Policy Research Working Paper 11302

The Effect of Survey Mode on Data Quality

Experimental Evidence from Nigeria

Yannick Markhof
Philip Wollburg
Amparo Palacios-Lopez
Pauline Castaing
Akiko Sagesaka
Ivette Contreras

WORLD BANK GROUP
Development Economics
Development Data Group
February 2026

Policy Research Working Paper 11302

Abstract

This paper uses a large-scale experiment in rural Nigeria to study the role of survey mode—in-person versus over the phone—in survey measurement and data quality. The experimental design isolates mode effects from other common sources of errors in surveys and covers 20 outcome measures across topics such as health, labor, shocks, wellbeing, and food security. The findings indicate consistent mode effects across outcomes, with phone responses differing from in-person responses by 17–18 percent at the median. These effects are large relative to other errors in phone surveys, such as under-coverage of households without phones. A within-respondent design enables capturing the full, respondent-level distribution of mode effects and finds them to vary much more than the averages reveal. Respondents with higher education levels are less prone to mode effects, whereas mode effects sharply increase in prevalence as respondents face more answer options. As the reliance on phone surveys in low- and middle-income countries grows, these findings indicate areas with large potential for data quality gains and have first-order implications for economic research in low- and middle-income countries.

This paper is a product of the Development Data 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 pwollburg@worldbank.org. A verified reproducibility package for this paper is available at http://reproducibility.worldbank.org, click **here** for direct access.

The Effect of Survey Mode on Data Quality: Experimental Evidence from Nigeria

Yannick Markhof[1] ⓡ Philip Wollburg ⓡ Amparo Palacios-Lopez ⓡ Pauline Castaing ⓡ Akiko Sagesaka ⓡ Ivette Contreras

Keywords: Data quality, Survey data, Phone survey, Measurement error, Survey mode

JEL codes: C81, C83, O12

[1] The author ordering was constructed through American Economic Association's randomization tool (confirmation code: zfsEk8v3EaGc). The authors are with the Survey Unit, Development Economics Data Group, World Bank. Markhof is with ETH Zurich. The authors may be contacted at pwollburg@worldbank.org. The authors are grateful for feedback from Kibrom Abay, Morgan Hardy, Robert Garlick, Talip Kilic, Travis Lybbert, and from participants and discussants at the 11th ECINEQ conference, the NEUDC 2025 conference, the 5th ICDE conference, the 28th IARIW-WB-CIEM conference, the World Bank conference on Harnessing High-Frequency Survey Data for Development Research in the Polycrisis Era, the World Bank conference on Better Data for Better Jobs and Lives, the 2024 IPA/GPRL Researcher Gathering at Northwestern University, and the IPA-GATES 2024 meeting on Recent Advances in Survey Methods for Low- and Middle-Income Countries. The authors also thank the National Bureau of Statistics in Nigeria for their management of survey implementation, and the team of enumerators for their dedication to data collection. We are grateful to Kevin McGee for his guidance in the sampling strategy for the experiment. We gratefully acknowledge support from the RSB grant "Experimental Evidence on Survey Mode Effects in Nigeria". This paper was produced with financial support from the 50x2030 Initiative to Close the Agricultural Data Gap.

1 Introduction

High-quality socioeconomic data is fundamental for advancing research and for guiding, targeting, and monitoring humanitarian and development interventions. In low- and middle-income countries (LMICs), surveys remain a critical source of representative data and a key tool for "ground truthing" socioeconomic outcomes. In recent years, phone surveys have become an increasingly important mode of data collection in these settings, particularly following the COVID-19 pandemic, which temporarily forced a shift from in-person to remote interviews (Gourlay et al., 2021a; Zezza et al., 2023). Their appeal is clear: phone surveys are less costly than traditional face-to-face surveys and can be implemented quickly and flexibly to provide timely and higher-frequency data. As such, phone surveys offer a pragmatic response to two converging pressures. On the one hand, the growing number and frequency of crises and shocks in LMICs has increased demand for timely, high-frequency data to inform rapid response and to analyze household resilience and recovery (Headey and Barrett, 2015). On the other hand, constrained budgets in both research and international development have reinforced the push for cost-effective approaches to data collection. Against this backdrop, phone surveys have become a prominent and likely enduring feature of the data landscape in LMICs.

Yet, as phone surveys have proliferated, concerns about their data quality remain. Like all data sources, phone (and in-person) surveys are subject to errors from multiple sources, but the survey mode itself is increasingly recognized as an important source of measurement error and bias. Despite growing evidence on phone survey data quality, important gaps remain in understanding the magnitude, direction, and drivers of mode effects, systematic differences in responses attributable to the data collection mode rather than true differences in underlying outcomes, and in developing strategies to mitigate them.

Literature on mode effects is an active field in its nascency, with limited and somewhat inconclusive experimental evidence . Mode effects have been documented for outcomes such as agricultural yields and production (Anderson et al., 2024; Kilic et al., 2021), consumption (Abate et al., 2023), microenterprise data (Garlick et al., 2020), and contraceptive use (Greenleaf et al., 2020), while other outcomes such as dietary diversity and health appear less sensitive to survey mode (Abate et al., 2023; Lamanna et al., 2019; Markhof et al., 2025). Findings on direction vary: for example, Abate et al. (2023) report lower consumption and higher poverty in phone interviews, whereas Greenleaf et al. (2020) find overreporting of contraceptive use. However, most studies cannot fully separate mode effects from confounding factors such as coverage bias or questionnaire differences, nor do they systematically explore mechanisms driving mode effects. More research and replication are needed to establish best practices and guidance for how surveys in LMICs should navigate the issue.

This study provides experimental evidence on survey mode effects based on a large-scale, multi-mode study in rural Nigeria. As part of the experiment, we distributed phones to 937 households in randomly selected rural villages in Nigeria and interviewed them both in person and by phone, using identical questions and randomizing the order of interview modes. This design enables two complementary strategies to identify mode effects: a between-respondent design that compares first interviews across randomized phone and in-person groups, and a within-respondent design that differences outcomes across modes for the same respondent while randomizing interview order. Our setup allows identifying mode effects and

distinguishing them from other confounders and sources of errors including coverage errors,[2] respondent selection effects,[3] timing effects,[4] and ordering/sequencing effects.[5]

We find robust evidence for significant and economically meaningful mode effects across a wide range of outcome variables, including food security, labor and employment, non-farm enterprise ownership, health and subjective wellbeing, weather expectations, and shocks. Responses collected over the phone differ from the in-person estimates by 18% (between-respondent design) and 17% (within-respondent design) at the median. The similarity of the results from between- and within-respondent designs – both in terms of signs and magnitudes of mode effects – lends further credibility to the robustness of our findings. Mode effects are positive for 70% of outcomes, suggesting a tendency for respondents to provide more affirmative or frequent responses by phone.

Leveraging the within-respondent design, we quantify mode effect heterogeneity across individuals and outcomes. We find that respondents with higher education levels are less prone to mode effects. We further document that individual respondents more often answer questions inconsistently across modes than the average mode effects suggest in aggregate. Finally, we document systematic behavioral differences across survey modes: phone respondents are more likely to agree with statements, answer "yes" to binary questions, report extreme values, and provide rounded responses, while mode effects are more likely for variables with more response options.

We then compare mode effects to other sources of errors in surveys. We find that coverage errors from non-universal phone ownership also matter: phone ownership is non-random and associated with gender, age, education, and household wealth. These differences translate into significant coverage effects in about half of the outcomes examined. Coverage effects are typically smaller than mode effects, standing at 21% of the size of mode effects at the median across 20 outcomes. Mode effects and coverage effects commonly have opposite signs. In these cases, both error sources attenuate each other in aggregate and conceal at times substantial sampling and non-sampling error in phone estimates. We show that coverage errors can be effectively mitigated by distributing phones to non-owners and that phone endowment did not influence the responses provided.

Our paper makes a number of contributions. First, we provide the most comprehensive experimental evidence to date on the magnitude and mechanisms of mode effects in LMIC surveys, drawing comparisons across multiple topics and variable types. Using a within-design (Charness et al., 2012; List, 2025), we measure the full, respondent-level distribution of mode effects in a representative sample of the rural population in Africa's largest country. This allows us to identify patterns in mode effects that hold more generally across topics, variables, and respondents and lets us quantify the amount of heterogeneity in mode effect size and direction.

Second, our findings highlight the importance of identifying and mitigating (phone) survey measurement errors for improving data quality in LMICs. Phone surveys address the growing demand for large-scale, high-frequency survey data at affordable cost. At the same time, we show that fully leveraging this potential will require addressing their own set of data quality challenges. Previous literature has largely focused on sampling errors in phone surveys that can be effectively mitigated through re-weighting (Ambel et al., 2021;

[2] Differences between households with and without a phone or differences in response rates between modes (Ambel et al., 2021a).
[3] Differences between respondents in phone and in-person interviews (Glazerman et al., 2023).
[4] Real changes in outcomes over time.
[5] Response behavior being affected by a previously conducted interview (Caceres-Santamaria, 2021; Struminskaya and Bosnjak, 2021).

Brubaker et al., 2021) or phone distribution (as in this study). Further, coverage errors will likely decline in importance as mobile phone penetration expands (GSMA, 2024). Conversely, we know much less about the extent, nature, and mitigation of measurement errors in phone surveys. Unlike coverage errors, measurement errors can threaten the *internal* validity of estimates and thus remain first-order concerns even when external validity is not a primary objective, as in the case of randomized controlled trials. Our study shows that mode effects can match or exceed coverage errors in their size despite having thus far received much less attention from survey researchers and users of phone surveys. Further, we demonstrate that mode effects and coverage errors can be cumulative, or they can have opposite signs. In the latter case, aggregate bias will mask the true extent of internal and external validity issues in the data.

Third, our study substantially advances our understanding of survey measurement errors and identifies areas with possibly large data quality returns: (i) understanding the sources of behavioral differences among respondents by survey mode; (ii) mitigating measurement errors ex-ante through survey design (e.g., questionnaire design or respondent incentives) and fieldwork protocols (e.g., enumerator training and assignment to interviews, or real-time data quality checks); and (iii) using ex-post statistical adjustments for respondent-level measurement error (e.g., based on objective benchmarks and response quality checks, or by differencing out within-respondent measurement errors via longitudinal data collection). Future research could bring about large data quality gains by investigating these issues further.

The remainder of the paper is structured as follows. Section 2 outlines the conceptual framework on sources of error in household surveys. Section 3 describes the identification strategy and data. Section 4 presents the main results on survey mode effects, while Section 5 explores potential sources of these effects. Section 6 examines other survey errors. Section 7 discusses the findings, and Section 8 concludes.

2 Conceptual Framework

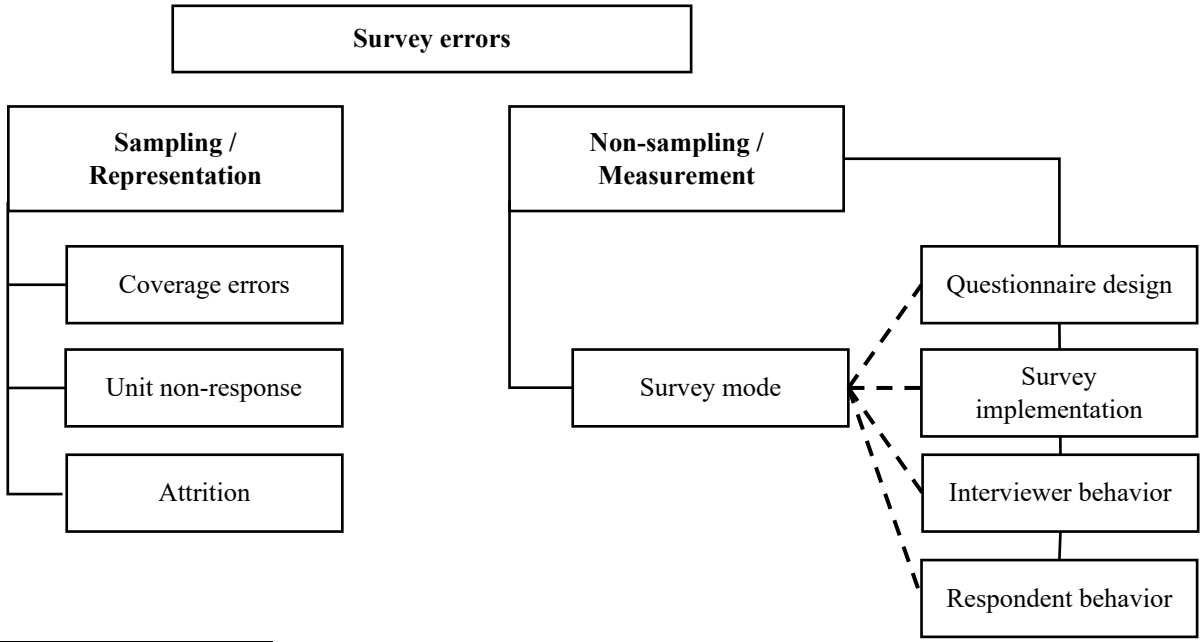
The concern for phone survey mode effects is part of a larger interest in the quality of data on low- and middle-income countries. Like all data sources, survey data is subject to errors originating from a range of sources, that is, an observed outcome measured through a survey for individual i, Y_i, is a function of the true outcome Y_i^* and error e_i:

$$Y_i = Y_i^* + e_i \quad (1)$$

If the error e_i is systematic across units, the observed Y_i is biased, while if the error is random, the variance of the estimates increases.[6] The survey mode, in our case phone versus in person, is one source of survey error. As conceptualized in the Total Survey Error framework, survey design choices at all stages of the survey lifecycle are sources of survey errors (Biemer, 2010). We can distinguish between, on the one hand, sampling or representation errors and, on the other hand, non-sampling or measurement errors.

Mode effects are classified as measurement errors. Other important sources of measurement errors are related interviewer behavior, respondent behavior, questionnaire design, as well as other aspects of survey implementation, which may interact with survey mode. These include, for example, survey length, which may induce respondent fatigue and inattention (Abay et al., 2021); incentives paid to respondents (Gibson et al., 2019); and repeated interviews at high frequencies which may cause ordering or sequencing effects (Schündeln, 2018), whereby the participation in a recent interview conditions survey responses in a subsequent interview. These survey errors may interact with and vary by survey mode and contribute to mode effects (Figure 1). For example, respondents may behave – and respond – differently when faced by an interviewer in person as opposed to hearing them over the phone or they may fatigue more easily over the phone.

Figure 1. Survey errors in phone surveys in LMICs



[6] For binary and categorical variables, all survey error, whether systematic or not, will lead to biased estimates (non-classical measurement error) (Abay et al., 2023).

Phone surveys in LMICs are also subject to sampling errors. Incomplete coverage of the target population (not everyone has access to a phone) and unit non-response (not everyone responds to their phone) are considered first-order concerns for the accuracy and representativeness of LMIC phone survey data (Ambel et al., 2021b; Brubaker et al., 2021; L'Engle et al., 2018). Phone surveys also tend to face higher rates of attrition than in-person surveys (Özler and Cuevas, 2019). Of course, in-person surveys also suffer from sampling and measurement errors (Abay et al., 2023; Weerdt et al., 2020a). Figure 1 summarizes the major sources of measurement and sampling errors relevant to phone survey implementation in LMICs.

