The effect of survey mode on economic measurement: Experimental evidence from Nigeria

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

Abstract

We provide causal evidence that survey mode—in-person versus over the phone—has first-order implications for the reliability of data in LMICs and policies based on such data. Across 20 common economic outcomes in a representative sample from rural Nigeria, survey mode causes a median difference of 17-18%. These effects are large relative to other errors in phone surveys, such as under-coverage of households without phones. Using a within-respondent design, we can measure the full, respondent-level distribution of mode effects and find them to vary systematically between respondents and much more than averages reveal. This directly affects the suitability of standard econometric adjustments for mode effects. It also causes the same pro-poor policy to identify vastly different beneficiaries depending on data collection mode. As demand for high-frequency, cost-effective data in LMICs grows, we discuss levers to mitigate mode-related measurement errors.

Keywords: Measurement error, Data quality, Survey mode, Survey data, Phone survey
JEL codes: C81, C83, O12

[1] The authors are with the Survey Unit, Development Economics Data Group, World Bank. Markhof is with ETH Zurich. 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. The author ordering was constructed through American Economic Association's randomization tool (confirmation code: zfsEk8v3EaGc).

1 Introduction

High-quality socioeconomic data is fundamental for advancing research and for guiding, targeting, and monitoring humanitarian and development interventions. Surveys remain a critical source of representative data and a key tool for "ground truthing" socioeconomic outcomes in low- and middle-income countries (LMICs). At the same time, mixed-mode and phone-based data collection have become increasingly central to economic measurement in these settings due to their lower cost, operational flexibility, and ability to support timely and higher-frequency data collection (Gourlay et al., 2021a; Headey and Barrett, 2015; Zezza et al., 2023). Their use has expanded rapidly in recent years, accelerated by the COVID-19 pandemic but also reflecting broader demand for scalable and cost-effective approaches to data collection amid increasingly frequent shocks and constrained research and policy budgets. As these survey modes become a more permanent feature of the data landscape in LMICs, an increasingly important question is how survey mode itself systematically shapes measured economic outcomes and, in turn, empirical inference and policy design.

Like all data sources, phone and in-person surveys are subject to errors from multiple sources, but survey mode itself is increasingly recognized as an important source of measurement error and bias. The literature on mode effects—systematic differences in responses attributable to the data collection mode rather than true differences in underlying outcomes—is still nascent, with limited and somewhat inconclusive experimental evidence.[1] Most studies cannot fully separate mode effects from confounding factors such as coverage bias or questionnaire differences, leaving uncertainty about the magnitude and direction of mode effects. Beyond these identification challenges, existing studies can only recover average mode effects, such that the individual-level prevalence, heterogeneity, and distributional consequences of survey mode remain unknown.

In this study, we provide causal evidence that survey mode generates pervasive, heterogeneous measurement error with first-order implications for economic measurement in LMICs and for policies based on it. We conduct a large-scale study in a representative sample of rural Nigeria, Africa's most populous country. As part of the experiment, we distributed phones to 937 households in randomly selected rural villages 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. Critically, this within-respondent design, which, to our knowledge has not been previously applied to study mode effects, allows us to observe mode effects at the individual level rather than only on average. Together, these two designs allow identifying mode effects and distinguishing

[1] Evidence on mode effects is mixed and outcome-specific. 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.

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

We find robust evidence for economically meaningful mode effects across a wide range of outcome variables capturing core dimensions of welfare, risk exposure, and economic behavior in LMICs and that are heavily used in research and policy. Responses collected over the phone differ from the in-person estimates by 18% (between-respondent design) and 17% (within-respondent design) at the median and range from 0% to 63% (excluding two identified outliers), across twenty outcomes. Results from the between- and within-respondent design suggest very similar conclusions in sign and size of mode effects, lending further credibility to the robustness of these results. Mode effects are positive for 70% of outcomes, suggesting a tendency for respondents to provide more affirmative or frequent responses by phone.

Leveraging our unique within-respondent design (List, 2025), we recover individual-level mode effects, yielding new insights that averages and conventional between-respondent designs miss.[3] First, there is substantial heterogeneity in mode effect size and sign. Second, because mode effects of different signs partially cancel out on average, survey mode matters for the responses of a much larger share of respondents than what averages let on. This heterogeneity implies that average mode effects may substantially understate the extent to which survey mode alters measured outcomes at the respondent level. Across outcomes and respondents, survey mode causally alters measured outcomes in 55% of cases. Third, observing mode effects at the respondent level allows us to pool more than 13,000 respondent-by-outcome pairs and identify systematic, as opposed to idiosyncratic, patterns in mode effects that hold more generally across topics for which economists collect (phone) survey data. We find that respondents with higher education levels are systematically less prone to mode effects. We further document recurrent 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.

In an applied exercise, we demonstrate how heterogeneous mode effects can distort policy targeting. We simulate a simple pro-poor targeting exercise in which a policymaker ranks households according to a welfare metric and targets the bottom quantile of the resulting distribution. Across three different welfare metrics (food consumption score, out-of-pocket health expenditure, and a composite deprivation index), only 14 –24% of beneficiaries overlap across survey modes when targeting the worst-off quintile. For the rest, receipt of assistance depends on whether eligibility is determined using phone or in-person survey data. Survey mode therefore materially alters beneficiary assignment, with direct implications for the effectiveness, efficiency, and equity of development policies.

We then benchmark mode effects against other sources of errors in surveys. Coverage errors from non-universal phone ownership also matter: phone ownership is non-random and associated with

[2] Coverage errors occur due to differences between households with and without a phone or differences in response rates between modes (Ambel et al., 2021a). Respondent selection effects comprise differences between respondents in phone and in-person interviews (Glazerman et al., 2023). Timing effects describe real changes in outcomes over time. Order effects occur when response behavior is affected by a previously conducted interview (Caceres-Santamaria, 2021; Struminskaya and Bosnjak, 2021).

[3] In several tests made possible by our experimental design, we confirm that these estimates satisfy the key identifying conditions for interpreting respondent-level mode effects causally: temporal stability and no ordering effects ("causal transience" in the framework of List (2025)).

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% the size of mode effects at the median across twenty 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 results present the most definitive and comprehensive experimental evidence to date on mode effects in surveys in LMICs and their implications for the data underpinning development economics and policy. Compared to the limited existing evidence, we are able to isolate the effect of survey mode from other confounding errors and use a within-design (Charness et al., 2012; List, 2025) to measure the full, respondent-level distribution of mode effects. This lets us quantify the amount of heterogeneity in mode effect size and direction, allows to identify patterns that hold more generally across topics, variables, and respondents, and demonstrates how survey mode choice distorts policy decisions. In a time of growing demand for large-scale, high-frequency socioeconomic data at affordable cost, our findings highlight the importance of identifying and mitigating mode-related *measurement* error for reliable empirical inference and policy design in LMICs. This complements evidence on *sampling* errors in mixed-mode surveys (e.g., Ambel et al., 2021; Brubaker et al., 2021) which are better understood with established tools for mitigation.

These results have several fundamental implications for economic research and data quality in LMICs. First, measurement, not sampling errors, are likely to present the binding data quality constraint for more cost-effective, scalable, and rapidly deployable socioeconomic survey data in LMICs. As access to mobile technology becomes universally available and cheap to provide, coverage errors can be effectively mitigated—for example, via re-weighting approaches (Ambel et al., 2021; Brubaker et al., 2021) or, as in our study, via phone distribution. Conversely, non-negligible, systematic, and easily missed mode-related measurement errors in survey data may pose a more persistent challenge to economic measurement and inference in LMICs.

Second, researchers and national data systems can navigate this challenge under certain conditions through the judicious and purposive deployment of phone surveys. Phone survey data can allow for unbiased inference of population averages, for example, in the estimation of average treatment effects in experimental research (where measurement errors in continuous outcomes cancel out in expectation between treatment arms) as well as in research designs that allow differencing-out the average effect of survey mode. The latter may be achieved through randomly chosen, mixed mode validation samples; or in the estimation of time trends through repeat observation within mode (as is the case for panel data). However, in any of these cases, the substantial heterogeneity in mode effects that we document will turn inference less precise and decrease statistical power. This may contribute to a trade-off between phone surveys' lower marginal costs and greater variance in estimates (see Anderson et al. 2024). Randomization also does not protect against attenuation in estimated treatment effects when the outcome is binary. For example, we find mode effects for common, binary economic outcomes of interest such as whether the respondent is employed, had an illness or injury episode, operated a non-farm enterprise, or reports having been exposed to an adverse shock.

