standard working hours might have declined the robocall. This is also confirmed by respondents in self-reported endline data, as discussed later in this section. Table A.9 in the Appendix compares the sample of first attempt for enumerator-led interviews, only conducted between 8 AM and 5 PM, to examine the response rates. As can be seen, with only the first attempts during usual working hours, the response rate in enumerator-led interviews (C or T2) reduces to 56%. While not as low as 34%, as in the case of robocalls, this is a considerable reduction from the 79% response rate in our current sample including multiple attempts, outside of the working hours.

The evidence above highlights the shortcomings of robocalls in terms of "learning" and "reactivity". While being cheaper (per attempt) and achieving much higher coverage, faster, they may not allow the same flexibility in interviews that enumerators do. Albeit subjective to the enumerator, this flexibility may (and in our data likely does) influence survey response, consent and completion to a considerable extent.

Self-reported reasons for non-response in IVR calls. Table 6 shows the self-reported experiences with the survey-mode and framing variation from the endline survey. Of our total randomization sample, 6,690 respondents answered questions related to the experiment at endline. Of all respondents, 34.2% reported having received an IVR call. The remaining 65% of the sample were called by enumerators, implying that our 30% - 70% ratio was more or less upheld during implementation. Of those that received robocalls, 23.3%—or 523 respondents—confirmed that they did not continue with the call. Given the low measured consent rate (compared to the $\sim$ 75% that reported continuing with the call), we believe that these are the respondents that immediately dropped the call, while the sample that initiated but then did not indicate consent by typing their response is much larger.

Of the endline sample that dropped the IVR call, the most common reasons not to continue were: "I was too busy" (51.4%), "I did not want to lose phone credit" (18.4%) and "I do not trust robocalls" (16.3%). Respondents also suggested that technological barriers played a role (options "could not hear message well", "had problems with phone/keypad", "does not know how to use keypad"), although these were not major concerns. 7.6% of respondents were not inclined to continue with the interview due to earlier participation in a similar interview, implying their disinterest in providing the same information again. Moreover, 6.9% of respondents clearly mentioned that they are not interested in such interviews, which could imply their disinterest in robocalls, or surveys in general. It is important to mention the bias in response stemming from the presentation of the questions. While all variables in Table 6 from row 3 onward are binary indicators, these were created from a single multiple choice question.[23]

[23]It may be the case that the order of the options might affect the response to these options. The order of the indicators in the table replicates the order of the choices in the questionnaire.

Table 6: Descriptive information on implementation of IVR calls

	(1) Mean	(2) # Obs.
Receive an IVR call	0.342	6,690
Did not continue with IVR call (said they discontinued IVR call)	0.233	2,249
Resp. was busy at the time of the call	0.514	523
Resp. did not want to lose phone credit	0.184	523
Resp. does not trust robocalls	0.163	523
Resp. did not feel comfortable with the voice used in the recording	0.021	523
Resp. could not hear the message well	0.031	523
Resp. had problems with phone/ keypad	0.015	523
Resp. does not know how to use the keypad	0.017	523
Resp. has already participated in similar interviews	0.076	523
Resp. was not interested in participating in such interviews	0.069	523
Interview was too long	0.023	523
It was not clear who is calling	0.029	523

Note: This table contains summary statistics from the endline survey. Column (1) displays the mean and column (2) the number of observations. All variables are binary indicators. Indicator 3 onwards were created from a single multiple choice question that asked individuals that indicated that indicated having received an IVR call (row 1) and said that they did not complete the IVR call (row 2) .

In the following section, we examine whether these results for response and consent also correlate with sample characteristics.

Representativeness. As explained in Section 3, it was not the goal of this project to generate a dataset which is representative of Pakistan's adult population. By using phone numbers from NGO beneficiaries, we deliberately focused on a vulnerable part of the population. However, the random assignment of different survey modes *after* the baseline interview allows us to examine which of these survey modes is most likely to preserve the sample composition reached at baseline.

To do so, we look at a set of sample characteristics over time and by survey mode. Table 7 displays the overall baseline means (see column (2)) as well as the differences between survey mode-specific means and baseline means for each survey wave (see columns (4) to (7)).[24] We show these comparisons for the subset of individuals who consented to take part in the survey. Throughout most of the comparisons, the differences between baseline means and IVR-means are insignificant. Although indicative of a negligible difference between the sample composition of the pure IVR sample and the overall baseline sample, this finding has to be treated with caution. Given the low response and consent rates for IVR calls, the sample size for IVR calls is small and measured differences are thus less likely to be significantly different from zero. In the following, we will thus not only look at statistical significance for mean differences in the

[24]This approach is similar to work by Brubaker et al. [2021], Gourlay et al. [2021], Lau et al. [2019] examining the representativeness of phone interviews as compared to nationally representative face-to-face interviews. We present the overall sample mean without differentiating by treatment arm. However, baseline means do not significantly differ across arms for the presented variables.

Table 7: Sample composition over time

				Difference			
	Base. N (1)	Base. mean (2)	Treatment (3)	FU1 (4)	FU2 (5)	FU3 (6)	Endline (7)
Female	12652	0.379	Enumerator-led	-0.066***	-0.082***	-0.049***	-0.031***
			IVR	0.058	0.050	0.088	-0.041***
			Mixed	-0.058***	-0.062***	-0.026***	-0.025***
18 - 50 years	12652	0.877	Enumerator-led	-0.021***	-0.023***	-0.019***	-0.015***
			IVR	0.086***	0.075**	0.123***	-0.022**
			Mixed	-0.001	-0.008	-0.001	0.001
No degree	12578	0.463	Enumerator-led	-0.047***	-0.054***	-0.034***	-0.026***
			IVR	-0.118*	-0.106	-0.063	-0.013
			Mixed	-0.024***	-0.025***	-0.010	-0.011*
Primary school education	12638	0.480	Enumerator-led	0.042***	0.044***	0.024***	0.016*
			IVR	0.102	0.044	0.098	0.006
			Mixed	0.020**	0.025***	0.012*	0.011
HH owns livestock or land	12652	0.651	Enumerator-led	0.023***	0.023***	0.011	0.015*
			IVR	-0.015	-0.008	-0.037	0.017
			Mixed	0.015*	0.020***	0.011	0.014**
Respondent income at BL	12652	884.785	Enumerator-led	26.737	31.047	19.483	28.765
			IVR	-5.513	-134.822	-211.575	18.584
			Mixed	26.133	53.830***	-5.327	20.264
Respondent worked (prev. 7d.) at BL	12652	0.430	Enumerator-led	0.001	0.002	0.000	0.007
			IVR	0.035	-0.073	-0.112	0.009
			Mixed	0.008	0.010	-0.003	0.011*
Up to 5 HH members	12652	0.177	Enumerator-led	-0.005	-0.002	-0.001	-0.007
			IVR	-0.050	-0.034	-0.088**	-0.012
			Mixed	-0.024***	-0.013**	-0.011**	-0.013**

Note: We display mean differences for a set of sample characteristics at baseline as compared to follow up waves and endline and by treatment arm. Column (1) shows the sample size, column (2) shows the overall baseline mean. Column (3) shows to which treatment arm the baseline value is compared. Columns (4) to (7) display the differences between the baseline mean and the treatment arm mean at FU1, FU2, FU3, and endline, respectively. Differences are calculated by subtracting the baseline sample mean from the mode- and time-specific mean. T-test significance levels are indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

IVR call sample, but also loosely interpret the direction and magnitude of the differences to the baseline sample.

Compared to a baseline value of roughly 38% of female participants, follow-up interviews conducted by enumerators or a mixture of both survey modes attract significantly less female respondents. As opposed to this, if anything, the share of women consenting to IVR follow-ups is higher than at baseline, and it is significantly higher than in enumerator-led follow-up calls (see Appendix Table A.10). This effect is reversed at endline, where all individuals are called via enumerator-led calls again, suggesting that the difference is indeed driven by a higher willingness of women to respond to IVR-based surveys rather than those conducted by an enumerator. A potential explanation for this pattern could be that women are more reluctant to share personal information with an unknown, potentially untrusted (male) individual rather than a recorded voice.

There are also (significantly) larger shares of younger and more educated individuals who consent to IVR calls as compared to the baseline sample. This is in line with the idea that younger and more educated individuals are more familiar with the relevant technologies, suggesting that less educated and older individuals are left behind when forced to reply to an interview using IVR technology. Interestingly, while the sample in the group of enumerator-led interviews is older than at baseline, i.e., the opposite of IVR, the bias towards more educated individuals is visible throughout all survey modes. This indicates that the sample composition in terms of education is not only related to the survey mode, but potentially also to the understanding of the importance of continuous participation and the willingness to contribute to the research undertaking.

The share of households who own livestock or land, an important proxy for wealth, is

significantly larger than at baseline for most of the enumerator-led and mixed method survey waves. This may be related to the fact that enumerator-led interviews take long and the opportunity cost of taking that time is comparably high for poorer individuals who rely on their daily income and working time.[25] Comparing the share of individuals who worked at baseline as well as the respondents' average income, there seem to be no significant differences between the baseline means and the mode-specific means over time. However, the signs of the differences support the outlined hypothesis.

Finally, there are no significant differences between the share of households with a maximum of five members at baseline and that of enumerator-led and IVR calls throughout most of the survey waves. However, the share of individuals with comparably smaller household sizes is significantly smaller in the mixed treatment arm, which, looking at the size of the coefficients for the IVR group, may well go back to the IVR part of the random alternation.

Overall, these results show that there is no clear advantage of either survey mode in terms of their ability to preserve the composition of the baseline sample. If anything, certain characteristics of the sample are correlated with a higher likelihood to consent to one mode, but not another. For instance, significantly more women consent to take part in IVR interviews than in enumerator-led interviews. The same is true for younger, potentially more technology-savvy individuals. We also see this difference, if not as stark, when comparing the group of individuals who received enumerator-led calls only against those who received a combination of both modes.[26] These findings are important in understanding the advantages and disadvantages of different survey modes and illustrate issues of sample representativeness beyond that related to phone ownership. However, they also give insights into how the use of different survey modes could help to reach different population groups. More research on this topic is needed in order to draw clear, generalizable conclusions.

6 Cost effectiveness

This section compares the costs across the three modes of data collection examined in this study: enumerator-led calls, robocalls and an alternating combination of the two. The comparison between telephonic data collection methods and face-to-face data collection became more prominent in recent literature, especially with the outbreak of the COVID-19 pandemic which called for rapid data collection for rapid action and policy adoption (Gourlay et al. 2021). However, only a handful of studies compare data collection costs across different types of telephonic modes of data collection (Leo et al. 2015, Lau et al. 2019), and, to our knowledge, no

[25]Note that there are no significant differences between the means of the three survey modes within the same survey wave. This is in line with the mentioned findings for a comparison of the enumerator-led vs. the mixed survey mode. For pure IVR calls, we would expect to find significant differences when comparing means to those of the two other survey modes, which we do not. Yet, the sign of the differences as well as the large magnitudes seem supportive of our hypothesis and statistical insignificance is most likely related to the small size of the IVR sample. For a complete illustration of the results, please refer to Appendix Table A.10.

[26]Differences between the means of the three treatment arms *within* survey wave are presented in Appendix Table A.10.

studies examine whether alternating the use of both enumerator and robocalls enhances cost-effectiveness. To contribute to this literature, we compare the costs per interview attempt and per completed interview for enumerator-led calls (C), robocalls (T1) and for the alternation between both modes (T2).[27] To do so, we exclude all costs not related to the data collection exercise (e.g., implementation costs for other experimental interventions that are part of this project) and include only costs associated with follow-up surveying where different data collection modes were used.[28] Given that all costs were incurred in the same year, we do not adjust for inflation and time value of money.

