INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI

RESEARCH PROGRAM ON Agriculture for Nutrition and Health
CGIAR
Led by IFPRI

IFPRI Discussion Paper 01976

November 2020

Telescoping Causes Overstatement in Recalled Food Consumption

Evidence from a Survey Experiment in Ethiopia

Gashaw T. Abate

Alan de Brauw

John Gibson

Kalle Hirvonen

Abdulazize Wolle

Markets, Trade, and Institutions Division
Development Strategy and Governance Division

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

AUTHORS

Gashaw T. Abate (g.abate@cgiar.org) is a Research Fellow in the Markets, Trade, and Institutions Division of the International Food Policy Research Institute (IFPRI), Washington DC.

Alan de Brauw (a.debrauw@cgiar.org) is a Senior Research Fellow in IFPRI's Markets, Trade, and Institutions Division, Washington DC.

John Gibson (jkgibson@waikato.ac.nz) is a Professor of Economics at the University of Waikato, New Zealand

Kalle Hirvonen (k.hirvonen@cgiar.org) is a Senior Research Fellow in IFPRI's Development Strategy and Governance Division, based in Addis Ababa.

Abdulazize Wolle (a.wolle@cgiar.org) is a Research Officer in IFPRI's Development Strategy and Governance Division, based in Addis Ababa.

Notices

Abstract

Telescoping errors occur if survey respondents misdate consumption or expenditure episodes by including events from outside the reference period in their recall. Concern about telescoping influenced the design of early Living Standards Measurement Study (LSMS) surveys, which used a two-visit interview format to allow a bounded recall. This design fell out of favor although not for evidence-based reasons. Recent guidelines to harmonize food data collection in low- and middle-income countries by using one-week recall increase the relevance of telescoping because errors spread over a shorter period will loom larger. To provide evidence on telescoping, we conducted a survey experiment in Ethiopia, randomly assigning a balanced sample – either a two-visit bounded recall or a single visit unbounded recall. The average value of reported food consumption is 16 percent higher in the unbounded single visit recall relative to the two-visit bounded recall. Put differently, in this experiment, telescoping errors amount, on average, to an entire extra day worth of consumption being included in the report for the last seven days. Most of the error is explained by difference in reporting of spending on less frequently consumed, protein-rich foods, so apparent diet diversity and dietary quality indicators are likely to be overstated when using unbounded recall.

Keywords: Diet quality, Food consumption, Household surveys, Recall, Telescoping

JEL Codes: C81, D12, I32

Acknowledgements

We thank the team at NEED for excellent survey coordination, particularly Abinet Tekle, Betelhem Lakew, Alemayehu Deme, and Abraha Weldegerima. We are also grateful for the survey supervisors and enumerators for their hard work in interviewing the respondents. None of this work would have been possible without the generosity of the households that took part in these surveys. We thank them all sincerely. Thanks to Kibrom A. Abay, Kaleab Baye, Kathleen Beegle, Calogero Carletto, Joachim De Weerdt, Tesfaye Hailu, Dean Jolliffe, Talip Kilic, Otto Toivanen, Rob Vos, and Alberto Zezza for useful comments. This work was undertaken as part of, and funded by, the CGIAR Research Program on Agriculture for Nutrition and Health (A4NH). The opinions expressed here belong to the authors and do not necessarily reflect those of A4NH or CGIAR.

1. Introduction

Monitoring progress towards meeting the first two Sustainable Development Goals (SDGs), to end poverty and hunger, requires accurate measurement. In low- and middle-income countries, the large consumption share for food means that any comprehensive assessment of welfare requires accurate food consumption data, typically obtained through surveys. Yet, our understanding of errors in these survey measures remain incomplete. Recently, this understanding has improved due to a series of survey experiments; key findings are described by De Weerdt, Gibson, and Beegle (2020). Special issues of journals that focus on survey measurement and survey experiments are introduced by McKenzie and Rosenzweig (2012) and Zezza, Carletto, Fiedler, Gennari, and Jolliffe (2017). These experiments generally show that survey design can have a large influence on measurements of concepts related to household consumption, labor use, and agricultural production. Moreover, the error structures appear to be complex and do not follow the classical assumptions of errors that are purely random. Given that survey designs exist that can reduce the occurrence of these errors, such designs may be quite important to measuring indicators of progress towards the SDGs.

Surprisingly, a long-discussed type of error in consumption surveys has not been addressed by the survey experiments covered in this recent literature. This error is *telescoping*, which is mis-dating by either recalling more distant events as occurring more recently (forward telescoping) or pushing recent events further back in time (backwards telescoping). Mahalanobis and Sen (1953) gave early attention to telescoping after finding that the food consumption of Indian households reported with a one-week unbounded recall appeared to greatly exceed that of households for whom the foods that were consumed had been weighed. Concern about telescoping influenced design of the early Living Standards Measurement Study (LSMS) surveys, which adopted a two-visit interview format partly to allow a bounded recall.[1] What Deaton (1997, p. 26) called the standard format of LSMS surveys was to have "two visits, roughly two weeks apart, and the interviewer asks how much was spent on each food item 'since my last visit.'"

The idea for this format is that the first visit to the household provides a distinct start to the recall period in the mind of the respondent, so that when the recall questions are asked in the second visit

[1] Using two visits also let very long LSMS interviews be broken into two more reasonable blocks of time.

it is for the period bound by the initial visit to the household (Grootaert, 1986). While the first use of bounded recall is usually attributed to Neter and Waksberg (1964), and was in quite a different context to the LSMS; surveying infrequent expenditures on dwelling alterations and repairs and using an unbounded recall on the first visit so that expenses reported then could be conveyed to respondents in the subsequent interview to help prevent them being reported again.[2] Other studies with benchmarks for assessing the extent of telescoping are also for high value and infrequent purchases, like computers (Morwitz, 1997). It is unclear if reports for frequently consumed and low value items like food would exhibit the same response to bounding, especially as there is no easy way (and the LSMS surveys did not try) to gather consumption data in the first interview that can be relayed to respondents in the second interview, to ensure such foods are not reported again.

It is conceivable that a bounding visit could help respondents remember whether episodes that involved low value items occurred within the recall period. The difficulty of this memory task is shown by the following comment from a discussant of one of the first papers to suggest that telescoping errors may explain why expenditures measured with a one-week unbounded recall appeared higher than what other survey approaches showed:

> *I confess that if somebody asked me to give an account of my expenditure for the last seven days I should not remember whether it was seven or eight days since I bought an article such as face powder* (Cole & Utting, 1956, p. 389).

However, whether bounding helps respondents to remember the occurrence of episodes largely depends on the cognitive process that respondents use to answer these questions. A plausible model is that when there are a large number of items, respondents do not try to remember and count, and instead use a rate-based estimation strategy where they multiply the frequency of occurrence by the length of the reference period (e.g., 2 eggs per day, ergo 14 eggs for 7 days). If rate-based estimation is used to answer questions about food consumption, then there is no reason to expect that a bounding visit would help to improve the accuracy of these data.

Perhaps because of the lack of evidence on the importance of telescoping errors in food consumption data and on whether bounding could reduce these errors, the two-visit LSMS format

[2] For a useful review of the early literature on this topic, see Dex (1995).

was never as widely used as the quote from Deaton (1997) suggests. In LSMS surveys prior to 1997, the two-visit LSMS format was used for Côte d'Ivoire, Ghana, Pakistan, and Vietnam but not in Jamaica, Nepal, South Africa, and many other countries. Moreover, even where it had been used, the two-visit format fell out of favor, with Vietnam abandoning it after their 1998 survey, and other countries following suit. Thus, by the year 2000, Deaton and Grosh (2000, p. 114) already note (in the LSMS books entitled *Designing Household Surveys for Developing Countries*) that the two-visit structure was being used less frequently. Going back to using unbounded recall was not due to some evidence of the two-visit format failing, rather, it reflects the practical matter that including bounding visits will tend to raise survey costs and complicate field work because of the need to return to the same households within a week or two. Absent any firm evidence that the two-visit format helped, it was easy to jettison the method for something simpler.

Yet researchers continue to speculate that telescoping errors affect patterns found in consumption survey data. For example, in a survey experiment in Tanzania, Beegle, De Weerdt, Friedman, and Gibson (2012) find unbounded 7-day recall yields higher estimates of consumption and lower poverty rates than does 14-day recall with the same list of items, and comes closer to matching their benchmark from a highly supervised individual diary. They note telescoping could contribute to this pattern, as bringing forward consumption that happened before the recall period matters more when that error is spread over just 7 days rather than over 14 days (Eisenhower, Mathiowetz, and Morganstein, 2004). In order to provide more evidence on telescoping, Beegle et al. (2012) suggest that future survey experiments could compare bounded and unbounded 7-day recall. A further analysis of this experiment found that macro- and micro-nutrient intakes measured from the 7-day recall most closely match those derived from the benchmark individual diaries, even as the 7-day recall is sensitive to having errors that vary with household characteristics (Ameye, De Weerdt, and Gibson, 2020). These authors suggest that 7-day recall may have offsetting errors – including from telescoping – that roughly balance, and note that an experiment on telescoping may help diagnose the source of the apparent success (in terms of matching the benchmark) with this design. In an experiment in Niger, 7-day unbounded recall indicated food consumption was 28% higher than what a 7-day diary showed; the authors speculate that telescoping could cause the 7-day recall to overstate the truth (Backiny-Yetna, Steele, and Djima, 2017).