In our experimental setup, we have two measurements for the same true outcome (Y_i^*), one by phone survey (Y_i^P) and one by in-person survey (Y_i^F):

$$Y_i^P = Y_i^* + e_i^P, \tag{2}$$

$$Y_i^F = Y_i^* + e_i^F \tag{3}$$

The error terms e_i^P and e_i^F represent the combined effect of all survey errors for phone and in-person surveys, respectively, including mode-related measurement error as well as other survey design factors (Figure 1).

To capture the mode effect, δ, we compare the two measurements for Y_i^*:

$$\delta_i \equiv Y_i^P - Y_i^F = e_i^P - e_i^F \tag{4}$$

Under the condition that all other survey design factors, and thus their associated error components, are held constant across modes, this difference isolates the mode effect, defined as the systematic difference in measurement error between phone and in-person interviews. In the next section, we outline how our experimental design ensures these conditions can be met.

Note that we do not have an external, objective benchmark for the true Y_i^*, so we cannot ascertain the values of e_i^P and e_i^F nor whether the phone or in-person measurements are closer to the true outcome.

3 Experimental Design and Data

3.1 Experimental design

To identify the effect of survey mode (phone versus in-person) on measured outcomes, we conducted a randomized survey experiment implemented as part of the fieldwork for Nigeria's General Household Survey Panel Wave 5 (GHS-Panel W5) 2023/24. This survey targeted 5,067 households in 519 enumeration areas to be interviewed nationwide, following a two-visit schedule: a post-planting interview (July–September 2023) and a post-harvest interview (January–March 2024). Among the 519 enumeration areas included in the GHS-Panel W5, 455 rural enumeration areas were eligible for the experiment and we randomly selected 106 for our experimental sample.[7]

[7] Enumeration areas eligible for the experiment were those where at least 60% of households were engaged in agricultural activities during the GHS-Panel Wave 4 2018/19. The focus on agricultural EAs was intended to ensure coverage of a broad set of outcomes relevant to rural livelihoods, including agricultural employment.

We targeted all 995 survey households in the selected enumeration areas, of which 49 could not be located during the post-planting visit, and 9 refused to participate. The final experimental sample includes 937 households across 105 enumeration areas. All were provided with mobile phones and sim cards during the post-planting visit and were assigned to in-person and phone interviews during the post-harvest period. These households present similar characteristics as those eligible but not selected for the experiment (Balance A1). Attrition from the target sample to the final experimental sample therefore did not compromise our experiment's representativeness of the broader agricultural population in Nigeria.

Households in the experimental sample were targeted to be interviewed both in-person (as part of the regular GHS-Panel post-harvest fieldwork) and over the phone just a few days apart. Respondents answered a set of identical survey questions across both modes. However, we randomized the order of the two interviews across households, creating two treatment groups. Treatment Group 1 (T1) includes 464 households that were first interviewed over the phone (T1-P) and then in-person (T1-F). Treatment Group 2 (T2) consists of 473 households that were first interviewed in-person (T2-F) and then over the phone (T2-P). Figure 2 summarizes our experimental design.

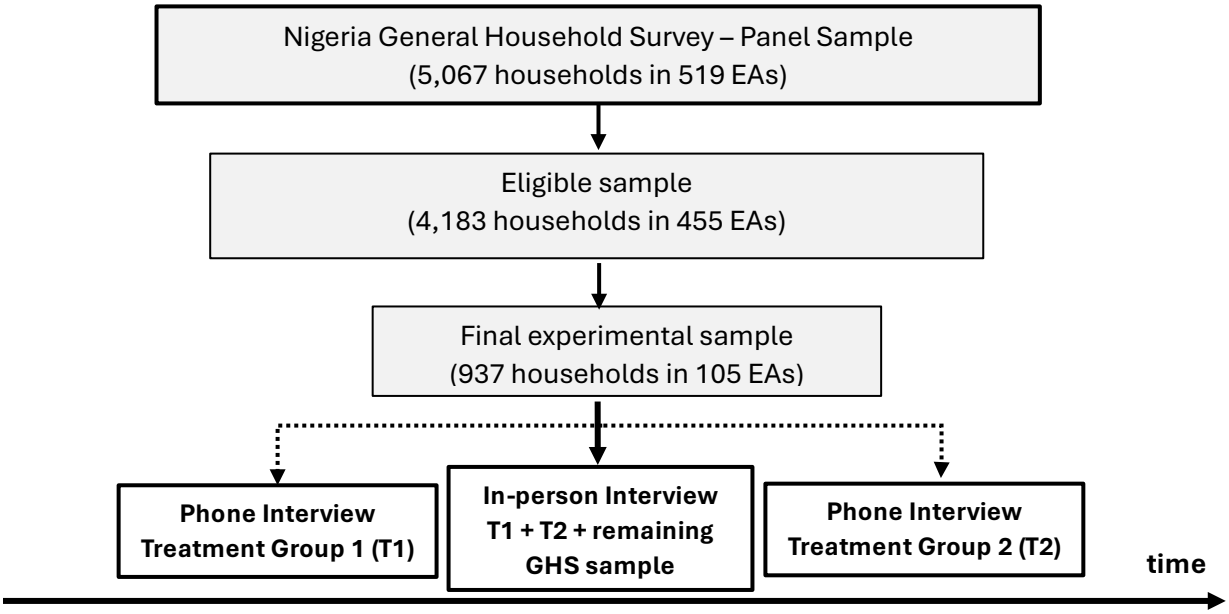
Within households, the primary respondent for the phone survey was the individual most knowledgeable about agricultural and work activities of the household.[8] In addition, we aimed to interview a second household member aged 15 or older over the phone.[9] Each phone respondent who completed the interview received 1,500 naira credited to their phone. This approach increased the likelihood of having matched respondents across both survey modes, allowing for a substantial number of direct comparisons of responses from the same individuals. Overall, 862 households had at least one member who completed both the phone and in-person interviews, and 852 households had the main respondents complete both interviews (Table A2). These two groups of households constitute the main analytical samples for the identification strategies outlined in section 3.2: the between-respondent design (852 interviews with households' main respondent, randomly split into T1-P and T2-F) and the within-respondent design (in-person and phone interviews in random order for 862 households with at least one identical respondent).[10]

[8] This choice was motivated by the design of a separate experiment in which the same households were called monthly between the post-planting and post-harvest visits to report on agricultural outcomes.

[9] The second respondent was the individual who answered the questions in the consumption module during the post-planting visit. If this person was also the main respondent, another adult household member was randomly selected instead. Of the experimental sample of 937 households, 862 households had at least one member complete a phone interview, and in 769 of these households, two members completed the interview. The post-harvest in-person interview was completed by 933 households.

[10] As such, the within design includes at least twice the number of interviews as households were covered (each household is interviewed in-person and over the phone). Even though we do not use respondent's second interviews (T1-F and T2-P) as part of the between-design, we still limit the sample to main respondents who answered both modes in order to mitigate possible imbalances in respondent, not just household, characteristics between the two treatment arms we compare (T1-P and T2-F).

Figure 2. Experimental design



3.2 Causal identification of mode effects

Based on this setup, we use two strategies to identify and estimate the mode effect.

Between-respondent design. The first strategy, which we refer to as the between-respondent design, compares outcomes across two groups of respondents, randomly assigned to be interviewed first either by phone or in-person: Treatment Group 1 (T1-P), which was first interviewed by phone and Treatment Group 2 (T2-F), which was first interviewed face-to-face. We express this with the following equation:

$$Y_{ih} = \beta_0 + \beta_1 M_{ih} + \varepsilon_{ih} \quad (5)$$

where Y_{ih} is the outcome of interest reported by individual *i* from household *h;* M_{ih} is a dummy indicating whether the interview of individual *i* was conducted over the phone (*M=1*) or in-person (*M = 0);* and ε_{ih} is the error term. β_0 is the intercept. The coefficient β_1 captures the average difference in reported outcomes between phone and in-person interviews.

Within-respondent design. The second strategy exploits the panel structure of the experiment, comparing the same respondents across modes to compute a within-respondent estimate of average mode effects. This approach increases statistical power, controls for time-invariant individual confounders, and allows us to measure mode effects at the respondent level. For each outcome of interest, we use the sample of respondents who provided responses in both the phone and in-person interviews (in random order). The outcome for individual *i* in interview $s \in \{1,2\}$ is modeled as:

$$Y_{ihs} = \alpha_{ih} + \theta_1 M_{ihs} + \varepsilon_{ihs} \quad (6)$$

where Y_{ihs} is the outcome of interest reported by individual *i* from household *h* in interview *s;* α_i is an individual effect; M_{ihs} is equal to 1 if interview *s* for individual *i* was conducted over the phone and 0 if in person*;* and ε_{ihs} is the error term. The coefficient θ_1 captures the average effect of phone (relative to in-person) interviews on the measured outcome, holding individual-specific factors constant.

For the coefficients θ_1 and β_1 to causally identify the mode effect, the following conditions should hold. These conditions are either controlled for by our experimental design or testable such that we do not rely on strong assumptions for identification.

(i) Random assignment. First, interview mode is randomly assigned and independent of both the true outcome and the residual errors, i.e., $M_{ih} \perp Y_{ih}^*, \varepsilon_{ih}$. For the between-respondent design (equation 5), we restrict the sample to one respondent per household, the main respondent of the in-person interview, as randomization is at the household level while interviews are with individual respondents. Restricting the sample to one respondent type per household ensures balance on observables across a range of household and respondent characteristics show no systematic differences between T1-P and T2-F (joint orthogonality p=0.7), with the only imbalance a slightly higher dependency ratio for in-person households (Table A3). For the within-respondent design (equation 6), we rely on the random sequencing of mode assignment to ensure that the mode is independent of unobserved, time-varying determinants of the outcome, i.e., $M_{ihs} \perp \varepsilon_{ihs}$. The mode order was randomized across households, and we confirm that the resulting assignment is balanced on observables at the household and respondent level (Table A3).

(ii) No timing effects. Second, true outcomes Y_{ih}^* need to remain stable between the phone and in-person interviews. In the between-respondent design, this could be violated if the phone interviews and in-person interviews of the two treatment groups were conducted at different points in time and the true outcome changed over time. To mitigate this concern, both surveys were conducted concurrently, with T1 phone interviews on average just four days earlier than T2 in-person interviews, and with sufficient overlap to make time-related confounding unlikely (Figure A1). In the within-respondent design, the random sequencing of mode ensures that survey timing for the average in-person interview does not differ from survey timing for the average phone-interview (Table A4). To causally interpret respondent-level mode effects, we further require outcomes to be stable between a respondent's two interviews. We keep the time that passes between both interviews of the same respondent to just under 6 days on average (Figure A2), an interval sufficiently short to make meaningful changes implausible for most outcomes. Table A7 confirms this with regressions of outcomes on interview date and mode effect, showing little role for timing.

(iii) Comparable survey design. Third, survey design related sources of errors (including both sampling and non-sampling) are held constant across treatment groups. Our experimental design ensures comparable conditions along key survey dimensions. Critically, by distributing mobile phones, we ensure equal coverage across modes.[11] Respondents are asked identical questions over the phone and in-person. As we restrict analysis to respondents' first interviews in the between design, we can rule out that spillovers from the first to the second interview (ordering effects) could confound the mode effect.

(iv) No ordering effects. Under the first three conditions, both $\hat{\beta}_1$ and $\hat{\theta}_1$ identify the average mode effect. In the within-respondent design, however, the same respondent answers the same question twice in the span of a few days, making it plausible that ordering or learning effects could influence responses in the second interview (Schündeln, 2018). Such effects may attenuate or inflate the mode effect for any given respondent. To be able to interpret respondent-level mode effects and conduct heterogeneity analysis, we therefore need to rule out such ordering effects. We can investigate ordering effects by testing whether measured outcomes Y_{ihs} systematically differ depending on whether an interview was conducted as the respondent's first or second interview (order being randomly assigned). We additionally include an interaction between mode and interview sequence:

[11] One potential concern could be attrition if respondents who drop out of the phone interview arm differ systematically from those who drop out of the in-person arm. Reassuringly, we find our final analytical sample to be balanced in respondent characteristics even after attrition and non-response (Table A3).

$$Y_{ihs} = \alpha_i + \gamma_1 M_{ihs} + \gamma_2 V_s + \gamma_3 M_{ihs} \times V_s + \varepsilon_{ihs} \quad (7)$$

with Y_{ihs}, α_i, M_{ihs}, and ε_{ihs} as before. V_s indicates whether this is the second interview of the respondent (V = 1) or not (V = 0)*;* $M_{ihs} \times V_s$ is their interaction; and ε_{ihs} is the error term. The estimate of γ_2 captures the ordering effect for in-person interviews, while $\gamma_2 + \gamma_3$ captures the ordering effect for respondents interviewed over the phone. The coefficient γ_3 tests whether the ordering effects differ by survey mode. We find strong evidence supporting that the order of interviews does not affect survey responses (see Tables A6 and A7).[12]

Under conditions (i) – (iii), interview mode is orthogonal to both the true outcome and all other sources of survey error and the observed difference in average outcomes by mode isolates the mode effect. This allows us to interpret the estimated coefficients $\hat{\beta}_1$and $\hat{\theta}_1$ as equal to δ from the conceptual model in equation 4.

With the addition of condition (iv), we can proceed to interpret and analyze respondent-level mode effects and use this to study heterogeneity. We create a stacked dataset at the respondent-outcome level which allows us to pool our analysis and identify factors that systematically matter across outcomes. We calculate mode effects for each respondent and outcome as the simple difference between measured outcomes over the phone and in-person. Second, for each respondent and outcome, we create a dummy variable indicating whether there is *any* mode effect (binary and categorical variables). This lets us estimate the probability that there is any mode effect across outcomes:

$$Y_{ihk} = \tau_1 + \varepsilon_{ihk} \quad (8)$$

where Y is now defined at the respondent-by-outcome level and $\hat{\tau}_1$ is the estimated probability of a mode effect (binary and categorical variables). To identify respondent characteristics that correlate with the emergence of mode effects and survey design factors mediating them, we estimate the following equation

$$Y_{ihk} = \beta_1 \boldsymbol{X_i} + \beta_2 \boldsymbol{H_i} + \beta_3 \boldsymbol{E_i} + \vartheta_k + \varepsilon_{ihk} \quad (9)$$

where *Y* is defined as in equation (8), $\boldsymbol{X}$ is a vector of respondent characteristics, $\boldsymbol{H}$ a vector of household characteristics,[13] $\boldsymbol{E}$ a vector of enumerator controls,[14] and ϑ_k is a set of outcome fixed effects. Additionally, we estimate a version of equation (9) that drops X, H, and E and replaces ϑ_k with a list of survey design characteristics, such as the number of answer options (outcome levels) a variable has.[15]

[12] In addition, we would expect anchoring effects to only attenuate estimated mode effect heterogeneity if respondents try to answer questions consistently between the two interviews.

[13] Household size and log total consumption.

[14] For respondents, we include sex, age, education, whether the respondent speaks the interview language at home, and whether they were part of phone surveys conducted in the previous few months. As enumerator controls, we include sex, age, whether they speak the interview language at home, and had worked on five or more surveys in the past (approximately the median) as there was barely any variation in enumerator education. As enumerators differ between the phone and in-person interview and either may give rise to mode effects, we include the same set of enumerator characteristics for both the in-person and phone survey enumerator. We also include a variable for whether enumerator and respondent had the same gender.

[15] For instance, binary variables have two levels, the locus of control questions have three levels (agree, neither agree nor disagree, disagree), the income adequacy question has four (well, fairly well, fairly, with difficulty), the weather extremes question five (extremely likely to extremely unlikely), and the subjective wellbeing questions have ten (step 1-10).

