Policy Research Working Paper 9999

Talk or Text?

Evaluating Response Rates by Remote Survey Method during COVID-19

Sofia Amaral
Lelys Dinarte-Diaz
Patricio Dominguez
Santiago M. Perez-Vincent
Steffanny Romero

WORLD BANK GROUP
Development Economics
Development Research Group
April 2022

Policy Research Working Paper 9999

Abstract

Researchers and policy makers face significant challenges in selecting a method to conduct remote surveys, especially when collecting sensitive information or during turbulent life stages of hard-to-reach groups. In the context of the COVID-19 lockdown, this study randomly selected about 600 adults in El Salvador to survey using two different tools: telephone interviews or a self-completion survey via WhatsApp. The findings show that phone-based surveys increase the rate of survey completion by 42 percentage points. Even larger effects are documented for women and older adults. Although the direct costs of phone-based surveys are substantially higher—doubling implementation cost—the estimates imply that when adjusted for the probability of completion, the costs of conducting phone-based surveys can be 25 percent lower.

This paper is a product of the Development Research Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at ldinartediaz@worldbank.org.

Talk or Text? Evaluating Response Rates by Remote Survey Method during COVID-19

*

Sofia A maral[†] Lelys D inarte-Diaz[‡] Patricio Dominguez[§]
Santiago M. Perez-Vincent[¶] Steffanny Romero[||]

Keywords: Phone surveys, WhatsApp surveys, Response rate, Survey Experiments

*We thank Nathan Fiala, Pablo Kolb, and seminar participants at ifo Institute and the Innovations for Poverty Action-Northwestern University Methods and Measurement Conference. We also thank the research assistance of Miguel Paniagua, and the implementation partners of Glasswing International. This work was supported by the IDB-COVID 19 Call for Research Projects and the World Bank Research Support Budget. This research project's protocol was reviewed and approved by the Institutional Review Board (IRB) at the *Universidad Francisco Gavidia* in El Salvador in April 2020 with the approval ID No. 003-2020. The authors have no conflicts of interest to report. The findings, interpretations, and conclusions expressed in this report are entirely those of the authors. They do not necessarily represent the views of the Inter-American Development Bank, its Board of Directors or the countries they represent, or those of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. All errors are due only to the authors.

[†]ifo Institute at the Ludwig Maximilian University of Munich and CESifo. Email: amaral@ifo.de

[‡]Development Research Group. The World Bank. Email: ldinartediaz@worldbank.org

[§]Pontificia Universidad Católica de Chile. Email: pdomingr@uc.cl

[¶]Inter-American Development Bank. Email: santiagoper@iadb.org

[||] Universidad del Rosario and The World Bank. Email:steffanny.romero@urosario.edu.co

1 Introduction

Given the recent acceleration of digitization worldwide, high mobile phone diffusion, and improved access to internet, researchers and policy makers are increasingly interested in understanding the specific trade-offs of alternative methods for surveying individuals who are program beneficiaries, target groups for research studies, and the population in general (Abay et al., 2021).[1] This is especially true when in-person interviews are not feasible—such as during conflicts, refugee crises, and pandemic outbreaks—or expensive enough to make researchers seek alternative methods for data collection. Relative to in-person surveys, phone-based or online (self-completion) surveys can substantially reduce implementation costs (Garlick et al., 2020; Leo et al., 2015; Dabalen et al., 2016; Dillon, 2012; Mahfoud et al., 2015; Lamanna et al., 2019). However, among hard-to-reach groups, the data quality and sampling implications of phone or self-completion surveys are not fully understood, as documented in the literature (Mahfoud et al., 2015; Lamanna et al., 2019).

In terms of varying data-collection quality due to survey methods, two issues stand out. A first-order one is a non-random response rate, which can bias estimates substantially. For instance, online surveys cannot offer valid information about populations with little access to internet. Similarly, phone-based surveys yield high response rates only for populations with high rates of phone ownership and connectivity. Thus, inferences based on these kinds of surveys should be interpreted cautiously when describing specific groups.[2] Besides differences in access, another important concern regards the cultural likelihood of a response: some groups are especially reluctant or unwilling to respond to a particular survey method (Hersh et al., 2021).

Another problematic aspect concerns the reporting of sensitive information. On the one hand, the survey method can affect cognitive burden of respondents. This is especially true for those from hard-to-reach and vulnerable populations, many of whom may lack experience with self-completion or automated surveys (Pariyo et al., 2019). Moreover, the survey mode can also affect how people report sensitive issues.

[1] Although several surveys in high-income countries are now phone-based, there is still lack of evidence on what is the most effective survey method in low- and middle-income countries, where such surveys are more recent.

[2] (Abraham et al., 2009) show that non-response bias leads to an extensive overestimation of the rate of volunteering in the U.S.

For example, evidence shows it is more likely for women to report intimate partner violence when the instrument guarantees full anonymity, relative to another with a lower level of anonymity (Aguero and Frisancho, 2021). Therefore, in light of the inherent problems, how can researchers and policy makers assess the convenience of each method?

In this paper, we evaluate the relative effectiveness of collecting data on sensitive topics through two remote survey methods: phone-based surveys and self-completion surveys over WhatsApp. We conduct a randomized-controlled experiment, embedded on a survey-mode evaluation of a positive-parenting and stress-management intervention for caregivers in El Salvador, to assess the relative convenience of each method. In particular, we randomly allocate respondents to either phone-based - Computer Assisted Telephone Interview (CATI), which was conducted by a live person – or self-completion surveys – surveys done over WhatsApp – and evaluate their relative effect. In order to provide specific guidance for researchers and policy makers, we compare different dimensions of analysis such as completion rate, changes in composition of respondents, quality of responses,[3] and implementation costs. We use mental health and attitudes toward the effectiveness of physical punishment of children as measures of sensitive information, due to the large incidence of these issues over time and their greater relevance during the pandemic crisis (WHO, 2020; Cullen et al., 2020; Kumar and Nayar, 2021).[4] To credibly identify the relative convenience of the two modes, we ensured that aside from the survey mode itself, both groups were treated equally.[5]

We document five main results. First, we find that phone surveys increase the response rate by 42 percentage points. This represents an increase of 140 percent relative to the response rate among those who were contacted via WhatsApp messages. To put the magnitude of this estimate in perspective, our back-of-

[3] As we explain in more detail later, we measure the quality of responses by checking the consistency of responses from the caregivers in a survey conducted at midline against those reported in the baseline (before the intervention) and endline (at the end of the intervention) surveys.

[4] For example, as of 2017, 792 million people (more than one in 10 people) worldwide lived with a mental health disorder (Saloni Dattani and Roser, 2021), an issue that is more relevant for individuals from the lowest income strata (Ridley et al., 2020). Moreover, close to 300 million (three in four) children aged two to four years worldwide experience violent discipline by their caregivers on a regular basis; 250 million (around six in 10) are punished by physical means (UNICEF, 2017).

[5] Specifically, we used the same instrument, provided the same monetary incentive and time to complete the survey, there were no costs of completing the survey by either method (such as data plan usage), and the maximum number of reminders was similar across groups.

the-envelope calculation on cost-effectiveness suggests that when including the differences in probability of survey completion, rather than doubling the implementation costs, phone-based surveys are 25 percent less costly than self-completion surveys.

Second, survey methods can change the reasons for non-response. Using reports of causes for non-response provided by enumerators, we show that amenability was lower among phone-based survey experiment participants in comparison to those with connectivity issues relative to the same proportion of non-respondents in the self-completion survey mode. This result is important for contexts with weak connectivity settings, such as low internet speed or bandwidth, and for surveys that use self-completion questionnaires with high respondent fatigue and therefore greater refusal rates (Savage and Waldman, 2008). This result also highlights that phone surveys may ease the experience of survey completion for participants, thereby raising response rates.

Third, given the magnitude of the effect, we perform several additional analyses to better understand the implications of our results. First, we analyze impact heterogeneity across groups and identify a set of interesting patterns. Relative to men, women are 15 percentage points more likely to answer the survey when contacted by phone. The magnitude of this estimate is substantial; it implies that in the case of women, completion rates can increase from 30 to 80 percent in the context of phone-based surveys. Furthermore, we detect that the magnitude of the effect increases with age. Older populations are much more likely to complete the survey when reached by phone. In particular, relative to respondents who are younger than 40 years old, response rates increase by 41 percentage points among those who are 40 years old or older. Importantly, contrary to the initial hypothesis, we do not detect significant responses based on whether the respondent shares the cellphone with relatives or access to digital resources. Altogether, these results show that response rates can increase substantially when targeting specific groups such as women or older people.

In addition to differences in completion rates, as a fourth result, we also examine the extent to which the survey method altered the group composition of respondents. In principle, the survey method can affect an observational outcome by altering the group composition of respondents or—alternatively—the willingness to report a particular outcome among respondents. This can be relevant for surveys on sensitive issues such

as mental distress or attitudes toward physical punishment, where the survey method can impact both. For instance, relative to online methods, phone-based surveys can bias the population group toward those who are more likely to answer the phone and at the same time modify the likelihood of reporting a sensitive issue. We explore this in our context and find suggestive evidence that the survey method does not affect the composition of respondents, at least when comparing an important outcome such as the likelihood of reporting a high level of stress.

Finally, we also measure the quality of the reporting of sensitive outcomes, such as stress and physical punishment of children, by the type of remote survey method. Specifically, we test if responses concerning this sensitive information are consistent by type of survey method. We find no differences in the self-reports of stress and perception of the effectiveness of physical punishment across survey waves by survey mode. This result adds to a growing literature implementing different survey methods within face-to-face interviews.

Our paper contributes to three strands of literature. First, we contribute to the growing literature on methodology and best practices for designing and conducting phone-based surveys in low- and middle-income countries, where such surveys are quite new. For example, the works of Himelein and Kastelic (2021); L'Engle et al. (2018) on sampling; Greenleaf et al. (2020); Henderson and Rosenbaum (2020); Lau et al. (2019) on survey mode; Ballivian et al. (2015); Dabalen et al. (2016); Gibson et al. (2019); Lau and di Tada (2018); Leo et al. (2015); Özler et al. (2021) on survey cost, non-response, attrition, and use of incentives; Glazerman et al. (2020) on questionnaire design; Demombynes et al. (2013); Brubaker et al. (2021) on sample representativeness; Abay et al. (2021) on survey fatigue; and more recently, the documentation of general practices for the design of high-frequency phone surveys in the context of a pandemic (Gourlay et al., 2021; Etang and Himelein, 2020; Himelein and Kastelic, 2021; Feng et al., 2018; Morse et al., 2016). In our setting, we show that the main finding—higher response rates for phone-based surveys relative to message-based ones—holds during peak stress periods in a hard-to-reach group. Our design also allow for a direct comparison of survey modes without confounding the results with each of the above mentioned dimensions. Moreover, our data allows us to present evidence on two potential mechanisms behind this difference; that is, amenability - or refusal to respond - and contactability - in our case, due to connectivity

problems (Pariyo et al., 2019; De Weerdt et al., 2020). Our results are aligned with those of Siddique et al. (2021). The authors show that awareness and compliance with health guidelines during COVID was higher when households were approached via phone calls when compared to SMS. As in our paper, effects are also larger for women.