Third, heterogeneity in respondent-level mode effects implies that average adjustments for survey mode may be insufficient for many applications in empirical economics, particularly those relying on subgroup comparisons or rankings. Aggregate corrections for survey mode will not fully

address, or even exacerbate, differences in measurement error across subgroups whose measurement error deviates substantially from the mean. This poses a challenge for applications such as conditional predictions made from regression models, distributional analysis, or estimates of treatment effect heterogeneity in RCTs. The targeting simulation we conduct—resulting in two materially different notions of the group that is most in need—provides one such example. In these cases, unbiased inference may require adjusting for mode effects within each subgroup of interest through (i) subgroup-specific mode corrections from adequately powered calibration samples, or (ii) testing for and including mode-by-covariate interactions.

Finally, realizing the potential of phone surveys for economic research and policy in LMICs requires more evidence on scalable strategies to improve measurement reliability and reduce survey errors. While recent evidence has advanced our understanding of the existence and extent of measurement errors in survey data in LMICs (e.g., Celhay et al., 2024; Meyer et al., 2015; Markhof et al., 2026), effectively mitigating them takes more causal evidence on their sources (respondents and enumerators), the exact mechanisms at play, the mediating role of survey mode, and the impact of scalable interventions that address key error sources. Our study provides a foundation for these efforts and reveals areas with large potential for data quality gains in LMICs. For example, this includes strategies such as cognitive testing of questionnaires across survey modes, adapting real-time survey management systems to track indicators of data quality and flag issues as they occur, and systematically incorporating experimentation into survey operations to uncover and mitigate survey design effects. As phone surveys and non-traditional data collection modes are likely to remain a lasting feature of official statistics and researchers' toolkits in LMICs, this growing research agenda is central to safeguarding reliable inference, accurate descriptive statistics, strong monitoring systems, and precise predictive models in development data and economic research.

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 respondent-level mode effects and heterogenous measurement error. Section 6 presents our applied policy exercise. 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:

$$Y_i = Y_i^* + e_i \tag{1}$$

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. The survey mode, in our case phone vs. 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 to 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 A1 from the Supplemental Appendix). 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.

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., 2020). Supplementary Figure A1 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.

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 without an external, objective benchmark for the true Y_i^*, 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 vs. in-person) on measured outcomes, we conduct 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 (EAs) 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 EAs were eligible for the experiment and we randomly selected 106 for our experimental sample.

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 A2 from the Supplemental Appendix 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. In addition, we aimed to interview a second household member aged 15 or older over the phone.[4] 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 (Supplementary Table B1). 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).[5]

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

[4] 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. 933 households completed the post-harvest in-person interview.

[5] 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).

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 modelled 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. T1-P and T2-F show no systematic differences across a range of household and respondent characteristics (joint orthogonality p=0.7), with the only imbalance being a slightly higher dependency ratio for in-person households (Table 1). 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 1).

(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 (Supplementary Figure A3, Panel A). 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 (Supplementary Table B2). 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 (Supplementary Figure A3, Panel B), an interval sufficiently short to make meaningful changes implausible for most outcomes. Table B3 from the Supplemental Appendix confirms this with regressions of outcomes on interview date, controlling for survey mode, and shows 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.[6] 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:

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

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 (Table B4 and B5 in the supplemental appendix).[7]

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

[6] We find our final analytical sample to be balanced on respondent characteristics also after attrition and non-response (Table 1).

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

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:

$$Y_{ihk} = \tau_1 + \varepsilon_{ihk} \tag{8}$$

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

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

where Y is defined as in equation (8), $\boldsymbol{X}$ is a vector of respondent characteristics, $\boldsymbol{H}$ a vector of household characteristics, $\boldsymbol{E}$ a vector of enumerator controls, and ϑ_k is a set of outcome fixed effects. [8] 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.

3.3 Outcomes of interest

Mode effects and other survey errors vary across different variables depending on such factors as the topic, the unit of observation, the variable type, and its complexity. As a distinct strength of this experiment compared to the existing evidence, we study a set of outcomes that capture core dimensions of welfare, risk exposure, and economic behavior in low- and middle-income countries, at the intersection of research and policy design. Substantively, these outcomes span health shocks, health care utilization, and out-of-pocket medical expenditures; labor supply and household enterprise activity across farm and non-farm sectors; food consumption and dietary diversity; and subjective wellbeing, relative economic position, and expectations, alongside detailed measures of realized and anticipated shocks, including weather-related risks. These variables are used in applied microeconomic research and across a range of data-generation contexts, such as randomized controlled trials, quasi-experimental designs, and high-frequency household panels. For example, health utilization and medical spending are key outcomes in studies of health insurance and financial protection (e.g. Ahmed et al., 2022; Dupas and Miguel, 2017; Finkelstein et al., 2012); labor supply and household enterprise activity are key margins in work on livelihoods, informality, and self-employment under credit and risk constraints (e.g. Blattman et al., 2014; McKenzie and Puerto, 2021; Somanathan et al., 2021); food consumption and dietary diversity feature prominently in evaluations of social assistance and transfer modalities (e.g. Cunha et al., 2019; Hoddinott et al., 2018); and subjective wellbeing and expectations have been shown to respond to income and economic security in experimental settings (Banerjee et al., 2015; Haushofer and Shapiro, 2016).

[8] Household characteristics include household size and log total consumption. 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, interview language spoken at home, and experience (worked on five or more surveys in the past) as there was barely any variation in enumerator education. As enumerators differ between the phone and in-person interview and either enumerator 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.

These indicators also often form the basis of policy decision-making by governments and international organizations. For example, food consumption and diet quality indicators are used operationally in food security monitoring systems and global reporting exercises such as the State of Food Security and Nutrition in the World (SOFI), informing the design and scaling of social assistance and humanitarian responses. Similarly, measures of health care utilization and out-of-pocket medical expenditures underpin monitoring of universal health coverage and financial-risk protection, including SDG indicator 3.8.2, and guide interventions and health financing reforms supported by international institutions.

Finally, 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 B7 in the Supplemental Appendix.

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.

Figure 1 shows the estimated mode effects from both the between-respondent (BR) and within-respondent (WR) designs, expressed as percentage changes relative to the in-person mean. Supplementary Table B8 reports the results in natural units. 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. 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.[9]

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%).

[9] The health expenditure variables 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 A4 in the supplemental appendix shows the distribution of 20 mode effect estimates for each design, expressed as percentage change compared to the in-person mean.

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 respondent characteristics that shape the presence and magnitude of mode effects.

5 Respondent-level mode effects and heterogenous measurement error

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 2, there is a lot of heterogeneity in the direction and size of respondent-level mode effects underlying the average mode effects we reported previously. 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: respondent-level mode effects of different signs cancel each other out. A case in point is the perceived likelihood of adverse impacts from 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 2 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 2). 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.

When looking at continuous variables only, the absolute magnitude of mode effects varies between different respondents (Figure 3). 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.[10] Mode effects exceed one standard deviation for 43% of respondents and distributions have long tails (Panel A). 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 (Panel B). Conversely, for a third of respondents, they differ by three or more steps.

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 effect are variable-specific. Focusing on the 16 binary and categorical outcome variables, Figure 4 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 outcomes. Respondents tend to provide inconsistent responses for about half of outcomes. For the median respondent, answers over the phone and in person differ for 56% of outcomes. Relative to the amount of spread we would expect by chance, the observed variance is 1.54 times as large.[11] This suggests that about a third (30-35%) of the observed spread in Figure 4 is down to systematic differences between respondents.

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 3). 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.

5.3 Respondent behavior and mode effects by survey design factors

Several patterns emerge from Figure 2 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

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

[11] We calculate the ratio of the observed variance in respondent-level inconsistencies to the expected variance (i.e., the variance we would expect even if respondents did not differ in how prone they are to mode effects). The main finding we report is for those respondents who answered all 16 binary and categorical outcomes. Pooling across all respondents that answered at least one outcome for both modes, we estimate a ratio of 1.43 (30%).

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 (Table 4).[12]

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 5). 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 B9) 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 5).

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 vs. in-person surveys.

6 Implications for policy targeting

How does data collection mode affect policy? Consider a policymaker whose goal is to support the most deprived part of the population by rolling out a new intervention. As resources are limited, the policymaker intends to target a fixed segment of the population (e.g., the bottom 20%). Who gets targeted is decided based on ranking households according to the pre-intervention outcome distribution. For this, the policymaker can draw on otherwise identical data that was either collected via a phone survey or an in-person survey of the potential target population. Formally:

$$T_i^M = \mathbb{1}\left[Y_i^M \leq Q_p^M\right], M \in \{P, F\}$$

where T_i^M is an indicator equal to one if household i is selected as a beneficiary under mode M, Y_i^M is the measured outcome of this household under mode M, and Q_p^M is the p-th quantile of the outcome distribution under mode M.

[12] 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 agreement of reported responses exceeds 'as if random' (1/k, with k response options) 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'.

