Telescoping Error in Recalled Food Consumption: Evidence from a Survey Experiment in Ethiopia

Gashaw T. Abate, Alan de Brauw, John Gibson, Kalle Hirvonen, and Abdulazize Wolle

Abstract

Telescoping errors occur if survey respondents misdate events from outside the reference period and include them 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 bound food consumption recall. This design fell out of favor although not for evidence-based reasons. To measure the extent of telescoping bias on food consumption measures, a survey experiment was conducted in Addis Ababa, Ethiopia, randomly assigning households to either a two-visit bounded recall or a single visit unbounded recall. The average value of reported food consumption is 16 percent higher (95 percent CI: 7.4–25.9) in the unbounded single visit recall relative to the two-visit bounded recall. Most of the error is explained by differences in reported spending on less frequently consumed, protein-rich foods, so apparent food security indicators based on household diet diversity are likely overstated with unbounded recall.

JEL classification: C81, D12, I32

Keywords: diet quality, food consumption, household surveys, recall, telescoping
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 the general understanding of errors in the survey measures of food consumption remains incomplete, bringing into question how progress towards the SDGs can be measured accurately.

Recently, this understanding has improved due to a series of survey experiments; key findings are described by De Weerdt, Gibson, and Beegle (2020). 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. As survey designs exist that can reduce the occurrence of these errors, such designs may be quite important to measuring indicators of progress toward 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 misdating 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 weighted. 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. What Deaton (1997, 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.”

However, the two-visit format fell out of favor when Vietnam abandoned it after their 1998 survey, and other countries followed suit. Thus, by 2000, Deaton and Grosh (2000, 114) noted (in the LSMS books entitled Designing Household Surveys for Developing Countries) the two-visit structure was being used less frequently. Going back to using unbounded recall was not the result of evidence of either the unimportance of telescoping or of the failure the two-visit format. Instead, it reflected 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. Without any firm evidence that the two-visit format helped, it was easy to jettison the method for something simpler and less costly.

Yet researchers continue to speculate that telescoping errors affect patterns found in consumption survey data. For example, in a survey experiment in Tanzania, Beegle et al. (2012) find unbounded 7-day recall yields higher estimates of consumption and lower poverty rates than 14-day recall with the same list of items and comes closer to matching their benchmark from a highly supervised 14-day 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). To provide more evidence on telescoping, Beegle et al. (2012) suggest that future survey experiments could compare bounded and unbounded seven-day recall.

A further analysis of the experiment in Tanzania found that macro- and micro-nutrient intakes measured from the seven-day recall most closely match those derived from the benchmark individual diaries,

---

1 Special issues of journals that focus on survey measurement and survey experiments are introduced by McKenzie and Rosenzweig (2012) and Zezza et al. (2017).
2 Using two visits also let very long LSMS interviews be broken into two more reasonable blocks of time.
3 Early LSMS surveys with the two-visit LSMS format include those for Côte d’Ivoire, Ghana, Pakistan, and Vietnam. However, this approach was not used in Jamaica, Nepal, South Africa, and several other countries.
even as the seven-day recall is sensitive to having errors that vary with household characteristics (Ameye, De Weerdt, and Gibson 2021). These authors suggest that seven-day recall may have offsetting errors—including those from telescoping—that roughly balance, and they 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, seven-day unbounded recall indicated that food consumption was 28 percent higher than what a seven-day diary showed; the authors speculate that telescoping could cause the seven-day recall to overstate the true value of consumption (Backiny-Yetna, Steele, and Djima 2017).

This paper provides evidence on telescoping from a survey experiment in Addis Ababa, Ethiopia, that randomly assigned either two-visit bounded recall or single-visit unbounded recall to surveyed households. In the two-visit format, a survey supervisor visited the household prior to the actual survey. During this first visit, the supervisor only informed the household that an enumerator would visit the household exactly seven days later. No data were collected during this visit, nor did the supervisor prime households to start thinking about their food consumption during the next seven days. The purpose of the supervisor visit was to introduce a salient recall marker. In the second visit, respondents were asked to recall consumption of food items since the visit by the supervisor. The sample was part of an endline study 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 three to four months prior to the present experiment. The unbounded and bounded recall groups were balanced, not only on household characteristics, but also on their food consumption in the previous survey.