Another relevant analysis of a survey experiment finds food consumption data from a 7-day unbounded recall survey are subject to incidence errors within several important food groups (the respondents entirely forget to report any consumption) that are offset by errors of overstatement in value of what was consumed, conditional on reporting any consumption (Friedman, Beegle, De Weerdt, and Gibson, 2017). Thus, the good performance of the 7-day unbounded recall (in matching data from the benchmark individual diaries in Tanzania) may be from happenstance of incidence errors and value errors canceling each other out. This analysis also found that the overstatement of consumption, conditional on incidence, was more than twice as high for infrequently purchased versus frequently purchased foods or for self-produced foods that are seldom consumed versus those that are frequently consumed. These frequency-related patterns may be due to respondents using a rate-based rule-of-thumb estimate for frequently purchased or consumed items, while they try to remember and count episodes for infrequently consumed foods (Chang and Krosnick, 2003; De Weerdt, Gibson, and Beegle, 2020). With these different modes of answering survey questions, telescoping error would matter more for infrequently consumed foods.

In addition to these speculations resulting from previous survey experiments, three developments make this an opportune time to experimentally examine effects of telescoping on surveyed food consumption. First, there has been a move away from surveys using an abstract construct like the 'usual month' where respondents recall how many months per year purchases are made for each food, how often those purchases are made per month, and the typical spending per occasion (with similar questions for self-production). These questions greatly increase the time taken for survey interviews and add education-related inequality to reported consumption due to the cognitively demanding nature of what is asked, while failing to accurately measure either means or variance-based indicators like inequality statistics (Beegle et al., 2012). With the switch to asking about consumption in an actual recent period, telescoping should matter more than it did when using the hypothetical 'usual month' construct. Second, recent FAO and the World Bank (2018) guidelines for food data collection in household surveys recommend using a 7-day recall, which is shorter than what was often used in the past (where 14-days or one-month recall was asked). As noted above, with a shorter recall period, any errors due to telescoping will loom larger. Finally, more diverse diets due to rising affluence and urbanization may make reports more susceptible to telescoping errors; when many people were poor, with monotonous diets based on the cheapest local staple, survey reports of food consumption could rely on rate-based estimates. Now, with growing numbers who can afford occasional luxuries like meat and fish, that are still eaten sufficiently infrequently that respondents need to remember and count rather than using rate-based estimates, answers to food recall questions may be more affected by telescoping.

To provide evidence on telescoping, we conducted a survey experiment in Addis Ababa, Ethiopia that randomly assigned either two-visit bounded recall or single visit unbounded recall. The sample was part of an end-line survey for evaluating a randomized video-based intervention related to fruit and vegetable consumption, and as part of this evaluation the food consumption of the subjects had been surveyed about 3-4 months prior to our experiment. The unbounded and bounded recall groups were balanced not only on household characteristics but also on their prior food consumption, as measured 3-4 months earlier. We find the value of food consumption appears to be 16 percent higher among the group of households to whom the unbounded recall was administered, relative to the bounded recall group. In effect, an entire extra day of consumption is included in the report for the last seven days. Moreover, this overstatement is not evenly distributed. The effect is particularly prominent for protein-rich foods like meat and eggs that are typically less frequently consumed. As a result, there also are implications for standard indicators of household dietary diversity and diet quality that are derived from consumption survey data, as these indictors may be overstated when unbounded recall is used. Further, as budget shares of specific classes of foods may be affected, income elasticities of demand for those foods could be biased, in turn potentially affecting macroeconomic models using those elasticities (e.g., Pauw, Thurlow, and Ecker, 2017).

We are aware of two other recent but unpublished survey experiments comparing bounded and unbounded recall. Both introduced other design variations as well, such as length of the food list, length of the recall period, frequency of interviewer visits, and type of data capture (diary versus recall). Durazo, Herawati, Pattinasarany, and Jolliffe (2017) found that for Indonesian households, a 7-day unbounded recall indicated that per capita food consumption was 24% higher than what a bounded recall showed, using a recall list with 94 items. This effect was larger than for some of the other design variants tested, such as cutting the food list from 229 items to 126 (which caused just a 2% drop in apparent consumption). Sharp, Buffiere, Himelein, and Gibson (2019) found no difference in food consumption for a Marshall Islands sample given a 7-day unbounded recall and another sample where an initial visit was made to the household. However, for the sample with two visits the recall questions continued to use the "In the last 7 days…" wording, and the gap between visits varied; visit 2 was 8 days after visit 1 for 53% of the sample, 9 or more days after for 14%, and 7 or less days after for 33%. Thus, the latter experiment did not really implement a bounded recall although it does highlight the logistical challenges of this survey design.

2. The survey experiment and data

We designed the survey experiment to understand the implications of telescoping for food consumption measurement by systematically contrasting responses from unbounded and bounded recalls. For the unbounded recall we adopted the common approach used in food consumption measurement, of requiring a respondent to report on the household's food consumption for each item from a list of 128 food items, asking about consumption within the reference or recall period (the last 7 days). For the bounded recall, we introduced a salient recall marker by visiting sample households exactly 7 days prior to the actual survey and in the second visit the respondents were asked to recall consumption of food items since the initial, bounding, visit. The household visits were conducted by survey supervisors, who wore a uniform to distinguish them from other visitors so as to make the visit equally notable and memorable for all sample households in the subgroup.

The survey experiment was implemented as an add-on to an end-line evaluation survey of a randomized controlled trial designed to assess the impact of video-based behavioral change communication on fruit and vegetable consumption in Addis Ababa, Ethiopia (Abate, Baye, de Brauw, and Hirvonen, 2019). The study sample was formed from 930 households randomly selected from six sub-cities, 20 *woredas* (districts), and 40 *ketenas* (neighborhoods; or clusters of households) within Addis Ababa.[3] To ensure the experiment on bounded versus unbounded recall did not affect the outcomes of the impact evaluation, we cross-randomized study samples into the bounded and unbounded consumption recall subgroups. The two groups differ only by the question format of the consumption module (Table 1). For all other aspects, the survey designs for the two groups were identical: the method of data capture, designated respondent, and the number of food items in the recall list.

[3] Melesse, van den Berg, de Brauw, and Abate (2019) provide a a detailed description of the sampling strategy.

The survey experiment took place between January 24 and February 11, 2020, and attempted to re-interview all 930 households who had been interviewed in September 2019.[4] Out of the households who had been interviewed in September 2019, 35 households were not interviewed in January-February.[5] The response rates are comparable across the two groups: 97% and 95% for the unbounded and bounded subgroups, respectively.

The survey module on consumption targeted the household member who was most knowledgeable of the household's food shopping and preparation as the primary respondent. More than 90 percent of the respondents were women. The food consumption module consisted of 128 food items.[6] For each item, we asked the respondent whether their household had consumed the item in the past 7 days (or, for the bounded group whether they had consumed it since the recent visit by a survey team member). If the answer was yes, the respondent was asked the number of days in which the item was consumed and the total quantity consumed during the recall period.[7]

We converted the reported quantities into local currency units (Ethiopian birr) using retail price data collected by the Central Statistical Agency (CSA) of Ethiopia on a monthly basis.[8] We also converted consumption in kilograms into calorie and protein equivalents using conversion factors provided in the Ethiopian food composition tables (EPHI, undated). These calculations account for refuse by using the edible portion estimates provided by the United States Department of Agriculture (USDA, 2013).

Extreme values in household per capita consumption variables (birr, kcal and protein) used in the analyses were winsorized to the 99th percentile.[9] After dropping five households with implausibly

[4] Previous work in Ethiopia shows how food consumption in urban areas is affected by religious fasting (Hirvonen, Taffesse, and Worku, 2016). To this end, we note that there was no major Orthodox or Muslim fasting when our survey experiment took place.
[5] Sixteen households refused the interview, 15 could not be found in their house during the survey visit, survey enumerators were unable to track 3 households, and sadly 1 respondent had passed away.
[6] The selection of these food items was based on the 2016 Household Consumption Expenditure Survey (HCES) data collected by the Central Statistical Agency of Ethiopia. HCES is a nationally and regionally representative sample and the 2016 sample had about 3,800 households in Addis Ababa. We selected all food items that were consumed by at least one percent of the 2016-HCES households located in Addis Ababa.
[7] This module only considered food consumed in the house. The survey instrument had another module for measuring foods consumed outside the house but these data are not considered in this study.
[8] More specifically, we used the CSA retail price data for February 2020, restricting the price observations to Addis Ababa.
[9] Results are robust to including the non-winsorized values; see the online appendix.

large per capita consumption values,[10] the final data set for analysis includes 890 households.[11] The average household in the bounded group consumed food valued at 275 birr per person during the 7 day period (equivalent to about USD$8.50 per person per week at market exchange rates). The average *daily* consumption reported was 1,640 kilocalories per capita, including 45 grams of protein per capita.[12]

We also considered household dietary diversity, often used as an indicator of household food security (Hoddinott and Yohannes, 2002). First, we computed the household dietary diversity score (HDDS) of Swindale and Bilinsky (2006) by grouping the 128 food items in our consumption module into 12 food groups: cereals; roots and tubers; vegetables; fruits; meat, poultry and offal; eggs; fish and seafood; pulses, legumes and nuts; milk and milk products; oil and fats; sugar and honey; and miscellaneous foods. The HDDS is a sum of all food groups from which the household consumed food items during the 7-day recall period, with a minimum of 1 and maximum of 12. Second, we constructed the food consumption score (FCS) developed by the WFP (2008). The FCS combines dietary diversity and consumption frequency by grouping the consumed food items into nine groups[13] and allocating more weight to protein rich foods. The weighted FCS index ranges between 0 and 112, with higher scores indicating a better food security situation.