3.3 Outcomes of interest

The biases in survey data that we investigate likely depend on the nature of the data collected, such as the topic, the unit of observation, and the type of data collected. A distinct strength of our experiment compared to the existing evidence is the ability to analyze sampling and measurement errors across a range of variables that reflect these factors. We selected outcome variables that are either key variables of interest during crises and shocks where phone surveys may be used as a rapidly deployable data collection mode or because they lend themselves well to longitudinal, high-frequency data collection in a mixed mode survey system. The reliability of these variables when measured over the phone compared to in-person is thus of particular interest and relevance. Further, we explicitly selected variables to include both household- and respondent-level variables, as well as different variable types (binary, continuous, count, or categorical). All 20 retained outcomes are presented in Table B1 in Appendix B.

4 Main Estimates of Mode Effects

We first present evidence on the prevalence, magnitude, and direction of survey mode effects. The results offer insights into average differences in responses between phone and in-person interviews. We find consistent evidence of meaningful survey mode effects across a broad range of outcomes.

Table 1 shows the estimated mode effects from both the between-respondent (BR) and within-respondent (WR) designs. Differences between phone and in-person responses are statistically detectable across nearly all topical domains and variable types covered in the questionnaire. The two specifications yield mode effect estimates that are economically meaningful and highly consistent across the within- and between designs, reinforcing the robustness of our results (Figure 3). Responses collected over the phone differ by zero to 63% relative to the in-person mean, with the median mode effect at 18% for between-design estimates and 17% for within-design estimates. The only outcome for which we do not detect a statistically significant mode effect on average is the perceived likelihood of adverse future weather events. In contrast, the health expenditure variables (incidence and amount) are outliers in the distribution of mode effects (health expenditure incidence: 199% and 198% for between and within design respectively; health expenditure amount: 204% and 255% for between and within design, respectively). Figure A3 in the appendix shows the distribution of 20 mode effect estimates for each design, expressed as percentage change compared to the in-person mean.

Mode effect sizes differ somewhat by topical domain: the largest mode effects are observed in the health domain. Besides the outlier values for incidence and value health expenditures, respondents also report significantly higher COVID-19 vaccination rates (between-respondent: 63%, within-respondent: 62%) and more frequent contact with health practitioners (BR: 56%; WR: 42%). Substantial differences are also found in reporting of household enterprise operations (BR: 61%; WR: 50%). More moderate differences are found for well-being (BR: 11%–18%; WR: 12%–17%), food consumption scores (BR: 24%; WR: 19%), and exposure to shocks (BR: 18%; WR: 17%). The magnitude of mode effects in the locus of control domain varies across statements, with larger effects for beliefs that life is shaped by external forces such as fate or other household members (BR: 33%–43%; WR: 34%–42%) and smaller effects for perceived self-control over life outcomes (BR: 7%; WR: 7%). The smallest mode effects are observed for dietary diversity scores (BR: 4%; WR: 2%).

Figure 3.Main estimates of survey mode effects

Estimated mode effects are statistically significant and positive for 70% of outcomes, with only 5 out of 20 estimates being statistically significant and negative. Respondents interviewed over the phone systematically report a higher incidence of health events, greater household enterprise activity, higher food consumption, greater exposure to shocks, and stronger agreement with locus of control statements. These patterns suggest a general tendency for phone interviews to elicit more affirmative or frequent reporting. Notable exceptions to this pattern appear for subjective well-being outcomes: phone respondents report lower well-being for themselves, their friends, and neighbors, but express greater optimism about their own future. Additional exceptions include income adequacy and hours worked in agriculture, both of which are reported at lower levels in phone interviews.

The presence of substantial mode effects across a wide range of outcomes underscores the need for a deeper understanding of their underlying drivers and the potential levers available to survey designers to mitigate them. In what follows, we leverage respondent-level estimates of mode effects to identify systematic patterns related to survey design features, as well as household and enumerator characteristics that shape the presence and magnitude of mode effects.

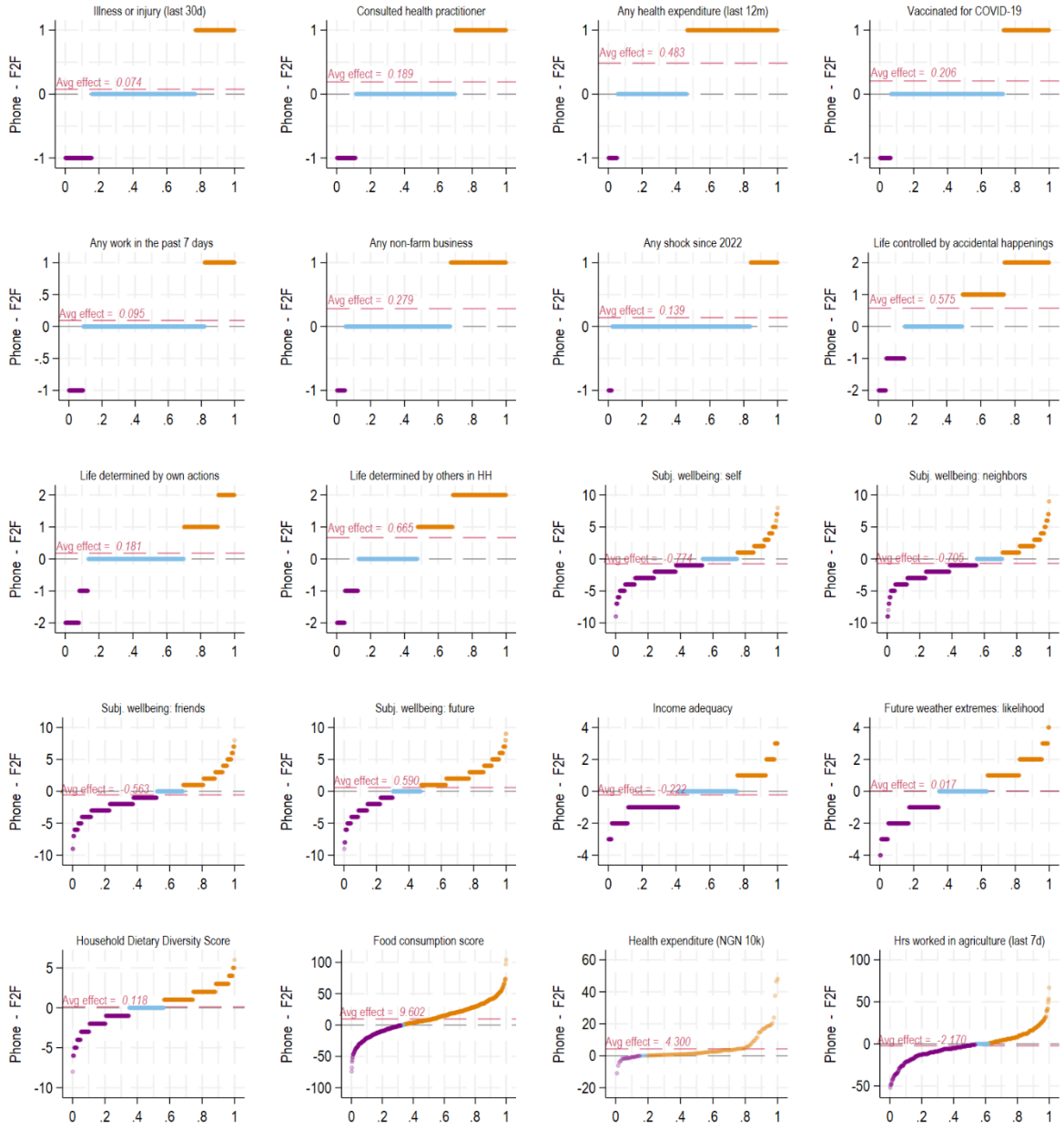
5 Respondent-Level Mode Effects and Heterogeneity

The within-respondent design makes it possible to observe respondent-level mode effects, that is, the difference in response between the phone and in-person interviews, for the same respondent, across our set of outcome variables. This setup allows us to decompose the average mode effects to examine the distribution of respondent-specific mode effects as well as to explore their heterogeneity by household and respondent characteristics, survey design and implementation features.

5.1 Distribution of respondent-level mode effects

As we visualize in Figure 4, there is a lot of heterogeneity in the direction and size of respondent-level mode effects underlying the average mode effects reported in section 4. Across outcome variables, there is a significant share of respondents whose responses differ between phone and in-person interviews – but not necessarily in the same direction as the average mode effect. For example, we find that, on average, respondents are 7 percentage points (18%) more likely to report having been ill or injured in the last 30 days when interviewed by phone compared to in person. Decomposing this average effect, we show that 23% of respondents report having been ill or injured when interviewed by phone but not when interviewed in person. At the same time, 16% of respondents report an illness in-person but not over the phone. This diverging pattern, which we find to some extent in all outcome variables, is masked in average mode effect estimates as respondent-level mode effects of different signs cancel each other out. A case in point is the question of the likelihood of adverse impacts of future weather extremes, the only outcome for which we did not find a statistically significant average mode effect. The fact that we do not detect an average mode effect comes about, as Figure 4 shows, simply because respondent-level mode effects with positive and negative signs cancel each other out almost perfectly. In reality, 71% of respondents do not give a consistent answer across both modes. Overall, the share of inconsistent responses between modes is 55%, ranging from 18% to 59% for binary variables and from 43% to 84% for categorical variables (Table A9). Contrasting these sign-invariant mode effects measures with mode effects estimated in aggregate shows that mode effects are much more prevalent than what aggregates suggest.

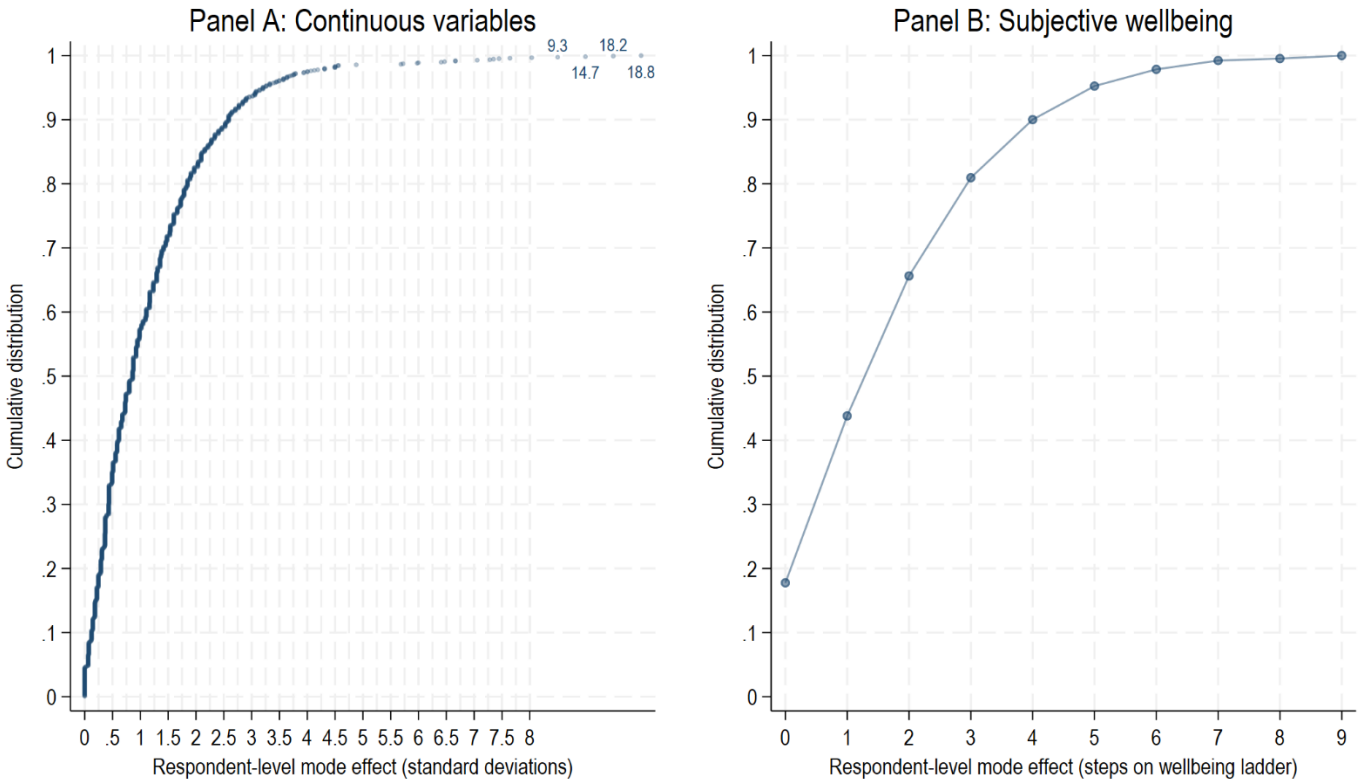
Figure 4. Distribution of respondent-level mode effects by outcome



Note: Distribution of respondent-level mode effects. Each marker represents the mode effect for one respondent, calculated as the difference between the respondent's phone survey answer and their in-person survey answer. Markers above (below) the zero line mean that the outcome value measured in the phone survey was higher (lower) than when collected in-person. Markers on the zero line reflect a consistent answer across both modes (no mode effect). The average of the respondent-level mode effects yields the within-estimate of the mode effect, the estimate of which is shown as the red dotted line.

When looking at continuous variables only, the absolute magnitude of mode effects varies between different respondents (Figure 5). Mode effects are within 0.1 standard deviations of the in-person mean for 9% of respondents and within 0.2 standard deviations for 15% of respondents when considering all continuous outcomes.[16] Mode effects exceed one standard deviation for 43% of respondents and distributions have long tails. Such dispersion is also visible in a number of discrete, multi-valued outcomes in our sample. For example, looking at the subjective wellbeing questions, answers over the phone and answers in person differ by no more than one step on the 10-step subjective wellbeing scale in 43% of cases. Conversely, for a third of respondents, they differ by three or more steps.

Figure 5. Distribution of mode effects at respondent level



Note: Panel A plots the cumulative distribution of respondent-level mode effects (expressed in standard deviations of the outcome when measured in-person) for all continuous outcome variables (food consumption score, health expenditure, hours worked in agriculture). Panel B does the same for the subjective wellbeing which is measured on a scale between 0 and 10.

