Policy Research Working Paper 9636

Assessing Response Fatigue in Phone Surveys

Experimental Evidence on Dietary Diversity in Ethiopia

Kibrom A. Abay
Guush Berhane
John Hoddinott
Kibrom Tafere

WORLD BANK GROUP
Development Economics
Development Research Group
April 2021

Policy Research Working Paper 9636

Abstract

The COVID-19 pandemic has spurred interest in the use of remote data collection techniques, including phone surveys, in developing country contexts. This interest has sparked new methodological work focusing on the advantages and disadvantages of different forms of remote data collection, the use of incentives to increase response rates, and how to address sample representativeness. By contrast, attention given to associated response fatigue and its implications remains limited. This study designed and implemented an experiment that randomized the placement of a survey module on women's dietary diversity in the survey instrument. The study also examines potential differential vulnerabilities to fatigue across food groups and respondents. The findings show that delaying the timing of mothers' food consumption module by 15 minutes leads to 8–17 percent decrease in the dietary diversity score and a 28 percent decrease in the number of mothers who consumed a minimum of four dietary groups. This is driven by underreporting of infrequently consumed foods; the experimentally induced delay in the timing of mothers' food consumption module led to decreases of 40 and 11 percent in the reporting of consumption of animal source foods and fruits and vegetables, respectively. The results are robust to changes in model specification and pass falsification tests. Responses by older and less educated mothers and those from larger households are more vulnerable to measurement error due to fatigue.

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

Assessing Response Fatigue in Phone Surveys: Experimental Evidence on Dietary Diversity in Ethiopia

Kibrom A. Abay, Guush Berhane, John Hoddinott, Kibrom Tafere*

*Kibrom A. Abay: International Food Policy Research Institute; Guush Berhane: International Food Policy Research Institute; John Hoddinott: Cornell University and International Food Policy Research Institute; Kibrom Tafere: Development Research Group, The World Bank.

Key words: COVID-19, response fatigue, phone survey, dietary diversity, Ethiopia.

JEL Codes: I30, I38, O10, Q18

Acknowledgment: This study was supported by funding from the CGIAR Research Program on Policies, Institutions, and Markets (PIM), which is led by IFPRI, and Research Support Budget (RSB) from the Development Research Group of The World Bank. Berhane and Hoddinott thank the Bill and Melinda Gates Foundation for their funding for work on the 2019 survey and Berhane for funding in 2020/21 (Award No. OPP1162182). Abay acknowledges funding from the Partnership for Economic Policy (PEP), which is financed by the Department for International Development (DFID) of the United Kingdom (UK Aid) and the International Development Research Centre (IDRC) of Canada. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the institutions the authors are affiliated to. The authors are grateful to Abraha Weldegerima and Mehari Abay, who helped in collecting and cleaning the data.

1. Introduction

Outbreaks of pandemics and conflicts make monitoring welfare outcomes such as food security particularly important to respond more effectively to these crises. However, such events (e.g., COVID-19 and Ebola) create substantial obstacles to using traditional methods that employ face-to-face (FTF) interviews.[1] This, together with increased penetration of mobile phones and continued improvements in access to the internet, has spurred interest in remote data collection using tools such as web sites, online polls, text messages, and phone surveys. For example, the World Bank has transformed the traditional multi-country and multi-round FTF Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) program into a high-frequency monthly phone survey following the outbreak of the COVID-19 pandemic.[2] Alongside this increased interest and use of surveys that do not rely on FTF interviews has been an outpouring of methodological guides and experimentation on aspects of remote data collection (e.g., Dillon, 2012; Dabalen et al., 2016; Lau et al., 2019; Angrist et al., 2020).[3] Focal points of discussion in this literature are respondent access to forms of remote interviewing techniques (for example, non-random differences in ownership of mobile phones or internet connectivity), levels and differences in response rates across platforms such as computer-assisted telephone interviewing (CATI), interactive voice response (IVR), and short message service (SMS), the value of providing incentives to respondents to participate in and complete these interviews, and the implications of all these considerations for sample representativeness.

To the best of our knowledge, there has not been much discussion about the impact of the length of time required from respondents to answer questions and the response fatigue associated with these remote methods, beyond noting concerns that long surveys might generate higher rates of non-response. Dabalen et al. (2016) are typical in this regard, noting that "As a rule of thumb, interviews during mobile phone surveys are each restricted to about 15–30 minutes. The suggested duration of an interview is based on common practice in recent household surveys relying on mobile phones. There is no other supporting evidence" (Dabalen et al., 2016, p. 59). It is not hard to surmise why the duration of remote methods such as CATI might affect the accuracy of

[1] Currently, most studies aiming to monitor the impact of the COVID-19 pandemic on various welfare outcomes in low-income countries rely on phone surveys. Similarly, phone surveys were effectively used to monitor the Ebola crisis in West Africa (e.g., Etang and Himelein, 2020).

[2] https://www.worldbank.org/en/programs/lsms/brief/lsms-launches-high-frequency-phone-surveys-on-covid-19

[3] See also, more recent living blog posts at J-PAL (Kooper and Sautmann, 2020) and 3ie (Mani and Barooah, 2020).

responses being provided. In FTF interviews, enumerators can use visual cues to see whether respondents are beginning to tire. They can suggest taking a short break, perhaps getting a drink or a brief stretch to allow the respondent (and the enumerator) to re-fresh. Such cues are not available when interviews are conducted remotely. Further, when using remote methods such as CATI, it is harder to ensure that the interview is taking place in an environment where the respondent is not subject to distractions (such as children or elders calling for help), the likelihood of which might increase with longer interviews. In localities where time charges for mobile use are high relative to incomes, respondents might feel that they need to rush their responses as interviews drag on, even if the call time is being paid for. Yet while work on FTF interviews has shown that seemingly small changes in survey methods and designs (such as the placement of questions) can distort stylized descriptive statistics as well as econometric analysis and associated conclusions (e.g., Kilic and Sohnesen, 2019; De Weerdt et al., 2020),[4] this issue appears to have received less attention even when the pandemic has forced many researchers to shift to remote data collection methods.