Assume that, as data collection is costly, the policymaker will only ever have data from one source at their disposal. Our goal is to compare who is selected as a beneficiary depending on whether the policymaker happens to draw on phone or in-person survey data. The key metric we are interested in is the extent of overlap in beneficiary populations between the phone-based and in-person based targeting approach. Formally

$$J(p) = \frac{\sum_i T_i^P \cdot T_i^F}{\sum_i \mathbb{1}\left[T_i^P + T_i^F \geq 1\right]}$$

where the numerator counts households targeted under both modes and the denominator counts households targeted under at least one mode. $J(p) = 1$ indicates perfect agreement and $J(p) = 0$ indicates no overlap.

How will data collection mode affect beneficiary selection? In the simplest case of no mode effects, phone and in-person survey data will produce identical outcome distributions and selected beneficiaries under either approach will coincide exactly ($J(p) = 1$). Similarly, if survey mode produces a shift in the distribution that is homogenous across the population, the relative position of each household in the distribution will remain unaffected and both data sources will indicate the same beneficiaries. However, if survey mode changes measured outcomes in ways that differ across households, the ranking of a household may change depending on the data source used. The extent of such re-ranking, and ultimately differences in beneficiary lists, depends on the heterogeneity of mode effects among households near the targeting threshold. When heterogeneity is relatively small, only a few marginal households may switch beneficiary status across modes. When it is large, the ranking is more substantially scrambled. In the extreme case where the ranking under one data source is completely uninformative for the ranking under the other source, any overlap in targeted populations will be coincidental and amount to

$$J^B(p) = \frac{p}{2 - p}$$

which gives the expected overlap when rankings under the two modes are statistically independent.[13] This provides us with a benchmark against which to compare the actual, empirical overlap in beneficiaries across the two data collection modes.

We conduct this exercise for three different hypothetical programs and sectors: a food security intervention targeting the most food insecure households, a financial protection intervention targeting households with the greatest out-of-pocket health expenditures, and a social protection intervention targeting households based on a multidimensional deprivation index.[14]

---

[13] Under statistical independence, knowing a household's targeting status under one mode is uninformative about its status under the other. Since both modes target a share p of the same population, $P(T_i^P = 1) = P(T_i^F = 1) = p$. Independence implies $P(T_i^P = 1 \cap T_i^F = 1) = p^2$. By inclusion-exclusion, $P(T_i^P = 1 \cup T_i^F = 1) = 2p - p^2 = p(2 - p)$. The expected overlap is then $\frac{p^2}{p(2-p)} = \frac{p}{2-p}$.

[14] The deprivation index emulates the common practice of using composite welfare indices to proxy for living standards, including subjective assessments of wellbeing (Ranucci et al., 2027). Our outcomes do not include durable assets or dwelling characteristics which are often part of conventional proxy means tests. Instead, the deprivation index takes the following inputs covering a broader range of welfare indicators: respondent not in wage employment, respondent had illness or injury in last 30 days, respondent households has been affected by an adverse

Figure 5 visualizes the overlap in selected beneficiaries between phone- and in-person survey data relative to the uninformative benchmark. This allows for three main observations. First, to a great extent, who is selected as a beneficiary depends on survey mode. Across the three simulated programs and reasonable program sizes, targeting under different survey modes overlaps only slightly more than what would be expected if both data sources captured entirely independent information. For example, only 14% of beneficiaries of an intervention targeting the most food insecure quintile would be identical between phone and in-person data; 14% of an intervention targeting the quintile with highest out-of-pocket expenditures; and 24% for an intervention targeting the worst-off quintile based on a multidimensional deprivation index. This compares to a benchmark overlap of 11% that we would expect by pure chance. Second, beneficiary selection is similarly sensitive to the underlying data source for more narrow programs (e.g., targeting the worst-off 10%) or broader programs (e.g., targeting the bottom/top half) (Supplementary Table B10).[15] Third, for narrow programs targeting only the worst-off households, the composite deprivation index performs somewhat better than targeting on a single outcome. In light of the variation in mode effects we find for the same respondents across outcomes (Section 5.1), this may reflect slightly less noise in a composite measure than in any variable individually.

These patterns are a direct consequence of the large and heterogeneous mode effects we find in Section 5. Since survey mode shifts measured outcomes differentially across households, household rankings—and therefore beneficiary lists—are substantially reshuffled depending on the data source. As policymakers typically draw on a single data source when determining program eligibility, the choice of data collection mode may substantially determine who receives assistance and who does not. Our exercise reveals this usually unobserved counterfactual.

7 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, larger and wealthier household significantly more likely to own a mobile phone (Table 6). Outcomes of interest differ between households that were previous phone owners and those that were not.

shock in last 2 years, self-reported income adequacy, expected likelihood of adverse future weather shock exposure, respondent subjective wellbeing.

[15] High prevalence of zero out-of-pocket health expenditures means overlap between both data sources is somewhat higher at high inclusion thresholds. Differences in the discreteness of reported data and the prevalence of zeros are a characteristic of the data generating process for each survey mode and part of what we want to capture here.

Previous phone ownership is also associated with significant differences in the outcomes of interest (Supplementary Table B11, Figure 6): 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 (Supplementary Figure A5). Further, the effective bias that coverage effects will introduce to general population estimates diminishes as the absolute size of the group of non-phone-owning households shrinks.

A caveat to this comparison is our limited ability to realistically assess unit non-response in phone surveys. Non-response in the phone surveys was very low in this experiment at 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.[16] 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-Panel 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).

Endowment effects

Turning to measurement errors resulting from survey design and implementation choices, we can assess whether providing mobile phones 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 (Supplementary Table B6). 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

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

ordering effects affecting survey responses (Supplementary Table B4 and Supplementary Table B5).

8 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. Finally, by leveraging a within-respondent design, we uncover substantial heterogeneity in mode-related measurement error and examine the underlying mechanisms. This offers deeper insight into how and for whom survey mode matters and how it affects development policy.

Our findings suggest that mode effects can be economically meaningful. They are in line with and can even exceed effect sizes of large-scale interventions or programs. For instance, the size of mode effects we find is approximately commensurate with estimates of the effect of Ethiopia's pilot community based health insurance on outpatient healthcare utilization (Mebratie et al., 2019), the impact of winning a cash grant through a national business plan competition in Nigeria on operating a non-farm enterprise after three years (McKenzie, 2017), or of business trainings plus cash grants in Sri Lanka on the short-run likelihood to run a business (de Mel et al., 2014).

The mode effects we find also have first-order implications for policy which we demonstrate in a targeting exercise covering welfare metrics in three different sectors: food security, health/financial risk protection, and social protection. In each of these cases, the simple fact of whether phone or in-person survey data is used to determine program eligibility results in the selection of two materially different beneficiary populations. The choice of data collection mode in such programs is typically dictated by contextual constraints: budget (phone surveys are generally cheaper), urgency (phone surveys can be deployed rapidly when no current data exists), feasibility (natural disasters or active conflict may preclude in-person data collection), and population coverage (phone surveys might miss populations with low mobile penetration). Our exercise reveals a counterfactual that is usually unobserved in practice but exceedingly consequential in our application: who appears eligible for assistance is substantially determined by survey mode. This suggests potential for large inclusion and exclusion errors in means-tested social assistance depending on whether data was collected over the phone or in person.

While our study offers the most rigorous evidence on mode effects to date, there are several important caveats and limitations to our study. 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,

mitigating differences in survey length between both modes. However, 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 therefore 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, they have to be shorter, and sometimes 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[17]. 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, albeit one of the world's most populous, 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. Our findings 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, differences and 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. Similarly, it strengthens experimental designs that can draw on calibration samples or that hold measurement error in continuous outcomes constant by virtue of randomization (provided treatment itself does

[17] For a review on emerging best practices in benchmarking survey experiments, see

not induce differential measurement error). In any of these cases, heterogeneous mode-related measurement error will, however, reduce precision and statistical power, even where average mode effects can be differenced out or otherwise accounted for. Randomization also does not protect against attenuation in the case of mismeasured binary or categorical outcomes which will pull estimated treatment effects toward zero.

More fundamentally, our findings suggest that heterogeneity in mode-related measurement error, rather than average mode effects alone, may be a central challenge for inference and policy applications relying on survey data. Standard adjustments, such as including a simple mode dummy in a regression, will not fully address, or even exacerbate, the confounding effect of survey mode for subgroups whose measurement error deviates substantially from the mean. Such scenarios are common in applied economics research and include, for example, conditional predictions made from regression analysis, distributional analysis, or estimates of treatment effect heterogeneity in RCTs.[18] In these cases, unbiased inference may require adjusting for survey mode effects within each subgroup of interest, for example via mode-by-covariate interactions or adequately powered calibration samples.