We differentiate between fixed costs and variable costs. Fixed costs are not sensitive to the number of interviews conducted and include costs for project and data management, enumerator training, and for supplies and stationary. In the case of enumerator-led calls, fixed costs include mainly the cost of hiring and training enumerators, while for IVR calls fixed costs include primarily the costs of recording the survey. Alternating randomly between two modes of data collection, as we did in T2, requires that the necessary structures for both modes are set up. Hence fixed costs for this treatment arm are calculated as the sum of costs incurred using both survey modes.[29] Variable costs increase with the number of interviews and include airtime costs for both enumerator-led and robocalls as well as daily enumerator rates for the former.[30] We exclude our own labor in drafting the questionnaire, programming, translation, and designing. The analysis does not permit a completely balanced comparison between the different modes because the questionnaire used in enumerator-led calls includes more questions than the one used in IVR interviews (as discussed in Section 3). On average, enumerator-led interviews contain 19 questions while IVR interviews contain 11 questions. The analysis therefore marginally understates the cost of IVR as compared to enumerator-led interviews.

Columns (1) and (2) of Table 8 illustrate, among others, the gross completion rates, total cost as well as the implied cost per interview for C and T1, i.e., enumerator-led and IVR calls in our data collection. Column (3) illustrates the same indicators in a setting where enumerator-led and IVR calls are used in alternation. We define the gross completion rate as the number of

[27]We define an attempted interview as any attempt to reach a targeted participant, irrespective of success (i.e., it includes calls that were not picked up, interviews not consented, non-completed interviews and completed interviews). We count one interview attempt per potential respondent. A completed interview is a consented interview in which the participant was presented with all the intended questions in the survey.

[28]In other words, we exclude variable baseline costs which are exclusively related to enumerator-led calls and all endline surveying costs. Baseline fixed costs of enumerator-led interviews are added to the fixed costs of follow-up surveys as those costs were essential to allow for follow-up surveying to take place.

[29]The implementing partner had uncured and reported costing data per mode of data collection and not per treatment arm. This means that we are not able to directly distinguish between the costs incurred for each treatment arm, but only by survey mode: enumerator-led phone interviews vs robocalls. Hence our approach to calculating the fixed cost of a random alternation between the two survey modes is based on the approximation of the overall fixed costs of both survey modes, rather than the marginal costs of adding IVR calls once the infrastructure for enumerator-led calls is already established or vice versa.

[30]We approximate the variable costs for the mixed treatment arm by allocating the share of attempted/completed IVR interviews multiplied by the total variable costs incurred for calls made via IVR. We do the same for enumerator-led calls. Vice versa, variable costs in C and T1 are calculated by multiplying the share of IVR/enumerator-led calls made in T1/C by the total amount of fixed costs using the respective survey mode.

Table 8: Costs in control group and per treatment arm

	(1) C	(2) T1	(3) T2
Number of questions	19 (average)	11	
Gross completion rate	72.1%	0.2%	53.5%
Overall cost			
Total fixed costs (in USD)	16,626.02	2,582.01	19,208.02
Total variable costs (in USD)	2028.49	237.02	2249.93
Total costs (in USD)	18654.51	2819.03	21457.95
Cost per interview			
Variable cost per attempted interview	0.31	0.06	0.25
Variable cost per completed interview	0.43	29.63	0.46
Total cost per attempted interview	2.88	0.74	2.34
Total cost per completed interview	3.99	352.38	4.37

Note: Cost data was available for only one of the implementing NGOs. Accordingly, our cost effectiveness analysis is based on the costs and gross completion rates for this NGO only. Gross completion rates are defined as the number of completed interviews divided by the number of unique numbers attempted. Total variable costs shown in this table are approximated for each treatment arm using the number of *completed* interviews conducted with the respective survey modes in the respective treatment group. We use this estimate to calculate total costs by treatment arm. Using the number of attempted calls instead gives very similar estimates of total (variable) costs. Costs per attempted call are based on the estimated variable costs using call attempts.

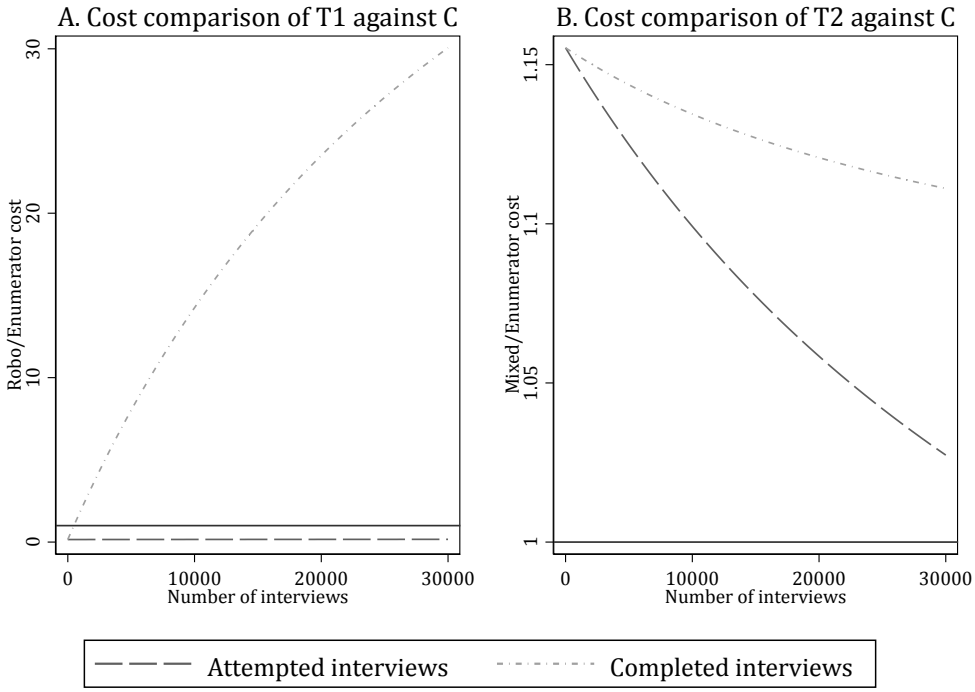
completed interviews divided by the number of unique phone numbers attempted.[31] The fixed costs of setting up an enumerator-led phone data collection are almost 6.5 times as high as those of setting up an IVR-based data collection. By definition, the fixed costs of alternating the use of both modes are even higher (and are a summation of costs of both modes). Since the total number of interviews conducted differs substantially by survey mode, we calculate the variable cost per interview for a better comparison. The last section of Table 8 shows that the variable cost per attempted interview in T2 is more than 4 times that of an attempted interview in a purely IVR-based data collection and only 0.8 times that of an attempted interview in an enumerator-led data collection, indicating that, similar to the fixed costs, variable costs per attempted interview are the lowest for robocalls.

In the last row of Table 8, the total costs per completed interview are presented. Results show that the relatively low gross completion rate for robocalls makes this mode the most expensive in terms of completed interview costs. The total cost of one completed IVR interview is almost 88 times the total cost required to complete one enumerator-led interview and 80 times the cost of completing one interview in T2.

To put the above into perspective, graph A of Figure 3 plots the total costs of conducting robocalls normalized to the cost of enumerator-led calls for different numbers of attempted and completed interviews. Graph B plots the cost of iterating the use of both modes to the cost of only enumerator-led interviews. Figure 3 aims at determining the sample size at which one mode of data collection becomes cheaper than the other, taking into account interview completion rate and economies of scale. Graph A shows that robocall attempts are cheaper than enumerator-

[31] Other authors call this ratio the response rate. We deviate from this wording due to the definition of response rate we used in the main part.

Figure 3: Costs relative to enumerator-led calls



Note: Figure A illustrates the cost of conducting robocalls as a proportion of the costs for enumerator-led calls depending on the number of attempted and completed interviews. Figure B does the same for a random alternation of both survey modes as compared to enumerator-led interviews. The horizontal lines in both graphs show the threshold at which the cost ratio of the two respective methods is equal to one, i.e., at which both methods are equally expensive.

led call attempts for any number of attempts. However, the cost of a completed interview using IVR as compared to using enumerator-led calls gradually increases with the number of planned completed interviews, rendering enumerator-led interviews more cost effective than IVR interviews for any planned sample size larger than 481 completed interviews. This suggests that robocalls are cheaper than enumerator-led interviews for comparably small data collections. However, given a gross completion rate of 0.2% to robocalls, more than 200,000 potential respondents would be needed to end up with less than 500 completed interviews. This number seems to be disproportionate in a setting where researchers aim for a relatively small sample size.

Similarly, Graph B shows that using a mix of both modes is always less cost-effective than enumerator-led interviews when looking at completed interviews. These findings indicate that, given the cost structure and the gross completion rates underlying this project, enumerator-led calls are far more cost-effective than robocalls. Moreover, enhancing cost-effectiveness via an iteration between both modes, to exploit the low cost of robocall technology as well as the higher gross completion rates of enumerator-led calls, is not feasible under the presented circumstances.[32]

[32]Note that the aim of T2 was to collect roughly 1/3 of the data via robocalls and 2/3 via enumerator-led calls. Given the low gross completion rates to IVR calls, the final distribution was biased almost exclusively towards enumerator-led calls. It is possible to vary the share of expected completed interviews conducted using robocalls to make the combination of IVR and enumerator-led calls more cost effective. However, this is not a promising approach given the low gross completion rates to IVR in our setting. Results outlining these findings

A comparison of our work to existing literature suggests that the cost effectiveness of phone survey modes depends to a large extent on the context-specific differences in gross completion rates and (implied) costs per completed interview.

Gourlay et al. [2021] shows that the cost per completed enumerator-led interview for the World Bank Living Standards Measurement Study (LSMS) ranges between 7.84 USD in Burkina Faso and 13.11 USD in Nigeria with around 1,950 completed interviews in each case. This is roughly in line with the estimated cost of 8.75 USD per completed enumerator-led interview for a sample of 2,000 completed interviews in our setting (see Appendix Table A.11). Similarly, gross completion rates of 63% to 90% are comparable to a gross completion rate of 72% in our setting.

The cost of and gross completion rate to IVR calls in our setting differ a lot as compared to other settings and, most strikingly, robocalls are never a cost-effective and feasible approach. Both Lau et al. [2019] and Ballivian et al. [2015] compare the cost of completed enumerator-led vs. IVR interviews. With much lower attrition rates for IVR calls than in our setting, the average cost of one completed IVR interview amounts to 68% of the cost of a completed enumerator-led interview for a sample of 1,500 completed interviews in Honduras and Peru (Ballivian et al. 2015). It would be nearly 3 times as large in our case. Costs for completed interviews are cheaper in the case of robocalls, with 17 USD as compared to 31.35 USD at a sample size of 1,500 in our setting. On the other hand, at a price of 11.52 USD per completed interview, enumerator-led interviews are much cheaper in our setting as compared to a price of 25 USD in Ballivian et al. [2015]. In Nigeria, Lau et al. [2019] shows that the price of a completed robo-interview is around 43% that of a completed enumerator-led interview for 3,000 completed interviews. This is far cheaper than in our case, and most likely driven by a gross completion rate to robocalls which, at 3%, is more than 10 times higher than in our setting. Studies in high-income countries even find gross completion rates of up to 65% for robocalls, making IVR a far more attractive technology (Tsoli et al. 2018, Andersson et al. 2014, Daher et al. 2017).