The results indicate that the value of food consumption is 16 percent (95 percent CI: 7.4–25.9) higher among the group of households to whom the unbounded recall was administered, relative to the bounded recall group. In effect, on average and relative to the bounded recall group, an entire extra day of consumption is included in the report for the previous seven days. This difference between the two recall groups is not evenly distributed. It is particularly prominent for protein-rich foods like meat and eggs that are typically less frequently consumed. As a result, there are also implications for standard indicators of household food security and diet quality derived from consumption survey data, as these indicators may be overstated when unbounded recall is used.

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 abstract constructs like the “usual month,” in which respondents are asked to recall how many months per year purchases are made for each type of food, how often those purchases are made per month, and the typical spending per occasion (with similar questions for self-production). These questions substantially increase the time taken for survey interviews and add education-related inequality to reported consumption due to the cognitively demanding nature of those questions, 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—shorter than what was often used in the past (where 14-days or 1-month recall was asked). With a shorter recall period, a telescoping error will loom larger as it is amortized over a shorter period (Eisenhower, Mathiowetz, and Morganstein 2004). Finally, more diverse diets that result from 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 respondents using a rate-based estimation strategy where they multiply the frequency of occurrence by the length of the reference period (e.g., 2 loaves of bread

4 Unfortunately, it is not possible to go one step farther and measure the effect of this reporting error on measures of inequality, as the survey did not ask a non–food consumption module required to complete the consumption aggregate used in inequality measurement.
per day, ergo 14 loaves of bread for 7 days). Now, there are growing numbers of people who can afford occasional luxuries like meat and fish. These foods are still eaten sufficiently infrequently that respondents need to remember and count them rather than using rate-based estimates when answering the food recall questions. For this reason survey data may be more affected by telescoping.

The higher reported food consumption of the group surveyed with an unbounded recall can be interpreted 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. However, it should be noted that estimates in this paper cannot be benchmarked against any “true value” of food consumption. As mentioned by De Weerdt, Gibson, and Beegle (2020), it is not clear if the true value is always available in household consumption surveys. Even the “gold standard” diary-based method to collect consumption-expenditure data can be prone to measurement error unless it is highly supervised, making it prohibitively expensive in large-scale surveys (Beegle et al. 2012; Brzozowski, Crossley, and Winter 2017). At a minimum, the findings contest the earlier decision to switch from the two-visit format to unbounded recall in LSMS and other consumption surveys administered in low- and middle-income countries.

This paper proceeds as follows. Section 2 reviews the literature related to telescoping, and section 3 describes the experiment and the data used for analysis. Section 4 provides the main results along with heterogeneity and robustness analyses. Section 4 also explores alternative interpretations, namely that bounded recall leads to lower reported food consumption because of declining compliance among households. Section 4 concludes with a discussion of the feasibility of constructing adjustment factors to correct for telescoping bias. Section 5 explores cost implications for surveys using a two-visit format rather than a single-visit recall. Section 6 concludes and outlines research and policy implications of the findings.

2. Previous Work Related to Telescoping

In a two-visit survey, the first visit to the household can provide a distinct start to the recall period in the mind of the respondent. When the recall questions are asked in the second visit, the first visit then bounds the recall period for the respondent (Grootaert 1986). While the first use of bounded recall is usually attributed to Neter and Waksberg (1964), it was in quite a different context to surveys like the LSMS; the survey was of infrequent spending on alterations and repairs of dwellings, and it used 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 from being reported again. Other studies with benchmarks for assessing the extent of telescoping are also for high-value and infrequent purchases, such as 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 to gather consumption data in the first interview that can be relayed to respondents in the second interview, to ensure that 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 and Utting 1956, 389).

---

5 For a useful review of the early literature on this topic, see Dex (1995).
However, whether bounding helps respondents to remember the occurrence of episodes largely depends on the cognitive process they 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. If rate-based estimation is used to answer questions about food consumption, there is less reason to expect that a device to aid memory, such as a bounding visit, would help to improve the accuracy of these data.