Table 2 shows that the bounded and unbounded groups are similar in terms of household characteristics. We also note that the (random) assignment into the video experiment study arms is orthogonal to the (random) allocation into the telescoping experiment groups. Finally, the sub-samples given bounded and unbounded recall were balanced not only on household characteristics but also on their baseline food consumption collected in September 2019 (i.e., 3-4 months before the survey experiment).

[10] Specifically, we considered calorie consumption above 5,000 kcal per adult equivalent implausible.

[11] Three of these households were in the bounded group and two in the unbounded group. Results are robust to including these five households in the estimation sample; see the online appendix.

[12] The corresponding values in per adult equivalent unit terms are 327 birr in 7 days, 1,946 kilocalories per day and 53 grams of protein per day.

[13] The FCS food groups are: main staples (weight: 2); pulses (3); vegetables (1); fruits (1); meat, eggs, fish (4); dairy products (4); sugar (0.5); oil/butter (0.5); and condiments (0).

3. Results

We quantify the difference in reported per capita consumption values across the two groups using ordinary least squares (OLS). In our most basic model, we regress both the per capita food consumption value and its logarithm on a binary treatment variable valued 1 if the household was randomly selected into the unbounded recall group, and zero if in the bounded recall group. We then also control for differences in basic household characteristics (household size, a binary variable to indicate male-headed households, the head's education in years), the household's treatment status in the video experiment and unobserved characteristics between sub-cities (fixed effects for each sub-city). The standard errors in all regressions are clustered at the *ketena* level.[14]

Figure 1 shows the full distributions of (log) household weekly per capita food consumption measured in birr for both bounded and unbounded recall groups. Relative to the bounded recall group, the estimated food consumption distribution for the unbounded group is shifted to the right, indicating larger reported food consumption values across the board.

Table 3 then shows the coefficient estimates for the difference in unbounded and bounded recall for household weekly per capita food consumption measured in birr. The estimated coefficients quantify the difference in the consumption outcome when the consumption module was based on an unbounded recall relative to when a bounded recall module was used. In columns 1 and 2, the outcome variable is the household per capita food consumption value in birr, whereas it is the logarithm of the value in columns 3 and 4. Columns 1 and 3 do not use additional covariates, whereas columns 2 and 4 control for additional covariates as described above. As the differences between the unadjusted and adjusted regressions are negligible, we focus our reporting and discussion on the adjusted regression results. When we discuss percentage differences derived from the coefficients in semi-log regressions they are based on $100 \times \left(e^{\hat{\beta}-0.5\hat{V}(\hat{\beta})}-1\right)$ with confidence intervals from the approximate unbiased variance estimator of van Garderen and Shah (2002).

The regression coefficient in column 2 in Table 3 shows that unbounded recall increases the reported per capita consumption value by 54 birr relative to bounded recall (p-value<0.001; 95%

[14] Our household sample groups into 6 sub-cities and 40 *ketenas*.

CI: 25.5 – 83.3). Considering the mean of the bounded group, this impact of using an unbounded recall is equivalent to a 20 percent increase in apparent food consumption per capita. In the semi-log regression models, the corresponding difference between the unbounded and bounded recall is 16 percent (p-value<0.001; 95% CI: 7.45 – 25.95). As noted in the introduction, these estimates are large and are roughly equivalent to adding an entire day of consumption value to the report of the household 7-day consumption total. The most plausible source of this apparently higher rate of food consumption is forward telescoping where consumption episodes occurring more than 7 days ago are included in the report for the last 7-days.

In Table 4 we use per capita calorie and protein intakes as the dependent variable, in columns 1-2 and 3-4, respectively. Unbounded recall results in 8.8 percent higher reported per capita calorie intakes compared to bounded recall (p-value<0.01; 95% CI: 3.10 – 14.7), which is considerably lower than the estimated difference when the consumption value is expressed in birr terms. Meanwhile the corresponding difference in protein intake is 16 percent (p-value<0.001; 95% CI: 7.91 – 24.5), which is similar to the estimated impact on the birr value reported in Table 3. That apparent protein intake is more sensitive to differences in survey design than is calorie intake agrees with a finding from Ameye, De Weerdt, and Gibson (2020), who note that previous findings on the fragility of calorie-based hunger estimates to variations in the design of food consumption surveys (De Weerdt, Beegle, Friedman and Gibson, 2016) are likely to understate the fragility of survey estimates when a richer consideration of nutrition that focuses on macro- and micro-nutrients is used.

Next, we look at how the type of recall survey influences the two indicators of diet diversity and diet quality, HDDS and FCS. The coefficient in Column 2 of Table 5 shows that the households in the unbounded recall group report consuming from 0.3 more food groups than do the households in the bounded recall group (p-value<0.01; 95% CI: 0.11 – 0.47). Considering that the mean HDDS in the bounded group is 9.1 food groups, this represents a 3 percent increase in HDDS when unbounded recall is used. The effect on FCS is slightly larger in magnitude: the mean FCS in the unbounded recall group is 4.3 units (p-value<0.01; 95% CI: 1.62 – 6.93) or 6-percent higher than in the bounded recall group.

In Table 6, we examine the impact by food group. In Panel A, the dependent variable has a value of one if the household consumed from the food group and zero otherwise. Households in the unbounded recall group are 8 percentage points (or 11 percent) more likely to report having consumed meat, poultry, fish or eggs in the past 7 days ($p<0.01$). In contrast, the difference between the bounded and unbounded subsamples is not statistically significant for fruit and dairy food groups. All other food groups were consumed virtually by all households and thus we do not have sufficient variation in the outcome variable to estimate coefficients using this method. In Panel B, the dependent variable measures the number of days the household consumed from the food group. We find the unbounded recall increases reported consumption frequency of meat and eggs by 0.75 days ($p<0.001$) and of fruit by 0.32 days ($p<0.05$). Meanwhile, the difference for pulses and dairy food groups is not statistically different from zero. In Panel C, we consider food consumption measured in birr. The choice of recall has a particularly large influence on the reported consumption values in the 'meat and eggs' food group. The mean per capita consumption in this group is 110 birr higher ($p<0.01$) when the unbounded recall is used. Considering the mean in the bounded group of 272 birr, this finding translates into a 40 percent increase in the apparent value of consumption when unbounded recall is used. Using unbounded recall also increases the reported consumption of staple crops by 31 birr – or 10 percent ($p<0.01$).

Overall, the impact of using a bounded versus an unbounded recall is most apparent for protein-rich foods (e.g., meat, poultry, eggs). Amongst our sample, these foods are typically less regularly consumed than are calorie-rich foods (e.g., maize, wheat, teff). Therefore, one implication of the results in Tables 4, 5, and 6 (given that diet diversity and quality scores also rise with consumption of the protein-rich foods) is that the telescoping effect is driven by infrequently consumed food items. This pattern has also been suggested by previous studies (e.g. Friedman et al, 2017), although without the benefit of an experiment designed to get at the telescoping effect.

To more directly explore this potential variation in terms of consumption frequency, we shift to food item level data to calculate the mean weekly per capita food consumption for each item, separately for the unbounded and bounded recall groups. We use the two means to calculate a ratio between mean per capita consumption in the unbounded group relative to the bounded group for each food item. We then compare this ratio to the mean number of days each food item was reportedly consumed by all households in our sample. After dropping food items that were consumed by less than 5 percent of the sample,[15] we are left with 70 food items. Figure 2 shows a scatter plot of the ratio of the mean weekly per capita food consumption in the unbounded sample relative to the bounded sample against the mean number of days that the item was consumed by all households. A linear regression line weighted by the price of the food item shows that the telescoping error is larger for infrequently consumed than for frequently consumed foods. For foods consumed nearly every day of the week, the predicted ratio is close to one indicating that the telescoping error is close to zero.

Heterogeneity by household characteristics

The degree of telescoping error *could* vary with household and respondent characteristics, if these affect the nature of the reporting task or the ability to do this task. To explore this possibility, we sequentially interacted our binary treatment variable with three control variables: household size, a binary variable to indicate male-headed households, and the head's education in years (Table 7). The coefficient on the interaction term is insignificant when we interact the treatment variable with the binary variable capturing male-headed household (column 2) or with the variable capturing head's education level in years (column 3). In contrast, the interaction term is significant ($p<0.05$) and negative for household size, indicating that the magnitude of the telescoping error decreases with household size. To illustrate, we plot predictive margins for the two recall groups across household size (Figure 3). We observe that the difference between the two recall types is larger for smaller households and decreases as the household size increases. One interpretation of this pattern is that consumption of a larger household relies on more transactions (e.g., purchases) or has more episodes than occurs for smaller households, over a fixed period like the last 7 days. With more episodes, a respondent for a larger household may be more likely to use rate-based rule-of-thumb reporting rather than recalling and counting (Gibson and Kim, 2007). Consequently, survey devices like a bounding visit that are designed to help make the memory task easier would be expected to have more impact on the reporting by smaller households than by larger ones.

[15] Five percent of the sample is about 45 households. There were 58 food items that were consumed by less than 45 households. For these food items, the ratio of unbounded recall to unbounded recall becomes excessively sensitive to extreme values because the two means are calculated from less than 45 observations. For this reason, we omit these food items from the sample for the scatterplot and regression presented in Figure 2.