Importantly, our experimental design provides a novel way to simultaneously measure the effects of different survey modes and control for components that can confound the effects. Our experimental design relies on a sample that was subject to the same sampling method, contact rate, incentives, survey length, survey structure and reminders. This advantage is important since by design it allows to overcome some of the shortcomings in the existing literature considering that each of these dimensions could on its own alter response rates, group composition and data quality beyond the true effect of the survey mode.[6]

Second, our findings contribute to a rich literature that examines the effects of self-administration on data quality and reporting related to sensitive survey items, such as intimate-partner violence (Aguero and Frisancho, 2021; Cullen, 2020); contraceptive use (Greenleaf et al., 2020); and other health issues (Kays et al., 2012).[7] Initial research suggests that interviews conducted via text messaging on mobile devices essentially mimic self-administration in terms of higher reporting of sensitive behaviors and higher data quality in general (Brenner and DeLamater, 2014; Cocco and Tuzzi, 2013; Schober and Conrad, 2015). Interviews using text messages also remove the potential for interviewer effects on responses due to observable characteristics of the interviewers, on which there is ample research (Schaeffer et al., 2010). In theory, one can therefore hypothesize that survey designs involving interviewers and mobile phones will simultaneously reduce respondent burden (due to decreased questionnaire length) and increase the quality of reporting on sensitive survey items; we test this hypothesis with the present study. Specifically, we focus on dimensions that have not been considered so far (perceptions of physical-punishment effectiveness and mental health). Overall, our results suggest there are no significant differences in response quality between phone-based surveys

[6] In Table A6 in the Appendix we provide a summary of the literature and the design features in each study.

[7] Within this strand, a subset of studies analyzes response and measurement error from surveys in general (not only phone-based or message-based ones). See, for example, the works from Friedman et al. (2017); Gibson et al. (2015); De Weerdt et al. (2016).

and self-administered instruments through phone messages.

Finally, existing evidence documents heterogeneity in data quality obtained from different survey modes by some respondent characteristics (Bardasi et al., 2011; Beegle et al., 2012; Gigliotti and Dietsch, 2014; LeFevre et al., 2020).[8] For example, Bardasi et al. (2011) document that responses by proxy of employment rather than self-report have differential effects on employment rates by gender in Tanzania. Similarly, Beegle et al. (2012) experimentally test alternative survey designs in face-to-face surveys that measure household consumption. The authors find that under-reporting is particularly apparent for illiterate households and for urban respondents. We contribute to this literature by experimentally testing survey mode – instead of survey design – and variation in response rate by respondents' characteristics that are standard in the literature – gender and age – and by others that are novel and relevant due to the policy implications of the findings – phone availability and access of electronic-transfer technologies.

The remainder of the paper is organized as follows: Section 2 describes the research design, including details on the recruitment process, and experimental design. Section 3 presents the data collected and our implementation process. In section 4, we describe our main empirical approach. Section 5 summarizes the main results obtained from specifications presented in section 4, and section 6 describes our robustness checks. Finally, in section 7 we briefly conclude.

2 Research Design

We embedded a survey-mode experiment in a midline survey of the impact evaluation of a positive-parenting and stress-management intervention for caregivers in El Salvador (Amaral et al., 2021). The midline survey sought to assess if caregivers who received the intervention were incorporating its concepts. We took the opportunity of the survey to conduct the survey-mode experiment and generate evidence on the most convenient method (CATI vs. self-completion) to collect data on sensitive topics. We used the results of this

[8] See De Weerdt et al. (2020) for a literature review on how survey designs can influence socioeconomic data collection in developing countries and, as a consequence, how this also affects the data and results of subsequent analyses.

experiment to choose the survey mode for the impact evaluation's endline survey (conducted a few months later).

As we show in Figure 1, we enrolled a total of 6,258 individuals for the main impact evaluation.[9] These individuals were recruited through three channels: Facebook, the implementing partner's network, and WhatsApp messages sent to customers of Tigo, the largest cell phone provider in El Salvador.[10] These individuals received a link to an enrollment survey in August 2020.[11] Only 4,718 individuals met the eligibility criteria for the impact evaluation and were allowed to enroll in the study.[12] After enrollment, we contacted them for a baseline survey between August and September 2020, and collected baseline data from 3,103 individuals (66 percent) for the impact evaluation.[13] In October 2020, we conducted the midline survey with a group of 599 randomly selected individuals out of the 4,718 enrolled in the impact evaluation.

The survey-mode experiment aimed to evaluate the relative convenience of the two survey modes based on four criteria: (i) response rate, (ii) changes in composition of respondents, (iii) quality of responses, and (iv) cost.[14] We calculate the response rate as the percentage of contacted individuals who complete the survey. We assess changes in the composition of respondents by measuring changes in their report of stress and attitudes toward violent parenting practices. We measure the quality of responses by checking the consistency of responses in the midline survey against those reported in the baseline and endline surveys. Finally, we use data on the monetary cost of contacting individuals through the two channels under analysis.

To conduct the survey-mode experiment, we first randomly selected 300 individuals from each treat-

[9]These were selected by Tigo by random digit selection and by also making sure that individuals were not residing in the same address or vicinity of the same address. This avoids a potential treatment contamination concern.

[10]Table A7 presents data on the channel used to recruit our sample. Most participants in the midline survey (and the impact evaluation) were recruited through Tigo's client network. A similar recruitment process could be replicated in contexts with a high mobile-phone ownership rate, such as El Salvador, where 81 percent of the population owns a mobile phone.

[11]The enrollment survey questions and consent form are available in this link.

[12]The eligibility criteria to participate in the impact evaluation were being between 18 and 45 years of age, living with at least one child eight years old or younger, having access to a mobile phone, and providing consent to receive digital messages and to participate in the study (Amaral et al., 2021).

[13]As explained in Amaral et al. (2021), 34 percent of enrolled individuals did not complete the baseline survey for different reasons, including did not provide a correct phone number, were not able to be reached after several attempts, and decided not to participate in the survey.

[14]The following section describes in more detail how we measure these three variables.

ment arm in the intervention and then we cross randomized and assigned the participants into two groups in the midline survey - see Table A1.[15] Each of these groups was asked to complete the survey through an alternative mode, as follows:

- **Group 1 — CATI — 300 individuals:** This group received a call from a live person asking them to complete a phone-based survey (CATI). At the outset, they were informed about the monetary incentive (US$ 2.50 in Tigo Money) to complete the survey. On average, participants received 3,7 (3,4) calls per day over three days – including the picked up survey call. Enumerators were trained to conduct surveys on sensitive topics. Following recommendations by Glazerman et al. (2020), we developed a contact-attempt tracking log for each participant.

- **Group 2 — Self-completion — 299 individuals:** This group received WhatsApp messages[16] inviting them to complete an online survey. The message included a link to the self-completion survey with a note on the incentive to complete the survey (US$ 2.50 in Tigo Money) and the deadline for completion (three days). Tigo provided free access to the survey. On average, participants received 2.9 (2.6) calls per day over three days from trained enumerators who reminded them to complete the self-completion survey. Thus we made the same effort to reach participants in each group. These contact attempts were tracked in the same individual log as Group 1.

To credibly identify the relative convenience of the two modes, we not only randomly assigned individuals to one of the two groups but also ensured that aside from the survey mode itself, both groups were treated equally. The instrument was the same, all individuals received the same monetary incentive to complete the survey, both groups had three days to complete it, there were no costs of completing the survey by either method (such as data plan usage), and the maximum number of reminders was similar across groups. These features are non-trivial and allow for a causal interpretation of the effect of the two survey modes without confounding with other potential aspects of survey design.

[15]While the initial sample was 600 participants, one person dropped out of the study, so we ended up with 299 individuals from the treatment group and 300 from the control group.

[16]Throughout this paper, we use the terms "CATI" and "phone-based survey" to refer to the first treatment arm, and "self-completion" and "WhatsApp" to refer to the second group.

We assessed the external validity of our experiment and the representativeness of our sample by comparing them with the sample of the 2019 Household and Multipurposes Survey (EHPM), a nationally representative survey in El Salvador that is conducted through face-to-face interviews. Of the approximately 75,000 individuals interviewed in the EHPM, 15,680 (21 percent) meet our eligibility criteria (See Table A2 in the Appendix).[17] When we restrict the EHPM sample to individuals who meet our criteria, we find no meaningful differences in age, gender, and age of the oldest child with our sample (See Table A3 in the Appendix).[18] We also verify that the randomization of individuals in the Phone/WhatsApp intervention was not correlated with the characteristics and treatment assignment of the impact evaluation studied in Amaral et al. (2021). Therefore, we show in Figure A2 that there is no correlation across gender, age or mobile phone usage.

3 Data and Summary Statistics

3.1 Data Sources

We use four sources of data to assess the relative convenience of the two survey modes: administrative records provided by the survey firm, responses to the enrollment survey, responses to the midline survey, and responses to the parenting intervention's baseline and endline surveys.[19]

Administrative records provided by the survey firm. This data includes information on whether the individuals completed or did not complete the midline survey, the reason for non-completion (for those who did not complete it), and the number of reminder calls made to each person.

Enrollment survey. The enrollment survey included questions on respondent's gender and age, mobile phone ownership status, whether they lived with and cared for a child eight years old or younger, and the

[17] The eligibility criteria are described in footnote 12.

[18] We only find differences between the EHPM and the survey-mode experiment samples in educational-level variables; the survey-mode experiment sample is more educated. This difference could also stem from the fact that only 60 percent of our sample completed the baseline survey. Therefore, educational attainment data is available only for a subsample of 358 (out of 599) individuals.

[19] In Figure 1 we present the different stages of data collection that we describe in this section.

age of the eldest child under eight years old under their supervision. We used this information to assess respondents' eligibility to participate in the parenting intervention.[20]

Midline survey. The midline survey consisted of 22 questions, divided into two sets. The first set of questions asked about respondents' knowledge of concepts and the use of techniques taught in the parenting intervention.[21] The second set included questions on respondents' stress levels and their attitudes toward violent parenting. The average completion time of the survey was 10 minutes. As explained in the previous section, some individuals responded to the survey over the phone and others completed an online form assessed via Whatsapp. All answers were self-reported.[22]

Intervention's baseline and endline surveys. For the impact evaluation of the parenting intervention, we conducted baseline and endline surveys (Amaral et al., 2021). We use information from these surveys regarding respondents' stress levels and their attitudes toward violent parenting practices.[23]

3.2 Outcomes

For the survey-mode experiment, we focus our analysis on the following main outcomes: completion rate and consistency of responses to sensitive questions.[24]

A. Completion rate: It consists of a dummy variable that takes the value of 1 if the survey was completed (either by phone or WhatsApp) within the three days it was open (through October 19, 2020). Both surveys were designed to make it impossible to skip any question. Information on completion status was obtained from administrative reports provided by the survey firm.

[20] See eligibility criteria for the impact evaluation in footnote 6.

[21] In Amaral et al. (2021), we use participants' responses about the use and knowledge of different stress-management and positive-parenting techniques and concepts to assess their assimilation of the content provided in the parenting intervention. In addition to this midline survey, we conducted three other surveys with different samples of individuals. Each survey asked about techniques taught in the previous two to three weeks.

[22] Following information protection protocols, data collected from all the other surveys was stored on a private server of the survey firm. Only the project staff and researchers had access to this.

[23] To collect baseline data, we sent a text message to caregivers inviting them to self-complete the baseline survey by clicking on a link included in the message.

[24] See Table A8 for further details on the description and data used to estimate the main outcomes of this survey-mode experiment.

B. Sensitive items: To measure changes in the composition of respondents, we use items from the second set of questions in the midline survey to construct two sensitive outcomes.[25] We asked the same questions in baseline and endline surveys for the impact evaluation.