Friedman et al. (2017) find that food consumption data from a seven-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. Thus, the good performance of the seven-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.

There are at least 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 et al. (2017) 6 randomized a number of different survey experiments on consumption into questionnaires in Indonesia; among other things, they found that with a seven-day unbounded recall, per capita food consumption was 24 percent 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 results in just a 2 percent drop in apparent consumption). Sharp et al. (2022) found no difference in food consumption for a Marshall Islands sample given a seven-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 seven days . . .” wording, and the gap between visits varied; visit 2 was eight days after visit 1 for 53 percent of the sample, nine or more days after for 14 percent of the sample, and seven or less days after for 33 percent of the sample. Thus, the latter experiment did not really implement a bounded recall although it does highlight the logistical challenges of this survey design.

3. The Survey Experiment, Data and Methods

The survey experiment was designed to study the implications of telescoping for food consumption measurement by systematically contrasting responses from unbounded and bounded recalls. For the unbounded recall, the common approach to food consumption measurement was used, 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). 7 For the bounded recall, a salient

---

6 This study has been presented at an international conference but, at the time of writing (May 2022), is not available as an online manuscript.

7 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. All food items consumed by at least 1 percent of the 2016-HCES households located in Addis Ababa were included on the list.
Table 1. Survey Experiment Design

<table>
<thead>
<tr>
<th></th>
<th>Bounded recall</th>
<th>Unbounded recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method of data capture</td>
<td>CAPI</td>
<td>Computer-assisted personal interviewing (CAPI)</td>
</tr>
<tr>
<td>Total number of survey modules</td>
<td>10 modules</td>
<td>10 modules</td>
</tr>
<tr>
<td>Module of order of the food</td>
<td>3rd module</td>
<td>3rd module</td>
</tr>
<tr>
<td>consumption module</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference period in the food</td>
<td>7-day recall</td>
<td>7-day recall</td>
</tr>
<tr>
<td>consumption module</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Designated respondent in the food</td>
<td>Household member who decides on food purchase and/or preparation</td>
<td>Household member who decides on food purchase and/or preparation</td>
</tr>
<tr>
<td>consumption module</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food consumption measurement</td>
<td>128 food items (frequency and quantity consumed)</td>
<td>128 food items (frequency and quantity consumed)</td>
</tr>
<tr>
<td>Question format in the food</td>
<td>Consumption of food item since the visit by a survey team member*</td>
<td>Consumption of food item during the last 7 days</td>
</tr>
<tr>
<td>consumption module</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of households</td>
<td>440</td>
<td>450</td>
</tr>
</tbody>
</table>

Source: Authors.

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.

recall marker was introduced by visiting sample households seven 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.8 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 two groups differ only by the question format (wording) 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, the number of food items in the recall list, and the total number as well as the order of survey modules.

The survey experiment was implemented as an add-on to an endline 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. To this end, the study team produced two types of videos with different information content and randomly allocated the sample of households into three groups: Control (no video screening), Video (video screening with standard informational content) and Video+ (video screening with advanced informational content). S1 in the supplementary online appendix, available with this article at the *World Bank Economic Review* website, provides more details about the video intervention, and its impact evaluation results are reported in Abate *et al.* (2021). To ensure that the experiment on bounded versus unbounded recall did not affect the outcomes of the impact evaluation (and vice versa), the study cross-randomized study samples into the bounded and unbounded consumption recall subgroups.

The study sample is representative of households in Addis Ababa and is formed from 930 households randomly selected from six sub-cities, 20 woredas (districts), and 40 ketenas (neighborhoods; or clusters of households) within Addis Ababa.9 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

8 For 99 percent of the 440 households in the bounded recall group, the duration between the two visits is recorded as exactly 7 days. For the remaining 1 percent (6 households), dates of the visits were not recorded.

9 Melesse *et al.* (2019) provide a detailed description of the sampling strategy. Hirvonen, Abate, and de Brauw (2020) show that the household demographics in this sample are comparable to other representative surveys conducted in Addis Ababa.
The interviews for both recall groups took place in parallel. Out of the households that had been interviewed in September 2019, 35 households were not interviewed in January–February. The response rates are comparable across the two groups: 97 percent and 95 percent for the unbounded and bounded subgroups, respectively.