Robustness checks

The results of several robustness checks are reported in the (online) appendix. First, we test the sensitivity of our estimates to outliers, by adding the five households with implausibly large food consumption back into the sample, and by using non-winsorized consumption values; both changes result in similar coefficients as reported in Column 4 of Table 3 (Table A1 in the appendix). Second, column 2 in Table A2 shows that the results are similar if we used the median regression based on the least absolute deviation procedure that is less sensitive to outliers or other extreme values than is OLS (Koenker and Bassett, 1978).[16] Moreover, while the estimated impact of the unbounded recall based on the quantile regressions seems slightly larger for richer households, the difference in the impact estimated at the 25th and at the 75th food consumption quintiles is not statistically different from zero (Table A2 in the appendix).

Finally, we conduct a robustness check on Figure 2. Rather than omitting food items that were consumed by less than 5 percent of sample households, an alternative way to deal with these imprecisely estimated values is to winsorize small and large values of the ratio of the mean consumption values. Specifically, Figure A1 in the appendix replicates Figure 2 based on 113 food items, winsorized at the 5th and 95th percentiles.[17] As before, we see that the telescoping error is driven by the infrequently consumed food items.

Alternative interpretations

We interpret the higher reported food consumption of the group surveyed with an unbounded recall as reflecting the impact of telescoping errors. This pattern fits with a widely discussed hypothesis about the memory task that is required of survey respondents. In line with this hypothesis, any features of survey design that make this memory task easier, such as the use of a bounding visit to demarcate the beginning of the recall period in the mind of the respondent, should yield data that are closer to the truth.

[16] Note that in using quantile regression the assumption underlying randomization necessarily changes; when randomizing, the assumption is that the average value of the outcome of interest would be the same for both groups in the absence of the treatment. In a quantile regression, the causal effect estimated is on the distribution of outcomes at the group level, rather than at the quantile specifically, as we cannot assume that individual observations would not change ranking due to the treatment (e.g., Meager, 2020).
[17] 15 food items that were neither consumed by unbounded nor bounded recall groups were dropped.

However, we acknowledge an alternative explanation for the lower reported food consumption of the households who are given the two-visit, bounded, recall. It is possible that visiting the same household twice, in quick succession, leads to declining compliance (Schündeln, 2018) and less cooperative respondents may under-report consumption in order to finish the interview sooner. For example, there is a Yes or No screening question for each of the 128 food items, asking whether there was any consumption within the last 7 days (or since the visit by the survey supervisor) and uncooperative respondents might say No when the true answer is Yes. In this case, the two-visit, bounded, recall might result in food consumption data that are further from the truth.

We believe that this alternative explanation is unlikely, for four reasons. First, the initial visit that the survey supervisor made to the household was very quick, usually taking less than five minutes. So respondents in the two-visit bounded recall group did not bear a time cost much greater than what the respondents in the one-visit unbounded recall group experienced. Second, the patterns of the bounding visit having more effect on reported consumption of small households and of rarely consumed foods can be explained by telescoping (in conjunction with a hypothesis about using rate-based rule-of-thumb reporting when the number of episodes is larger) but are less easily explained by reduced cooperation which should have effects across-the-board. Third, declining cooperation should especially show up for foods that occur late in the list of the 128 foods as respondents learn the pattern of the questions and start answering No when the true answer is Yes for the screening question, but we see no such effect of position in the food list.[18] Fourth, when we use the benchmark for assessing data quality—Benford's Law—that Schündeln (2018) uses in his study of declining respondent compliance in each successive survey visit, there is no difference between the two groups in how the pattern of digits differs from Benford's Law (see Appendix B). Thus, there is no evidence of less cooperative respondents in the two-visit bounded recall group.

Adjustment factors and mean-reverting error patterns

There are far more food consumption surveys carried out every year than there are experiments that can inform about the sensitivity of results to different design choices. Thus, a common request

[18] We tested this point by regressing the ratio of Yes responses between unbounded to bounded recall groups on the order of the item in the food list (i.e., #1 through to #128). The estimated coefficient on the food item order variable was 0.0017 with a White (1980) adjusted 95% confidence interval ranging between -0.0015 and 0.0049 (p-value = 0.291). In other words, the position in the food list is not correlated with the ratio of Yes responses between unbounded to bounded recall groups.

of survey experiments is to provide adjustment factors that may let analysts take data collected in different ways and line them up on a comparable basis. For example, the widely used Deininger and Squire (1996) database of inequality estimates has Gini coefficients from both expenditure surveys and income surveys, with expenditure-based Gini coefficients being 6.6 points lower, on average; thus, some analysts combine the two types of Gini by adding 6.6 points to expenditure-based ones. In the current setting, unbounded recall gives the equivalent of an entire extra day of consumption in the report for what is ostensibly the last 7 days of food consumption. Therefore, an analyst might be tempted to annualize this estimate by multiplying by (365/8) rather than by (365/7) in order to make up for the overstatement that results from telescoping.[19]

Evidence from survey experiments suggests that such adjustment factors are rarely available. One reason is that errors are mean-reverting, contrary to classical assumptions of measurement errors being uncorrelated with anything of interest. For example, Pradhan (2009) considers evidence from the SUSENAS survey in Indonesia where some households get a short list of broadly defined items for their consumption recall while others get a far longer list of narrowly defined items. Using fewer questions yields lower consumption but the fraction by which consumption is under-estimated increases as consumption rises, so without data on actual consumption one cannot devise a simple correction factor to line up data from the two survey designs. Other consumption surveys also show this mean-reverting error pattern (Gibson, Beegle, De Weerdt, and Friedman, 2015).

The ideal way to test for mean reverting error is to regress a noisy measure on the true measure, for the same household. If errors are random, the slope coefficient of this regression should be 1 (and the intercept will be 0), while with mean-reverting errors the slope coefficient is less than 1. The design of our experiment does not give us two measures for the same household, so instead we follow the approach of Gibson et al (2015) and take the mean of the (log) per capita household food consumption across each sampling unit (*ketena*), separately for both recall groups. We then regress the *ketena* level means for the unbounded recall group (the noisier measure) on the means for the bounded recall group. The slope coefficient is 0.201 with a White (1980) adjusted 95% confidence interval ranging between -0.030 and 0.433. Consequently, the null hypothesis that the

[19] This approach uses a naïve extrapolation to annualize and even if the sample is staggered over the weeks in the year so that extrapolation to annual means and totals is correct, any variance-based measures—including the share of observations in the lower tail, such as a poverty rate or a hunger rate—will be overstated because short-term shocks that are subsequently reversed (at least partially) in the rest of the year are ignored (Gibson, 2020).

errors are random (i.e., coefficient equals 1) is firmly rejected ($p<0.0001$) in favor of the alternative hypothesis of a mean reverting error structure. This error structure matters because with mean-reverting errors the bias in coefficient estimates can be in either direction rather than just attenuation as happens with classically mis-measured right-hand side variables (Abay, Abate, Barrett and Bernard, 2019) and the typical econometric approach for mitigating errors in variables bias, of using instrumental variables, is unlikely to be successful (Gibson et al, 2015).

4. Cost implications

One unexpected finding from recent survey experiments is that design variations that are expected to save either time or money often do neither, while potentially giving worse quality data (De Weerdt, Gibson, and Beegle, 2020). For example, in the Tanzania experiment analyzed by Beegle et al. (2012) cutting the number of items on the food recall list from 58 to 17 reduced interview time by just 15% on average (going from 49 minutes to 41 minutes). Likewise, in an experiment in Indonesia, cutting the length of the food recall list from 229 items to just 21 reduced the average interview time (for food at home) by just nine minutes (Durazo et al., 2017).

It appears that in the setting of our experiment, in Addis Ababa, using the unbounded recall as a way to save time and money has a similarly modest effect on costs, while opening up the results to the impact of telescoping errors, as we show above. The cost comparison for the two ways of implementing the recall survey in the case of our experiment is straightforward, as the two arms are identical in all aspects of the survey design and administration (Table 1), except that households in the bounded group were visited by a survey supervisor 7 days prior to the actual survey in order to establish the recall marker in the mind of the respondents. In practice, this resulted in an additional 7 days of field work by survey supervisors prior to the actual survey commencement, in order to visit households who were scheduled to be interviewed during the first week of the survey. Once the actual survey begins, we simply adjust the regular appointment visits by survey supervisors in order for them to be making visits 7 days prior to the actual survey for households in the bounded group. Thus, the field expenses associated with an extra 7 days of fieldwork by supervisors prior to the actual survey were the only additional costs due to bounded recall. Our *back-of-the-envelope* estimates show that these extra costs were US$3.60 per household, which is 6.5% higher than the cost for the unbounded recall.

We acknowledge that the costs of using the two-visit format in order to implement bounded recall may be higher in other settings. For example, in surveys of rural communities where transport costs would be higher, arranging two visits to the household would have a bigger impact on both overall costs and the time demands on supervisors. Indeed, in the experiment in the Marshall Islands (Sharp et al., 2019), where the overall cost was more than US$1,000 per household partly due to expensive boat travel to atoll locations, using the two-visit recall format versus single-visit unbounded recall appears to have increased costs by at least 30 percent. However, the bounded approach can be feasible in other rural surveys, especially with resident enumerators. For example, the Ethiopian household consumption expenditure survey (HCES), a nationally representative survey and the official source of poverty statistics in the country, is based on a two-visit structure (without bounded recall). At least in Ethiopia, the timing of the HCES visits can potentially be adjusted with minimal cost implications.

Clearly more experiments are needed to understand both the cost and benefit implications of the two-visit format in different settings. However, at least in urban areas the cost savings from using an unbounded recall may not be very great. Yet it is in urban areas where unbounded recall surveys may be most susceptible to telescoping errors; it is urban households whose diets are likely to include infrequently consumed non-staple foods such as meat and eggs that appear to be most prone to having overstated reports of consumption due to telescoping. Thus, a benefit-cost comparison of the two-visit format may be most favorable in urban areas.