- *Stress:* To measure participants' perception of stress in the midline survey, we asked them, "How stressed did you feel last week?" The response options were *Very Stressed, Moderately Stressed, Mildly Stressed, Not Stressed at All*. An individual has "High Stress" if she responds, "Very Stressed." In the baseline and endline surveys, we measured participants' stress using the Depression, Anxiety, and Stress Scale (DASS-21) instrument (Lovibond and Lovibond, 1996). We use the seven items from the stress subscale. Each item is measured on a scale of 0–3 points (*Never, Rarely, Almost Always, or Always*). We estimate the total stress score following Lovibond and Lovibond (1996); that is, a participant has "High Stress" if the score is greater than 18.

- *Perceptions of effectiveness of physical violence toward children:* We asked participants, "Do you think physical punishment is an effective method to discipline children?" The response options were *No, it is never effective; It is rarely effective; It is sometimes effective; It is effective most of the time; It is always effective*. The same question and response options were included in baseline, midline, and endline surveys.

C. Consistency in self-reports of sensitive items: To measure consistency in the reporting of either of the two sensitive outcomes at baseline or in the endline survey, we generated dummy indicators that take the value of 1 if the answer in the midline survey is equal to the one reported at baseline and in the endline survey, respectively.[26]

D. Sociodemographic information: In addition to the outcomes described above, we merged the midline survey with information collected during the enrollment of participants or baseline survey concerning their age, gender, education, age of children in her care, municipality of residence, and mobile phone us-

[25] More details on the construction of these outcomes—sources of data, items (original scale and translation in Spanish), and description—are in Table A9 in the Appendix.
[26] See Table A8 in the Appendix for more details.

age.[27]

3.3 Descriptive Statistics and Balancing Tests

Table 1 shows descriptive statistics of socio-demographic information for the full sample and separated by treatment arm. Panel A presents summary statistics of characteristics available for the full sample. On average, caregivers in our sample are 31 years of age,[28] 52 percent are women, and have an oldest child who is 4.6 years old. In terms of phone exclusivity, only 23 percent of participants share a phone with other relatives within the household. Finally, 3 of each 10 participants use a money-transfer technology (Tigo Money). Only 60 and 44 percent of caregivers completed baseline and endline surveys, respectively.

Panel B in Table 1 shows descriptive statistics for variables that are available only for individuals who completed the baseline survey. We find that most caregivers (64 percent) in our sample do not have tertiary education (bachelor's degree or higher). We also show that 28 percent of the sample were either unemployed or lost their jobs during the pandemic.[29] Both treatment arms are balanced in most socio-demographic characteristics except for gender; there are more women in the WhatsApp (self-completion) group.[30] Since our sample involves participants with and without a baseline survey completed we also test for potential selection in pre-determined characteristics taking into account this feature. These results are presented in Figure A2 and show that there are no significant differences across arms.

Table 2 presents summary statistics of our main outcomes. In panel A, we show that on average, the completion rate was 51 percent and the number of attempts to reach participants (reminders) was greater than three. Moreover, between 60 and 68 percent of participants in our study were consistent in their responses of sensitive information at baseline, and between 58 and 62 percent at endline. Among reasons for

[27] In Table A10 in the Appendix, we present the list of variables that were collected during the enrollment and baseline surveys and merged to the participants of the midline survey.

[28] There are two missing observations in the survey-mode experiment sample for which we were not able to obtain information on age.

[29] As we show in Figure A2 in the Appendix section, we reject the null hypothesis that the completion of the baseline survey was driven by other individual characteristics such as participant's age, gender, and availability of digital resources.

[30] As we discuss in the next section, to account for this difference, we include the gender indicator as a control in all estimations.

non-completion, 67 percent of participants refused to respond and one-third reported connectivity problems.

Finally, we present some descriptive statistics of the sensitive information measures (stress and perceptions about effectiveness of physical punishment) in panel B. On average, caregivers in our sample in the midline survey display low levels of mental health—about 43 percent exhibit an above-normal level of stress—and more than half of participants report that physical punishment is an effective method to discipline a child. These levels are very similar in the baseline survey (31 percent report high stress and 48 percent report that physical punishment is effective), but they are significantly lower in the endline survey (only 17 percent of participants report high levels of stress and only 37 percent agree that punishment is an effective method of discipline).

4 Empirical Strategy

To identify whether response rates, composition of respondents, and measures of consistency in the reporting of sensitive issues are affected by survey mode, we exploit the random assignment of each participant to a particular survey method and estimate the following regression:

$$Y_i = \delta Phone_i + \sum_{j}^{n} X_{ij} + M_i + \epsilon_i \tag{1}$$

where Y_i stands for the response rate, sensitive issue outcome, or consistency of reports measured for individual i. $Phone_i$ is a survey mode indicator that takes the value 1 if individual was assigned to CATI or 0 if she was assigned to the WhatsApp group. $\sum_j^n X_{ji}$ is a vector of j time-invariant participant characteristics, which include age and gender. We also add municipality fixed effects (M_i) to control for time-invariant shocks at the municipal level and to improve the precision of the estimates. The coefficient associated with the survey mode indicator in Equation (1), $\hat{\delta}$, captures the impact of using the phone-based survey mode and represents our coefficient of interest (ITT). ϵ_i is an idiosyncratic error term. Since the randomization occurred at the individual level and we do not expect clusters in our sample, we report heteroskedasticity-

robust standard errors.

We also examine treatment-effect heterogeneity to learn more about potential drivers of the main effects. In practice, we estimate the following model:

$$Y_i = \delta_1 Phone_i + \delta_2 Phone_i \times H_{ki} + \sum_j^n \beta_j X_{ji} + M_i + \epsilon_i \quad (2)$$

where H_{ki} is a variable that measures k type of heterogeneity, and all other variables are defined as in Equation (1). We separately explore four types of heterogeneity: age (an indicator for whether the participant is older than a specific age), gender (an indicator variable for whether the respondent is a woman), phone exclusivity (an indicator variable for whether the respondent shares a phone with a relative), and use of electronic-transfer technology (an indicator variable for whether the respondent uses Tigo Money). We hypothesize that older and female respondents are more likely to complete a phone-based survey relative to using a message-based technology. Moreover, sharing a phone with a relative (23 percent of our sample) might reduce the probability that an individual completes a phone-based survey relative to receiving a message that can be completed later. However, the quality or consistency of the responses might be affected, since the relative (rather than the main respondent) might complete the survey. Finally, the use of electronic-transfer technology might facilitate the reception of incentives and reminders, and therefore increase the completion rates of phone-based surveys.

In specification (2), δ_1 delivers the ITT estimate for the main respondent (for example, for women), and δ_2 provides the difference in the ITT effect between the two groups (for example, women and men). The ITT estimate for the reference group is given by the sum of δ_1 and δ_2. The interpretation of the coefficients is analogous for the other two indicator variables. We estimate Equation (2) by OLS and report heteroskedasticity-robust standard errors.

5 Results

5.1 Main Results

Table 3 presents regression results focusing on the change in the likelihood of completing the survey due to the mode of data collection. Column (1) presents the model with no controls. Columns (2) and (3) include socio-demographic characteristics (age and sex) as controls. To address potential concerns that phone availability is correlated with income, we include educational level as a proxy in the model in column (4).[31] Also, we include municipality fixed effects to account for potential time-invariant characteristics within municipalities. These results are included in columns (3) and (4). The main model presented in Equation (1) is estimated in column (3).[32] In sum, our estimations indicate that respondents who were interviewed on the phone were 42 percentage points more likely to complete the survey, relative to those who were reached through WhatsApp. The average response rate of this group was around 30 percent. These differences between the two survey modes are stable after including individual controls and municipality fixed effects. The total response rate for the phone-based survey was similar to rates estimated in similar experiments in Bangladesh, Liberia, and Kenya (Pariyo et al., 2019; Morse et al., 2016; Lamanna et al., 2019).

An interesting feature of our experiment is our collection of information about the potential explanations for failure to respond that survey modes may provide. During the contact attempts, enumerators reported one of two potential reasons for lack of completion: contactability – due to connectivity issues – and refusal to respond by the caregiver. Using this information, we note an interesting fact: in addition to changing the probability of completing the survey, phone-based surveys change the proportion of respondents who refuse to answer, relative to those who had connectivity issues. Among those who were contacted by phone, the proportion of refusal vs. connectivity issues was 0.62 and 0.38, respectively. Similarly, among

[31] The categories for educational level are primary, high school, or tertiary education—the omitted category. Since educational level is available only for participants who completed the baseline survey, we impute 0 for the missing values.

[32] We are not using the model from column (4) as our main model because we do not have information for all the observations for educational level. Instead, we are imputing the average for those with missing data. We include that model in our main results table to show that the estimations do not change after including such variables.

those who were contacted by message, the refusal vs. connectivity issues partition was 0.69 and 0.31. On the one hand, this result is expected since completing a digital survey can require more connectivity settings (internet speed, bandwidth, phone data) than a phone-based one (basic cell phone reception). Alternatively, it suggests that lack of response due to refusal to respond can be reduced by implementing phone-based surveys, as documented by Savage and Waldman (2008).[33]

5.2 Heterogeneous Impacts on Completion Rates by Respondent's Characteristics

There are many reasons to believe the treatment effect is heterogeneous across the population. Although our sample size is relatively small, we explore heterogeneity across key dimensions to assess if the treatment effects estimated on completion rate may differ by respondent characteristics. We measure changes in the likelihood of completing the survey by interacting our treatment status variable with an indicator for some of the relevant baseline categories. We selected these covariates based on their potential availability since researchers can easily get that information before running the survey; this could help them to better target resources toward specific groups.[34]

Table 4 and Figure 2 summarize impacts of the phone-based survey on survey completion across some key categories: age, gender, phone exclusivity, and use of electronic-transfer technology (Tigo Money). Column (1) in Table 4 shows heterogeneous results for women relative to men, column (2) presents differences in impacts on individuals sharing the phone relative to those with exclusive use, and column (3) shows results for respondents who use Tigo Money compared to those who do not.

We document three main results. First, women are 15 percentage points more responsive to completing the survey through a phone call, relative to men assigned to the same survey mode. The total treatment

[33] In the context of choice experiments, the authors find that the quality of the online respondents' answers declines throughout a series of experiments.

[34] Unfortunately, due to the way baseline information was collected for the experiment in which we embedded this project, our database lacks additional baseline characteristics such as educational level, employment status, and other socioeconomic variables that are available only for a partial portion of our sample.

effect on women is around 50 percentage points.[35] Second, we explore heterogeneity by respondent's availability of digital resources. Evidence on the gap in ownership of mobile phones by gender or income is limited, and this heterogeneity analysis addresses efforts to link phone ownership among women and poor people to obtain information from them (LeFevre et al., 2020). To proxy for respondent's availability of digital resources, we use indicators of whether the respondent shares a phone with someone else and if she has Tigo Money, and we test for heterogeneity based on these two variables. As we show in columns (2) and (3) in Table 4, there are no differences relative to those with different availability of digital resources—that is, relative to those with phone exclusivity or without Tigo Money. Yet, the sign and the magnitude of the interaction between having Tigo Money and being assigned to the phone-based survey are consistent with the idea that respondents who are more prone to use digital resources are less responsive to a phone-based survey than to a survey delivered via WhatsApp.

We also examine treatment effect heterogeneity by age. As we emphasized in the introduction, phone usage and survey completion can differ substantially by age (Gigliotti and Dietsch, 2014). Thus we can expect that the effect of receiving a phone call on answering the survey increases with age. We test that hypothesis by estimating specification (2)—for example, by interacting our treatment indicator with an "old group" indicator for whether the respondent is older than a certain age i, where $i \in \{22, 23, ..., 43\}$. Figure 2 summarizes the age heterogeneity by modifying the age threshold for the "old group" indicator we include in each separate regression. We can see that coefficients are increasing with the age threshold that defines the old-group indicator variable, suggesting that older respondents are more responsive to the treatment. For example, although we observe no difference when comparing respondents older and younger than 25 years old, respondents who are older than 40 years old are 41 percentage points more likely to answer the survey.