Next, we compare respondent-specific mode effects across outcome variables for the same respondents. This enables us to assess whether mode effects are correlated within individuals, that is, whether some respondents exhibit systematic, multi-variable discrepancies between phone and in-person interviews, or whether mode effects are variable-specific. Focusing on the 16 binary and categorical outcome variables, Figure 6 plots the percentage of outcomes with a mode effect for each respondent. No respondent answered all questions consistently across modes. At the same time, fewer than 10% of respondents give an inconsistent answer for 80% or more of the outcomes. Respondents tend to provide inconsistent responses

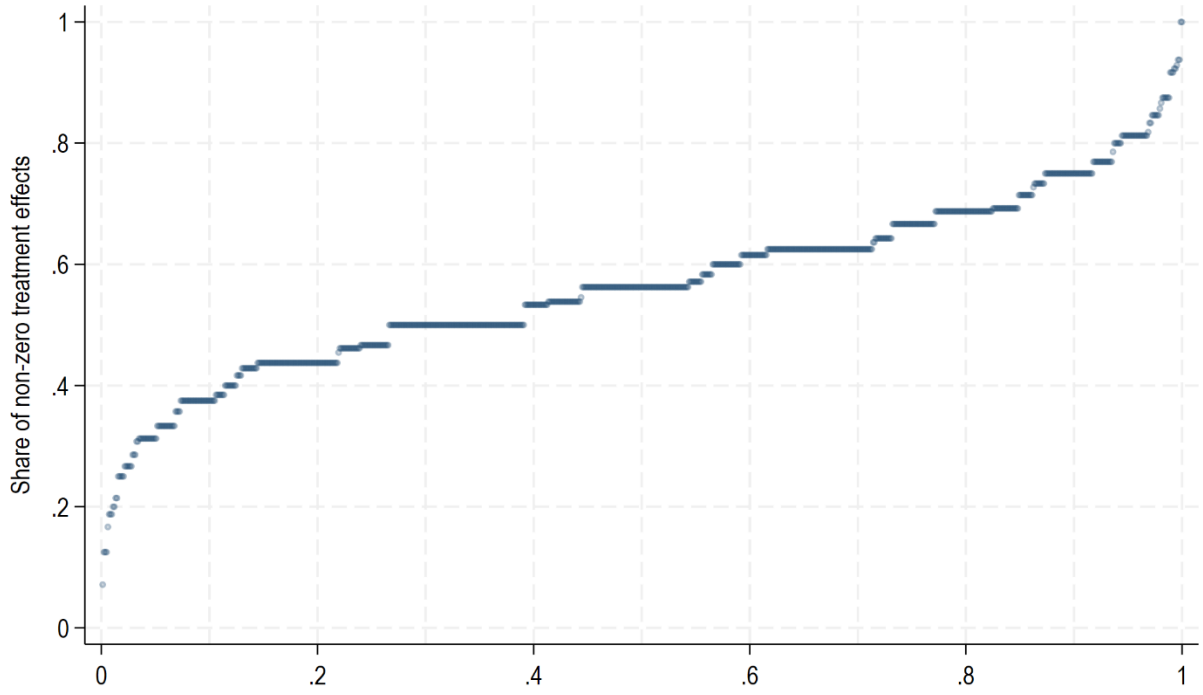
[16] Food consumption score, health expenditure, hours worked in agriculture.

for about half of the outcomes. For the median respondent, answers over the phone and in person differ for 56% of the outcomes. It therefore appears that many respondents report inconsistently sometimes and consistently other times, rather than there being a group with mode effects and one without mode effects.[17]

5.2 Mode effects by respondent and enumerator characteristics

We investigate whether variation in the likelihood of mode effects is systematically correlated with respondent characteristics (Table 2).[18] We find answers from more educated respondents to be less susceptible to mode effects. Answers from a respondent with primary education are about 4 percentage points less likely than from a respondent with no primary education to differ over the phone and in person. Point estimates continue to be consistent with the pattern of mode effects becoming less likely among more educated respondents. Compared to those without formal education, we estimate that respondents with secondary education are 5 percentage points less likely and those with tertiary education 7 percentage points less likely to give two different answers when interviewed in-person and over the phone. We do not find the likelihood of mode effects to be related to respondent gender, age, and familiarity with the interview language.

Figure 6. Share of non-zero mode effects (binary and categorical variables only)



[17] Naturally, if respondents satisfice such as pick answers at random, we would expect to see a degree of chance agreement between survey modes even if reporting under one mode is just noise.
[18] As enumerator assignment was not randomized and the set of enumerators differed between in-person and phone interviews, we control for a number of enumerator characteristics in these regressions.

5.3 Respondent behavior and mode effects by survey design factors

Several patterns emerge from Figure 4 on how respondent behavior varies by survey mode. First, respondents are consistently more likely to agree with statements over the phone than in person. This pattern holds across all binary variables and is especially pronounced for the locus of control questions, which offer respondents three response options (agree, neither agree nor disagree, disagree).

Second, the likelihood of inconsistent responses between modes increases sharply with the number of discrete answer options. While 36% of respondents give inconsistent answers for binary questions, this share increases to 58% for the three-level locus of control questions, 65% for income adequacy (four levels), 71% for weather extremes (five levels), and 82% for subjective wellbeing (ten levels). On average, each additional response option is associated with a 5 percentage point increase in the likelihood of inconsistency (Figure 7).[19]

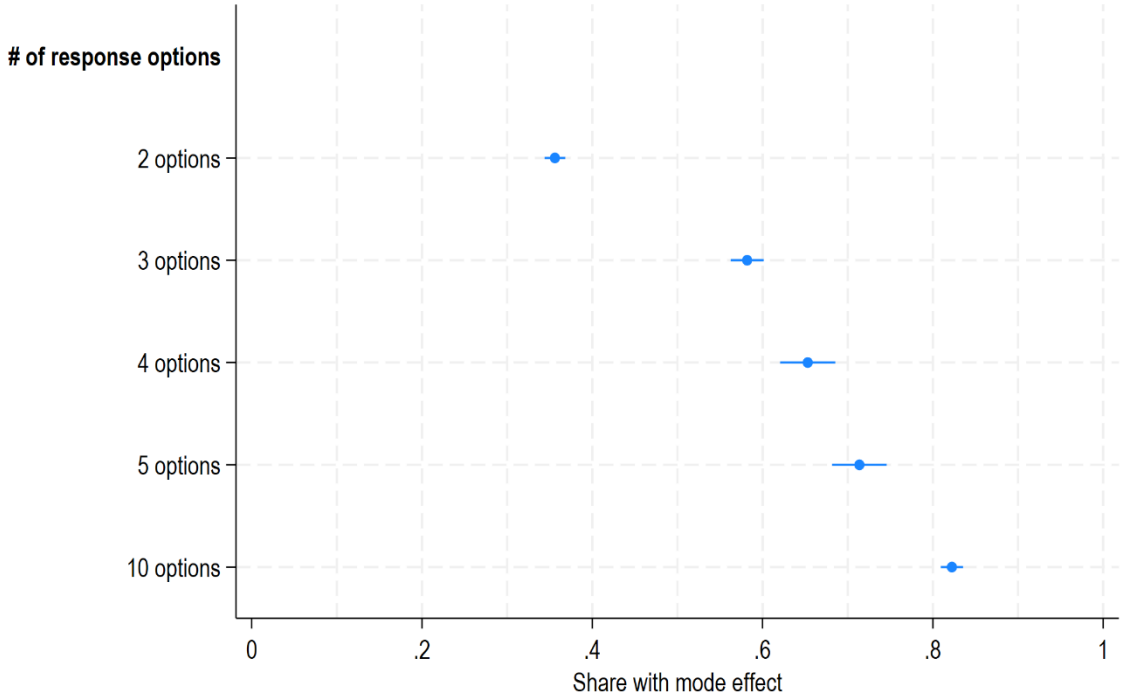
Third, respondents tend to report more negatively about their current economic situation by phone, for both subjective wellbeing and income adequacy, while paradoxically being more optimistic about their future economic situation.

We investigate these patterns more formally in a series of regressions (Table 3). Phone respondents are significantly more likely to answer "yes" to binary outcomes (+19.4 p.p.) and to agree with locus of control statements (+34.7 p.p.) relative to their in-person responses. Multinomial logit results (Supplementary Table A5) show that this shift largely comes from reductions in "neither agree nor disagree" (-22.5 p.p.) and "disagree" (-9.6 p.p.) responses, indicating a strong tendency to confirm statements when interviewed by phone. Mode effects are also evident in other response behaviors. Respondents are more likely to give extreme values on subjective wellbeing questions when interviewed by phone (+16.7 p.p.) and to report rounded values in continuous measures such as health expenditures and agricultural work hours (+11.8 p.p.) (Table 3).

Taken together, these results document systematic differences in response patterns by survey mode. While these patterns are correlational and not causal, they are consistent with differences in cognitive processing or engagement between phone and in-person interviews. However, absent a verifiable and objective benchmark for the outcome variables, we cannot conclusively determine the bias relative to the true outcome value in phone versus in-person surveys.

[19] Of course, more response options naturally allow for more heterogeneity in responses and a lower likelihood of chance agreement even if respondents report 'as if random.' The chance of agreement when responding 'as if random' is 1/k, with k response options. The agreement of reported responses exceeds 'as if random' at all points by between 8 and 14 percentage points, but the decline in agreement tracks the decline in the probability of agreement when responding 'as if random.'

Figure 7. Share of inconsistent responses between the phone and in-person interview, by number of response options



Note: The figure plots the share of inconsistent answers by the same respondent when interviewed in person and over the phone, by the number of discrete answer options (or 'levels') of the variable. Binary variables have two levels, the locus of control questions have three levels (agree, neither agree nor disagree, disagree), the income adequacy question has four (well; fairly well; fairly; with difficulty), the weather extremes question five (extremely likely to extremely unlikely), and the subjective wellbeing questions have ten (step 1-10).

6 Mode Effects and Other Survey Errors

Phone and in-person surveys are subject to a range of potential biases stemming from sampling and survey design choices (section 2). In our experimental design, we have sought to minimize the impact of these errors to be able to credibly identify the mode effect (section 3). However, we can observe some of these errors in our data and compare their size to the mode effect.

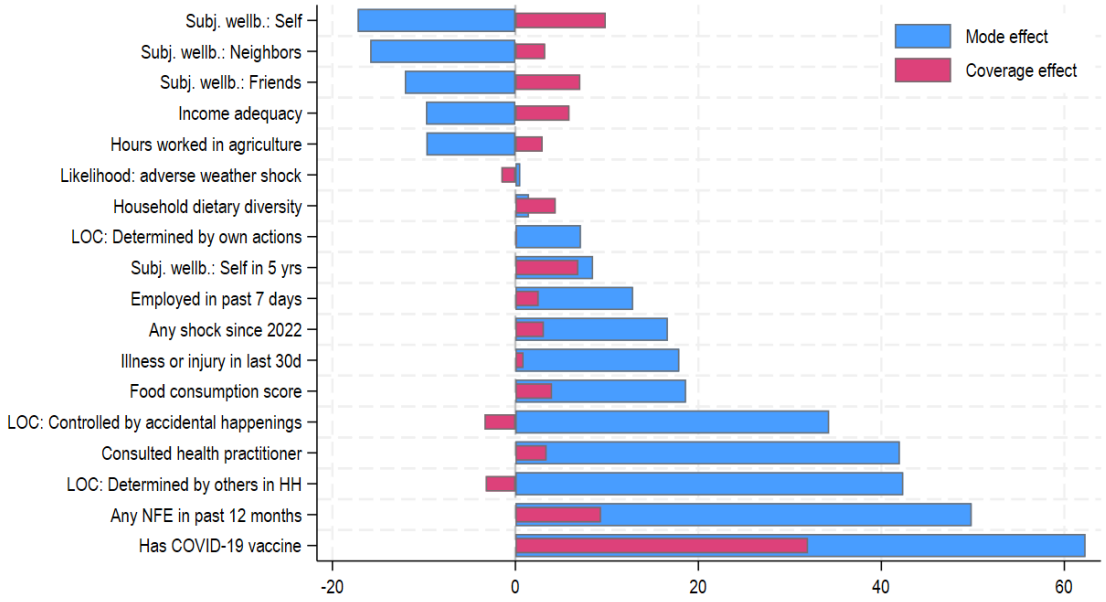
Coverage errors

Absent universal phone ownership in LMICs, coverage error is a first-order concern for phone surveys in these countries (Ambel et al., 2021b; Brubaker et al., 2021; Gourlay et al., 2021b). In this experiment, we distributed mobile phones to obtain representative estimates by phone and eliminate coverage differences when comparing phone and in-person responses. This setup allows us to estimate coverage errors by comparing the responses of selected households with previous phone ownership to those without (which would be ineligible to participate in a regular phone survey that does not provide mobile phones).

Phone ownership is non-random with men, older and more educated individuals, and larger and wealthier households significantly more likely to own a mobile phone (Table A10). Outcomes of interest differ

between households that were previous phone owners and those that were not. Previous phone ownership is also associated with significant differences in the outcomes of interest (Table A11, Figure 8): households that were previous phone owners have significantly higher health expenditures (70%) and score higher across the various subjective wellbeing variables (3-7%); they also report higher income adequacy (6%), greater household dietary diversity (4.4%) and food consumption score (4%), and they are more likely to run a household business (9.4%). We find a statistically significant coverage effect in about half of outcome variables. For eight out of twenty outcomes, mode effects and coverage effects have opposite signs. In these cases, both error sources attenuate each other when viewed in aggregate. Paradoxically, this may lead to estimates that appear less biased in aggregate but that conceal substantial underlying sampling and non-sampling errors. In most cases, the coverage effects are smaller in absolute size than mode effects, standing at 21% of the mode effect at the median (Figure A4).

Figure 8. Coverage effect size relative to mode effects



A caveat to this comparison is our limited ability to realistically assess unit non-response in phone surveys. Non-response was very low in this experiment (8%), thanks in part to generous incentives and the provision of mobile phones. In regular phone surveys, unit non-response is generally much higher, even in list-based phone surveys.[20] For example, Gourlay et al. (2021) report non-response rates of 35% for the LSMS-High Frequency Phone Survey (HFPS) in Nigeria, which, like our experimental sample, relied on the Nigeria GHS sample of households. As a result, effective sample differences between regular phone and in-person surveys would be greater, with greater coverage and non-response errors than we document (see also Ambel et al., 2021).

[20] In list-based phone surveys, survey implementers have access to a list of households or individuals along with contact phone numbers.

Endowment effects

Turning to measurement errors resulting from survey design and implementation choices, we can assess whether providing mobile phones to households affects reporting behavior. The concern is that this kind of intervention would lead to changes in the reporting or behavior. Households that received a phone might adjust their answers to please interviewers or in anticipation of receiving additional gifts. The endowment of a phone may itself cause changes in behavior. To assess this issue, we compare the outcome variables between households eligible for the experiment but not selected and selected households, leveraging the data collected in-person for both groups (Table A1). We find no significant differences in outcomes between selected households that received a phone and eligible but non-selected households, suggesting that receiving a phone has not influenced the responses provided.

Ordering and anchoring effects

Participating in two interviews within a short period of time might lead to behavioral changes among respondents. This may be due to learning or because respondents 'anchor' answers of the second interview in answers of the first interview (Schündeln, 2018). These are concerns for our identification, but they are also broader survey design concern whenever multiple (high frequency) longitudinal measures are taken. However, as we show in this paper, we find no evidence of such ordering effects affecting survey responses (Tables A6 and A7).

7 Discussion

We present strong experimental evidence of systematic differences between survey responses collected by phone and in person. Our findings extend the existing evidence on survey mode effects in LMICs in three ways. First, while prior studies have focused on specific outcomes, our study covers a wide set of development indicators. This broader scope allows us to identify systematic patterns in mode effects across different types of variables. Second, our experimental design isolates the causal impact of survey mode by accounting for coverage, respondent selection, and ordering biases, and enables direct comparisons between sources of error. We show that the magnitude of mode effects often exceeds that of other survey errors. Finally, by leveraging a within-respondent design, we uncover substantial respondent-level heterogeneity and examine the mechanisms underlying mode effects, offering deeper insight into how and for whom survey mode matters.