To our knowledge there exists no literature that compares the two mentioned modes of data collection to an approach where a random iteration between the two is used.

In summary, the cost of enumerator-led phone interviews seems to be fairly comparable across settings. However, tremendous differences in gross completion rates make pure IVR data collections more or less feasible depending on the context. An important aspect in determining expected gross completion rates is the familiarity of respondents with similar technologies, which should be taken into consideration when planning the roll-out of IVR-supported data collections. Once the suitability of robocalls is confirmed, a potential strategy to increase gross completion rates endogenously is to provide incentives for participation. However, the additional cost of doing so needs to be incorporated when calculating the cost-effectiveness of such strategy. Finally, looking at currently quite low gross completion rates for robocalls in developing countries, the gains from incentivizing participation have to be enormous for

are available upon request.

this to make robocalls cost-effective, taking into consideration the huge amount of potential respondents necessary to reach the desired sample size–even at much larger gross completion rates than in our case.

7 Conclusion

In an attempt to understand and explain interview response behavior, our study tested the performance of two survey technologies that have become increasingly widespread - enumerator-led CATI calls and IVR calls. Contributing to the scarce literature comparing the performance of either technology in low-income countries, our results show that enumerator-led interviews outperform IVR interviews in all regards in the context of rural Pakistan. IVR calls perform worse not only in terms of the likelihood of responding to the call or consenting to the interview, but also in terms of interview completion rates and the likelihood to respond to sensitive questions. In a novel attempt to test whether alternating both approaches within waves combines the cost effectiveness of recorded interactive calls with the (hypothesized) higher engagement of enumerator-led telephonic interviews, our study finds that purely enumerator-led calls are superior in both regards. Iterating between enumerator-led and robocalls moderately mediates the overall negative effects of IVR calls on response behavior. Howver, we show that robocalls are not a cost-effective alternative to enumerator-led interviews despite their lower fixed costs, even when alternated with enumerator calls. This is due to the extremely low gross completion rates of robocalls.

Our evidence suggests that the effectiveness of innovations in data collections need to be carefully tested and adjusted to the local setting. We present such tests and attempts for adjustments, but eventually the results indicate that the fundamental challenges faced in rural, poor settings are reflected in the lack of capacities to quickly adopt innovative technologies. While in other, more high-income settings, IVR calls have been found to be an effective means of health monitoring (for immunosuppressive individuals or individuals with highly transmissible infections, who are either faced with the risk of infection, or transmitting the infection, upon clinical visits) and have been shown to be more effective in extracting sensitive information than enumerator-led calls, participants in our study did not adopt the technology.

Based on self-reported information collected in an enumerator-led endline survey, we find suggestive evidence that mistrust in robocalls, fear of losing phone credit while participating, and a general mistrust in robocalls are the main reasons for not having participated in these calls. In this regard, the adaptability of the enumerators (e.g., to respondents' barriers in terms of time or trust) was a clear advantage, which led to much higher response, consent and completion rates for enumerator-led calls.

Policy research is progressively moving towards telephonic information gathering– methods that are safer (implying less in person contact), faster, and allow a large coverage. Especially in the sphere of public and private health monitoring, the nature of these technologies is an added advantage. For instance, remotely situated individuals, or contact-wary individuals, are

less likely to suffer from the burden of health information inequality, exacerbated by the lack of quality surveillance, with the greater reach of mobile-based technologies. However, evidence from our study shows that, while effective in certain settings, these technologies can be ineffective in others. With a heavily under-capacitated and overburdened medical infrastructure, as in the case of rural Pakistan, the role of mobile-based data collection and monitoring will remain pertinent. With our study, we show that the robocall technology is indeed cheap and easy to administer, yet unlikely to resonate where low levels of literacy and high levels of distrust prevent its uptake.

These results also suggest important avenues for future research. While enumerator-led phone surveys seem to be an effective data collection tool even in rural, low-income settings as in our study, contextual factors may lower the performance of IVR as a remote data collection mode. Future work should focus on untangling the barriers to the uptake of this technology, some of which are evident in our study. Over multiple waves of data collection, the IVR technology may be adapted to "learn" what timings are suitable for improving response. Alternatively, where multiple waves of data are collected, investment into an enumerator-led awareness raising call at the start of the data collection could facilitate information on the benefits of robocalls, and lead to lower mistrust. Finally, understanding the audience is key. Our results show that the younger, more educated respondents were likelier to respond to robocalls, as compared to a less educated, elder audience. Robocalls could therefore be tested as an effective technology in interviews that are largely targeted at a younger, more educated (and potentially less technologically fazed) population. With sufficiently high pick-up rates, the role of framing in improving IVR interview completion rates should be tested in more detail.

References

Catherine RH Aicken, Sebastian S Fuller, Lorna J Sutcliffe, Claudia S Estcourt, Voula Gkatzidou, Pippa Oakeshott, Kate Hone, S Tariq Sadiq, Pam Sonnenberg, and Maryam Shahmanesh. Young people's perceptions of smartphone-enabled self-testing and online care for sexually transmitted infections: qualitative interview study. *BMC public Health*, 16(1):974, 2016.

Claes Andersson, Susanne Danielsson, Gunilla Silfverberg-Dymling, Gunnel Löndahl, and Björn Axel Johansson. Evaluation of interactive voice response (ivr) and postal survey in follow-up of children and adolescents discharged from psychiatric outpatient treatment: a randomized controlled trial. *SpringerPlus*, 3(1):1–3, 2014.

Peter M Aronow, Alexander Coppock, Forrest W Crawford, and Donald P Green. Combining list experiment and direct question estimates of sensitive behavior prevalence. *Journal of Survey Statistics and Methodology*, 3(1):43–66, 2015.

M Niaz Asadullah, Elisabetta De Cao, Fathema Zhura Khatoon, and Zahra Siddique. Measuring gender attitudes using list experiments. *Journal of Population Economics*, 34(2):367–400, 2021.

Amparo Ballivian, João Pedro Azevedo, Will Durbin, J Rios, J Godoy, and C Borisova. Using mobile phones for high-frequency data collection. *Mobile Research Methods*, 21, 2015.

Graeme Blair, Kosuke Imai, and Yang-Yang Zhou. Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, 110(511):1304–1319, jul 2015. doi: 10.1080/01621459.2015.1050028. URL `https://doi.org/10.1080%2F01621459.2015.1050028`.

Christopher Blattman, Julian Jamison, Tricia Koroknay-Palicz, Katherine Rodrigues, and Margaret Sheridan. Measuring the measurement error: A method to qualitatively validate survey data. *Journal of Development Economics*, 120:99–112, 2016.

Ulf Böckenholt, Sema Barlas, and Peter GM Van Der Heijden. Do randomized-response designs eliminate response biases? an empirical study of non-compliance behavior. *Journal of Applied Econometrics*, 24(3):377–392, 2009.

Joshua Brubaker, Talip Kilic, and Philip Wollburg. Representativeness of individual-level data in covid-19 phone surveys. 2021.

Jody D. Ciolino, Alicia Diebold, Jessica K. Jensen, Gerald W. Rouleau, Kimberly K. Koloms, and Darius Tandon. Choosing an imbalance metric for covariate-constrained randomization in multiple-arm cluster-randomized trials. 20, 2019.

Caitlin E. Coombes and Megan E. Gregory. The current and future use of telemedicine in infectious diseases practice. *Current infectious disease reports*, 21:41, October 2019. ISSN 1523-3847. doi: 10.1007/s11908-019-0697-2.

Ross Corkrey and Lynne Parkinson. Interactive voice response: Review of studies 1989–2000. *Behavior Research Methods, Instruments, & Computers*, 34(3):342–353, aug 2002. doi: 10.3758/bf03195462.

Jana Daher, Rohit Vijh, Blake Linthwaite, Sailly Dave, John Kim, Keertan Dheda, Trevor Peter, and Nitika Pant Pai. Do digital innovations for hiv and sexually transmitted infections work? results from a systematic review (1996-2017). *BMJ open*, 7(11):e017604, 2017.

Robert Garlick, Kate Orkin, and Simon Quinn. Call me maybe: Experimental evidence on frequency and medium effects in microenterprise surveys. *The World Bank Economic Review*, 34(2):418–443, 2020.

S Glazerman, M Rosenbaum, R Sandino, and L Shaughnessy. Remote surveying in a pandemic: handbook. *Innovation for Poverty Action*, 2020.

Sydney Gourlay, Talip Kilic, Antonio Martuscelli, Philip Wollburg, and Alberto Zezza. High-frequency phone surveys on covid-19: Good practices, open questions. *Food Policy*, 105: 102153, 2021.

Abigail R Greenleaf, Aliou Gadiaga, Georges Guiella, Shani Turke, Noelle Battle, Saifuddin Ahmed, and Caroline Moreau. Comparability of modern contraceptive use estimates between a face-to-face survey and a cellphone survey among women in burkina faso. *PloS one*, 15(5): e0231819, 2020.

Abigail R Greenleaf, Gerald Mwima, Molibeli Lethoko, Martha Conkling, George Keefer, Christiana Chang, Natasha McLeod, Haruka Maruyama, Qixuan Chen, Shannon M Farley, et al. Participatory surveillance of covid-19 in lesotho via weekly calls: Protocol for cell phone data collection. *JMIR Research Protocols*, 10(9):e31236, 2021.

Savanna Henderson and Michael Rosenbaum. Remote surveying in a pandemic: research synthesis. *Innovation for Poverty Action*, 2020.

Frauke Kreuter, Stanley Presser, and Roger Tourangeau. Social desirability bias in cati, ivr, and web surveys: the effects of mode and question sensitivity. *Public opinion quarterly*, 72 (5):847–865, 2008.

Charles Q Lau, Alexandra Cronberg, Leenisha Marks, and Ashley Amaya. In search of the optimal mode for mobile phone surveys in developing countries. a comparison of ivr, sms, and cati in nigeria. In *Survey Research Methods*, volume 13, pages 305–318, 2019.

Gerty JLM Lensvelt-Mulders, Joop J Hox, Peter GM Van der Heijden, and Cora JM Maas. Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research*, 33(3):319–348, 2005.

Benjamin Leo, Robert Morello, Jonathan Mellon, Tiago Peixoto, and Stephen T Davenport. Do mobile phone surveys work in poor countries? *Center for Global Development Working Paper*, (398), 2015.

K. F. Lock. *Rerandomization to improve covariate balance in randomized experiments Ph.D. thesis*. PhD dissertation, Harvard Univ. Cambridge, MA., 2011.

Elisa M Maffioli. Collecting data during an epidemic: A novel mobile phone research method. *Journal of International Development*, 32(8):1231–1255, 2020.

David S Metzger, Beryl Koblin, Charles Turner, Helen Navaline, Francesca Valenti, Sarah Holte, Michael Gross, Amy Sheon, Heather Miller, Philip Cooley, et al. Randomized controlled trial of audio computer-assisted self-interviewing: utility and acceptability in longitudinal studies. *American journal of epidemiology*, 152(2):99–106, 2000.

NIPS Pakistan and ICF. Pakistan Demographic and Health Survey 2017-18. Technical report, NIPS/Pakistan and ICF, Islamabad, Pakistan, 2019. URL `http://dhsprogram.com/pubs/pdf/FR354/FR354.pdf`.

Rachael Phadnis, Champika Wickramasinghe, Juan Carlos Zevallos, Stacy Davlin, Vindya Kumarapeli, Veronica Lea, Juliette Lee, Udara Perera, Francisco Xavier Solórzano, and Juan Francisco Vásconez. Leveraging mobile phone surveys during the covid-19 pandemic in ecuador and sri lanka: Methods, timeline and findings. *Plos one*, 16(4):e0250171, 2021.