Understanding whether and how fatigue in phone surveys affects the responses received is the focus of our experimental study. We examine responses to a module aimed at characterizing the dietary diversity of mothers. We use this measure for several reasons. Dietary diversity is a common indicator of quality of diets, which has been validated to predict nutritional outcomes (e.g., Ruel 2003; Steyn et al., 2006; Kennedy et al., 2007; Moursi et al., 2008). Dietary diversity is usually operationalized using a simple count of foods or food groups consumed over a given reference period, mostly over the last 24 hours which makes it well suited for CATI. Because of their simplicity, dietary diversity indicators are also sometimes used to assess overall household food security (e.g., Hoddinott and Yohannes, 2002). Dietary diversity indicators are also commonly employed to examine nutritional transition, food system transformations and the effects of shocks on household food security.[5]

To quantify the overall and differential implications of fatigue, we conduct an experiment by randomizing the order of this important module on mothers' dietary diversity. We do this in the

[4] Kilic and Sohnesen's (2019) work show that a small change in questionnaire design and survey length yields important differences in poverty predictions. Other examples on the implication of survey designs include Caeyers et al. (2012), Beegle et al. (2012), Gibson et al. (2015), and Ameye et al. (2021).

[5] Discussions on nutrition transition and associated measures and indicators can be found at Popkin (2003) and Pingali and Sunder (2017).

context of a longitudinal study that tracks the impact of the COVID-19 pandemic on households' food security in rural Ethiopia in which respondents were mothers of a young child. This study includes FTF interviews fielded in March and August 2019, subsequent CATI surveys conducted in June 2020 and December 2020. These rich data allow us to control for a wide array of confounding factors including household and enumerator fixed effects, other temporal factors while also allowing us to assess pre-intervention trends. In the first (baseline) phone survey, we follow the usual approach and kept mothers' and children's dietary diversity modules around the middle of the survey instrument for all respondents. In the second phone survey, for randomly selected respondents, we moved the mothers' dietary diversity module upfront, creating an approximately 15 minute gap between the treatment group (who answer these questions early in the interview) and the control group (who answer these questions at the same time as in the baseline). This allows us to assess whether responding to mothers' food consumption questions later in the interview by the control group leads to differences in the reporting of dietary diversity due to fatigue. We kept the placement of a second module on dietary diversity, on children, unchanged thus allowing us to use that measure for a falsification test.

We also consider several extensions. First, we consider whether the reporting of frequently consumed food groups is less prone to measurement error due to fatigue than infrequently purchased food groups as the latter are vulnerable to other sources of measurement error, including recall bias and telescoping (e.g., Beegle et al., 2012; Gibson et al., 2015; Abate et al., 2020). Second, we consider whether different types of respondents exhibit varying vulnerability to response fatigue. This is particularly important in the context of phone surveys, which require substantial cognitive resources to instantly recall recent consumption episodes. For instance, recalling diets consumed in the last 24 hours entails varying cognitive burden on respondents with varying age and education. Similarly, the breadth of consumption items consumed often vary across households, which imply varying cognitive burden to recall consumption in the last 24 hours. If response fatigue is randomly distributed across the distribution of the true measure of interest and uncorrelated with relevant observable underlying factors, such errors introduce the usually predictable biases from classical measurement error (e.g., Bound et al., 2001). However, if response fatigue exhibits a systematic pattern across respondents and associated observables, the inferential biases introduced by misreporting conform with non-classical measurement error (e.g., Bound et al., 2001; Gibson et al., 2015; Abay et al., 2019). As characterizing the nature of response

fatigue is crucial to inform the inferential consequences of response fatigue, we assess whether the placement of the women's food consumption module differs by maternal and household characteristics.

We find significant impacts of fatigue: respondents who received the dietary diversity module earlier in the survey report significantly higher dietary diversity score. Delaying the arrival of the dietary diversity module by about 15 minutes leads to 8-17 percent underestimation in dietary diversity score. This is a striking result given the relatively short difference in the timing of the food consumption module between the treatment and control groups. The impact of fatigue appears to be more pronounced for food groups that are infrequently consumed, consistent with evolving evidence that these food groups are prone to other sources of measurement error. We also find significant and intuitive heterogeneity in respondents' vulnerability to fatigue. For example, older women and less educated women are prone to fatigue than younger and more educated ones. Moving the module on diet diversity had a larger effect on responses provided by mothers in larger households, consistent with the idea that women in larger households might face more distractions when answering phone surveys or are simply more vulnerable to the effects of fatigue. Relatedly, this systematic underreporting of dietary diversity for larger households may contribute to the inverse relationship between food demand and household size documented in developing countries (Lanjouw and Ravallion, 1995; Deaton and Paxson, 1998; Gibson, 2002; Gibson and Kim, 2007).

Although our findings come from a specific rural women sample and phone surveys in rural Ethiopia, the implication and relevance of our findings are likely to extend to FTF surveys and other settings involving phone surveys. While the impact of fatigue in choice experiments and diaries is well-documented (e.g., Bradley and Daly, 1994; Savage and Waldman, 2008; Silberstein and Scott, 2011; Beegle et al., 2012; Hess et al., 2012; Schündeln, 2018; Battistin et al., 2020), the effect of fatigue in the usual multi-module and long rural household survey remains understudied. Two recent exceptions include Laajaj and Macours (2020) and Ambler et al. (2021) which formally test for the effect of fatigue on measuring skills and number of rural activities, respectively.[6] However, we are not aware of any study examining the impact and implication of response fatigue

[6] Laajaj and Macours (2020) find no evidence of response fatigue, while Ambler et al. (2021) detect significant fatigue effects in households' reporting of productive rural activities. Ambler et al. (2021) show that household members listed one position ahead of others report (are associated with) 2.2% higher number of productive activities.

in phone surveys. Thus, this paper provides the first empirical evidence of substantial response fatigue in phone surveys in a developing country context.

The rest of this paper proceeds as follows: The next section describes the data and survey design. Section 3 reports summary statistics and descriptive results. Section 4 presents our empirical strategy. Section 5 discusses our main results and heterogeneity analysis. Section 6 concludes.

2. Data and Experimental Design

We use data from two rounds of phone surveys (CATI) collected in June 2020 and December 2020. These build on previous FTF surveys conducted in March and August 2019 to understand the impact of the nutrition sensitive components of Ethiopia's Productive Safety Nets Programme (PSNP) on the nutritional outcomes of mothers and children in the four regions (Tigray, Amhara, Oromia, SNNP) where it was operational.