To choose the primary respondent, the survey module on consumption targeted the household member who was most knowledgeable about the household’s food shopping and preparation. More than 90 percent of the respondents were women. For each item, the study asked the respondent whether their household had consumed the item in the past seven days (or, for the bounded group whether they had consumed it since the visit by a survey supervisor). 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.

Data

Reported quantities were converted into local currency units (Ethiopian birr) using retail price data that are collected every month by the Central Statistical Agency (CSA) of Ethiopia. The quantities were also converted consumption in kilograms into calorie and protein equivalents using conversion factors provided in the Ethiopian food composition tables (EPHI 1981). These calculations account for the inedible portion of the weight 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. After dropping five households with implausibly large per capita consumption values, the final data set for analysis includes 890 households. 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.

As an alternative measure of consumption, a further outcome used is household dietary diversity, often used as an indicator of household food security (Hoddinott and Yohannes 2002). The household dietary diversity score (HDDS) of Swindale and Bilinsky (2006) was computed by first grouping the 128 food items in this study’s 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. As an alternative measure of household dietary diversity, the food consumption score (FCS) developed by the WFP (2008) was computed. The FCS combines dietary diversity and consumption frequency by grouping the consumed food items into nine groups and allocating more weight to protein-rich foods.

---

10 Previous work in Ethiopia shows how food consumption in urban areas is affected by religious fasting (Hirvonen, Taffesse, and Worku 2016). To this end, the study made sure that there was no major Orthodox or Muslim fasting period during the survey experiment.
11 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.
12 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.
13 More specifically, the study used the CSA retail price data for February 2020, restricting the price observations to Addis Ababa.
14 Specifically, calorie consumption above 5,000 kcal per adult equivalent was considered implausible.
15 Three of these households were in the bounded group and two in the unbounded group.
16 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.
17 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).
The weighted FCS index ranges between 0 and 112, with higher scores indicating a better food security situation.

The bounded and unbounded groups are similar in terms of household characteristics (table 2). Further, the t(random) assignment into the video experiment study arms is orthogonal to the (random) allocation into the telescoping experiment groups. The subsamples given bounded and unbounded recall were also balanced on both household characteristics, and their baseline food consumption data were collected in September 2019 (i.e., three to four months before the survey experiment).

Methods
The difference in reported per capita consumption values across the two groups is quantified using ordinary least squares (OLS). In the most basic model, both the per capita food consumption value and its logarithm are regressed on a binary treatment variable valued 1 if the household was randomly selected into the unbounded recall group, and 0 if selected into the bounded recall group. The next regressions 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). When percentage differences derived from the coefficients in semi-log regressions are discussed, they are based on $100 \times (e^{\hat{\beta} - 0.5V(\hat{\beta})} - 1)$ with confidence intervals from the approximate unbiased variance estimator of van Garderen and Shah (2002). The standard errors in all regressions are clustered at the ketena level.

18 In the equation, $\hat{\beta}$ refers to the estimated coefficient and $\hat{V}$ to the estimated variance.
19 The household sample groups into 6 sub-cities and 40 ketenas.
**Figure 1.** Distribution of (ln) Weekly Food Consumption per Capita (in Birr), by Recall Type

Source: Authors’ calculations on Addis Ababa survey data.

Note: Kernel density estimates. N = 890 households.

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

Source: Authors’ calculations on Addis Ababa survey data.

Note: The unit of observation is the 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 (ketena) level and reported in parentheses. Statistical significance denoted with $+ p < 0.10$, $* p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

### 4. Results

As a first step, the study illustrates the full distributions of (log) household weekly per capita food consumption measured in birr for both bounded and unbounded recall groups (fig. 1). 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.