Catherine Porter, Marta Favara, Alan Sánchez, and Douglas Scott. The impact of covid-19 lockdowns on physical domestic violence: Evidence from a list randomization experiment. *SSM-population health*, 14:100792, 2021.

Bryn Rosenfeld, Kosuke Imai, and Jacob N Shapiro. An empirical validation study of popular survey methodologies for sensitive questions. *American Journal of Political Science*, 60(3): 783–802, 2016.

Jilian A Sacks, Elizabeth Zehe, Cindil Redick, Alhoussaine Bah, Kai Cowger, Mamady Camara, Aboubacar Diallo, Abdel Nasser Iro Gigo, Ranu S Dhillon, and Anne Liu. Introduction of mobile health tools to support ebola surveillance and contact tracing in guinea. *Global Health: Science and Practice*, 3(4):646–659, 2015.

Stergiani Tsoli, Stephen Sutton, and Aikaterini Kassavou. Interactive voice response interventions targeting behaviour change: a systematic literature review with meta-analysis and meta-regression. *BMJ Open*, 8:e018974, 2017. doi: 10.1136/bmjopen-2017-018974.

Stergiani Tsoli, Stephen Sutton, and Aikaterini Kassavou. Interactive voice response interventions targeting behaviour change: a systematic literature review with meta-analysis and meta-regression. *BMJ open*, 8(2):e018974, 2018.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Alberto Zezza, Antonio Martuscelli, Philip Wollburg, Sydney Gourlay, and Talip Kilic. High-frequency phone surveys on covid-19: Good practices, open questions. *Food Policy*, 105: 102153, 2021.

A Appendix

A.1 Tables

Table A.1: Survey-mode variation - Balance table, **estimation sample**: T1 & T2 versus C

	(1) Control		(2) T1		(3) T2		T-test Difference	
Variable	N/[Clusters]	Mean/SE	N/[Clusters]	Mean/SE	N/[Clusters]	Mean/SE	(1)-(2)	(1)-(3)
Individual with more than 5 imputations at baseline	4430	0.123	1899	0.114	5688	0.121	0.009	0.002
	[1036]	(0.008)	[849]	(0.010)	[1074]	(0.008)		
Female	4430	0.381	1899	0.374	5688	0.379	0.007	0.001
	[1036]	(0.015)	[849]	(0.017)	[1074]	(0.014)		
Age	4430	38.294	1899	38.101	5688	38.002	0.193	0.293
	[1036]	(0.201)	[849]	(0.288)	[1074]	(0.181)		
Household size	4430	8.893	1899	9.006	5688	9.080	-0.112	-0.186*
	[1036]	(0.094)	[849]	(0.121)	[1074]	(0.094)		
Household owns either land or livestock	4430	0.649	1899	0.659	5688	0.651	-0.010	-0.002
	[1036]	(0.011)	[849]	(0.013)	[1074]	(0.010)		
Income in the past 7 days w	4430	872.126	1899	898.544	5688	898.797	-26.418	-26.670
	[1036]	(31.040)	[849]	(38.821)	[1074]	(30.890)		
Worked in the past 7 days	4430	0.427	1899	0.432	5688	0.436	-0.005	-0.008
	[1036]	(0.012)	[849]	(0.015)	[1074]	(0.012)		
Village avg. Respondent participates and doesn't stop interview	4430	0.478	1899	0.473	5688	0.479	0.004	-0.001
	[1036]	(0.008)	[849]	(0.009)	[1074]	(0.008)		
Somehousehold members fell sick past 14 days]	4430	0.150	1899	0.146	5688	0.144	0.004	0.006
	[1036]	(0.007)	[849]	(0.009)	[1074]	(0.006)		
# of household members with COVID-like symptoms [past 14 days]	4430	0.093	1899	0.091	5688	0.098	0.002	-0.005
	[1036]	(0.007)	[849]	(0.009)	[1074]	(0.007)		
# of days household members traveled for visit [past 7 days]	4430	1.115	1899	1.128	5688	1.131	-0.013	-0.015
	[1036]	(0.056)	[849]	(0.067)	[1074]	(0.057)		
# of days worked outside home [past 7 days]	4430	2.094	1899	2.040	5688	2.102	0.054	-0.008
	[1036]	(0.070)	[849]	(0.086)	[1074]	(0.068)		
F-test of joint significance (F-stat)							0.848	1.021
F-test, number of observations							6329	10118

Note: This table shows balance statistics across the randomization variables for the comparison of pure IVR and Mixed group versus pure CATI in the technology-based intervention. Columns 1, 3 and 5 display the number of observations. The number of clusters (villages) are displayed in brackets. Columns 2, 4 and 6 display the mean of the baseline variables in the two groups. Standard deviations are displayed in parentheses. Columns 7-8 show the estimated difference in means which is obtained from regressing the variable of interest on the treatment dummy. Standard errors are clustered at the village level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Variables marked with (w) were winsorized. All variables were scrutinized for outliers using the $90_th/50_th$ ratio. Based on the ratio values, variables were categorized into high (ratio $\geq$ 20),medium (ratio $\geq$= 10 and $\leq$ 20), and moderate (ratio $>$ 0.5 and $\leq$ 10).

Table A.2: Total attempts per contact by wave for enumerator-led calls

Total attempts per individual	Wave 1	Wave 2	Wave 3
1	4,321	5,494	5,404
2	2,346	2,430	2,642
3	5,049	6,123	5,082
4	592	664	980
5	190	65	295
6	-	12	174
7	-	-	-
8	-	-	32
Total	12,498	14,788	14,609

Notes: Table A.2 presents the total attempts made per individual in each wave of data collection.

Table A.3: Questions asked in IVR interviews

Module ID	Module	Question	Answer choices	Outcome	Enumerator-led questionnaire
Introductory messages			**Case 1**: I am Imam ..., Starting with the Holy name of Allah, The Most Beneficent, the Merciful. In a previous interview, you agreed to be contacted again for short interviews about you and your household members' health. I encourage your participation in these interviews as these will contribute to a greater understanding of the health status of households in Pakistan and hence in better assessing your needs.	Pick-up the call	
			Case 2: My name is Dr ... I am a medical Dr in the field of .... In a previous interview, you agreed to be contacted again for short interviews about your and your household members' health. I encourage your participation in these interviews as these will contribute to a greater understanding the health status of households in Pakistan and hence in better assessing your needs.		
			Case 3 and 4 My name is ... and I work for (NGO name). In a previous interview, you agreed to be contacted again for short interviews about your and your household members' health. I encourage your participation in these interviews as these will contribute to a greater understanding the health status of households in Pakistan and hence in better assessing your needs.		
	Call status	This service is completely free of cost.	Press 1 if you like to proceed. Press 2 and end the call if you like to be called at a different time. Press 9 and end the call if you refuse to proceed.	Consent to interview	
r	Health status	What is your current health status?	Press 1 for good health, Press 2 for fair health, Press 3 for bad health, Press 8 if you do not know, Press 9 if you do not want to answer		x
		How many people including yourself in your household fell sick in the last 14 days?	Press 0 if no one; Press 1 if 1 member; Press 2 if 2 members; Press 3 if 3 members; Press 4 if 4 members; Press 5 if 5 members or more; Press 8 if you do not know; Press 9 if you do not want to answer	Respond to all sensitive questions, Respond to all sensitive health questions	x
		How many of the individuals that got sick had any, some or all of symptoms from the following list: Fever, Fatigue, and/ or a dry cough	Press 0 if no one; Press 1 if one member; Press 2 if 2 members; Press 3 if 3 members; Press 4 if 4 members; Press 5 if 5 members or more; Press 8 if you do not know; Press 9 if you do not want to answer; This question is not applicable if you had no sick family members in the past 14 days, please press 7 if it is the case	Respond to all sensitive questions, Respond to all sensitive health questions	x
		How many of the sick individuals were ever chronically ill before?	Press 0 if no one; Press 1 if one member; Press 2 if 2 members; Press 3 if 3 members; Press 4 if 4 members; Press 5 if 5 members or more; Press 8 if you do not know; Press 9 if you do not want to answer; This question is not applicable if you had no sick family members in the past 14 days, please press 7 if it is the case		x
		How many of the sick individuals have now fully recovered?	Press 0 if no one; Press 1 if 1 member; Press 2 if 2 members; Press 3 if 3 members; Press 4 if 4 members; Press 5 if 5 members or more; Press 8 if you do not know; Press 9 if you do not want to answer; This question is not applicable if you had no sick family members in the past 14 days, please press 7 if it is the case		x
sd	Mobility and social distancing behavior	In the last 7 days, for how many days did you leave your village to visit another place or a person outside?	Press 0 if you did not leave village; Press 1 if you left the village for 1 day a week; Press 2 if you left the village for 2 days a week; Press 3 if you left the village for 3 days a week; Press 4 if you left the village for 4 days a week; Press 5 if you left the village for 5 days a week; Press 6 if you left the village for 6 days a week; Press 7 if you left the village every day; Press 8 if you do not know; Press 9 if you do not want to answer	Respond to all sensitive questions, Respond to all sensitive non-health questions	x
		In the last 7 days, have you attended any social gatherings, for example family gatherings, visits, a wedding, a community meeting, or a religious festival?	Press 0 if no; Press 1 if yes; Press 8 if you do not know; Press 9 if you do not want to answer	Respond to all sensitive questions, Respond to all sensitive non-health questions	x
i	Labor supply	Have you done any paid work outside your home in the past 7 days?	Press 0 if you did not work outside; Press 1 if you worked outside 1 day a week; Press 2 if you worked outside 2 days a week; Press 3 if you worked outside 3 days a week; Press 4 if you worked outside 4 days a week; Press 5 if you worked outside 5 days a week; Press 6 if you worked outside 6 days a week; Press 7 if you worked outside 7 days a week; Press 8 if you do not know; Press 9 if you do not want to answer; This question is not applicable if you did not work at all in the past 14 days, please press 7 if it is the case.	Respond to all sensitive questions, Respond to all sensitive non-health questions	x
k	Covid knowledge	What do you think are the three common symptoms of the on-going corona virus?	Press 1 for fever, vomiting, and fatigue; Press 2 for nausea, vomiting and diarrhea; Press 3 for fatigue, fever, and dry cough; Press 4 for none of the above; Press 5 for all of the above; Press 8 if you do not know; Press 9 if you do not want to answer		x
End message			We are very grateful for your precious time and cooperation. You will be receiving a similar call from us again within 12-14 days. Take care. Goodbye!	Complete interview	

Note: Table A.3 summarizes all questions that were asked during IVR interviews.

Table A.4: Tertiary outcomes, **conditional on receiving consent and the question being asked**

#	(1) Outcome	(2) Description
Panel A - Tertiary outcomes		
8	Respond to health question 1	Fraction of sample, that completed the interview, responding to the question on the number of household members who fell sick in the past 14 days
9	Respond to health question 2	Fraction of sample, that completed the interview, responding to the question on the number of household members that had COVID-like symptoms in the past 14 days
10	Respond to non-health question 1	Fraction of sample, that completed the interview, responding to the question on whether they left the village in the past 7 days
11	Respond to non-health question 2	Fraction of sample, that completed the interview, responding to the question on whether they have attended a social gathering in the past 7 days
12	Respond to non-health question 3	Fraction of sample, that completed the interview, responding to the question on whether they have done a paid work outside home in the past 7 days
13	# of total questions answered	Number of total questions answered (out of 9 total) conditional on sample, that completed the interview
14	# of total sensitive questions answered	Number of total sensitive questions answered (out of 5 total) conditional on sample, that completed the interview
15	# of total health questions answered	Number of total health questions answered (out of 2 total) conditional on sample, that completed the interview
16	# of total non-health questions answered	Number of total non-health questions answered (out of 3 total) conditional on sample, that completed the interview

Note: Table A.4 displays tertiary outcome variables of interest. Outcomes 8 to 12 are dummy variables. Outcomes 13 to 16 are count variables. Related Tables: Table 3.