Addressing measurement error and mode effects in phone and mixed-mode surveys, therefore, 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

Our results also point to concrete steps for improving the design and implementation of phone– and in-person–surveys, potentially mitigating mode effects. 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,

[18] While randomization ensures that, in expectation, treatment assignment is uncorrelated with mode-related errors, it does not guard against such measurement error in the mediator variable by which treatment effects are stratified. To the extent that mode effects render measurement of the mediator noisy, this will attenuate estimated treatment effect heterogeneity.

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 to 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.

9 Conclusion

Phone and mixed-mode surveys offer important advantages for economic measurement, including lower costs, operational flexibility, and the ability to support timely and higher-frequency data collection. These benefits are particularly valuable in a 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 and the reliability of economic measurement.

In a representative sample from rural Nigeria, we find strong evidence of economically meaningful survey mode effects across a wide range of outcomes relevant for economic research and

development policy. Using a within-design, we document substantial heterogeneity in respondent-level mode-related measurement error. These findings suggest that heterogeneity in mode-related measurement error may be particularly consequential for applications relying on subgroup comparisons, rankings, or distributional analysis. As phone and mixed-mode data proliferates, our findings offer novel guidance regarding their reliable collection, analysis, and interpretation.

We conclude by highlighting several promising areas for future research. First, more work is needed to unpack the mechanisms driving mode effects, including the role of incentives and enumerators. Qualitative research could also help shed light on how respondents behave and report information across modes. Second, benchmarking data collected through different modes against a reliable measure of truth is crucial to understanding how reported outcomes from each mode may deviate from the true outcome. Third, future empirical studies should explore 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. More attention should also be given to how mode effects change over time and what this means for longitudinal surveys. Finally, while real-time survey management systems are increasingly used in fieldwork, their potential to detect and reduce survey and measurement 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

Ahmed, T., Roberton, T., Vergeer, P., Hansen, P.M., Peters, M.A., Ofosu, A.A., Mwansambo, C., Nzelu, C., Wesseh, C.S., Smart, F., 2022. Healthcare utilization and maternal and child mortality during the COVID-19 pandemic in 18 low-and middle-income countries: An interrupted time-series analysis with mathematical modeling of administrative data. PLoS Med. 19, e1004070.

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

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., Udry, C., 2015. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. Science 348, 1260799. https://doi.org/10.1126/science.1260799
Biemer, P.P., 2010. Total Survey Error. Design, Implmentation, and Evaluation. Public Opin. Q. 74, 817–848.
Blattman, C., Fiala, N., Martinez, S., 2014. Generating Skilled Self-Employment in Developing Countries: Experimental Evidence from Uganda *. Q. J. Econ. 129, 697–752. https://doi.org/10.1093/qje/qjt057
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
Cunha, J.M., Giorgi, G.D., Jayachandran, S., 2019. The Price Effects of Cash Versus In-Kind Transfers. Rev. Econ. Stud. 86, 240–281.
de Mel, S., McKenzie, D., Woodruff, C., 2014. Business training and female enterprise start-up, growth, and dynamics: Experimental evidence from Sri Lanka. J. Dev. Econ. 106, 199–210. https://doi.org/10.1016/j.jdeveco.2013.09.005
Dupas, P., Miguel, E., 2017. Impacts and determinants of health levels in low-income countries, in: Handbook of Economic Field Experiments. Elsevier, pp. 3–93.
Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J.P., Allen, H., Baicker, K., Oregon Health Study Group, 2012. The Oregon Health Insurance Experiment: Evidence from the First Year*. Q. J. Econ. 127, 1057–1106. https://doi.org/10.1093/qje/qjs020
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
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

Haushofer, J., Shapiro, J., 2016. The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya*. Q. J. Econ. 131, 1973–2042. https://doi.org/10.1093/qje/qjw025
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
Hoddinott, J., Sandström, S., Upton, J., 2018. The Impact of Cash and Food Transfers: Evidence from a Randomized Intervention in Niger. Am. J. Agric. Econ. 100, 1032–1049. https://doi.org/10.1093/ajae/aay019
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
McKenzie, D., 2017. Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition. Am. Econ. Rev. 107, 2278–2307. https://doi.org/10.1257/aer.20151404
McKenzie, D., Puerto, S., 2021. Growing markets through business training for female entrepreneurs: A market-level randomized experiment in Kenya. Am. Econ. J. Appl. Econ. 13, 297–332.
Mebratie, A.D., Sparrow, R., Yilma, Z., Abebaw, D., Alemu, G., Bedi, A.S., 2019. The impact of Ethiopia's pilot community based health insurance scheme on healthcare utilization and cost of care. Soc. Sci. Med. 220, 112–119. https://doi.org/10.1016/j.socscimed.2018.11.003
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).
Ranucci, I., Abay, K.A., Tiberti, L., 2027. Perceived need and measured well-being: How well do subjective rankings capture relative poverty? J. Dev. Econ. 185, 103943. https://doi.org/10.1016/j.jdeveco.2026.103943
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

Somanathan, E., Somanathan, R., Sudarshan, A., Tewari, M., 2021. The Impact of Temperature on Productivity and Labor Supply: Evidence from Indian Manufacturing. J. Polit. Econ. 129, 1797–1827. https://doi.org/10.1086/713733

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

Weerdt, J.D., Gibson, J., Beegle, K., 2020. 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