While our study offers the most rigorous evidence on mode effects to date, there are several important caveats and limitations to our study which qualify the interpretation and implications of our results for survey design and analysis. On the one hand, there are some survey design aspects which we were unable to fully control, relating to enumerators, survey length, and incentives. Different enumerators conducted in-person and phone interviews. While enumerator training followed a similar protocol and although we observe some enumerator characteristics, we cannot fully disentangle the extent to which enumerator effects may contribute to the measured mode differences. Likewise, there are some differences in survey questionnaires and administration. While both modes included an identical set of core questions, the in-person interviews were embedded within the regular post-harvest survey of the Nigeria GHS panel wave 5. As a result, they formed part of a substantially longer interview, typically lasting an average of 238 minutes. Most of the outcome variables analyzed here were placed early in the interview, but we cannot causally assess the impact of overall survey length or question positioning on responses, as this would require

experimental variation such as randomized question or module order. We cannot exclude that differential survey length may affect the in-person data and contribute to mode effects. Finally, each phone respondent received an incentive of 1,500 Nigerian naira (approximately US$1.75) in airtime upon interview completion, which they did not receive for in-person interviews. In the case of the in-person interviews, households received a food flask as souvenir valued at 6,150 Nigerian naira (or US$5.13). This was critical to achieving high response rates in phone surveys and minimizing non-response bias. However, it is conceivable that incentives induce respondents to behave and answer systematically differently across modes. All respondents received the same amount, preventing a causal investigation of the effect of incentives. These differences do reflect real-world conditions and constraints in survey implementation in LMICs. Phone interviews often have different enumerators; have to be shorter; and may require incentives. Our results suggest that there is a need to pay close attention to comparability of conditions.

On the other hand, we cannot conclusively determine which survey mode (phone or in-person) more accurately reflects the true underlying value (Y*). This is because we have no objective, external measurement of any of the outcomes against which to benchmark the survey data.[21] Likewise, the surveys did not include cognitive tests, such as reversed binary questions, to detect lack of attention or acquiescence bias affecting either mode. Cognitive or qualitative interviews could also serve to discover the underlying mechanisms giving rise to mode effects, but these were out of the scope of the present experiment. Instead, we offer evidence from indirect indicators of survey reliability, specifically, agreement patterns (phone survey respondents agree more often), the prevalence of extreme values (phone survey respondents are more likely to provide extreme values), the number of response options (the likelihood of a mode difference increases with the number of response options), and rounding (phone survey respondents are more likely to round). These tests suggest that respondents may be less cognitively engaged or otherwise behave differently when interviewed over the phone, which may affect the reliability of phone surveys. However, both the in-person and phone surveys likely suffer from some degree of measurement error, and we cannot ultimately quantify the extent. Finally, our study presents a data point from a single country, and more replication is needed to draw more robust inferences about mode effects in surveys in LMICs.

Implications for data analysis and inference

Systematic measurement error in survey data is a longstanding concern in empirical research, with the potential to bias inference, distort descriptive statistics, weaken monitoring tools, and undermine predictive models. A growing body of research has emphasized the need to diagnose and reduce errors in household surveys in LMICs (Abay et al., 2023; Weerdt et al., 2020a). The findings from this literature suggest that it is prudent for researchers and analysts to assume the presence of non-negligible systematic measurement error in most survey data – and indeed in most data sources in LMICs (Torchiana et al., 2025; Wollburg et al., 2024).

Our findings on mode effects reinforce this view and point to several implications for data analysis and inference, particularly for mixed-mode datasets and phone surveys. When analyzing data collected through the same mode, for example through phone surveys, inference can remain valid under certain conditions. If measurement error is uncorrelated with the outcome of interest or remains stable over time, trends can still be reliably assessed and associations plausibly estimated. This makes regular, repeated phone surveys a valuable tool for trend analysis, provided survey design and implementation are kept consistent. However, this assumption of stable, uncorrelated measurement error may not always hold in practice. Further, when measurement error is correlated across variables and affects both dependent and independent variables, this

[21] For a review on emerging best practices in benchmarking survey experiments against the truth, see Weerdt et al. (2020b).

can bias estimates and attenuate relationships within and across modes. The sign and severity of this bias depend on the structure and correlation of the underlying errors, which are typically unobserved and may differ across contexts.

Comparing data across survey modes or pooling observations from phone and in-person interviews comes with a different set of challenges. Systematic differences between modes can lead to substantial bias when used to estimate levels, trends, or treatment effects. While survey mode can be controlled for in regression models, this only adjusts for average differences between modes. It does not account for respondent-level misreporting or for differential measurement error across groups, outcomes, or contexts. Our results show that these respondent-level mode effects are widespread and highly variable. This challenges the assumption that a simple mode dummy suffices to adjust for bias and underscores the importance of testing for mode effects when integrating data across survey modes.

Addressing measurement error and mode effects in phone and mixed-mode surveys requires moving beyond simple regression adjustments toward more ambitious strategies that combine rigorous survey design with analytical correction techniques. Most statistical approaches to correcting measurement error, such as IV-based estimators or measurement-error-robust methods, rely on auxiliary data, repeated measures, or assumptions about the structure of the error (Abay et al., 2023). Longitudinal surveys or within-respondent designs, like the one used in this study, can provide the repeated observations needed to estimate and correct for such errors. More generally, large-scale or recurring survey programs can embed cross-mode validation subsamples, experimental modules, or explicit error-measuring items to improve identification. Statistical ex-post error correction techniques have already shown promise in addressing other survey errors, particularly coverage and non-response biases (Ambel et al., 2021b; Brubaker et al., 2021; Himelein, 2014). Similar approaches could be extended to mitigate bias from mode effects. While these strategies require additional resources and careful planning, they provide a credible path to improving the accuracy and policy relevance of data collected through phone and mixed-mode surveys.

Implications for survey design and data collection

Ex-ante, survey design-based strategies to limit mode effects are more promising. Our results also point to concrete steps for improving the design and implementation of phone – and in-person – surveys. In mixed-mode survey systems, where there is a special interest in comparable survey data, survey design and implementation should be standardized and harmonized as much as possible across modes. This includes aligning questionnaires, enumerator training, and field protocols. Incentives may be necessary to reduce non-response and attrition in higher-frequency phone survey data collection – but their effect on phone survey data is not yet well understood. In the absence of evidence-based guidance, randomizing incentive amounts would allow survey designers to assess how incentives shape responses in their surveys. Using the same enumerators across both modes, whenever possible, improves comparability and reduces the risk of interviewer effects distorting cross-mode comparisons and differencing out enumerator effects in analysis. A different option would be to rely on resident enumerators for higher frequency in person data collection, though this is unlikely to be feasible at scale.

When it comes to survey questionnaires, harmonization in question phrasing is a necessary condition for obtaining comparable estimates. But questionnaires also need to be appropriate for phone surveys, which are more limited with respect to which questions can be asked and understood (Gourlay et al., 2021b; Zezza et al., 2023). Asking very difficult questions over the phone will likely lead to measurement error, including mode effects, even if they are comparable across modes and it is arguably better to refrain from asking those kinds of questions over the phone. Our findings show that question harmonization alone is not sufficient. More careful testing and piloting is required, including through cognitive interviews specifically

for phone surveys. While this goes beyond what most (phone) survey operations currently do, the potential gains in data quality make it worthy of consideration.

Real-time survey management systems could be adapted to play a larger role in reducing mode effects and other sources of measurement error in phone surveys. Beyond their current use for monitoring field progress, these tools could be expanded to track indicators of respondent engagement and interviewer behavior that are closely tied to mode-related biases, such as unusually fast response times, excessive agreement in responses, or frequent rounding and heaping in numeric answers. Automated detection of such patterns would allow supervisors to intervene during data collection, for example by providing immediate feedback to interviewers or re-contacting respondents flagged for low-quality responses. This can be paired with built-in low-cost quality checks, such as reversed items, cognitive probes, attention filters, or repeated questions. These techniques provide diagnostics on measurement error and respondent engagement, identify problematic questions, respondent or enumerator behavior. Systematically incorporating experimentation into survey operations can likewise help to uncover and mitigate survey design effects. For example, randomizing the order of survey modules or questions, where compatible with survey objectives, can reduce ordering effects and improve the robustness of responses across both phone and in-person modes.

Finally, conducting phone surveys together with in-person surveys, for example as part of a mixed-mode system, supports benchmarking and anchoring phone survey responses, providing a reference point against which phone-based estimates can be calibrated. While mixed-mode systems introduce additional complexity in managing and accounting for mode differences, they ultimately provide a pathway to higher-quality data by reducing measurement and representation errors.

Addressing mode effects remains an active area of research, as building a cumulative evidence base will require further experimentation, validation across contexts and outcomes, and replication in diverse survey systems. Such work is essential for developing evidence-based standards for survey practice and for ensuring the production of high-quality, comparable data that underpin credible empirical research.

8 Conclusion

Phone surveys offer several advantages over traditional in-person surveys: they are less costly, faster to deploy, and more flexible, enabling data collection during emergencies or in hard-to-reach areas. These benefits are particularly valuable in the context of growing demand for timely, high-frequency data and an increasingly constrained budget for research and international development. However, it is critical to ensure that innovations and adoption of new survey methods do not come at the expense of data quality. While a growing body of evidence has explored the quality of phone survey data, important gaps remain in assessing and understanding the effect of survey mode on reported outcomes.

This study provides evidence on survey mode effects based on a large-scale, multi-mode experiment in rural Nigeria. We find strong evidence of significant mode effects across a wide range of outcomes, with phone responses differing from in-person responses by 17%–18% at the median. We further document that these effects are substantial in comparison to other common sources of error in surveys, like coverage bias. Leveraging a within-respondent design, we show that respondents more often answer questions inconsistently across modes than the average mode effect estimates suggest. As regards the mechanisms, we find that respondents with higher education levels are less prone to mode effects, while the age and experience of enumerators also play a role. Mode effects are particularly pronounced for variables with more response options. Finally, we identify systematic behavioral differences between modes, with phone

respondents exhibiting a higher tendency to agree with statements, answer “yes” to binary questions, and provide rounded or extreme values to numeric questions.

We conclude by highlighting several promising areas for future research. First, since the literature on survey mode effects in low- and middle-income countries is rather nascent, conducting similar survey experiments in alternative contexts is needed to test the external validity of our findings. Second, more work is needed to unpack the mechanisms driving mode effects, including the role of incentives and enumerators, which can be pursued in future survey experiments. Qualitative research can also help shed light on how respondents behave and report information across modes. Third, future research can document the accuracy of the data collected through different modes, should there be opportunities to obtain a reliable measure of truth – which can be pursued as part of a survey experiment or obtained from a different data source, depending on the outcome of interest. Fourth, future research can also investigate how these mode-related biases could be mitigated through statistical adjustment methods, such as response propensity models and calibration techniques, and through careful survey design. Fifth, more attention should be paid to understanding whether and how mode effects change over time, given the interest in implementing phone surveys with a longitudinal design. Finally, while real-time survey management systems are increasingly used in fieldwork, their potential to detect and reduce survey errors like mode effects remains understudied.

References

Abate, G.T., de Brauw, A., Hirvonen, K., Wolle, A., 2023. Measuring consumption over the phone: Evidence from a survey experiment in urban Ethiopia. J. Dev. Econ. 161, 103026. https://doi.org/10.1016/j.jdeveco.2022.103026

Abay, K.A., Berhane, G., Hoddinott, J., Hirfrfot, K.T., 2021. Assessing Response Fatigue in Phone Surveys : Experimental Evidence on Dietary Diversity in Ethiopia. Policy Res. Work. Pap. Ser., Policy Research Working Paper Series.

Abay, K.A., Wossen, T., Abate, G.T., Stevenson, J.R., Michelson, H., Barrett, C.B., 2023. Inferential and Behavioral Implications of Measurement Error in Agricultural Data. Annu. Rev. Resour. Econ. 15, 63–83. https://doi.org/10.1146/annurev-resource-101422-090049

Ambel, A., McGee, K., Tsegay, A., 2021a. Reducing Bias in Phone Survey Samples. Effectiveness of Reweighting Techniques Using Face-to-Face Surveys as Frames in Four African Countries (No. 9676), Policy Research Working Paper. World Bank, Washington D.C.

Ambel, A., McGee, K., Tsegay, A., 2021b. Reducing Bias in Phone Survey Samples. Effectiveness of Reweighting Techniques Using Face-to-Face Surveys as Frames in Four African Countries (No. 9676), Policy Research Working Paper. World Bank, Washington D.C.

Anderson, E., Lybbert, T.J., Shenoy, A., Singh, R., Stein, D., 2024. Does survey mode matter? Comparing in-person and phone agricultural surveys in India. J. Dev. Econ. 166, 103199. https://doi.org/10.1016/j.jdeveco.2023.103199

Biemer, P.P., 2010. Total Survey Error. Design, Implmentation, and Evaluation. Public Opin. Q. 74, 817–848.

Brubaker, J., Kilic, T., Wollburg, P., 2021. Representativeness of individual-level data in COVID-19 phone surveys: Findings from Sub-Saharan Africa. PLOS ONE 16, e0258877. https://doi.org/10.1371/journal.pone.0258877

Caceres-Santamaria, A.J., 2021. The Anchoring Effect [WWW Document]. URL https://research.stlouisfed.org/publications/page1-econ/2021/04/01/the-anchoring-effect (accessed 6.6.24).

Celhay, P., Meyer, B.D., Mittag, N., 2024. What leads to measurement errors? Evidence from reports of program participation in three surveys. J. Econom. 238, 105581. https://doi.org/10.1016/j.jeconom.2023.105581

Charness, G., Gneezy, U., Kuhn, M.A., 2012. Experimental methods: Between-subject and within-subject design. J. Econ. Behav. Organ. 81, 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Garlick, R., Orkin, K., Quinn, S., 2020. Call Me Maybe: Experimental Evidence on Frequency and Medium Effects in Microenterprise Surveys. World Bank Econ. Rev. 34, 418–443. https://doi.org/10.1093/wber/lhz021

Gibson, D.G., Wosu, A.C., Pariyo, G.W., Ahmed, S., Ali, J., Labrique, A.B., Khan, I.A., Rutebemberwa, E., Flora, M.S., 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 Glob. Health 4. https://doi.org/10.1136/bmjgh-2019-001604

Glazerman, S., Grépin, K.A., Mueller, V., Rosenbaum, M., Wu, N., 2023. Do referrals improve the representation of women in mobile phone surveys? J. Dev. Econ. 162, 103077. https://doi.org/10.1016/j.jdeveco.2023.103077

Gourlay, S., Kilic, T., Martuscelli, A., Wollburg, P., Zezza, A., 2021a. Viewpoint: High-frequency phone surveys on COVID-19: Good practices, open questions. Food Policy 105, 102153. https://doi.org/10.1016/j.foodpol.2021.102153

Gourlay, S., Kilic, T., Martuscelli, A., Wollburg, P., Zezza, A., 2021b. Viewpoint: High-frequency phone surveys on COVID-19: Good practices, open questions. Food Policy 105, 102153. https://doi.org/10.1016/j.foodpol.2021.102153