Table A.5: Summary statistics of primary and secondary outcomes

	(I) All		(II) C		(III) T1		(IV) T2			
							Enumerator-led		IVR	
	Mean (1)	Obs. (2)	Mean (3)	Obs. (4)	Mean (5)	Obs. (6)	Mean (7)	Obs. (8)	Mean (9)	Obs. (10)
Panel A - Primary outcomes										
Pick-up the call	0.66	30,306	0.79	10,353	0.34	5,695	0.78	11,113	0.37	3,145
Consent to interview, **conditional on call pick-up**	0.83	19,956	0.97	8,180	0.07	1,910	0.96	8,714	0.05	1,152
Complete interview, **conditional on consent**	0.95	17,096	0.96	8,180	0.08	142	0.96	8,714	0.07	60
Respond to all questions, **conditional on consent**	0.90	17,096	0.91	8,180	0.06	142	0.91	8,714	0.03	60
Panel B - Secondary outcomes, ***conditional on consent and question asked in both survey modes***										
Respond to all sensitive questions	0.97	8,895	0.99	4,170	0.07	142	0.99	4,523	0.05	60
Respond to all sensitive health questions	0.90	1,214	1.00	514	0.43	142	1.00	498	0.35	60
Respond to all sensitive non-health questions	0.97	8,385	0.99	3,913	0.07	142	0.99	4,270	0.08	60

Note: Table A.5 contains descriptive statistics of the primary and secondary outcomes on the estimation sample, pooling across the three FU waves. Columns (1) and (2) contain information for all interviews. Columns (3) and (4) summarize outcomes of C, columns (5) and (6) of T1, and columns (7) and (8) of T2. Columns (1), (3), (5), and (7) display the variable mean value. Columns (2), (4), (6), and (8) present the number of observations. Related tables: Table A.12.

Table A.7: Survey-mode variation - Estimates for "response rate" indicators, single sensitive questions, **conditional on consent and question being asked**

	(1) Respond to health question 1	(2) Respond to health question 2	(3) Respond to non-health question 1	(4) Respond to non-health question 2	(5) Respond to non-health question 3
T1	-0.405***	-0.529***	-0.785***	-0.901***	-0.895***
	(0.042)	(0.044)	(0.036)	(0.027)	(0.026)
T2	-0.048***	-0.057***	-0.020***	-0.024***	-0.027***
	(0.009)	(0.010)	(0.003)	(0.004)	(0.004)
Statistical tests (p-values)					
Equal	0.000	0.000	0.000	0.000	0.000
Jointly zero	0.000	0.000	0.000	0.000	0.000
Mean	0.998	0.998	0.998	0.994	0.994
SD	0.044	0.044	0.039	0.074	0.075
Obs.	1,214	1,214	4,343	4,343	4,244
R^2	0.243	0.339	0.536	0.560	0.542

Note: Sample restricted to pure IVR and enumerator-led calls only. Sample description for indicator 1- Sample that responded to call; Sample description for indicators 2 and 3- sample that responded to call and consented to interview. Panel A - Control group = all male-voice recorded IVR calls. Panel B - Control group = male enumerator recorded IVR calls. Control variables: randomization variables, mixed arm, wave and IP dummies. Health question 1 asks about any sickness cases in the household in the past 14 days. Health question 2 asks about COVID-like symptoms for household members who were sick in the past 14 days. Non-health question 1 asks the respondent to report whether they left the village in past 7 days. Non-health question 2 asks whether the respondents attended any social gatherings in the past 7 days. Non-health question 3 asks whether the respondent did any paid work outside home in past 7 days. Significance levels are indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are shown for each estimation below the coefficient.

Table A.6: Survey-mode variation - Response (overall and item), consent and completion upon consented sample

	(1) Pick-up the call	(2) Consent to interview	(3) Complete interview	(4) Respond to all questions
T1	-0.451***	-0.888***	-0.947***	-0.907***
	(0.009)	(0.007)	(0.003)	(0.004)
T2	-0.097***	-0.108***	-0.112***	-0.113***
	(0.006)	(0.004)	(0.004)	(0.005)
Statistical tests (p-values)				
Equal	0.000	0.000	0.000	0.000
Jointly zero	0.000	0.000	0.000	0.000
Mean	0.790	0.965	0.956	0.912
SD	0.407	0.183	0.204	0.283
Obs.	30,306	19,956	19,956	19,956
R^2	0.128	0.441	0.467	0.369

Note: Sample description for indicator 1- full sample, indicator 2- only sample that picked up the call, indicator 3- only sample that picked up the call and consented to interview, indicator 4- only sample that picked up the call, consented to interview and were asked all modules named r, i, sd and k. The control group are pure enumerator-led calls. Control variables: randomization variables, wave and IP dummies. Significance levels are indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are shown for each estimation below the coefficient. Statistical tests refer to the difference in the coefficients of the two treatment arms.

Table A.8: Survey-mode variation - Estimates for "length" indicators, **conditional on consent and question being asked**

	(1) # of total questions answered	(2) # of total sensitive questions answered	(3) # of total health questions answered	(4) # of total non-health questions answered
T1	0.668***	-0.209	-0.934***	-1.095***
	(0.238)	(0.136)	(0.079)	(0.080)
T2	-0.012	-0.022	-0.105***	-0.017
	(0.020)	(0.015)	(0.018)	(0.011)
Statistical tests (p-values)				
Equal	0.004	0.168	0.000	0.000
Jointly zero	0.015	0.120	0.000	0.000
Mean	2.214	1.654	1.996	1.501
SD	1.237	0.689	0.088	0.507
Obs.	16,348	8,895	1,214	8,385
R^2	0.009	0.008	0.326	0.068

Note: Sample consists of pure IVR and enumerator-led calls only. Sample description for indicator 1- sample that responded to the call and consented to the interview, Sample description for indicator 2- Sample that responded to call, consented to interview and were asked modules r, sd or i. Sample description for indicator 3- Sample that responded to call, consented to interview and were asked module r. Sample description for indicator 4- Sample that responded to call, consented to interview and were asked module i or sd. Control group = pure enumerator-led calls. Controls: randomization variables, wave and IP dummies. Significance levels are indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are shown for each estimation below the coefficient.

Table A.9: Attempts made outside "official" calling times for enumerator led calls

Wave	Share	Total
F0	14.11	90,770
F1	14.48	20,175
F2	14.49	12,498
F3	17.21	14,788
F4	18.01	14,609

Notes: Table A.9 displays the share of call attempts made outside the "official" calling time (8am to 5pm), sorted by data collection wave.

Table A.10: Sample composition over time

				Difference			
	Base. N	Base. mean	Comparison	FU2	FU3	FU4	Endline
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	12652	0.379	Enumerator-led - IVR	-0.123*	-0.132*	-0.137*	0.010
			Enumerator-led - Mixed	-0.008	-0.020*	-0.023*	-0.007
			IVR - Mixed	0.115*	0.112	0.114	-0.016
18 - 50 years	12652	0.877	Enumerator-led - IVR	-0.108***	-0.098***	-0.142***	0.007
			Enumerator-led - Mixed	-0.020**	-0.015*	-0.018**	-0.016*
			IVR - Mixed	0.088***	0.083**	0.124***	-0.023**
No degree	12578	0.463	Enumerator-led - IVR	0.070	0.052	0.029	-0.013
			Enumerator-led - Mixed	-0.024*	-0.029**	-0.025*	-0.014
			IVR - Mixed	-0.094	-0.081	-0.054	-0.002
Primary school education	12638	0.480	Enumerator-led - IVR	-0.060	0.000	-0.074	0.010
			Enumerator-led - Mixed	0.022	0.019	0.012	0.005
			IVR - Mixed	0.082	0.018	0.086	-0.005
HH owns livestock or land	12652	0.651	Enumerator-led - IVR	0.038	0.031	0.048	-0.002
			Enumerator-led - Mixed	0.008	0.003	0.001	0.001
			IVR - Mixed	-0.030	-0.028	-0.047	0.003
Respondent income at BL	12652	884.785	Enumerator-led - IVR	32.250	165.870	231.059	10.181
			Enumerator-led - Mixed	0.604	-22.783	24.810	8.501
			IVR - Mixed	-31.646	-188.652	-206.249	-1.679
Respondent worked (prev. 7d.) at BL	12652	0.430	Enumerator-led - IVR	-0.034	0.075	0.112	-0.002
			Enumerator-led - Mixed	-0.007	-0.008	0.003	-0.004
			IVR - Mixed	0.027	-0.083	-0.110	-0.002
Up to 5 HH members	12652	0.177	Enumerator-led - IVR	0.044	0.033	0.087**	0.004
			Enumerator-led - Mixed	0.018*	0.011	0.010	0.005
			IVR - Mixed	-0.026	-0.022	-0.077*	0.001

Note: We display mean differences for a set of sample characteristics across treatment arms and by survey wave. Column (1) shows the sample size, column (2) shows the overall baseline mean. Column (3) shows which treatment arms are compared against each other. Columns (4) to (7) display the mean differences between the respective treatment arms at FU2, FU3, FU4, and endline, respectively. T-test significance levels are indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.11: Cost per interview and necessary sample frame by survey mode

	robocalls		Enumerator-led calls		Mixed		Ratio	
Number of interviews	Cost per interview	Necessary sample frame	Cost per interview	Necessary sample frame	Cost per interview	Necessary sample frame	Robo/enumerator cost	Mixed/enumerator cost
1500	31.35	715500	11.52	2082	13.26	2805	2.72	1.15
2000	30.92	954001	8.75	2776	10.06	3740	3.53	1.15
3000	30.49	1431000	5.98	4164	6.86	5610	5.1	1.15
5000	30.14	2385000	3.76	6939	4.3	9349	8.02	1.14
10000	29.89	4770000	2.1	13878	2.38	18698	14.25	1.13
20000	29.76	9540000	1.27	27755	1.42	37395	23.51	1.12

Note: *Number of interviews* refers to the amount of completed interviews. *Cost per interview* refers to the cost of one completed interview in US dollars, accounting for the necessary attempts to finalize one interview, i.e., accounting for response rates. *Necessary sample frame* shows the size of the respondent pool necessary to reach the respective amount of completed interviews given the response rates. *Robo/enumerator cost* shows the ratio of costs for one completed robo interview as compared to one completed enumerator-led interview, *Mixed/enumerator cost* shows the ratio of costs for one completed interview in the treatment arm where both modes are used compared to one completed enumerator-led interview.