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 figures and tables

Figure 1. Main estimates of survey mode effects

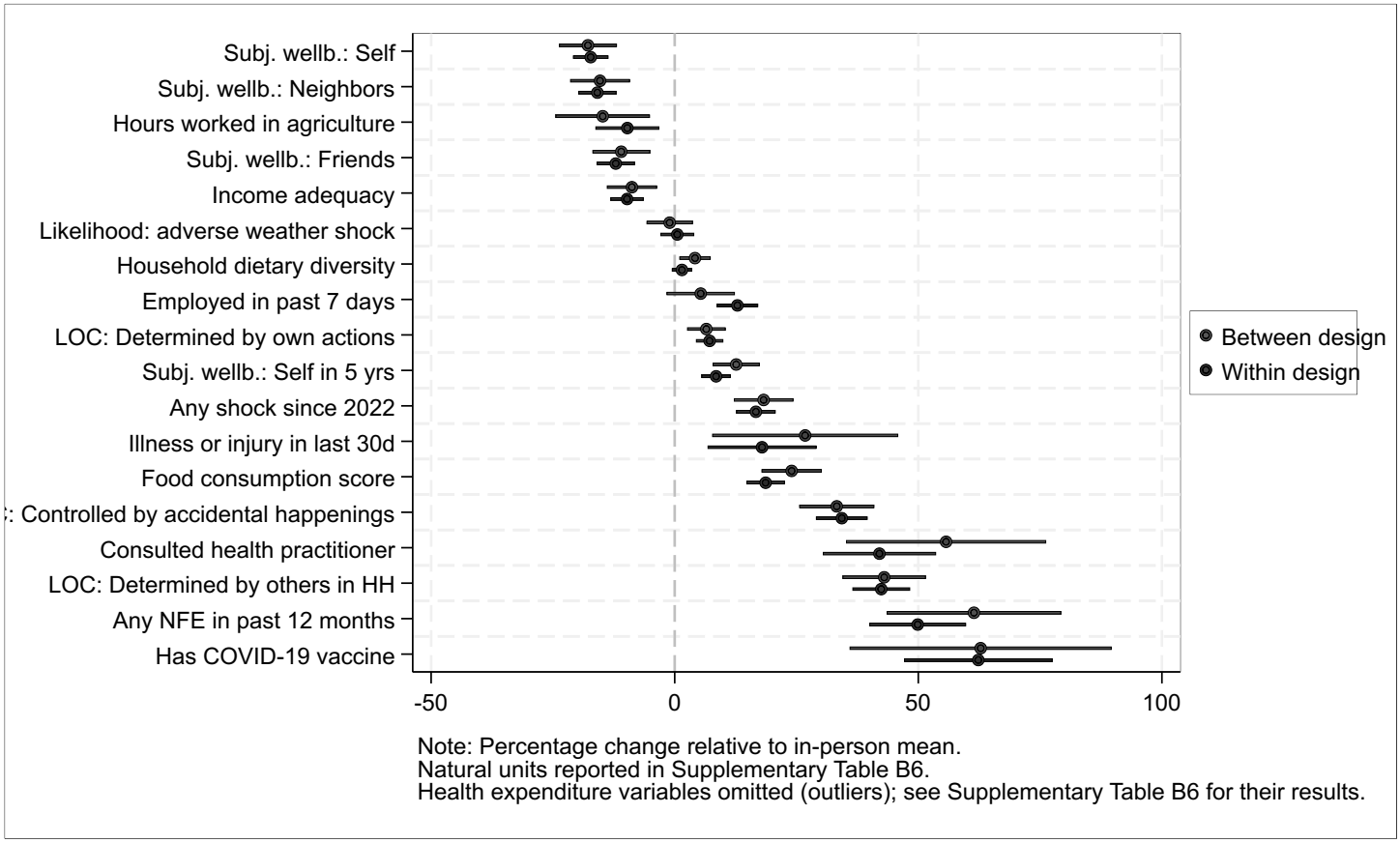
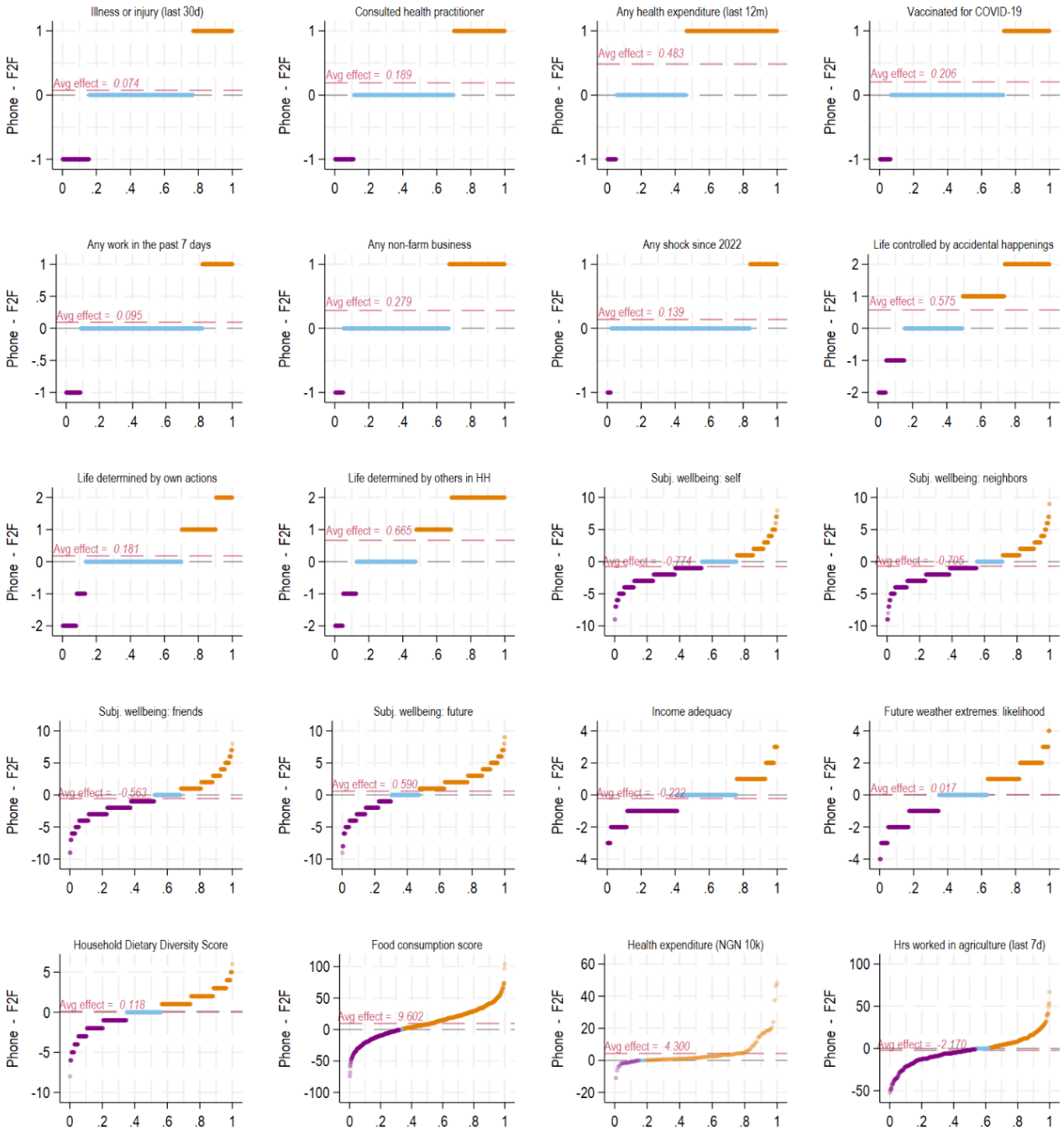
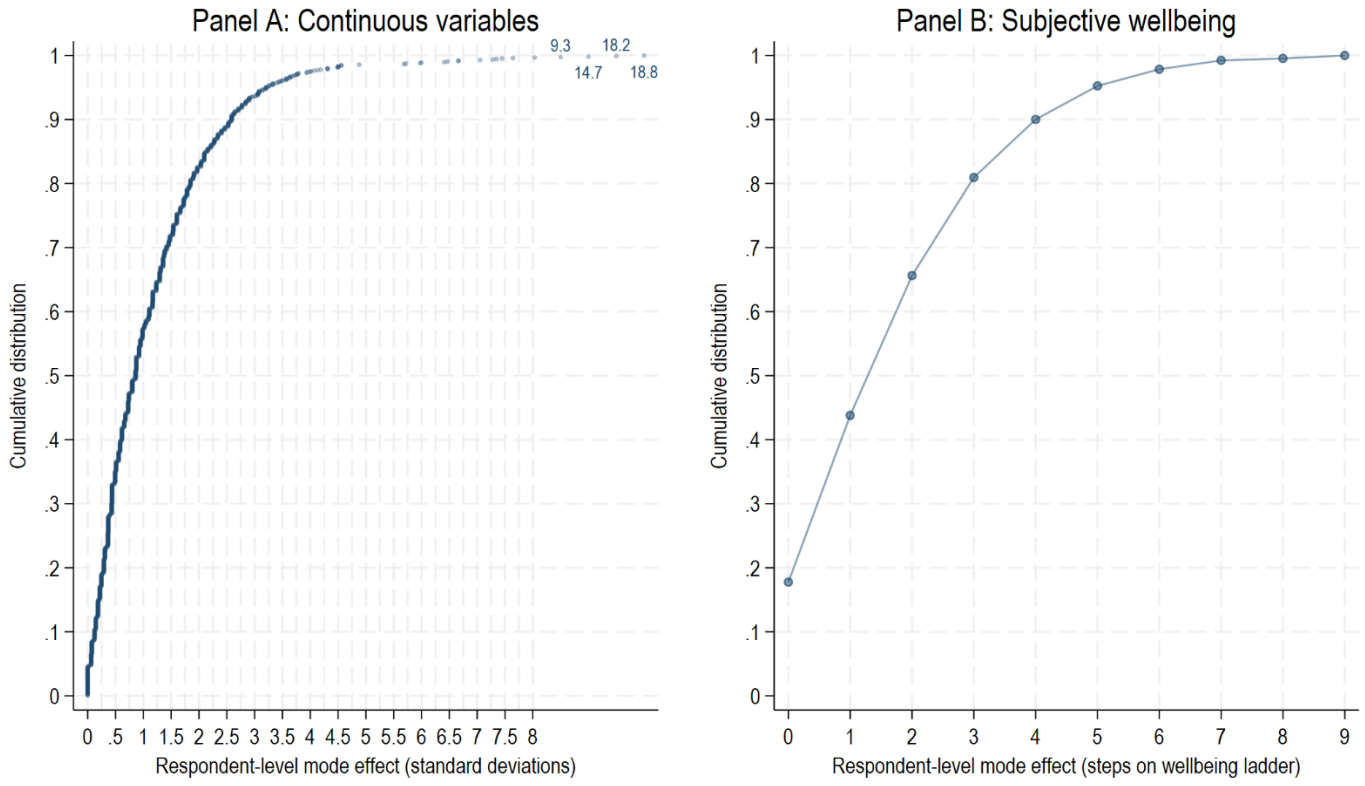