A stratified random sampling procedure was used to select sample households in the 2019 surveys. From a list of woredas (districts) in each region where the nutrition sensitive PSNP was operational, 22 woredas were randomly selected, using probability proportional to size (of population and program coverage). Within each woreda, three rural kebeles (sub-districts) were randomly selected and, within these, one enumeration area (EA) was randomly chosen. For each selected EA, a household list was constructed using the following selection criteria: having a child aged zero to 23 months; and being PSNP beneficiary or, if not a PSNP beneficiary, considered poor based on a subjective ranking scheme applied to both PSNP and non-PSNP households. From this list of eligible households, five PSNP and five non-PSNP households were randomly selected for interview. A total of 2,640 households from 264 EAs and 88 woredas were surveyed in March 2019, of which 2,551 were re-interviewed in August 2019 (Berhane et al., 2020).

In the first phone (CATI) survey conducted in June 2020, households were contacted using phone numbers provided in the August 2019 survey. About 54 percent of the August 2019 sample had access to a phone. We were also able to locate additional households who had acquired mobile phones between the August 2019 and June 2020 surveys. This yielded a list of 1,497 households (about 59 percent) of the 2,551 who had been surveyed FTF in August 2019 (Abay et al., 2020). Prior to the commencement of our second phone survey, war broke out in the Tigray region; because of the near total blackout in telecommunications, we were unable to contact the 378

households from that region. Consequently, in December 2020, we interviewed 1,109 out of the 1,119 households from Amhara, Oromia and SNNPR who also participated in the June 2020 round. In all rounds, the primary respondent was the mother or caregiver of the young child.

The long questionnaire administered in the FTF surveys was considerably shortened when we shifted to phone surveys. We retained only core modules that focused on household food security, maternal food consumption, child feeding and practices, access to nutrition and health services. These modules were shortened to minimize respondent burden while preserving the framing and comparability of questions across rounds. Questions were simplified to fit interview protocols using phones. Doing so reduced the time taken to administer the survey from the two+ hours needed for the FTF survey to a median duration of 26 minutes. However, feedback from our enumerators indicated that respondents were tiring towards the end of the phone survey, particularly when asked about long lists of items such as the 17 yes/no questions about food groups mothers have consumed the previous day.

We modified our December 2020 survey instrument to assess the effect of response fatigue on a measure of women's food security, dietary diversity. Specifically, we introduced a randomized assignment of respondents to one of two questionnaire types that differed only in the placement of the module on women's diet. About 50 percent of respondents were randomly assigned to the treatment group, moving the instrument on women's dietary diversity to a position approximately 15 minutes earlier in the interview. Mothers assigned to the treatment group were asked to respond to (in a yes/no format) a list of the food items they had consumed in the last 24 hours, at the start of the interview, following some background questions. Mothers assigned to the control group were asked the same set of questions in the middle of the interview. The placement of these questions for the control group was the same as it was in the June 2020 survey. To maintain balance by PSNP status, an important source of heterogeneity in our sample, and within administrative regions, randomization was stratified by PSNP beneficiary status and region.

3. Summary Statistics and Descriptive Results

The August 2019 FTF survey contains detailed background information about the sample households who were later interviewed in the phone (CATI) surveys. These background characteristics of respondents serve two important purposes, mainly to (1) assess the validity and balancing of the randomization and (2) facilitate the identification of differential vulnerability to

fatigue among different groups of respondents. In the CATI surveys, we also collected detailed information about the phone calls, including interview date and time, number of call attempts made, interviewer identifiers, interview duration and other features of the interview. These temporal features enable us to capture additional contexts of the interview. For instance, as noted by Di Maio and Fiala (2020), controlling for interviewer fixed effects can help address interviewer fatigue which could interact with response fatigue in ways that can affect the estimates on the outcome of interest.

Table 1 presents baseline balance tests between treatment and control groups. On average, except for mothers' age, there is no statistical difference between the two groups. We also conduct joint significance tests by regressing the treatment dummy on baseline characteristics listed in Table 1. The joint significance F-test statistic is 0.83 (prob>F=0.72), indicating that we cannot reject the null hypothesis that all coefficients associated with these regressors are jointly zero. The mean difference tests shown in Table 1 and the joint significance tests confirm that our randomized design is balanced on observables in the treatment and control groups.

4. Empirical Strategy

To identify whether and to what extent reported dietary diversity outcomes are affected by response fatigue in phone surveys, we estimate a respondent fixed effects (FE) model. Since the ordering of modules was randomly assigned to respondents, identification is relatively straightforward. We estimate the following equation:

$$Y_{mt} = \alpha_h + \beta_0 Round_t + \beta_1 Treatment_{mt} + \gamma X_{mt} + \varepsilon_{mt} \quad (1)$$

where Y_{mt} stands for the dietary diversity outcomes of mother *m* in round *t*. *Round* is survey round indicator that takes value 1 for the December 2020 survey and 0 for the June 2020 round. *Treatment* is a dummy variable equal to 1 for mothers' receiving the dietary diversity module early in the interview and 0 for those receiving the same module later. This variable takes the value 0 for all respondents in the baseline round. *X* is a vector of time variant observable mother characteristics and interview features, which include mothers' fasting status as well as temporal features of the survey that may affect dietary diversity outcomes. These temporal factors include interview day and time, interview duration, number of call attempts made and enumerator fixed effects. The

coefficient associated with the treatment indicator in equation (1), β_1, captures the impact of being asked questions about diet earlier in the interview. We cluster our standard errors at the EA level, the lowest sampling unit.

5. Results and Discussion

5.1. Main results

Table 2 presents fixed effects regression results focusing on mothers' dietary diversity and minimum dietary diversity. Odd columns provide results from a parsimonious specification while even columns show results based on more saturated specifications controlling for respondent fixed effects, time varying respondent characteristics and temporal features of the survey. We focus our discussion on results obtained from regressions with the full set of controls, noting that there are no important differences between the results reported in columns (1) and (2) and columns (3) and (4). Column (2) shows that mothers who were asked the diet diversity module early in the interview (treatment group) report consumption of 0.25 more food groups compared to those who were asked the module later in the interview, and were likely to be more fatigued. This impact is large, especially given the delay in the food consumption module for the control group was short. At the mean diversity score, these results are equivalent to an 8.4 percent reduction in maternal diet diversity for respondents whose food consumption module was delayed by 15 minutes.