Table A.12: Summary statistics of tertiary outcomes, conditional on consent and question being asked

	(I) All		(II) C		(III) T1		(IV) T2			
							Enumerator-led		IVR	
	Mean (1)	Obs. (2)	Mean (3)	Obs. (4)	Mean (5)	Obs. (6)	Mean (7)	Obs. (8)	Mean (9)	Obs. (10)
Respond to health question 1	0.93	1,214	1.00	514	0.58	142	1.00	498	0.52	60
Respond to health question 2	0.91	1,214	1.00	514	0.46	142	1.00	498	0.43	60
Respond to non-health question 1	0.96	4,343	1.00	1,984	0.21	142	1.00	2,157	0.27	60
Respond to non-health question 2	0.95	4,343	0.99	1,984	0.09	142	0.99	2,157	0.10	60
Respond to non-health question 3	0.95	4,244	0.99	1,929	0.10	142	0.99	2,113	0.12	60
# of total questions answered	2.21	16,348	2.21	7,819	2.89	142	2.20	8,327	2.90	60
# of total sensitive questions answered	1.64	8,895	1.65	4,170	1.44	142	1.63	4,523	1.43	60
# of total health questions answered	1.83	1,214	2.00	514	1.04	142	2.00	498	0.95	60
# of total non-health questions answered	1.47	8,385	1.50	3,913	0.40	142	1.50	4,270	0.48	60

Note: Table A.12 contains summary statistics of the tertiary outcomes on the estimation sample, pooling across the three FU waves. Columns (1) and (2) contain information for all interviews. Columns (3) and (4) summarize outcomes of C, columns (5) and (6) of T1, columns (7) and (8) for enumerator-led calls of T2, and columns (9) and (10) for IVR calls of T2. Columns (1), (3), (5), (7), and (9) display the variable mean value. Columns (2), (4), (6), (8), and (10) present the number of observations.
Related tables: Table A.5.

Table A.13: Pick-up rate for first call attempts and "official" calling time

	(I) All		(II) C		(III) T1		(IV) T2			
							Enumerator-led		IVR	
	Mean (1)	Obs. (2)	Mean (3)	Obs. (4)	Mean (5)	Obs. (6)	Mean (7)	Obs. (8)	Mean (9)	Obs. (10)
First attempt	0.55	30,306	0.65	10,353	0.34	5,695	0.64	11,113	0.37	3,145
Fisrt attempt within "official" calling time	0.49	30,306	0.56	10,353	0.34	5,695	0.55	11,113	0.37	3,145

Note: Table A.13 contains summary statistics of the pick-up rate with alternative definitions, pooling across the three FU waves. Row (1) uses only the first call attempts to calculate the pick-up rate, and row (2) further restricts to the "official" calling time (8am to 5pm). Columns (1) and (2) contain information for all interviews. Columns (3) and (4) summarize outcomes of C, columns (5) and (6) of T1, columns (7) and (8) for enumerator-led calls of T2, and columns (9) and (10) for IVR calls of T2. Columns (1), (3), (5), (7), and (9) display the variable mean value. Columns (2), (4), (6), (8), and (10) present the number of observations.