5. Discussion and Conclusions

In this paper, we report on a survey experiment in Addis Ababa conducted with a sample that were asked to recall 7 days of food consumption. The aim of the experiment was to help understand the effects of telescoping errors on food consumption measures. A randomly selected subsample had their 7-day recall bounded by a short visit from survey supervisors seven days before the actual enumeration (and the questions were then phrased in terms of consumption since the recent visit by the survey supervisor). Their reported food consumption is compared with that of a subsample whose food consumption was reported with an unbounded recall. The unbounded recall survey design is currently widely used, while the bounded recall design was used in the past in the early stages of the LSMS surveys. We find that reported per capita household food consumption expenditures are 16 percent higher among households in the unbounded group. The apparent per capita protein consumption for this subsample is also 16 percent higher and per capita calories nine percent higher than amongst the bounded recall group. These differences are most likely due to forward telescoping, which appears predominantly to occur for foods that are less frequently consumed, such as animal source foods.

The results have potentially important implications for three types of measurements. First, poverty measurement might be affected in two ways. If households given unbounded recall overstate food consumption, then they may appear less poor than they really are. For example, for some household marginally below the poverty line, their reported consumption that includes the telescoping error could put them just above the line, and so the true poverty rate would be understated. The strength of this effect depends partly on the type of poverty line: if it is either based on a global standard (e.g., $1.90 per capita per day) or it is a poverty line established long ago whose value is simply updated with some price index like a CPI, then the error is in the welfare measure but not in the threshold used to distinguish the poor from the non-poor. On the other hand, if the poverty line and the welfare measure are derived from the same survey that is subject to telescoping bias, as could be the case for a cost-of-basic needs food poverty line, then the errors might cancel out.[20]

Second, estimates of food insecurity, of dietary quality, and more generally of hunger, that rely on food consumption surveys are likely to be affected by these telescoping errors. There is already considerable debate about these measurements and the role of household consumption expenditure survey data (see, for example, De Weerdt et al., 2016 for a summary). Given that telescoping errors have a larger effect for a shorter recall period, the recent recommendation by the FAO and World Bank (2018) to harmonize food consumption surveys in low- and middle-income countries by using a one-week recall period may see a spurious rise in apparent diet quality and a fall in apparent hunger, because the bias due to telescoping will be relatively more important than it was when food consumption surveys were using longer recall periods.

Third, consider again the fact that infrequently consumed foods, including those higher in protein, appear most overstated in a 7-day recall relative to the effect on reported consumption of foods

[20] However, the usual approach to setting cost-of-basic needs food poverty lines relies only on a calorie target, and the results here show that calories are less overstated by telescoping bias than is overall food consumption. So once this feature of poverty lines is allowed for, the error in the welfare measure could be larger than the error in the threshold. This feature may be another reason to use food poverty lines that are derived from linear programming in order to consider dietary requirements for all macro- and micronutrients (Ameye, De Weerdt, and Gibson, 2020).

that are consumed more frequently (such as major staples). It is easy to therefore see that the share of protein rich foods in the overall food budget will be overstated, and that will tend to bias food demand elasticities (Deaton, 1997). As these elasticities are used in macro-level models that are attempting to measure food demand, the results of such models should be treated with caution when predicting demand for anything but staple foods.

There are two obvious ways in which the experiment here could be extended in order to learn more about these telescoping effects. The first is to include both urban and rural households in any future experiment, as the costs may be higher and the benefits lower, of implementing the bounded recall approach in rural areas. The second extension is to consider other ways to reduce telescoping errors without using the two-visit bounded recall approach. For example, if corroborating evidence that it is the infrequently consumed foods that are especially prone to telescoping is generated, survey experiments could examine other ways to ask questions about these foods even within the single visit format. For example, the standard design is to use a fixed interval, such as the last 7 days, and then ask for each food in turn whether a consumption episode occurred within that interval. For the infrequently consumed foods, it is evidently difficult for many respondents to correctly place the last consumption episodes into this interval. It may instead be more natural for some foods for respondents to answer in terms of when was the last occasion that the food was consumed, perhaps with some prompts about the sort of occasions when such foods might be consumed – especially for expensive and rarely consumed foods – and to then 'unfold' the questions from that remembered event.

References

Abate, G. T., Baye, K., de Brauw, A., & Hirvonen, K. (2019). Promoting fruit and vegetable intake in urban Ethiopia: An experiment using video-based communication. IFPRI-MTID project note, Washington D.C.: International Food Policy Research Institute (IFPRI).

Abay, K. A., Abate, G. T., Barrett, C. B., & 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*, 171-184.

Ameye, H., De Weerdt, J., & Gibson, J. (2020). Measuring macro-and micronutrient intake in multi-purpose surveys: evidence from a survey experiment in Tanzania. LICOS Discussion Paper Series, Discussion Paper 421/2020, Leuven: KU Leuven, Faculty of Economics and Business (FEB).

Backiny-Yetna, P., Steele, D., & Djima, I. Y. (2017). The impact of household food consumption data collection methods on poverty and inequality measures in Niger. *Food Policy, 72*(October), 7-19.

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, 98*(1), 3-18.

Chang, L., & Krosnick, J. A. (2003). Measuring the frequency of regular behaviors: Comparing the "typical week" to the "past week". *Sociological Methodology, 33*(1), 55-80.

Cole, D., & Utting, J. (1956). Estimating expenditure, saving and income from household budgets. *Journal of the Royal Statistical Society. Series A (General), 119*(4), 371-392.

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

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.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore, MD: Published for the World Bank by Johns Hopkins University Press.

Deaton, A., & Grosh, M. (2000). Consumption. In M. Grosh & P. Glewwe (Eds.), *Designing Household Survey Questionaires for Developing Countries: Lessons from 15 Years of Living Standards Measurement Study* (Vol. 1, pp. 91-133). Washington D.C.: World Bank.

Dex, S. (1995). The reliability of recall data: a literature review. *Bulletin de Methodologie Sociologique*, 49(1), 58-89.

Deininger, K., & Squire, L. (1996). A new data set measuring income inequality. *The World Bank Economic Review*, *10*(3), 565-591.

Durazo, J., Herawati, Y., Pattinasarany, G., & Jolliffe, D. (2017) Bridging surveys: A methodological experiment of food consumption data in Indonesia. Paper presented at the 15th European Association of Agricultural Economists Congress, Parma, Italy, August.

Eisenhower, D., Mathiowetz, N. A., & Morganstein, D. (2004). Recall error: Sources and bias reduction techniques. *Measurement Errors in Surveys*, 125-144.

EPHI. (undated). Expanded Food Composition Table for Use in Ethiopia. Addis Ababa Ethiopian Public Health Institute (EPHI).

FAO, & the World Bank. (2018). Food Data Collection in Household Consumption and Expenditure Surveys : Guidelines for Low- and Middle-Income Countries. Accessed on 21 June, 2020 from: https://openknowledge.worldbank.org/handle/10986/32503. Rome & Washington D.C.: Food and Agriculture Organization (FAO) and the World Bank.

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

Gibson, J. (2020). Measuring chronic hunger from diet snapshots. *Economic Development and Cultural Change*, *68*(3), 813-838.

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*(2), 473-489.

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

Grootaert, C. (1986). Measuring and analyzing levels of living in developing countries: An annotated questionnaire. *Living Standards Measurement Study (LSMS) Working Paper*, LSMS 24, Washington D.C.: The World Bank.

Hirvonen, K., Taffesse, A. S., & Worku, I. (2016). Seasonality and household diets in Ethiopia. *Public Health Nutrition*, 19(10), 1723-1730.

Hoddinott, J., & Yohannes, Y. (2002). Dietary diversity as a food security indicator. IFPRI-FCND discussion paper, 136, Washington, DC: International Food Policy Research Institute (IFPRI).

Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica, 46*(1), 33-50.

Mahalanobis, P., & Sen, S. (1953). On some aspects of the Indian national sample survey. *Bulletin de l'Institut International de Statistique, 34*(2), 5-14.

McKenzie, D., & Rosenzweig, M. (2012). Preface for symposium on measurement and survey design. *Journal of Development Economics, 1*(98), 1-2.

Meager, R. (2020). "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature," *Working Paper*, London School of Economics.

Melesse, M. B., van den Berg, M., de Brauw, A., & Abate, G. T. (2019). Understanding urban consumers' food choice behavior in Ethiopia: Promoting demand for healthy foods. IFPRI-ESSP working paper, 131, Washington D.C.: International Food Policy Research Institute (IFPRI).

Morwitz, V. G. (1997). It seems like only yesterday: The nature and consequences of telescoping errors in marketing research. *Journal of Consumer Psychology, 6*(1), 1-29.

Neter, J., & Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association, 59*(305), 18-55.

Pauw, K., Thurlow, J., Ecker, O., 2018. Micro-econometric and Micro-Macro Linked Models: Modeling Agricultural Growth and Nutrition Linkages: Lessons from Tanzania and Malawi, in: Henning, C., Badiane, O., Krampe, E. (Eds.), *Development Policies and Policy Processes in Africa*. Springer Open, Berlin, pp. 97-115.

Pradhan, M. (2009). Welfare analysis with a proxy consumption measure: evidence from a repeated experiment in Indonesia. *Fiscal Studies*, *30*(3-4), 391-417.