Next, regression estimates for the difference in unbounded and bounded recall for household weekly per capita food consumption, measured in birr, are computed (table 3). 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 natural 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
Table 4. Impact of Unbounded Recall on (log) Daily per Capita Calorie and Protein Intakes

<table>
<thead>
<tr>
<th>Dependent variable: (ln) calorie consumption</th>
<th>(ln) protein consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(1)</strong> (2) (3) (4)</td>
<td></td>
</tr>
<tr>
<td>Unbounded recall</td>
<td></td>
</tr>
<tr>
<td>0.081* (0.031)</td>
<td>0.084** (0.026)</td>
</tr>
<tr>
<td>Household level controls? No Yes</td>
<td></td>
</tr>
<tr>
<td>Sub-city fixed effects? No Yes</td>
<td></td>
</tr>
<tr>
<td>Observations 890</td>
<td>890</td>
</tr>
<tr>
<td>R² 0.008</td>
<td>0.189</td>
</tr>
<tr>
<td>Source: Authors' calculations on Addis Ababa survey data.</td>
<td></td>
</tr>
</tbody>
</table>
| Note: The unit of observation is the 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 (ketena) level and reported in parentheses. Statistical significance denoted with + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001.

and adjusted regressions are negligible, the study focuses its reporting and discussion on the adjusted regression results.

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 percent 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 percent 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 seven-day consumption total. The most plausible source of this apparently higher rate of food consumption is forward telescoping where consumption episodes occurring more than seven days earlier are included in the report for the last seven days.

Next, the analysis in table 3 is repeated with per capita calorie and protein intakes as the dependent variables, in columns (1)–(2) and 3–4, respectively (table 4). Unbounded recall results in 8.8 percent higher reported per capita calorie intakes compared to bounded recall (p-value < 0.01; 95 percent 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 percent CI: 7.91–24.5), which is similar to the estimated impact on the birr value reported in table 3. The fact 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 (2021), who note that previous findings on the fragility of calorie-based hunger estimates to variations in the design of food consumption surveys (e.g., De Weerdt et al. 2016) are likely to underestimate the fragility of survey estimates when a richer consideration of nutrition is used, one that focuses on macro- and micro-nutrients.

Next, the influence of the type of recall survey on the two indicators of food security and diet quality, HDDS and FCS, is examined (table 5). 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 percent CI: 0.11–0.47). Considering that the mean HDDS in the bounded group is 9.1 food groups, this estimate 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 percent CI: 1.62–6.93) or 6 percent higher than in the bounded recall group.

To further consider potential differences by food group, the following analysis uses consumption of specific food groups as dependent variables (table 6). In panel A, the dependent variable takes on a value of 1 if the household consumed from the food group and 0 otherwise. Households in the unbounded
Table 5. Impact of Unbounded Recall on Household Diet Diversity (HDDS) and Food Consumption Scores (FCS)

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>(1) Household diet diversity score</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbounded recall</td>
<td>0.299**</td>
<td>0.290**</td>
<td>4.371**</td>
<td>4.275**</td>
<td></td>
</tr>
<tr>
<td>(0.089)</td>
<td>(0.089)</td>
<td>(1.338)</td>
<td>(1.313)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household level controls?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Sub-city fixed effects?</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>890</td>
<td>890</td>
<td>890</td>
<td>890</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.009</td>
<td>0.202</td>
<td>0.011</td>
<td>0.176</td>
<td></td>
</tr>
<tr>
<td>Bounded group mean of the dependent variable</td>
<td>9.12</td>
<td>9.12</td>
<td>65.89</td>
<td>65.89</td>
<td></td>
</tr>
</tbody>
</table>

Source: Authors’ calculations on Addis Ababa survey data.
Note: The unit of observation is the 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 (ketena) 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