A.2 Figures

Figure A.1: Item response for all IVR questions, across waves

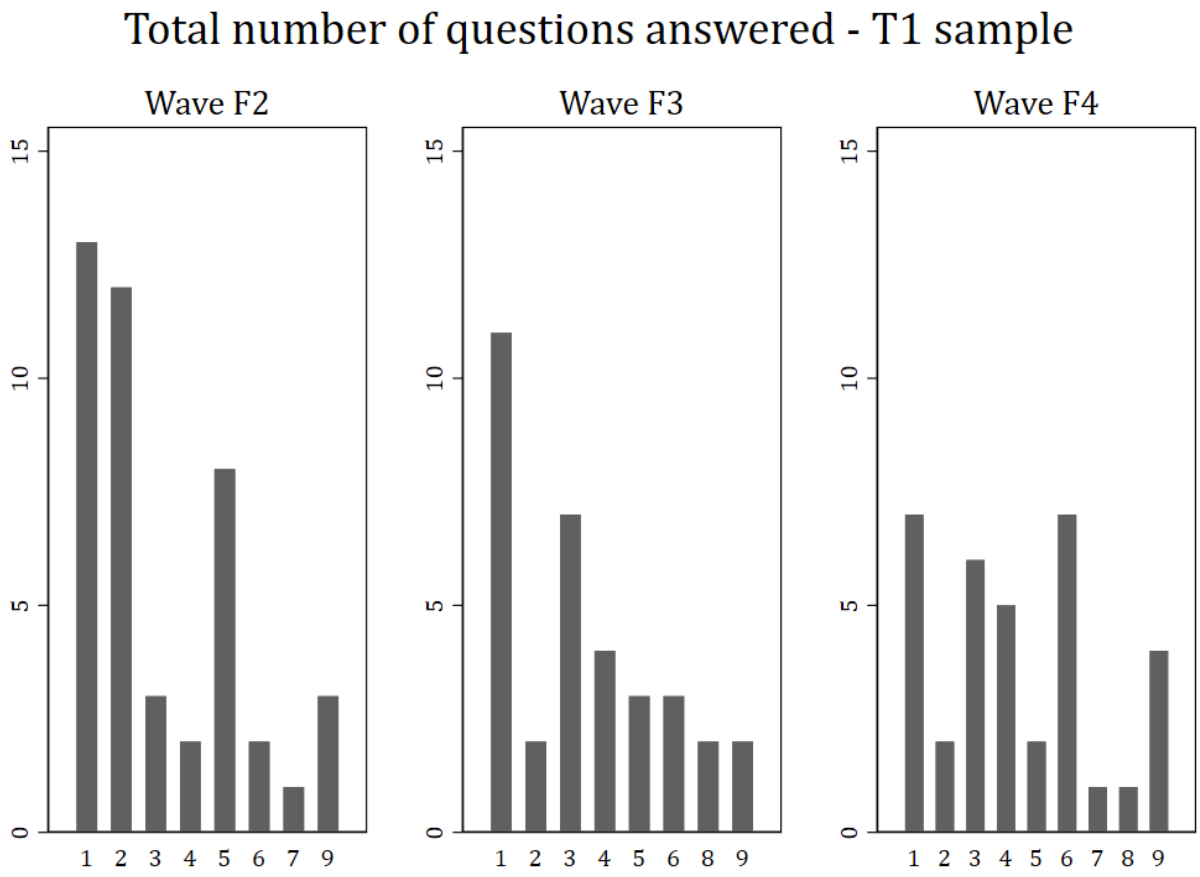


Figure A.2: Response rate for T1 and C arms across waves

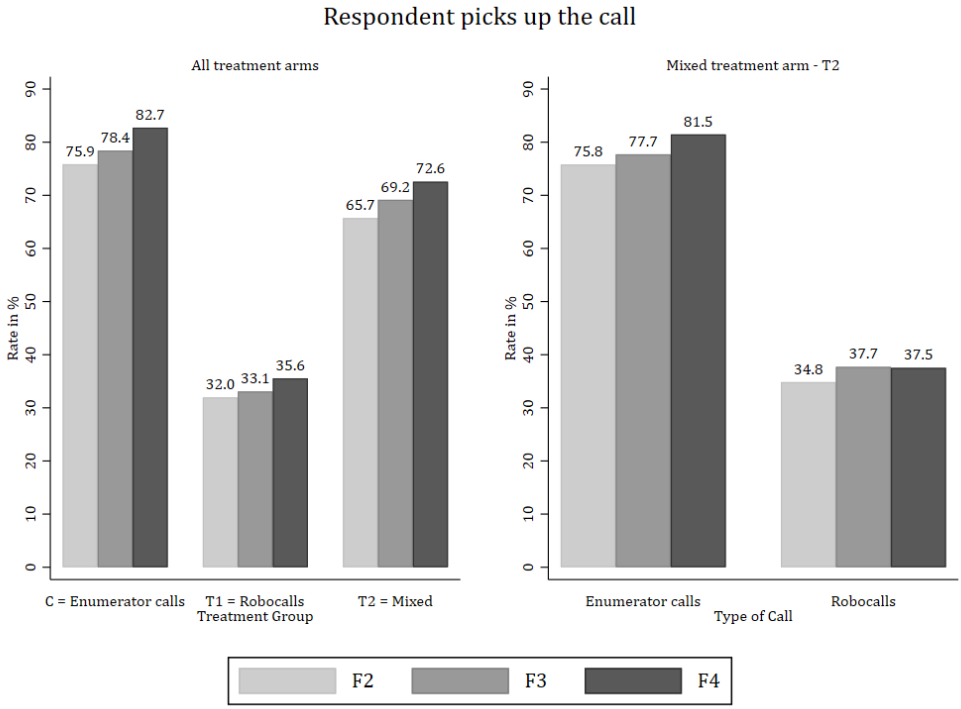


Figure A.3: Consent rate for T1 and C arms across waves

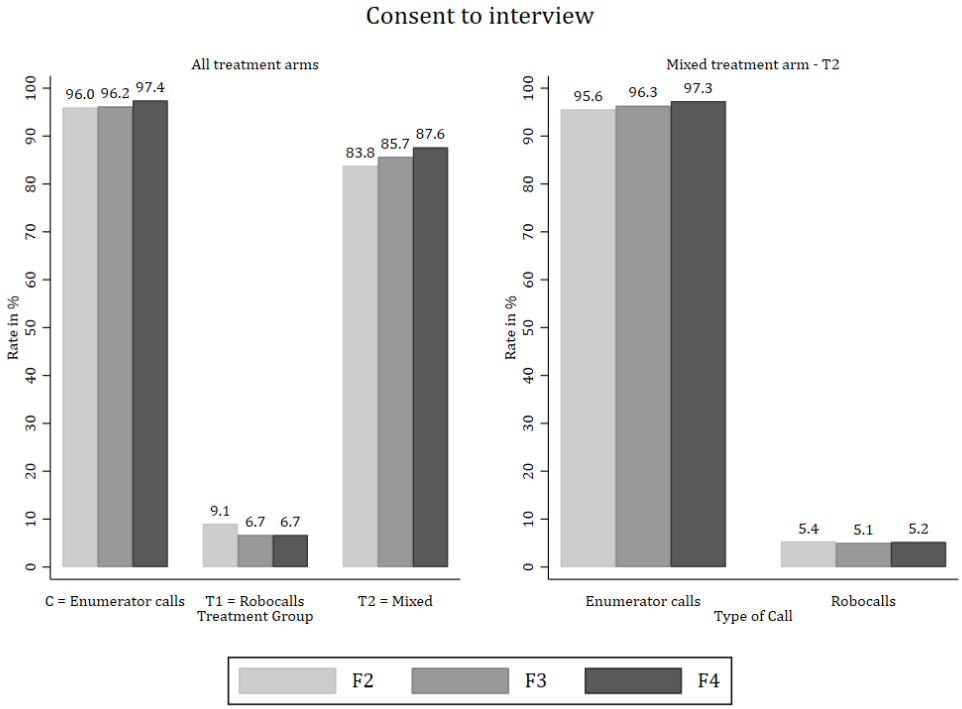


Figure A.4: Completion rate for T1 and C arms across waves

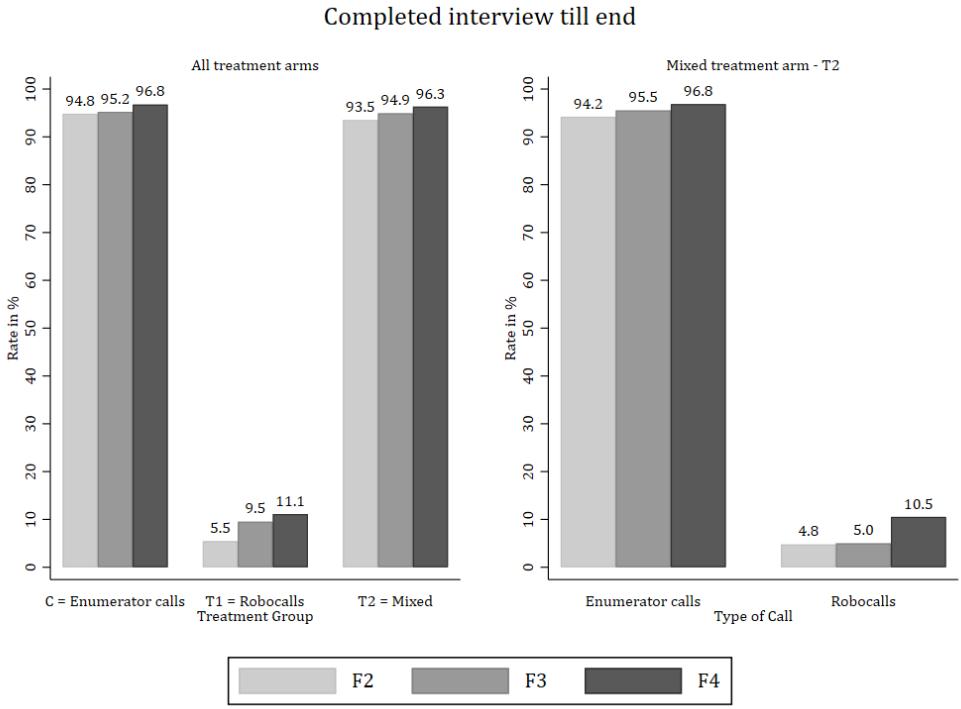
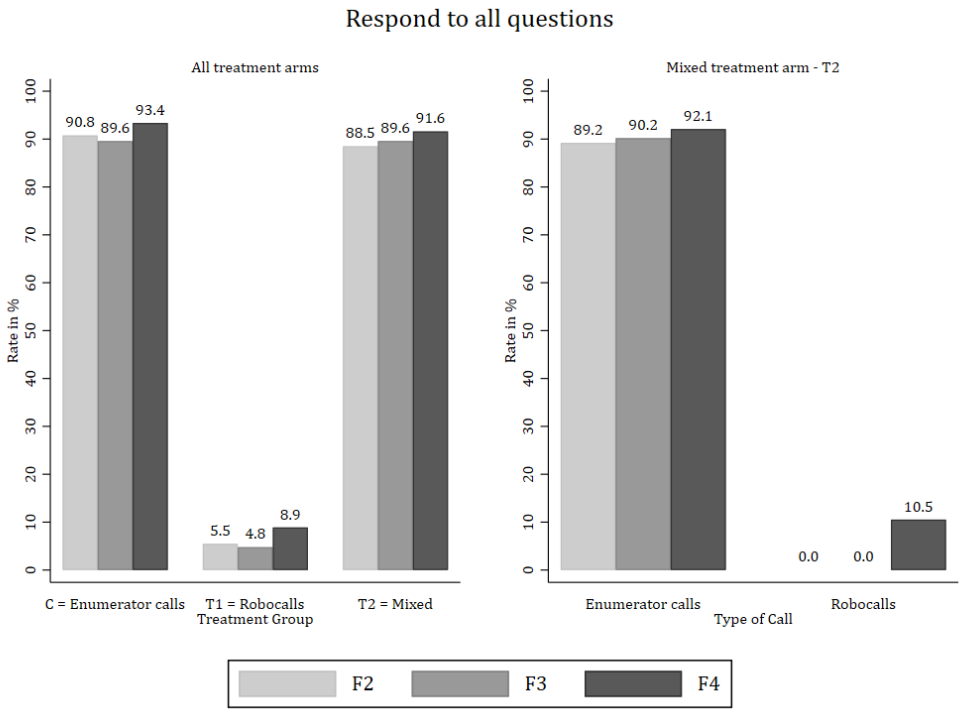


Figure A.5: Response rate to all questions for T1 and C arms across waves



B Online Appendix

B.1 Design

Randomization and estimation sample. Randomization for both random variations, survey-mode and framing variation, took place after the baseline survey, with individuals that has successfully completed a baseline interview. For both experimental variations, randomization was performed at the individual level, with stratification by NGO, followed by a re-randomization procedure to achieve balance on baseline values.[33]

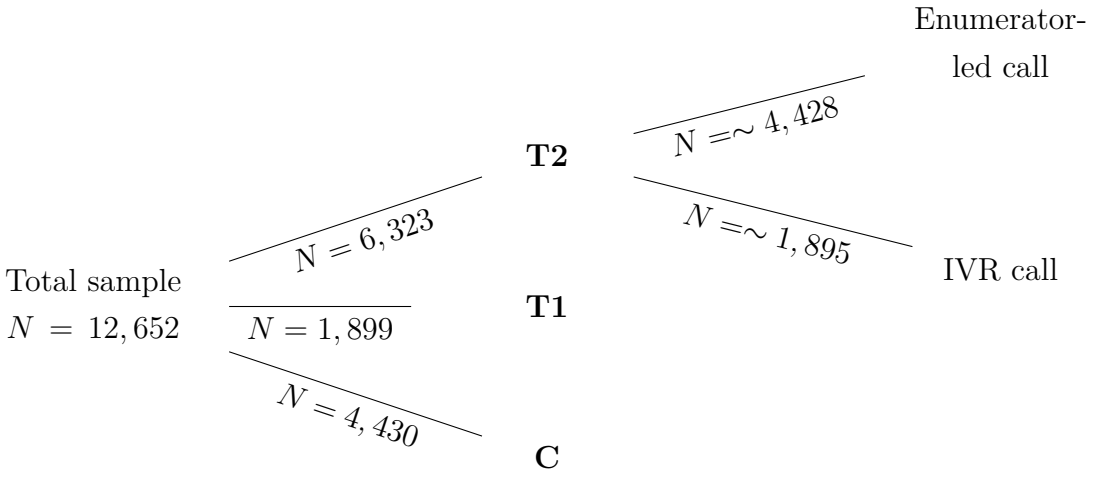
The randomization happened in the following order: First, for the *survey-mode variation*, individuals were randomly allocated into three groups - enumerator-led interview (C), T1, and T2. Second, for the *framing variation*, individuals in T1, and in the IVR part of T2 were randomized into four equally-sized groups, allocated to the four different introductory messages (male enumerator, female enumerator, doctor, religious leader). The random allocation of these introductory messages stayed constant across the data collection waves, i.e., across all three FU waves, one individual always received the same introductory message. Finally, for the *survey-mode variation*, the sample of T2 was split into ten equally sized groups. Groups 1 to 3 were allocated IVR calls in FU wave 1, groups 4 to 6 in wave 3, groups 7 to 9 were allocated IVR calls in FU wave 3, and groups 10 to 2 in FU wave 4.[34] Due to the technical issues in FU wave 1 with the service providers of IVR calls, individuals in FU wave 1 received only enumerator-led calls. Thus, group 3 never received an IVR call throughout the study period. For this reason, group 3 (10% of the sample of T2) is dropped. This results in the reduction of the sample from $N = 12,652$ (baseline sample) to $N = 12,017$ (estimation sample).

Figure O.1 reports the design of the survey-mode variation with the baseline sample (before the drop of group 10).

[33] We employed multivariate balance checks using Wilk's λ statistics (Lock 2011). Following Ciolino et al. [2019], a randomization was considered acceptable whenever the Wilk's λ statistics surpassed the threshold value 0.30. If this threshold was not exceeded, the randomization was repeated.

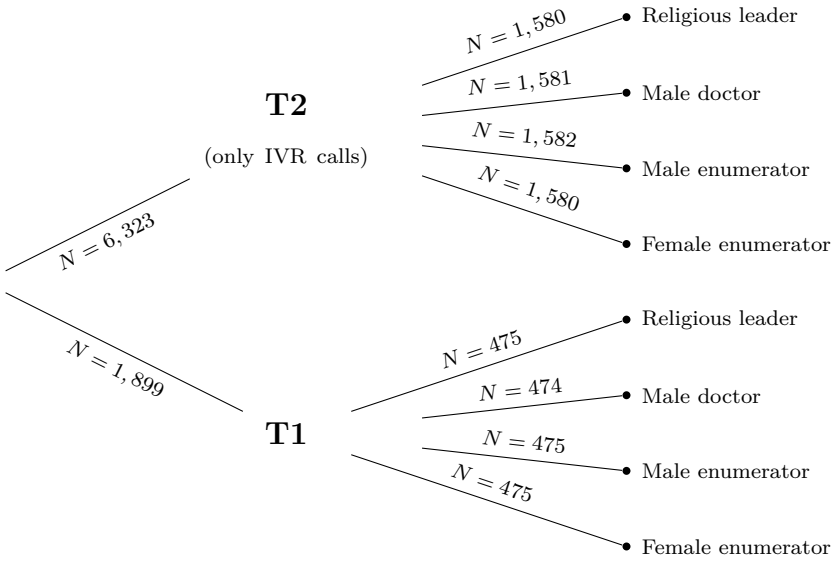
[34] Between each FU wave, *T2* individuals randomly assigned to IVR calls in wave $t-1$ were replaced to avoid that a given individual may receive two consecutive IVR calls.

Figure O.1: Survey-mode variation - Impact evaluation design (baseline sample)



Note: Figure O.1 summarizes the impact evaluation design of the survey-mode variation with the baseline sample.

Figure O.2: Framing variation - Impact evaluation design (baseline sample)



Note: Figure O.2 summarizes the impact evaluation design of the survey-mode variation with the baseline sample.

B.2 Tables

Table O.1: Balance table, **original randomization sample** for survey-mode variation: T1 & T2 versus C

	(1) Control		(2) T1		(3) T2		T-test Difference	
Variable	N/[Clusters]	Mean/SE	N/[Clusters]	Mean/SE	N/[Clusters]	Mean/SE	(1)-(2)	(1)-(3)
Individual with more than 5 imputations at baseline	4430	0.123	1899	0.114	6323	0.121	0.009	0.002
	[1036]	(0.008)	[849]	(0.010)	[1086]	(0.007)		
Female	4430	0.381	1899	0.374	6323	0.379	0.007	0.002
	[1036]	(0.015)	[849]	(0.017)	[1086]	(0.014)		
Age	4430	38.294	1899	38.101	6323	38.019	0.193	0.275
	[1036]	(0.201)	[849]	(0.288)	[1086]	(0.179)		
Household size	4430	8.893	1899	9.006	6323	9.073	-0.112	-0.180*
	[1036]	(0.094)	[849]	(0.121)	[1086]	(0.093)		
Household owns either land or livestock	4430	0.649	1899	0.659	6323	0.650	-0.010	-0.002
	[1036]	(0.011)	[849]	(0.013)	[1086]	(0.010)		
Income in the past 7 days w	4430	872.126	1899	898.544	6323	889.522	-26.418	-17.395
	[1036]	(31.040)	[849]	(38.821)	[1086]	(30.365)		
Worked in the past 7 days	4430	0.427	1899	0.432	6323	0.432	-0.005	-0.005
	[1036]	(0.012)	[849]	(0.015)	[1086]	(0.011)		
Village avg. Respondent participates and doesn't stop interview	4430	0.478	1899	0.473	6323	0.478	0.004	-0.001
	[1036]	(0.008)	[849]	(0.009)	[1086]	(0.008)		
Somehousehold members fell sick past 14 days]	4430	0.150	1899	0.146	6323	0.146	0.004	0.003
	[1036]	(0.007)	[849]	(0.009)	[1086]	(0.006)		
# of household members with COVID-like symptoms [past 14 days]	4430	0.093	1899	0.091	6323	0.099	0.002	-0.006
	[1036]	(0.007)	[849]	(0.009)	[1086]	(0.007)		
# of days household members traveled for visit [past 7 days]	4430	1.115	1899	1.128	6323	1.121	-0.013	-0.006
	[1036]	(0.056)	[849]	(0.067)	[1086]	(0.055)		
# of days worked outside home [past 7 days]	4430	2.094	1899	2.040	6323	2.076	0.054	0.018
	[1036]	(0.070)	[849]	(0.086)	[1086]	(0.067)		
F-test of joint significance (F-stat)							0.848	0.930
F-test, number of observations							6329	10753

Note: This table shows balance statistics across the randomization variables for the comparison of pure IVR and Mixed group versus pure CATI in the technology-based intervention. Columns 1, 3 and 5 display the number of observations. The number of clusters (villages) are displayed in brackets. Columns 2, 4 and 6 display the mean of the baseline variables in the two groups. Standard deviations are displayed in parentheses. Columns 7-8 show the estimated difference in means which is obtained from regressing the variable of interest on the treatment dummy. Standard errors are clustered at the village level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Variables marked with (w) were winsorized. All variables were scrutinized for outliers using the $90_th/50_th$ ratio. Based on the ratio values, variables were categorized into high (ratio $\geq$ 20),medium (ratio $\geq$= 10 and $\leq$ 20), and moderate (ratio $>$ 0.5 and $\leq$ 10).