[35] Although existing evidence has shown differences in the quality of data that can be collected from women using different survey modes, further research is required to understand differences in the quality of data or response rate by gender. For example, Lamanna et al. (2019) find that the use of mobile phones to collect nutrition data from rural women in Africa may result in differences of between 0 and 25 percent in nutrition estimates, relative to collecting this information in face-to-face interviews.

5.3 Effects on Group Composition

An additional challenge of the comparison across remote survey methods relates to the likelihood of altering the composition of the respondent group. That is, a potential concern is that one survey method at midline has a differentiated impact on future attrition than the other. This is important because there are two competing ways through which a survey method can modify an observational outcome of interest. For instance, in a context of low response rate, a change in the composition of the group of people who complete the survey can confound the impact of a program. Similarly, different survey methods can alter the likelihood of reporting a particular response, which can be especially salient in surveys about sensitive issues.[36] In our context, this matters because researchers want to know not only the extent to which a particular method affects the size of the group that completes the survey, but also whether it alters the group composition in ways that correlate with the outcome of interest. Of course, our research design does not allow us to causally disentangle both possibilities, but we address this issue by exploiting the experiment within which this survey is embedded. Specifically, we do so by comparing the effect of the survey method on the level of stress reported by respondents.

Table 5 measures the effect of the phone-based survey on self-reported stress level and perceptions of the effectiveness of physical punishment in two different time periods. In columns (1) and (2), we show the impact of the phone-based survey on reporting a high level of stress in the midline survey, while columns (5) and (6) report similar estimates at the endline survey. Similarly, for perceptions of the effectiveness of physical punishment, coefficients for the midline survey are in columns (5) and (6), and for the endline survey, in columns (7) and (8). Given the high level of non-response in both remote survey methods, we would like to know whether the coefficients in columns (1) and (2) capture either a change in willingness to

[36] We examine this issue in Table 5. Columns (1)–(4) show estimates of differences in the willingness to report sensitive information by type of survey mode during our remote survey experiment (midline survey). Using our measures of stress and perceptions of the effectiveness of physical punishment, we find that participants who completed the survey over the phone are less likely to report higher levels of stress, relative to those assigned to the WhatsApp survey mode. At the same time, we do not observe significant differences when comparing the outcome binary variable *Believe that physical punishment is effective*. At least in the case of stress, these findings show a large difference among those who completed the survey by phone.

report when interviewing by phone or a change in the group composition of those who answer the survey. For instance, people can be less likely to report a high level of stress on the phone. Likewise, people who are less likely to answer the phone can coincide with those who are less likely to report a high level of stress, regardless of the survey method.

Thus, we examine whether the survey method modifies the composition of the respondents by comparing differences in reporting a high level of stress at the midline and endline surveys. The advantage of using the endline survey as a benchmark is that we are comparing everyone under the same data-collection process. If differences in the phone-based survey method at the midline survey are driven by changes in the composition of respondents, we should observe similar disparity at the endline survey. In other words, by virtue of the randomization, the detection of a difference across groups at the endline will likely provide information about changes in the group composition rather than the survey method used in the midline, 6 to 8 weeks earlier. Unlike columns (1) and (2), in which the phone-based survey method substantially affects the likelihood of reporting a high level of stress, we observe no significant differences in columns (5) and (6). This suggests that differences in columns (1) and (2) are likely to be driven by changes in the likelihood of reporting a high level of stress based on the survey method rather than a change in the composition of respondents.

5.4 Effects on Consistency in Reporting Sensitive Outcomes

To measure the quality of the reporting of sensitive information, such as mental health and physical punishment, the ideal is to collect objective measures of sensitive items and compare them to what respondents are reporting. However, it was not possible to collect these measures of stress or physical punishment in this particular context.[37] In this sense, with the aim of providing a proxy for quality in the collection of sensitive information through different survey modes, we measure consistency in self-reports of stress and perceptions of the effectiveness of physical punishment. To do this, we compare changes in perceptions of

[37] Face-to-face interviews would be necessary to collect objective measures of stress. Also, information on physical punishment is usually under-reported in administrative records.

the effectiveness of physical punishment as a method of disciplining children. Then, we use this variable as a secondary outcome in our regressions.[38]

Results using specification (1) are presented in Table 6. Columns (1)–(4) compare consistency between midline and baseline surveys, and columns (5)–(8) compare consistency in self-reports between midline and endline surveys by the mode of the survey. For each outcome, we present models that include or exclude interviewer fixed effects. Overall, we find no differences in the self-reports of stress or the perception of the effectiveness of physical punishment across survey waves by survey mode.

6 Robustness Checks

In addition to assessing whether our main results were affected by the inclusion of control variables, we conducted several additional robustness checks. Results are summarized in Table A4 in the Appendix section. Column (1) presents the estimated coefficient of the main model, which includes age and sex as control variables, and municipality fixed effects. We first test that the absence of baseline information for some of the participants does not affect the results by including in column (1) indicators for completing baseline survey or type of recruitment that are equal to 1 if recruitment was done through Tigo Clients and 0 if through other channels. These results are in columns (2) and (3), respectively. We show that the estimated coefficient does not change after the addition of these control variables.

Second, since conventional p-values typically rely on approximations that assume that the pool of individuals is large enough that the test statistics follow a specific sampling distribution (Ding et al., 2016), we conduct an adjustment of the estimated standard errors using randomization inference. Results in brackets in column (4) show that the statistical significance did not change and the precision of the estimation improved. Lastly, we show that the inclusion of an educational-level variable as control does not change the main results, especially when we impute 0 for the missing values for caregivers who did not complete the

[38] An important caveat from this analysis is that we are measuring consistency in self-reports comparing responses that were reported five to six weeks between midline and baseline, and six to seven weeks between midline and endline, which can be relatively long periods of time. Yet, we are finding no differences between the two reports.

baseline survey – see column (5). In the case that we include the variable without imputation – see column (6) – the estimated coefficient is slightly lower but the statistical significance does not change.

Finally, we also conduct robustness checks of the effects of survey mode on secondary outcomes after including other control variables that can proxy for income level. Specifically, we test if the lack of differences in the consistency of reporting sensitive information by survey mode changes after the inclusion of educational level as control variable. As we show in Table A5, differences are still statistically insignificant after including this proxy for income level.

7 Discussion and Concluding Remarks

We provide empirical evidence on how a phone-based mode affects survey completion rates relative to a self-completion mode. Considering the rapid increase in access to digital technologies, this represents a key parameter for policy makers who confront a set of trade-offs when collecting data. We find that phone-based surveys increase response rates by 42 percentage points—even more if targeted to women and older generations. Considering the costs of data collection and the likelihood of failing to find a significant effect in underpowered studies, this result provides a key piece of information, especially in countries similar to El Salvador, where we ran this experiment.

Importantly, along with differences in completion rates, we also document differences in responses across groups. This can be relevant when collecting data on sensitive issues. For example, we observe that phone-based survey respondents were less likely to report a high level of stress. We find suggestive evidence that differences in stress level by mode of survey were due to differences in willingness to report as opposed to changes in the composition of the group that completed the survey. This is important if a research team is considering combining phone-based and self-completion surveys in some part of the data-collection process, and it wants to evaluate how much its decision can affect both the integrity of the data collected and its comparability with other sources of information.

Finally, policy makers would like to know the trade-off between the two alternative methods we compare. Based on our findings and a simple back-of-the-envelope calculation, we observe that after consider-

ing the costs of implementing each method and the number of attempts per completed survey (3.69 CATI vs. 2.9 WhatsApp), the direct cost for each completed survey by CATI is double that of each completed survey using WhatsApp (US$ 11.40 vs. US$ 6.20). However, when we adjust the cost by probability of completion (0.72 vs. 0.3), the relationship flips and the cost of CATI becomes 25 percent lower (US$ 16 vs. US$ 21.20).

References

Abay, K. A., Berhane, G., Hoddinott, J. F., and Tafere, K. (2021). *Assessing response fatigue in phone surveys: Experimental evidence on dietary diversity in Ethiopia*, volume 2017. Intl Food Policy Res Inst.

Abraham, K. G., Helms, S., and Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the united states. *American Journal of Sociology*, 114(4):1129–1165.

Aguero, J. and Frisancho, V. (2021). Measuring violence against women with experimental methods.

Amaral, S., Dinarte-Diaz, L., Dominguez, P., and Perez-Vincent, S. (2021). *Helping Families Help Themselves? Heterogeneous Effects of a Digital Parenting Program*. World Bank Publications.

Ballivian, A., Azevedo, J. P., Durbin, W., Rios, J., Godoy, J., and Borisova, C. (2015). Using mobile phones for high-frequency data collection. *Mobile Research Methods*, 21.

Bardasi, E., Beegle, K., Dillon, A., and Serneels, P. (2011). Do labor statistics depend on how and to whom the questions are asked? results from a survey experiment in tanzania. *The World Bank Economic Review*, 25(3):418–447.

Beegle, K., De Weerdt, J., Friedman, J., and Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from tanzania. *Journal of development Economics*, 98(1):3–18.

Brenner, P. S. and DeLamater, J. D. (2014). Social desirability bias in self-reports of physical activity: is an exercise identity the culprit? *Social Indicators Research*, 117(2):489–504.

Brubaker, J., Kilic, T., and Wollburg, P. (2021). Representativeness of individual-level data in covid-19 phone surveys.

Cocco, M. and Tuzzi, A. (2013). New data collection modes for surveys: a comparative analysis of the influence of survey mode on question-wording effects. *Quality & quantity*, 47(6):3135–3152.

Cullen, C. (2020). Method matters: Underreporting of intimate partner violence in nigeria and rwanda. *World Bank Policy Research Working Paper*, (9274).

Cullen, W., Gulati, G., and Kelly, B. D. (2020). Mental health in the covid-19 pandemic. *QJM: An International Journal of Medicine*, 113(5):311–312.

Dabalen, A., Etang, A., Hoogeveen, J., Mushi, E., Schipper, Y., and von Engelhardt, J. (2016). *Mobile phone panel surveys in developing countries: a practical guide for microdata collection*. World Bank Publications.

De Weerdt, J., Beegle, K., Friedman, J., and Gibson, J. (2016). The challenge of measuring hunger through survey. *Economic Development and Cultural Change*, 64(4):727–758.

De Weerdt, J., Gibson, J., and Beegle, K. (2020). What can we learn from experimenting with survey methods? *Annual Review of Resource Economics*, 12:431–447.

Demombynes, G., Gubbins, P., and Romeo, A. (2013). Challenges and opportunities of mobile phone-based data collection: Evidence from south sudan. *World Bank Policy Research Working Paper*, (6321).

Dillon, B. (2012). Using mobile phones to collect panel data in developing countries. *Journal of international development*, 24(4):518–527.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 655–671.

Etang, A. and Himelein, K. (2020). Monitoring the ebola crisis using mobile phone surveys. In *Data collection in fragile states*, pages 15–31. Palgrave Macmillan, Cham.

Feng, S., Grépin, K. A., and Chunara, R. (2018). Tracking health seeking behavior during an ebola outbreak via mobile phones and sms. *NPJ digital medicine*, 1(1):1–8.

for Health Initiative: NCD Mobile Phone Survey, B. P. D. (2020a). Executive summary: Zambia ncd mobile phone survey 2017. Technical report, Bloomberg Philanthropies.

for Health Initiative: NCD Mobile Phone Survey, B. P. D. (2020b). Ncd mobile phone survey in the philippines. Technical report, Bloomberg Philanthropies.