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A. Dependent variable: = 1 if consumed from the food group, 0 otherwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unbounded recall</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>0.011</td>
<td>0.082**</td>
<td>0.034</td>
<td>n/a</td>
</tr>
<tr>
<td>R²</td>
<td>(n/a)</td>
<td>(0.028)</td>
<td>(0.029)</td>
<td>(0.028)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bounded group mean of the dependent variable</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>0.79</td>
<td>0.120</td>
<td>0.103</td>
<td>n/a</td>
</tr>
<tr>
<td>Panel B. Dependent variable: Number of days consumed from the food group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unbounded recall</td>
<td>n/a</td>
<td>0.046</td>
<td>n/a</td>
<td>0.322**</td>
<td>0.750***</td>
<td>0.197</td>
<td>n/a</td>
</tr>
<tr>
<td>R²</td>
<td>(n/a)</td>
<td>(0.159)</td>
<td>(0.154)</td>
<td>(0.147)</td>
<td>(0.167)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bounded group mean of the dependent variable</td>
<td>6.96</td>
<td>5.54</td>
<td>6.96</td>
<td>3.49</td>
<td>2.38</td>
<td>2.18</td>
<td>6.99</td>
</tr>
<tr>
<td>Panel C. Dependent variable: Birr value consumed from the food group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unbounded recall</td>
<td>31.30**</td>
<td>0.76</td>
<td>9.82</td>
<td>3.64</td>
<td>110.71**</td>
<td>5.73</td>
<td>16.93+</td>
</tr>
<tr>
<td>R²</td>
<td>(3.70)</td>
<td>(4.04)</td>
<td>(8.89)</td>
<td>(6.68)</td>
<td>(40.24)</td>
<td>(6.82)</td>
<td>(9.16)</td>
</tr>
<tr>
<td>Bounded group mean of the dependent variable</td>
<td>304.8</td>
<td>97.3</td>
<td>197.1</td>
<td>86.5</td>
<td>272.3</td>
<td>53.2</td>
<td>183.1</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations on Addis Ababa survey data.
Note: The unit of observation is the 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 (ketena) level and reported in parentheses. Statistical significance denoted with +p < 0.10, *p < 0.05, **p < 0.01, ***p < 0.001. n/a = not applicable; insufficient amount of variation in the dependent variable.

recall group are 8 percentage points (or 11 percent) more likely to report having consumed meat, poultry, fish, or eggs in the past seven days (p < 0.01). The estimated coefficients are positive for both fruit and dairy food groups but are not statistically different from 0. All other food groups were consumed virtually by all households, and thus there is not sufficient variation in the outcome variable to estimate coefficients using this method. In panel B, the dependent variable is the number of days the household reports consuming item from the food group. The unbounded recall increases the 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 legumes and dairy food groups is not statistically different from 0. In panel C, the dependent variable is food group 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). Among sample households, these foods are typically less regularly consumed than are calorie-rich staple foods (e.g., maize, wheat, teff). Therefore, one implication of the results in tables 4–6 (given that diet diversity and quality scores also rise with consumption of the protein-rich foods) is that the telescoping effect could be 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 test the hypothesis that the telescoping effect is driven by infrequently consumed food items, the mean weekly per capita food consumption for each food item is calculated, separately for the unbounded and bounded recall groups. The two means are then used to calculate a ratio between mean per capita consumption in the unbounded group relative to the bounded group for each food item. After dropping food items that were consumed by less than 5 percent of the sample, the study is left with 70 food items. For those items, the ratio is compared in a scatter plot to the mean number of days each food item was reportedly consumed by all households in the sample (fig. 2). 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 1, indicating that for these foods the telescoping error is close to 0.

Heterogeneity by Household Characteristics
The degree of telescoping error could vary with household and respondent characteristics, if such characteristics affect the nature of the reporting task or the ability to do this task. To explore this possibility, the binary treatment variable was sequentially interacted with three control variables: a binary variable to indicate male-headed households, the head’s education in years, and household size (table 7). The coefficient on the interaction term is insignificant when the treatment variable is interacted with the binary variable capturing male-headed household (column (1)) or the variable capturing head’s education level in years (column (2)). In contrast, the interaction term is statistically significant ($p < 0.05$) and negative for household size, indicating the magnitude of the telescoping error decreases with household size (column (3)). However, this finding is partly driven by very large households containing nine or more household members. When households with nine or more members (about 6.5 percent of the total sample) are omitted from the sample (column (4)), the coefficient on the interaction gets smaller and is no longer statistically different from 0 ($p = 0.240$).