Figure 2. Distribution of respondent level mode effects by outcome



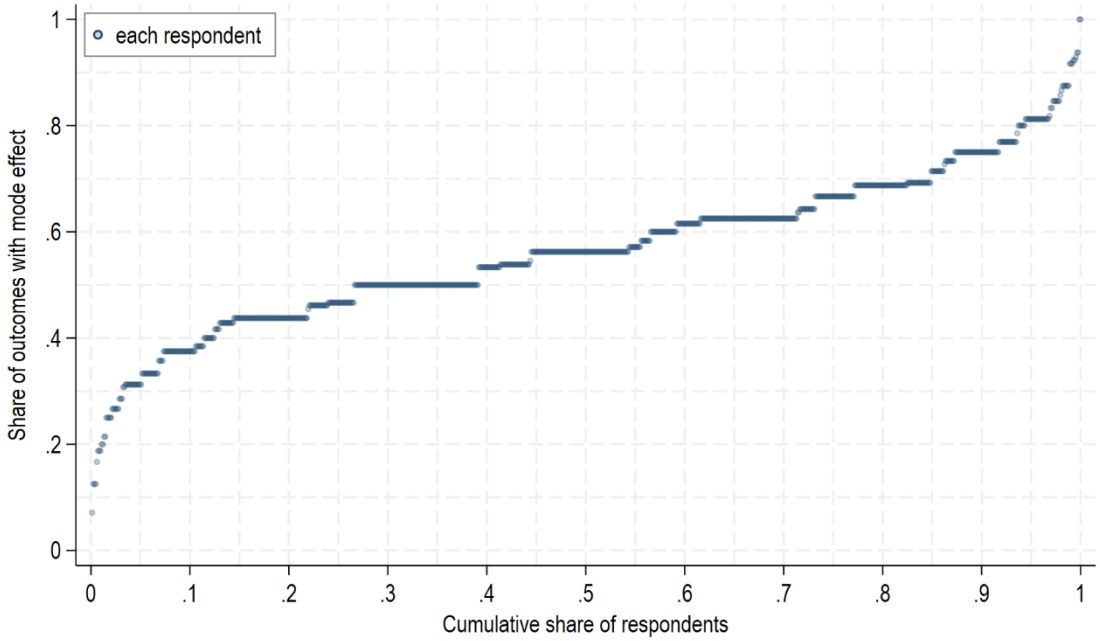
Note: Distribution of respondent-level mode effects. Each marker represents the mode effect for one respondent, calculated as the difference between the respondent's phone survey answer and their in-person survey answer. Markers above (below) the zero line mean that the outcome value measured in the 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.

Figure 3. 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 subjective wellbeing which is measured on a scale between 0 and 10.

Figure 4. Share of non-zero mode effects (binary and categorical variables only)



Note: For each respondent, the figure plots the share of outcomes with a mode effect, i.e., the share of outcomes for which the respondent answered inconsistently between the phone and in-person interview. Respondents are ordered along the x-axis in ascending order of mode effect prevalence. Each marker represents one respondent.

Figure 5: Overlap in targeted beneficiaries between survey modes across different

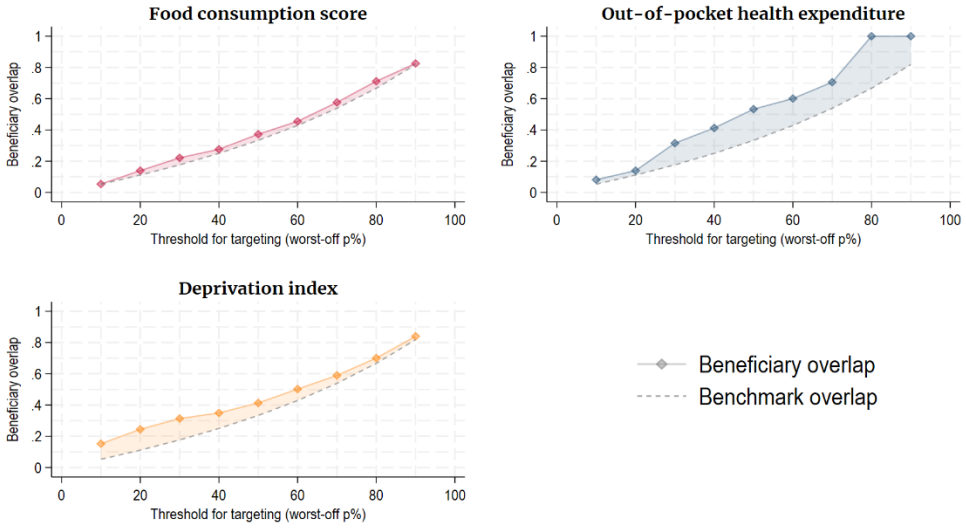


Figure 6: Coverage effect size relative to mode effects

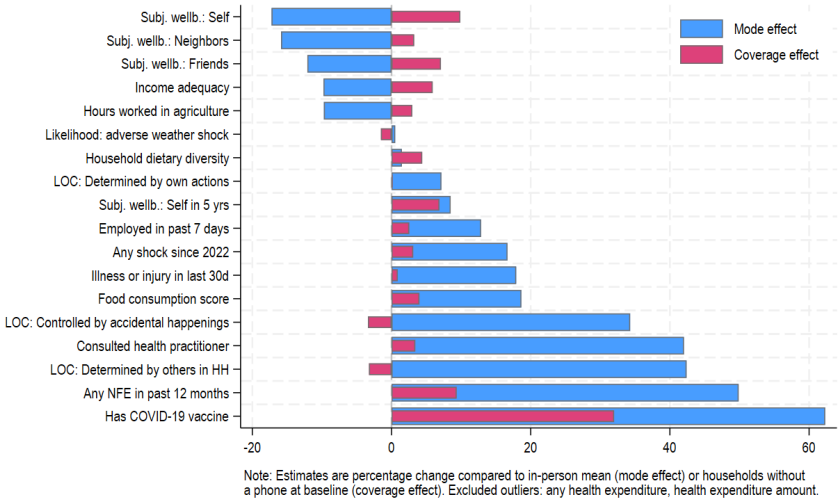


Table 1. Descriptive statistics and balance of characteristics for the between-respondent sample

	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 2. Share of inconsistent answers for binary and categorical variables

	Share in %	Observations
Pooled across all variables	55.24	13272
	(0.488)	
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)	

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.

Table 3. 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 4. Share of inconsistent responses between the phone and in-person interview, by number of response options

	(1)	(2)	(3)	(4)	(5)
	2 options	3 options	4 options	5 options	10 options
Share of respondents with a mode effect	0.3561	0.5819	0.6530	0.7137	0.8224
	(0.0062)	(0.0099)	(0.0165)	(0.0164)	(0.0067)
Observations	5,922	2,490	830	765	3,265

Note: The table shows the share of inconsistent answers by the same respondent when interviewed in person and over the phone, depending on the number of discrete answer options (or) 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).

Table 5. Response patterns

	(1)	(2)	(3)	(4)	(5)	(6)
	Yes to binary	Agree with LOC	Extreme value to subj. question	Low extreme value to subj. question	High extreme value to subj. question	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.

Table 6. 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.