Columns 3 and 4 of Table 2 show results for minimum dietary diversity dummy of mothers (MDD-W). This is defined as the consumption of foods from five or more categories on the previous day (FAO and FH360, 2016; WHO, 2010). The randomized placement of questions has no effect on this outcome. That said, Figure 1 shows the distribution of the number of food groups consumed at baseline (June 2020) and after the experiment was implemented as part of the December 2020 follow-up. In both cases, the mass of the distribution lies to the left of the five food group cut-off for the MDD-W indicator. If, however, we were to consider a different dichotomous measure, consumption of four food groups or more (columns 5 and 6), we see that mothers in the treatment group are 8.1 percentage points more likely to report meeting this threshold. Asking questions about maternal diets later in the interview reduces the percentage of women meeting the four-food group minimum by 28 percent. In turn, this suggests that the impact of response fatigue on outcomes where continuous variables have been converted to dichotomous

outcomes will depend on both the magnitude of the impact of fatigue on responses and where the mass of the distribution of the underlying continuous variable lies relative to the threshold.

The impacts of response fatigue reported in Table 2 are based on a dietary diversity index constructed from a series of yes/no responses to questions on maternal consumption of 10 food groups. In low-income settings where diversity of foods consumed remains largely monotonous, response fatigue may entail differential impacts by food groups depending on how frequently a given food group is consumed. For example, response fatigue is expected to have greater impact on animal source foods (meat, milk, and egg) and fruits and vegetables as they are less frequently consumed in such contexts. With this in mind, we aggregated the 10 food groups into three categories: staples, beans, and nuts; animal source foods; and fruits and vegetables. We re-estimated equation (1) separately for these categories of food groups with the outcome variables being dummy variables equal to 1 if a mother reported to have consumed from a specific food category in the last 24 hours and 0 otherwise. We report these disaggregated results in Table 3. The results in columns 1, 3 and 5 are based on a parsimonious model that includes just treatment and survey round dummies, whereas columns 2, 4 and 6 include controls for mother characteristics and temporal features of the survey in addition to treatment and survey round indicators.

We find that treatment households are 8.6 percentage points more likely to report consuming animal sourced foods, and vegetable and fruits. Delaying the food consumption module by 15 minutes leads to a 40 percent decrease in the number of mothers who report consuming animal source food (mean = 22 percent) and 11 percent decrease in mothers that report consuming vegetables and fruits (mean = 76 percent). To the contrary, there is little impact of response fatigue on food groups that are monotonously consumed such as staples, beans, and nuts. These results are insensitive to inclusion of respondent, survey, and enumerator controls. One intuitive explanation for these patterns could relate to respondent incentives and responses to lengthy interviews. For instance, if each additional question involving "yes" response brings (or is perceived to bring) follow-up questions this may encourage time-constrained and fatigued respondents to respond "no" and, likely to be more so, for less frequent items (De Weerdt et al., 2020).

5.2. Heterogenous impacts by respondent and household characteristics

Different respondents may exhibit varying vulnerability to response fatigue for several reasons. First, recalling diets consumed in the last 24 hours entails varying cognitive burden on respondents with different age and education profiles. For example, older and less educated mothers may be more vulnerable to response fatigue than younger and more educated mothers, because the latter are likely to be mobile-savvy and familiar with long-running telephone conversations. Second, the breadth of consumption items that households usually consume vary across households, which imply varying cognitive burden to recall consumption in the last 24 hours. For instance, maternal diets for poorer households may be monotonous, which require less cognitive burden to recall. Third, mothers with more familial responsibilities may be more vulnerable to fatigue because larger families may involve more complex intrahousehold diet distribution as well as greater demands that may distract attention from a more accurate recollection of events. To uncover potential differential vulnerability to response fatigue, we re-estimated equation (1), disaggregating by respondent and household characteristics. We report results using our full set of controls in Table 4; results excluding these covariates are similar and are available on request.

The first two columns of Table 4 provide results disaggregated by mother's median age (29 years). Older mothers, but not younger mothers, reported 0.5 more food groups when asked about these food groups earlier in the interview, which is a 17 percent decrease for women whose food consumption module was delayed by approximately 15 minutes (and thus were more fatigued). Next, we split the sample by median level of education (3 years of schooling) (columns 3 and 4 of Table 4). Moving the food consumption module closer to the beginning of the interview increases the number of reported food groups by mothers with below median education level by 0.21 groups, which is equivalent to a 7.5 percent decrease for the group that was administered mothers' food consumption module later.

Results from splitting the sample by median household size (5 members) are reported in columns (5) and (6) of Table 4. Moving the module earlier increased the number of food groups reported by women in larger households by 0.46 groups, a 15 percent fall for respondents who answered the diet diversity questions later, but had no effect on reporting by women in smaller households. Note that this result is similar to work using in-person surveys which often report negative correlations between household size and food expenditures (Lanjouw and Ravallion; 1995; Deaton and Paxson, 1998; Gibson, 2002; Gibson and Kim, 2007). Finally, we used principal component analysis to construct a household wealth score based on pre-pandemic holdings of

durable goods, then disaggregated households into those below and above the median wealth score (columns 7 and 8). Moving the food consumption module earlier in the questionnaire had no impact on reporting by poor households but reduced reported consumption by women in less poor households whose women's module was delayed by 0.37 food groups, a 12 percent reduction. These differential impacts are consistent with the fact that very poor households in rural Ethiopia consume monotonous diets of staples and pulses and as seen in Table 3, reporting of these is less susceptible to question placement in the survey instrument. As also seen in Table 3, more diversified diet are more likely to be reported when the module appears earlier in the interview; this is consistent with the impact of module placement on reported women's dietary diversity in less-poor households.