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

Sharp, M., Buffiere, B., Himelein, K., & Gibson, J. (2019). Effects of Data Collection Methods on Estimated Household Consumption and Poverty, and on Survey Costs: Evidence from an Experiment in the Marshall Islands. Paper Presented at the IARIW-World Bank Conference on New Approaches to Defining and Measuring Poverty in a Growing World, Washington DC, November.

Swindale, A., & Bilinsky, P. (2006). Household dietary diversity score (HDDS) for measurement of household food access: indicator guide. Washington, DC: Food and Nutrition Technical Assistance Project, Academy for Educational Development, Washington D.C.: FANTA FHI 360.

USDA. (2013). *National Nutrient Database for Standard Reference, Release 28*.

van Garderen, K., & Shah, C. (2002). Exact interpretation of dummy variables in semilogarithmic equations. *The Econometrics Journal, 5*(1), 149-159.

White, H. (1980). A Heteroskedasticity-Consistent Covariance-Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica, 48*(4), 817-838.

WFP. (2008). Food consumption analysis: Calculation and use of the food consumption score in food security analysis, Rome: World Food Programme (WFP), Vulnerability Analysis and Mapping Branch (ODAV).

Zezza, A., Carletto, C., Fiedler, J. L., Gennari, P., & Jolliffe, D. (2017). Food counts. Measuring food consumption and expenditures in household consumption and expenditure surveys (HCES). Introduction to the special issue. *Food Policy, 72*, 1-6.

Tables and figures

Table 1. Survey experiment design

	Unbounded recall	**Bounded recall**
Method of data capture	Computer-assisted personal interviewing (CAPI)	CAPI
Reference period	7-day recall	7-day recall
Designated respondent	Household member who decides on food purchase and/or preparation	Household member who decides on food purchase and/or preparation
Consumption measurement	128 food items (frequency and quantity consumed)	128 food items (frequency and quantity consumed)
Question format	Consumption of food item **during the last 7 days**	Consumption of food item **since the recent visit by a survey team member***
Number of households	450	440

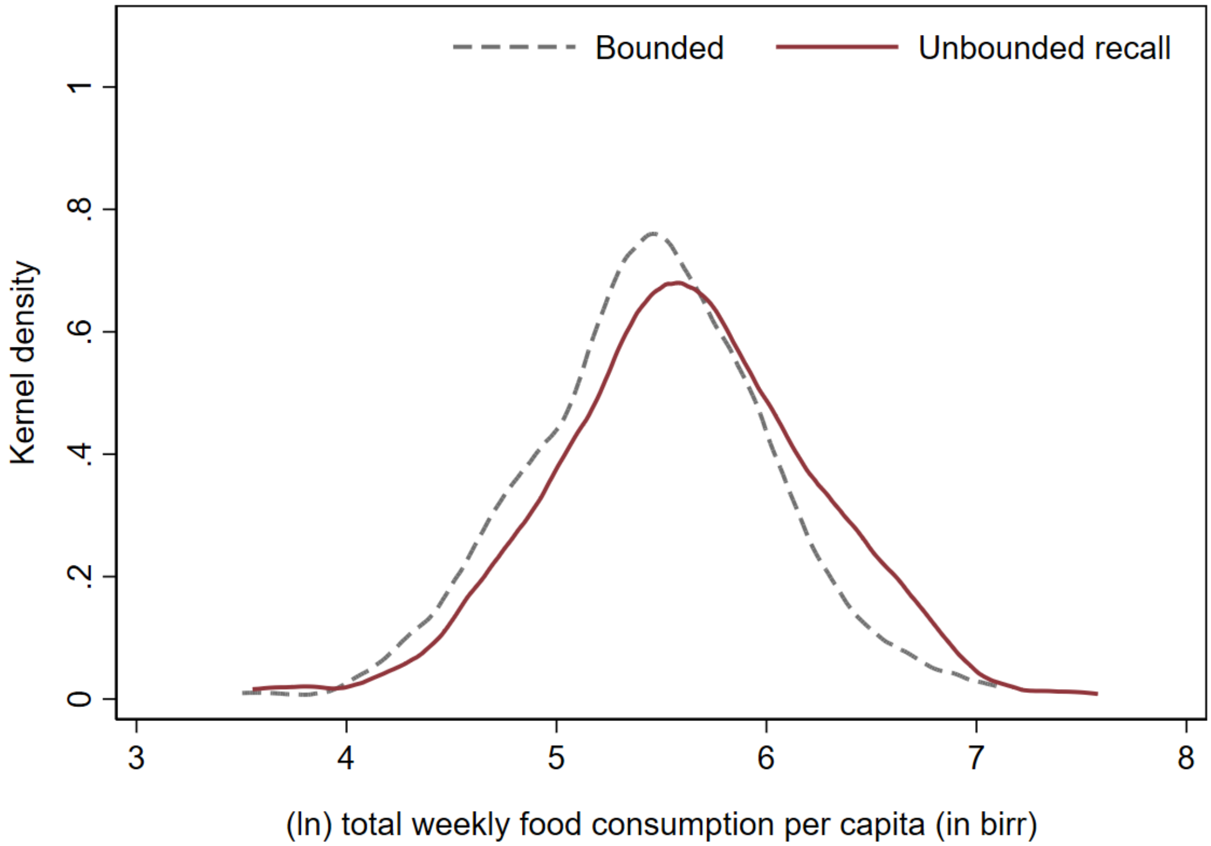
Note: * Households in the bounded recall group were visited by survey supervisors exactly 7 days before the actual data collection. The supervisors wore a uniform (a white t-shirt and hat) supplied by the research team while visiting households in the bounded recall group so as to differentiate themselves from other visitors. The enumerators were trained to specifically remind and confirm the visit by a survey team member with a white t-shirt and hat just before administering the food consumption module for households in the bounded treatment group.

Table 2. Household characteristics, by recall type

Variable	Bounded recall Mean/[SE]	Unbounded recall Mean/[SE]	Difference	t-test p-value
Female respondent	0.925	0.911	0.014	0.443
	[0.015]	[0.015]		
Household size	4.566	4.549	0.017	0.907
	[0.126]	[0.109]		
Household size in adult equivalent units	3.881	3.837	0.044	0.730
	[0.109]	[0.094]		
Male-headed household	0.566	0.551	0.015	0.696
	[0.036]	[0.029]		
Head's education in years	6.345	6.491	-0.146	0.548
	[0.275]	[0.278]		
Household asset index	-0.130	0.103	-0.233	0.193
	[0.168]	[0.143]		
Other treatment: Control	0.305	0.353	-0.048	0.096
	[0.015]	[0.016]		
Other treatment: Video	0.343	0.331	0.012	0.632
	[0.014]	[0.014]		
Other treatment: Video+	0.352	0.316	0.036	0.264
	[0.018]	[0.018]		
Weekly food consumption per capita in	311.281	323.739	-12.458	0.313
before the experiment (in September 2019)	[10.096]	[13.650]		
Number of households:	440	450		
Clusters:	40			

Note: Unit of observation is household. Standard errors (SE) are clustered at enumeration area level. Difference in means between the groups tested with a t-test (null-hypothesis: difference in means = 0).

Figure 1. Distribution of (ln) weekly food consumption per capita (in birr), by recall type



Note: Kernel density estimates. N = 890 households.

Table 3. Impact of unbounded recall on weekly per capita food consumption (in birr)

	(1)	(2)	(3)	(4)
Dependent variable:	**food consumption (birr)**		**(ln) food consumption (birr)**	
Unbounded recall	56.13**	54.41***	0.156***	0.152***
	(15.88)	(14.31)	(0.044)	(0.039)
Household level controls?	No	Yes	No	Yes
Sub-city fixed effects?	No	Yes	No	Yes
Observations:	890	890	890	890
R^2	0.020	0.158	0.017	0.196
Bounded group mean of the dependent variable	275.3	275.3	n/a	n/a

Note: Unit of observation is household. Household level controls include household size (number of members), indicator variable for male-headed households, head's education in years, and treatment status in the video experiment. Standard errors are clustered at the enumeration area level and reported in parentheses. Statistical significance denoted with + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001.

Table 4. Impact of unbounded recall on (log) daily per capita calorie and protein intakes

	(1)	(2)	(3)	(4)
Dependent variable:	**(ln) calorie consumption**		**(ln) protein consumption**	
Unbounded recall	0.081*	0.084**	0.147***	0.148***
	(0.031)	(0.026)	(0.040)	(0.035)
Household level controls?	No	Yes	No	Yes
Sub-city fixed effects?	No	Yes	No	Yes
Observations	890	890	890	890
R^2	0.008	0.189	0.017	0.159

Note: Unit of observation is household. Household level controls include household size (number of members), indicator variable for male-headed households, head's education in years, and treatment status in the video experiment. Standard errors are clustered at the enumeration area level and reported in parentheses. Statistical significance denoted with + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5. Impact of unbounded recall on household diet diversity (HDDS) and food consumption scores (FCS)

	(1)	(2)	(3)	(4)
Dependent variable:	**household diet diversity score**		**food consumption score**	
Unbounded recall	0.299**	0.290**	4.371**	4.275**
	(0.089)	(0.089)	(1.338)	(1.313)
Household level controls?	No	Yes	No	Yes
Sub-city fixed effects?	No	Yes	No	Yes
Observations	890	890	890	890
R^2	0.009	0.202	0.011	0.176
Bounded group mean of the dependent variable	9.12	9.12	65.89	65.89