Supplemental Appendix A. Additional Figures

Figure A1. Survey error framework for surveys in LMICs

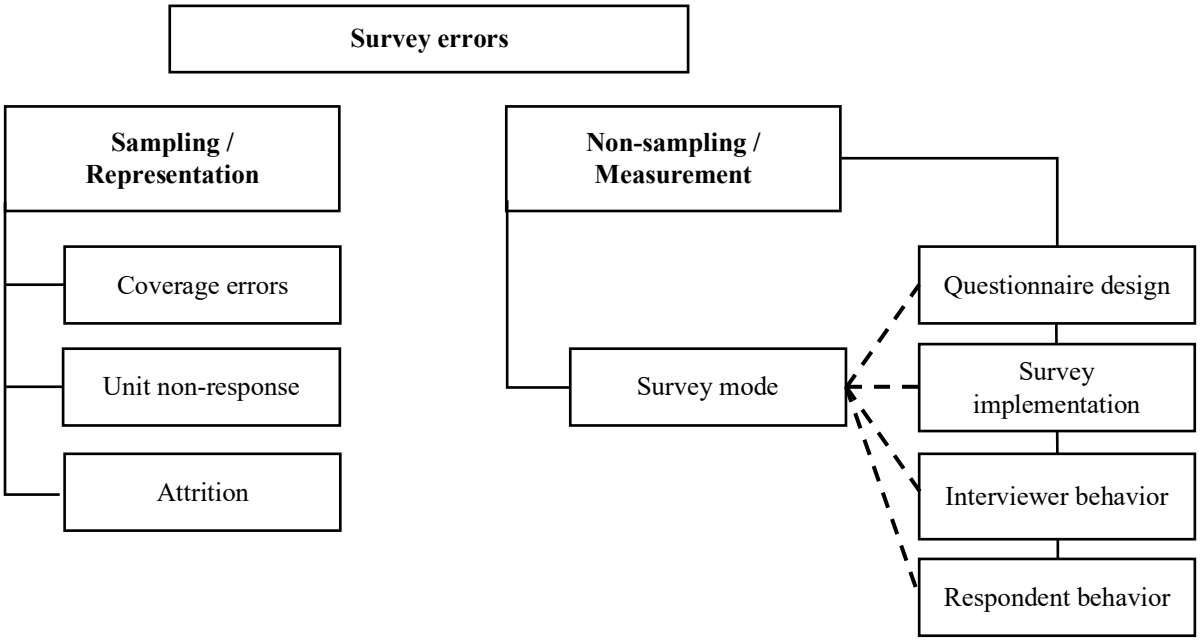


Figure A1. Experimental Design

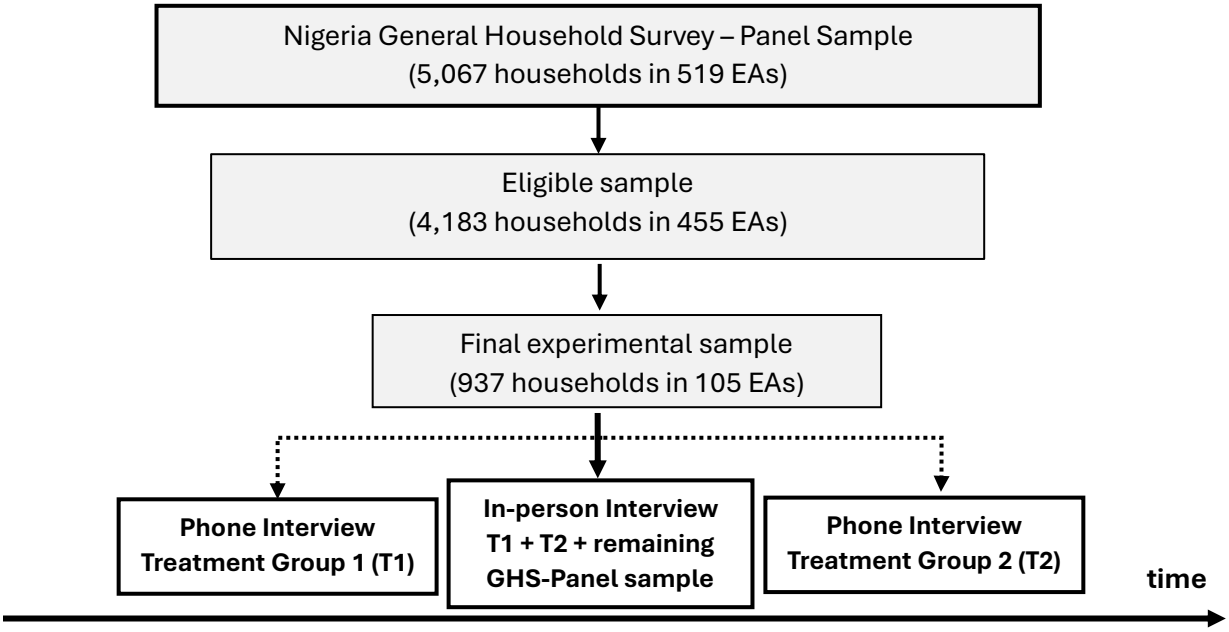


Figure A2. Dates of first-time interviews across survey modes (Panel A) and time between two interviews (Panel B)