---

20 Dairy products are also protein rich and are infrequently consumed, yet the coefficient estimates on dairy products are not statistically different from 0 for any of the outcomes studied in table 6. However, one must be cautious about interpreting the lack of rejection of the null hypothesis as “no effect.” The coefficients are all positive, small relative to the average value of each outcome in the bounded group, and somewhat imprecisely estimated. It could be that the experiment simply lacked statistical power to estimate positive coefficients with $p$-values in the standard rejection range.

21 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, these food items are omitted from the sample for the scatterplot and regression presented in fig. 2.

22 Variables measuring the gender and education level of the household head come from an earlier survey conducted among the same households, which is detailed in Melesse et al. (2019). The gender of the respondent to the consumption survey was recorded, but the respondent code cannot be linked back to the household roster from previous surveys.
Figure 2. Consumption Frequency and Telescoping Error

Source: Authors’ calculations on Addis Ababa survey data.
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 1, 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 percent confidence interval (CI).

Table 7. Regression Results from Interaction Models

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>(1) (ln total weekly food consumption per capita)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbounded recall</td>
<td>0.171**</td>
<td>0.181**</td>
<td>0.310**</td>
<td>0.264**</td>
</tr>
<tr>
<td></td>
<td>(0.051)</td>
<td>(0.058)</td>
<td>(0.088)</td>
<td>(0.091)</td>
</tr>
<tr>
<td>Unbounded recall * Male-headed household</td>
<td>−0.038</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.064)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unbounded recall * Head’s education in years</td>
<td>−0.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unbounded recall * Household size</td>
<td></td>
<td>−0.035*</td>
<td></td>
<td>−0.023</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.017)</td>
<td></td>
<td>(0.020)</td>
</tr>
<tr>
<td>Household level controls?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sub-city fixed effects?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations:</td>
<td>890</td>
<td>890</td>
<td>890</td>
<td>832</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.193</td>
<td>0.193</td>
<td>0.196</td>
<td>0.193</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations on Addis Ababa survey data.
Note: The unit of observation is the household. The dependent variable in all columns is (ln) total weekly food consumption per capita (in birr). 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. Column (4) removes households with 9 or more members. Standard errors are clustered at the enumeration area (ketena) level and reported in parentheses. Statistical significance denoted with + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Robustness Checks
The results of several robustness checks are reported in S2 in the supplementary online appendix. First, the sensitivity of the main estimates to outliers is tested, 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 S2.1 in the supplementary
online appendix). Second, column 2 in table S2.2 shows that the results are similar if median regression based on the least absolute deviation procedure is used; the least absolute deviation procedure is less sensitive to outliers or other extreme values than OLS (Koenker and Bassett 1978). 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 0 (table S2.2). Third, the survey experiment was carried out during an endline evaluation of a video intervention that encouraged households to consume more fruit and vegetables. While regressions do control for treatment status in the video intervention, it is possible that the video treatment amplified the effect of (un)bounding in the survey experiment. To explore this possibility, the treatment status in the video intervention is interacted with the binary variable representing membership in the unbounded recall group. Whether log food expenditures, calorie consumption, or protein consumption is used as the dependent variable, none of the coefficient estimates on the interaction terms are statistically different from 0, and they are all relatively small and imprecisely estimated, indicating that the video experiment does not influence the findings (table S2.3). Fourth, table S2.4 replicates table 3 using an ANCOVA estimator, by adding household per capita consumption measured in September 2019 as an additional control variable. The coefficients are very similar to those reported in table 3.

Finally, two robustness checks are conducted on consumption frequency and telescoping error (fig. 2). First, the regression line in the figure is re-estimated without the weight on more expensive food items (fig. S2.1). The slope of the unweighted regression line remains negative but is slightly less steep than the one in fig. 2. Second, 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, fig. S2.2 replicates fig. 2 based on 113 food items, winsorized at the 5th and 95th percentiles. As before, the telescoping error is observed as driven by the infrequently consumed food items.

**Alternative Interpretations**

This study interprets the results as being the result of forward telescoping. The higher reported food consumption of the group surveyed with an unbounded recall is in line with concerns about the cognitive burden of the memory task imposed on survey respondents. Therefore, any features of the 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.

However, this study acknowledges 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 seven 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.