Greenleaf, A.R., Gadiaga, A., Guiella, G., Turke, S., Battle, N., Ahmed, S., 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, e0231819. https://doi.org/10.1371/journal.pone.0231819
GSMA, 2024. The Mobile Economy Sub-Saharan Africa 2024. GMSA.
Headey, D., Barrett, C.B., 2015. Opinion: Measuring development resilience in the world's poorest countries. Proc. Natl. Acad. Sci. 112, 11423–11425.
Himelein, K., 2014. Weight Calculations for Panel Surveys with Subsampling and Split-off Tracking. Stat. Public Policy 1, 40–45. https://doi.org/10.1080/2330443X.2013.856170
Kilic, T., Moylan, H., Ilukor, J., Mtengula, C., Pangapanga-Phiri, I., 2021. Root for the tubers: Extended-harvest crop production and productivity measurement in surveys. Food Policy 102, 102033. https://doi.org/10.1016/j.foodpol.2021.102033
Lamanna, C., Hachhethu, K., Chesterman, S., Singhal, G., Mwongela, B., Ng'endo, M., Passeri, S., Farhikhtah, A., Kadiyala, S., Bauer, J.-M., Rosenstock, T.S., 2019. Strengths and limitations of computer assisted telephone interviews (CATI) for nutrition data collection in rural Kenya. PLOS ONE 14, e0210050. https://doi.org/10.1371/journal.pone.0210050
L'Engle, K., Sefa, E., Adimazoya, E.A., Yartey, E., Lenzi, R., Tarpo, C., Heward-Mills, N.L., Lew, K., Ampeh, Y., 2018. Survey research with a random digit dial national mobile phone sample in Ghana: Methods and sample quality. PLOS ONE 13, e0190902. https://doi.org/10.1371/journal.pone.0190902
List, J.A., 2025. The Experimentalist Looks Within: Toward an Understanding of Within-Subject Experimental Designs. Working Paper Series. https://doi.org/10.3386/w33456
Markhof, Y., Dietrich, S., Vincent, R.C., 2026. Lies in Disguise: Respondent Behavior and Survey Measurement Uncertainty. Working Paper.
Markhof, Y., Wollburg, P., Zezza, A., 2025. Beyond the records: Data quality and COVID-19 vaccination progress in low- and middle-income countries. J. Dev. Econ. 174, 103449. https://doi.org/10.1016/j.jdeveco.2024.103449
Meyer, B.D., Mok, W.K.C., Sullivan, J.X., 2015. Household Surveys in Crisis. J. Econ. Perspect. 29, 199–226. https://doi.org/10.1257/jep.29.4.199
Özler, B., Cuevas, P.F., 2019. Reducing attrition in phone surveys. World Bank Blogs. URL https://blogs.worldbank.org/impactevaluations/reducing-attrition-phone-surveys (accessed 11.6.20).
Schündeln, M., 2018. Multiple Visits and Data Quality in Household Surveys. Oxf. Bull. Econ. Stat. 80, 380–405. https://doi.org/10.1111/obes.12196
Struminskaya, B., Bosnjak, M., 2021. Panel Conditioning: Types, Causes, and Empirical Evidence of What We Know So Far, in: Lynn, P. (Ed.), Wiley Series in Probability and Statistics. Wiley, pp. 272–301. https://doi.org/10.1002/9781119376965.ch12
Torchiana, A.L., Rosenbaum, T., Scott, P.T., Souza-Rodrigues, E., 2025. Improving Estimates of Transitions from Satellite Data: A Hidden Markov Model Approach. Rev. Econ. Stat. 107, 426–441.
Weerdt, J.D., Gibson, J., Beegle, K., 2020a. What Can We Learn from Experimenting with Survey Methods? Annu. Rev. Resour. Econ. 12, 431–447. https://doi.org/10.1146/annurev-resource-103019-105958
Weerdt, J.D., Gibson, J., Beegle, K., 2020b. What Can We Learn from Experimenting with Survey Methods? Annu. Rev. Resour. Econ. 12, 431–447. https://doi.org/10.1146/annurev-resource-103019-105958
Wollburg, P., Markhof, Y., Bentze, T., Ponzini, G., 2024. Substantial impacts of climate shocks in African smallholder agriculture. Nat. Sustain. 1–10. https://doi.org/10.1038/s41893-024-01411-w
Zezza, A., McGee, K., Wollburg, P., Assefa, T., Gourlay, S., 2023. From necessity to opportunity: lessons for integrating phone and in-person data collection. Eur. Rev. Agric. Econ. 50, 1364–1400. https://doi.org/10.1093/erae/jbad017

Main tables

Table 1. Survey mode effects across outcomes

	A. Between-design estimates				B. Within-design estimates			
	Phone Interview	Constant	N	Percentage change	Phone Interview	Constant	N	Percentage change
Illness or injury in last 30 days	0.107**	0.398***	852	26.78	0.0743**	0.415***	1668	17.92
	(0.0339)	(0.0236)			(0.0243)	(0.0171)		
Consulted health practitioner in last 30 days	0.233***	0.419***	852	55.71	0.189***	0.451***	1668	42.02
	(0.0333)	(0.0238)			(0.0240)	(0.0172)		
Any health expenditure in last 12 months	0.464***	0.234***	721	198.8	0.483***	0.236***	1430	204.1
	(0.0331)	(0.0243)			(0.0231)	(0.0159)		
Total health expenditures (in NGN10'000)	4.119***	2.079***	359	198.1	4.300***	1.688***	256	254.8
	(0.707)	(0.411)			(0.793)	(0.226)		
Respondent vaccinated for COVID-19	0.202***	0.322***	852	62.80	0.206***	0.331***	1668	62.32
	(0.0332)	(0.0225)			(0.0238)	(0.0163)		
Subjective wellbeing ladder step: self	-0.797***	4.475***	848	-17.80	-0.774***	4.489***	1654	-17.24
	(0.142)	(0.0843)			(0.0997)	(0.0610)		
Subjective wellbeing ladder step: neighbors	-0.674***	4.396***	835	-15.33	-0.705***	4.440***	1618	-15.87
	(0.142)	(0.0793)			(0.0971)	(0.0584)		
Subjective wellbeing ladder step: friends	-0.510***	4.648***	837	-10.97	-0.563***	4.656***	1620	-12.09
	(0.143)	(0.0777)			(0.101)	(0.0575)		
Subjective wellbeing ladder step: self in 5 yrs	0.863***	6.833***	843	12.62	0.590***	6.941***	1638	8.496
	(0.157)	(0.1000)			(0.112)	(0.0714)		
Income adequacy	-0.197**	2.248***	850	-8.784	-0.222***	2.267***	1660	-9.777
	(0.0606)	(0.0419)			(0.0434)	(0.0308)		
Life controlled by accidental happenings	0.561***	1.688***	850	33.26	0.575***	1.676***	1660	34.29
	(0.0577)	(0.0358)			(0.0416)	(0.0261)		
Life determined by own actions	0.163***	2.514***	850	6.490	0.181***	2.522***	1660	7.167
	(0.0480)	(0.0334)			(0.0341)	(0.0243)		
Life determined by others in HH	0.680***	1.581***	850	42.99	0.665***	1.569***	1660	42.40
	(0.0577)	(0.0357)			(0.0417)	(0.0259)		
Household Dietary Diversity Score	0.326**	7.828***	678	4.159	0.118	7.956***	1356	1.483
	(0.122)	(0.0907)			(0.0895)	(0.0578)		
Food Consumption Score Raw	12.07***	50.31***	678	24.00	9.602***	51.46***	1356	18.66
	(1.373)	(1.023)			(0.959)	(0.623)		
Any work in the past 7 days	0.0422	0.790***	836	5.337	0.0946***	0.735***	2474	12.87
	(0.0271)	(0.0199)			(0.0165)	(0.0126)		
Hours worked/helped in household agriculture in last 7 days	-3.442**	23.29***	481	-14.78	-2.170*	22.31***	912	-9.726
	(1.234)	(0.894)			(0.901)	(0.643)		
Any HE in the last 12 months	0.328***	0.533***	732	61.42	0.279***	0.559***	1442	49.88
	(0.0329)	(0.0281)			(0.0230)	(0.0185)		
Likelihood of negative effect on HH	-0.0327	3.165***	822	-1.032	0.0170	3.192***	1530	0.532
	(0.0758)	(0.0533)			(0.0555)	(0.0399)		
Any shock since 2022	0.151***	0.827***	753	18.25	0.139***	0.835***	1494	16.67
	(0.0218)	(0.0206)			(0.0148)	(0.0136)		

Note: OLS regressions of various outcomes on a dummy for the survey mode of the interview. Sample (A): First-time respondents. Sample (B): Identical respondents interviewed via phone or face-to-face in random order. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 2. Heterogeneity in mode effects by respondent characteristics

	(1) Any mode effect (dummy)
Respondent female	0.0039
	(0.0206)
Respondent age	0.0002
	(0.0004)
Primary	-0.0396***
	(0.0136)
Secondary	-0.0547***
	(0.0157)
Tertiary	-0.0671***
	(0.0179)
Resp. speaks interview lang. at home	-0.0093
	(0.0121)
Resp. prev phone interview	-0.0207
	(0.0235)
Household size	0.0005
	(0.0018)
Household log total consumption	-0.0128
	(0.0101)
Household part of HFPS	-0.0078
	(0.0104)
Outcome FE	YES
Enumerator controls	YES
Observations	12,455

Note: OLS regressions at the respondent-by-outcome level. The dependent variable is a dummy for whether phone and in-person answer by the same respondent were different. The sample pools across all binary or categorical outcomes. Enumerator controls include, for the in-person and phone enumerator, respectively: gender, age, and experience (five or more surveys completed) of the enumerator; a dummy for whether the enumerator speaks the interview language at home; an enumerator social desirability index (based on the BIDR-16, Hart et al. (2015)); and a dummy for whether the enumerator and respondent have the same gender. Clustered standard errors at the respondent level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 3. Response patterns

	(1) Yes to binary	(2) Agree with LOC	(3) Extreme value to subj. question	(4) Low extreme value to subj. question	(5) High extreme value to subj. question	(6) Rounded value to continuous outcome
Phone interview	0.196***	0.347***	0.192***	0.107***	0.085***	0.118***
	(0.009)	(0.014)	(0.011)	(0.009)	(0.006)	(0.025)
Constant	0.524***	0.323***	0.057***	0.021***	0.036***	0.200***
	(0.008)	(0.009)	(0.005)	(0.004)	(0.004)	(0.016)
Observations	11,844	4,980	6,530	6,530	6,530	1,168

Note: All columns report results from OLS regressions estimated at the respondent-by-outcome level. Column 1 uses a dummy for answering yes to a binary question. Column 2 uses a dummy for agreeing with a locus of control statement. Columns 3 to 5 use dummies for providing an extreme response (1 or 10) to a wellbeing question, an extreme low value, and an extreme high value, respectively. Column 6 uses a dummy for rounding responses—to the nearest 10 (hours worked) or 10,000 (health expenditures). Clustered standard errors at the respondent level are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix A. Supplementary Exhibits

Table A1. Balance test for households eligible to the experiment

	(1) In the experimental group		(2) Eligible but not in the experimental group		(1)-(2) Pairwise t-test	
Variable	N/Clusters	Mean/(SE)	N/Clusters	Mean/(SE)	N/Clusters	Mean difference
Baseline characteristics						
Household size	937	5.905	2309	6.202	3246	-0.297
	105	(0.238)	258	(0.140)	349	
Dependency ratio	937	0.929	2309	1.022	3246	-0.093**
	105	(0.032)	258	(0.023)	349	
Female household head	937	0.228	2309	0.214	3246	0.014
	105	(0.020)	258	(0.012)	349	
Age of household head	937	53.471	2309	53.326	3246	0.145
	105	(0.676)	258	(0.430)	349	
Owns non cultivated land	937	0.841	2309	0.819	3246	0.022
	105	(0.022)	258	(0.014)	349	
Agricultural household	937	0.914	2309	0.890	3246	0.024
	105	(0.016)	258	(0.011)	349	
Outcomes of interest						
Household Dietary Diversity Score	933	7.953	2279	7.805	3212	0.148
	105	(0.084)	248	(0.051)	349	
Food Consumption Score Raw	933	52.026	2279	51.542	3212	0.484
	105	(0.879)	248	(0.601)	349	
Any HE in the last 12 months	933	0.542	2279	0.513	3212	0.030
	105	(0.025)	248	(0.017)	349	
Any shock since 2022	933	0.830	2279	0.821	3212	0.009
	105	(0.019)	248	(0.013)	349	
Any health expenditure in last 12 months	933	0.234	2279	0.214	3212	0.020
	105	(0.024)	248	(0.015)	349	
Total health expenditures (in NGN10'000)	218	1.672	487	1.941	705	-0.268
	66	(0.157)	155	(0.139)	219	

Significance: ***=.01, **=.05, *=.1. Errors are clustered at the enumeration area level.

Table A2. Number of households in final analytical samples

	Total	Treatment Group T1	Treatment Group T2
Experimental sample	937	464	473
Within-respondent design			
At least one adult interviewed in person	933		
At least one adult interviewed over the phone	864		
Final sample (at least one adult interviewed in both modes)	862	423	439
Between-respondent design			
Main respondent interviewed in person	906		
Main respondent interviewed over the phone	863		
Final sample (main respondent interviewed in both modes)	852	420	432

Table A3. Descriptive statistics and balance of characteristics for the between-respondent sample

	F2F interview first		Phone interview first		Pairwise t-test	
	N	Mean/(SE)	N	Mean/(SE)	N	Mean difference
Respondent variables						
Female	432	0.289	420	0.290	852	-0.001
	432	(0.022)	420	(0.022)	852	
Age	432	49.514	420	49.026	852	0.488
	432	(0.717)	420	(0.739)	852	
Currently married	429	0.704	407	0.703	836	0.001
	429	(0.022)	407	(0.023)	836	
Completed secondary school	432	0.481	410	0.429	842	0.052
	432	(0.024)	410	(0.024)	842	
Respondent is household head	432	0.824	420	0.812	852	0.012
	432	(0.018)	420	(0.019)	852	
Individual owns a phone at the Pre-Planting Visit	427	0.724	414	0.742	841	-0.018
	427	(0.022)	414	(0.022)	841	
Household demographics						
Household size	432	5.731	420	5.576	852	0.155
	432	(0.165)	420	(0.172)	852	
Dependency ratio	432	1.001	420	0.900	852	0.100*
	432	(0.041)	420	(0.034)	852	
Female household head	432	0.234	420	0.221	852	0.012
	432	(0.020)	420	(0.020)	852	
Age of household head	432	53.204	420	52.874	852	0.330
	432	(0.701)	420	(0.702)	852	
Household living conditions						
Owns non cultivated land	432	0.840	420	0.848	852	-0.007
	432	(0.018)	420	(0.018)	852	
Agricultural household	432	0.905	420	0.921	852	-0.016
	432	(0.014)	420	(0.013)	852	
Total annual consumption expenditures pc	432	4.76e+05	419	4.84e+05	851	-7103.838
	432	(18004.272)	419	(16539.063)	851	
Has mud wall	432	0.375	420	0.360	852	0.015
	432	(0.023)	420	(0.023)	852	
Has mud/straw/sand/dirt floor	432	0.245	420	0.243	852	0.003
	432	(0.021)	420	(0.021)	852	
Omnibus F-test p-value					0.700	

Significance: ***=.01, **=.05, *=.1. Errors are clustered at the household level.