Table O.2: Balance table, **original randomization sample** framing variation: female enumerator versus any male IVR

Variable	(1) Any male IVR call N/[Clusters]	(1) Any male IVR call Mean/SE	(2) Female enumerator IVR call N/[Clusters]	(2) Female enumerator IVR call Mean/SE	T-test Difference (1)-(2)
Individual with more than 5 imputations at baseline	6165	0.119	2057	0.120	-0.001
	[1075]	(0.008)	[882]	(0.009)	
Female	6165	0.379	2057	0.375	0.004
	[1075]	(0.014)	[882]	(0.016)	
Age	6165	38.004	2057	38.142	-0.138
	[1075]	(0.180)	[882]	(0.273)	
Household size	6165	9.052	2057	9.075	-0.023
	[1075]	(0.090)	[882]	(0.127)	
Household owns either land or livestock	6165	0.655	2057	0.646	0.009
	[1075]	(0.010)	[882]	(0.013)	
Income in the past 7 days w	6165	891.853	2057	890.866	0.987
	[1075]	(30.285)	[882]	(37.411)	
Worked in the past 7 days	6165	0.432	2057	0.432	-0.000
	[1075]	(0.012)	[882]	(0.014)	
Village avg. Respondent participates and doesn't stop interview	6165	0.476	2057	0.481	-0.006
	[1075]	(0.008)	[882]	(0.009)	
Somehousehold members fell sick past 14 days]	6165	0.148	2057	0.142	0.006
	[1075]	(0.006)	[882]	(0.009)	
# of household members with COVID-like symptoms [past 14 days]	6165	0.096	2057	0.098	-0.002
	[1075]	(0.007)	[882]	(0.010)	
# of days household members traveled for visit [past 7 days]	6165	1.112	2057	1.155	-0.043
	[1075]	(0.054)	[882]	(0.067)	
# of days worked outside home [past 7 days]	6165	2.068	2057	2.067	0.001
	[1075]	(0.068)	[882]	(0.082)	
F-test of joint significance (F-stat)					0.324
F-test, number of observations					8222

Note: This table shows balance statistics across the randomization variables for the comparison of female enumerator IVR calls with any male IVR calls in the framing-based intervention. Columns 1, 3 display the number of observations. The number of clusters (villages) are displayed in brackets. Columns 2, 4 display the mean of the baseline variables in the two groups. Standard deviations are displayed in parentheses. Column 5 shows the estimated difference in means which is obtained from regressing the variable of interest on the treatment dummy. Standard errors are clustered at the village level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Variables marked with (w) were winsorized. All variables were scrutinized for outliers using the $90_th/50_th$ ratio. Based on the ratio values, variables were categorized into high (ratio $\geq$ 20),medium (ratio $\geq$= 10 and $\leq$ 20), and moderate (ratio $>$ 0.5 and $\leq$ 10).

Table O.3: Balance table, **original randomization sample** framing variation: religious leader, doctor versus any male enumerator IVR

	(1)		(2)		(3)		T-test	
	Male enumerator IVR call		Religious leader IVR call		Male doctor IVR call		Difference	
Variable	N/[Clusters]	Mean/SE	N/[Clusters]	Mean/SE	N/[Clusters]	Mean/SE	(1)-(2)	(1)-(3)
Individual with more than 5 imputations at baseline	2055	0.117	2056	0.118	2054	0.121	-0.001	-0.004
	[861]	(0.009)	[867]	(0.010)	[880]	(0.009)		
Female	2055	0.381	2056	0.376	2054	0.379	0.006	0.003
	[861]	(0.016)	[867]	(0.016)	[880]	(0.016)		
Age	2055	38.018	2056	37.780	2054	38.212	0.238	-0.194
	[861]	(0.271)	[867]	(0.279)	[880]	(0.272)		
Household size	2055	9.103	2056	9.019	2054	9.034	0.084	0.069
	[861]	(0.128)	[867]	(0.124)	[880]	(0.120)		
Household owns either land or livestock	2055	0.660	2056	0.647	2054	0.657	0.013	0.003
	[861]	(0.013)	[867]	(0.013)	[880]	(0.013)		
Income in the past 7 days w	2055	884.986	2056	891.275	2054	899.301	-6.289	-14.314
	[861]	(39.062)	[867]	(37.667)	[880]	(36.328)		
Worked in the past 7 days	2055	0.431	2056	0.434	2054	0.430	-0.003	0.001
	[861]	(0.015)	[867]	(0.015)	[880]	(0.015)		
Village avg. Respondent participates and doesn't stop interview	2055	0.474	2056	0.477	2054	0.476	-0.002	-0.002
	[861]	(0.008)	[867]	(0.009)	[880]	(0.009)		
Somehousehold members fell sick past 14 days]	2055	0.154	2056	0.146	2054	0.144	0.008	0.010
	[861]	(0.009)	[867]	(0.009)	[880]	(0.009)		
# of household members with COVID-like symptoms [past 14 days]	2055	0.098	2056	0.091	2054	0.100	0.006	-0.002
	[861]	(0.010)	[867]	(0.009)	[880]	(0.011)		
# of days household members traveled for visit [past 7 days]	2055	1.101	2056	1.122	2054	1.113	-0.021	-0.013
	[861]	(0.065)	[867]	(0.065)	[880]	(0.067)		
# of days worked outside home [past 7 days]	2055	2.059	2056	2.074	2054	2.071	-0.016	-0.012
	[861]	(0.085)	[867]	(0.084)	[880]	(0.083)		
F-test of joint significance (F-stat)							0.206	0.243
F-test, number of observations							4111	4109

Note: This table shows balance statistics across the randomization variables for the comparison of religious leader and doctor IVR calls with male enumerator IVR calls in the framing-based intervention. Columns 1, 3, 5 display the number of observations. The number of clusters (villages) are displayed in brackets. Columns 2, 4, 6 display the mean of the baseline variables in the two groups. Standard deviations are displayed in parentheses. Columns 7-8 show the estimated difference in means which is obtained from regressing the variable of interest on the treatment dummy. Standard errors are clustered at the village level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. The value displayed for t-tests are the differences in the means across the groups. The value displayed for F-tests are the F-statistics. Variables marked with (w) were winsorized. All variables were scrutinized for outliers using the $90_t h/50_t h$ ratio. Based on the ratio values, variables were categorized into high (ratio $\geq$ 20),medium (ratio $\geq$= 10 and $\leq$ 20), and moderate (ratio $>$ 0.5 and $\leq$ 10).

B.3 Analysis of sub-treatment arms within T2 group

We restrict the sample to the T2 arm only for this analysis. Recall that, across the three FU waves, respondents in T2 received a robocall in one wave only (enumerator calls in the other two waves), randomly chosen to be the FU2, FU3 or FU4 wave. In the results presented below, the reference group is that of respondents who received a robocall in FU4.

Table O.4: Estimates for main indicators - T2 only

	(1) Pick-up the call	(2) Consent to interview	(3) Complete interview	(4) Respond to all questions
Received IVR in FU2	0.008	0.009	0.002	-0.016*
	(0.010)	(0.008)	(0.007)	(0.009)
Received IVR in FU3	0.005	0.001	0.007	-0.009
	(0.010)	(0.009)	(0.007)	(0.009)
Statistical tests (p-values)				
Equal	0.806	0.309	0.486	0.393
Jointly zero	0.743	0.462	0.572	0.165
Mean	0.689	0.853	0.943	0.903
SD	0.463	0.354	0.231	0.296
Obs.	14,258	9,866	8,774	8,774
R^2	0.019	0.016	0.006	0.004

Note: Sample consists of T2 arm only. The reference group is those respondents who received a robocall in FU4. Sample description for indicator 1- full sample, indicator 2- only sample that picked up the call, indicator 3- only sample that picked up the call and consented to interview, indicator 4- only sample that picked up the call, consented to interview and were asked all modules named r, i, sd and k. Control variables: randomization variables, wave and IP dummies.

As can be seen, the results show that besides the variable "Responds to all questions", there is no other significant result. This weakly significant result may suggest that there is a "priming" effect, i.e., the first enumerator-led interviews "primed" respondents to be more willing to complete interviews until the end. We know enumerator-led calls perform better than robocalls in terms of response rates, i.e. respondents are more likely to respond to all questions. So, someone who received 3 enumerator calls in a row (at baseline and in the first two follow-up waves, i.e. group 3 as defined above) "got used to" completing interviews. Consequently, by the time the receive the robocall, they are more familiar with the questions, and are "more used to" following through with the interview until the end.

C Structured Ethics Appendix

Policy Equipoise There is, in our opinion, no reasonable expectation that one arm of the study produces more benefits to participants than any other arm. None of the treatment arms was superior to the other w.r.t. the participants' net benefits.

Role of researchers with respect to implementation The research team (the authors of this study) had direct decision making power over whether and how to implement the activities tested in this study. IRB approval was obtained on July 28th, 2020 from Research and Development Solutions, Islamabad, Pakistan (IRB00010843). The research team did not directly intervene with the participants, it did however give instructions to endorse one or more of the interventions. No formal explanation of the experiment was provided since it may have influenced the results, but information about the data collection was shared. Informed consent was acquired for the data collection.

Potential harms to participants or non-participants from the interventions or policies The intervention being studied poses no potential harm to participants or non-participants. Participants' access to future services or policies did not changed because of participation in the study.

Potential harms to research participants or research staff from data collection (e.g., surveying, privacy, data management) or research protocols (e.g., random assignment) Our goal was to ensure that the data collection and/or research procedures adherent to privacy, confidentiality, risk-management, and informed consent protocols with regard to human subjects.

We do not think that research staff was at risk to be exposed to potential harms from conducting the data collection that are beyond "normal" risks. All data collection was remotely managed.

Financial and reputational conflicts of interest The researchers had no financial conflicts of interest with regard to the results of the research. The researchers have also no potential reputational conflicts of interest.

Intellectual freedom There were no contractual limitations on the ability of the researchers to report the results of the study.

Feedback to participants or communities The research team is acknowledging the need to share the evidence and, once the findings are in a final, peer-reviewed version, will be elaborate on how to best forward and communicate the results to the participants.

Foreseeable misuse of research results We anticipate no foreseeable and plausible risk that the results of the research will be misused and/or deliberately misinterpreted by interested parties to the detriment of other interested parties.

Other Ethics Issues to Discuss No other issues to discuss. The authors are available for further clarifications.