5.3 Additional robustness checks

In addition to assessing whether our results were affected by the inclusion of control variables, we conducted two additional robustness checks. The first is a falsification test. Our survey instrument also included a module on child diet diversity. The randomization of the placement of the questions on mothers' food consumption had no effect on the placement of children's food consumption module; this was kept in the same place in the June and December survey rounds. Thus, our treatment should not affect reported dietary diversity outcomes for children. Based on standard practice (FAO and FH360, 2016; WHO, 2010), we construct a continuous dietary diversity score and indicator variable for minimum dietary diversity score of children, a variable that assumes a value of 1 if the number of food groups consumed in the 24 hours preceding the interview was equal to or greater than 4, 0 otherwise. Estimating equation (1) using these outcome variables for children, we find that the randomized placement of mothers' food consumption module had no effect on reported dietary diversity of children (Table 5). These results are also robust to inclusion of controls for respondent characteristics, temporal features of the survey and enumerator fixed effects.

Second, we replaced the June 2020 baseline phone survey with the FTF survey conducted in August 2019, the latter was collected before the start of the COVID-19 pandemic. Results reported in Supplementary Appendix Table A1 show that placement of mothers' food consumption module earlier in the interview increased reported dietary diversity. Table A2 shows that this is

driven by underreporting of infrequently consumed food items, mainly animal source foods, while Table A3 replicates our falsification test, showing that reported dietary diversity of children was unaffected.

6. Concluding Remarks

The COVID-19 pandemic has spurred interest in the use of remote data collection techniques, including phone surveys (CATI), in developing country contexts. This interest has sparked new methodological work focusing on the advantages and disadvantages of different forms of remote data collection, the use of incentives to increase response rates and how to address sample representativeness. By contrast, to the best of our knowledge, attention given to associated response fatigue and its implications is limited.

We designed and implemented an experiment that randomized the placement of a survey module on women's dietary diversity. The treated group was randomly assigned to a survey instrument where the mothers' food consumption module was placed at the beginning of the interview, approximately 15 minutes earlier than those assigned to the control group. The food consumption module is well-suited to do this experiment as it requires respondents to recall food groups they consumed in the 24 hours preceding the interview, which entails considerable cognitive burden on them. As a result, responses to such burdensome modules coming later in the interview (after the respondent is cognitively taxed) are prone to measurement errors due to response fatigue.

We find that delaying the timing of mothers' food consumption module by 15 minutes leads to 8-17 percent decrease in the dietary diversity score and a 28 percent decrease in the number of mothers who meet the minimum dietary diversity (defined at 4 food groups or higher). This result is mainly driven by underreporting of infrequent food groups, including animal source food and fruits and vegetables. A 15-minute delay in the timing of mothers' food consumption module leads to 40 percent and 11 percent decrease in the number of mothers who report consumption of animal source foods, and fruits and vegetables, respectively. By contrast, we find no impact on the consumption of more frequently consumed foods such as staples, beans, and nuts. To probe whether these results are driven by our randomized variation in the placement of mothers' consumption module, we conducted a series of falsification tests using children's dietary diversity

outcomes, whose relevant module was unaffected by the randomization. We do not find any statistically significant impact on children's dietary diversity outcomes.

We also conducted a range of heterogeneity tests to investigate potential differential vulnerability to fatigue using observable information on mothers' age and education, household size, and household wealth. We find that responses by older mothers (above median age) and mothers with lower level of schooling (below the median years of schooling) are more vulnerable to measurement error due to fatigue. Mothers from larger households (above median household size) are also more prone to response fatigue. As we noted in the introduction, if response fatigue is randomly distributed across the distribution of the true measure of interest and uncorrelated with relevant observable underlying factors, such errors introduce the usually predictable biases from classical measurement error (e.g., Bound et al., 2001). But if response fatigue exhibits a systematic pattern across respondent characteristics and associated observables, the inferential biases introduced by misreporting conform with non-classical measurement error (e.g., Bound et al., 2001; Gibson et al., 2015; Abay et al., 2019).

We end with some remarks on the implications of these results. First, these results suggest that comparisons of descriptive statistics across studies and countries on key welfare metrics such as food security and dietary quality may be confounded by factors as simple as the placement of these modules in a given survey instrument. Second, variations in vulnerability to response fatigue by respondent characteristics suggest that some of the errors in phone survey-based analysis cannot be treated as classical measurement error. These non-classical errors are likely to render inferential biases on relationships and impacts involving these fatigue-prone phone survey data such as underestimating income elasticities for dietary quality. Third, our findings also highlight the important trade-offs between volume of information collected (length of surveys) and ensuring the quality of data that needs to be taken into account when designing phone surveys. They suggest that questions that impose significant cognitive burden or those that assume strategic importance to specific research agenda should be asked towards the start of the interview to reduce potential biases due to fatigue. Although our findings come from a specific phone survey among rural women in Ethiopia, the implication and relevance of our findings are likely to extend to in-person surveys and other outcomes.

References

Abate, G.T., De Brauw, A., Gibson, J., Hirvonen, K., Wolle, A., 2020. Telescoping Causes Overstatement in Recalled Food Consumption: Evidence from a Survey Experiment in Ethiopia. IFPRI Discussion Paper 01976.

Abay, K., Abate, G., Barrett, C., Bernard, T. 2019. Correlated non-classical measurement errors, 'second best' policy inference, and the inverse size-productivity relationship in agriculture. *Journal of Development Economics* 139(1): 171-184.

Abay, K.A., Berhane, G., Hoddinott, J. Tafere, K. 2020. COVID-19 and Food Security in Ethiopia: Do Social Protection Programs Protect? Policy Research Working Paper No. 9475. World Bank, Washington, DC.

Ambler, K., Herskowitz, S., Maredia, M., 2020. Are We Done Yet? Response Fatigue and Rural Livelihoods. IFPRI Discussion Paper 01980.

Ameye, H., De Weerdt, J., Gibson, J., 2021. Measuring macro-and micronutrient intake in multi-purpose surveys: evidence from a survey experiment in Tanzania. *Food Policy*, Forthcoming.

Angrist, N., Bergman, P., Evans., D.K., Hares, S., Jukes, M.C.H., Letsomo, M., 2020. Practical lessons for phone-based assessments of learning. *BMJ Global Health* 2020;5:e003030. doi:10.1136/bmjgh-2020-003030.

Arimond, M., M.T. Ruel., 2004. Dietary diversity is associated with child nutritional status: evidence from 11 demographic and health surveys. *The Journal of Nutrition* 134 (10): 2579-2585.