Table A4. Survey mode and interview dates

	Date of interview
Phone interview	0.0094
	(0.2489)
Constant	23424.1186***
	(0.3362)
Observations	2546

Note: OLS regressions of the interview date on a dummy for the survey mode of the interview. Sample: Phone and in-person interviews from identical respondents. Clustered standard errors at individual level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A5. Mode effects by locus of control response options

	(1) Pooled	(2) Accidental happenings	(3) Determined by others	(4) Determined by self
Disagree	-0.096***	-0.121***	-0.195***	0.020
	(0.013)	(0.021)	(0.021)	(0.015)
Neither agree nor disagree	-0.225***	-0.246***	-0.183***	-0.239***
	(0.013)	(0.019)	(0.018)	(0.022)
Agree	0.321***	0.367***	0.378***	0.219***
	(0.011)	(0.013)	(0.013)	(0.023)
Observations	4,980	1,660	1,660	1,660

Note: Change in the probability to disagree, neither agree nor disagree, and agree to locus of control questions when asking over the phone instead of in person. Coefficients are average marginal effects from a multinomial logit regression of all answers to locus of control questions on a dummy for phone survey mode. Clustered standard errors at the respondent level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A6. Test of ordering effects across outcomes

	(1) Categorical + Binary	(2) Categorical	(3) Binary	(4) Continuous	(5) Categorical + Binary	(6) Continuous
Second interview	0.0203	0.0154	0.0244	0.0031	0.0354	-0.0092
	(0.0200)	(0.0108)	(0.0351)	(0.0177)	(0.0284)	(0.0234)
Second interview * Phone interview					-0.0339	0.0176
					(0.0494)	(0.0350)
Phone interview					0.0906***	0.0957***
					(0.0316)	(0.0253)
Outcome FE	YES	YES	YES	YES	YES	YES
Observations	26,544	11,844	14,700	3,880	26,544	3,880

Note: OLS regressions at the respondent-by-outcome level on a dummy for whether the interview was the respondent's second, a phone interview dummy, and their interaction. Continuous outcomes are logged such that coefficients across outcomes are interpretable on the same scale as a 100 * beta percent changes in the outcome. Standard errors are clustered at the respondent level. *** p<0.01, ** p<0.05, * p<0.1

Table A7. Ordering effects by outcome

	Phone interview* Second interview	Phone interview	Second interview	constant	N
Illness or injury in last 30 days	-0.0774	0.113***	0.0323	0.399***	1668
	(0.0538)	(0.0343)	(0.0342)	(0.0238)	
Consulted health practitioner in last 30 days	-0.105*	0.242***	0.0627	0.420***	1668
	(0.0527)	(0.0335)	(0.0345)	(0.0240)	
Any health expenditure in last 12 months	-0.0386	0.485***	0.0357	0.231***	1146
	(0.0543)	(0.0366)	(0.0362)	(0.0244)	
Total health expenditures (in NGN10'000)	-1.405	4.672***	-0.294	1.866***	212
	(1.743)	(1.128)	(0.560)	(0.507)	
Respondent vaccinated for COVID-19	0.0105	0.200***	0.0191	0.322***	1668
	(0.0558)	(0.0336)	(0.0326)	(0.0227)	
Subjective wellbeing ladder step: self	0.0874	-0.819***	0.0321	4.473***	1654
	(0.224)	(0.143)	(0.122)	(0.0853)	
Subjective wellbeing ladder step: neighbors	0.0229	-0.719***	0.0690	4.407***	1618
	(0.207)	(0.142)	(0.117)	(0.0810)	
Subjective wellbeing ladder step: friends	-0.0444	-0.540***	0.0117	4.650***	1620
	(0.214)	(0.145)	(0.115)	(0.0793)	
Subjective wellbeing ladder step: self in 5 yrs	-0.530*	0.855***	0.261	6.814***	1638
	(0.247)	(0.159)	(0.142)	(0.102)	
Income adequacy	-0.0473	-0.198**	0.0454	2.245***	1660
	(0.0927)	(0.0612)	(0.0618)	(0.0423)	
Life controlled by accidental happenings	0.0152	0.567***	-0.00683	1.679***	1660
	(0.0883)	(0.0585)	(0.0523)	(0.0362)	
Life determined by own actions	0.0337	0.163***	0.0154	2.514***	1660
	(0.0690)	(0.0488)	(0.0486)	(0.0339)	
Life determined by others in HH	-0.00589	0.669***	-0.0332	1.585***	1660
	(0.0875)	(0.0586)	(0.0519)	(0.0362)	
Household Dietary Diversity Score	-0.0675	0.172	0.0939	7.837***	974
	(0.231)	(0.148)	(0.134)	(0.0911)	
Food Consumption Score Raw	-0.457	10.54***	1.203	50.32***	974
	(2.404)	(1.643)	(1.507)	(1.032)	
Any work in the past 7 days	0.0168	0.0450	0.00142	0.788***	1602
	(0.0432)	(0.0277)	(0.0289)	(0.0202)	
Hours worked/helped in household agriculture in last 7 days	2.561	-3.726*	-0.733	23.64***	670
	(2.379)	(1.487)	(1.541)	(1.061)	
Any HE in the last 12 months	-0.0468	0.316***	0.0181	0.535***	1198
	(0.0559)	(0.0353)	(0.0408)	(0.0284)	
Likelihood of negative effect on HH	0.149	-0.0591	0.00492	3.190***	1530
	(0.112)	(0.0789)	(0.0800)	(0.0556)	
Any shock since 2022	-0.0253	0.150***	0.0232	0.831***	1292
	(0.0308)	(0.0220)	(0.0287)	(0.0206)	

Note: OLS regressions of various standardized outcomes on a dummy for the survey mode of the interview fully interacted with a dummy for whether the respondent had been previously interviewed (either over the phone or in person). Sample: Identical respondents interviewed via phone or face-to-face in random order. Robust standard errors in parentheses. *** p<0.01 ** p<0.05 * p<0.1.

Table A8. Survey mode effects after controlling for the date of the interview

	A. Between-respondent sample				B. Within-respondent sample			
	Phone interview	Interview date	constant	N	Phone interview	Interview date	constant	N
Illness or injury in last 30 days	0.121***	0.00317	-73.96	852	0.0743**	0.000973	-22.38	1668
	(0.0347)	(0.00180)	(42.05)		(0.0243)	(0.00106)	(24.83)	
Consulted health practitioner in last 30 days	0.252***	0.00410*	-95.54*	852	0.189***	0.00203	-46.99	1668
	(0.0338)	(0.00182)	(42.62)		(0.0239)	(0.00106)	(24.75)	
Any health expenditure in last 12 months	0.475***	0.00241	-56.19	721	0.482***	0.00220*	-51.33*	1430
	(0.0336)	(0.00172)	(40.19)		(0.0231)	(0.00105)	(24.53)	
Total health expenditures (in NGN10'000)	4.385***	0.0651	-1523.1	359	4.312***	0.0266	-620.9	256
	(0.826)	(0.0547)	(1280.6)		(0.798)	(0.0297)	(695.8)	
Respondent vaccinated for COVID-19	0.208***	0.00139	-32.27	852	0.206***	-0.0000263	0.946	1668
	(0.0340)	(0.00167)	(39.06)		(0.0238)	(0.000995)	(23.31)	
Subjective wellbeing ladder step: self	-0.853***	-0.0125	296.7	848	-0.774***	-0.00907*	216.9*	1654
	(0.143)	(0.00664)	(155.5)		(0.0996)	(0.00390)	(91.38)	
Subjective wellbeing ladder step: neighbors	-0.733***	-0.0126	299.3	835	-0.704***	-0.00579	140.1	1618
	(0.142)	(0.00657)	(154.0)		(0.0971)	(0.00386)	(90.49)	
Subjective wellbeing ladder step: friends	-0.554***	-0.00956	228.6	837	-0.562***	-0.00642	155.0	1620
	(0.142)	(0.00633)	(148.2)		(0.101)	(0.00427)	(99.97)	
Subjective wellbeing ladder step: self in 5 yrs	0.811***	-0.0111	267.4	843	0.590***	-0.00866	209.7	1638
	(0.159)	(0.00745)	(174.5)		(0.112)	(0.00470)	(110.2)	
Income adequacy	-0.228***	-0.00686*	163.0*	850	-0.222***	-0.00274	66.45	1660
	(0.0614)	(0.00299)	(70.12)		(0.0434)	(0.00179)	(42.03)	
Life controlled by accidental happenings	0.547***	-0.00330	78.92	850	0.575***	-0.00342*	81.87*	1660
	(0.0592)	(0.00277)	(64.94)		(0.0415)	(0.00160)	(37.47)	
Life determined by own actions	0.173***	0.00212	-47.10	850	0.181***	0.00166	-36.44	1660
	(0.0487)	(0.00240)	(56.28)		(0.0341)	(0.00134)	(31.46)	
Life determined by others in HH	0.676***	-0.000826	20.92	850	0.665***	-0.000457	12.27	1660
	(0.0593)	(0.00271)	(63.41)		(0.0417)	(0.00183)	(42.87)	
Household Dietary Diversity Score	0.265*	-0.0124*	298.9*	678	0.118	-0.00690	169.5	1356
	(0.126)	(0.00598)	(140.0)		(0.0894)	(0.00401)	(93.86)	
Food Consumption Score Raw	12.17***	0.0188	-389.3	678	9.602***	-0.0218	561.3	1356
	(1.447)	(0.0684)	(1602.5)		(0.959)	(0.0399)	(935.2)	
Any work in the past 7 days	0.0518	0.00215	-49.50	836	0.0945***	0.00121	-27.65	2474
	(0.0282)	(0.00122)	(28.58)		(0.0165)	(0.000767)	(17.97)	
Hours worked/helped in household agriculture in last 7 days	-2.625*	0.143*	-3316.2*	481	-2.213*	0.0767	-1774.1	912
	(1.305)	(0.0678)	(1587.6)		(0.898)	(0.0424)	(993.1)	
Any HE in the last 12 months	0.312***	-0.00367*	86.39*	732	0.279***	-0.00169	40.15	1442
	(0.0337)	(0.00157)	(36.75)		(0.0230)	(0.00110)	(25.70)	
Likelihood of negative effect on HH	-0.0231	0.00217	-47.78	822	0.0169	0.00358	-80.66	1530
	(0.0775)	(0.00459)	(107.4)		(0.0555)	(0.00274)	(64.25)	
Any shock since 2022	0.155***	0.000908	-20.44	753	0.139***	0.00127	-28.87	1494
	(0.0226)	(0.00138)	(32.32)		(0.0147)	(0.000705)	(16.52)	

Note: OLS regressions of various outcomes on a dummy for the survey mode of the interview and the date of the interview. Sample (A) includes first-time respondents, with robust standard errors in parentheses. Sample (B) includes identical respondents interviewed both by phone and in person, with standard errors clustered at the respondent level. *** p<0.01 ** p<0.05 * p<0.1.

Table A9. Share of inconsistent answers for binary and categorical variables

	Share in %	Observations
Illness or injury in last 30 days	38.85***	834
	(1.689)	
Consulted health practitioner in last 30 days	41.25***	834
	(1.706)	
Respondent vaccinated for COVID-19	33.57***	834
	(1.636)	
Any health expenditure in last 12 months	58.88***	715
	(1.841)	
Any work in the past 7 days	26.76***	1237
	(1.259)	
Any HE in the last 12 months	37.86***	721
	(1.808)	
Any shock since 2022	18.21***	747
	(1.413)	
Subjective wellbeing ladder step: self	78.48***	827
	(1.430)	
Subjective wellbeing ladder step: neighbors	84.05***	809
	(1.288)	
Subjective wellbeing ladder step: friends	83.83***	810
	(1.295)	
Subjective wellbeing ladder step: self in 5 yrs	82.66***	819
	(1.324)	
Income adequacy	65.30***	830
	(1.653)	
Life controlled by accidental happenings	66.14***	830
	(1.644)	
Life determined by own actions	43.49***	830
	(1.722)	
Life determined by others in HH	64.94***	830
	(1.657)	
Likelihood of negative effect on HH	71.37***	765
	(1.635)	
Pooled across all variables	55.24***	13272
	(0.488)	

Note: Share of variables with mode effects for all binary and categorical variables separately, and for all binary and categorical variables pooled together. Clustered standard errors at the respondent level are in parentheses. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$.

Table A10. Coverage effects, respondent and household characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Female	Age	Completed secondary school	Respondent is household head	Household size	Female household head	Log annual household consumption
Individual owns a phone	-0.143***	4.349***	0.199***	0.189***			
	(0.0256)	(0.895)	(0.0254)	(0.0254)			
Household owns a phone					0.659**	-0.070*	0.104**
					(0.2994)	(0.0370)	(0.0499)
Constant	0.584***	41.672***	0.333***	0.400***	5.130***	0.288***	12.815***
	(0.0199)	(0.7368)	(0.0195)	(0.0198)	(0.2689)	(0.0334)	(0.0448)
Observations	1,572	1,572	1,523	1,572	862	862	861

Note: OLS regressions of various respondent and household characteristics on phone ownership at baseline. For respondent characteristics, phone ownership is identified at the individual level. For household level characteristics, it is identified at the household level. Sample: All phone survey respondents (individual characteristics) or all households with a phone and in-person interview (household characteristics). Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A11. Coverage effects, main outcomes, household-level phone ownership

	Household phone ownership	constant	N	Had phone vs had no phone, % change
Illness or injury in last 30 days	0.00402	0.449***	1668	0.896
	(0.0337)	(0.0301)		
Consulted health practitioner in last 30 days	0.0181	0.531***	1668	3.415
	(0.0326)	(0.0290)		
Any health expenditure in last 12 months	0.00479	0.474***	1430	1.011
	(0.0287)	(0.0253)		
Total health expenditures (in NGN10'000)	1.714**	2.458***	256	69.72
	(0.571)	(0.256)		
Respondent vaccinated for COVID-19	0.111***	0.347***	1668	31.98
	(0.0318)	(0.0275)		
Subjective wellbeing ladder step: self	0.376**	3.805***	1654	9.886
	(0.132)	(0.116)		
Subjective wellbeing ladder step: neighbors	0.130	3.985***	1618	3.255
	(0.126)	(0.111)		
Subjective wellbeing ladder step: friends	0.294*	4.142***	1620	7.099
	(0.130)	(0.116)		
Subjective wellbeing ladder step: self in 5 yrs	0.472**	6.863***	1638	6.877
	(0.156)	(0.140)		
Income adequacy	0.122*	2.060***	1660	5.913
	(0.0526)	(0.0453)		
Life controlled by accidental happenings	-0.0683	2.017***	1660	-3.386
	(0.0534)	(0.0472)		
Life determined by own actions	0.00361	2.609***	1660	0.138
	(0.0420)	(0.0372)		
Life determined by others in HH	-0.0632	1.951***	1660	-3.239
	(0.0541)	(0.0482)		
Household Dietary Diversity Score	0.342**	7.747***	1356	4.412
	(0.112)	(0.0970)		
Food Consumption Score Raw	2.186	54.54***	1356	4.007
	(1.163)	(1.022)		
Any work in the past 7 days	0.0195	0.767***	2474	2.541
	(0.0232)	(0.0209)		
Hours worked/helped in household agriculture in last 7 days	0.617	20.73***	912	2.977
	(1.154)	(0.996)		
Any HE in the last 12 months	0.0609*	0.651***	1442	9.354
	(0.0299)	(0.0262)		
Likelihood of negative effect on HH	-0.0494	3.239***	1530	-1.525
	(0.0680)	(0.0601)		
Any shock since 2022	0.0275	0.883***	1494	3.115
	(0.0190)	(0.0171)		

Note: OLS regressions of various outcomes on a dummy for whether the household owned a phone before data collection. Sample: Phone and in-person interviews from identical respondents. Clustered standard errors at individual level in parentheses. *** p<0.01 ** p<0.05 * p<0.1.