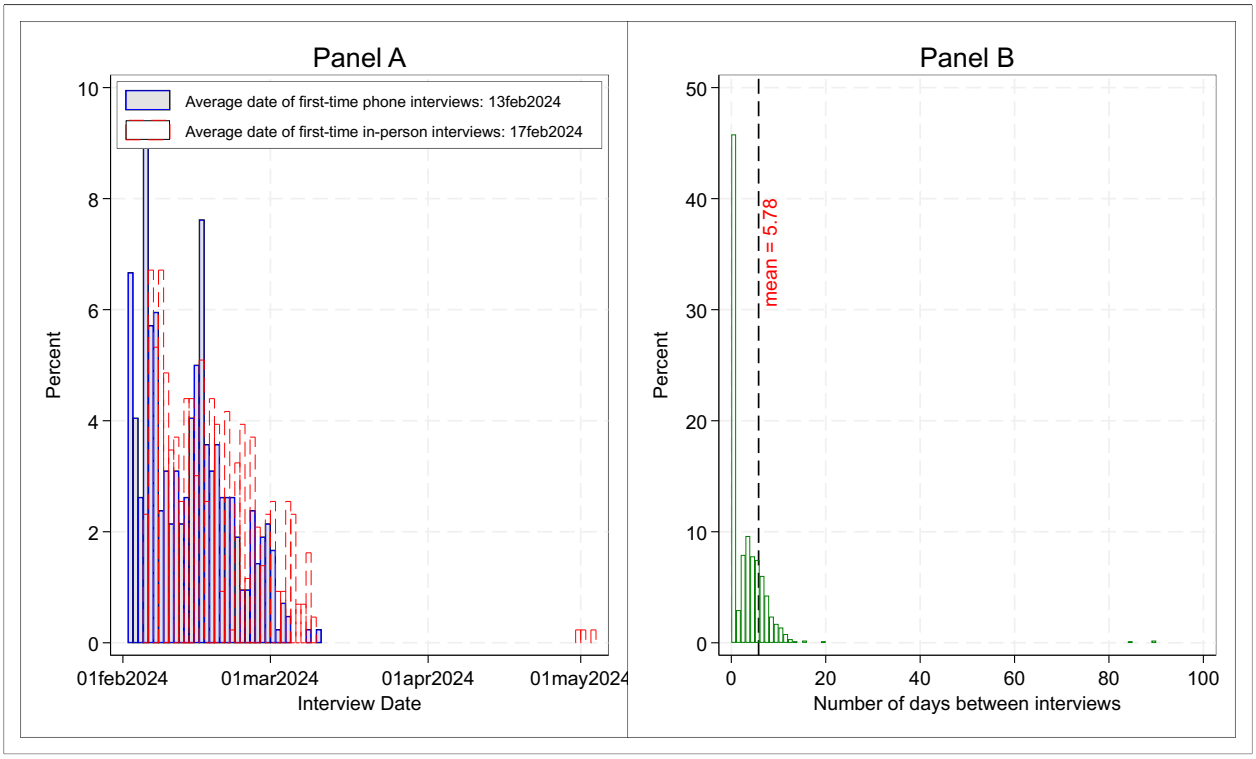


Figure A4. Distribution of mode effects estimates across the two designs

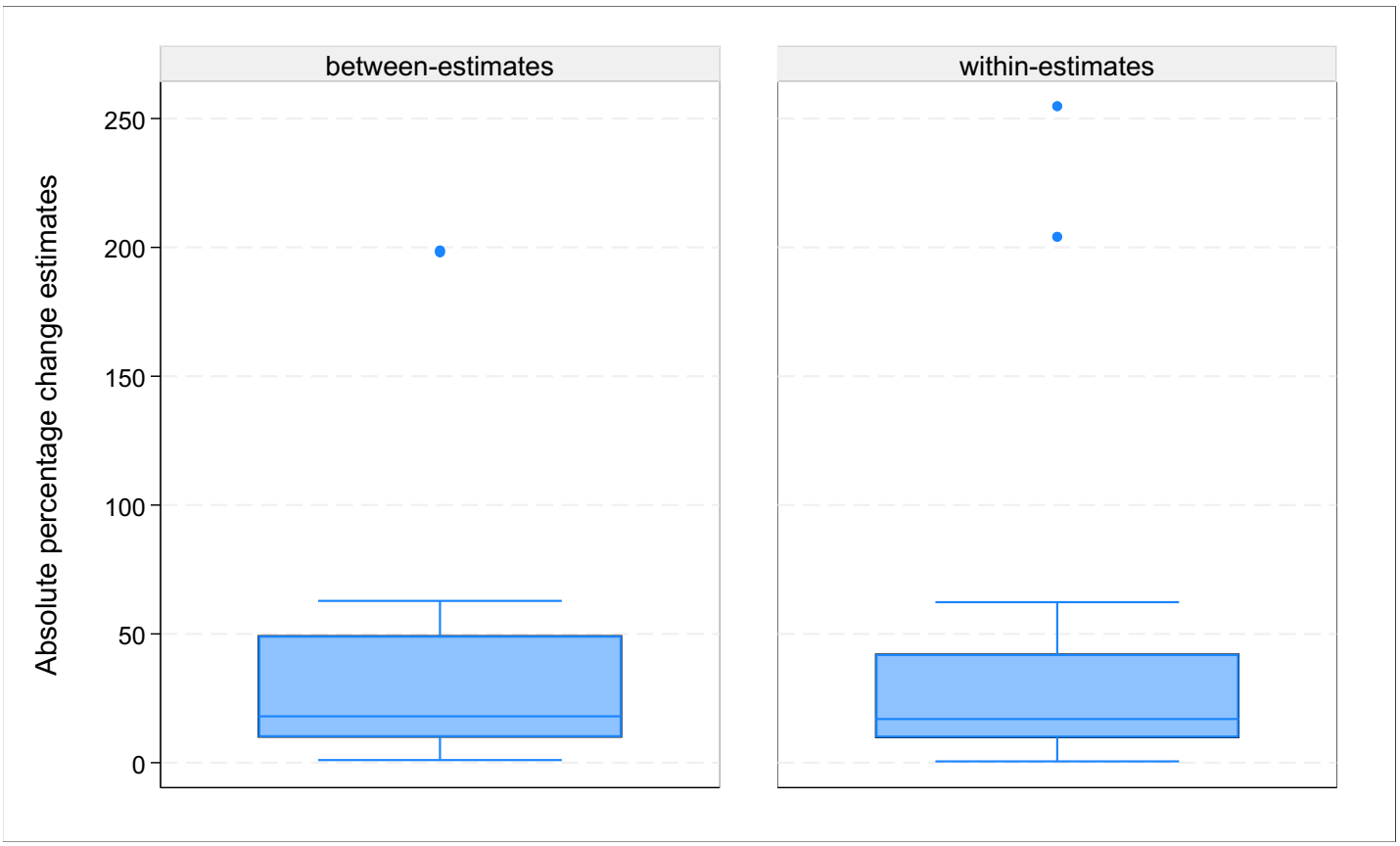
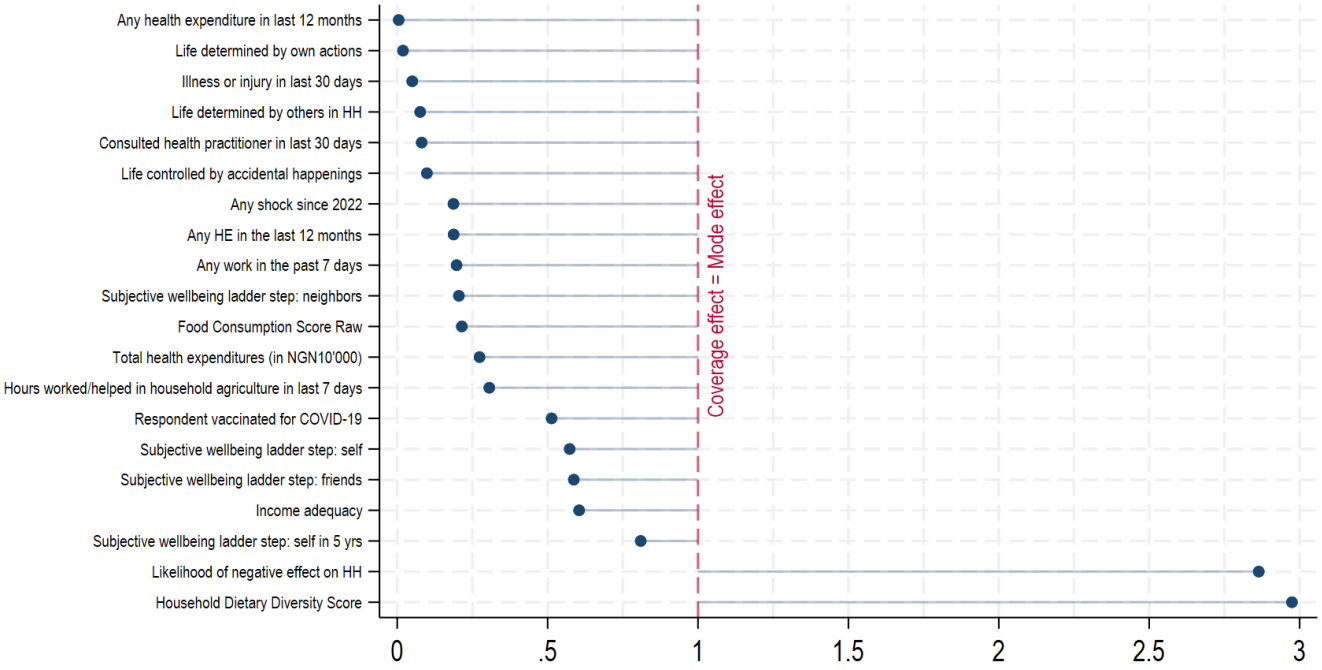


Figure A5. Coverage effect size relative to mode effect size



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.

Supplemental Appendix B. Additional Tables

Table B1. 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 B2. 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 B3. Temporal outcome stability

	Phone interview	Interview date	constant	N
Illness or injury in last 30 days	0.0743**	0.000973	-22.38	1668
	(0.0243)	(0.00106)	(24.83)	
Consulted health practitioner in last 30 days	0.189***	0.00203	-46.99	1668
	(0.0239)	(0.00106)	(24.75)	
Any health expenditure in last 12 months	0.482***	0.00220*	-51.33*	1430
	(0.0231)	(0.00105)	(24.53)	
Total health expenditures (in NGN10'000)	4.312***	0.0266	-620.9	256
	(0.798)	(0.0297)	(695.8)	
Respondent vaccinated for COVID-19	0.206***	-0.0000263	0.946	1668
	(0.0238)	(0.000995)	(23.31)	
Subjective wellbeing ladder step: self	-0.774***	-0.00907*	216.9*	1654
	(0.0996)	(0.00390)	(91.38)	
Subjective wellbeing ladder step: neighbors	-0.704***	-0.00579	140.1	1618
	(0.0971)	(0.00386)	(90.49)	
Subjective wellbeing ladder step: friends	-0.562***	-0.00642	155.0	1620
	(0.101)	(0.00427)	(99.97)	
Subjective wellbeing ladder step: self in 5 yrs	0.590***	-0.00866	209.7	1638
	(0.112)	(0.00470)	(110.2)	
Income adequacy	-0.222***	-0.00274	66.45	1660
	(0.0434)	(0.00179)	(42.03)	
Life controlled by accidental happenings	0.575***	-0.00342*	81.87*	1660
	(0.0415)	(0.00160)	(37.47)	
Life determined by own actions	0.181***	0.00166	-36.44	1660
	(0.0341)	(0.00134)	(31.46)	
Life determined by others in HH	0.665***	-0.000457	12.27	1660
	(0.0417)	(0.00183)	(42.87)	
Household Dietary Diversity Score	0.118	-0.00690	169.5	1356
	(0.0894)	(0.00401)	(93.86)	
Food Consumption Score Raw	9.602***	-0.0218	561.3	1356
	(0.959)	(0.0399)	(935.2)	
Any work in the past 7 days	0.0945***	0.00121	-27.65	2474
	(0.0165)	(0.000767)	(17.97)	
Hours worked/helped in household agriculture in last 7 days	-2.213*	0.0767	-1774.1	912
	(0.898)	(0.0424)	(993.1)	
Any HE in the last 12 months	0.279***	-0.00169	40.15	1442
	(0.0230)	(0.00110)	(25.70)	
Likelihood of negative effect on HH	0.0169	0.00358	-80.66	1530
	(0.0555)	(0.00274)	(64.25)	
Any shock since 2022	0.139***	0.00127	-28.87	1494
	(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. The sample includes identical respondents interviewed both by phone and in person. The two interviews are conducted on average 5.78 days apart. Standard errors are clustered at the respondent level. *** p<0.01 ** p<0.05 * p<0.1.

Table B4. 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 B5. 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 B6. 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 B7. 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

Table B8. 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 B9. 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 B10. Overlap in selected beneficiaries between different data collection mode for different simulated interventions

	Targeting bottom/top p%					
	10%	20%	30%	40%	50%	N
Food consumption score	0.05	0.14	0.22	0.28	0.37	678
Out-of-pocket health expenditure	0.08	0.14	0.32	0.41	0.53	707
Deprivation index	0.15	0.24	0.31	0.35	0.41	592
Expected overlap (benchmark)	**0.05**	**0.11**	**0.18**	**0.25**	**0.33**	

Note: The table shows the results of a targeting simulation in which households are ranked according to either food consumption score, out-of-pocket health expenditures, or a multidimensional deprivation index using either phone survey or in-person survey data. The worst-off p% are then selected as beneficiaries of a policy intervention. The values shown is the fraction of beneficiary households that are selected irrespective of survey mode (as a share of all households selected by at least one mode). That is, it shows the share of households selected in both data collection modes, conditional on a household being identified as a beneficiary using at least one of the modes. The final row provides the benchmark overlap we would expect to see by pure chance even if the ranking via one mode is completely uninformative for the ranking via the other. The deprivation index takes the following inputs: respondent is not in wage employment, respondent had illness or injury in last 30 days, respondent households has been affected by an adverse shock in last 2 years, self-reported income adequacy, expected likelihood of adverse future weather shock exposure, respondent subjective wellbeing. The sample comprises of identical respondents who gave an in-person and phone interview and had no missing data for the outcome of interest for either mode.

Table B11. 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.