This alternative explanation, however, 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

---

23 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 it cannot be assumed that individual observations would not change ranking due to the treatment (e.g., Meager 2022).

24 Fifteen food items that were neither consumed by unbounded nor bounded recall groups were dropped.
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 there appears to be no such effect of position in the food list.25 Fourth, when the benchmark is used for assessing data quality—Benford’s Law—which 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 (S.3 in the supplementary online appendix; regression results are in tables S3.1 and S3.2). 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 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 seven days of food consumption. Therefore, an analyst might be tempted to annualize this estimate by multiplying by (365/8) rather than by (365/7) to make up for the overstatement that results from telescoping.26

Evidence from other 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 in which 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 underestimated increases as consumption rises, so without data on actual consumption it is not possible to 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 et al. 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 this experiment does not provide two measures for the same household, so instead the approach of Gibson et al. (2015) is followed by taking the mean of the (log) per capita household food consumption across

---

25 The study 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 percent 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.

26 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).
each sampling unit (ketena), separately for both recall groups. The ketena level means for the unbounded recall group (the noisier measure) are then regressed on the means for the bounded recall group. The slope coefficient is 0.201 with a White (1980) adjusted 95 percent confidence interval ranging between −0.030 and 0.433. Consequently, the null hypothesis that the 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 mismeasured right-hand side variables (Abay et al. 2019). As a result, simple adjustment factors to account for telescoping errors cannot be easily devised, and the typical econometric approach for mitigating errors in variables bias, of using instrumental variables, is unlikely to be successful with mean-reverting errors (Gibson et al. 2015). Therefore, preventing these errors occurring during the data collection seems the only viable option (De Weerdt, Gibson, and Beegle 2020).

5. 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 lower-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 percent 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 this 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 shown above. The cost comparison for the two ways of implementing the recall survey in the case of the present 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 seven days prior to the actual survey to establish the recall marker in the minds of respondents. In practice, this procedure resulted in an additional seven days of field work by survey supervisors prior to the actual survey commencement, so households scheduled to be interviewed during the first week of the survey could be visited. Once the actual survey begins, the study simply adjusts the regular appointment visits by survey supervisors for them to make visits seven days prior to the actual survey for households in the bounded group. Thus, the field expenses associated with an extra seven days of field work by supervisors prior to the actual survey were the only additional costs due to bounded recall. A back-of-the-envelope calculation suggests these extra costs averaged US$3.60 per household, which is 6.5 percent higher than the cost for the unbounded recall.

In other settings, the cost of using a two-visit format to implement bounded recall may be higher. For example, in surveys of rural communities where transportation 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. 2022), 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 substantial. 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 nonstaple foods such as meat and eggs that appear to be most prone to overstated reports of consumption due to telescoping. Thus, a benefit-cost comparison of the two-visit format may be most favorable in urban areas.

6. Discussion and Conclusions

This paper reports on a survey experiment in Addis Ababa conducted with a sample of households that were asked to recall seven 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 seven-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 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. The paper finds 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 9 percent higher than among 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, all of which are of considerable importance for measuring progress towards attaining SDGs related to poverty, food insecurity, and inequality. 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 Consumer Price Index (CPI), then the error is in the welfare measure but not in the threshold used to distinguish the poor from the nonpoor. Given that, at least pre-COVID, consumption trends appear to include more luxuries in many countries, these trends could lead to an overestimation of poverty reduction. 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.27

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

27 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 to consider dietary requirements for all macro- and micronutrients (Ameye, De Weerdt, and Gibson 2021).
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 to be the most overstated in a seven-day recall relative to the effect on reported consumption of foods that are consumed more frequently (such as major staples). It is easy therefore to see that the share of protein-rich foods in the overall food budget will be overstated, and this 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 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 is generated showing that infrequently consumed foods are especially prone to telescoping, 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 seven 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 to have respondents 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.

**Data Availability**

Raw data for this article are available at [https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/Z51Y1T](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/Z51Y1T). Replication data are available at IFPRI’s dataverse: [https://dataverse.harvard.edu/dataverse/IFPRI](https://dataverse.harvard.edu/dataverse/IFPRI).

**References**