Friedman, J., Beegle, K., De Weerdt, J., and Gibson, J. (2017). Decomposing response error in food consumption measurement: Implications for survey design from a randomized survey experiment in tanzania. *Food Policy*, 72:94–111.

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

Gibson, D. G., Wosu, A. C., Pariyo, G. W., Ahmed, S., Ali, J., Labrique, A. B., Khan, I. A., Rutebemberwa, E., Flora, M. S., and Hyder, A. A. (2019). Effect of airtime incentives on response and cooperation rates in non-communicable disease interactive voice response surveys: randomised controlled trials in bangladesh and uganda. *BMJ global health*, 4(5):e001604.

Gibson, J., Beegle, K., De Weerdt, J., and Friedman, J. (2015). What does variation in survey design reveal about the nature of measurement errors in household consumption? *Oxford Bulletin of Economics and Statistics*, 77(3):466–474.

Gigliotti, L. and Dietsch, A. (2014). Does age matter? the influence of age on response rates in a mixed-mode survey. *Human Dimensions of Wildlife*, 19(3):280–287.

Glazerman, S., Rosenbaum, M., Sandino, R., and Shaughnessy, L. (2020). Remote surveying in a pandemic: Handbook. *Innovations for Poverty Action*.

Gourlay, S., Kilic, T., Martuscelli, A., Wollburg, P., and Zezza, A. (2021). High-frequency phone surveys on covid-19: Good practices, open questions. *Food Policy*, 105:102153.

Greenleaf, A. R., Gadiaga, A., Guiella, G., Turke, S., Battle, N., Ahmed, S., and Moreau, C. (2020). 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.

Heath, R., Mansuri, G., Sharma, D., Rijkers, B., and Seitz, W. (2017). Measuring employment: Experimental evidence from ghana. Technical report, Working paper.

Henderson, S. and Rosenbaum, M. (2020). Remote surveying in a pandemic: research synthesis. *Innovation for Poverty Action*.

Hersh, S., Nair, D., Komaragiri, P. B., and Adlakha, R. K. (2021). Patchy signals: capturing women's voices in mobile phone surveys of rural india. *BMJ Global Health*, 6(Suppl 5):e005411.

Himelein, K. and Kastelic, J. G. (2021). The socio-economic impacts of ebola in liberia: Results from a high frequency cell phone survey, round 5. Technical report.

Kays, K., Gathercoal, K., and Buhrow, W. (2012). Does survey format influence self-disclosure on sensitive question items? *Computers in Human Behavior*, 28(1):251–256.

Kumar, A. and Nayar, K. R. (2021). Covid 19 and its mental health consequences.

Lamanna, C., Hachhethu, K., Chesterman, S., Singhal, G., Mwongela, B., Ng'endo, M., Passeri, S., Farhikhtah, A., Kadiyala, S., Bauer, J.-M., et al. (2019). Strengths and limitations of computer assisted telephone interviews (cati) for nutrition data collection in rural kenya. *PloS one*, 14(1):e0210050.

Lau, C. and di Tada, N. (2018). Identifying non-working phone numbers for response rate calculations in africa. *Survey Practice*, 11(2):3269.

Lau, C. Q., Cronberg, A., Marks, L., and Amaya, A. (2019). 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.

Lau, C. Q., Johnson, E., Amaya, A., LeBaron, P., and Sanders, H. (2018). High stakes, low resources: what mode (s) should youth employment training programs use to track alumni? evidence from south africa. *Journal of International Development*, 30(7):1166–1185.

LeFevre, A. E., Shah, N., Bashingwa, J. J. H., George, A. S., and Mohan, D. (2020). Does women's mobile phone ownership matter for health? evidence from 15 countries. *BMJ global health*, 5(5):e002524.

Leo, B., Morello, R., Mellon, J., Peixoto, T., and Davenport, S. T. (2015). Do mobile phone surveys work in poor countries? *Center for Global Development Working Paper*, (398).

Lindhjem, H. and Navrud, S. (2011). Are internet surveys an alternative to face-to-face interviews in contingent valuation? *Ecological economics*, 70(9):1628–1637.

Lovibond, S. H. and Lovibond, P. F. (1996). *Manual for the depression anxiety stress scales*. Psychology Foundation of Australia.

L'Engle, K., Sefa, E., Adimazoya, E. A., Yartey, E., Lenzi, R., Tarpo, C., Heward-Mills, N. L., Lew, K., and Ampeh, Y. (2018). Survey research with a random digit dial national mobile phone sample in ghana: methods and sample quality. *PloS one*, 13(1):e0190902.

Maffioli, E. M. (2019). Relying solely on mobile phone technology: Sampling and gathering survey data in challenging settings.

Mahfoud, Z., Ghandour, L., Ghandour, B., Mokdad, A. H., and Sibai, A. M. (2015). Cell phone and face-to-face interview responses in population-based surveys: how do they compare? *Field methods*, 27(1):39–54.

Morse, B., Grépin, K. A., Blair, R. A., and Tsai, L. (2016). Patterns of demand for non-ebola health services during and after the ebola outbreak: panel survey evidence from monrovia, liberia. *BMJ global health*, 1(1):e000007.

Özler, B., Çelik, Ç., Cunningham, S., Cuevas, P. F., and Parisotto, L. (2021). Children on the move: Progressive redistribution of humanitarian cash transfers among refugees. *Journal of Development Economics*, 153:102733.

Pariyo, G. W., Greenleaf, A. R., Gibson, D. G., Ali, J., Selig, H., Labrique, A. B., Al Kibria, G. M., Khan, I. A., Masanja, H., Flora, M. S., et al. (2019). Does mobile phone survey method matter? reliability of computer-assisted telephone interviews and interactive voice response non-communicable diseases risk factor surveys in low and middle income countries. *PloS one*, 14(4):e0214450.

Pattnaik, A., Mohan, D., Chipokosa, S., Wachepa, S., Katengeza, H., Misomali, A., and Marx, M. A. (2020). Testing the validity and feasibility of using a mobile phone-based method to assess the strength of implementation of family planning programs in malawi. *BMC health services research*, 20(1):1–9.

Ridley, M., Rao, G., Schilbach, F., and Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science*, 370(6522).

Saloni Dattani, H. R. and Roser, M. (2021). Mental health. *Our World in Data*. https://ourworldindata.org/mental-health.

Savage, S. J. and Waldman, D. M. (2008). Learning and fatigue during choice experiments: a comparison of online and mail survey modes. *Journal of Applied Econometrics*, 23(3):351–371.

Schaeffer, N. C., Dykema, J., and Maynard, D. W. (2010). Interviewers and interviewing. *Handbook of survey research*, 2:437–471.

Schober, M. F. and Conrad, F. G. (2015). Improving social measurement by understanding interaction in survey interviews. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):211–219.

Siddique, A., Rahman, T., Pakrashi, D., Islam, A., and Ahmed, F. (2021). Raising health awareness in rural communities: A randomized experiment in bangladesh and india.

UNICEF (2017). A familiar face: Violence in the lives of children and adolescents.

West, B. T., Ghimire, D., and Axinn, W. G. (2015). Evaluating a modular design approach to collecting survey data using text messages. In *Survey research methods*, volume 9, page 111. NIH Public Access.

WHO (2020). Mental health and psychosocial considerations during the covid-19 outbreak, 18 march 2020. Technical report, World Health Organization.

8 Tables and Figures

Table 1: Summary Statistics and Balance Tests of Sample Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Full Sample				WhatsApp	Phone	P-val.Dif.
Variable	Obs.	Mean	St.Dev.	Min	Max	Mean	Mean	(6) vs (7)
Panel A: Characteristics of full sample								
Age (years)	597	31.22	6.65	16	45	31.24	31.21	(0.960)
Female (%)	599	0.52	0.50	0	1	0.572	0.473	(0.016)**
Oldest child age (years)	599	4.56	2.70	0	13	4.602	4.517	(0.699)
Shares phone (%)	599	0.23	0.42	0	1	0.234	0.220	(0.681)
Has Tigo Money (%)	599	0.30	0.46	0	1	0.291	0.300	(0.809)
Completed baseline survey (%)	599	0.60	0.49	0	1	0.589	0.607	(0.653)
Completed endline survey (%)	599	0.44	0.50	0	1	0.431	0.453	(0.590)
Observations						299	300	
Panel B. Characteristics of subsample								
Education level								
Basic (1-9 grades, %)	358	0.21	0.41	0	1	0.239	0.176	(0.143)
High school (10-12 grades, %)	358	0.43	0.50	0	1	0.426	0.429	(0.963)
Bachelor or higher (%)	358	0.37	0.48	0	1	0.335	0.396	(0.237)
Employment status pre and post pandemic								
Always unemployed (%)	358	0.20	0.40	0	1	0.190	0.203	(0.746)
Always employed (%)	358	0.49	0.50	0	1	0.449	0.522	(0.165)
Lost job (%)	358	0.08	0.28	0	1	0.098	0.071	(0.370)
Found job (%)	358	0.23	0.42	0	1	0.264	0.203	(0.175)
Observations						176	182	

Notes: This table shows descriptive statistics of samples' characteristics in columns (1)-(5). Balance tests comparing mean characteristics of control and treatment groups and p-value for the difference in means are presented in columns (6)-(8). Panel A presents descriptive statistics using the full sample of the experiment and Panel B shows descriptive statistics only for the sub-sample of participants that completed baseline survey. Description of each variable is presented in table A10 in the Appendix. "WhatsApp" refers to the self-completion survey treatment and "CATI" to the phone-based survey. Robust standard errors are in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 2: Summary Statistics of Main Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Full Sample					WhatsApp	CATI	P-val.Dif.
Variable	Obs.	Mean	St.Dev.	Min	Max	Mean	Mean	(6) vs (7)
Panel A: Outcomes								
Completed survey (%)	599	0.51	0.50	0	1	0.298	0.720	(0.000)***
Consistency with baseline survey								
Stress high (%)	207	0.60	0.49	0	1	0.574	0.619	(0.535)
Physical punishment is effective (%)	207	0.68	0.47	0	1	0.735	0.647	(0.207)
Consistency with follow up survey								
Stress high (%)	169	0.62	0.49	0	1	0.566	0.638	(0.376)
Physical punishment is effective (%)	169	0.58	0.50	0	1	0.642	0.552	(0.275)
Reasons to not complete the survey								
Refusal to response (%)	294	0.67	0.47	0	1	0.686	0.619	(0.275)
Connectivity problems (%)	294	0.33	0.47	0	1	0.314	0.381	(0.275)
Panel B: Sensitive information measures using data from...								
Knowledge incorporation survey								
Stress high (%)	305	0.43	0.50	0	1	0.584	0.370	(0.001)***
Physical punishment is effective (%)	305	0.54	0.50	0	1	0.539	0.546	(0.912)
Baseline survey								
Stress high (%)	358	0.31	0.46	0	1	0.295	0.324	(0.558)
Physical punishment is effective (%)	358	0.48	0.50	0	1	0.500	0.456	(0.407)
Follow - up survey								
Stress high (%)	265	0.17	0.38	0	1	0.202	0.147	(0.243)
Physical punishment is effective (%)	265	0.37	0.48	0	1	0.341	0.390	(0.413)

Notes: This table shows summary statistics of the main outcomes used in this paper. Panel A shows statistics of main outcomes, which were constructed as described in table A8 in the Appendix. Panel B presents statistics on measures of sensitive information using different sources (knowledge incorporation surveys, baseline, and follow up data). Construction of these sensitive items and data details are presented in table A9 in the Appendix. Differences in number of observations to estimate the consistence measures are due to availability of data in each survey. For example, 305 participants completed the midline survey and 358 has baseline information in our sample. However, to estimate consistence, we only have information in both samples for 207 respondents.