Battistin, E., De Nadai, M., Krishnan, N., 2020. The insights and illusions of consumption measurement: evidence from a large-scale randomization. Policy Research Working Paper 9255. World Bank, Washington DC.

Beegle, K., De Weerdt, J., Friedman, J., Gibson, J., 2012. Methods of household consumption measurement through surveys: experimental results from Tanzania. *Journal of Development Economics*, 99(1), 3–18.

Berhane, G., Golan, J., Hirvonen, K., Hoddinott, J., Kim, S., Taffesse, A.S., Abay, K. Assefa, T.W., Habte, Y., Abay, M.H., Koru, B., Tadesse, F., Tesfaye, H., Wolle, A., and F. Yimer, 2020. Evaluation of the nutrition-sensitive features of the fourth phase of Ethiopia's Productive Safety Net Program. ESSP Working Paper 140. Washington, DC: International Food Policy Research Institute (IFPRI). https://doi.org/10.2499/p15738coll2.133685.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5. Elsevier, Amsterdam, pp. 3705–3843.

Bradley, M., Daly, A., 1994. Use of the logit scaling approach to test for rank order and fatigue effects in stated preference data. *Transportation* 21(1), 167–184.

Caeyers, B., De Weerdt, J., Chalmers, N., 2012. Improving consumption measurement and other survey data through CAPI: evidence from a randomized experiment. *Journal of Development Economics,* 98(1):19–33.

Dabalen, A., Etang, A., Hoogeveen, J., Mushi, E., Schipper, Y., von Engelhardt, J., 2016. Mobile Phone Panel Surveys in Developing Countries: A Practical Guide for Microdata Collection. Directions in Development. Washington, DC: World Bank. doi:10.1596/978-1-4648-0904-0.

Deaton, A., Paxson, C., 1998. Economies of Scale, Household Size, and the Demand for Food. *Journal of Political Economy* 106(5):897–930.

De Weerdt, J., Gibson, J., Beegle, K., 2020. What can we learn from experimenting with survey methods? *Annual Review of Resource Economics, 12*(1), https://doi.org/10.1146/annurev-resource-103019-105958.

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

Di Maio, M., Fiala, F., 2020. Be Wary of Those Who Ask: A Randomized Experiment on the Size and Determinants of the Enumerator Effect. *The World Bank Economic Review* 34, 654–669.

Etang, A., Himelein, K., 2020. Monitoring the Ebola crisis using mobile phone surveys. In J. Hoogeveen & U. Pape (Eds.), Data collection in fragile states. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030- 25120-8_2.

FAO (Food and Agriculture Organization) and FHI 360. 2016. Minimum Dietary Diversity for Women: A guide to measurement. Rome: Food and Agriculture Organization (FAO) of the United Nations and FHI 360.

Gibson, J., 2002. Why Does the Engel Method Work? Food Demand, Economies of Size and Household Survey Methods. *Oxford Bulletin of Economics and Statistics* 64(4):341–60.

Gibson, J., Kim, B., 2007. Measurement error in recall surveys and the relationship between household size and food demand. *American Journal of Agricultural Economics* 89, 473–489.

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

Hess, S., Hensher, D.A., Daly, A., 2012. Not bored yet–revisiting respondent fatigue in stated choice experiments. *Transportation Research Part A: Policy and Practice* 46, 626–644.

Hoddinott, J., Yohannes, Y., 2002. Dietary Diversity as a Food Security Indicator. IFPRI Discussion Paper 136: Washington, DC: International Food Policy Research Institute (IFPRI).

Kennedy, G.L., Pedro, M.R., Seghieri, C., Nantel, G., Brouwer, I., 2007. Dietary diversity score is a useful indicator of micronutrient intake in non-breast-feeding Filipino children. *The Journal of Nutrition* 137 (2): 472-477.

Kilic T, Sohnesen T., 2019. Same question but different answer: experimental evidence on questionnaire design's impact on poverty measured by proxies. *Review of Income and Wealth* 65(1):144-165.

Kopper S, Sautmann A., 2020. Best practices for conducting phone surveys. Available: https://www. povertyactionlab. org/ blog/ 3- 20- 20/best- practices- conducting- phone-surveys.

Lajaaj, R., Macours, K., 2019. Measuring skills in developing countries. *Journal of Human Resources.* forthcoming.

Lau, C. Q., Cronberg, A., Marks, L., Amaya, A., 2019. In search of the optimal mode for mobile phone surveys in developing countries: A comparison of IVR, SMS, and CATI in Nigeria. *Survey Research Methods*, *13*(3), 305-318. https://doi.org/10.18148/srm/2019.v13i3.7375

Lanjouw, P., Ravallion, M., 1995. Poverty and Household Size. *Economic Journal* 105, 1415–34.

Mani, S., Barooah., 2020. Phone surveys in developing countries need an abundance of caution. Posted April 9, 2020. https://www.3ieimpact.org/blogs/phone-surveys-developing-countries-need-abundance-caution

Moursi, M.M., Arimond, M., Dewey, K.W., Trèche, S., Ruel, M.T., Delpeuch, F., 2008. Dietary diversity is a good predictor of the micronutrient density of the diet of 6-to 23-month-old children in Madagascar. *The Journal of Nutrition* 138 (12): 2448-2453.

Pingali, P., Sunder, N., 2017. Transitioning toward nutrition-sensitive food systems in developing countries. *Annual Review of Resource Economics* 9:439–59.

Popkin, B.M., 2003. The nutrition transition in the developing world. *Development Policy Review* 21, 581–597.

Ruel, M.T., 2003. Operationalizing dietary diversity: a review of measurement issues and research priorities. *The Journal of Nutrition* 133 (11): 3911S-3926S.

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

Steyn, N., Nel, J., Nantel, G., Kennedy, G., Labadarios, G., 2006. Food variety and dietary diversity scores in children: are they good indicators of dietary adequacy? *Public Health Nutrition* 9 (5): 644-650.

Schündeln, M., 2018. Multiple visits and data quality in household surveys. *Oxford Bulletin of Economics and Statistics* 80(2): 380-405.

Silberstein, A. R., Scott, S., 2011. *Expenditure Diary Surveys and Their Associated Errors, in Measurement Errors in Surveys*. Wiley-Blackwell, 2011, chapter 16, pp. 303 {326.