Figure A1. Dates of first-time interviews across survey mode

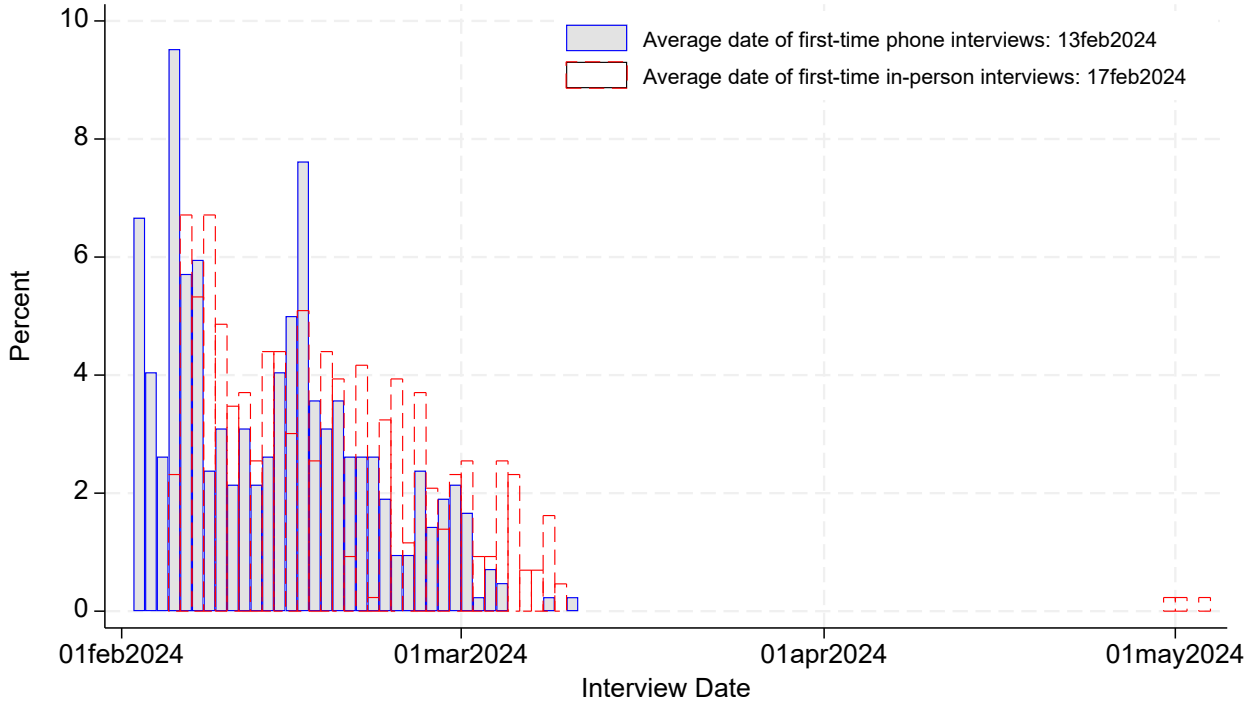


Figure A2. Number of days between interviews of same respondents

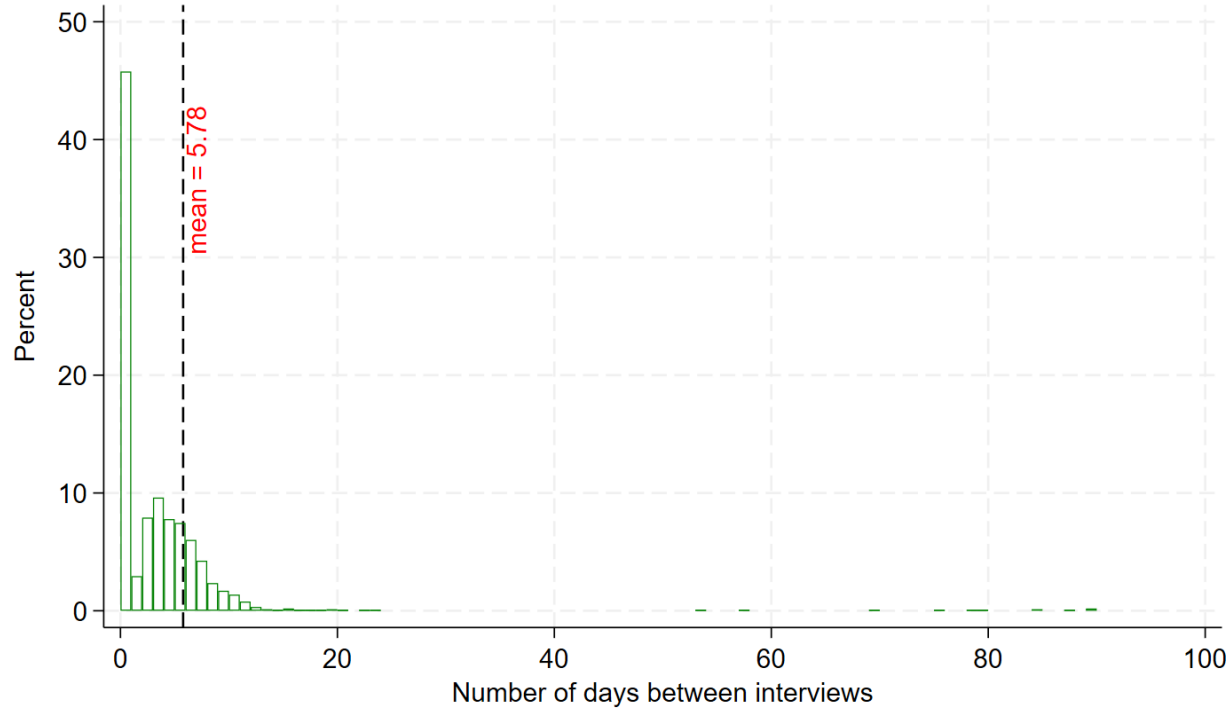


Figure A3. Distribution of mode effects estimates across the two designs

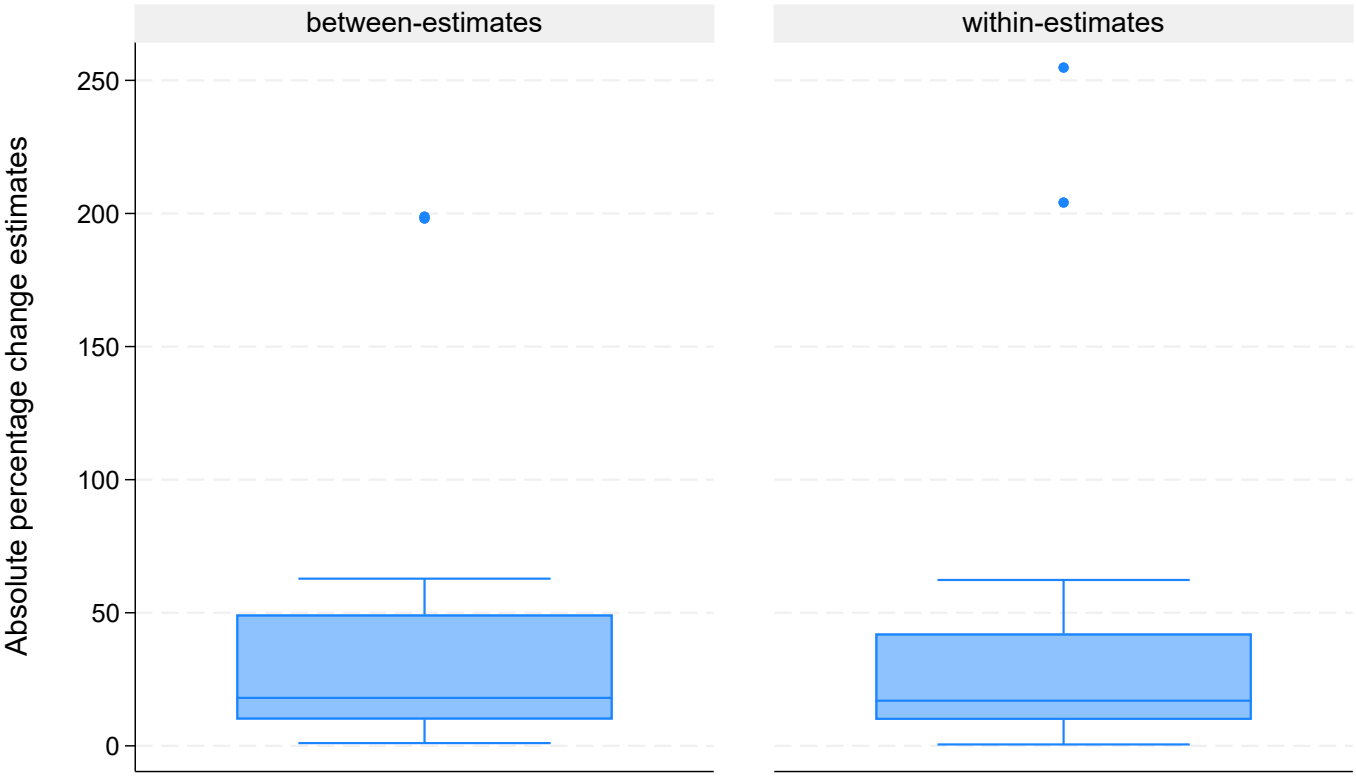
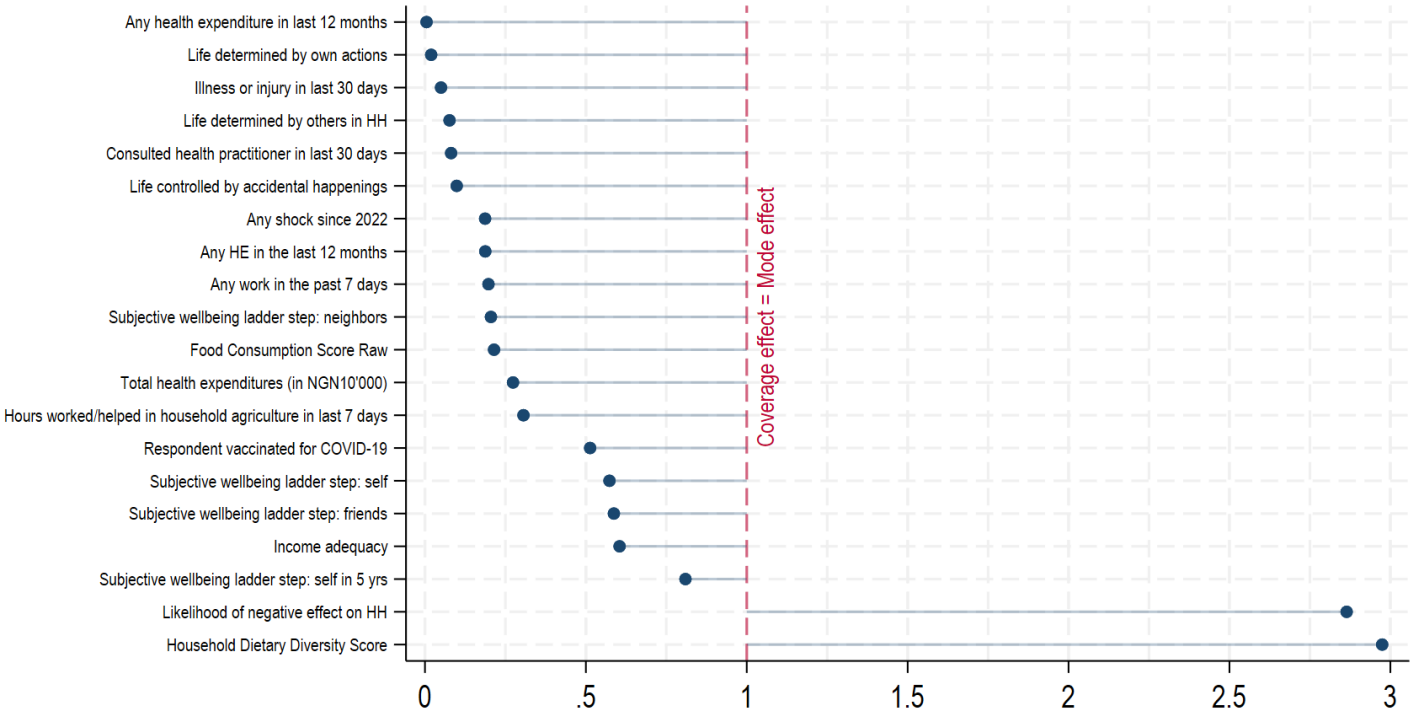


Figure A4: Coverage effect size relative to mode effect size



Note: Absolute size of coverage effect (% deviation of outcome for households owning a phone from that of households not owning a phone at baseline) relative to the mode effect (% deviation of phone interviews from in-person interviews). Values smaller than one indicate coverage effect < mode effect; and values greater than one indicate that coverage effect > mode effect.

Appendix B. Codebook and Variables

Table B1. Codebook of variables

Domains	Label	Type	Level	Part of "extra-module"
Health	During the last 4 weeks, did you suffer from an illness or injury?	Binary	Individual	Yes
	During the past 4 weeks have you consulted a health practitioner, dentist, traditional healer, drugstore, chemist, pharmacy or a Patent Medicine Vendor or visited a health centre?	Binary	Individual	Yes
	Over the past 12 months, did the household purchase or pay for any health expenditure?	Binary	Household	No
	In total, how much did the household spend on health expenditure in the past 12 months?	Count	Household	No
	Have you been vaccinated for COVID-19 (coronavirus)?	Binary	Individual	Yes
Subjective wellbeing	I would now like to know about your general wellbeing. Imagine ten steps, at the top of the ladder (Number 10) are the people who are the best off, those who have the most money, most education, and best jobs. At the bottom (Number 1) are the people who are the worst off, those who have the least money, least education, worst jobs, or no job. On which step are you today? On which step are most of your neighbors today? On which step are most of your friends today? On which step do you expect to be in 5 years?	Categorical	Individual	Yes
Income	Considering the level of your current household income, would you say that you are living... well…fairly well…fairly…with difficulty	Categorical	Household	Yes
Locus of control	Please indicate the degree to which you agree with each of the following statements. Disagree/ Neither Agree nor disagree/ Agree. "To a great extent, my life is controlled by accidental happenings." "My life is determined by my own actions." "I feel like what happens in my life is mostly determined by others in my household."	Categorical	Individual	Yes
Weather extremes expectations	How likely is it that extreme weather events will negatively affect you and your household financially during the next 12 months?	Categorical	Household	Yes
Labor	In the last seven days, did you do any work for someone who is not a member of this household for payment in cash or in-kind?	Binary	Individual	No
	In the last seven days, did you work in a non-farm household business that you operate, for one or more hours?	Binary	Individual	No
	In the last seven days, did you work on household farming, raising livestock, fishing or forestry activities, for one or more hours?	Binary	Individual	No
	How many hours did you work or help on a household farming, raising livestock, fishing or forestry activities in the last seven days?	Count	Individual	No
Food consumption score	Over the past 7 days, how many days did you or others in your household consume any [...]? Grains and Flours; Starchy roots tubers and plantains; pulses, nuts and seeds; etc.	Count	Household	No
Household dietary diversity score	Over the past 7 days, how many days did you or others in your household consume any [...]? Grains and Flours; Starchy roots tubers and plantains; pulses, nuts and seeds; etc.	Count	Household	No
Household Enterprise	During the past 12 months, has anyone in your household… YES/NO for 8 categories of activities.	Binary	Household	No
Shock occurrence	Since January 2022, has your household been affected by...? YES/NO for 28 categories of shocks.	Binary	Household	No