Table 3: Differences in response rates by mode of data collection

	(1)	(2)	(3)	(4)
	Outcome: Survey completed			
Phone survey	0.422***	0.425***	0.427***	0.417***
	(0.037)	(0.037)	(0.038)	(0.038)
Mean of Dep.Var (WhatsApp)	0.298	0.299	0.299	0.299
R-squared	0.178	0.182	0.212	0.242
Observations	599	597	597	597
Controls (Age, Sex)	No	Yes	Yes	Yes
Controls (Education)	No	No	No	Yes
Municipality FE	No	No	Yes	Yes

Notes: This table shows the estimated impacts of the phone-based survey relative to WhatsApp-based survey on completion rate. "Phone survey" is a dummy variable equal to one if the survey was per call and zero if was assigned to the WhatsApp group. The dependent variable in columns (1)-(4) is a dummy equal to one if the survey was completed (See more details in Table A8). The control variables in columns (2)-(3) include age in years and sex of the respondent. We also add the educational level of the respondent for primary, high-school or tertiary education -the omitted category- in column (4). We impute zero for the missing values in control variables in column (4). To account for potential time-invariant characteristics within municipalities, we estimate a model with municipality fixed effects and present results in columns (3) and (4). Robust standard errors are in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 4: Heterogeneity Analysis

	(1)	(2)	(3)
	Outcome: Survey completed		
i. Phone survey	0.348***	0.420***	0.456***
	(0.057)	(0.045)	(0.045)
ii. Female	-0.030		
	(0.055)		
iii. Female x Phone Survey	0.153**		
	(0.077)		
iv. Shares phone		-0.009	
		(0.065)	
v. Shares phone x Phone survey		0.032	
		(0.091)	
vi. Has Tigo money			0.130**
			(0.062)
vii. Has Tigo money x Phone survey			-0.101
			(0.085)
Total effect on women ($[i] + [iii]$)	0.501***		
Total effect on people who share phone ($[i] + [v]$)		0.451***	
Total effect on people that have Tigo money ($[i] + [vii]$)			0.356***
Mean of Dep.Var (WhatsApp)	0.299	0.299	0.299
R-squared	0.217	0.212	0.219
Observations	597	597	597
Controls (Age, Sex)	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes

Notes: This table shows the heterogeneous impacts of the phone-based survey – relative to WhatsApp-based survey – on completion rate by gender, phone exclusivity, and use of electronic transfer technology. Female is an indicator variable for whether the respondent is a woman, "Shares phone" is a dummy variable for whether the respondent shares a phone with a relative, and "Has Tigo Money" is an indicator variable for whether the respondent uses Tigo Money. "Phone survey" consists of a dummy variable equal to one if the survey was per call and zero if was assigned to the WhatsApp group. The dependent variable in columns (1)-(3) is a dummy equal to one if the survey was completed (See more details in Table A8 in the Appendix). The control variables include age in years and sex of the respondent. To account for potential time-invariant characteristics within municipalities, we estimate a model with municipality fixed effects. Robust standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Willingness to report versus group composition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Midline Survey				Endline Survey			
	Stress high		Physical punishment is effective		Stress high		Physical punishment is effective	
Phone survey	-0.202***	-0.249***	0.014	0.052	-0.052	-0.049	0.038	0.057
	(0.065)	(0.073)	(0.067)	(0.073)	(0.050)	(0.048)	(0.063)	(0.061)
Mean of Dep. Var (WhatsApp)	0.584	0.618	0.539	0.500	0.202	0.202	0.341	0.341
R-squared	0.154	0.216	0.087	0.266	0.136	0.195	0.142	0.204
Observations	304	207	304	207	265	265	265	265
Controls (Age,Sex)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dep. Var Baseline Level	No	Yes	No	Yes	No	Yes	No	Yes

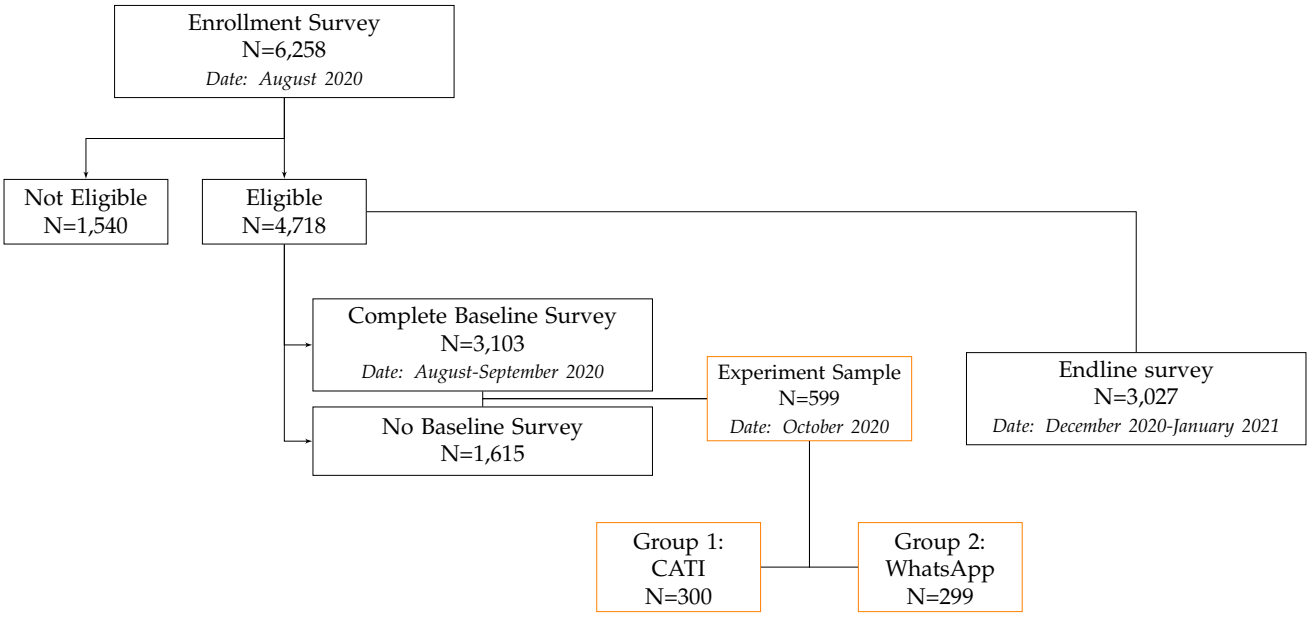
Notes: This table shows the estimated impacts of the phone-based survey relative to WhatsApp-based survey on sensitive outcomes. "Phone survey" is a dummy variable equal to one if the survey was per call and zero if was assigned to the WhatsApp group. The dependent variable in columns (1)-(2) and columns (5)-(6) is a dummy equal to one if self-report of stress was high in the midline and endline survey, respectively; in columns (3)-(4) and columns (7)-(8) is the perception about physical punishment effectiveness in the midline and endline survey, respectively. The control variables include age in years and sex of the respondent. To account for potential time-invariant characteristics within municipalities, we include municipality fixed effects. Robust standard errors are in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 6: Consistency in Responses on Sensitive Information with Additional Surveys

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Consistency with Baseline Survey				Consistency with Endline Survey			
	Stress high		Physical punishment is effective		Stress high		Physical punishment is effective	
Phone survey	0.032	-0.053	-0.066	-0.103	0.067	-0.083	-0.076	-0.022
	(0.079)	(0.137)	(0.074)	(0.129)	(0.090)	(0.154)	(0.089)	(0.157)
Mean of Dep. Var (WhatsApp)	0.574	0.574	0.735	0.735	0.566	0.566	0.642	0.642
R-squared	0.129	0.132	0.159	0.166	0.163	0.177	0.153	0.177
Observations	207	207	207	207	169	169	169	169
Controls (Age, Sex)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interviewer FE	No	Yes	No	Yes	No	Yes	No	Yes
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

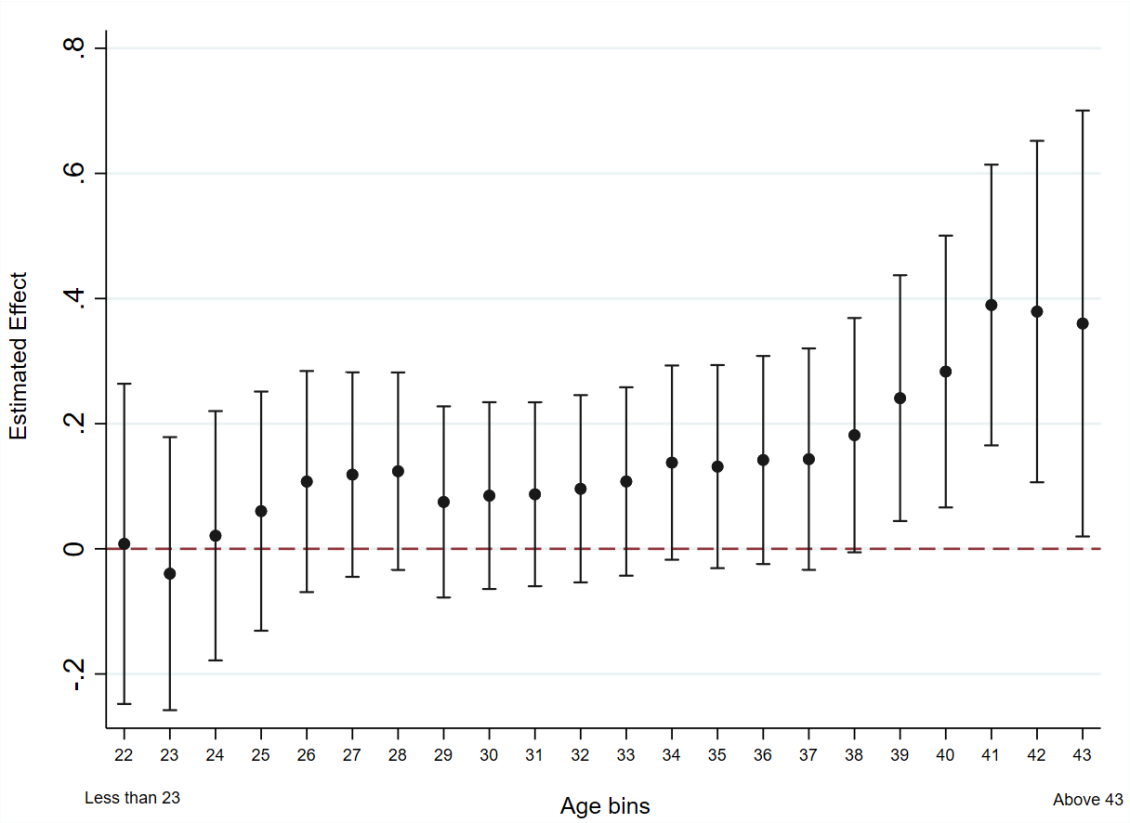
Notes: This table shows the estimated impacts of the phone-based survey relative to WhatsApp-based survey on consistency in responses on sensitive information. "Phone survey" is a dummy variable equal to one if the survey was per call and zero if was assigned to the WhatsApp group. The dependent variables is a dummy equal to one if the participant was consistent on her self-report of stress and her perception about physical punishment effectiveness between pairs of surveys (See more details in Table A8). The control variables include age in years and sex of the respondent. To account for potential time-invariant characteristics within municipalities, we include municipality fixed effects. As robustness check, we include education level as control variable. Results are presented in Table A5 in the Appendix. Robust standard errors are in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Figure 1: Sample Selection and Experimental Design



Notes: This figure shows the sample selection and randomization procedure applied in this survey-mode experiment design (orange boxes). From the total of enrolled and eligible caregivers that participated in the impact evaluation in Amaral et al. (2021), we randomly selected a subsample of 599 caregivers. Half of this subsample were assigned to group 1 (CATI or phone-based survey) and the rest to group 2 (WhatsApp or self-completion survey).