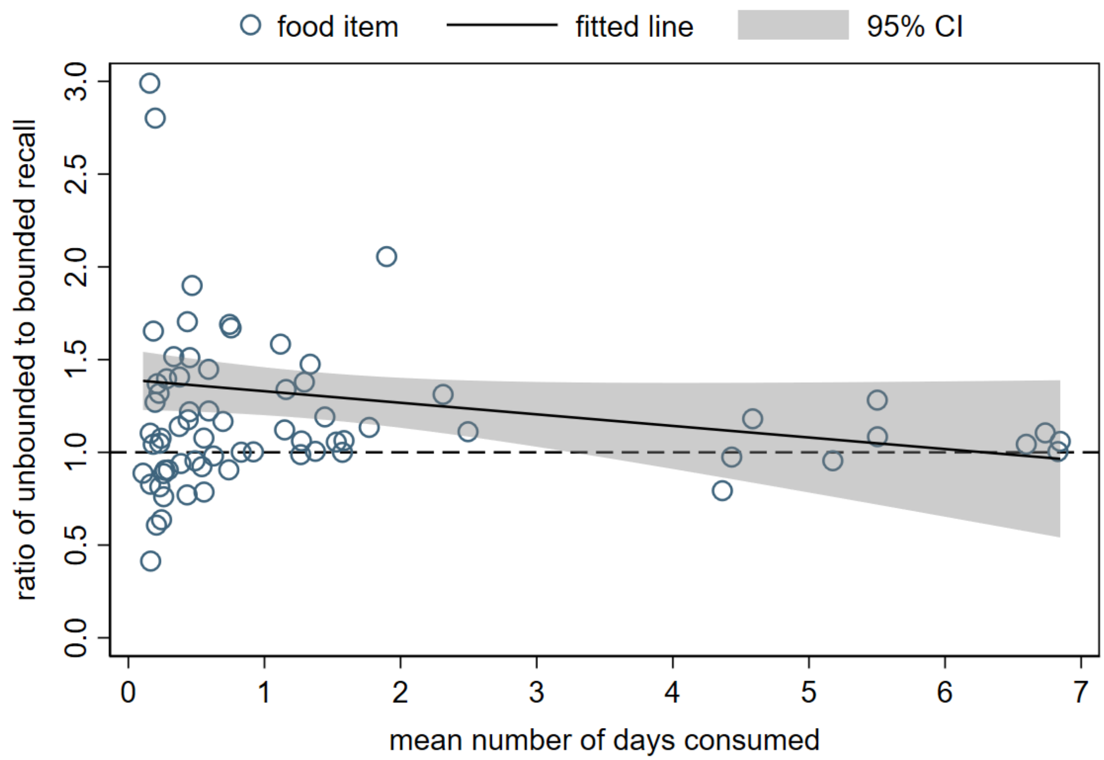
Note: Unit of observation is household. Household level controls include household size (number of members), indicator variable for male-headed households, head's education in years, and treatment status in the video experiment. Standard errors are clustered at the enumeration area level and reported in parentheses. Statistical significance denoted with + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001.

Table 6. Impact of unbounded recall on consumption of different food groups

	(1) Staples	(2) Legumes	(3) Vegetables	(4) Fruit	(5) Meat & eggs	(6) Dairy	(7) Other
Panel A. Dependent variable: =1 if consumed from the food group, zero otherwise							
Unbounded recall	n/a	n/a	n/a	0.011	0.082**	0.034	n/a
				(0.028)	(0.029)	(0.028)	
*R*2	n/a	n/a	n/a	0.089	0.120	0.103	n/a
Bounded group mean of the dependent variable	1.00	0.99	1.00	0.79	0.69	0.55	1.00
Panel B. Dependent variable: Number of days consumed from the food group							
Unbounded recall	n/a	0.046	n/a	0.322*	0.750***	0.197	n/a
		(0.159)		(0.154)	(0.147)	(0.167)	
*R*2	n/a	0.042	n/a	0.125	0.173	0.115	n/a
Bounded group mean of the dependent variable	6.96	5.54	6.96	3.49	2.38	2.18	6.99
Panel C. Dependent variable: birr value consumed from the food group							
Unbounded recall	31.30**	0.76	9.82	3.64	110.71**	5.73	16.93+
	(9.70)	(4.04)	(8.89)	(6.68)	(40.24)	(6.82)	(9.16)
*R*2	0.304	0.112	0.190	0.127	0.096	0.095	0.124
Bounded group mean of the dependent variable	304.8	97.3	197.1	86.5	272.3	53.2	183.1

Note: Unit of observation is household; N = 890. All regressions include household level controls (household size, indicator variable for male-headed households, head's education in years, and treatment status in the video experiment) and sub-city fixed effects. Standard errors are clustered at the enumeration area level and reported in parentheses. Statistical significance denoted with + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001.

Figure 2. Consumption frequency and telescoping error



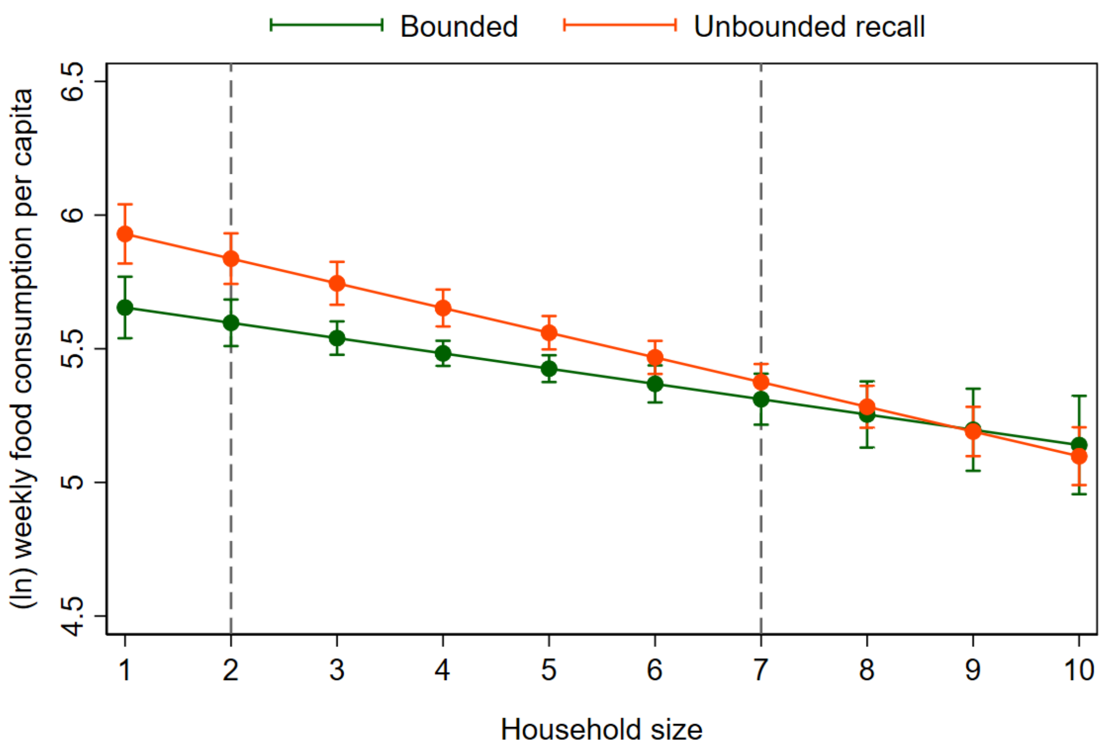
Note: N = 70 food items; food items consumed by less than 5 percent of the households were dropped. The vertical axis measures the mean weekly per capita food consumption in the unbounded sample relative to the bounded sample. These two means are equal when the ratio equals one, marked by the dashed horizontal line. The horizontal axis measures the mean number of days the item was consumed by all households. The fitted line (solid black line) is based on a weighted linear regression that puts more weight on more expensive food items. The shaded area around the fitted line is the 95% confidence interval (CI).

Table 7. Regression results from interaction models

	(1)	(2)	(3)
Dependent variable:	**Treatment (Unbounded recall=1)**		
Interacted variable:	**Household Size**	**Male-headed household (binary)**	**Head's education in years**
Unbounded recall	0.310**	0.171**	0.181**
	(0.088)	(0.051)	(0.058)
Interaction term	-0.035*	-0.038	-0.005
	(0.017)	(0.064)	(0.006)
Household level controls?	Yes	Yes	Yes
Sub-city fixed effects?	Yes	Yes	Yes
Observations:	890	890	890
R^2	0.196	0.193	0.193

Note: Unit of observation is household. Household level controls include household size (number of members), indicator variable for male-headed households, head's education in years, and treatment status in the video experiment. Standard errors are clustered at the enumeration area level and reported in parentheses. Statistical significance denoted with + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001.

Figure 3. Predictive margins between households in bounded and unbounded recall groups, by household size



Note: Predictive margins for bounded and unbounded recalls across household size. The vertical capped lines show the 95% confidence intervals for the point estimates. The grey vertical dashed lines represent the 10th and 90th percentile of the household size distribution.