WHO (World Health Organization), 2010. Indicators for assessing infant and young child feeding practices: Part 2 Measurement.

Table 1: Balance of baseline characteristics

	No obs	Mean Control	No obs	Mean Treatment	Mean difference
Male headed household (dummy)	555	0.933	554	0.931	0.002
Age of household head(dummy)	555	37.286	554	37.827	-0.54
Education of household head (years)	555	3.616	554	3.599	0.017
Age of the mother (years)	551	29.216	553	28.429	0.787**
Education of mother (years)	555	3.117	554	3.255	-0.137
Fasting mother (dummy)	555	0.139	554	0.125	0.014
Age of the child (months)	555	30.773	554	31.191	-0.418
Household size	555	5.782	554	5.679	0.103
Livestock assets (TLU)	555	3.303	554	3.42	-0.117
Corrugated iron roof (dummy)	555	0.551	554	0.554	-0.003
Access to electricity (dummy)	555	0.427	554	0.403	0.024
Farm size (ha)	555	0.9	554	0.96	-0.06
Poor housing condition (dummy)	555	0.193	554	0.184	0.009
Food gap (in months)	555	2.485	554	2.558	-0.073
Food insecure household	555	0.773	554	0.756	0.017
Mothers' dietary diversity (June 2020)	555	2.814	554	2.744	0.071
Mothers' minimum dietary diversity (June 2020)	555	0.25	554	0.231	0.019
Mother consumed staples (June 2020)	555	0.944	554	0.926	0.018
Mother consumed animal source food (June 2020)	555	0.339	554	0.303	0.035
Mother consumed vegetable fruits (June 2020)	555	0.683	554	0.67	0.013
Children's dietary diversity (June 2020 survey)	554	1.939	554	2.002	-0.063
Amhara region	555	0.339	554	0.338	0.001
Oromia region	555	0.342	554	0.341	0.001
SNNP region	555	0.319	554	0.321	-0.002

Notes: Most of these baseline characteristics are collected in the in-person survey in August 2019. We rely on the in-person survey for those detailed household and mother characteristics. Outcomes related to mothers' and children's diets are based on information collected in the June 2020 phone survey because it is our main baseline survey. Significance levels: $^{*}p < 0.10, ^{**}p < 0.05, ^{***}p < 0.01$.

Table 2: Impact of early placement on maternal diet diversity score, respondent fixed effects estimates

	(1) Diet diversity score	(2) Diet diversity score	(3) Minimum diet diversity dummy (five and above)	(4) Minimum diet diversity dummy (five and above)	(5) Minimum diet diversity dummy (four and above)	(6) Minimum diet diversity dummy (four and above)
Treatment: Early placement	0.229***	0.252***	0.022	0.025	0.072**	0.081**
	(0.083)	(0.083)	(0.023)	(0.023)	(0.035)	(0.034)
Round	-0.029	-0.127	-0.022	-0.030	-0.002	-0.053
	(0.065)	(0.080)	(0.017)	(0.020)	(0.027)	(0.033)
Controls	No	Yes	No	Yes	No	Yes
Interview day	No	Yes	No	Yes	No	Yes
Enumerator fixed effect	No	Yes	No	Yes	No	Yes
Mean of dependent variable	2.985	2.985	0.090	0.090	0.292	0.292
R-squared	0.01	0.08	0.00	0.06	0.01	0.06
No. observations	2,234	2,234	2234	2234	2,234	2,234

Notes: Controls include a dummy variable indicating whether the mother was fasting, duration of interview, time of interview, a dummy variable if interview was conducted in the afternoon, and the number of call attempts. Interview day and enumerator are day of the week and enumerator dummy variables, respectively. Standard errors clustered at the EA level in parentheses: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3: Fatigue effects on dietary diversity of mothers, by food groups, respondent fixed effects estimates

	(1) Staples, beans and nuts	(2) Staples, beans and nuts	(3) Animal source foods	(4) Animal source foods	(5) Vegetables and fruits	(6) Vegetables and fruits
Treatment: Early placement	-0.014	-0.019	0.081**	0.086***	0.092***	0.086***
	(0.012)	(0.012)	(0.032)	(0.031)	(0.030)	(0.029)
Round	0.029***	0.031***	-0.032	-0.041	-0.013	-0.053*
	(0.010)	(0.012)	(0.024)	(0.029)	(0.025)	(0.028)
Controls	No	Yes	No	Yes	No	Yes
Interview day	No	Yes	No	Yes	No	Yes
Enumerator fixed effect	No	Yes	No	Yes	No	Yes
Mean of dependent variable	0.981	0.981	0.216	0.216	0.755	0.755
R-squared	0.01	0.06	0.01	0.06	0.01	0.07
No. observations	2,234	2,234	2,234	2,234	2,234	2,234

Standard errors clustered at the EA level in parentheses: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Notes: See Table 2.

Table 4: Heterogeneous effects of treatment on mothers' diet diversity, respondent fixed effects estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Maternal age		Maternal education		Household size		Household wealth	
	Below median	Above median	Below median	Above median	Below median	Above median	Below median	Above median
Treatment: Early placement	0.060	0.514***	0.213*	0.134	0.065	0.462***	0.173	0.378***
	(0.115)	(0.135)	(0.113)	(0.123)	(0.119)	(0.127)	(0.106)	(0.132)
Round	-0.051	-0.242**	-0.194*	0.050	-0.010	-0.247**	-0.118	-0.183
	(0.109)	(0.110)	(0.103)	(0.121)	(0.117)	(0.104)	(0.112)	(0.124)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interview day	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Enumerator fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of dependent variable	2.979	2.991	2.832	3.167	2.945	3.023	2.786	3.194
R-squared	0.11	0.10	0.07	0.13	0.09	0.11	0.08	0.11
No. observations	1,170	1,050	1,223	1,001	1,098	1,136	1,146	1,088

Standard errors clustered at the EA level in parentheses: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Notes: See Table 2.