Figure 2: Heterogeneous Effects by Age



Notes: This figure shows the heterogeneous impacts of the phone-based survey—relative to the WhatsApp-based survey—on completion rate by respondent's age. We report the estimated coefficient from the interaction between the treatment indicator ("Phone survey") and a dummy that takes the value of 1 if the respondent's age was older than age i, where $i \in \{22, 23, ..., 43\}$. In the estimations, we include sex of the respondent as a control variable and municipality fixed effects to account for potential time-invariant characteristics within municipalities. Confidence intervals indicate statistical significance at 5%.

Appendix Tables and Figures

Table A1: Distribution of sample by previous treatment assignment

	Survey mode experiment	
	Phone survey	WhatsApp
Treatment	150	149
Control	150	150
P-value	0.928	

Notes: This table presents the distribution of the survey experiment sample drawn from the sample of the main impact evaluation being analysed in Amaral et al. (2021). Within each treatment arm, we randomly selected a group of 300 individuals (for a total of 600) and randomly assigned them to one of the survey-mode treatments. "*P*-value" indicate the *p*-value of the test for differences in treatment status (Treatment or Control) in the main impact evaluation by assignment to any survey-mode (Phone-based survey or WhatsApp).

Table A2: Construction of the Comparable Nationally Representative Sample

	Percentage	Observations
	100%	75045
Eligibility by age of children	51%	38,097
Eligibility by age (own)	47%	17,733
Eligibility of main caregiver	98%	17,315
Eligibility of access to mobile phone	91%	15,680
Eligibility Sample	21%	15,680

Notes: This table presents the total sample from the 2019 Household Survey (EHPM) that meet the eligibility criteria for the impact evaluation. *Eligibility by Age of Children* indicates if the individual lives together with at least one child 8 years old or younger. *Eligibility by age* is met if the individual's age is between 18-65 years. *Eligibility of main caregiver* indicates if the person is either a parent or a grandparent of a child within the household. Finally, *Eligibility by access to mobile phone* was created using the information on whether the person has a mobile phone. Columns (1) and (2) show the percent and number of individuals in the EHPM sample that meets each of the consecutive eligibility criteria. For example, 17,733 individuals (47 percent) are adults between 18-65 years of age living together with at least one child 8 years old or younger.

Table A3: Differences in means

Variable	EHPM Sample (Mean) (1)	Survey-mode experiment Sample (Mean) (2)
Age (years)	29.181	31.221
Female (%)	0.566	0.521
Oldest child age (years)	4.605	4.553
Observations	15,680	597
Education level		
None	0.014	
Basic (1-9 grades, %)	0.514	0.207
High school (10-12 grades, %)	0.281	0.427
Bachelor or higher (%)	0.075	0.366
Observations	15,680	358

Notes: This table shows the average characteristics of the sample in the Household Survey (EHPM) that meets the eligibility criteria (column 1) and the sample in the survey-mode experiment (column 2). Panels A and B show characteristics that are available for the full survey-mode experiment sample and only for those that completed baseline, respectively.

Table A4: Robustness checks

	(1)	(2)	(3)	(4)	(5)	(6)
	Outcome: Survey completed					
Phone survey	0.427***	0.423***	0.425***	0.427***	0.417***	0.347***
	(0.038)	(0.038)	(0.039)	(0.038)	(0.038)	(0.053)
				[0.000]		
Mean of Dep. Var (WhatsApp)	0.299	0.299	0.299	0.299	0.299	0.386
R-squared	0.212	0.234	0.212	0.212	0.242	0.217
Observations	597	597	597	597	597	358
Controls (Age, Sex)	Yes	Yes	Yes	Yes	Yes	Yes
Controls (Education)	No	No	No	No	Yes	Yes
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Controlling for baseline completion		✓				
Controlling for mode of initial recruitment			✓			
Computing randomized inference p-values				✓		
Imputing missing values					✓	

Notes: This table shows the estimated impacts of the phone-based survey relative to WhatsApp-based survey on completion rate. "Phone survey" is a dummy variable equal to one if the survey was per call and zero if was assigned to the WhatsApp group. The dependent variable is a dummy equal to one if the survey was completed. Column (1) presents the estimated coefficient of the main model (same as in column (3) in Table 3, which includes age and sex as control variables, and municipality fixed effects). Columns (2) and (3) include indicators for completing baseline survey or type of recruitment (=1 if Tigo, = 0 if other channel), respectively. Column (4) present results after adjusting standard errors using randomization inference, which are presented in squared-brackets. Column (5) includes education level as control variable, imputing zero for the missing values for caregivers that did not complete the baseline survey. Finally, column (6) presents estimated coefficients of the same model as in column (5) but with no imputation of missing values on the education level variable. Robust standard errors are in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

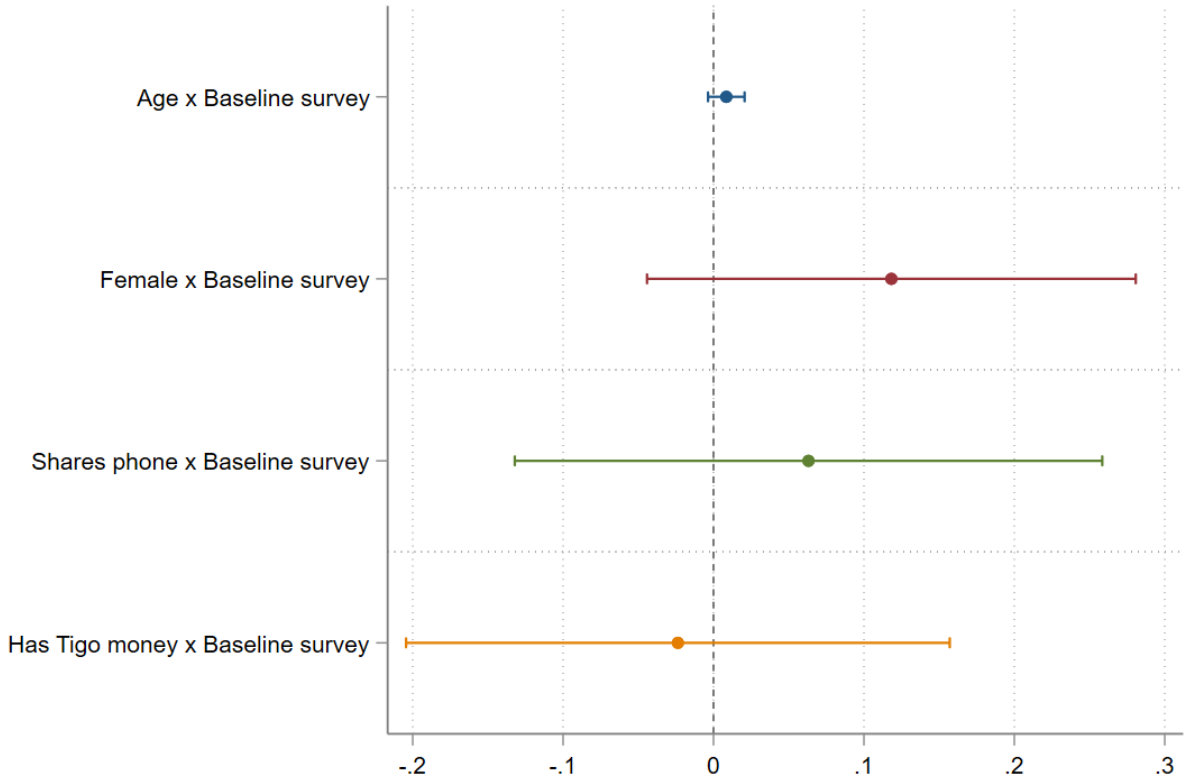
Table A5: Consistency in Responses on Sensitive Information with Additional Surveys

(Including education level as control variable)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Consistency with Baseline Survey				Consistency with Endline Survey			
	Stress high		Physical punishment is effective		Stress high		Physical punishment is effective	
Phone survey	0.030	-0.071	-0.064	-0.114	0.073	-0.087	-0.075	-0.017
	(0.079)	(0.142)	(0.074)	(0.127)	(0.090)	(0.157)	(0.089)	(0.158)
Mean of Dep.Var (WhatsApp)	0.574	0.574	0.735	0.735	0.566	0.566	0.642	0.642
R-squared	0.142	0.147	0.171	0.177	0.173	0.188	0.156	0.178
Observations	207	207	207	207	169	169	169	169
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interviewer FE	No	Yes	No	Yes	No	Yes	No	Yes
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

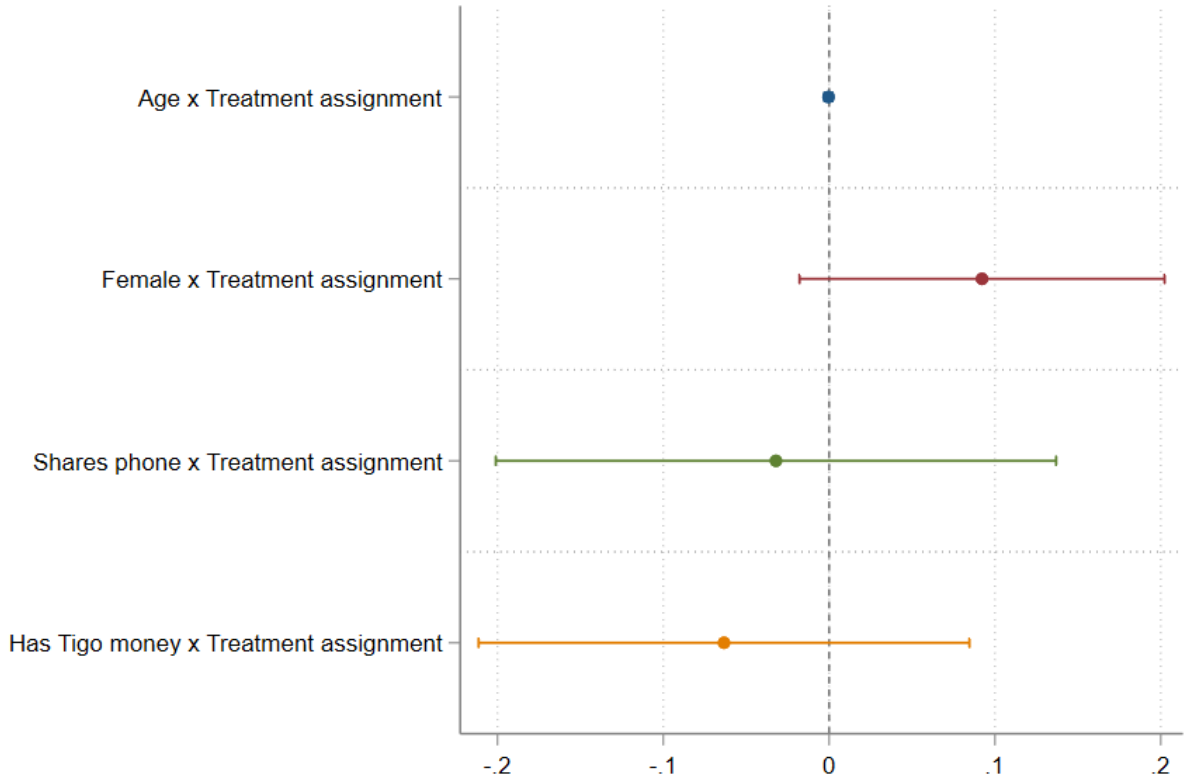
Notes: This table shows the estimated impacts of the phone-based survey relative to WhatsApp-based survey on consistency in responses on sensitive information. "Phone survey" is a dummy variable equal to one if the survey was per call and zero if was assigned to the WhatsApp group. The dependent variables is a dummy equal to one if the participant was consistent on her self-report of stress and her perception about physical punishment effectiveness between pairs of surveys (See more details in Table A8). The control variables include age in years, sex of the respondent and the educational level of the respondent for primary, high-school or tertiary education -the omitted category-. To account for potential time-invariant characteristics within municipalities, we include municipality fixed effects. Robust standard errors are in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Figure A1: Baseline survey completion and individual characteristics



Notes: This figure shows that results from tests of the completion of balance survey were driven by individual characteristics, such as age, gender, and availability of digital resources. We report the estimated coefficient from the interaction between the indicator "Has baseline survey completed" and the age of the individual or indicators for being a woman, sharing the phone, or having Tigo Money. Confidence intervals indicate statistical significance at 5%.