Online appendices

Appendix A. Alternative Specifications

Table A1. Replicating Column 4 in Table 3 but including the five households with implausibly high consumption values and using non-winsorized consumption data

	(1)	(2)
Dependent variable:	**(ln) food consumption (birr)**	
Unbounded recall	0.143**	0.155**
	(0.041)	(0.045)
Household level controls?	Yes	Yes
Sub-city fixed effects?	Yes	Yes
Observations	895	890
R^2	0.207	0.293

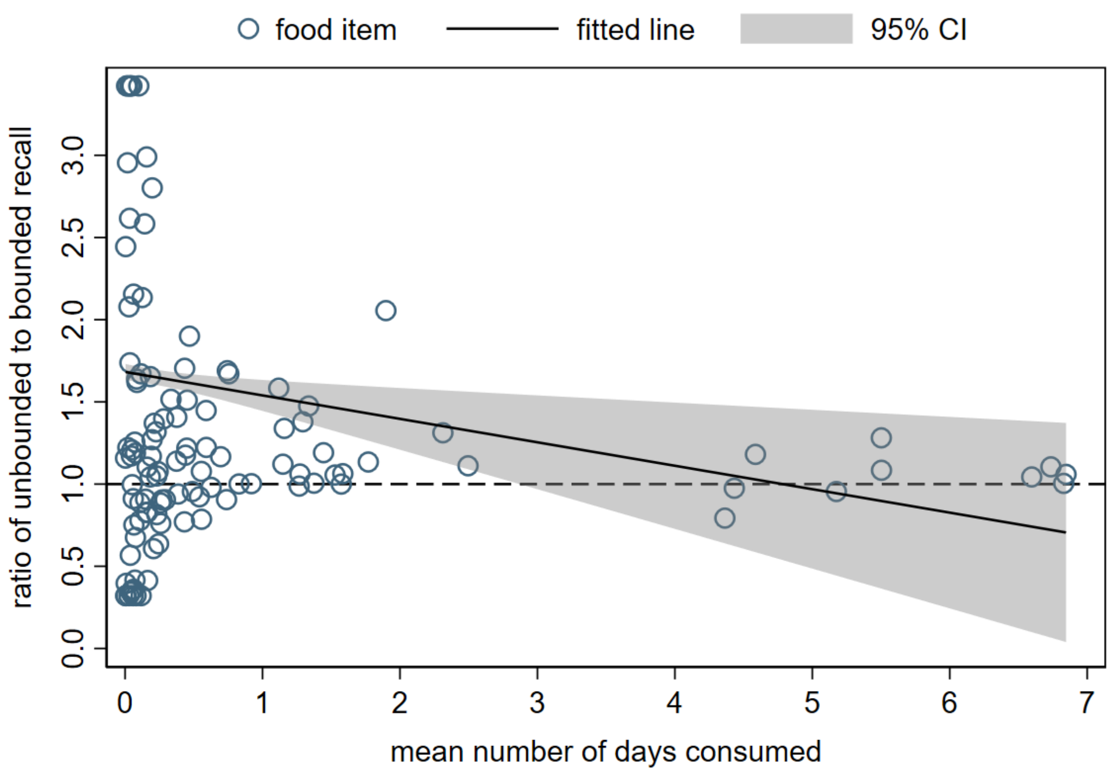
Note: Column (1) includes 5 households with implausibly high food consumption values. Column (2) is based on non-winsorized food consumption values. Unit of observation is household. Household level controls include household size (number of members), indicator variable for male-headed households, head's education in years, and treatment status in the video experiment. Standard errors are clustered at the enumeration area level and reported in parentheses. Statistical significance denoted with + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A2. Replicating Column 4 in Table 3 using quantile regressions

	(1)	(2)	(3)	(4)
quintile:	***25th***	***50th***	***75^{th}***	***25th vs 75th***
Unbounded recall	0.120*	0.145**	0.192***	0.072
	(0.051)	(0.045)	(0.051)	(0.061)
Household level controls?	Yes	Yes	Yes	Yes
Sub-city fixed effects?	Yes	Yes	Yes	Yes
Observations	890	890	890	890

Note: Unit of observation is household. Dependent variable is log household per capita food expenditure (in birr) over a 7-day period. Columns (1) to (3) are estimated using a quantile regression; column (4) using an inter-quantile regression. Bootstrapped standard errors (500 repetitions) are reported in parentheses. Household level controls include household size (number of members), indicator variable for male-headed households, head's education in years, and treatment status in the video experiment. Standard errors are clustered at the enumeration area level and reported in parentheses. Statistical significance denoted with + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure A1. Replicating Figure 2 but using a larger set of food items and winsorizing the ratio of unbounded recall to bounded recall value



Note: N = 113 food items; food items that were not consumed by both groups were dropped. The vertical axis measures the mean weekly per capita food consumption in the unbounded sample relative to the bounded sample. These two means are equal when the ratio equals one, marked by the dashed horizontal line. The horizontal axis measures the mean number of days the item was consumed by all households. The fitted line (solid black line) is based on a weighted linear regression that puts more weight on more expensive food items. The shaded area around the fitted line is the 95% confidence interval (CI).

Appendix B. Benford's Law Analysis

As noted in the text, we might be concerned that the bounded recall and unbounded recall groups exhibit different degrees of compliance with the survey which could affect their responses related to food consumption. Differences in compliance may influence the estimated difference between the two groups in their reported food consumption, providing an alternative to the telescoping error explanation for why reported food consumption of the two groups differs. Some support for the idea that visiting the same household repeatedly affects survey compliance comes from Schündeln (2018), who argued that in data collection in Ghana, each successive visit to the same household (for up to ten visits in one month) reduced data quality. The evidence for this effect used the law coined after Benford (1938), according to which the distribution of first-digits in many numerical data sets are approximately distributed based on the following probability (P):

$$P(d) = \log_{10}(d+1) - \log_{10}(d),$$

where $d \in \{1,...,9\}$ refers to the first-digit of the observation. Benford's law has been used to detect fraud or data fabrication for example in accounting, survey data and scientific research (e.g. Stein, 1975; Nigrini, 1996; Judge & Schechter, 2009). A particularly useful property is that Benford's law is scale invariant, so in our case it holds irrespective of the units in which the food consumption data were measured. Table B1 reports the observed first-digit distributions in our data and compares them to the distribution predicted by Benford's law.[21] In both cases, we can reject the null-hypothesis of observed distributions following Benford's law.[22] However, like Schündeln (2018), we are less interested in absolute differences from the predicted distribution. Instead, our focus is on relative differences, to see if data quality differs significantly between the two recall groups in a way that would be consistent with diminished compliance amongst the two-visit bounded recall group, where this diminished compliance might account for their lower reported food consumption. Following Schündeln (2018), we compute normalized Euclidean distances between the observed first-digit distribution and the one predicted by Benford's law.[23] We first

[21] These distributions we calculated using a user-written Stata routine devised by Jann (2007).

[22] We need not necessarily be concerned about the rejection of Benford's law in the individual food consumption variables; Kaiser (2019) finds that individual level outcomes deviate substantially from Benford's law in several well used data sets, but aggregate income at the household level

[23] The Euclidian distance is calculated as the square root of the sum of squared differences between the observed percentage and the percentage predicted by the Benford's law. We further normalize the calculated distances by taking a Z-score: subtracting the mean distance and dividing this by the standard deviation calculated using the pooled data.

calculate these distances separately for each food item and for both recall groups. We then test whether the food item specific average Euclidean distances for the two recall group are statistically different by regressing the mean distance on our binary treatment variable. Table B2 shows that coefficient on the treatment variable is insignificant irrespective of whether we use all food items (column 1) or restrict the data to more frequently consumed items (column 3). The result holds if we control for food item fixed effects (columns 2 and 4). Consequently, we cannot reject the hypothesis that the reports provided by the two recall groups reflect similar compliance (or that there is no differential degree of data fabrication between the two groups).

Table B1. Predicted and observed first-digit distribution, Bounded and Unbounded Recall Groups

First digit	Prevalence predicted by Benford's law	Prevalence observed in the bounded recall group	Prevalence observed in the unbounded recall group
1	30.10	32.10	31.37
2	17.61	21.33	21.12
3	12.49	14.10	13.74
4	9.69	7.96	8.14
5	7.92	6.43	7.00
6	6.70	6.97	7.82
7	5.80	4.65	4.49
8	5.12	3.52	3.40
9	4.58	2.94	2.93
Observations	**N/A**	**11,947**	**11,346**

Table B2. Testing differences in Euclidean distance to the distribution predicted by Benford's law

	(1)	(2)	(3)	(4)
Sample:	**All food items**		**Infrequently consumed food items omitted**	
Unbounded recall	-0.064	-0.064	0.051	0.033
	(0.133)	(0.064)	(0.099)	(0.037)
Food item fixed effects?	No	Yes (N=113)	No	Yes (N=75)
Observations:	226	226	150	150

Note: Dependent variable is Euclidean distance to the distribution predicted by Benford's law. Unit of observation is food item (one for each recall group). Coefficients measure Z-scores. Columns 1 and 2 are based on all food items that households in both recall groups reported to have consumed. Columns 3 and 4 are based on food items consumed by at least 20 households in each recall group. Heteroskedasticity adjusted standard errors are reported in parentheses. Statistical significance denoted with + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

References for Appendix B

Benford, F. (1938). The law of anomalous numbers. *Proceedings Of the American Philosophical Society*, 551-572.

Jann, B. (2007). DIGDIS: Stata module to analyze the distribution of digits.

Judge, G., & Schechter, L. (2009). Detecting problems in survey data using Benford's Law. *Journal of Human Resources, 44*(1), 1-24.

Kaiser, M. (2019). Benford's Law as an indicator of survey reliability—Can we trust our data? *Journal of Economic Surveys,* 33(5), 1602-1618.

Nigrini, M. J. (1996). A taxpayer compliance application of Benford's law. *The Journal of the American Taxation Association, 18*(1), 72.

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

ALL IFPRI DISCUSSION PAPERS

All discussion papers are available **here**

They can be downloaded free of charge

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
www.ifpri.org

IFPRI HEADQUARTERS
1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org