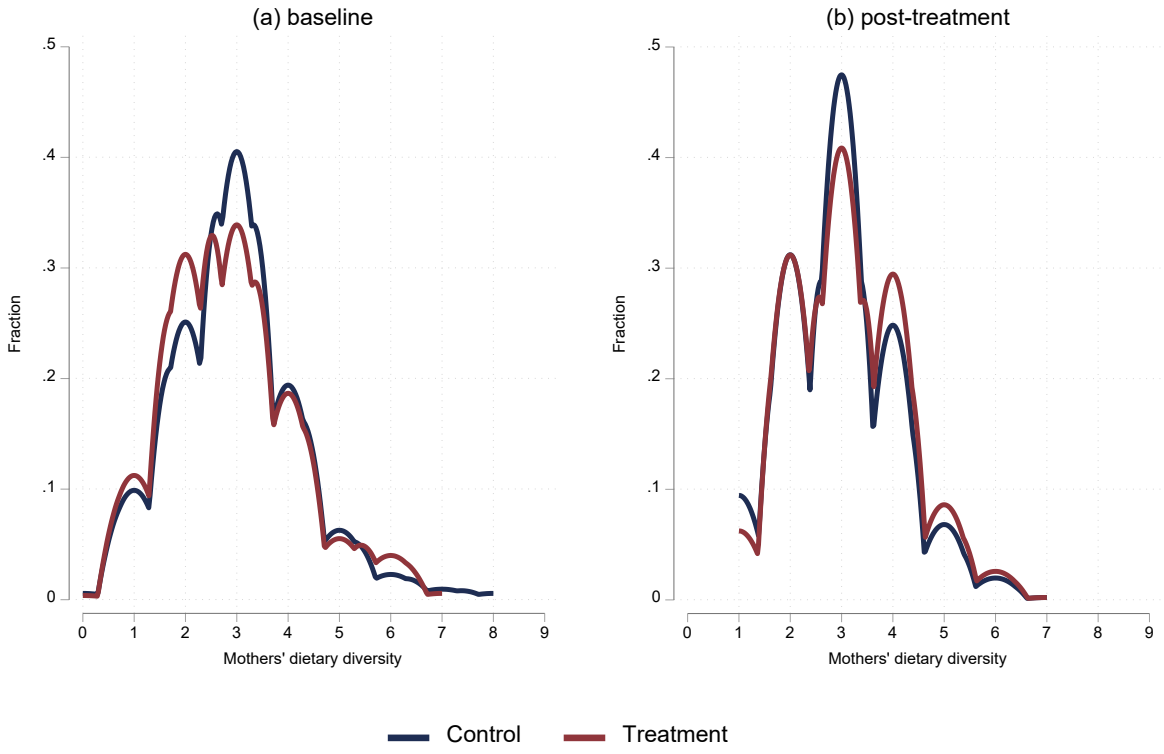
Table 5: Falsification test: Effect of treatment on child diet diversity, respondent fixed effects estimates

	(1) Diet diversity score	(2) Diet diversity score	(3) Minimum diet diversity dummy	(4) Minimum diet diversity dummy
Treatment: Early placement	0.077	0.088	0.048	0.048
	(0.097)	(0.098)	(0.042)	(0.043)
Round	0.080	0.120	0.019	0.010
	(0.070)	(0.093)	(0.030)	(0.039)
Controls	No	Yes	No	Yes
Interview day	No	Yes	No	Yes
Enumerator fixed effect	No	Yes	No	Yes
Mean of dependent variable	2.690	2.690	0.213	0.213
R-squared	0.01	0.06	0.01	0.05
No. observations	1,763	1,763	1,763	1,763

Standard errors clustered at the EA level in parentheses: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$

Notes: See Table 2.

Figure 1: Distribution of dietary diversity by survey round and treatment status



Appendix

Table A1: Effects on diet diversity of mothers, using in-person 2019 as baseline, respondent fixed effects estimates

	(1) Diet diversity score	(2) Diet diversity score	(3) Minimum diet diversity dummy	(4) Minimum diet diversity dummy
Treatment	0.208**	0.201**	0.085**	0.081**
	(0.096)	(0.095)	(0.039)	(0.039)
Round	0.146*	0.205**	0.027	0.047
	(0.081)	(0.085)	(0.033)	(0.035)
Controls	No	Yes	No	Yes
Mean of dependent variable	2.905	2.905	0.276	0.276
R-squared	0.03	0.04	0.02	0.03
No. observations	2228	2228	2228	2228

Standard errors clustered at the EA level in parentheses: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$

Notes: controls included are a dummy variable indicating whether the mother was fasting and food security status of the household.

Table A2: Effects on diet diversity of mothers by food groups, using in-person 2019 as baseline, respondent fixed effects estimates

	(1) Staples, beans and nuts	(2) Staples, beans and nuts	(3) Animal source foods	(4) Animal source foods	(5) Vegetables and fruits	(6) Vegetables and fruits
Treatment	0.016	0.017	0.097***	0.092***	0.040	0.040
	(0.015)	(0.015)	(0.035)	(0.035)	(0.036)	(0.036)
Round	0.049***	0.057***	-0.150***	-0.146***	0.076**	0.100***
	(0.012)	(0.013)	(0.031)	(0.032)	(0.029)	(0.029)
Controls	No	Yes	No	Yes	No	Yes
Mean of dependent variable	0.964	0.964	0.270	0.270	0.724	0.724
R-squared	0.05	0.05	0.04	0.05	0.03	0.04
No. observations	2228	2228	2228	2228	2228	2228

Standard errors clustered at the EA level in parentheses: $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$

Notes: controls included are a dummy variable indicating whether the mother was fasting and food security status of the household

Table A3: Effects on diet diversity of children under 36 months, using in-person 2019 as baseline, respondent fixed effects estimates

	(1) Diet diversity score	(2) Diet diversity score	(3) Minimum diet diversity dummy	(4) Minimum diet diversity dummy
Treatment	-0.099	-0.098	0.020	0.018
	(0.111)	(0.111)	(0.033)	(0.033)
Round	0.977***	1.031***	0.161***	0.180***
	(0.084)	(0.085)	(0.025)	(0.025)
Controls	No	Yes	No	Yes
Mean	2.279	2.279	0.140	0.140
R-square	0.29	0.29	0.13	0.14
N	1872	1872	1872	1872

Standard errors clustered at the EA level in parentheses: $^{*}p<0.10, ^{**}p<0.05, ^{***}p<0.01$

Notes: controls included are a dummy variable indicating whether the mother was fasting and food security status of the household.