Figure A2: Treatment assignment and individual characteristics



Notes: This figure displays coefficients and 95% confidence intervals of a regression on treatment assignment to the Phone or WhatsApp treatment arm and baseline characteristics of respondents interacted with the treatment assignment analysed in Amaral et al. (2021).

Table A6: Summary of the experimental design features in the literature

Authors	Compared Methods	Sample Size	Sample Recruitment	Same contact rate?	Same incentives?	Same survey length?	Same survey structure?	Same attempts?	Country	Troubled context
This paper	**CATI, SMS**	**599**	**RDD**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**El Salvador**	**Yes**
Lau, Cronberg, Marks and Amaya (2019)	CATI, IVR, SMS,FTF	2016421	RDD, National surveys	No	No	No	No	No	Nigeria	No
West, Ghimire and Axinn (2015)	CATI, SMS, SMS modular design	450	National surveys	Yes	No incentives	Yes	Yes	No reminders	Nepal	No
Pariyo,Greenleaf, Gibson, Ali, Selig, Labrique, Al Kibria, Khan, Masanja, Flora, Ahmed and Hyder (2019)	CATI, IVR	2196	RDD	Yes	No	Yes	No	Yes	Bangladesh and Tanzania	No
Lindhjem and Navrud (2011)	E-mail web link, FTF	1043	Survey firm	Yes	Yes	Yes	Yes	No reminders	Norway	No
Garlick, Orkin, and Quinn (2020)	Monthly - Weekly FTF, Weekly CATI	895	National surveys	Yes	Yes	Yes	Yes	Yes	South Africa	No
Cullen (2020)	FTF, ACASI, List Experiment	8273	National surveys	No	–	No	No	No reminders	Rwanda and Nigeria	No
Bloomberg Philanthropies Data for Health Initiative: NCD Mobile Phone Survey (2020a)	IVR, SMS	6056	RDD	Yes	Yes	Yes	Yes	Yes	Zambia	No
Maffioli (2019)	IVR, CATI: Sequential	2265	RDD	Yes	No	No	No	No	Ivory Coast	Yes
Health, Mansuri, Rijkers, Seitz, and Sharma (2017)	FTF, CATI	1579	National surveys	Yes	No	Yes	Yes	Yes	Ghana	No
Lau, Johnson, Amaya, LeBaron, and Sanders (2018)	SMS, Web link	2603	Social enterprise	Yes	Yes	No	No	Yes	South African	No
Lamanna, Hachhethu, Chesterman, Singhal, Mwongela, Ng'endo, Passeri, Farhikhtah, Kadiyala, Bauer, and Rosenstock (2019)	CATI, FTF	1821	Recruitment team	Yes	No incentives	No	No	No reminders	Kenya	No
Bloomberg Philanthropies Data for Health Initiative: NCD Mobile Phone Survey (2020b)	SMS, Web link	3673	RDD	Yes	Yes	Yes	Yes	Yes	Philippines	No
Pattnaik, Mohan, Chipokosa, Wachepa, Katengeza, Misomali, and Marx (2020)	CATI,FTF: Sequential	1057	Social enterprise	Yes	–	No	No	–	Malawi	No
Brenner, and DeLamater (2014)	SMS, Web link	156	RDD, Midwestern university	Yes	No	Yes	Yes	–	United States	No

This table presents a summary of the features of each experiment we identified in the literature. We include the description of each feature and code as Yes/No if the study contains a determined feature. The features coded are: (i) survey modes compared; (ii) sample size; (iii) if the mode of recruitment was the same across survey modes; (iv) if the contact mode was the same across survey modes; (v) if the incentives were the same by survey mode; (vi) if the survey length was the same by survey mode; (vii) if the survey structure was the same by survey mode; and, (vii) if the attempts to contact participants were the same by survey mode. We also display the setting in which the study took place by coding the country and if the participants were part of a troubled health or political context. "ACASI" stands for Audio Computer-Assisted Self-Interviewing; "IVR" for "Interactive Voice Response; "CATI" for Computer-Assisted Telephone Interviewing; "SMS" Short Message Service, and "FTF" for Face-to-Face. Moreover, "RDD" stands for Random Digital Dial.

Survey tools appendix

Table A7: Summary of Enrollment Sample

	Mode of Data Collection			**Total**
	Facebook	Glasswing Communities	Tigo Clients	
	Female Caregiver			
	45	9	259	313
	Male Caregiver			
	27	2	257	286
Total	72	11	516	599

Notes: This table presents the distribution of caregivers in the survey-mode experiment sample by recruitment mode and participant's gender. Almost the full sample (86 percent) was recruited from the Tigo Clients dataset, 12 percent from Facebook, and only 2 percent from the NGO's Communities program.

Table A8: Outcomes List and Description

Outcome	Source	Outcome construction
Survey completed (completion rate)	Monitoring data	Dummy indicator that takes the value of 1 if the KI survey was completed (either by phone or SMS), and 0 if it was incomplete after 3 days.
Consistency on tolerance towards child punishment as in baseline (follow up) survey. Outcome defined as ″physical punishment is effective″	Survey instruments (KI, baseline, and follow up surveys)	Dummy indicator that takes the value of 1 if the response in the KI survey was the same as in the baseline (follow up) survey, 0 otherwise.
Consistency on high stress level as in baseline (follow up). Outcome defined as ″High stress″	Survey instruments (KI, baseline, and follow up surveys)	Dummy indicator that takes the value of 1 if the individual reported having high stress level in the KI survey and in the baseline (follow up) survey, 0 otherwise.

Notes: This table describes the outcomes used in the main analysis of the survey-mode experiment. It also presents the source of the data and a description of how the outcome was constructed. "KI" stands for Knowledge Incorporation Survey.

Table A9: Survey Instruments and Construction of Sensitive Outcomes

Outcome	Source(s)	Outcome construction	Item – Original version	Item – Spanish version
High stress	Knowledge Incorporation Survey (KI)	Dummy indicator that takes the value of 1 if the individual responded "Very stressed," 0 otherwise	How stressed did you feel during the past week?	¿Qué tan estresado se ha sentido esta última semana?
			Response options: Very stressed, Moderately stressed, Mildly stressed, Not stressed at all	Opciones de respuesta: Muy estresado, Estresado, Poco estresado, Nada estresado
High stress	Baseline/ Follow up Surveys– (DASS-21)	Each of the 7 items in the category of stress in the DASS-21 are measured on a scale of 0–3 points (*Never, Rarely, Almost Always or Always*). We estimate the total stress score following Lovibond and Lovibond (1996). The outcome is a dummy indicator that takes the value of 1 if the individual has "High Stress," that is, if the score was greater than 18.	During the past week, how often did you experienced the following?	Durante la semana pasada, ¿que tan frecuente experimento lo siguiente?
			1. I found it hard to wind down	1. Me costo mucho relajarse
			2. I tended to over-react to situations	2. Reaccione de forma exagerada en ciertas situaciones
			3. I felt that I was using a lot of nervous energy	3. Senti que estaba muy nervioso
			4. I found myself getting agitated	4. Note que estaba muy agitado/a
			5. I found it difficult to relax	5. Me costo mucho calmarme
			6. I was intolerant of anything kept me from getting on with what was doing	6. Fui intolerante con lo que me distraia o desconcentraba
			7. I felt that I was rather touchy	7. Senti que estaba muy irritable
			Response options: Never, Rarely, Almost Always, Always	Opciones de respuesta: Nunca, Algunas Veces, Casi Siempre, Siempre
Physical punishment is effective	KI, Baseline and Follow Up Surveys	Dummy indicator that takes the value of 1 for all response options except "No, it's never effective"	Do you think physical punishment is an effective method to discipline children?	¿Piensa que el castigo físico es un método de corrección efectivo?
			Response options: No, it's never effective; It's rarely effective; It's sometimes effective; It's effective most of the time; It's always effective	Opciones de respuesta: No, nunca es efectivo; Casi nunca es efectivo; En ocasiones es efectivo; Casi siempre es efectivo; Si, siempre es efectivo

Notes: This table describes the measures of sensitive outcomes used in the consistency analysis. It also presents the questions –and their respective response options– used to construct the measures, survey sources, and a description of how the sensitive outcome was constructed.

Table A10: Questions used to collect socio-demographic information from caregivers

	Response options			
Panel A: Full Sample				
Gender	Male	Female		
Age	Digit			
How many children aged 8 or younger are you responsible for?	Digit			
What is the age of the eldest child at your care?	Digit			
Do you share the use of this phone number with someone else in your household?	Yes	No		
Do you have a Tigo account associated with this phone number?	Yes	No		
Panel B: Subsample with baseline information				
Highest educational level achieved	Kindergarden (4-5) Preparatoria 1er grado 2do grado 3er grado 4to grado 5to grado 6to grado 7mo grado 8vo grado 9no grado 1er ano bachto. 2do ano bachto. 3er ano bachto. Tecnico sup. incompl. Tecnico sup. compl. Univers. incompl. Univers. compl.			
During the last 6 months, have you been unemployed or self-employed?,	Yes, employed	Yes, unemployed	No	Dont know
Were you employed with a salary at the time of the announcement of the COVID-19 quarantine?	Yes	No	Dont know	Does not reply

Notes: This table lists the questions—and their respective response options—used to collect socio-demographic characteristics of caregivers. Panel A presents the characteristics available for the full sample of the survey-mode experiment. Characteristics that are available only for the group of caregivers that completed the baseline survey are presented in Panel B.

Table A11: Questions Used to Enroll Caregivers

No.	Question	Response options	Eligibility criteria
1	Name		
2	Gender	Male / Female	
3	Cell phone number		
4	Mobile company	Tigo / Claro / Movistar / Digicel	
5	Do you have Tigo money account in the cell phone number?	Yes / No	Tigo
6	Authorize Glasswing to provide your number to Tigo El Salvador so that they can contact you in order to activate your Tigo Money account and be able to receive the benefits	Yes / No	Yes
7	E-mail		
8	Municipality of residence	List of municipalities	
9	Age		18-45
10	Do you have a girl or boy 8 years old or less under your care?	Yes / No	Yes
11	Do the boys (or girls) live in the same house with you?	Yes / No	Yes
12	Of the children under the age of 8 in your care, how old is the oldest child?	Less than 1 / 1 - 8	
13	Do you share this phone with another adult in your household?	Yes / No	

Notes: This table lists the questions –and their respective response options– used to enroll caregivers. Our minimum age was 18, however we have 5 younger observations.